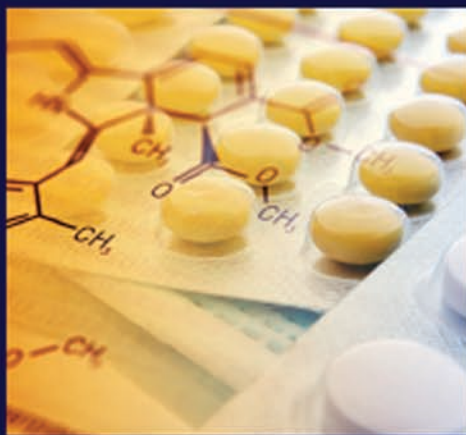


FOURTH EDITION

ENCYCLOPEDIA OF Biopharmaceutical Statistics

VOLUME 1-4



EDITED BY

Shein-Chung Chow



CRC Press
Taylor & Francis Group

Encyclopedia of
Biopharmaceutical
Statistics,
4th Edition

Volume I – IV

Encyclopedias from the Taylor & Francis Group

Print	Online	Print	Online
Agriculture		Encyclopedia of Energy Engineering and Technology, 2nd Ed., 4 Vols. Pub'd. 12/11/14	
Encyclopedia of Agricultural, Food, and Biological Engineering, 2nd Ed., 2 Vols. K10554 (978-1-4398-1111-5)	Pub'd. 10/21/10 K11382 (978-1-4398-2806-9)	K14633 (978-1-4665-0673-2)	KE16142 (978-1-4665-0674-9)
Encyclopedia of Animal Science, 2nd Ed., 2 Vols. K10463 (978-1-4398-0932-7)	Pub'd. 2/1/11 K10528 (978-0-415-80286-4)	Encyclopedia of Optical and Photonic Engineering, 2nd Ed., 5 Vols. Pub'd. 9/22/15	
Encyclopedia of Biotechnology in Agriculture and Food DK271X (978-0-8493-5027-6)	Pub'd. 7/16/10 DKE5044 (978-0-8493-5044-3)	K12323 (978-1-4398-5097-8)	K12325 (978-1-4398-5099-2)
Business and Computer Science		Environment	
Encyclopedia of Computer Science & Technology, 2nd Ed., 2 Vols. Pub'd. 12/21/2016	K21573 (978-1-4822-0819-1)	Encyclopedia of Environmental Management, 4 Vols. K11434 (978-1-4398-2927-1)	Pub'd. 12/13/12 K11440 (978-1-4398-2933-2)
Encyclopedia of Information Assurance, 4 Vols. AU6620 (978-1-4200-6620-3)	Pub'd. 12/21/10 AUE6620 (978-1-4200-6622-7)	Encyclopedia of Environmental Science and Engineering, 6th Ed., 2 Vols. Pub'd. 6/25/12	K10243 (978-1-4398-0442-1)
Encyclopedia of Information Systems and Technology, 2 Vols. K15911 (978-1-4665-6077-2)	Pub'd. 12/29/15 K21745 (978-1-4822-1432-1)	Encyclopedia of Natural Resources, 2 Vols. K12418 (978-1-4398-5258-3)	Pub'd. 7/23/14 K12420 (978-1-4398-5260-6)
Encyclopedia of Library and Information Sciences, 4th Ed. K15223 (978-1-4665-5259-3)	Pub'd. 11/13/17 K15224 (978-1-4665-5260-9)	Medicine	
Encyclopedia of Software Engineering, 2 Vols. AU5977 (978-1-4200-5977-9)	Pub'd. 11/24/10 AUE5977 (978-1-4200-5978-6)	Encyclopedia of Biomaterials and Biomedical Engineering, 2nd Ed. Pub'd. 5/28/08	H7802 (978-1-4200-7802-2)
Encyclopedia of Supply Chain Management, 2 Vols. K12842 (978-1-4398-6148-6)	Pub'd. 12/21/11 K12843 (978-1-4398-6152-3)	Encyclopedia of Biomedical Polymers and Polymeric Biomaterials, 11 Vols. Pub'd. 4/2/15	HE7803 (978-1-4200-7803-9)
Encyclopedia of U.S. Intelligence, 2 Vols. AU8957 (978-1-4200-8957-8)	Pub'd. 12/19/14 AUE8957 (978-1-4200-8958-5)	K14324 (978-1-4398-9879-6)	K14404 (978-1-4665-0179-9)
Encyclopedia of Wireless and Mobile Communications, 2nd Ed., 3 Vols. Pub'd. 12/18/12	K14731 (978-1-4665-0956-6)	Concise Encyclopedia of Biomedical Polymers and Polymeric Biomaterials, 2 Vols. Pub'd. 8/14/17	K14313 (978-1-4398-9855-0)
Chemistry, Materials and Chemical Engineering		Encyclopedia of Biopharmaceutical Statistics, 4th Ed. H100102 (978-1-4398-2245-6)	Publishing 2018 HE10326 (978-1-4398-2246-3)
Encyclopedia of Chemical Processing, 5 Vols. DK2243 (978-0-8247-5563-8)	Pub'd. 11/1/05 DKE499X (978-0-8247-5499-0)	Encyclopedia of Clinical Pharmacy DK7524 (978-0-8247-0752-1)	Pub'd. 11/14/02 DKE6080 (978-0-8247-0608-1)
Encyclopedia of Chromatography, 3rd Ed. 84593 (978-1-4200-8459-7)	Pub'd. 10/12/09 84836 (978-1-4200-8483-2)	Encyclopedia of Dietary Supplements, 2nd Ed. H100094 (978-1-4398-1928-9)	Pub'd. 6/25/10 HE10315 (978-1-4398-1929-6)
Encyclopedia of Iron, Steel, and Their Alloys, 5 Vols. K14814 (978-1-4665-1104-0)	Pub'd. 1/6/16 K14815 (978-1-4665-1105-7)	Encyclopedia of Medical Genomics and Proteomics, 2 Vols. DK2208 (978-0-8247-5564-5)	Pub'd. 12/29/04 DK501X (978-0-8247-5501-0)
Encyclopedia of Plasma Technology, 2 Vols. K14378 (978-1-4665-0059-4)	Pub'd. 12/12/2016 K21744 (978-1-4822-1431-4)	Encyclopedia of Pharmaceutical Science and Technology, 4th Ed., 6 Vols. Pub'd. 7/1/13	H100233 (978-1-84184-819-8)
Encyclopedia of Supramolecular Chemistry, 2 Vols. DK056X (978-0-8247-5056-5)	Pub'd. 5/5/04 DKE7259 (978-0-8247-4725-1)	Routledge Encyclopedias	HE10420 (978-1-84184-820-4)
Encyclopedia of Surface & Colloid Science, 3rd Ed., 10 Vols. K20465 (978-1-4665-9045-8)	Pub'd. 8/27/15 K20478 (978-1-4665-9061-8)	Encyclopedia of Public Administration and Public Policy, 3rd Ed., 5 Vols. Pub'd. 11/6/15	K16418 (978-1-4665-6909-6)
Engineering		Routledge Encyclopedia of Modernism Y137844 (978-1-135-00035-6)	K16434 (978-1-4665-6936-2)
Dekker Encyclopedia of Nanoscience and Nanotechnology, 3rd Ed., 7 Vols. Pub'd. 3/20/14	K14119 (978-1-4398-9134-6)	Routledge Encyclopedia of Philosophy Online RU22334 (978-0-415-24909-6)	Pub'd. 11/1/00
K14120 (978-1-4398-9135-3)		Routledge Performance Archive Y148405 (978-0-203-77466-3)	Pub'd. 11/12/12

Encyclopedia titles are available in print and online.

To order, visit <https://www.crcpress.com>

Telephone: 1-800-272-7737

Email: orders@taylorandfrancis.com

Encyclopedia of
Biopharmaceutical
Statistics,
4th Edition

Volume I – IV

edited by
Shein-Chung Chow, PhD



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2018 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed on acid-free paper

International Standard Book Number-13: 978-1-4987-3395-3 (Set), 978-0-8153-6314-9 (Vol. 1)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

Brief Content

Volume I

Acceptance Sampling.....	1	Biologics	233
Active Control Non-Inferiority Trials: Hypotheses and False Positive Rate.....	8	Biomarker Development with Applications to Genetics Data: Statistical Tests.....	243
Active Control Trials.....	14	Biomarker: Clinical Trials	251
Active Controlled Clinical Trials: Noninferiority Analysis	23	Biopharmaceutics.....	254
Adaptive Design Clinical Trials: Case Studies.....	28	Biosimilar Studies: Sample Size.....	274
Adaptive Design Methods in Clinical Trials	36	Biosimilar Studies: Three-Arm Design.....	277
Adaptive Survival Trials.....	50	Biosimilarity Assessment	281
Adverse Event Reporting.....	53	Biosimilarity of Follow-On Biologics.....	304
Alpha Spending Function	59	Bivariate Zero-Inflated Poisson Populations: Statistical Test for Homogeneity	316
Ames Test	67	Blinded Sample Size Re-Estimation.....	323
Analysis of $2 \times K$ Tables	75	Blinding.....	336
Analysis of Clustered Binary Data.....	82	Bootstrap, The	342
Analysis of Clustered Categorical Data.....	88	Botanical Drug Product Development: Scientific Issues	348
Analysis of Heritability	98	Bracketing Design	359
Analysis of Variance (ANOVA).....	104	Bridging Studies.....	361
Analytical Similarity Assessment.....	116	Bridging Studies: Statistical Methods.....	366
ANCOVA Approach: Premarketing Shelf Life Determination with Multiple Factors	125	Calibration	370
Assay Development	130	Canadian Health Products and Food Branch (HPFB) and Therapeutic Products Directorate (TPD).....	378
Assay Validation.....	138	Cancer Chemotherapy: Maximum Tolerable Dose	384
Assay Validation: Evaluation of Linearity	146	Cancer Trials	390
Bayesian Designs: Phase II Oncology Clinical Trials	154	Carcinogenicity Studies of Pharmaceuticals	398
Bayesian Estimate: Concordance Correlation Coefficient.....	164	Carry-Forward Analysis.....	416
Bayesian Methods: Meta-Analysis.....	173	Case-Control Studies: Inference in	422
Bayesian Methods: Stability Analysis	180	Center Weighting: Multicenter Trials.....	435
Bayesian Model: Prior Effective Sample Size	186	Clinical Data Management.....	444
Bayesian Statistics	190	Clinical Endpoint.....	451
Bayesian Statistics: Medical Device Clinical Trials	198	Clinical Endpoint BE Studies: Generic Drugs	455
Binary 2×2 Crossover Trials	206	Clinical Inspection: Statistical Process.....	462
Bioassay.....	213	Clinical Pharmacology	469
Bioavailability and Bioequivalence.....	216	Clinical Research: Benefit and Risk Assessment.....	472
Bioinformatics	224	Clinical Research: Comparing Variabilities	486
		Clinical Research: Outlier Detection.....	502
		Clinical Research: Reproducibility Probability	510

Volume I (cont'd.)

Clinical Trial Design: Statistical Assurance.....	523
Clinical Trial Designs with Prospective Patient Population Enrichment by Response.....	526
Clinical Trial Process.....	532
Clinical Trial Simulation.....	536
Clinical Trial Simulations: Earlier Development Phases.....	539
Clinical Trial Simulations: Later Development Phases.....	544
Clinical Trial: Combination Drug Development.....	552
Clinical Trial: Failure-Time Model.....	556
Clinical Trial: <i>N</i> -of-1 Design Analysis.....	564
Clinical Trials.....	572
Clinical Trials: Controversial Issues.....	580

Volume II

Clinical Trials: Interval-censored Failure Time Data.....	589
Clinical Trials: Investigating Quality-of-Life.....	597
Clinical Trials: Meta-Analysis using R.....	603
Clinical Trials: Response-Adaptive Repeated Measurement Designs.....	612
Clinical Trials: Selection of Control.....	622
Clinical Trials: Vaccine Development.....	628
Cluster Randomized Trials: Sample Size.....	645
Cluster Trials.....	648
Clustered Study Designs: Power Analysis.....	658
Companion Diagnostics.....	672
Comparability Studies: Statistical Considerations.....	682
Confidence Interval and Hypothesis Testing.....	699
Confounding and Interaction.....	703
Content Uniformity.....	709
Continual Reassessment Methods.....	712
Contract Research Organization (CRO): Early 2000s.....	716
Contract Research Organization (CRO).....	721
Controlled Clinical Trials: <i>P</i> -Values, Evidence, and Multiplicity Considerations.....	725
Correlated Probit Model.....	741

Cost-Effectiveness Analysis.....	751
Covariate-Adjusted Adaptive Dose-Finding: Early Phase Clinical Trials.....	760
Covariates: Adjustment for.....	772
Crossover Design.....	776
Crossover Design: Rank-Based Robust Analysis.....	784
Cutoff Designs.....	798
Data Mining and Biopharmaceutical Research.....	805
Data Monitoring Committees (DMC).....	811
Diagnostic Device Evaluation.....	822
Diagnostic Imaging.....	829
Diagnostic Procedure: Sensitivity and Specificity.....	836
Disease Modification Effects: Design and Analysis.....	839
Dissolution Profile Comparison: Statistical Methods.....	844
Dissolution Profiles Comparison: Model-Independent Approaches.....	854
Dosage Units Using Small and Large Sample Sizes: Uniformity.....	859
Dose Finding Clinical Trials Using R.....	866
Dose Proportionality.....	876
Dose Response Analysis in Clinical Trials.....	879
Dose Response Study Design.....	885
Dose-Ranging Studies: Design Considerations.....	891
Dropout.....	901
Drug Development.....	908
Drug Interchangeability: Biosimilar Products.....	914
Drug Interchangeability: Criteria and Design.....	922
Drug Interchangeability: Meta-Analysis for Safety Monitoring.....	928
Drug Interchangeability: Generic Products, Prescribability and Switchability.....	935
Ecologic Inference.....	940
ED ₅₀ /ED ₉₀	952
Enrichment Design.....	960
Enrichment Design with Patient Population Augmentation.....	964
Equivalence Trials.....	970
Ethnic Factors.....	976
Expiration Dating Period.....	979

Exploratory Factor Analysis.....	985	Kruskal-Wallis Test for Ordered	
Extra Variation Models.....	992	Categorical Data: Power Calculation	1245
Factor Analysis.....	1008	Kullback–Leibler Divergence for	
Factorial Designs	1014	Evaluating Equivalence	1249
False Discovery Rate (FDR)	1028	Laboratory Analyses	1258
GEE in Cancer Research	1036	Latent Class Analysis.....	1266
Generalizability Probability in		Lilly Reference Ranges	1276
Clinical Research	1042	Local Influence Analysis	1292
Generalized Estimating Equation	1046	LOCF: Validity	1296
Generalized Inference	1054	Logistic Regression	1303
Genetic Linkage and Linkage		Logistic Regression Process: Predictive	
Disequilibrium Analysis	1057	Model Building in Clinical Research	1309
Global Database and System	1069	Logistic Regression: Three-Point	
Good Clinical Practice (GCP)	1075	Designs	1315
Good Laboratory Practice (GLP)	1082	McNemar’s Test.....	1320
Good Programming Practice (GPP)	1088	MCP-Mod: Dose–Response Testing	
Good Statistics Practice (GSP).....	1095	and Estimation	1326
Good Validation Practice (GVP)	1101	Measuring Agreement.....	1335
Group Sequential Methods	1105	MedDRA: Impact on Pharmaceutical	
Group Sequential Tests and Variance		Development.....	1342
Heterogeneity: Clinical Trials	1114	Medical Devices	1348
Human Drug Abuse Trials	1121	Meta-Analysis of Therapeutic Trials.....	1362
Hypothesis Testing.....	1124	Microarray Gene Expression	1377
Imputation with Item Nonrespondents	1128	Microdosing Studies	1394
Imputation: Clinical Research	1135	Minimization Procedure.....	1401
<i>In Vitro</i> Testing: Bioequivalence	1141	Minimum Effective Dose.....	1406
<i>In Vitro</i> Testing: Dissolution Profile		Mixed Effects Models	1409
Comparison	1148	Mixed-Outcome Data	1414
<i>In Vitro</i> Testing: Micronucleus Test.....	1155	MMRM with Missing Data.....	1422
Individual Bioequivalence.....	1160	Modified Large Sample Method.....	1427
		Multicenter Trials	1434
		Multicollinearity	1438
		Multidimensional Data Analysis:	
		Penalized Regression Methods	1448
		Multinational Clinical Trials	1454
		Multiple Comparisons.....	1461
		Multiple Endpoints	1467
		Multiple-Dose Bioequivalence Studies.....	1477
		Multiple-Stage Designs: Phase II	
		Cancer Trials.....	1481
		Multiplicative Intensity Models	1494
		Multiplicity in Clinical Trials.....	1502
		Multi-Regional Clinical Trials:	
		Consistency Test	1511
		Multiregional Clinical Trials: Qualitative	
		Consistency of Treatment Effects	1522
		Multivariate Meta-Analysis	1528
		Noninferiority Trials: Type I Error Rates.....	1536
		Nonparametric Regression	1543

Volume III

Instrument Development and	
Validation.....	1167
Integrated Summary Report	1179
Intention-to-Treat Analyses (ITT).....	1183
Interchangeability: Generic Drugs	1187
Interim Analysis.....	1195
International Conference on	
Harmonization (ICH).....	1202
Item Response Theory.....	1207
IVRS/IWRS for Randomization	1216
Japan: Ministry of Health, Labour and	
Welfare and Pharmaceutical	
Administration	1224
Kaplan–Meier Estimator	1231
Kappa Coefficients: Medical Research	1237

Volume III (cont'd.)

Odds Ratio	1554	Propensity Score Analysis and Its Application in Regulatory Settings	1823
Oncology Clinical Trials: Efficient Dose-Finding Designs	1560	Propensity Score: Methodology and Application to Clinical Studies	1825
Oncology Trials: Development Strategy	1572	Proportion of Treatment Effect	1830
Onset of Action	1585	Proportional Hazards Regression Model	1834
Optimal Futility Interim Design	1593	Protocol Development	1841
Ordered Categorical Data: Test for	1602	P-Values	1845
Ordered Multiple Class Receiver Operating Characteristic (ROC) Analysis	1608	QT Analysis	1855
Ordinal Data: Crossover Design Analysis	1613	Quality by Design in Biopharmaceutical Development	1865
Parallel Design	1621	Randomization	1875
Patient Compliance	1627	Randomized Trials with Clusters: Design and Analysis	1880
Patient-Reported Outcomes	1633	Rank Regression: Stability Analysis	1888
Percentile Charts on Correlated Measures	1643	Regulatory Statistics	1893
Personalized Medicine: Validation of Diagnostic Medical Devices	1648	Release Targets	1903
Pharmaceutical Development: Statistical Designs	1656	Reliability	1908
Pharmacodynamic Issues	1665	Repeated Measurement Designs: Missing Values	1918
Pharmacodynamics with Covariates	1673	Repeated Measures Data with Missing Values Analysis: Overview of Methods	1924
Pharmacodynamics with No Covariates	1685	Reproductive and Developmental Studies	1931
Pharmacoeconomics	1693	Response Surface Methodology	1939
Pharmacoeconomics: Cost-Effectiveness Analysis	1707	Response-Adaptive Designs	1950
Phase I Cancer Clinical Trials	1715	Risk Ratio Analysis	1956
Phase II Clinical Development	1721	ROC Curve	1962
Phase III Clinical Trials with Response- Adaptive Allocation	1727	Sample Size Calculation: Nonparametrics	1970
Placebo Effect	1735	Sample Size Calculation: Survival Data	1980
Population Bioequivalence	1740	Sample Size Determination	1987

Volume IV

Population PK/PD Analysis	1747	Sample Size Determination: Three-arm Ratio of Mean Differences Equivalence Test	2002
Postmarketing Adverse Drug Event Signaling	1756	Sample Size Estimation: Generalized Estimating Equations (GEE) Method	2010
Postmarketing Surveillance	1769	Sample Size for Multiregional Clinical Trials	2015
Power	1777	Sample Size Re-estimation Based on Observed Treatment Difference	2021
Precision Medicine: Statistical Issues	1779	Sample Size Re-estimation Design with Robust Power	2026
Prediction Trees	1782	Sample Size: Negative Binomial With Unequal Follow-Up Times	2031
Premarket Applications: Registries	1791	Scaled Average Bioequivalence (SABE)	2038
Principal Component Analysis	1796	Screening Design	2045
Process Validation	1802	Seamless Adaptive Trial Design: Dose Selection Criteria	2047
Profile Analysis	1817		

Semi-Parametric Time-Varying		
Regression Models.	2058	
Sequential Estimation: Additive Hazards		
Rate Model with Staggered Entry.	2063	
Signal Detection in Pharmacovigilance.	2068	
Slope Approach: Assessment of Dose		
Proportionality and Linearity Under a		
Crossover Design.	2084	
Spatio-Temporal Modeling.	2089	
Specifications.	2099	
SROC Curve.	2101	
Stability Analysis: Frozen Drug		
Products.	2111	
Stability Matrix Designs.	2116	
Statistical Genetics.	2122	
Statistical Principles for Clinical Trials.	2130	
Statistical Process Control.	2135	
Statistical Significance.	2142	
Structural Equation Model.	2147	
Stuart–Maxwell Test.	2154	
Subgroup Analysis.	2159	
Subject-Treatment Interaction.	2167	
Surrogate Endpoint.	2174	
Survival Analysis.	2179	
Targeted Clinical Trials.	2184	
Testing for Qualitative Interaction.	2192	
Therapeutic Equivalence.	2199	
Thorough QT Studies.	2206	
Tiered Approach of Analytical Similarity		
Assessment: Statistical Methods.	2219	
Titration Design.	2225	
Toxicological Studies.	2230	
Traditional Chinese Medicine (TCM):		
Clinical Development.	2244	
Traditional Chinese Medicine (TCM):		
General Considerations.	2252	
Traditional Chinese Medicine (TCM):		
Unified Approach for Evaluation.	2265	
Translational Medicine: Concepts, Statistical		
Methods, and Related Issues.	2270	
Trend Estimation.	2283	
Two-Stage Adaptive Design: Analysis.	2289	
Two-Stage Design: Phase II Cancer		
Clinical Trials.	2296	
US Food and Drug Administration.	2305	
USP Tests.	2311	
Validation of Quantitative and		
Qualitative Assays.	2313	
Weibull Model: Group Sequential Design.	2325	
Z-Score.	2329	



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Encyclopedia of Biopharmaceutical Statistics

Editor-in-Chief

Shein-Chung Chow, PhD

*Professor, Department of Biostatistics and Bioinformatics
Duke University School of Medicine, Durham, North Carolina, U.S.A.*



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Contributors

- Marilyn A. Agin** / *Pfizer Global Research and Development, Ann Arbor, Michigan, U.S.A.*
Chul Ahn / *University of Texas Medical School, Houston, Texas, U.S.A.*
Mohamed Alosch / *Office of Biostatistics, CDER, FDA, Silver Spring, Maryland, U.S.A.*
Glen Andrews / *Department of Biostatistics, Pfizer, Inc., San Diego, California, U.S.A.*
C. Ané / *Ecole Nationale Vétérinaire de Toulouse, Toulouse, France*
He Bai / *Department of Drug & Cosmetics Registration, China Food and Drug Administration, Beijing, China*
Shan Bai / *Eli Lilly and Company, Indianapolis, Indiana, U.S.A.*
Heejung Bang / *Division of Biostatistics, Department of Public Health Sciences, University of California, Davis, California, U.S.A.*
Avner Bar-Hen / *INA-PG/INRA Biometrie, Paris, France*
Mary K. Barrick / *CDRH, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
Richard Baumgartner / *Biometrics Research, MRL, Rahway, New Jersey, U.S.A.*
Guenther Baus / *Aventis Pharma Deutschland GmbH, Frankfurt, Germany*
Henri Begleiter / *Department of Psychiatry, HSCB, State University of New York, Downstate Medical Center, Brooklyn, New York, U.S.A.*
James S. Bergum / *Bristol-Myers Squibb Company, New Brunswick, New Jersey, U.S.A.*
Rahul Bhattacharya / *University of Calcutta, Kolkata, India*
Rafia Bhore / *Department of Clinical Sciences, University of Texas, Southwestern Medical Center, Dallas, Texas, U.S.A.*
Atanu Biswas / *Indian Statistical Institute, Kolkata, India*
Bipasa Biswas / *Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
Laurent Bordes / *Laboratoire de Mathématiques Appliquées, Université de Pau et des Pays de l'Adour, Pau, France*
Björn Bornkamp / *Clinical Development & Analytics, Novartis Pharma AG, Basel, Switzerland*
William A. Brennenman / *The Procter & Gamble Company, Mason, Ohio, U.S.A.*
Frank Bretz / *Statistical Methodology and Consulting, Novartis, Basel, Switzerland; Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Wien, Austria*
Christelle Breuils / *Département STID, IUT de Metz, Metz, France*
Philippe Broet / *INSERM U472, Villejuif Cedex, France*
Brent D. Burch / *U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
Greg Campbell / *CDRH, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
Robert Capen / *Merck & Co., Inc., West Point, Pennsylvania, U.S.A.*
Joseph C. Cappelleri / *Pfizer Inc., Groton, Connecticut, U.S.A.*
Joseph L. Carrasco / *Bioestadística, Universitat de Barcelona, Barcelona, Spain*
Patricia B. Cerrito / *University of Louisville, Louisville, Kentucky, U.S.A.*
Ivan S. F. Chan / *Merck Research Laboratories, West Point, Pennsylvania, U.S.A.*

- Cheng-Tao Chang** / *Novo Nordisk Inc., Princeton, New Jersey, U.S.A.*
- Jeffrey T. Chang** / *Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina, U.S.A.*
- Mark Chang** / *Veristat LLC, Southborough, Massachusetts, U.S.A.*
- Suming W. Chang** / *Berlex Laboratories, Inc., Montville, New Jersey, U.S.A.*
- Yu-Wei Chang** / *Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, Connecticut, U.S.A.*
- Rick Chappell** / *University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.*
- Chi-Tian Chen** / *Institute of Population Health Sciences, National Health Research Institutes, Miaoli, Taiwan*
- Chi-wan Chen** / *Office of Biostatistics/Office of Pharmacoepidemiology and Biostatistical Sciences, Center of Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Chun-Houh Chen** / *Institute of Statistical Science, Academia Sinica, Taipei, Taiwan*
- Cong Chen** / *Merck & Co., Inc., West Point, Pennsylvania, U.S.A.*
- Ding-Geng (Din) Chen** / *Department of Biostatistics, Gillings School of Global Health, University of North Carolina, Chapel Hill, North Carolina, U.S.A.; Department of Statistics, University of Pretoria, Pretoria, South Africa; School of Social Work, University of North Carolina, Chapel Hill, North Carolina, U.S.A.*
- Gang Chen** / *Johnson & Johnson Pharmaceutical Research & Development, Raritan, New Jersey, U.S.A.*
- James J. Chen** / *National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, Arkansas, U.S.A.*
- Jie Chen** / *Abbott Laboratories, Abbott Park, Illinois, U.S.A.*
- Meng Chen** / *Duke University School of Medicine, Durham, North Carolina, U.S.A.*
- Mon-Gy Chen** / *Department of Biostatistics, Amgen, Thousand Oaks, California, U.S.A.*
- Wen Jen Chen** / *Office of Biostatistics/Office of Pharmacoepidemiology and Biostatistical Sciences, Center of Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Wen-Hung Chen** / *North Potomac, Maryland, U.S.A.*
- Ying Qing Chen** / *Fred Hutchinson Cancer Research Center, Seattle, Washington, U.S.A.*
- Bin Cheng** / *Department of Biostatistics, Columbia University, New York, New York, U.S.A.*
- Robert D. Chew** / *Pfizer Global Research and Development, Groton, Connecticut, U.S.A.*
- Eric Chi** / *Amgen Inc., Thousand Oaks, California, U.S.A.*
- George Y. H. Chi** / *Johnson & Johnson Pharmaceutical Research & Development, Raritan, New Jersey, U.S.A.*
- Chieh Chiang** / *Institute of Population Health Sciences, National Health Research Institutes, Miaoli, Taiwan*
- Vernon M. Chinchili** / *Penn State College of Medicine, Hershey, Pennsylvania, U.S.A.*
- Keumhee Carriere Chough** / *Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada*
- Shein-Chung Chow** / *Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, U.S.A.*
- Christy Chuang-Stein** / *Pfizer Global Research and Development, Groton, Connecticut, U.S.A.*
- Antonio Ciampi** / *McGill University, Monreal, Québec, Canada*
- Mario Comelli** / *Università di Pavia, Pavia, Italy*

- D. Concorde** / *Ecole Nationale Vétérinaire de Toulouse, Toulouse, France*
- John R. Cook** / *Merck Research Laboratories, West Point, Pennsylvania, U.S.A.*
- Can Cui** / *Duke University, Durham, North Carolina, U.S.A.*
- Lu Cui** / *Data and Statistical Science, AbbVie Inc., North Chicago, Illinois, U.S.A.*
- Cyril Dalmasso** / *INSERM U472, Villejuif Cedex, France*
- Qianyu Dang** / *CDER, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Kalyan Das** / *Department of Statistics, University of Calcutta, Kolkata, India*
- Jean-Jacques Daudin** / *Applied Mathematics and Computer Sciences, INA-PG/INRA, Paris, France*
- Jill Dawson** / *Corporate Communications Consultant, Danville, California, U.S.A.*
- A. R. de Leon** / *Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada*
- Charmaine B. Dean** / *Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada*
- Qiqi Deng** / *Department of Biostatistics and Data Sciences, Boehringer-Ingelheim Pharmaceuticals, Inc., Ridgefield, Connecticut, U.S.A.*
- Mehul Desai** / *Division of Cardiovascular and Renal Products, Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Marc DeSomer** / *PPD, Wilmington, North Carolina, U.S.A.*
- Xiaoyu Dong** / *Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Vladimir Dragalin** / *GlaxoSmithKline, Collegeville, Pennsylvania, U.S.A.*
- Michael F. Drummond** / *Centre for Health Economics, York, U.K.*
- Satya D. Dubey** / *Office of Biostatistics, CDER, FDA, Silver Spring, Maryland, U.S.A.*
- Amylou C. Dueck** / *Mayo Clinic, Scottsdale, Arizona, U.S.A.*
- Valerie Durkalski** / *Division of Biostatistics and Epidemiology, Medical University of South Carolina, Charleston, South Carolina, U.S.A.*
- Susan S. Ellenberg** / *U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.*
- Laszlo Endrenyi** / *Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada*
- Chunpeng Fan** / *Department of Biostatistics and Programming, Sanofi US Inc., Bridgewater, New Jersey, U.S.A.*
- Douglas E. Faries** / *Eli Lilly and Company, Indianapolis, Indiana, U.S.A.*
- Patrick J. Farrell** / *School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada; McLaughlin Center for Population Health Risk Assessment, University of Ottawa, Ottawa, Ontario, Canada and Risk Sciences International, Ottawa, Ontario, Canada*
- C. Feng** / *Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York, U.S.A.*
- Dai Feng** / *Biometrics Research, MRL, Rahway, New Jersey, U.S.A.*
- Jason Fine** / *Department of Statistics and Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.*
- Guifang Fu** / *Departments of Statistics, Pennsylvania State University, University Park, Pennsylvania, U.S.A.*
- Gary L. Gadbury** / *Department of Statistics, Kansas State University, Manhattan, Kansas, U.S.A.*
- Paul Gallo** / *Biostatistics, Novartis Pharmaceuticals Corporation, East Hanover, New Jersey, U.S.A.*

- Ying Geng** / *National Institutes for Food and Drug Control, Beijing, China*
- Elizabeth S. Garrett-Mayer** / *Hollings Cancer Center, Medical College of South Carolina, Charleston, South Carolina, U.S.A.*
- Michael L. Garriott** / *(Retired) Eli Lilly and Company, Indianapolis, Indiana, U.S.A.*
- Theo Gasser** / *University of Zurich, Zurich, Switzerland*
- Shiva Gautam** / *Harvard Medical School, Boston, Massachusetts, U.S.A.*
- Patrick Genyn** / *Johnson & Johnson PRD, Raritan, New Jersey, U.S.A.*
- Robert A. Gerber** / *Pfizer Inc., New London, Connecticut, U.S.A.*
- A. Lawrence Gould** / *Merck Research Laboratories, Blue Bell, Pennsylvania, U.S.A.*
- Christopher A. Gravel** / *School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada; McLaughlin Center for Population Health Risk Assessment, University of Ottawa, Ottawa, Ontario Canada; and Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada*
- Gerry Gray** / *Data-Fi LLC, Silver Spring, Maryland, U.S.A.*
- Sander Greenland** / *Department of Epidemiology and Department of Statistics, University of California, Los Angeles, California, U.S.A.*
- Stella Grosser** / *Office of Biostatistics, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Ralitza V. Gueorguieva** / *Department of Epidemiology and Public Health, Yale University, New Haven, Connecticut, U.S.A.*
- Stephen Hall** / *Division of Clinical Pharmacology, Department of Medicine, Indiana University, Indianapolis, Indiana, U.S.A.*
- Ian R. Harris** / *Southern Methodist University, Dallas, Texas, U.S.A.*
- Dieter Hauschke** / *Byk Gulden Pharmaceuticals, Konstanz, Germany*
- Lan He** / *National Institutes for Food and Drug Control, Beijing, China*
- Li He** / *Merck & Co., Inc., West Point, Pennsylvania, U.S.A.*
- Xuming He** / *Department of Statistics, University of Illinois, Champaign, Illinois, U.S.A.*
- Jay Herson** / *Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, U.S.A.*
- Shahram Heshmat** / *Department of Public Health, University of Illinois at Springfield, Springfield, Illinois, U.S.A.*
- Joseph Heyse** / *Merck Research Laboratories, West Point, Pennsylvania, U.S.A.*
- Shuyen Ho** / *Global Data Operations, PAREXEL International, Durham, North Carolina, U.S.A.*
- Thomas Hoffman** / *Aventis Pharma Deutschland GmbH, Hattersheim, Germany*
- Wherly P. Hoffman** / *Department of Biometrics, Elan Pharmaceuticals Inc., South San Francisco, California, U.S.A.*
- Minghuang Hong** / *Sun Yat-sen University Cancer Centre, Guangzhou, China*
- A. D. Horne** / *U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.*
- Chin-Fu Hsiao** / *Institute of Population Health Sciences, National Health Research Institutes, Miaoli, Taiwan*
- Eric Hsieh** / *Division of Biometry, Department of Agronomy, National Taiwan University, Taipei, Taiwan*
- Alice T. M. Hsuan** / *Aventis Pharmaceuticals, Inc., Bridgewater, New Jersey, U.S.A.*
- Feifang Hu** / *University of Virginia, Charlottesville, Virginia, U.S.A.*
- Zhaowei Hua** / *Takeda Pharmaceuticals Inc., Cambridge, Massachusetts, U.S.A.*
- Dalong (Patrick) Huang** / *Office of Biostatistics, Office of Translational Sciences, CDER, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*

- Rong Huang** / *Statistics Canada, Ottawa, Ontario, Canada*
- Yangxin Huang** / *University of Rochester, Rochester, New York, U.S.A.*
- Jurgen Hummel** / *PPD, Bellshill, U.K.*
- H. M. James Hung** / *Division of Biometrics I, Office of Biostatistics, U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.*
- Mohammad F. Huque** / *Office of Biostatistics, CDER, FDA, Silver Spring, Maryland, U.S.A.*
- Abdulkadir Hussein** / *University of Windsor, Windsor, Ontario, Canada*
- William J. Huster** / *Eli Lilly and Company, Indianapolis, Indiana, U.S.A.*
- Irving K. Hwang** / *Irving Consulting Group, Pluckemin, New Jersey, U.S.A., and University of Medicine and Dentistry of New Jersey (UMDNJ), New Brunswick, New Jersey, U.S.A.*
- Boris Iglewicz** / *Temple University, Philadelphia, Pennsylvania, U.S.A.*
- John P. A. Ioannidis** / *Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece*
- Anastasia Ivanova** / *Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.*
- Hongyu Jiang** / *Harvard University, Boston, Massachusetts, U.S.A.*
- Qi Jiang** / *Global Biostatistical Sciences, Amgen Inc., Thousand Oaks, California, U.S.A.*
- Yuying Jin** / *Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Kevin Jones** / *Department of Psychiatry, HSCB, State University of New York, Brooklyn, New York, U.S.A.*
- Jungnam Joo** / *National Heart, Lung and Blood Institute, Bethesda, Maryland, U.S.A.*
- Sin-Ho Jung** / *Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, U.S.A.*
- Seung-Ho Kang** / *Department of Applied Statistics, Yonsei University, Seoul, South Korea*
- Chunlei Ke** / *Global Biostatistical Sciences, Amgen Inc., Thousand Oaks, California, U.S.A.*
- Roswitha E. Kelly** / *Office of Biostatistics/Office of Pharmacoepidemiology and Biostatistical Sciences, Center of Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Gary King** / *Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts, U.S.A.*
- Michael J. Klepper** / *Integrated Safety Systems, Inc., Cary, North Carolina, U.S.A.*
- Ronald K. Knickerbocker** / *Pfizer Global Research and Development, Ann Arbor, Michigan, U.S.A.*
- Masha Kocherginsky** / *Department of Health Studies, The University of Chicago, Chicago, Illinois, U.S.A.*
- Djokouri A. Kouassi** / *Merck Research Laboratories, Summit, New Jersey, U.S.A.*
- J. Kowalski** / *Division of Oncology Biostatistics, Johns Hopkins University, Baltimore, Maryland, U.S.A.*
- Helena Chmura Kraemer** / *Stanford University School of Medicine, Stanford, Connecticut, U.S.A.*
- Paul F. Kramer** / *Bristol-Myers Squibb Company, Plainsboro, New Jersey, U.S.A.*
- Daniel Krewski** / *School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada; McLaughlin Center for Population Health Risk Assessment, University*

of Ottawa, Ottawa, Ontario, Canada; Risk Sciences International, Ottawa, Ontario, Canada; and School of Epidemiology, Public Health, and Preventive Medicine, University of Ottawa, Ottawa, Ontario, Canada

Anant M. Kshirsagar / *University of Michigan, Ann Arbor, Michigan, U.S.A.*

James T. Kuznicki / *The Procter & Gamble Co., Cincinnati, Ohio, U.S.A.*

P. A. Lachenbruch / *U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.*

Pan-Yu Lai / *Allergan, Inc., Irvine, California, U.S.A.*

Mani Y. Lakshminarayanan / *Biostatistics and Research Decision Sciences, Merck & Co., Inc., Rahway, New Jersey, U.S.A.*

Joseph Lau / *Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, Massachusetts, U.S.A.*

C. Gordon Law / *Pfizer Inc., Groton, Connecticut, U.S.A.*

Allen Lee / *Merck Research Laboratories, Summit, New Jersey, U.S.A.*

Bee Leng Lee / *Department of Mathematics, San José State University, San José, California, U.S.A.*

Cindy Lee / *Eli Lilly and Company, Indianapolis, Indiana, U.S.A.*

Hye-Seung Lee / *Pediatrics Epidemiology Center, Department of Pediatrics, University of South Florida, Tampa, Florida, U.S.A.*

Shiowjen Lee / *U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*

T. Y. Lee / *Independent Consultant, Morristown, New Jersey, U.S.A.*

Yonghee Lee / *Forest Research Institute, Jersey City, New Jersey, U.S.A.*

F. Léger / *Université Paul-Sabatier and Institut Claudius-Regaud, Toulouse, France*

Robert G. Lehr / *Bayer Healthcare Pharmaceuticals, Montville, New Jersey, U.S.A.*

Xiudong Lei / *Division of Biostatistics, School of Public Health, University of California, Berkeley, California, U.S.A.*

Benjamin E. Leiby / *Department of Pharmacology and Experimental Therapeutics, Thomas Jefferson University, Philadelphia, Pennsylvania, U.S.A.*

Jeannie-Marie S. Leoutsakos / *Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, Maryland, U.S.A.*

Chin-Shang Li / *St. Jude Children's Research Hospital, Memphis, Tennessee, U.S.A.*

Chi-Rong Li / *Chung Shan Medical University, Taichung, Taiwan, R.O. China*

David Li / *Solvay Pharmaceuticals, Inc., Marietta, Georgia, U.S.A.*

Heng Li / *Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*

Hongying Li / *Departments of Statistics, University of Florida, Gainesville, Florida, U.S.A.*

Hung-Ir Li / *Allergan, Inc., Irvine, California, U.S.A.*

Juan Li / *Amgen Inc., Thousand Oaks, California, U.S.A.*

Junfang Li / *Aventis Pharma, Ltd., Tokyo, Japan*

Lang Li / *Department of Medicine, Indiana University, Indianapolis, Indiana, U.S.A.*

Meijuan Li / *Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*

Ning Li / *Sanofi-Aventis, China*

Qin Li / *Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*

Xuefeng Li / *Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*

- Meijuan Li** / *Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Yuanyuan Liang** / *Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, San Antonio, Texas, U.S.A.*
- Chen-Tuo Liao** / *National Taiwan University, Taipei, Taiwan, R.O. China*
- Jason J. Z. Liao** / *Merck & Co., Inc., North Wales, Pennsylvania, U.S.A.*
- Daphne T. Lin** / *Office of Biostatistics/Office of Pharmacoepidemiology and Biostatistical Sciences, Center of Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Karl K. Lin** / *U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Lawrence I-Kuei Lin** / *Baxter Healthcare Corporation, Round Lake, Illinois, U.S.A.*
- Melody H. Lin** / *Office for Human Research Protections, Department of Health and Human Services, Rockville, Maryland, U.S.A.*
- Shang P. Lin** / *Department of Health Studies, The University of Chicago, Chicago, Illinois, U.S.A.*
- Tsai-Lien Lin** / *U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Xue Lin** / *US Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Karl K. Lin** / *Office of Biostatistics/Office of Pharmacoepidemiology and Biostatistical Sciences, Center of Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Min Lin** / *Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Stacy R. Lindborg** / *Eli Lilly and Company, Indianapolis, Indiana, U.S.A.*
- Aiyi Liu** / *National Institute of Child Health and Human Development, Bethesda, Maryland, U.S.A.*
- Cameron S. Liu** / *Ortho Biotech Oncology Research & Development, a Unit of Cougar Biotechnology, Los Angeles, California, U.S.A.*
- Jen-pei Liu** / *Division of Biometry, Department of Agronomy, and Institute of Epidemiology, National Taiwan University, Taipei, Taiwan; and Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan*
- Yi Liu** / *Takeda Pharmaceuticals Inc., Cambridge, Massachusetts, U.S.A.*
- Mo Liu** / *National Clinical Research Center for Digestive Diseases, Beijing Friendship Hospital, Capital Medical University, Beijing, China*
- Wen-Wei Liu** / *Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, U.S.A.*
- Bing Lu** / *Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, U.S.A.*
- Nelson Lu** / *Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Jingyu Julia Luan** / *Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Kung-Jong Lui** / *Department of Mathematics and Statistics, San Diego State University, San Diego, California, U.S.A.*
- Herman G. Luus** / *Quintiles ClinData, Bloemfontein, South Africa*
- C. J. Lynch** / *U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.*
- Petra Macaskill** / *Screening and Test Evaluation Program (STEP), Sydney School of Public Health, University of Sydney, Sydney, New South Wales, Australia*
- James MacDougall** / *Genzyme, Cambridge, Massachusetts, U.S.A.*

- David Machin** / *Children's Cancer and Leukaemia Group, University of Leicester, Leicester, U.K.*
- Ying C. MacNab** / *Department of Health Care and Epidemiology, University of British Columbia, Vancouver, British Columbia, Canada*
- Terri Madison** / *STATPROBE, Ann Arbor, Michigan, U.S.A.*
- Iain J. McGilveray** / *McGilveray Pharmacon, Inc., Nepean, Ontario, Canada*
- Helen McGough** / *Human Subjects Division, University of Washington, Seattle, Washington, U.S.A.*
- Devan V. Mehrotra** / *Merck Research Laboratories, West Point, Pennsylvania, U.S.A.*
- Kerrie L. Mengersen** / *School of Mathematical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia*
- Laura J. Meyerson** / *Biogen, Inc., Cambridge, Massachusetts, U.S.A.*
- G. Molengerghs** / *Leuven Biostatistics and Statistical Bioinformatics Centre, Department of Public Health, Leuven, Belgium*
- Sean D. Mooney** / *Buck Institute for Age Research, Novato, California, U.S.A.*
- Dirk F. Moore** / *Department of Biostatistics, University of Medicine and Dentistry of New Jersey School of Public Health, New Jersey, U.S.A.*
- Jorge G. Morel** / *The Procter & Gamble Company, Winton Hill Business Center, Cincinnati, Ohio, U.S.A.*
- Satoshi Morita** / *Yokohama City University Medical Center, Yokohama, Kanagawa, Japan*
- Shahrul Mt-Isa** / *Biostatistics and Research Decision Sciences (BARDS) Europe, MSD Research Laboratories, MSD, Hoddesdon, U.K.*
- John R. Murphy** / *Eli Lilly and Company, Indianapolis, Indiana, U.S.A.*
- William R. Myers** / *The Procter & Gamble Company, Mason, Ohio, U.S.A.*
- Kiyohito Nakai** / *Ministry of Health, Labour and Welfare, Tokyo, Japan*
- Christos T. Nakas** / *Laboratory of Biometry, University of Thessaly, N. Ionia, Magnesia, Greece*
- Mamoru Narukawa** / *Division of Pharmaceutical Medicine, Kitasato University Graduate School, Tokyo, Japan*
- Nagaraj K. Neerchal** / *Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, Maryland, U.S.A.*
- Kai W. Ng** / *Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong*
- Hoang Q. Nguyen** / *Department of Biostatistics, The University of Texas, M.D. Anderson Cancer Center, Houston, Texas, U.S.A.*
- Lei Nie** / *Division of Biometrics VI, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Baoming Ning** / *National Institutes for Food and Drug Control, Beijing, China*
- Hiroshi Nishiyama** / *Eli Lilly Japan K.K, Kobe, Japan*
- Clet Niyikiza** / *Eli Lilly and Company, Indianapolis, Indiana, U.S.A.*
- Earl Nordbrock** / *Nordbrock Consulting, Draper, Utah, U.S.A.*
- Maria Overbeck-Larisch** / *University of Applied Sciences, Darmstadt, Germany*
- Myunghee Cho Paik** / *Mailman School of Public Health, Columbia University, New York City, New York, U.S.A.*
- Misook Park** / *Office of Biostatistics, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Taesung Park** / *Seoul National University, Seoul, South Korea*

- Sofia Paul** / *Eli Lilly and Company, Carmel, Indiana, U.S.A.*
- Gene Pennello** / *Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Cynthia Pennise** / *Baxter Healthcare Corporation, New Providence, New Jersey, U.S.A.*
- Thomas Permutt** / *U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.*
- Allan Phillips** / *Wythe Research, Berkshire, U.K.*
- Phillip Phiri** / *Merck Research Laboratories, Summit, New Jersey, U.S.A.*
- Tamara Pinkett** / *Biostatistics, IQVIA, Durham, North Carolina, U.S.A.*
- Keya T. Pitts** / *Bristol-Myers Squibb Company, Princeton, New Jersey, U.S.A.*
- Marianne Plaunt** / *STATPROBE, Ann Arbor, Michigan, U.S.A.*
- Annpey Pong** / *Merck Research Laboratories, Rahway, New Jersey, U.S.A.*
- Bernice Porjesz** / *Department of Psychiatry, HSCB, State University of New York, Downstate Medical Center, Brooklyn, New York, U.S.A.*
- John S. F. Preisser** / *Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, North Carolina, U.S.A.*
- Edward F. C. Pun** / *Agouron Pharmaceuticals, Inc., San Diego, California, U.S.A.*
- Chunfu Qiu** / *Sanofi-Aventis Pharmaceuticals, Bridgewater, New Jersey, U.S.A.*
- Sam G. Raney** / *Office of Research and Standards, Office of Generic Drugs, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Elena Rantou** / *Office of Biostatistics, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- M. Mushfiqur Rashid** / *U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Erik Rasmussen** / *Amgen, Thousand Oaks, California, U.S.A.*
- Sarah J. Ratcliffe** / *Department of Biostatistics and Epidemiology, Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, U.S.A.*
- Stephane Robin** / *Applied Mathematics and Computer Sciences, INA-PG/INRA, Paris, France*
- Cindy Rodenberg** / *The Procter & Gamble Co., Cincinnati, Ohio, U.S.A.*
- William F. Rosenberger** / *Department of Statistics, George Mason University, Fairfax, Virginia, U.S.A.*
- Mark Rothmann** / *U.S. Food and Drug Administration, White Oaks, Maryland, U.S.A.*
- Valentin Rousson** / *University of Zurich, Zurich, Switzerland*
- Estelle Russek-Cohen** / *U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Mary D. Sammel** / *Department of Biostatistics and Epidemiology, Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, U.S.A.*
- Werner Sanns** / *University of Applied Sciences, Darmstadt, Germany*
- Nicola Sartori** / *Dipartimento di Statistica, Università Cà Foscari di Venezia, Venezia, Italy*
- Pradeep M. Sathe** / *U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.*
- Robert Schall** / *Quintiles ClinData, Bloemfontein, South Africa*
- Kenneth B. Schechtman** / *Washington University School of Medicine, St. Louis, Missouri, U.S.A.*

- Thomas H. Scheike** / *Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark*
- Enrique F. Schisterman** / *National Institute of Child Health and Human Development, Bethesda, Maryland, U.S.A.*
- Dan Schnell** / *The Procter & Gamble Co., Cincinnati, Ohio, U.S.A.*
- Timothy L. Schofield** / *Merck Research Laboratories, West Point, Pennsylvania, U.S.A.*
- Stephen Senn** / *Department of Statistics, University of Glasgow, Glasgow, U.K.*
- Vinod P. Shah** / *U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.*
- Jun Shao** / *University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.*
- Meiyu Shen** / *Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Xiaoming Sheng** / *University of Utah, Salt Lake City, Utah, U.S.A.*
- Jian Qing Shi** / *School of Mathematics and Statistics, Newcastle University, U.K.*
- Brian Smith** / *Eli Lilly and Company, Indianapolis, Indiana, U.S.A.*
- Steven Snapinn** / *Merck Research Laboratories, West Point, Pennsylvania, U.S.A.*
- Changhong Song** / *Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Fuyu Song** / *Center for Food and Drug Inspection, China Food and Drug Administration, Beijing, China; Peking University Health Science Center, Peking University Clinical Research Institute, Beijing, China*
- Catharine B. Stack** / *Glenmore, Pennsylvania, U.S.A.*
- Thomas R. Stiger** / *Pfizer Inc., Groton, Connecticut, U.S.A.*
- Elizabeth Stojanovski** / *School of Mathematics and Physical Sciences, University of Newcastle, Callaghan, New South Wales, Australia*
- Xiaogang Su** / *Department of Statistics and Actuarial Science, University of Central Florida, Orlando, Florida, U.S.A.*
- Jianguo Sun** / *University of Missouri, Columbia, Missouri, U.S.A.*
- Rolf Sundberg** / *Stockholm University, Stockholm, Sweden*
- Lev Sverdlov** / *Merck Research Laboratories, Summit, New Jersey, U.S.A.*
- Vladimir Svetnik** / *Biometrics Research, MRL, Rahway, New Jersey, U.S.A.*
- Masahiro Takeuchi** / *Kitasato University Graduate School, Tokyo, Japan*
- Susan Talbot** / *Global Biostatistical Sciences, Amgen Limited, Uxbridge, U.K.*
- Say Beng Tan** / *Singapore Clinical Research Institute, Centre of Quantitative Biology and Medicine, Duke-NUS Graduate Medical School, Singapore*
- Sze Huey Tan** / *Division of Clinical Trials & Epidemiological Sciences, National Cancer Centre, Singapore; Centre of Quantitative Biology and Medicine, Duke-NUS Graduate Medical School, Singapore*
- Yoko Tanaka** / *Eli Lilly and Company, Indianapolis, Indiana, U.S.A.*
- Dejun Tang** / *Novartis Pharmaceuticals, East Hanover, New Jersey, U.S.A.*
- Man Lai Tang** / *Department of Mathematics, Hong Kong Baptist University, Hong Kong*
- W. Tang** / *Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York, U.S.A.*
- Yongqiang Tang** / *Shire, Lexington, Massachusetts, U.S.A.*
- Zhongwen Tang** / *Novartis, East Hanover, New Jersey, U.S.A.*
- Zhaoyang Teng** / *Takeda Pharmaceuticals Inc., Cambridge, Massachusetts, U.S.A.*
- Peter F. Thall** / *Department of Biostatistics, The University of Texas, M.D. Anderson Cancer Center, Houston, Texas, U.S.A.*

- Laura Thompson** / *Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Naitee Ting** / *Department of Biostatistics and Data Sciences, Boehringer-Ingelheim Pharmaceuticals, Inc., Ridgefield, Connecticut, U.S.A.*
- J. Tiwari** / *U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.*
- Laszlo Tothfalusi** / *Department of Pharmacodynamics, Semmelweis University, Budapest, Hungary*
- William M. K. Trochim** / *Department of Policy Analysis and Management, Cornell University, Ithaca, New York, U.S.A.*
- Siu-Keung Tse** / *Department of Management Sciences, City University of Hong Kong, Hong Kong, China*
- Yi Tsong** / *Office of Biostatistics, Office of Translational Sciences, CDER, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Hsiao-Hui Tsou** / *Institute of Population Health Sciences, National Health Research Institutes, Miaoli, Taiwan*
- Dongsheng Tu** / *NCIC Clinical Trials Group, Queen's University, Kingston, Ontario, Canada*
- X. M. Tu** / *Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York, U.S.A.*
- Merlin L. Utter** / *Pfizer Inc., Pearl River, New York, U.S.A.*
- Mårten Vågerö** / *AstraZeneca R&D Södertälje, Södertälje, Sweden*
- K. Van Steen** / *Montefiore Institute – Systems and Modeling Unit; Bioinformatics and Modeling, Université de Liège, Liège; and Department of Human Genetics, Catholic University of Leuven, Leuven, Belgium*
- Julia A. Varshavsky** / *Eli Lilly and Company, Indianapolis, Indiana, U.S.A.*
- R. Lakshmi Vishnuvajjala** / *Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- S. D. Walter** / *Department of Clinical Epidemiology and Biostatistics, Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada*
- Hansheng Wang** / *Peking University, Beijing, People's Republic of China*
- Hongwei Wang** / *Merck & Co., Inc., Rahway, New Jersey, U.S.A.*
- Kongming Wang** / *Department of Statistics, Pfizer Inc., New York, New York, U.S.A.*
- Sue-Jane Wang** / *Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- William W. B. Wang** / *Merck Research Laboratories, West Point, Pennsylvania, U.S.A.*
- Xiaoming Wang** / *Shanghai University of Finance and Economics, Shanghai, China*
- Yong-Cheng Wang** / *Division of Biometrics I, Office of Biostatistics, Office of Pharmacoeconomics and Statistical Science, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.*
- Greg C. G. Wei** / *Pfizer Global Research and Development, Groton, Connecticut, U.S.A.*
- Thomas R. Weihrauch** / *Bayer AG, Wuppertal, Germany*
- Peter Westfall** / *Area of Information Systems and Quantitative Sciences, Texas Tech University, Lubbock, Texas, U.S.A.*
- Michael G. Wilson** / *Butler University, Indianapolis, Indiana, U.S.A.*
- Robert L. Wong** / *Abbott Laboratories, Parsippany, New Jersey, U.S.A.*
- Jiang-Ming Johnny Wu** / *Biostatistics, DOV Pharmaceutical, Inc., Somerset, New Jersey, U.S.A.*

- Jianrong Wu** / *Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee, U.S.A.*
- Jincao Wu** / *Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Rongling Wu** / *Departments of Public Health Science and Departments of Statistics, Pennsylvania State University, University Park, Pennsylvania, U.S.A.*
- Chien-Hua Wu** / *Department of Applied Mathematics, Chung-Yuan Christian University, Chung-Li, Taiwan*
- Yuh-Jenn Wu** / *Department of Applied Mathematics, Chung Yuan Christian University, Taoyuan, Taiwan*
- Liming Xiang** / *Nanyang Technology University, Singapore*
- Yaji Xu** / *Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Yunling Xu** / *Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Zhiheng Xu** / *Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Aileen L. Yam** / *Independent Consultant*
- Xin Yan** / *Merck Research Laboratories, Blue Bell, Pennsylvania, U.S.A.*
- Xu Yan** / *Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Bo Yang** / *Biometrics, Vertex Pharmaceuticals, Boston, Massachusetts, U.S.A.*
- Chunyan Yang** / *Hong Kong Baptist University, Hong Kong; and Yunnan University, Kunming, Yunnan, China*
- Jing Yang** / *Gilead Sciences, Foster City, California, U.S.A.*
- Harry Yang** / *MedImmune, LLC, Gaithersburg, Maryland, U.S.A.*
- Jingjing Ye** / *Division of Biometrics V, Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Laura Yee** / *Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Constantin T. Yiannoutsos** / *Division of Biostatistics, Indiana University School of Medicine, Indianapolis, Indiana, U.S.A.*
- Lisa Ying** / *Bristol-Myers Squibb Company, Princeton, New Jersey, U.S.A.*
- Gloria H. Yiu** / *The Procter & Gamble Co., Cincinnati, Ohio, U.S.A.*
- Tinghui Yu** / *Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Guojun Yuan** / *LFB USA, Inc., Framingham, Massachusetts, U.S.A.*
- Mengdie Yuan** / *Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Lilly Q. Yue** / *Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Hak-Keung Yuen** / *Department of Management Sciences, City University of Hong Kong, Hong Kong, China*

- Langche Zeng** / *Department of Political Science, University of California, San Diego, California, U.S.A.*
- Na Zeng** / *National Clinical Research Center for Digestive Disease, Beijing Friendship Hospital, Capital Medical University, Beijing, China*
- Donghui Zhang** / *Department of Biostatistics and Programming, Sanofi US Inc., Bridgewater, New Jersey, U.S.A.*
- J. Zhang** / *Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, U.S.A.*
- Lanju Zhang** / *Data and Statistical Science, AbbVie Inc., North Chicago, Illinois, U.S.A.*
- Shu Zhang** / *Sepracor, Inc., Marlborough, Massachusetts, U.S.A.*
- Ying Zhang** / *AIG American General, Neptune, New Jersey, U.S.A.*
- Zhiwei Zhang** / *Department of Statistics, University of California, Riverside, Riverside, California, U.S.A.*
- Joanne Zhang** / *Office of Biostatistics, Office of Translational Sciences, CDER, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Lanju Zhang** / *Data and Statistical Sciences, AbbVie Inc., North Chicago, Illinois, U.S.A.*
- Jun Zhao** / *Merck Research Laboratories, Summit, New Jersey, U.S.A.*
- Zhigen Zhao** / *Department of Statistics, Temple University, Philadelphia, Pennsylvania, U.S.A.*
- Gang Zheng** / *National Heart, Lung and Blood Institute, Bethesda, Maryland, U.S.A.*
- Jiayin Zheng** / *Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, U.S.A.*
- Yanbing Zheng** / *Data and Statistical Sciences, AbbVie Inc., North Chicago, Illinois, U.S.A.*
- Bob Zhong** / *Janssen Pharmaceutical Companies, Johnson and Johnson, Malvern, Pennsylvania, U.S.A.*
- Jinglin Zhong** / *Division of Biometrics IV, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*
- Jihao Zhou** / *Department of Biostatistics, Allergan, Inc., Irvine, California, U.S.A.*
- Qingning Zhou** / *University of Missouri, Columbia, Missouri, U.S.A.*
- Yijie Zhou** / *Data and Statistical Science, AbbVie Inc., North Chicago, Illinois, U.S.A.*
- Li Zhu** / *Amgen Inc., Thousand Oaks, California, U.S.A.*



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Contents

<i>Contributors</i>	xiii
<i>Topical Table of Contents</i>	xxxv
<i>Preface</i>	xlvi
<i>About the Editor-in-Chief</i>	xlix

Volume I

Acceptance Sampling / William A. Brenneman and William R. Myers	1
Active Control Non-Inferiority Trials: Hypotheses and False Positive Rate / Gang Chen, George Y. H. Chi, and Yong-Cheng Wang	8
Active Control Trials / George Y. H. Chi, Gang Chen, Mark Rothmann, and Ning Li	14
Active Controlled Clinical Trials: Noninferiority Analysis / Sue-Jane Wang, H. M. James Hung, and Yi Tsong	23
Adaptive Design Clinical Trials: Case Studies / Shiohjen Lee and Min Lin	28
Adaptive Design Methods in Clinical Trials / Shein-Chung Chow and Annpey Pong	36
Adaptive Survival Trials / William F. Rosenberger and Lanju Zhang	50
Adverse Event Reporting / Cynthia Pennise	53
Alpha Spending Function / Jihao Zhou and Glen Andrews	59
Ames Test / Wherly P. Hoffman and Michael L. Garriott	67
Analysis of $2 \times K$ Tables / Shiva Gautam	75
Analysis of Clustered Binary Data / Valerie Durkalski	82
Analysis of Clustered Categorical Data / Kalyan Das	88
Analysis of Heritability / Brent D. Burch and Ian R. Harris	98
Analysis of Variance (ANOVA) / Maria Overbeck-Larisch and Werner Sanns	104
Analytical Similarity Assessment / Shein-Chung Chow, Fuyu Song, and He Bai	116
ANCOVA Approach: Premarketing Shelf Life Determination with Multiple Factors / Yi Tsong, Chi-wan Chen, Wen Jen Chen, Roswitha E. Kelly, Daphne T. Lin, and Karl K. Lin	125
Assay Development / Timothy L. Schofield	130
Assay Validation / Timothy L. Schofield	138
Assay Validation: Evaluation of Linearity / Jen-pei Liu and Eric Hsieh	146
Bayesian Designs: Phase II Oncology Clinical Trials / Sze Huey Tan, David Machin, and Say Beng Tan	154
Bayesian Estimate: Concordance Correlation Coefficient / Dai Feng, Richard Baumgartner, and Vladimir Svetnik	164
Bayesian Methods: Meta-Analysis / Elizabeth Stojanovski and Kerrie L. Mengersen	173
Bayesian Methods: Stability Analysis / Jie Chen, Jinglin Zhong, and Lei Nie	180
Bayesian Model: Prior Effective Sample Size / Satoshi Morita and Peter F. Thall	186
Bayesian Statistics / Julia A. Varshavsky and Stacy R. Lindborg	190
Bayesian Statistics: Medical Device Clinical Trials / Gerry W. Gray, Xuefeng Li, Laura Thompson, Yunling Xu, Xu Yan, and Lilly Q. Yue	198
Binary 2×2 Crossover Trials / Li Zhu, Juan Li, and Eric Chi	206
Bioassay / Anant M. Kshirsagar	213
Bioavailability and Bioequivalence / Mengdie Yuan and Jingyu Julia Luan	216
Bioinformatics / Sean D. Mooney and Jeffrey T. Chang	224

Volume I (cont'd.)

Biologics / P. A. Lachenbruch, A. D. Horne, C. J. Lynch, J. Tiwari, and Susan S. Ellenberg	233
Biomarker Development with Applications to Genetics Data: Statistical Tests / Gang Zheng and Jungnam Joo	243
Biomarker: Clinical Trials / Mark Chang	251
Biopharmaceutics / Iain J. McGilveray	254
Biosimilar Studies: Sample Size / Seung-Ho Kang	274
Biosimilar Studies: Three-Arm Design / Seung-Ho Kang	277
Biosimilarity Assessment / Shein-Chung Chow and Fuyu Song	281
Biosimilarity of Follow-On Biologics / Shein-Chung Chow and Jen-pei Liu	304
Bivariate Zero-Inflated Poisson Populations: Statistical Test for Homogeneity / Siu-Keung Tse and Hak-Keung Yuen	316
Blinded Sample Size Re-Estimation / Chien-Hua Wu	323
Blinding / Shein-Chung Chow and Jun Shao	336
Bootstrap, The / Shein-Chung Chow and Jun Shao	342
Botanical Drug Product Development: Scientific Issues / Annpey Pong and Shein-Chung Chow	348
Bracketing Design / John R. Murphy	359
Bridging Studies / Jen-pei Liu	361
Bridging Studies: Statistical Methods / Yuh-Jenn Wu, Chieh Chiang, Chi-Tian Chen, Hsiao-Hui Tsou, and Chin-Fu Hsiao	366
Calibration / Shein-Chung Chow	370
Canadian Health Products and Food Branch (HPFB) and Therapeutic Products Directorate (TPD) / Dongsheng Tu	378
Cancer Chemotherapy: Maximum Tolerable Dose / Jen-pei Liu	384
Cancer Trials / Clet Niyikiza and Douglas E. Faries	390
Carcinogenicity Studies of Pharmaceuticals / Karl K. Lin	398
Carry-Forward Analysis / Naitee Ting	416
Case-Control Studies: Inference in / Gary King and Langche Zeng	422
Center Weighting: Multicenter Trials / Paul Gallo	435
Clinical Data Management / Terri Madison and Marianne Plaunt	444
Clinical Endpoint / Sofia Paul	451
Clinical Endpoint BE Studies: Generic Drugs / Stella Grosser, Misook Park, Sam G. Raney, and Elena Rantou	455
Clinical Inspection: Statistical Process / Fuyu Song, Minghuang Hong, and Shein-Chung Chow	462
Clinical Pharmacology / Brian Smith	469
Clinical Research: Benefit and Risk Assessment / Susan Talbot, Shahrul Mt-Isa, Qi Jiang, and Chunlei Ke	472
Clinical Research: Comparing Variabilities / Yonghee Lee, Hansheng Wang, and Shein-Chung Chow	486
Clinical Research: Outlier Detection / Siu-Keung Tse and Liming Xiang	502
Clinical Research: Reproducibility Probability / Shein-Chung Chow	510
Clinical Trial Design: Statistical Assurance / Shuyen Ho	523
Clinical Trial Designs with Prospective Patient Population Enrichment by Response / Anastasia Ivanova, Marc DeSomer, and Jurgen Hummel	526
Clinical Trial Process / Paul F. Kramer	532
Clinical Trial Simulation / Hung-Ir Li and Pan-Yu Lai	536
Clinical Trial Simulations: Earlier Development Phases / Mark Chang	539
Clinical Trial Simulations: Later Development Phases / Mark Chang	544

Clinical Trial: Combination Drug Development / <i>H. M. James Hung</i>	552
Clinical Trial: Failure-Time Model / <i>Chin-Shang Li</i>	556
Clinical Trial: N-of-1 Design Analysis / <i>Can Cui and Shein-Chung Chow</i>	564
Clinical Trials / <i>Kenneth B. Schechtman</i>	572
Clinical Trials: Controversial Issues / <i>Shein-Chung Chow and Fuyu Song</i>	580

Volume II

Clinical Trials: Interval-censored Failure Time Data / <i>Jianguo Sun, Qingning Zhou, and Ding-Geng (Din) Chen</i>	589
Clinical Trials: Investigating Quality-of-Life / <i>Patricia B. Cerrito</i>	597
Clinical Trials: Meta-Analysis using R / <i>Ding-Geng (Din) Chen</i>	603
Clinical Trials: Response-Adaptive Repeated Measurement Designs / <i>Keumhee Carriere Chough and Yuanyuan Liang</i>	612
Clinical Trials: Selection of Control / <i>Irving K. Hwang</i>	622
Clinical Trials: Vaccine Development / <i>Ivan S. F. Chan, William W. B. Wang, and Joseph Heyse</i> ...	628
Cluster Randomized Trials: Sample Size / <i>Sin-Ho Jung</i>	645
Cluster Trials / <i>John S. F. Preisser and Bing Lu</i>	648
Clustered Study Designs: Power Analysis / <i>J. Zhang, C. Feng, W. Tang, X. M. Tu, and J. Kowalski</i> ..	658
Companion Diagnostics / <i>Gene Pennello and Jingjing Ye</i>	672
Comparability Studies: Statistical Considerations / <i>Jason J.Z. Liao, Li He, and Robert Capen</i> ...	682
Confidence Interval and Hypothesis Testing / <i>Ivan S. F. Chan and Devan V. Mehrotra</i>	699
Confounding and Interaction / <i>Catharine B. Stack</i>	703
Content Uniformity / <i>John R. Murphy</i>	709
Continual Reassessment Methods / <i>Xue Lin</i>	712
Contract Research Organization (CRO): Early 2000s / <i>T. Y. Lee</i>	716
Contract Research Organization (CRO) / <i>Tamara Pinkett and Jill Dawson</i>	721
Controlled Clinical Trials: P-Values, Evidence, and Multiplicity Considerations / <i>Mohammad F. Huque, Mohamed Alos, Satya D. Dubey, and Rafia Bhore</i>	725
Correlated Probit Model / <i>Ralitza V. Gueorguieva</i>	741
Cost-Effectiveness Analysis / <i>Heejung Bang</i>	751
Covariate-Adjusted Adaptive Dose-Finding: Early Phase Clinical Trials / <i>Peter F. Thall and Hoang Q. Nguyen</i>	760
Covariates: Adjustment for / <i>Thomas Permutt</i>	772
Crossover Design / <i>Stephen Senn</i>	776
Crossover Design: Rank-Based Robust Analysis / <i>M. Mushfiqur Rashid</i>	784
Cutoff Designs / <i>Joseph C. Cappelleri and William M. K. Trochim</i>	798
Data Mining and Biopharmaceutical Research / <i>Patricia B. Cerrito</i>	805
Data Monitoring Committees (DMC) / <i>Jay Herson</i>	811
Diagnostic Device Evaluation / <i>Meijuan Li, Tinghui Yu, Qin Li, Gene Pennello, R. Lakshmi Vishnuvajjala, Bipasa Biswas, Changhong Song, and Lilly Q. Yue</i>	822
Diagnostic Imaging / <i>Suming W. Chang and Annpey Pong</i>	829
Diagnostic Procedure: Sensitivity and Specificity / <i>Jiayin Zheng and Shein-Chung Chow</i>	836
Disease Modification Effects: Design and Analysis / <i>Djokouri A. Kouassi and Annpey Pong</i>	839
Dissolution Profile Comparison: Statistical Methods / <i>Harry Yang</i>	844
Dissolution Profiles Comparison: Model-Independent Approaches / <i>Yanbing Zheng and Lanju Zhang</i>	854
Dosage Units Using Small and Large Sample Sizes: Uniformity / <i>Meiyu Shen and Yi Tsong</i>	859
Dose Finding Clinical Trials Using R / <i>Naitee Ting and Ding-Geng (Din) Chen</i>	866

Volume II (cont'd.)

Dose Proportionality / <i>C. Gordon Law</i>	876
Dose Response Analysis in Clinical Trials / <i>James MacDougall</i>	879
Dose Response Study Design / <i>Naitee Ting</i>	885
Dose-Ranging Studies: Design Considerations / <i>Qiqi Deng and Naitee Ting</i>	891
Dropout / <i>Cindy Rodenberg, Dan Schnell, and Cameron S. Liu</i>	901
Drug Development / <i>Naitee Ting</i>	908
Drug Interchangeability: Biosimilar Products / <i>Laszlo Endrenyi, Shein-Chung Chow, and Laszlo Tothfalusi</i>	914
Drug Interchangeability: Criteria and Design / <i>Shein-Chung Chow, Fuyu Song, and Meng Chen</i> . .	922
Drug Interchangeability: Meta-Analysis for Safety Monitoring / <i>Wen-Wei Liu and Shein-Chung Chow</i>	928
Drug Interchangeability: Generic Products, Prescribability and Switchability / <i>Laszlo Endrenyi, Shein-Chung Chow, and Laszlo Tothfalusi</i>	935
Ecologic Inference / <i>Sander Greenland</i>	940
ED₅₀/ED₉₀ / <i>Yangxin Huang</i>	952
Enrichment Design / <i>Jen-pei Liu</i>	960
Enrichment Design with Patient Population Augmentation / <i>Yijie Zhou, Bo Yang, Lanju Zhang, and Lu Cui</i>	964
Equivalence Trials / <i>Jen-pei Liu</i>	970
Ethnic Factors / <i>Melody H. Lin and Helen McGough</i>	976
Expiration Dating Period / <i>Annpey Pong</i>	979
Exploratory Factor Analysis / <i>Joseph C. Cappelleri and Robert A. Gerber</i>	985
Extra Variation Models / <i>Jorge G. Morel and Nagaraj K. Neerchal</i>	992
Factor Analysis / <i>Mary D. Sammel, Sarah J. Ratcliffe, and Benjamin E. Leiby</i>	1008
Factorial Designs / <i>Junfang Li</i>	1014
False Discovery Rate (FDR) / <i>Avner Bar-Hen, Philippe Broet, Cyril Dalmasso, Jean-Jacques Daudin, and Stephane Robin</i>	1028
GEE in Cancer Research / <i>Sin-Ho Jung</i>	1036
Generalizability Probability in Clinical Research / <i>Shein-Chung Chow</i>	1042
Generalized Estimating Equation / <i>Myunghee Cho Paik and Hye-Seung Lee</i>	1046
Generalized Inference / <i>Chen-Tuo Liao and Chi-Rong Li</i>	1054
Genetic Linkage and Linkage Disequilibrium Analysis / <i>Kongming Wang, Bernice Porjesz, Henri Begleiter, and Kevin Jones</i>	1057
Global Database and System / <i>Alice T. M. Hsuan and Patrick Genyn</i>	1069
Good Clinical Practice (GCP) / <i>Mamoru Narukawa and Masahiro Takeuchi</i>	1075
Good Laboratory Practice (GLP) / <i>Ying Geng, Baoming Ning, and Dejiang Tan</i>	1082
Good Programming Practice (GPP) / <i>Aileen L. Yam</i>	1088
Good Statistics Practice (GSP) / <i>Shein-Chung Chow</i>	1095
Good Validation Practice (GVP) / <i>Ying Geng and Lan He</i>	1101
Group Sequential Methods / <i>Cheng-Tao Chang</i>	1105
Group Sequential Tests and Variance Heterogeneity:	
Clinical Trials / <i>Keumhee Carriere Chough, Abdulkadir Hussein, and Taesung Park</i>	1114
Human Drug Abuse Trials / <i>Qianyu Dang</i>	1121
Hypothesis Testing / <i>Devan V. Mehrotra and Ivan S. F. Chan</i>	1124
Imputation with Item Nonrespondents / <i>Hansheng Wang and Shein-Chung Chow</i>	1128
Imputation: Clinical Research / <i>Hansheng Wang and Shein-Chung Chow</i>	1135

<i>In Vitro</i> Testing: Bioequivalence / Hansheng Wang, Ying Zhang, Jun Shao, and Shein-Chung Chow . .	1141
<i>In Vitro</i> Testing: Dissolution Profile Comparison / Yi Tsong, Pradeep M. Sathe, and Vinod P. Shah	1148
<i>In Vitro</i> Testing: Micronucleus Test / Wherly P. Hoffman, Michael L. Garriott, and Cindy Lee . . .	1155
Individual Bioequivalence / Shein-Chung Chow and Jen-pei Liu	1160

Volume III

Instrument Development and Validation / Cindy Rodenberg, James T. Kuznicki, and Gloria H. Yiu	1167
Integrated Summary Report / Laura J. Meyerson	1179
Intention-to-Treat Analyses (ITT) / Ronald K. Knickerbocker.	1183
Interchangeability: Generic Drugs / Jiayin Zheng, Shein-Chung Chow, and Mengdie Yuan.	1187
Interim Analysis / Paul Gallo	1195
International Conference on Harmonization (ICH) / Jen-pei Liu and Shein-Chung Chow.	1202
Item Response Theory / Wen-Hung Chen and Joseph C. Cappelleri	1207
IVRS/IWRS for Randomization / Mon-Gy Chen	1216
Japan: Ministry of Health, Labour and Welfare and Pharmaceutical Administration / Kiyohito Nakai	1224
Kaplan–Meier Estimator / Hongyu Jiang and Jason Fine	1231
Kappa Coefficients: Medical Research / Helena Chmura Kraemer.	1237
Kruskal–Wallis Test for Ordered Categorical Data: Power Calculation / Chunpeng Fan and Donghui Zhang	1245
Kullback–Leibler Divergence for Evaluating Equivalence / Vladimir Dragalin	1249
Laboratory Analyses / Michael J. Klepper.	1258
Latent Class Analysis / Elizabeth S. Garrett-Mayer and Jeannie-Marie S. Leoutsakos.	1266
Lilly Reference Ranges / Michael G. Wilson	1276
Local Influence Analysis / Nicola Sartori.	1292
LOCF: Validity / Bin Cheng and Shein-Chung Chow	1296
Logistic Regression / Rick Chappell and Jason Fine	1303
Logistic Regression Process: Predictive Model Building in Clinical Research / Mo Liu and Shein-Chung Chow	1309
Logistic Regression: Three-Point Designs / Mårten Vågerö and Rolf Sundberg	1315
McNemar’s Test / Robert G. Lehr.	1320
MCP-Mod: Dose–Response Testing and Estimation / Björn Bornkamp and Naitee Ting	1326
Measuring Agreement / Lawrence I-Kuei Lin	1335
MedDRA: Impact on Pharmaceutical Development / Keya T. Pitts	1342
Medical Devices / Greg Campbell, Lilly Q. Yue, Gene Pennello, and Mary K. Barrick	1348
Meta-Analysis of Therapeutic Trials / Joseph C. Cappelleri, John P. A. Ioannidis, and Joseph Lau. . .	1362
Microarray Gene Expression / James J. Chen and Chun-Houh Chen	1377
Microdosing Studies / Fuyu Song and Shein-Chung Chow	1394
Minimization Procedure / Dongsheng Tu.	1401
Minimum Effective Dose / Keumhee Carriere Chough.	1406
Mixed Effects Models / Donghui Zhang, Chunpeng Fan, and A. Lawrence Gould	1409
Mixed-Outcome Data / A. R. de Leon.	1414
MMRM with Missing Data / Annpey Pong, Jun Zhao, Allen Lee, Phillip Phiri, and Lev Sverdlov. . .	1422
Modified Large Sample Method / Yonghee Lee.	1427
Multicenter Trials / William J. Huster	1434
Multicollinearity / K. Van Steen and G. Molengerghs	1438

Volume III (cont'd.)

Multidimensional Data Analysis: Penalized Regression Methods / Keumhee Carriere Chough and Xiaoming Wang	1448
Multinational Clinical Trials / Thomas R. Weihrauch	1454
Multiple Comparisons / Mani Y. Lakshminarayanan	1461
Multiple Endpoints / Mario Comelli	1467
Multiple-Dose Bioequivalence Studies / Vernon M. Chinchili	1477
Multiple-Stage Designs: Phase II Cancer Trials / Masha Kocherginsky and Shang P. Lin	1481
Multiplicative Intensity Models / Keumhee Carriere Chough and Abdulkadir Hussein	1494
Multiplicity in Clinical Trials / Peter Westfall and Frank Bretz	1502
Multi-Regional Clinical Trials: Consistency Test / Jiayin Zheng, Lisa Ying, Na Zeng, and Shein-Chung Chow	1511
Multiregional Clinical Trials: Qualitative Consistency of Treatment Effects / Yoko Tanaka and Hiroshi Nishiyama	1522
Multivariate Meta-Analysis / Elizabeth Stojanovski and Kerrie L. Mengersen	1528
Noninferiority Trials: Type I Error Rates / Seung-Ho Kang	1536
Nonparametric Regression / Kongming Wang and Theo Gasser	1543
Odds Ratio / Dongsheng Tu	1554
Oncology Clinical Trials: Efficient Dose-Finding Designs / Erik Rasmussen, Qi Jiang, and Chunlei Ke	1560
Oncology Trials: Development Strategy / Zhaoyang Teng, Yi Liu, Jing Yang, Zhaowei Hua, and Mark Chang	1572
Onset of Action / Thomas R. Stiger and Christy Chuang-Stein	1585
Optimal Futility Interim Design / Zhongwen Tang	1593
Ordered Categorical Data: Test for / Shein-Chung Chow, Siu-Keung Tse, and Chunyan Yang	1602
Ordered Multiple Class Receiver Operating Characteristic (ROC) Analysis / Christos T. Nakas and Constantin T. Yiannoutsos	1608
Ordinal Data: Crossover Design Analysis / Kung-Jong Lui	1613
Parallel Design / Mani Y. Lakshminarayanan	1621
Patient Compliance / Kenneth B. Schechtman	1627
Patient-Reported Outcomes / Amylou C. Dueck and Joseph C. Cappelleri	1633
Percentile Charts on Correlated Measures / Xuming He and Kai W. Ng	1643
Personalized Medicine: Validation of Diagnostic Medical Devices / Meijuan Li, Gene Pennello, Jincao Wu, Laura Yee, Zhiwei Zhang, Gerry Gray, Yuying Jin, Yaji Xu, Jingjing Ye, and Zhiheng Xu	1648
Pharmaceutical Development: Statistical Designs / Annpey Pong and Shein-Chung Chow	1656
Pharmacodynamic Issues / Cheng-Tao Chang and Robert L. Wong	1665
Pharmacodynamics with Covariates / Cheng-Tao Chang and Robert L. Wong	1673
Pharmacodynamics with No Covariates / Cheng-Tao Chang and Robert L. Wong	1685
Pharmacoeconomics / Joseph Heyse, John R. Cook, and Michael F. Drummond	1693
Pharmacoeconomics: Cost-Effectiveness Analysis / Shahram Heshmat	1707
Phase I Cancer Clinical Trials / Seung-Ho Kang and Chul Ahn	1715
Phase II Clinical Development / Naitee Ting and Guojun Yuan	1721
Phase III Clinical Trials with Response-Adaptive Allocation / Atanu Biswas and Rahul Bhattacharyya	1727
Placebo Effect / Thomas R. Stiger and Edward F. C. Pun	1735
Population Bioequivalence / Naitee Ting, Greg C. G. Wei, and William W. B. Wang	1740

Volume IV

Population PK/PD Analysis / <i>D. Concordet, C. Ané, and F. Léger</i>	1747
Postmarketing Adverse Drug Event Signaling / <i>Yi Tsong</i>	1756
Postmarketing Surveillance / <i>Patricia B. Cerrito</i>	1769
Power / <i>Kenneth B. Schechtman</i>	1777
Precision Medicine: Statistical Issues / <i>Fuyu Song and Shein-Chung Chow</i>	1779
Prediction Trees / <i>Antonio Ciampi</i>	1782
Premarket Applications: Registries / <i>Nelson Lu, Yunling Xu, and Lilly Q. Yue</i>	1791
Principal Component Analysis / <i>Aiyi Liu and Enrique F. Schisterman</i>	1796
Process Validation / <i>James S. Bergum and Merlin L. Utter</i>	1802
Profile Analysis / <i>Bin Cheng and Shein-Chung Chow</i>	1817
Propensity Score Analysis and Its Application in Regulatory Settings / <i>Lilly Q. Yue</i>	1823
Propensity Score: Methodology and Application to Clinical Studies / <i>Lilly Q. Yue and Heng Li</i>	1825
Proportion of Treatment Effect / <i>Cong Chen and Hongwei Wang</i>	1830
Proportional Hazards Regression Model / <i>Bee Leng Lee and Jason Fine</i>	1834
Protocol Development / <i>Robert D. Chew</i>	1841
P-Values / <i>Stephen Senn</i>	1845
QT Analysis / <i>Lang Li, Stephen Hall, and Mehul Desai</i>	1855
Quality by Design in Biopharmaceutical Development / <i>Harry Yang</i>	1865
Randomization / <i>Shein-Chung Chow and Jun Shao</i>	1875
Randomized Trials with Clusters: Design and Analysis / <i>Sin-Ho Jung</i>	1880
Rank Regression: Stability Analysis / <i>Annpey Pong, Ying Qing Chen, and Xiudong Lei</i>	1888
Regulatory Statistics / <i>Estelle Russek-Cohen, Tsai-Lien Lin, and Shiohwen Lee</i>	1893
Release Targets / <i>Greg C. G. Wei</i>	1903
Reliability / <i>Valentin Rousson and Theo Gasser</i>	1908
Repeated Measurement Designs: Missing Values / <i>Keumhee Carriere Chough, Yuanyuan Liang, Rong Huang, and Xiaoming Sheng</i>	1918
Repeated Measures Data with Missing Values Analysis: Overview of	
Methods / <i>Keumhee Carriere Chough, Yuanyuan Liang, and Taesung Park</i>	1924
Reproductive and Developmental Studies / <i>James J. Chen</i>	1931
Response Surface Methodology / <i>William R. Myers</i>	1939
Response-Adaptive Designs / <i>Feifang Hu and Anastasia Ivanova</i>	1950
Risk Ratio Analysis / <i>Kung-Jong Lui</i>	1956
ROC Curve / <i>Robert G. Lehr and Annpey Pong</i>	1962
Sample Size Calculation: Nonparametrics / <i>Hansheng Wang and Shein-Chung Chow</i>	1970
Sample Size Calculation: Survival Data / <i>Qi Jiang, Steven Snapinn, and Boris Iglewicz</i>	1980
Sample Size Determination / <i>Lilly Q. Yue, David Li, and Shan Bai</i>	1987
Sample Size Determination: Three-arm Ratio of Mean Differences Equivalence Test / <i>Yu-Wei Chang, Yi Tsong, Xiaoyu Dong, and Zhigen Zhao</i>	2002
Sample Size Estimation: Generalized Estimating Equations (GEE) Method / <i>Sin-Ho Jung</i>	2010
Sample Size for Multiregional Clinical Trials / <i>Yuh-Jenn Wu, Chieh Chiang, Chi-Tian Chen, Hsiao-Hui Tsou, and Chin-Fu Hsiao</i>	2015
Sample Size Re-estimation Based on Observed Treatment Difference / <i>Lu Cui</i>	2021
Sample Size Re-estimation Design with Robust Power / <i>Lu Cui</i>	2026
Sample Size: Negative Binomial With Unequal Follow-Up Times / <i>Yongqiang Tang</i>	2031
Scaled Average Bioequivalence (SABE) / <i>Laszlo Endrenyi and Laszlo Tothfalusi</i>	2038
Screening Design / <i>John R. Murphy</i>	2045

Volume IV (cont'd.)

Seamless Adaptive Trial Design: Dose Selection Criteria / Jiayin Zheng and Shein-Chung Chow . . .	2047
Semi-Parametric Time-Varying Regression Models / Thomas H. Scheike	2058
Sequential Estimation: Additive Hazards Rate Model with Staggered Entry / Laurent Bordes and Christelle Breuils	2063
Signal Detection in Pharmacovigilance / Patrick J. Farrell, Christopher A. Gravel, and Daniel Krewski	2068
Slope Approach: Assessment of Dose Proportionality and Linearity Under a Crossover Design / Bin Cheng and Shein-Chung Chow	2084
Spatio-Temporal Modeling / Ying C. MacNab and Charmaine B. Dean	2089
Specifications / John R. Murphy	2099
SROC Curve / S. D. Walter and Petra Macaskill	2101
Stability Analysis: Frozen Drug Products / Shein-Chung Chow and Jun Shao	2111
Stability Matrix Designs / Earl Nordbrock	2116
Statistical Genetics / Rongling Wu, Guifang Fu, and Hongying Li	2122
Statistical Principles for Clinical Trials / Allan Phillips	2130
Statistical Process Control / William R. Myers and William A. Brennenman	2135
Statistical Significance / Dieter Hauschke, Robert Schall, and Herman G. Luus	2142
Structural Equation Model / Joseph L. Carrasco	2147
Stuart–Maxwell Test / Jiang-Ming Johnny Wu	2154
Subgroup Analysis / Patricia B. Cerrito	2159
Subject-Treatment Interaction / Gary L. Gadbury	2167
Surrogate Endpoint / Shu Zhang and Edward F. C. Pun	2174
Survival Analysis / Dirk F. Moore	2179
Targeted Clinical Trials / Jen-pei Liu	2184
Testing for Qualitative Interaction / Xin Yan and Xiaogang Su	2192
Therapeutic Equivalence / Jen-pei Liu	2199
Thorough QT Studies / Joanne Zhang, Dalong (Patrick) Huang, and Yi Tsong	2206
Tiered Approach of Analytical Similarity Assessment: Statistical Methods / Yi Tsong	2219
Titration Design / Marilyn A. Agin and Edward F. C. Pun	2225
Toxicological Studies / Guenther Baus and Thomas Hoffman	2230
Traditional Chinese Medicine (TCM): Clinical Development / Fuyu Song and Shein-Chung Chow	2244
Traditional Chinese Medicine (TCM): General Considerations / Shein-Chung Chow, Annpey Pong, and Yu-Wei Chang	2252
Traditional Chinese Medicine (TCM): Unified Approach for Evaluation / Bin Cheng and Shein-Chung Chow	2265
Translational Medicine: Concepts, Statistical Methods, and Related Issues / Siu-Keung Tse and Shein-Chung Chow	2270
Trend Estimation / Jian Qing Shi and Man Lai Tang	2283
Two-Stage Adaptive Design: Analysis / Shein-Chung Chow and Min Lin	2289
Two-Stage Design: Phase II Cancer Clinical Trials / Sin-Ho Jung	2296
US Food and Drug Administration / Stacy R. Lindborg and Susan S. Ellenberg	2305
USP Tests / John R. Murphy	2311
Validation of Quantitative and Qualitative Assays / Bob Zhong, Chunfu Qiu, and Dejun Tang . . .	2313
Weibull Model: Group Sequential Design / Jianrong Wu	2325
Z-Score / Jiang-Ming Johnny Wu	2329

Topical Table of Contents

Regulatory Requirement

Regulatory Agency

Canadian Health Products and Food Branch (HPFB) and Therapeutic Products Directorate (TPD) / Dongsheng Tu.	378
US Food and Drug Administration / Stacy R. Lindborg and Susan S. Ellenberg	2305
International Conference on Harmonization (ICH) / Jen-pei Liu and Shein-Chung Chow.	1202
Japan: Ministry of Health, Labour and Welfare and Pharmaceutical Administration / Kiyohito Nakai	1224

Good Pharmaceutical Practices

Data Monitoring Committees (DMC) / Jay Herson	811
Good Clinical Practice (GCP) / Mamoru Narukawa and Masahiro Takeuchi	1075
Good Laboratory Practice (GLP) / Ying Geng, Baoming Ning, and Dejiang Tan	1082
Good Statistics Practice (GSP) / Shein-Chung Chow	1095
Good Programming Practice (GPP) / Aileen L. Yam	1088
Good Validation Practice (GVP) / Ying Geng and Lan He	1101
Regulatory Statistics / Estelle Russek-Cohen, Tsai-Lien Lin, and Shiohjen Lee	1893

Pre-Marketing

Development

Premarket Applications: Registries / Nelson Lu, Yunling Xu, and Lilly Q. Yue.	1791
--	------

Statistical Concepts

Alpha Spending Function / Jihao Zhou and Glen Andrews.	59
Bayesian Statistics / Julia A. Varshavsky and Stacy R. Lindborg.	190
Bootstrap, The / Shein-Chung Chow and Jun Shao	342
Clinical Research: Outlier Detection / Siu-Keung Tse and Liming Xiang	502
Clinical Trial Designs with Prospective Patient Population Enrichment by Response / Anastasia Ivanova, Marc DeSomer, and Jurgen Hummel.	526
Clinical Trials: Selection of Control / Irving K. Hwang	622
Confidence Interval and Hypothesis Testing / Ivan S. F. Chan and Devan V. Mehrotra	699
Confounding and Interaction / Catharine B. Stack.	703
Continual Reassessment Methods / Xue Lin	712
Controlled Clinical Trials: P-Values, Evidence, and Multiplicity Considerations / Mohammad F. Huque, Mohamed Alosch, Satya D. Dubey, and Rafia Bhore	725
Dropout / Cindy Rodenberg, Dan Schnell, and Cameron S. Liu.	901
Extra Variation Models / Jorge G. Morel and Nagaraj K. Neerchal.	992

Statistical Concepts (cont'd.)

False Discovery Rate (FDR) / Avner Bar-Hen, Philippe Broet, Cyril Dalmasso, Jean-Jacques Daudin, and Stephane Robin	1028
Generalizability Probability in Clinical Research / Shein-Chung Chow	1042
Generalized Estimating Equation / Myunghee Cho Paik and Hye-Seung Lee	1046
Generalized Inference / Chen-Tuo Liao and Chi-Rong Li	1054
Hypothesis Testing / Devan V. Mehrotra and Ivan S. F. Chan	1124
Integrated Summary Report / Laura J. Meyerson	1179
Intention-to-Treat Analyses (ITT) / Ronald K. Knickerbocker	1183
Interim Analysis / Paul Gallo	1195
Item Response Theory / Wen-Hung Chen and Joseph C. Cappelleri	1207
Kaplan–Meier Estimator / Hongyu Jiang and Jason Fine	1231
Kappa Coefficients: Medical Research / Helena Chmura Kraemer	1237
Kruskal-Wallis Test for Ordered Categorical Data: Power Calculation / Chunpeng Fan and Donghui Zhang	1245
Kullback–Leibler Divergence for Evaluating Equivalence / Vladimir Dragalin	1249
LOCF: Validity / Bin Cheng and Shein-Chung Chow	1296
Logistic Regression / Rick Chappell and Jason Fine	1303
McNemar’s Test / Robert G. Lehr	1320
MCP-Mod: Dose–Response Testing and Estimation / Björn Bornkamp and Naitee Ting	1326
Measuring Agreement / Lawrence I-Kuei Lin	1335
Minimization Procedure / Dongsheng Tu	1401
Minimum Effective Dose / Keumhee Carriere Chough	1406
Modified Large Sample Method / Yonghee Lee	1427
Multicollinearity / K. Van Steen and G. Molengerghs	1438
Multiple Comparisons / Mani Y. Lakshminarayanan	1461
Multiple Endpoints / Mario Comelli	1467
Multiregional Clinical Trials: Qualitative Consistency of Treatment Effects / Yoko Tanaka and Hiroshi Nishiyama	1522
Noninferiority Trials: Type I Error Rates / Seung-Ho Kang	1536
Nonparametric Regression / Kongming Wang and Theo Gasser	1543
Odds Ratio / Dongsheng Tu	1554
Onset of Action / Thomas R. Stiger and Christy Chuang-Stein	1585
Ordered Categorical Data: Test for / Shein-Chung Chow, Siu-Keung Tse, and Chunyan Yang ..	1602
Placebo Effect / Thomas R. Stiger and Edward F.C. Pun	1735
Power / Kenneth B. Schechtman	1777
Principal Component Analysis / Aiyi Liu and Enrique F. Schisterman	1796
Propensity Score Analysis and Its Application in Regulatory Settings / Lilly Q. Yue	1823
Proportion of Treatment Effect / Cong Chen and Hongwei Wang	1830
P-Values / Stephen Senn	1845
Reliability / Valentin Rousson and Theo Gasser	1908
Repeated Measurement Designs: Missing Values / Keumhee Carriere Chough, Yuanyuan Liang, Rong Huang, and Xiaoming Sheng	1918
Response Surface Methodology / William R. Myers	1939
Risk Ratio Analysis / Kung-Jong Lui	1956

ROC Curve / Robert G. Lehr and Annpey Pong	1962
SROC Curve / S. D. Walter and Petra Macaskill	2101
Statistical Principles for Clinical Trials / Allan Phillips	2130
Statistical Significance / Dieter Hauschke, Robert Schall, and Herman G. Luus	2142
Stuart–Maxwell Test / Jiang-Ming Johnny Wu	2154
Subgroup Analysis / Patricia B. Cerrito	2159
Subject-Treatment Interaction / Gary L. Gadbury	2167
Surrogate Endpoint / Shu Zhang and Edward F. C. Pun	2174
Survival Analysis / Dirk F. Moore	2179
Testing for Qualitative Interaction / Xin Yan and Xiaogang Su	2192
Therapeutic Equivalence / Jen-pei Liu	2199
Trend Estimation / Jian Qing Shi and Man Lai Tang	2283
Z-Score / Jiang-Ming Johnny Wu	2329

Non-Clinical Development

Drug Discovery and Development

Drug Development / Naitee Ting	908
Oncology Trials: Development Strategy / Zhaoyang Teng, Yi Liu, Jing Yang, Zhaowei Hua, and Mark Chang	1572
Pharmaceutical Development: Statistical Designs / Annpey Pong and Shein-Chung Chow. . .	1656

Laboratory Development

Dose Proportionality / C. Gordon Law	876
Laboratory Analyses / Michael J. Klepper	1258
Lilly Reference Ranges / Michael G. Wilson	1276

Analytical Method Development and Validation

Assay Development / Timothy L. Schofield	130
Assay Validation / Timothy L. Schofield	138
Assay Validation: Evaluation of Linearity / Jen-pei Liu and Eric Hsieh	146
Bioassay / Anant M. Kshirsagar	213
Calibration / Shein-Chung Chow	370
Instrument Development and Validation / Cindy Rodenberg, James T. Kuznicki, and Gloria H. Yiu	1167
Validation of Quantitative and Qualitative Assays / Bob Zhong, Chunfu Qiu, and Dejun Tang . .	2313

Process Validation

Comparability Studies: Statistical Considerations / Jason J. Z. Liao, Li He, and Robert Capen. . .	682
Process Validation / James S. Bergum and Merlin L. Utter	1802
Statistical Process Control / William R. Myers and William A. Brenneman	2135

Quality Control and Assurance

Acceptance Sampling / William A. Brenneman and William R. Myers	1
--	---

Quality Control and Assurance (cont'd.)

Specifications / John R. Murphy	2099
Release Targets / Greg C. G. Wei	1903
Quality by Design in Biopharmaceutical Development / Harry Yang.....	1865

USP Tests

Content Uniformity / John R. Murphy	709
Dissolution Profile Comparison: Statistical Methods / Harry Yang	844
Dissolution Profiles Comparison: Model Independent Approaches / Yanbing Zheng and Lanju Zhang.....	854
Dosage Units Using Small and Large Sample Sizes: Uniformity / Meiyu Shen and Yi Tsong ...	859
In Vitro Testing: Dissolution Profile Comparison / Yi Tsong, Pradeep M. Sathe, and Vinod P. Shah.....	1148
In Vitro Testing: Micronucleus Test / Wherly P. Hoffman, Michael L. Garriott, Cindy Lee. ...	1155
USP Tests / John R. Murphy	2311

Stability Analysis

ANCOVA Approach: Premarketing Shelf Life Determination with Multiple Factors / Yi Tsong, Chi-wan Chen, Wen Jen Chen, Roswitha E. Kelly, Daphne T. Lin, and Karl K. Lin	125
Bayesian Methods: Stability Analysis / Jie Chen, Jinglin Zhong, and Lei Nie.....	180
Bracketing Design / John R. Murphy	359
Expiration Dating Period / Annpey Pong.....	979
Stability Analysis: Frozen Drug Products / Shein-Chung Chow and Jun Shao	2111
Stability Matrix Designs / Earl Nordbrock.....	2116
Rank Regression: Stability Analysis / Annpey Pong, Ying Qing Chen, and Xiudong Lei	1888

Pre-Clinical Development
Animal Study

Ames Test / Wherly P. Hoffman and Michael L. Garriott	67
Analysis of Heritability / Brent D. Burch and Ian R. Harris	98

Toxicology Study

Carcinogenicity Studies of Pharmaceuticals / Karl K. Lin.....	398
ED₅₀/ED₉₀ / Yangxin Huang	952
Reproductive and Developmental Studies / James J. Chen	1931
Toxicological Studies / Guenther Baus and Thomas Hoffman.....	2230

Bioavailability and Bioequivalence

Bioavailability and Bioequivalence / Mengdie Yuan and Jingyu Luan Julia.	216
Clinical Endpoint BE Studies: Generic Drugs / Stella Grosser, Misook Park, Sam G. Raney, and Elena Rantou.	455

Drug Interchangeability: Meta-Analysis for Safety Monitoring / Wen-Wei Liu and Shein-Chung Chow	928
Drug Interchangeability: Criteria And Design / Shein-Chung Chow, Fuyu Song, and Meng Chen	922
Drug Interchangeability: Generic Products, Prescribability and Switchability / Laszlo Endrenyi, Shein-Chung Chow, and Laszlo Tothfalusi	935
Drug Interchangeability: Meta-Analysis for Safety Monitoring / Wen-Wei Liu and Shein-Chung Chow	928
In Vitro Testing: Bioequivalence / Hansheng Wang, Ying Zhang, Jun Shao, and Shein-Chung Chow	1141
Individual Bioequivalence / Shein-Chung Chow and Jen-pei Liu	1160
Interchangeability: Generic Drugs / Jiayin Zheng, Shein-Chung Chow, and Mengdie Yuan ...	1187
Multiple-Dose Bioequivalence Studies / Vernon M. Chinchili	1477
Population Bioequivalence / Naitee Ting, Greg C. G. Wei, and William W. B. Wang	1740
Profile Analysis / Bin Cheng and Shein-Chung Chow	1817
Scaled Average Bioequivalence (SABE) / Laszlo Endrenyi and Laszlo Tothfalusi.	2038
Slope Approach: Assessment of Dose Proportionality and Linearity Under a Crossover Design / Bin Cheng and Shein-Chung Chow	2084
Pharmacokinetics and Pharmacodynamics	
Clinical Pharmacology / Brian Smith	469
Pharmacodynamic Issues / Cheng-Tao Chang and Robert L. Wong	1665
Pharmacodynamics with No Covariates / Cheng-Tao Chang and Robert L. Wong	1685
Population PK/PD Analysis / D. Concorde, C. Ané and F. Léger	1747
Sample Size Re-estimation Design with Robust Power / Lu Cui	2026
Bioinformatics	
Bioinformatics / Sean D. Mooney and Jeffrey T. Chang	224
Biomarker Development with Applications to Genetics Data: Statistical Tests / Gang Zheng and Jungnam Joo	243
Data Mining and Biopharmaceutical Research / Patricia B. Cerrito	805
Genetic Linkage and Linkage Disequilibrium Analysis / Kongming Wang, Bernice Porjesz, Henri Begleiter, and Kevin Jones.	1057
Microarray Gene Expression / James J. Chen and Chun-Houh Chen	1377
Statistical Genetics / Rongling Wu, Guifang Fu, and Hongying Li	2122
Clinical Development	
Randomization and Blinding	
Blinding / Shein-Chung Chow and Jun Shao	336
IVRS/IWRS for Randomization / Mon-Gy Chen	1216
Randomization / Shein-Chung Chow and Jun Shao	1875
Sample Size Requirement	
Bayesian Model: Prior Effective Sample Size / Satoshi Morita and Peter F. Thall.	186

Sample Size Requirement (cont'd.)

Blinded Sample Size Re-Estimation / Chien-Hua Wu	323
Cluster Randomized Trials: Sample Size / Sin-Ho Jung	645
GEE in Cancer Research / Sin-Ho Jung	1036
Kruskal-Wallis Test for Ordered Categorical Data: Power Calculation / Chunpeng Fan and Donghui Zhang	1245
Sample Size Estimation: Generalized Estimating Equations (GEE) Method / Sin-Ho Jung	2010
Sample Size for Multiregional Clinical Trials / Yuh-Jenn Wu, Chieh Chiang, Chi-Tian Chen, Hsiao-Hui Tsou, and Chin-Fu Hsiao	2015
Sample Size Calculation: Nonparametrics / Hansheng Wang and Shein-Chung Chow	1970
Sample Size Calculation: Survival Data / Qi Jiang, Steven Snapinn, and Boris Iglewicz	1980
Sample Size Determination / Lilly Q. Yue, David Li and Shan Bai	1987
Sample Size Re-estimation Based on Observed Treatment Difference / Lu Cui	2021
Sample Size: Negative Binomial With Unequal Follow-Up Times / Yongqiang Tang	2031
Sample Size Re-estimation Design with Robust Power / Lu Cui	2026

Clinical Trial

Active Control Non-Inferiority Trials: Hypotheses and False Positive Rate / Gang Chen, George Y. H. Chi, and Yong-Cheng Wang	8
Active Control Trials / George Y. H. Chi, Gang Chen, Mark Rothmann, and Ning Li	14
Bayesian Designs: Phase II Oncology Clinical Trials / Sze Huey Tan, David Machin, and Say Beng Tan	154
Biomarker: Clinical Trials / Mark Chang	251
Cancer Chemotherapy: Maximum Tolerable Dose / Jen-pei Liu	384
Cancer Trials / Clet Niyikiza and Douglas E. Faries	390
Clinical Inspection: Statistical Process / Fuyu Song, Minghuang Hong, and Shein-Chung Chow	462
Clinical Research: Benefit and Risk Assessment / Susan Talbot, Shahrul Mt-Isa, Qi Jiang, and Chunlei Ke	472
Clinical Trial Process / Paul F. Kramer	532
Clinical Trial: Combination Drug Development / H. M. James Hung	552
Clinical Trials / Kenneth B. Schechtman	572
Clinical Trials: Controversial Issues / Shein-Chung Chow and Fuyu Song	580
Cluster Trials / John S.F. Preisser and Bing Lu	648
Equivalence Trials / Jen-pei Liu	970
Group Sequential Tests and Variance Heterogeneity:	
Clinical Trials / Keumhee Carriere Chough, Abdulkadir Hussein, and Taesung Park	1114
Human Drug Abuse Trials / Qianyu Dang	1121
Microdosing Studies / Fuyu Song and Shein-Chung Chow	1394
Multicenter Trials / William J. Huster	1434
Multinational Clinical Trials / Thomas R. Weihrauch	1454
Multiple-Stage Designs: Phase II Cancer Trials / Masha Kocherginsky and Shang P. Lin	1481
Multiplicative Intensity Models / Keumhee Carriere Chough and Abdulkadir Hussein	1494
Multiplicity in Clinical Trials / Peter Westfall and Frank Bretz	1502
Phase I Cancer Clinical Trials / Seung-Ho Kang and Chul Ahn	1715

Phase II Clinical Development / Naitee Ting and Guojun Yuan	1721
Protocol Development / Robert D. Chew	1841

Classical Design

Center Weighting: Multicenter Trials / Paul Gallo	435
Clinical Endpoint / Sofia Paul.	451
Clustered Study Designs: Power Analysis / J. Zhang, C. Feng, W. Tang, X. M. Tu, and J. Kowalski	658
Cutoff Designs / Joseph C. Cappelleri and William M. K. Trochim.	798
Dose Response Study Design / Naitee Ting	885
Dose-Ranging Studies: Design Considerations / Qiqi Deng and Naitee Ting	891
Enrichment Design with Patient Population Augmentation / Yijie Zhou, Bo Yang, Lanju Zhang, and Lu Cui.	964
Extra Variation Models / Jorge G. Morel and Nagaraj K. Neerchal.	992
Factorial Designs / Junfang Li.	1014
Optimal Futility Interim Design / Zhongwen Tang	1593
Ordinal Data: Crossover Design Analysis / Kung-Jong Lui.	1613
Parallel Design / Mani Y. Lakshminarayanan.	1621
Randomized Trials with Clusters: Design and Analysis / Sin-Ho Jung	1880
Screening Design / John R. Murphy	2045
Surrogate Endpoint / Shu Zhang and Edward F. C. Pun.	2174

Innovative Trial Design

Adaptive Design Clinical Trials: Case Studies / Shiohjen Lee and Min Lin.	28
Adaptive Design Methods in Clinical Trials / Shein-Chung Chow and Annpey Pong	36
Adaptive Survival Trials / William F. Rosenberger and Lanju Zhang	50
Clinical Trial: N-of-1 Design Analysis / Can Cui and Shein-Chung Chow	564
Clinical Trials: Response-Adaptive Repeated Measurement Designs / Keumhee Carriere Chough and Yuanyuan Liang	612
Covariate-Adjusted Adaptive Dose-Finding: Early Phase Clinical Trials / Peter F. Thall and Hoang Q. Nguyen.	760
Enrichment Design / Jen-pei Liu.	960
Phase III Clinical Trials with Response-Adaptive Allocation / Atanu Biswas and Rahul Bhattacharya.	1727
Response-Adaptive Designs / Feifang Hu and Anastasia Ivanova.	1950
Seamless Adaptive Trial Design: Dose Selection Criteria / Jiayin Zheng and Shein-Chung Chow	2047
Targeted Clinical Trials / Jen-pei Liu	2184
Two-Stage Adaptive Design: Analysis / Shein-Chung Chow and Min Lin	2289
Two-Stage Design: Phase II Cancer Clinical Trials / Sin-Ho Jung.	2296

Data Management

Clinical Data Management / Terri Madison and Marianne Plaunt.	444
Global Database and System / Alice T. M. Hsuan and Patrick Genyn	1069
Mixed-Outcome Data / A. R. de Leon.	1414

Statistical Methods for Data Analysis

Adverse Event Reporting / Cynthia Pennise	53
Analysis of $2 \times K$ Tables / Shiva Gautam	75
Analysis of Clustered Binary Data / Valerie Durkalski	82
Analysis of Clustered Categorical Data / Kalyan Das	88
Analysis of Variance (ANOVA) / Maria Overbeck-Larisch and Werner Sanns	104
Covariates: Adjustment for / Thomas Permutt	772
Bayesian Estimate: Concordance Correlation Coefficient / Dai Feng, Richard Baumgartner, and Vladimir Svetnik	164
Bayesian Methods: Meta-Analysis / Elizabeth Stojanovski and Kerrie L. Mengersen	173
Bivariate Zero-Inflated Poisson Populations: Statistical Test for Homogeneity / Siu-Keung Tse and Hak-Keung Yuen	316
Case-Control Studies: Inference in / Gary King and Langche Zeng	422
Clinical Research: Comparing Variabilities / Yonghee Lee, Hansheng Wang, and Shein-Chung Chow	486
Clinical Trial: Failure-Time Model / Chin-Shang Li	556
Clinical Trial Design: Statistical Assurance / Shuyen Ho	523
Clinical Trials: Interval-censored Failure Time Data / Jianguo Sun, Qingning Zhou, and Ding-Geng (Din) Chen	589
Clinical Trials: Meta-Analysis using R / Ding-Geng (Din) Chen	603
Correlated Probit Model / Ralitza V. Gueorguieva	741
Cost-Effectiveness Analysis / Heejung Bang	751
Crossover Design: Rank-Based Robust Analysis / M. Mushfiqur Rashid	784
Disease Modification Effects: Design and Analysis / Djokouri A. Kouassi and Annpey Pong ...	839
Dose Finding Clinical Trials Using R / Naitee Ting and Ding-Geng (Din) Chen	866
Dose Response Analysis in Clinical Trials / James MacDougall	879
Ecologic Inference / Sander Greenland	940
Exploratory Factor Analysis / Joseph C. Cappelleri and Robert A. Gerber	985
Factor Analysis / Mary D. Sammel, Sarah J. Ratcliffe, and Benjamin E. Leiby	1008
Good Statistics Practice (GSP) / Shein-Chung Chow	1095
Group Sequential Methods / Cheng-Tao Chang	1105
Hypothesis Testing / Devan V. Mehrotra and Ivan S. F. Chan	1124
Imputation: Clinical Research / Hansheng Wang and Shein-Chung Chow	1135
Imputation with Item Nonrespondents / Hansheng Wang and Shein-Chung Chow	1128
Latent Class Analysis / Elizabeth S. Garrett-Mayer and Jeannie-Marie S. Leoutsakos	1266
Local Influence Analysis / Nicola Sartori	1292
Logistic Regression: Three-Point Designs / Mårten Vågerö and Rolf Sundberg	1315
Mixed Effects Models / Donghui Zhang, Chunpeng Fan, and A. Lawrence Gould	1409
MMRM with Missing Data / Annpey Pong, Jun Zhao, Allen Lee, Phillip Phiri, and Lev Sverdlov	1422
Multidimensional Data Analysis: Penalized Regression Methods / Keumhee Carriere Chough and Xiaoming Wang	1448
Multiplicative Intensity Models / Keumhee Carriere Chough and Abdulkadir Hussein	1494
Multivariate Meta-Analysis / Elizabeth Stojanovski and Kerrie L. Mengersen	1528

Oncology Clinical Trials: Efficient Dose-Finding Designs / Erik Rasmussen, Qi Jiang, and Chunlei Ke	1560
Ordered Multiple Class Receiver Operating Characteristic (ROC) Analysis / Christos T. Nakas and Constantin T. Yiannoutsos	1608
Percentile Charts on Correlated Measures / Xuming He and Kai W. Ng.	1643
Prediction Trees / Antonio Ciampi	1782
Proportional Hazards Regression Model / Bee Leng Lee and Jason Fine.	1834
QT Analysis / Lang Li, Stephen Hall, and Mehul Desai.	1855
Semi-Parametric Time-Varying Regression Models / Thomas H. Scheike.	2058
Sequential Estimation: Additive Hazards Rate Model with Staggered Entry / Laurent Bordes and Christelle Breuils.	2063
Spatio-Temporal Modeling / Ying C. MacNab and Charmaine B. Dean	2089
Structural Equation Model / Joseph L. Carrasco	2147
Thorough QT Studies / Joanne Zhang, Dalong (Patrick) Huang, and Yi Tsong.	2206
Repeated Measures Data with Missing Values Analysis: Overview of Methods / Keumhee Carriere Chough, Yuanyuan Liang, and Taesung Park	1924
Weibull Model: Group Sequential Design / Jianrong Wu.	2325

Post-Marketing

Post Marketing Surveillance

Patient-Reported Outcomes / Amylou C. Dueck and Joseph C. Cappelleri	1633
Postmarketing Adverse Drug Event Signaling / Yi Tsong	1756
Postmarketing Surveillance / Patricia B. Cerrito	1769
Signal Detection in Pharmacovigilance / Patrick J. Farrell, Christopher A. Gravel, and Daniel Krewski.	2068

Outcome Research

Logistic Regression Process: Predictive Model Building in Clinical Research / Mo Liu and Shein-Chung Chow	1309
Pharmacoeconomics / Joseph Heyse, John R. Cook, and Michael F. Drummond	1693

Quality of Life

Clinical Trials: Investigating Quality-of-Life / Patricia B. Cerrito.	597
---	-----

Miscellaneous

CRO

Contract Research Organization (CRO): Early 2000s / T. Y. Lee.	716
Contract Research Organization (CRO) / Tamara Pinkett and Jill Dawson.	721

Medical Dictionary

MedDRA: Impact on Pharmaceutical Development / Keya T. Pitts	1342
---	------

Compliance

Patient Compliance / Kenneth B. Schechtman	1627
---	------

Bridging Study and Multi-Regional Clinical Trial

Bridging Studies / Jen-pei Liu	361
Bridging Studies: Statistical Methods / Yuh-Jenn Wu, Chieh Chiang, Chi-Tian Chen, Hsiao-Hui Tsou, and Chin-Fu Hsiao.	366
Ethnic Factors / Melody H. Lin and Helen McGough.	976
Multi-Regional Clinical Trials: Consistency Test / Jiayin Zheng, Lisa Ying, Na Zeng, and Shein-Chung Chow	1511

Vaccine Development

Clinical Trials: Vaccine Development / Ivan S. F. Chan, William W. B. Wang, and Joseph Heyse.	628
---	-----

Medical Devices

Bayesian Statistics: Medical Device Clinical Trials / Gerry W. Gray, Xuefeng Li, Laura Thompson, Yunling Xu, Xu Yan, and Lilly Q. Yue	198
Companion Diagnostics / Gene Pennello and Jingjing Ye	672
Diagnostic Device Evaluation / Meijuan Li, Tinghui Yu, Qin Li, Gene Pennello, R. Lakshmi Vishnuvajjala, Bipasa Biswas, Changhong Song, and Lilly Q. Yue	822
Diagnostic Imaging / Suming W. Chang and Annpey Pong	829
Diagnostic Procedure: Sensitivity and Specificity / Jiayin Zheng and Shein-Chow Chung	836
Medical Devices / Greg Campbell, Lilly Q. Yue, Gene Pennello, and Mary K. Barrick	1348

Biosimilar Drug Products

Analytical Similarity Assessment / Shein-Chung Chow, Fuyu Song, and He Bai	116
Biologics / P. A. Lachenbruch, A. D. Horne, C. J. Lynch, J. Tiwari, and Susan S. Ellenberg	233
Biosimilar Studies: Sample Size / Seung-Ho Kang	274
Biosimilar Studies: Three-Arm Design / Seung-Ho Kang	277
Biosimilarity Assessment / Shein-Chung Chow and Fuyu Song	281
Biosimilarity of Follow-On Biologics / Shein-Chung Chow and Jen-pei Liu	304
Drug Interchangeability: Biosimilar Products / Laszlo Endrenyi, Shein-Chung Chow, and Laszlo Tothfalusi	914
Tiered Approach of Analytical Similarity Assessment: Statistical Methods / Yi Tsong	2219

Translational Medicine

Translational Medicine: Concepts, Statistical Methods, and Related Issues / Siu-Keung Tse and Shein-Chung Chow.	2270
---	------

Precision Medicine

Precision Medicine: Statistical Issues / Fuyu Song and Shein-Chung Chow	1779
Personalized Medicine: Validation of Diagnostic Medical Devices / Meijuan Li, Gene Pennello, Jinciao Wu, Laura Yee, Zhiwei Zhang, Gerry Gray, Yuying Jin, Yaji Xu, Jingjing Ye, and Zhiheng Xu	1648

Traditional Chinese Medicine

Botanical Drug Product Development: Scientific Issues / <i>Annpey Pong and Shein-Chung Chow</i>	348
Traditional Chinese Medicine (TCM): Clinical Development / <i>Fuyu Song and Shein-Chung Chow</i>	2244
Traditional Chinese Medicine (TCM): General Considerations / <i>Shein-Chung Chow, Annpey Pong, and Yu-Wei Chang</i>	2252
Traditional Chinese Medicine (TCM): Unified Approach for Evaluation / <i>Bin Cheng and Shein-Chung Chow</i>	2265



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Preface

In the past decade, there has been explosive growth of literature in biopharmaceutical research and development of new medicines for the treatment of critical diseases that are severe and/or life-threatening. In biopharmaceutical research and development, biopharmaceutical statistics plays an extremely important role in ensuring not only the efficacy and safety of the medicine under investigation but also that the pharmaceutical product possess good drug characteristics such as identity, strength, purity, quality, safety, stability, and reproducibility. The aims of this encyclopedia are to (1) provide a comprehensive and unified presentation of designs and analyses used at different stages of biopharmaceutical research and development, (2) give a well-balanced summary of current regulatory requirements, and (3) describe recently developed statistical methods in the pharmaceutical sciences.

Since the first edition was published in 2000, this encyclopedia has been well received by pharmaceutical scientists/researchers and is now widely used as a reference source in biopharmaceutical research and development. The second edition, published in 2003 included for the first time an online version for easy access by pharmaceutical scientists/researchers and biostatisticians. The third edition was published in 2010 included additional 89 new entries to reach a total of 230 entries in this encyclopedia. The third edition was also available online through the Informa website. This 4th edition can be distinguished from the first three editions in the following ways:

1. Revised and updated entries reflect changes and recent development in regulatory requirements for drug review/approval process and statistical design and methodology.
2. Additional topics have been included to provide a more complete and comprehensive reference source. To name a few, this edition includes important topics such as multiple-stage adaptive trial design in clinical research, biomarker development for target clinical trial for precision medicine, translational medicine, design and analysis of biosimilar drug development, quantitative methods for traditional Chinese medicine development, big data analytics, and real world evidence for clinical research and development. An additional 78 new entries have been contributed by experts in their respective field. As a result, there are now a total of 308 entries in this encyclopedia. This 4th edition will be available online through the Taylor & Francis website.
3. Categorized table of content by stages of biopharmaceutical development provides an easy access to relevant topics in specific stages of biopharmaceutical development.

Although most of the contributors to this encyclopedia are biostatisticians with extensive practical experience in biopharmaceutical research and development, we have attempted to make each entry self-contained and self-explanatory to the reader who is not an expert in the subject matter of each topic. I believe this edition will benefit scientists and researchers from the pharmaceutical and/or biotechnology industry, policymakers, members of regulatory agencies (e.g., EU EMA and US FDA) and government bodies (e.g., US NIH), as well as academia by serving as an extremely useful guide on study protocol and statistical methodology, or indeed the matters of quality control, clinical trial design, risk-benefit analyses, big data analytics, interpretation of results, and formal assessment and appraisal of drugs and pharmaceutical products.

I would like to thank David Grubbs of Taylor & Francis for providing me with the opportunity to work on this edition, and specifically Stephanie DeRosa for her outstanding efforts in coordinating with the contributors during the preparation of this edition. I am deeply indebted to the U.S. Food and Drug Administration (FDA) and many major pharmaceutical companies, including Amgen, Merck, Eli Lilly, Pfizer, Bayer, Novartis, and Bristol-Myers Squibb for their support. I would like to express my gratitude to Dr. Eric Chi (Amgen), Dr. Jiayin Zheng and Dr. Ralph Corey (Duke University School of Medicine), Dr. Laszlo Endrenyi (University of Toronto), and Dr. Lisa LaVange (FDA) for their constant encouragement and support during the process of the preparation of this edition.

Finally, the views expressed in this edition are those of the contributors and not necessarily those of their respective companies or organizations. Any comments and suggestions will be very much appreciated for the preparation of future editions.

Shein-Chung Chow, PhD, Editor



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

About the Editor

Shein-Chung Chow, PhD.

Shein-Chung Chow, PhD. is currently an Associate Director, Office of Biostatistics, U.S. Food and Drug Administration (FDA). Dr. Chow is also a Professor at the Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, and an Adjunct Professor at Duke-NUS, Singapore and North Carolina State University, Raleigh, NC. Dr. Chow is also the Founding Director of the Global Clinical Trial and Research Center, Tianjin, China. Prior to joining Duke University, he was the Director of TCOG (Taiwan Cooperative Oncology Group) Statistical Center and the Executive Director of National Clinical Trial Network Coordination Center. Prior to that, Dr. Chow also held various positions in the pharmaceutical industry such as Vice President, Biostatistics, Data Management, and Medical Writing at Millennium Pharmaceuticals, Inc., Cambridge, MA; Executive Director, Statistics and Clinical Programming at Covance, Inc., Director and Department Head at Bristol-Myers Squibb Company, Plainsboro, NJ. Dr. Chow is the Editor-in-Chief of the Journal of Biopharmaceutical Statistics, Open Access Drug Designing, and the Journal of Biosimilars. Dr. Chow is also the Editor-in-Chief of the Biostatistics Book Series at Chapman and Hall/CRC Press of Taylor & Francis Group. He was elected Fellow of the American Statistical Association in 1995 and an elected member of the ISI (International Statistical Institute) in 1999. He was appointed as a special government employee (SGE) by the FDA in June 2015. In this capacity, Dr. Chow serve as an Oncologic Drug Advisory Committee (ODAC) voting member and a statistical expert consultant to the FDA. Dr. Chow is the author or co-author of over 300 methodology papers and 29 books including Design and Analysis of Bioavailability and Bioequivalence Studies, Statistical Design and Analysis in Pharmaceutical Science, Design and Analysis of Clinical Trials, Design and Analysis of Animal Studies in Pharmaceutical Development, Sample Size Calculations in Clinical Research, Adaptive Design Methods in Clinical Trials, Translational Medicine, Design and Analysis of Bridging Studies, Biosimilars: Design and Analysis of Follow-on Biologics., Quantitative Methods for Traditional Chinese Medicine Development, Biosimilar Drug Product Development, and Statistical Methods for HIV/AIDS Research.

Dr. Chow received a B.S. in mathematics from National Taiwan University, Taiwan, and a PhD. in statistics from the University of Wisconsin, Madison, Wisconsin.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Encyclopedia of Biopharmaceutical Statistics, 4th Edition

Volume I

Acceptance–Clinical Trials

Pages 1—588

Acceptance–
Ames Test

Analysis–Assay

Assay–Bioassay

Bioavailability–
Bivariate

Blinded–Cancer Trials

Carcinogenicity–
Clinical Pharmacology

Clinical Research–
Clinical Trials



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Acceptance Sampling

William A. Brenneman and William R. Myers
The Procter & Gamble Company, Mason, Ohio, U.S.A.

Abstract

This entry summarizes the process and significance of acceptance sampling, which is defined as a statistical tool used to help make decisions concerning whether or not a batch (or lot) of product should be released for consumer consumption or use.

Keywords: Acceptance criteria; Acceptance sampling; Sampling plans.

INTRODUCTION

In general, acceptance sampling is a statistical tool used to help make decisions concerning whether or not a batch (or lot) of product should be released for consumer consumption or use. Acceptance sampling, which will be discussed in this paper, should not be confused with the sampling plans and acceptance criteria mentioned in the United States Pharmacopeia 25 and National Formulary 20.^[1] While acceptance sampling has been used extensively in many industries over the past 60 years, in far too many cases, inspection sampling and quality have been thought of as being synonymous. This is simply not the case because quality stems from a well-designed product produced from a well-designed and capable process. Once the product has been produced and the quality of the product has been fully determined, acceptance sampling can only give us a level of assurance that the product quality is at a certain level and that an Acceptable Quality Level (AQL) will be exposed to consumers. For this reason, the focus in the quality arena has changed from the inspection of finished product quality to the continuous improvement of the product design and process performance.

Although most quality efforts should be put toward the proactive approach found in a continuous quality improvement program, acceptance sampling still plays a practical role. Situations for which acceptance sampling may be useful include incoming inspection with a new supplier for which little is known regarding its capability or quality history; new processes where statistical control is attempted; processes where the quality is relatively poor; and processes that experience flare-ups or special causes (unusual or unexpected result). Furthermore, acceptance sampling is required in current good manufacturing practice (cGMP) in pharmaceutical research and development.^[2] However, the most attractive feature of acceptance sampling is that it allows one to balance two risks: the risk to the producer and the risk to the consumer. These risks need to be fully

understood in order to appreciate acceptance sampling methodology.

An example will be used in order to provide the basic ideas behind acceptance sampling, giving special attention to the producer and the consumer risks. It turns out that these two types of risks uniquely determine a plan, and a thorough understanding of these basic principles will help in understanding more complex plans. Double sampling plans as well as more complex sampling plans are discussed briefly. Variables sampling plans are introduced, and the derivation of their acceptance regions is discussed. Finally, some common misconceptions associated with acceptance sampling plans are presented, and thoughts concerning the use of ANSI/ASQC^[3] sampling procedures are given. Theoretical justification for acceptance sampling is outside the scope of the entry but can be found in Wallis,^[4] Schilling,^[5] and Stephens.^[6] For a historical look at inspection sampling, see Sherman.^[7] Acceptance sampling is reviewed extensively from a Bayesian viewpoint by Wetherill,^[8] Hald,^[9] Moskowitz and Berry,^[10] and Thyregod.^[11] Bayesian cost models are reviewed by Hald,^[12] Guenther,^[13] Lorenzen,^[14] and Wadsworth and Olsen.^[15] Situations where the fraction nonconforming is very low are discussed in Hahn,^[16] Pesotchinsky,^[17] and Moreno and Reeves.^[18] Suich^[19] talks about the effect of inspection errors on the performance of acceptance sampling plans. Finally, the debate concerning all (100%) or nothing (0%) inspection is given by Vander Weil and Vardeman.^[20]

ACCEPTANCE SAMPLING TERMINOLOGY

There are two types of risks that go into creating or evaluating any acceptance sampling plan. The producer's risk is the probability that a lot of product that is at an AQL is rejected by the plan. The AQL is the maximum percent nonconforming that can be considered satisfactory as a process average. On the other hand, the consumer's

risk is the probability that a lot with unacceptable quality is accepted by the plan. The unacceptable quality level is frequently referred to as the Lot Tolerance Percent Defective (LTPD). These four quantities uniquely determine an acceptance sampling plan and can be thought of as the properties (or characteristics) of the plan. In formal hypothesis testing, the terminology used for the producer and the consumer risks are the type I and type II error rates, respectively. In the acceptance sampling literature, plans based on binary data (pass/fail) are referred to as *attribute* acceptance sampling plans, whereas plans based on continuous data are referred to as *variable* acceptance sampling plans. This common terminology will be used throughout. The attribute data could be for the physical evaluation of drug product (e.g., capsule damage, empty capsules, cracked or broken tablets, and coating defects), packaging material (e.g., illegible print) or incoming raw material from a supplier. Examples of variable data are tablet weight, tablet hardness, and bottle measurements.

ACCEPTANCE SAMPLING PLANS FOR ATTRIBUTES

Acceptance sampling plans can be determined from the four quantities: producer's risk, AQL, consumer's risk, and LTPD. For example, consider the creation of a single sampling plan for which the response of interest is an attribute. A single sampling plan then requires determining a sample size n and an acceptance number c . The plan is carried out by counting the number of nonconforming (or defective) units d in a sample of size n and then accepting the lot if $d \leq c$. The sample size and the acceptance number are chosen so that the plan will have the specified properties: producer's risk (α), AQL ($p_1 = \text{AQL}/100$), consumer's risk (β), and LTPD ($p_2 = \text{LTPD}/100$). Assuming that the binomial distribution is an appropriate distribution for modeling the sampling procedure, the sample size n and the acceptance number c are the solutions to the two equations:

$$1 - \alpha = \sum_{d=0}^c \frac{n!}{d!(n-d)!} p_1^d (1 - p_1)^{n-d} \quad (1)$$

$$\beta = \sum_{d=0}^c \frac{n!}{d!(n-d)!} p_2^d (1 - p_2)^{n-d} \quad (2)$$

To be specific, suppose a single sampling plan for cracked tablets calls for a producer's risk of 0.05, an AQL of 1%, a consumer's risk of 0.10, and an LTPD of 5%. Then a sample size of 132 tablets and an acceptance number of 3 are required. To carry out the plan, one would randomly sample 132 tablets from a lot and count the number of cracked tablets found. In practice, this is often done on line. If the number of cracked tablets found is three or less, then the lot is accepted. Otherwise, the lot is rejected.

ACCEPTANCE SAMPLING GRAPHS

An important graph to construct for any sampling plan is called the Operating Characteristic (OC) curve. The OC curve depicts the probability of accepting a lot for a given true, but unknown, defect level. The defect level p is placed on the horizontal axis and the probability of acceptance is placed on the vertical axis. For a single sampling plan for attributes, the probability of acceptance is the probability that d is less than or equal to c , and is given by the expression:

$$P_a = P\{d \leq c\} = \sum_{d=0}^c \frac{n!}{d!(n-d)!} p^d (1 - p)^{n-d} \quad (3)$$

Fig. 1 shows the OC curve for the sampling plan where the sample size is 132 and the acceptance number is 3. The OC curve shows how well a sampling plan does at discriminating between lots with various defect levels. For example, if the true defect level is 2%, then the proposed acceptance sampling plan will have approximately a 73% probability of accepting the lot. That is, if 100 lots from a process that produces 2% defective products are evaluated by the sampling plan, then we can expect, on average, to accept 73 of the lots and reject 27 of them. For a discussion on the effect of a sample size n and an acceptance number c on the OC curve, see Montgomery^[21] or Stephens.^[6]

Another important graph is one that depicts the Average Outgoing Quality (AOQ). The AOQ is the average lot quality that one would expect over time from a process producing fraction defective p . The AOQ only makes sense when the rejected lots are subject to 100% inspection and all non-conforming items are either removed or replaced with conforming items. Such sampling programs are called *rectification inspection programs* because the action of 100% inspection affects the final product quality. Because the only time defective items are released is when

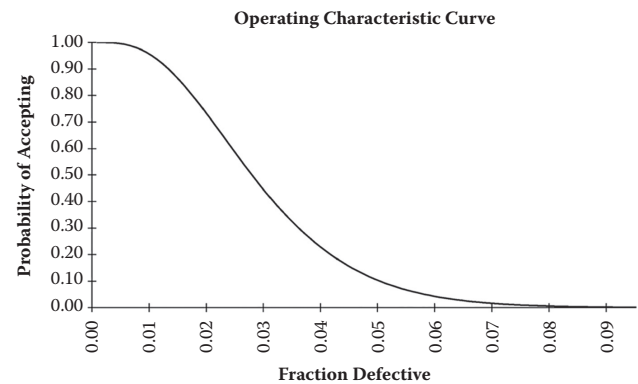


Fig. 1 The OC curve for a single acceptance sampling plan with producer's risk = 5%, AQL = 1%, consumer's risk = 10%, and LTPD = 5%. Sample size, $n = 132$; and acceptance number, $c = 3$

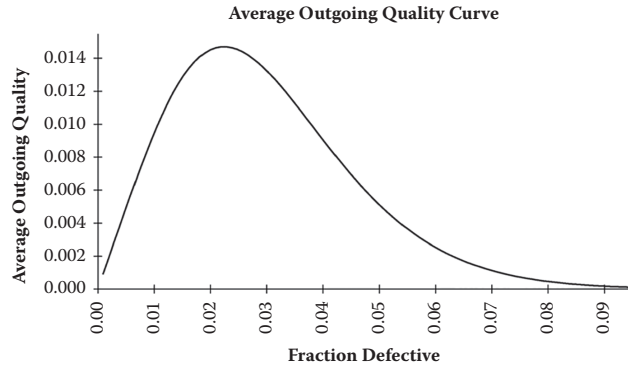


Fig. 2 The AOQ curve for a single acceptance sampling plan with producer's risk = 5%, AQL = 1%, consumer's risk = 10%, and LTPD = 5%. Sample size, $n = 132$; acceptance number, $c = 3$; and lot size, $N = 100,000$

a lot is accepted, it is easy to see that the formula for AOQ is given by:

$$AOQ = \frac{P_a p(N - n)}{N} \quad (4)$$

The AOQ curve for the $c = 3$, $n = 132$ plan where the lot size is $N = 100,000$ units is given in Fig. 2.

Note that the AOQ curve obtains a maximum of 0.015 when the fraction defective is 0.023. This maximum is called the Average Outgoing Quality Limit (AOQL), and it represents the poorest average quality level achievable under a rectification inspection program. This means that no matter what quality level the process is producing, the average defect rate over a large number of lots will be no greater than 1.5%.

The last graph that needs mentioning is the Average Total Inspection (ATI) graph. This graph shows the average number of samples that need to be taken under a rectification inspection program. If lots contain no defective items, then the number of samples per lot will be n , while if the lots contain all defective items then the number of samples will be N . If the proportion of defective items is between these two extremes, then the number of samples will be between n and N . In this case, the ATI per lot is given by the equation:

$$ATI = n + (1 - P_a)(N - n) \quad (5)$$

The ATI graph for the $c = 3$, $n = 132$ plan where the lot size is $N = 100,000$ units is given in Fig. 3.

DOUBLE SAMPLING: NOT AN AD HOC EXTENSION OF SINGLE SAMPLING

Up to this point, we have discussed single sampling plans, where a decision on the lot is based on one sample. However, there are formal double sampling plans whereby,

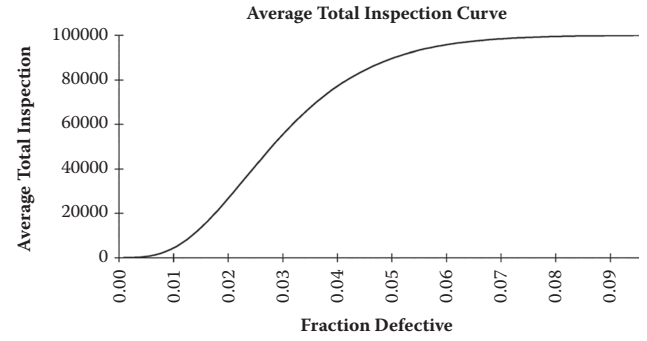


Fig. 3 The ATI curve for a single acceptance sampling plan with producer's risk = 5%, AQL = 1%, consumer's risk = 10%, and LTPD = 5%. Sample size, $n = 132$; acceptance number, $c = 3$; and lot size, $N = 100,000$

based on the results observed in the initial sample, a second sample may be required. A double sampling plan consists of the following components:

- Sample size of the first sample (n_1);
- Acceptance number for first sample (c_1);
- Rejection number for first sample (r_1);
- Sample size of the second sample (n_2);
- Acceptance number for the first and second samples combined (c_2); and
- Rejection number for the first and second samples combined (r_2 ; $r_2 = c_2 + 1$).

Fig. 4 provides a schematic of the double sampling acceptance plan. The producer's risk, AQL, consumer's risk, and LTPD are the necessary input, as in the case of single sampling, in order to develop a double sampling plan.

The main advantage of double sampling over single sampling is that in most cases, double sampling will reduce the average number of samples (Average Sample Number, ASN) taken. The greatest potential for reducing the average number of samples taken in a double sampling plan is when sampling is stopped (called *curtailment*) when the total number of defects is greater than the second rejection number. To see the relationship between ASN for single and double sampling plans, consider the ASN curve

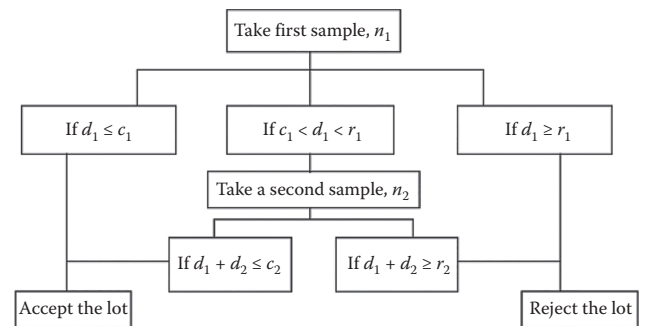


Fig. 4 Operating flow chart for double acceptance sampling program

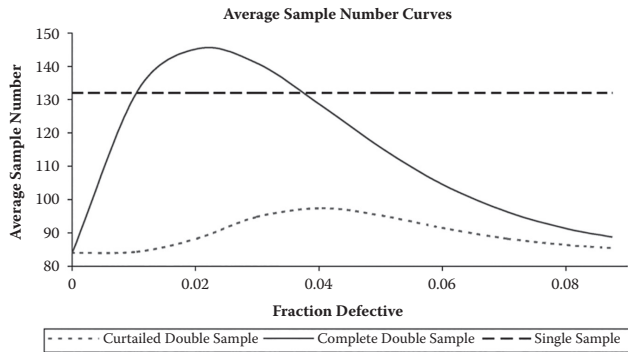


Fig. 5 The ASN curve for complete double sampling, curtailed double sampling, and single sampling plans with properties: producer's risk = 5%, AQL = 1%, consumer's risk = 10%, and LTPD = 5%

graphed in Fig. 5. When curtailment is followed, gains in sample size reduction can be made over single sampling plans, but when complete inspection is performed, the gains may not be realized. Another advantage of double sampling plans is that they may give the user a psychological advantage knowing that if the first sample does not allow for acceptance, then the second sample may. The disadvantage of double sampling plans is that they are more complicated to administer.

Double sampling is not an ad hoc extension of single sampling. For example, let us assume that a quality assurance organization instituted a single sampling plan, but decided to take a second sample after the number of nonconforming units observed was greater than the predefined acceptance number. This type of ad hoc double sampling plan results in an inflation of the consumer's risk. Double sampling plans can be extended to multiple sampling plans, whereby more than two samples could possibly be taken. An even greater extension to double and multiple sampling plans is the sequential sampling plan. Schilling^[5] and Stephens^[6] provide an in-depth discussion on these types of sampling plans.

ACCEPTANCE SAMPLING PLAN FOR VARIABLES

Variable acceptance sampling plans are used when the quality characteristic can be measured on a continuous scale. Examples of variables are tablet weight, tablet hardness, and bottle measurements. Variables contain much more information about the quality level than do attributes. For example, if the weight of a tablet is to be between 2 and 6 mg, then reporting that a tablet weighs 2.1 mg is more informative than reporting that the tablets weight is within specification. The fact that much more information can be gained from a variable over an attribute is reason enough to argue that one should never artificially change a variable into an attribute.

The increase in information of a variable over an attribute is the main advantage of variables sampling over

attributes sampling and leads to much smaller sample sizes for variables plans over attributes plans. Consider the example again where it is necessary to have a plan with a producer's risk of 0.05, an AQL of 1%, a consumer's risk of 0.10, and an LTPD of 5%. The variable sampling plan requires only 70 samples, whereas the corresponding attributes sampling plan requires 132 samples.

The main advantage of attributes sampling over variables sampling is that in general, attributes sampling is less time-consuming and requires less expensive testing equipment. Also, the amount of record-keeping and the calculations for variables are often thought of as being more difficult to manage than for attributes. Occasionally, the advantage of simplicity of attribute plans can offset the reduction in sample size due to variables. Another advantage of attribute sampling is that the mathematical assumptions about how quality varies from one item to another are fulfilled as long as a random sample is taken from the lot. On the other hand, the mathematical assumption of variables plans is that quality varies from one item to another in the form of a normal distribution. Whether or not the normal distribution adequately fits the underlying distribution of a variable quality characteristic should be determined through appropriate statistical methods.

Designing a variables sampling plan is similar to designing an attribute sampling plan in that the following four quantities need to be specified: producer's risk (α), AQL ($p_1 = \text{AQL}/100$), consumer's risk (β), and LTPD ($p_2 = \text{LTPD}/100$). From these values, the sample size n and the constant k , which are then used to carry out the plan, can be determined. To carry out the plan, a sample of size n is taken and then the sample mean $\bar{x}$ and sample standard deviation s are calculated. Then for a quality characteristic with a single specification limit, say an USL, the criteria for acceptability is that $\bar{x} + ks \leq \text{USL}$. The formula for k is given by:

$$k = \frac{z_{p_2} z_\alpha + z_{p_1} z_\beta}{z_\alpha + z_\beta} \quad (6)$$

where z_p is the standard normal quantile of the order $(1 - p)$. That is, if $z < N(0,1)$, then $p = \text{Pr}(Z > z_p)$. The formula for n depends on whether or not the variance is assumed to be known or unknown. In general, the recommendation is to assume that the variance is unknown and needs be estimated from the sample because in practice, the variance can change from lot to lot. However, if one has a very large amount of historical data and can show that the variance is constant from lot to lot, then one may wish to assume that the variance is known by using an estimate from the historical data. The result of assuming that the variance is known is a reduction in sample size. When the variance is assumed to be *known*, the formula for n is given by

$$n = \left(\frac{z_\alpha + z_\beta}{z_{p_1} + z_{p_2}} \right)^2 \quad (7)$$

and when the variance is assumed to be *unknown*, the formula for n is given by

$$n = \left(1 + \frac{k^2}{2}\right) \left(\frac{z_\alpha + z_\beta}{z_{p_1} - z_{p_2}}\right) \quad (8)$$

SINGLE-SIDED SPECIFICATION LIMITS

For single-sided specifications, the acceptance regions can be characterized algebraically.

Sigma Known Testing Criteria

- USL: Accept if $\bar{x} + k\sigma \leq \text{USL}$; reject otherwise.
- LSL: Accept if $\bar{x} - k\sigma \geq \text{LSL}$; reject otherwise.

Sigma Unknown Testing Criteria

- USL: Accept if $\bar{x} + k\sigma \leq \text{USL}$; reject otherwise.
- LSL: Accept if $\bar{x} - k\sigma \geq \text{LSL}$; reject otherwise.

TWO-SIDED SPECIFICATION LIMITS

When there are two-sided specification limits, more care needs to be taken in finding the acceptance region. Simply taking the intersection of an LSL acceptance region and a USL acceptance region will not always give the correct acceptance region.

A two-sided acceptance region is most easily shown on a graph. Fig. 6 shows the acceptance region for the plan, assuming unknown variance, with a producer's risk of 0.05, an AQL of 1%, a consumer's risk of 0.10, and an LTPD of 5%. For each sample taken, the sample mean and the sample standard deviation are plotted, and if the point lies below the curve and above the horizontal axis, then the lot is accepted.

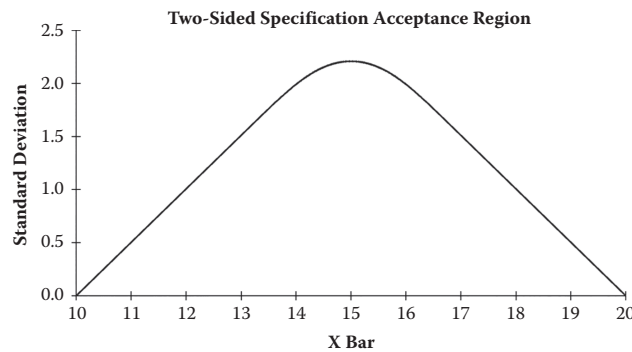


Fig. 6 The acceptance region for a variables sampling plan with two-sided specifications (LSL = 10 and USL = 20) and producer's risk = 5%, AQL = 1%, consumer's risk = 10%, and LTPD = 5%

Wallis^[4] provides the mathematical algorithm for finding the acceptance region boundary. The first step is to compute k and n from Eqs. 6 and 7, or Eq. 8. In this case, the p_1 and p_2 values represent the fraction of items outside of the interval LSL-USL. The second step is to find the number p_0 such that $z_{p_0} = k$. The third step is to divide p_0 into two parts p'_0 and p''_0 such that $p_0 = p'_0 + p''_0$. The fourth step is to calculate:

$$\bar{x} = \frac{\text{USL} * z_{p'_0} + \text{LSL} * z_{p''_0}}{z_{p'_0} + z_{p''_0}} \quad \text{and} \quad s = \frac{\text{USL} - \text{LSL}}{z_{p'_0} + z_{p''_0}} \quad (9)$$

which are points that lie on the acceptance region boundary; another point on the acceptance region boundary is $(\bar{x}', s)$ where $\bar{x}' = \text{USL} + \text{LSL} - \bar{x}$.

COMMON ACCEPTANCE SAMPLING MISCONCEPTIONS

One of the most common misconceptions about acceptance sampling plans concerns the misinterpretation of the AQL. Many users of sampling plans interpret the AQL as if it were the LTPD. For example, a user of a sampling plan with an AQL equal to 1% and a producer's risk equal to 5% will often make the incorrect statement that one is 95% confident that the accepted lots have a defect level less than 1%. The correct confidence statement about the defect level of accepted lots needs to be based on the LTPD and the consumer's risk. For example, if the sampling plan just mentioned has an LTPD equal to 5% and a consumer risk equal to 10%, then one can say with 90% confidence that the lots that are accepted have a defect level less than 5%.

Another misconception about sampling plans is that sample size should be proportional to lot size. One incorrect rule of thumb for developing sampling plans is that the sample size should be equal to 10% of the lot size with an acceptance number of zero. The presumption is that following this type of rule provides plans with identical properties for any lot size. However, this is not the case as depicted in Fig. 7, where we have OC curves for two zero acceptance sampling plans that follow the 10% rule. The first plan is derived by assuming a lot size of 10,000 with a sample size of 1,000 and the second plan is derived by assuming a lot size of 1,000 with a sample size of 100. In order for these two plans to have identical properties, their associated OC curves would need to be identical. It is clear from the graph that this is not the case and that the sampling plan with a sample size of 1,000 is much more sensitive to rejecting "poor-quality" lots.

Another misconception encountered is the square root of $N + 1$ (or square root of N) rule for determining the sample size for a sampling plan, where N is the lot size. There is no apparent statistical justification for this rule to be used in general.

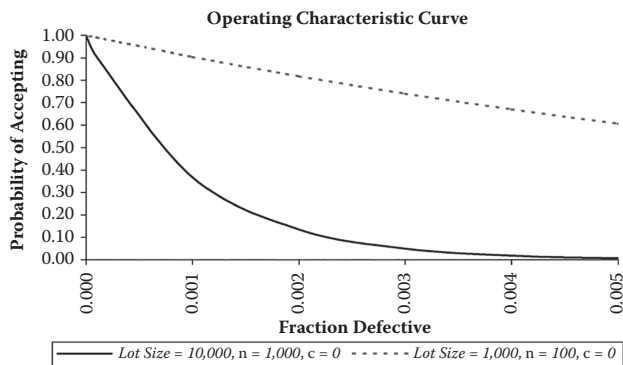


Fig. 7 The OC curve for two zero acceptance sampling plans for which sample size is equal to 10% of lot size

The final misconception is that in general, lot size should be used to determine a sampling plan. Theoretically, because all lots are of finite size, the hypergeometric distribution is the one that provides the correct model. The hypergeometric distribution does take into account the lot size and so any acceptance sampling plan developed under the hypergeometric distribution will need to take lot size into consideration. However, in most applications, the binomial distribution approximates the hypergeometric distribution so closely that lot size does not play a *significant* role in determining a sampling plan. A general guideline to determine if lot size should be taken into consideration is to calculate n/N , where n is the sample size and N is the lot size, and to see if this quantity is less than or equal to 0.1. That is, if the sample size is greater than 10% of the lot size, then the lot size should be considered when determining a sampling plan. However, a reasonable approximation is obtained even beyond this guideline.

THOUGHTS ON THE ANSI/ASQC SAMPLING PROCEDURES

Another popular method for developing an acceptance sampling plan is to use the ANSI/ASQC Z1.4: Sampling Procedures and Tables for Inspection by Attributes,^[3] formally the MIL STD 105E. This set of tables allows one to derive an acceptance sampling plan that is based on batch size, inspection level, and AQL. Experience has shown that many of those who implement these plans do not completely understand their properties. These sampling plans are AQL-oriented, in that they focus attention on the producer's risk. This means that as long as the process average of non-conforming units is at or below the pre-specified AQL, the probability of acceptance will be high (at or above 90%). However, these sampling plans do not directly address the consumer's risk and, consequently, do not necessarily reject "poor-quality" lots with high probability. This should be clearly understood by those who use these sampling plans. In order to understand

the consumer's risk, one needs to observe the OC curve for the given sampling plan. There are OC curves available in the ANSI/ASQC Z1.4. Also, a common misuse of ANSI/ASQC Z1.4 is failure to use the switching rule. The general concept of the switching rule is that if lots are continuously accepted, then a "reduced" sampling plan is instituted. However, if lots are continuously rejected, then a "tightened" sampling plan is established. By invoking the switching rules, the consumer risk will be reduced. Consequently, switching rules should be implemented when using these procedures. Finally, we want to mention that tables for deriving variables sampling plans can be found in the ANSI/ASQC Z1.9: Sampling Procedures and Tables for Inspection by Variables for Percent Non-conforming.^[22]

CONCLUSION

Acceptance sampling still has a role in the pharmaceutical industry. In reality, inspection is seen to have value when compared to the catastrophic results of performing no inspection when an unsatisfactory level of quality exists. In addition, acceptance sampling is an audit tool to ensure that the output of a process conforms to requirements. Therefore, it is essential for those implementing acceptance sampling plans to fully understand their properties and to be aware of their common misconceptions. Experience shows that the most common misconception is that the AQL has the interpretation of the LTPD. This may lead to overconfidence in the acceptance sampling plans' ability to reject lots of poor quality. This certainly applies when one selects an acceptance sampling plan from the ANSI/ASQC Sampling Procedure.

REFERENCES

1. *United States Pharmacopeia 25 and National Formulary 20*; United States Pharmacopoeial Convention: Rockville, MD, 2001.
2. FDA. *Current Good Manufacturing Practice (cGMP) Regulations: Division of Manufacturing and Product Quality (HFD-320)*; 1996.
3. *ANSI/ASQC Z1.4-1993: American National Standard: Sampling Procedures and Tables for Inspection by Attributes*; ASQ Quality Press: Milwaukee, WI, 1993.
4. Wallis, W.A. *Use of Variables in Acceptance Inspection for Percent Defective. Techniques of Statistical Analysis*; McGraw-Hill: New York, 1947; 3–93.
5. Schilling, E.G. *Acceptance Sampling in Quality Control*; Marcel Dekker: New York, 1982.
6. Stephens, K.S. *The Handbook of Applied Acceptance Sampling: Plans, Procedures and Principles*; ASQ Quality Press: Milwaukee, WI, 2001.
7. Sherman, W.H. *Inspection Through the Years. ASQC 50th Annual Quality Congress Proceedings, 1996*; 571–574.

8. Wetherill, G.C. Bayesian solution of single sample inspection. *Technometrics* **1960**, 2, 341–352.
9. Hald, A. Bayesian single sampling attribute plans for continuous prior distributions. *Technometrics* **1968**, 10 (4), 667–683.
10. Moskowitz, H.; Berry, W. A Bayesian algorithm for determining optimal single sample acceptance plans for product attributes. *Manag. Sci.* **1976**, 22 (11), 1238–1250.
11. Thyregod, P. Bayesian single sampling acceptance plans for finite lot sizes. *J. R. Stat. Soc. Ser. B* **1974**, 36, 305–319.
12. Hald, A. The compound hypergeometric distribution and a system of single sampling inspection plans based on prior distribution and costs. *Technometrics* **1960**, 2, 275–340.
13. Guenther, W.C. On the determination of single sampling attribute plans based upon a linear cost model and a prior distribution. *Technometrics* **1971**, 13 (3), 483–498.
14. Lorenzen, T.J. Minimum cost sampling plans using Bayesian methods. *Nav. Res. Logist. Q.* **1985**, 32, 57–69.
15. Wadsworth, H.M.; Olsen, M.E. A cost model for acceptance sampling. *AIIE Trans.* **1976**; 349–355.
16. Hahn, G.J. Estimating the percent non-conforming in the accepted product after zero defect sampling. *J. Qual. Technol.* **1986**, 18 (3), 182–188.
17. Pesotchinsky, L. Plans for very low fraction non-conforming. *J. Qual. Technol.* **1987**, 19 (4), 191–196.
18. Moreno, C.W.; Reeves, A.T. Acceptance sampling for small fraction defectives. *Pharm. Technol.* **1979**, 67 (3), 41–51.
19. Suich, R. The effects of inspection errors on acceptance sampling for nonconformities. *J. Qual. Technol.* **1990**, 22 (4), 314–318.
20. Vander Wiel, S.A.; Vardeman, S.B. A discussion of all-or-none inspection policies. *Technometrics* **1994**, 36 (1), 102–109.
21. Montgomery, D.C. *Introduction to Statistical Quality Control*; John Wiley: New York, 1997.
22. *ANSI/ASQC Z1.9-1993: American National Standard: Sampling Procedures and Tables for Inspection by Variables for Percent Non-conforming*; ASQ Quality Press: Milwaukee, WI, 1993.

Active Control Non-Inferiority Trials: Hypotheses and False Positive Rate

Gang Chen and George Y. H. Chi

Johnson & Johnson Pharmaceutical Research & Development, Raritan, New Jersey, U.S.A.

Yong-Cheng Wang

Division of Biometrics I, Office of Biostatistics, Office of Pharmacoepidemiology and Statistical Science, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.

Abstract

Hypotheses and False Positive Rate in Active Control Non-Inferiority Trials. The hypotheses in non-inferiority trials involve two unknown parameters. Statistical inferences on non-inferiority hypotheses, ideally, should be based on randomized data collected in the current trial. This entry reviews some of the ethical issues in active control non-inferiority trials and presents statistical models for performing this type of trial.

INTRODUCTION

The purpose of clinical trials in drug development is usually to demonstrate that a new treatment is better than placebo or best supportive care. The hypotheses in such clinical trials involve only one parameter of interest, e.g., treatment effect (relative to placebo or best supportive care). For ethical reason, however, when there are available known effective treatments for the disease, investigators should use one of these active treatments instead of a placebo as the control therapy (declaration of Helsinki^[1]). More and more active controls have been used in clinical trials without a placebo. This naturally leads to an interest in demonstrating the effectiveness of a new treatment through an active control non-inferiority trial. Readers may refer to Chi, Chen, and Rothmann^[2] and Temple^[3] for a discussion of the major issues in the design and analysis of non-inferiority trials. The objective of an active control non-inferiority trial is to demonstrate the efficacy or equivalence of a new treatment through the test of preserving a given fraction or margin of the control effect.

An active control is an intervention such as a drug, therapy, or medical device whose effectiveness has previously been established, mostly based on historical, randomized clinical trials. In an active control non-inferiority trial, the assumption that the active control is effective is fundamental. When this assumption does not hold, a harmful (worse than placebo) drug may be approved based on the results

of non-inferiority test. In current non-inferiority trials, this fundamental assumption has been used either in the formulation or in the inference of the hypotheses. Since the true active-control effect is unknown, the assessment of the fundamental assumption is usually based on historical randomized trials that have demonstrated positive control effect. However, the conclusion of those trials may be falsely positive, and the false positive rate for claiming the control effect in those trials is usually controlled at a given significance level, say 5%. In other words, the assumption may be wrong with a certain probability. A wrong assumption may lead to an invalid non-inferiority trial design and a false positive conclusion. The false positive rate associated with the objective of a non-inferiority trial may not be correctly assessed if the assumption is wrong. We discuss in this paper various possible hypotheses formulated in non-inferiority trials. The false positive rate for concluding non-inferiority is evaluated through a simulation. The hypotheses appropriately associated with the objective of a non-inferiority trial are suggested.

HYPOTHESES IN ACTIVE CONTROL NON-INFERIORITY TRIALS

Since there are two parameters involved in non-inferiority hypotheses, the hypotheses can be formulated in different ways. The appropriateness of the hypotheses is dependent on the objective of the non-inferiority trial. In this section, we present the possible hypotheses in non-inferiority trials and discuss their appropriateness in terms of the parameter spaces under the null and the alternative hypotheses.

The views expressed in this entry are those of the authors and not necessarily those of the U.S. Food and Drug Administration.

Parameter Space

In order to understand non-inferiority hypotheses and the associated probability of the type I error, in this section, we present parameter spaces under the null and the alternative non-inferiority hypothesis. Let T and C denote the new treatment and active control, respectively. The concepts and issues are discussed here within the framework of a randomized trial comparing the survival experience of a new treatment T to a single active control C in treating patients with a serious disease. Let the true hazard ratio of T relative to C be denoted by $HR(T/C)$. We define the fraction loss of control effect as

$$1 - \delta = [HR(T/C) - 1] / [HR(P/C) - 1] = \theta_t / \theta_c$$

where $\theta_t = HR(T/C) - 1$, $\theta_c = HR(P/C) - 1$, and P stands for the reference placebo. In the definition of the fraction loss $(1 - \delta)$, θ_c represents the control effect (relative to placebo) and θ_t the treatment effect relative to the control. Fig. 1 shows the partition of two coordinate systems $(HR(P/C), HR(T/C))$ and (θ_c, θ_t) , respectively. Coordinate system (θ_c, θ_t) is a location shift of the coordinate system $(HR(P/C), HR(T/C))$.

In Fig. 1, the relationship between the two treatments, T and C , is presented in each subspace. The notation “>” used in Fig. 1 and the following figures stands for “better than.” Since the primary interest in a non-inferiority trial is to test whether the new treatment T is effective, an ideal parameter space under the alternative in non-inferiority hypotheses should contain the subspace that $T > P$ (T is better than P).

Superiority Hypotheses for Treatment and Control Effect

If there is a placebo arm in the trial, the effectiveness of the new treatment T can be demonstrated through comparison with the placebo using the following superiority hypotheses:

$$H_0 : \mu_t \leq 1 \quad \text{vs.} \quad H_a : \mu_t > 1 \quad (1)$$

where $\mu_t = HR(T/P)$ denotes the hazard ratio of treatment vs. placebo.

When placebo cannot be used in a trial due to ethical concern, the effectiveness of a new treatment T needs to be

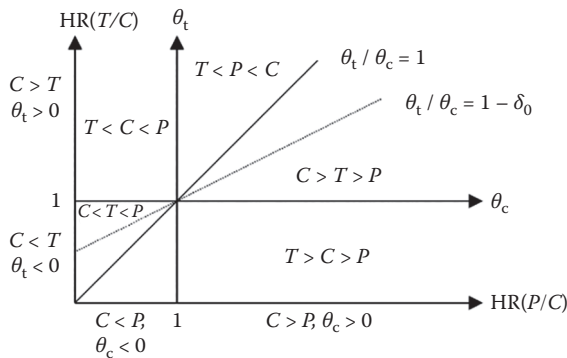


Fig. 1 The partition of the parameter space

demonstrated through proving either that T is superior to the active control C , or that T is non-inferior to C . The term “non-inferior” means that the new treatment T preserves a fraction of the control effect or the new treatment is not worse than control by a given (fixed) margin, assuming the control is effective. Unless the new treatment represents a significant therapeutic advance, in general it would be difficult for it to show superiority to the active control.

The control effect θ_c is usually assessed in randomized, historical trials with superiority hypotheses:

$$H_0 : \theta_c \leq 0 \quad \text{vs.} \quad H_a : \theta_c > 0 \quad (2)$$

If the effectiveness of the control has been demonstrated in those historical trials, one may consider a non-inferiority design of a trial.

Ratio Non-Inferiority Hypotheses

A non-inferiority hypothesis generally involves two parameters, i.e., treatment effect θ_t and control effect θ_c . The direct translation of the non-inferiority objective to statistical hypotheses is the following ratio hypotheses:

$$H_0 : \theta_t / \theta_c \geq 1 - \delta_0 \quad \text{vs.} \quad H_a : \theta_t / \theta_c < 1 - \delta_0 \quad (3)$$

where δ_0 ($0 \leq \delta_0 \leq 1$) denotes a given level of fraction retention.

Fig. 2 presents the parameter space under the null and the alternative of the ratio hypotheses in (3).

In some clinical settings, one may fix the control effect in non-inferiority hypotheses, i.e., $\theta_c = M_1 > 0$, and formulate a fixed margin non-inferiority hypotheses. The validity of the fixed margin hypotheses depends on the assumption $\theta_c > 0$. If this assumption does not hold, the fixed margin hypotheses are not valid for a non-inferiority trial. In the fixed margin setting, the ratio non-inferiority hypotheses can be simplified to hypotheses with one parameter:

$$H_0 : \theta_t / M_1 \geq 1 - \delta_0 \quad \text{vs.} \quad H_a : \theta_t / M_1 < 1 - \delta_0$$

where $\theta_t = HR(T/C) - 1$.

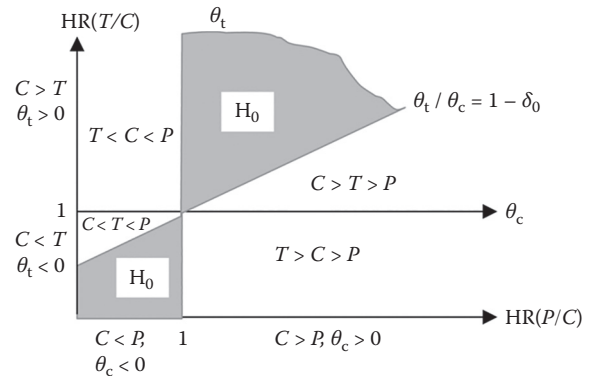


Fig. 2 The parameter space under the null and alternative hypotheses in (3)

Since $M_1 > 0$, the above hypotheses are equivalent to

$$\begin{aligned} H_0 : \text{HR}(T/C) &\geq 1 + M \quad \text{vs.} \\ H_a : \text{HR}(T/C) &< 1 + M \end{aligned}$$

where $M = M_1(1 - \delta_0) > 0$ is the margin defined based on control effect.

Fixed margin hypotheses have been used in some oncology trials with the response rate as the primary endpoint, as well as in thrombolysis trials.^[7] The interpretation of the fixed margin hypotheses and fraction retention hypotheses are different. A detailed discussion is given in Refs. [2,4].

Linear Transformation

Since it is difficult to draw a statistical inference on such ratio hypotheses, a transformation of the ratio hypothesis is often needed. In the paper by Rothmann et al.,^[4] the following linear hypotheses were formulated:

$$\begin{aligned} H_0 : \theta_t - (1 - \delta_0)\theta_c &\geq 0 \quad \text{vs.} \\ H_a : \theta_t - (1 - \delta_0)\theta_c &< 0 \end{aligned} \quad (4)$$

The assumption that $\theta_c > 0$ is used to show how the linear hypotheses in (4) are evolved from the definition of fraction retention (ratio hypotheses). This assumption is not used in the inference of the linear hypotheses.^[4,5] The parameter spaces under the null and the alternative hypothesis in (4) are presented in Fig. 3.

Fig. 3 illustrates that the parameter space under the alternative hypothesis in (4) contains an inappropriate subset $C < T < P$, which should be located in the parameter space under the null hypothesis in non-inferiority trials. The claim of non-inferiority based on the rejection of the null hypotheses in (4) leads to an incorrect assessment of the false positive rate.

It is questionable to make the assumption $\theta_c > 0$ in the formulation of non-inferiority hypotheses. Fig. 4 presents the parameter spaces under the null and the alternative hypothesis in (4) with the assumption that $\theta_c > 0$.

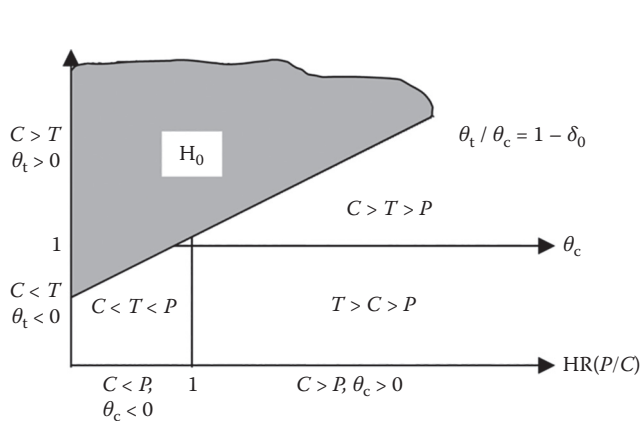


Fig. 3 The parameter space under the null and alternative hypotheses in (4)

Assuming $\theta_c > 0$, the parameter space under the hypotheses has been reduced, i.e., the left column ($\theta_c < 0$) in the parameter space has been truncated. The shaded area is the parameter space under the null hypothesis in (4) with the assumption that $\theta_c > 0$. The non-inferiority hypotheses with this reduced parameter space ($\theta_c > 0$) are questionable since the data for the statistical inference of the hypotheses are sampled from distributions with population parameters for the control and the treatment over the entire parameter space ($\text{HR}(P/C) > 0$, $\text{HR}(T/C) > 0$). The reduction of the parameter space leads to an inappropriate statistical inference for a non-inferiority trial.

If $\theta_c > 0$ does not hold, the hypotheses in (4) are not appropriate for non-inferiority trials since the parameter space under the alternative contains the subset that $T < C < P$, i.e., the new treatment is inferior to placebo. This subset should be located in the parameter space under the null in non-inferiority trials.

Other Type of Transformation

To avoid using the assumption that $\theta_c > 0$, one may have another type of transformation of a non-inferiority hypothesis. For example, Wang, Chen, and Chi^[6] considered the following transformation of non-inferiority hypotheses:

$$\begin{aligned} H_0 : \ln(\theta_t/\theta_c + k)^2 &\geq \eta \quad \text{vs.} \\ H_a : \ln(\theta_t/\theta_c + k)^2 &< \eta \end{aligned} \quad (5)$$

where $k > 0$ and $\eta > 0$. The details of the hypotheses test are not discussed in this article.

Union Non-Inferiority Hypotheses

As discussed above, both linear and ratio hypotheses are not appropriate for a non-inferiority trial in terms of the partition of their parameter spaces under the null and

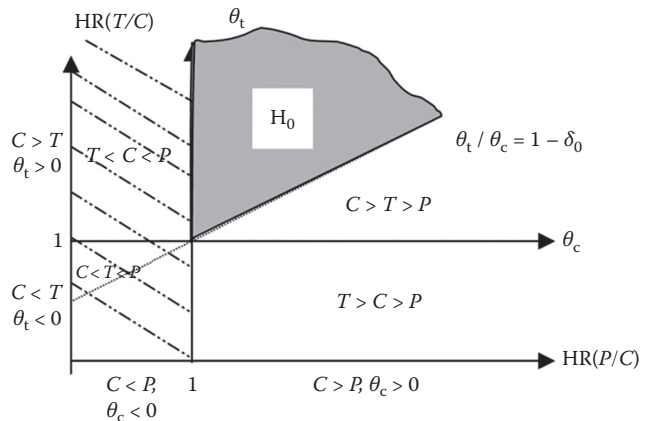


Fig. 4 The parameter space under the null and alternative hypotheses in (4) with the assumption $\theta_c > 0$

the alternative hypothesis. The following union hypotheses for non-inferiority trials may be considered:

$$\begin{aligned} H_0 &: \{\theta_t - (1 - \delta_0)\theta_c \geq 0\} \cup \{\theta_c \leq 0\} \quad \text{vs.} \\ H_a &: \{\theta_t - (1 - \delta_0)\theta_c < 0\} \cap \{\theta_c > 0\} \end{aligned} \quad (6)$$

Fig. 5 presents the parameter spaces under the null and the alternative hypothesis in (6).

The hypotheses in (6) fit the objective of non-inferiority trials. The parameter space under the alternative contains both subsets $P < T < C$ and $T > C > P$, i.e., the new treatment is either better than the active control or preserves a fraction of the control effect (non-inferior to the active control).

If the purpose of a non-inferiority trial is only to demonstrate that the experimental treatment is effective, the hypotheses may be formulated as follows:

$$\begin{aligned} H_0 &: \{\theta_t/\theta_c \geq 1 - \delta_0\} \cup \{\{\theta_c \leq 0\} \cap \{\theta_t/\theta_c \geq 1\}\} \\ \text{vs.} \\ H_a &: \{\theta_t/\theta_c < 1 - \delta_0\} \cap \{\{\theta_c > 0\} \cup \{\theta_t/\theta_c < 1\}\} \end{aligned} \quad (7)$$

Fig. 6 presents the parameter spaces under the null and the alternative hypothesis in (7).

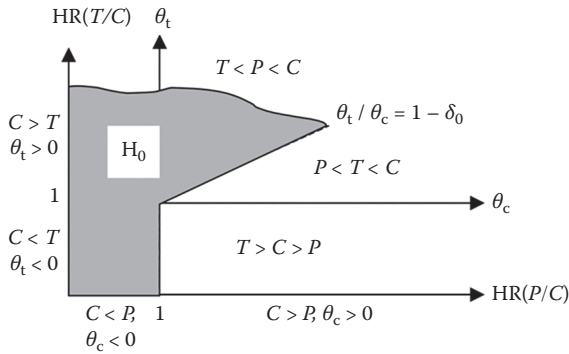


Fig. 5 The parameter space under the null and alternative hypotheses in (6)

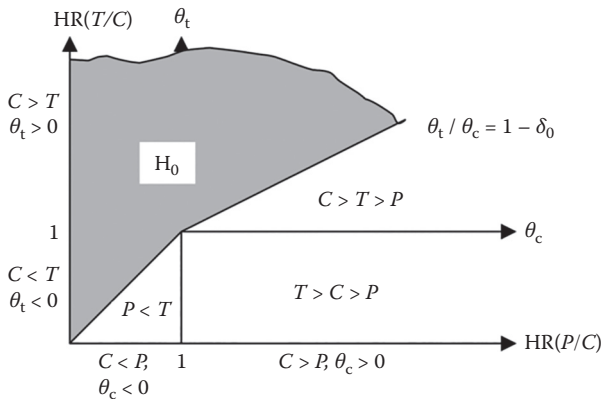


Fig. 6 The parameter space under the null and alternative hypotheses in (7)

The hypotheses in (7) may be ideal for testing the efficacy of the new treatment through a non-inferiority trial, since the parameter space under the alternative contains the entire subset $T > P$.

Of the above non-inferiority hypotheses, the hypotheses in (6) may be appropriately associated with the objective of a non-inferiority trial. The rejection of the null in (6) leads to the conclusion that the new treatment is effective. However, since the statistical inference for both hypotheses (6) and (7) involves a direct comparison of treatment with placebo, which is not available in the trial, the validity of the statistical inference is the major concern.

FALSE POSITIVE RATE CONSIDERATION

Since the assessment of the control effect θ_c is based on data collected from randomized historical trials, the evaluation of the false positive rate associated with the non-inferiority test procedure for claiming non-inferiority should consist of two components. The first component is the probability of wrong rejection of the null in the non-inferiority hypotheses in current non-inferiority trial. The second component is the probability that the control is not effective as assessed in historical trials. In current non-inferiority trials, however, the probability that the control is not effective has not been taken into the consideration in the non-inferiority test procedure. More discussion is given in the following section. As an example, in section “Simulation,” the simulation results demonstrate the magnitude of the false positive rate inflation associated with the non-inferiority test procedure when using the linear non-inferiority hypotheses in (4) under the assumption that the control is effective.

Non-Inferiority Test Procedure

In a non-inferiority trial, linear hypothesis (4) ($H_0^{(4)}$) has often been formulated for testing a fraction retention of control effect. The non-inferiority test procedure can be described as follows. The first step is to assess the control effect, i.e., to test the hypothesis in (2) ($H_0^{(2)}$). If the $H_0^{(2)}$ is rejected at a prespecified α_1 level, e.g., 0.05, we conclude that the control is effective and then further test the non-inferiority hypothesis in $H_0^{(4)}$ in a non-inferiority trial. If $H_0^{(4)}$ is rejected at a prespecified α level, e.g., 0.05, we conclude “non-inferiority.” The non-inferiority test procedure is described in Fig. 7.

False Positive Rate

To assess the false positive rate, we denote by α_1 a prespecified Type I error rate for testing $H_0^{(2)}$, and α a prespecified Type I error rate for testing $H_0^{(4)}$. According to Fig. 7, the

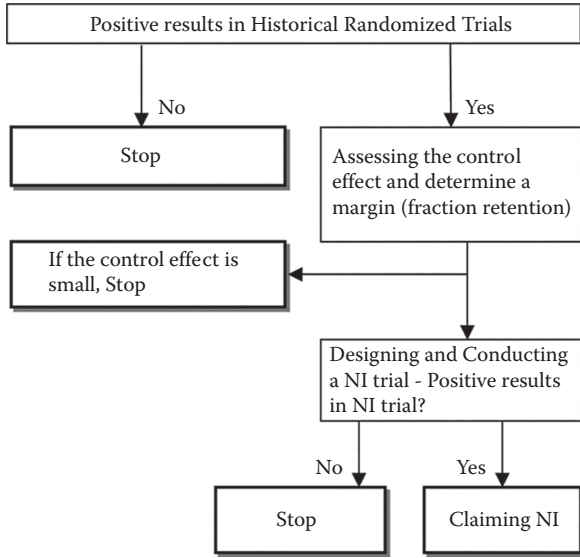


Fig. 7 Flowchart for a non-inferiority trial test procedure

false positive rate associated with the non-inferiority test procedure can be defined as below.

$$\begin{aligned}\alpha^* &= \Pr(\text{reject } H_0^{(4)} | H_0^{(4)} \text{ and control is truly positive}) \\ &\quad + \Pr(\text{reject } H_0^{(4)} | H_0^{(4)} \text{ and control is truly negative}) \\ &= \alpha + \alpha_0\end{aligned}$$

Simulation

The simulation in this section is to demonstrate the false positive rate associated with the non-inferiority test procedure when the type I error rate under linear hypothesis $H_0^{(4)}$ is controlled at a prespecified level α .

The following assumptions are used in the simulation. Let $X_1, X_2, \dots, X_m$ be independent random variables with normal distribution $N(\theta_c, 1)$, and $Y_1, Y_2, \dots, Y_n$ be independent random variables with normal distribution $N(\theta_t, 1)$, where $\theta_t = (1 - \delta_0) \theta_c$ and $0 \leq \delta_0 \leq 1$ is a constant (fraction retention). The simulation is based on 500,000 runs and both $n = m = 1000$ and $\theta_c = 0$. We define α^* to be the false positive rate associated with the non-inferiority test procedure, and we assume $\alpha_1 = \alpha = 0.05$. The simulation results are listed in Table 1.

Table 1 The False Positive Rate α^* Associated with Non-inferiority Test Procedure

δ_0	α_1	α	α^*
0	0.0499	0.0501	0.0697
0.25	0.0498	0.0503	0.0656
0.5	0.0501	0.0503	0.0607
0.75	0.0504	0.0502	0.0560
1	0.0506	0.0504	0.0526

The results in Table 1 show that the false positive rate associated with the non-inferiority test procedure is much greater than the prespecified level of Type I error α under hypothesis $H_0^{(4)}$. This inflation is attributed to the nonzero Type I error rate α_1 under hypothesis $H_0^{(2)}$ for testing the control effect in historical trials. The higher significance level specified in testing control effect results in the lower false positive rate associated with the non-inferiority test procedure. The selection of the higher fraction retention (δ_0) also results in the lower false positive rate.

Superiority Test Based on Non-Inferiority Hypothesis

From Table 1, the false positive rate inflation relative to α (type I error rate) under hypothesis $H_0^{(4)}$ is at least 0.0025 even when $\delta_0 = 1$, i.e., the new treatment is superior to control. The reason for such inflation is that we use non-inferiority hypotheses for the superiority test $T > C$. The false positive rate associated with the non-inferiority test procedure is not only controlled for testing $T > C$ (treatment is superior to control), but also controlled for testing $C > P$ (control is effective). The following two figures illustrate the parameter spaces under the null and the alternative for the superiority hypotheses in (8) and the non-inferiority hypotheses in (9) for testing whether the treatment is superior to the control.

Fig. 8 presents the parameter space under the null and the alternative for the following superiority hypotheses:

$$H_0 : \theta_t \geq 0 \quad \text{vs.} \quad H_a : \theta_t < 0 \quad (8)$$

Fig. 9 presents the parameter space under the null and alternative for the following non-inferiority hypotheses:

$$\begin{aligned}H_0 : \{\theta_t \geq 0\} \cup \{\theta_c \leq 0\} \quad \text{vs.} \\ H_a : \{\theta_t < 0\} \cap \{\theta_c > 0\}\end{aligned} \quad (9)$$

The superiority of T over C can be demonstrate based on either hypothesis (8) or (9). The higher Type I error rate

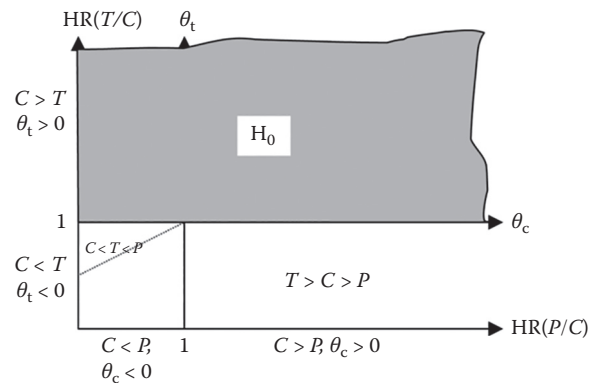


Fig. 8 The parameter space under the null and alternative hypotheses in (8)

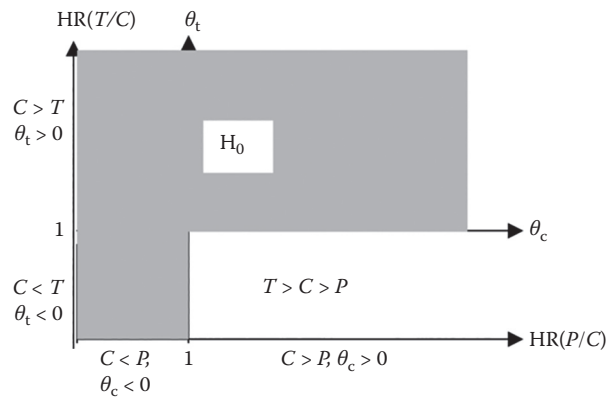


Fig. 9 The parameter space under the null and alternative hypotheses in (9)

associated with non-inferiority hypotheses (9) is because the parameter space under the null in (9) contains the subset $\{\theta_c \leq 0\} \cap \{\theta_t \leq 0\}$, which is located in the parameter space under the alternative in (8).

CONCLUSIONS AND DISCUSSION

The objective of an active control non-inferiority trial is to demonstrate the effectiveness or non-inferiority of a new treatment through comparison with an active control. The hypotheses in non-inferiority trials involve two unknown parameters. Statistical inferences on non-inferiority hypotheses, ideally, should be based on randomized data collected in the current trial. Due to ethical considerations, however, there is no placebo arm included in current non-inferiority trials, and the inference on the control effect is usually based on historical randomized data. There have been intensive discussions on those issues when involving nonconcurrent randomized trial data in the current statistical inference of non-inferiority hypotheses.^[2-5] The major issues include whether the patient population used in historical trials is similar to that used in the current non-inferiority trial in terms of patients characteristics, whether the control effect has changed over time (the constancy of the control effect), whether the best supportive care has been changed, etc. Therefore, the design of a non-inferiority trial should be done on a case-by-case basis. We need to have assurance that the type I error in the historical randomized trials for testing active-control effect is well controlled. We should have multiple historical randomized placebo controlled studies that show relatively consistent results.

If the claim of non-inferiority is based on testing the linear null hypothesis in (4) using the test statistic Z as

defined in,^[4] it generally inflates the false positive rate associated with the non-inferiority test procedure. This has been demonstrated from simulation. Since the control effect is tested based on historical trials, we cannot guarantee that the true control effect may not be negative. For a fraction retention hypothesis, the assumption that $\theta_c > 0$ is only used to show how the linear null hypothesis in (4) is evolved from the definition of fraction retention.^[4] However, it is critical to note that for fixed margin hypotheses, it is essential that the assumption, $\theta_c > 0$, hold; otherwise, one cannot define a positive margin. Therefore, fixed margin hypotheses would be appropriate only when either the control effect is very large, or the control effect cannot be negative, as in tumor response rate.

When designing a non-inferiority trial, the hypotheses in (4) can be formulated. The advantage of using the linear hypotheses in (4) is that statistical inference is simple. However, if the non-inferiority conclusion is based on the rejection of the null hypothesis in (4), the type I error rate should be appropriately adjusted to control the false positive rate associated with the non-inferiority test procedure at a pre-specified level. For example, based on the result given in Table 1, if the linear hypotheses in (4) are used to test 50% retention of the control effect, i.e., $\delta_0 = 0.5$, to control the false positive rate associated with the non-inferiority test procedure for the claim of “non-inferiority” at 0.05 level, the nominal α under $H_0^{(4)}$ should be prespecified at a significance level of $0.05/0.0609 = 0.041$.

REFERENCES

1. World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *J. Am. Med. Assoc.* **2000**, 284 (23), 3043–3045.
2. Chi, G.Y.H.; Chen, G.; Rothmann, M.; Li, N. Active control trials; *Encyclopedia of Biopharmaceutical Statistics*, 2nd Ed.; 2003.
3. Temple, R. Problems in interpreting active control equivalence trials. *Accountabil. Res.* **1996**, 4, 267–275.
4. Rothmann, M.; Li, N.; Chen, G.; Chi, G.Y.H.; Tsou, H.-H.; Temple, R. Design and analysis of non-inferiority mortality trials in oncology. *Stat. Med.* **2003**, 22, 239–264.
5. Holmgren, E.B. Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained. *J. Biopharm. Stat.* **1999**, 9 (4), 651–659.
6. Wang, Y.C.; Chen, G.; Chi, G. Design and analysis of active control non-inferiority trials with a survival endpoint. Submitted to *Stat. Med.*
7. White, H.D. Thrombolytic therapy and equivalence trials. *J. Am. Coll. Cardiol.* **1998**, 31, 494–496.

Active Control Trials

George Y. H. Chi and Gang Chen

*Johnson and Johnson Pharmaceutical Research and Development, Raritan,
New Jersey, U.S.A.*

Mark Rothmann

U.S. Food and Drug Administration, White Oaks, Maryland, U.S.A.

Ning Li

Sanofi-Aventis, China

Abstract

An active control is an intervention, such as a drug, a therapy, or a medical device, whose effectiveness has previously been established. This entry will discuss the basic issues associated with an active control trial, and some of the issues attendant with the current methods that are used in the design and analysis of active control trials.

Keywords: Active control trials; Non-inferiority hypotheses; Sample size; Superiority hypotheses.

INTRODUCTION

An active control is an intervention, such as a drug, a therapy, or a medical device, whose effectiveness has previously been established. Active controls have been used in trials with or without a placebo. For clinical drug trials with a placebo, sometimes called trials with a gold standard design, active controls usually play a secondary role with respect to demonstrating the effect of a new treatment. In placebo control trials involving psychotropic agents, active controls are intended for the purpose of verifying the assay sensitivity of the trials (Leber^[1,2]), i.e., the ability of the trial to demonstrate an effective drug to be effective (ICH-E10^[3]). In this article, active control trial with a gold standard design will not be considered; an active control trial refers specifically to a trial with a control that is known to be effective and without the presence of a placebo. Active control trials often arise when the objective of a trial is to investigate the effect of the new treatment on mortality or serious morbidity outcome for patients with certain serious diseases. For obvious ethical reason, when there are available known effective treatments for the disease, one should use one of these active treatments instead of a placebo as the control (Declaration of Helsinki^[4]). This article will discuss the basic issues associated with an active control trial, and some of the issues attendant with the current methods that are used in the design and analysis of active control trials. The concepts and issues will be discussed within the framework of a blinded, parallel, randomized trial comparing the survival experience of a new treatment, T , to a single active control, C , in treating patients with certain serious disease.

OBJECTIVE OF AN ACTIVE CONTROL TRIAL

The primary objective of an active control trial is to demonstrate the effectiveness of a new treatment, T . To show that a new treatment, T , is effective, traditionally, it is required to demonstrate that T is superior to C , the active control. This is analogous to a placebo control trial where the new treatment, T , has to show superiority to the placebo. It often happens that upon failing to demonstrate superiority in such a trial, the experimenter then concludes that the new treatment is equivalent, similar, or non-inferior to the active control. However, this reasoning is fallacious because in a superiority trial, failing to reject the null hypothesis of inferiority or no difference does not imply that the null hypothesis is true. Unless the new treatment represents a significant therapeutic advance, in general, it would be difficult for the new treatment to show superiority to the active control. This naturally leads to an interest in demonstrating the effectiveness of a new treatment through an active control non-inferiority (or a one-sided equivalence) trial, whereby the new treatment is shown to be equivalent or no worse than the active control by a certain equivalence or non-inferiority margin δ . The question then is how does one determine this margin δ ? One may refer to Temple,^[5] EMEA^[6] and FDA^[7] for a discussion of some of the problems involved in the interpretation of active control trials. In the following sections, we will discuss how the two current methods arrive at the determination of this margin and the problems associated with these methods.

THE BASIC ASSUMPTIONS IN AN ACTIVE CONTROL TRIAL

The first critical assumption implicit in any active control trial is that the active control is still effective in the current setting. This assumption cannot be verified short of the inclusion of a placebo in the active control trial. If one doesn't believe that the active control is effective in the current setting, then one should either use a different active control that is believed to be effective in the current setting, or else demonstrates that the new treatment is superior to the active control. Therefore, for this reason as well as for obvious ethical reasons, one should use the most effective active control currently available as the control. With such an active control, one can at least feel more assured that the active control is effective in the present setting.

The second critical assumption that is also implicit in any active control trial is that the trial has assay sensitivity. Assay sensitivity refers to the ability of a trial to distinguish an effective treatment from an ineffective treatment. Without assay sensitivity, a trial cannot even detect the effect of a control that is known to be effective. Thus, with the first assumption, the active control is assumed to be effective in the current trial setting, and with assay sensitivity, the trial should be able to detect the effect of this control. In addition, if the new treatment is ineffective, then this trial should be able to differentiate it from the control.

There are generally two current methods to the determination of a non-inferiority margin. The first approach is termed the fixed margin method, and the second approach is called the fraction retention method, or sometimes termed the synthesis method (Hung et al.^[8–10] and Ng^[11]). Both assumptions are needed in the two methods. However, as will be noted later, the first assumption is implicitly made in each method in a different way.

HYPOTHESES OF AN ACTIVE CONTROL TRIAL

Let the true hazard ratio of T relative to C be denoted by $HR(T/C)$, and let $hr(T/C)$ denote its estimator. Let P denote the reference or imputed “placebo,” and $HR(P/C)$ and $hr(P/C)$ denote the corresponding hazard ratios.

Superiority Hypotheses

If there is reason to believe that the new treatment, T , represents a true therapeutic advance and one wishes to demonstrate the superiority of T over C , then the null and alternative hypotheses are simply,

$$H_0 : \log HR(T/C) \geq 1 \quad \text{vs.} \quad H_a : HR(T/C) < 1.$$

In terms of log hazard, these hypotheses may be stated as,

$$\begin{aligned} H_0 : \log HR(T/C) &\geq 0 \quad \text{vs.} \\ H_a : \log HR(T/C) &< 0 \end{aligned} \quad (1)$$

The null hypothesis in Eq. 1 may be tested by the following test statistic:

$$\tau_0 = \frac{\log hr(T/C)}{SE[\log hr(T/C)]} \quad (2)$$

where SE stands for “standard error.” The null hypothesis in Eq. 1 is rejected if $\tau_0 < -1.96$ (Wald's test), or equivalently, if $\log hr(T/C) + 1.96SE[\log hr(T/C)] < 0$, at the one-sided 2.5% significance level. As in a placebo control trial, in an active control superiority trial, there is only one parameter that will be tested. In the case here, this parameter is the log-hazard ratio, $\log HR(T/C)$, of the new treatment relative to the control.

However, if the new treatment is only as effective as the control, then the null hypothesis in Eq. 1 is unlikely to be rejected. One may not conclude that the new treatment is non-inferior (or equivalent) to the control upon failing to reject this null hypothesis, since failing to reject the null hypothesis in Eq. 1 does not imply that the null hypothesis is true.

Non-inferiority Hypotheses with a Fixed δ Margin

When the new treatment is thought to be as effective as the control and the objective of the trial is to demonstrate the effectiveness of this new treatment by showing that it is non-inferior to the control, then the null and alternative hypotheses may be stated as follows:

$$\begin{aligned} H_0 : HR(T/C) &\geq 1 + \delta \quad \text{vs.} \\ H_a : HR(T/C) &< 1 + \delta. \end{aligned}$$

where $\delta > 0$ is a fixed non-inferiority margin.^[12]

Again, in terms of log hazard, these hypotheses may be stated as follows:

$$\begin{aligned} H_0 : \log HR(T/C) &\geq \delta \quad \text{vs.} \\ H_a : \log HR(T/C) &< \delta \end{aligned} \quad (3)$$

where $\delta > 0$ is a fixed non-inferiority margin (this δ is different from the δ above) (See Remark 3).

The null hypothesis in Eq. 3 may be tested by the following test statistic.

$$\tau_0(\delta) = \frac{[\log hr(T/C) - \delta]}{SE[\log hr(T/C)]} \quad (4)$$

The null hypothesis in Eq. 3 is rejected if $\tau_0(\delta) < -1.96$, or equivalently, if $\log hr(T/C) + 1.96SE[\log hr(T/C)] < \delta$,

at the one-sided 2.5% significance level. One may then conclude that the new treatment is non-inferior to the active control within this fixed margin δ . The question is how should this δ margin be determined?

If the margin δ is arbitrarily selected, then the rejection of the null hypothesis in Eq. 3 may actually lead to the conclusion that a new treatment is non-inferior to the control when, in reality, it is worse than a placebo. This can easily be seen from the following scenario. Assume that in the current trial, the true hazard ratio is known and is given by $HR(P/C)$. If the margin δ is specified so that it is greater than the effect of the control, i.e., $\delta > \log HR(P/C)$, then the rejection of the null hypothesis in Eq. 3 does not rule out the possibility that the new treatment can be worse than the placebo, i.e., $\log HR(P/C) < \log HR(T/C)$. This is easily seen from the diagram in Fig. 1.

From Fig. 1, it is clear that, in general, in an active control trial, one should not specify the margin δ to be an arbitrarily fixed number. In particular, one should avoid specifying a δ margin that is greater than the effect of the control. How can this be accomplished?

If the true hazard ratio, $HR(P/C)$, is known, then this can be done easily by defining the δ margin to be a fraction of the control effect, e.g., $\delta = 1/2 \log HR(P/C)$. However, the true control effect is never known even if there were a placebo present in the current trial. The true control effect can only be estimated. When there is a placebo in the current trial, then the true control effect may be estimated by $\log hr(P/C)$. When there is no placebo in the current trial, the true control effect needs to be estimated from historical data. In either case, one can only provide at best a probability statement regarding the likelihood of $\delta < \log HR(P/C)$. For example, if the true control effect, $\log HR(P/C)$, can be estimated from the historical data by the lower limit of the 95% confidence interval (CI) of the control effect, then the δ margin defined as a fixed number equal to half of this estimate of the control effect, i.e.,

$$\delta = 1/2 \{ [\log hr(P/C) - 1.96SE[\log hr(P/C)]] \} \quad (5)$$

provides a high probability that $\delta < \log HR(P/C)$. But here it is implicitly assumed that the historical control effect remains the same in the current trial setting. If this constancy assumption does not hold, then this probability may not be high either.

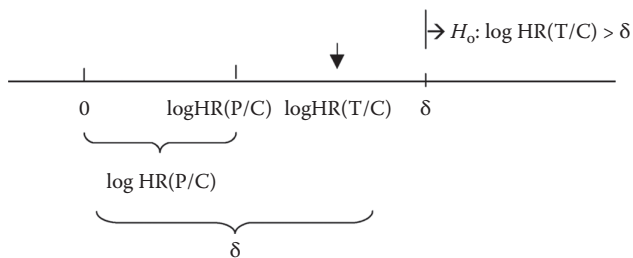


Fig. 1 Control effect, treatment effect, and margin δ

The test of the null hypothesis in Eq. 3 by $\tau_0(\delta)$, where δ is defined as in Eq. 5, has been called the two 95% CI-testing procedure. That is, the null hypothesis in Eq. 3 is rejected if

$$\log hr(T/C) + 1.96SE[\log hr(T/C)] < \frac{1}{2} \{ \log hr(P/C) - 1.96SE[\log hr(P/C)] \} \quad (6)$$

A somewhat similar two 90% CI-testing procedure has been used in thrombolytic trials (White^[13]).

Note that this δ in Eq. 5 is not truly a fixed number because it is defined in terms of an estimate of the control effect from the observed historical data. Thus, treating δ as a fixed number in the hypotheses in Eq. 3 is not exactly correct.

It is argued that although the two 95% CI-testing procedure offers a high probability that $\delta < \log HR(P/C)$, it may be extremely conservative because the true control effect, $\log HR(P/C)$, may be very much underestimated by the lower limit of the 95% CI of the control effect. But this argument is true only if the constancy assumption holds. An example given later will show that this may not be conservative enough when the constancy condition fails to hold.

On the other hand, using the point estimate, $hr(P/C)$, in the estimate of the true control effect may lead to a margin δ that is larger than the true control effect with relatively high probability. That is, if one defines

$$\delta = 1/2 \log hr(P/C) \quad (7)$$

This probability would be even higher when the constancy assumption does not hold. Testing the null hypothesis in Eq. 3 with the statistic $\tau_0(\delta)$, where δ is defined by Eq. 7, is equivalent to an asymmetric two CI-testing procedure where the control effect, $\log HR(P/C)$, is estimated by the point estimate, $\log hr(P/C)$, which can be thought of as the lower limit of the 0% CI of the control effect.

In the fixed margin method, the determination of the non-inferiority margin δ is based on the estimate of the historical control effect. Depending upon how this estimate is actually defined, the constancy assumption may or may not be assumed. Thus, for example, if the 95% confidence lower limit is used as the estimate of the control effect, then the constancy assumption is clearly not assumed and the historical control effect is actually discounted. If the point estimate is used, then the constancy assumption would be implicitly made and there would be no discounting of the historical control effect. The non-inferiority margin is then defined as half of this discounted historical control effect. Thus, implicitly through the way the control effect is estimated by the historical data, the fixed margin method assumes that the active control is effective in the current trial setting, and its effect size is estimated by a certain

degree of discount of the historical control effect. Thus, the choice of the non-inferiority margin δ in the fixed margin method as discussed above can be problematic, because it can be too liberal or too restrictive depending upon whether the historical control effect is assumed to hold true in the current trial setting or whether it is discounted by taking for example the 95% confidence lower limit of the estimate of the historical control effect.

Non-inferiority Hypotheses with Fraction of Control Effect to be Retained

From the above discussion, it is clear that for the margin δ to satisfy the condition that $\delta < \log \text{HR}(P/C)$ with high probability, one must define δ as a certain fraction, $0 < \phi_0 < 1$, of the control effect, $\log \text{HR}(P/C)$, which is unknown and needs to be estimated with data from historical placebo control trials. The idea of considering retention of a certain fraction of the control effect has been discussed in Hauck and Anderson,^[14] Holmgren,^[15] Koch and Tangen,^[16] and Rothmann et al.^[17] When the non-inferiority hypothesis is formulated in terms of a fraction of control effect to be retained, it provides a couple of advantages. First, with the specification of ϕ_0 , one does not need to pre-specify a δ margin. Secondly, the fraction of control effect to be retained can be specified according to the trial objective. For instance, if the current trial objective is to demonstrate that the new treatment is non-inferior to the active control, then the fraction specified should be closer to 1. If the current trial objective is to demonstrate that the new treatment is simply effective, then the fraction specified may be relatively smaller. Current practice often sets $\phi_0 = 0.5$.

The definition of the fraction, ϕ , of control effect retained by the new treatment is simply,

$$\phi = \frac{\log \text{HR}(P/C) - \log \text{HR}(T/C)}{\log \text{HR}(P/C)} \quad (8)$$

where the active control is assumed to be effective, i.e., $\log \text{HR}(P/C) > 0$, in the current trial. The non-inferiority hypotheses may be stated as follows.

$$H_0 : \phi \leq \phi_0 \quad \text{vs.} \quad H_a : \phi > \phi_0 \quad (9)$$

where ϕ_0 is the desired fraction of active control effect to be retained by the new treatment. Upon substituting the expression for ϕ in Eqs. 8 into 9 and after some algebra, the null and alternative hypotheses in Eq. 9 can be restated as,

$$\begin{aligned} H_0 : \log \text{HR}(T/C) &\geq (1 - \phi_0) \log \text{HR}(P/C) \quad \text{vs.} \\ H_a : \log \text{HR}(T/C) &< (1 - \phi_0) \log \text{HR}(P/C) \end{aligned} \quad (10)$$

The alternative hypothesis of interest in Eq. 10 can be interpreted as a desire to rule out a loss of more than

$(1 - \phi_0)$ of the control effect. The fraction ϕ_0 reflects, in a sense, the level of evidence one may demand of the new treatment. For example, when $\phi_0 = 1$, i.e., requiring 100% retention of the control effect, the hypotheses in Eq. 10 become the superiority hypotheses in Eq. 1. When $\phi_0 = 0$, i.e., requiring 0% retention of the control effect, the hypotheses in Eq. 10 become indirect comparisons to the reference “placebo.”

Comparing to Eq. 3, one may view the expression $(1 - \phi_0) \log \text{HR}(P/C)$ in Eq. 10 as the δ margin although, here, $(1 - \phi_0) \log \text{HR}(P/C)$ is not a pre-specified number. It involves the true control effect, $\log \text{HR}(P/C)$, which is unknown. In the two 95% CI-testing procedure, the expression $(1 - \phi_0) \log \text{HR}(P/C)$ in Eq. 10 is replaced by the estimate Eq. (Eq. 5), and the whole expression is considered as a fixed δ margin. In contrast to a superiority trial or a fixed margin method, in an active control non-inferiority trial, the hypotheses in Eq. 10 involve two parameters: the hazard ratios $\text{HR}(T/C)$ and $\text{HR}(P/C)$, as discussed in Chen et al.^[18]

Recognizing the fact that $\log \text{HR}(P/C)$ is an unknown parameter, the following statistic T is proposed for testing the null hypothesis in Eq. 10.

$$T = \frac{[\log \text{hr}(T/C)] - (1 - \phi_0)[\log \text{hr}(P/C)]}{\sqrt{\{SE^2[\log \text{hr}(T/C)] + SE^2[(1 - \phi_0) \log \text{hr}(P/C)]\}}} \quad (11)$$

T is claimed to be asymptotically normal and to have good convergence property^[6]. One then rejects H_0 at the one-sided 2.5% significance level if $T < -1.96$. The fraction retention approach of Rothmann et al.^[17] was applied in 2001 in the review and evaluation of the effectiveness of Xeloda in two active control studies FDA^[19] and also discussed in Rothmann et al.^[17] An application of this method can also be found in Wang et al.^[20] termed this approach the Synthesis Method, referring to the synthetic nature of the test statistic T in (Eq. 11), whereas Rothmann et al.^[17] called it the fraction retention method referring to the fraction retention hypothesis (Eq. 9), for which the linearized test statistic T was applied.

Since in an active control trial, there is no concurrent placebo, one cannot estimate the control effect, $\log \text{HR}(P/C)$, from the current trial. In Eq. 11, the estimate, $\log \text{hr}(P/C)$, needs to be estimated with data from historical placebo control trials. For this to be valid, one has to assume that the current control effect has not changed over time or, if it has changed, how much it has been reduced. For example, one may assume that the current control effect is a fraction, θ , of the historical control effect. Thus, in Eq. 11, one may include a fraction, θ , in front of the estimate of the control effect, $\log \text{hr}(P/C)$, which is based on the data from historical placebo control trials. However, in practice, information for such θ would be difficult to come by. The fraction retention method as proposed by Rothmann et al.^[17] assumes that the constancy assumption

holds. Thus, in the discussion below, we will assume that $\theta = 1$.

It should be noted that in Eq. 8, it is assumed that $\log HR(P/C) > 0$, that is, the Rothmann's fraction retention method assumes that the active control is effective. Furthermore, in the test statistic T in Eq. 11, the fraction of control effect lost, $(1 - \phi_0) \log HR(P/C)$, is unknown and is estimated by $(1 - \phi_0) \log hr(P/C)$ based on historical data. Thus, it also implicitly makes the constancy assumption without stating it. Moreover, the statistic T combines data from the current trial with historical data from other completed trials, and this raises question with regard to the validity of the statistical inference drawn based on the test of the hypotheses in Eq. 9 or Eq. 10 by T and causes difficulty in the interpretation of the result.

RELATIONSHIP BETWEEN THE FIXED δ -MARGIN HYPOTHESES AND FRACTION RETENTION HYPOTHESES

It should be noted that because the fraction retention approach uses the historical control data for the estimate of the control effect in its test statistic Eq. 11, whereas the fixed margin approach uses the historical control data in the specification of its non-inferiority margin, the two methods appear to be different at least in the manner in which the control effect is estimated. The former assumes the constancy assumption holds, whereas the latter discounts the historical control effect. However, one can cast these methods within the same fraction retention hypothesis framework given by Eq. 10 as described below. Within this same framework of fraction retention hypothesis Eq. 10, one can see that the liberalism or conservatism of these methods is easily seen from their corresponding test statistics used to test the fraction retention hypotheses give in Eq. 10. These test statistics differ simply in the standard errors used for their respective estimator, $\log hr(T/C) - (1 - \phi_0) \log hr(P/C)$. Note that all these methods are considering 50% retention of the control effect. Certainly, 50% retention is arbitrary and subjective, but one can use other threshold if appropriate.

The test of the fixed δ -margin null hypothesis in Eq. 3 using the statistics $\tau_0(\delta)$ with δ defined by Eq. 5 is equivalent to the two 95% CI-testing procedure (Eq. 6). It can easily be shown to be equivalent to using the following statistic t to test the fraction retention null hypothesis in Eq. 10:

$$t = \frac{\{\log hr(T/C) - (1 - \phi_0) \log hr(P/C)\}}{\sqrt{\{SE^2 \log hr(T/C)\} + \{SE^2 [(1 - \phi_0) \log hr(P/C)]\}}} \quad (12)$$

One rejects H_0 at the one-sided 2.5% significance level if $t < -1.96$. From the triangle inequality, one can easily deduce that

$$\begin{aligned} & \sqrt{\{SE^2 \log hr(T/C)\} + \{SE^2 [(1 - \phi_0) \log hr(P/C)]\}} \\ & \geq \sqrt{\{SE^2 \log hr(T/C) + SE^2 [(1 - \phi_0) \log hr(P/C)]\}} \end{aligned} \quad (13)$$

and hence, it follows that $T < t$, whenever the common numerator of both T and t , $\log hr(T/C) - (1 - \phi_0) \log hr(P/C)$, is negative. Therefore, the null hypothesis H_0 in Eq. 10 will be rejected by T whenever it is rejected by t at the same critical value, say, -1.96 . This implies that relative to testing the null hypothesis in Eq. 10, the statistic t is deflating the type I error.

The test of the fixed δ -margin null hypothesis in Eq. 3 using the statistic $\tau_0(\delta)$ with δ defined by Eq. 7 based on the point estimate, $\log hr(P/C)$, is a special case of an asymmetric two CI-testing procedure. It is easily seen to be equivalent to using the following statistic t^* to test the null hypothesis in Eq. 10:

$$t^* = \frac{\{\log hr(T/C) - (1 - \phi_0) \log hr(P/C)\}}{\sqrt{SE^2 \{\log hr(T/C)\}}} \quad (14)$$

Once again, note that $t^* < T < t$ whenever the common numerator of t^* , T and t , $\log hr(T/C) - (1 - \phi_0) \log hr(P/C)$ is negative. Thus, in testing the null hypothesis in Eq. 10 at the same critical value of -1.96 , the statistic t^* inflates the type I error. Thus, the statistic T protects against the conservatism of t and the liberalism of t^* relative to testing the fraction retention null hypothesis in Eq. 10. It is shown that for time-to-event endpoints, the statistic t^* deflates the type I error from 0.025 to as low as approximately 0.0028, while t^* can inflate the type I error to as much as 0.50.^[6]

On the other hand, it is of interest to point out that these test statistics can be shown to correspond conditionally to two asymmetric CI-testing procedures, where the control effect is estimated by the lower limit of a $100(1 - \gamma)\%$ CI, $0 < \gamma < 1$. The test statistic t^* corresponds to $\gamma = 1$, t corresponds to $\gamma = 0.05$, and T corresponds to some $\gamma = \gamma_T$, where $0.05 < \gamma_T < 1$ such that if ${}^c(1 - \gamma_T) < 1.96$ denote the critical value corresponding to γ_T , then

$$\begin{aligned} & SE[\log hr(T/C)] + ({}^c(1 - \gamma_T) / 1.96) SE[\log hr(P/C)] \\ & = \sqrt{SE^2 \left[\log hr \left(\frac{T}{C} \right) \right] + SE^2 \left[\log hr \left(\frac{P}{C} \right) \right]}. \end{aligned}$$

Although in general γ_T depends on the unknown standard error, $SE[\log hr(T/C)]$, but for time-to-event endpoints, such as survival, one may obtain an approximate estimate of γ_T at the design stage as shown in Ref. [6], because one can derive an approximate estimate of the unknown $SE[\log hr(T/C)]$ at the design stage and hence ${}^c(1 - \gamma_T)$, from which γ_T can be obtained. This will be illustrated in more detail in the later section on sample size determination. From this two CI-testing procedure perspective, one obtains an intuitive explanation as to why

the statistic T for testing the hypothesis in Eq. 3 protects against the conservatism of t and the liberalism of t^* , even though the fraction retention method does not quite fit into the two CI-testing procedure framework in general as indicated above except in the case for time-to-event endpoint.

This duality between the fraction retention hypothesis testing approach and the two CI-testing procedure approach exists because both approach uses the historical data to estimate the control effect, and this estimate is then used in either the test statistic or in the determination of the non-inferiority margin. It should be pointed out that all methods as discussed above actually use the same fraction retention level of 50%, and the control effect is estimated from historical data that came from studies that have been completed and known. If historical data is not available for the active control, then none of these methods would be applicable to an active control trial.

SOME REMARKS

Remark 1: As noted previously, for the same fraction retention level, the fixed margin method is considered to be more conservative than the fraction retention method when the constancy assumption holds. However, the constancy assumption often cannot be verified without a placebo in the current trial. Therefore, in the fixed margin approach, an ostensible reason to select a conservative estimate of the control effect size (e.g., using the lower limit of a 95% confidence interval) is to minimize the potential impact of a deviation from the constancy assumption that is assumed under the fraction retention method. Perhaps, what is not readily appreciated is the fact that in the fixed margin approach, simply using a conservative estimate of the control effect size may not totally avoid the impact of an actual deviation from the constancy assumption as illustrated in the following example.

Example: Consider a highly effective cancer treatment C for relapsed and refractory multiple myeloma patients. Based on two randomized historical trials, a pooled point estimate of the response rate for treatment A is approximately 40% with an associated 95% confidence interval of (36–45%). A non-inferiority trial is proposed to demonstrate that a new treatment T is non-inferior to the control treatment C. The non-inferiority margin was selected based on the lower limit of the 95% confidence interval of 36% in the response rate. Clinically, it is believed that the treatment effect in response rate can be compromised at most 40% in a non-inferiority trial (i.e., a 60% retention of the control effect). Based on the fixed margin method, the non-inferiority margin is determined to be $\delta = (1 - 0.60) \times 36\% = 14.4\%$. The non-inferiority hypothesis using this margin may be considered conservative according to current regulatory guidance. However, if the constancy assumption does not hold in this case, and the response rate

in the current non-inferiority trial is actually 25% say due to patient population differences and/or some other reasons, the non-inferiority margin of 14.4% would be quite large compared to the current control effect of 25%.

Remark 2: As noted earlier, the fraction retention threshold ϕ_0 determines the stringency of the non-inferiority margin, and in turn it will affect the sample size required for the non-inferiority trial. An important question regarding how stringent should the fraction retention threshold be for a given non-inferiority trial has never been resolved. It has been argued that the FDA drug regulations do not require that a new treatment has to be shown to be non-inferior to an active control in order to be approved. All that the regulations require is that the new treatment be superior to placebo. If that is the case, then what fraction retention threshold would be acceptable as a demonstration of superiority to a placebo that is not present? From an ethical perspective, for a trial with mortality or serious morbidity outcome, it would be unacceptable to treat patients with a potentially inferior new treatment, unless this new treatment is expected to provide some other substantive benefit that may offset such inferiority. However, the problem is that these potential benefits are yet to be demonstrated and quantified, and hence cannot be used in the actual determination of the non-inferiority margin, although they certainly may be considered in the final benefit/risk assessment. This raises the questions as to how liberal or strict should the fraction retention threshold be and how does one determine that?

Remark 3: The definition of the fraction, ϕ , of control effect to be retained by the new treatment can also be stated in terms of hazard ratios as follows.

$$\phi_h = \frac{HR(P/C) - HR(T/C)}{HR(P/C) - 1} \quad (15)$$

where the active control is assumed to be effective, i.e., $HR(P/C) - 1 > 0$, in the current trial. The non-inferiority hypotheses may be stated as follows.

$$H_0 : \phi_h \leq \phi_0 \quad \text{vs.} \quad H_a : \phi_h > \phi_0 \quad (16)$$

where ϕ_0 is the desired fraction of active control effect to be retained by the new treatment. Upon substituting the expression for ϕ_h in Eq. 15 into 16, and after some algebra, the null and alternative hypotheses in Eq. 16 can be restated as

$$\begin{aligned} H_0 : HR(T/C) &\geq 1 + (1 - \phi_0)[HR(P/C) - 1] \quad \text{vs.} \\ H_a : HR(T/C) &< 1 + (1 - \phi_0)[HR(P/C) - 1] \end{aligned} \quad (17)$$

When $0 < \phi < 1$, it can be shown easily that $\phi < \phi_h < 1$. Thus, if the non-inferiority null hypothesis in Eq. 10 is rejected, then it follows that the null hypothesis in Eq. 16 is rejected for the same value of ϕ_0 . Also, a test statistic

for those hypotheses in Eq. 16 analogous to Eq. 14 can be constructed. An application of such a test statistic can be found in Ref. [19].

To pose the non-inferiority hypothesis in Eq. 16 directly in terms of the hazard ratios is attractive in that it offers ease of interpretation. Wang et al.^[21] have proposed a ratio test statistic for testing the non-inferiority hypothesis in Eq. 16. However, unlike the authors' claim that it has greater power when compared to the linearized version, its power is comparable to the Rothmann's linearized test statistic for the hypothesis in Eq. 10. The reason is that the authors assumed equal variance under the null and alternative hypotheses in the derivation. However, when one allows unequal variance as is likely to be the case, the power is substantially reduced.

PLANNING FOR AN ACTIVE CONTROL TRIAL

The following are some important issues to be considered when contemplating an active control trial. See also the articles by Hung et al.^[8–10]

Validity of Active Control Trials

In order to have a valid statistical inference for an active control trial, several key assumptions have to be reasonably satisfied. The first assumption is that in the current trial setting, the active control is effective. Generally, this assumption is unverifiable. If this assumption is not true, then, for a superiority trial, one may only conclude that the new treatment is effective, but not that it is superior to the control. For a non-inferiority trial, this assumption may lead to the approval of a new treatment that is actually worse than a placebo. If possible, one should consider special design features and other concurrent trials that may provide evidence that the active control is effective. The second assumption is that the current trial has assay sensitivity. Assay sensitivity is defined as the ability of a trial to show an effective treatment to be effective. The ICH-E10 document on the Choice of Control^[3] introduces and discusses this concept in some detail. It essentially involves the quality of a trial including various aspects of trial design, conduct, and analysis. The third assumption is that the control effect in the current trial is either the same as or is some fraction of the historical control effect, and that there are placebo control historical trials on the basis of which the control effect and its variability can be reasonably assessed. In this regard, issues regarding quality, reliability, publication, and selection biases of these historical trials are of major concern and need to be dealt with. One should strive to make the current trial as similar as possible to the historical trials. If the active control effect has been diminished, then one needs to be able to specify the amount of reduction. The fourth assumption is that the trial is free of bias. An active control trial has potential

for bias toward the null without unblinding the trial. Strict standard operating procedures should be in place. As none of these assumptions can be truly verified, the validity of an active control trial is always subject to question.

The fraction retention method may have further validity issue as its test statistic is based on integrating data from the current trial with the data from some completed known historical control data. On the other hand, the fixed margin method tends to be conservative when the constancy assumption holds. However, when this assumption does not hold, then it is difficult to judge whether the specified margin would be liberal or conservative. For both methods, the actual fraction of the control effect to be retained has customarily been set at one-half. This retention level is arbitrary, and for a given situation, other levels may be appropriate and should be considered.

Therefore, in view of these issues, active control trials should be considered only when using a placebo would truly be unethical.

SOME DESIGN CONSIDERATIONS

As there is no placebo in the current active control trial, the control effect needs to be estimated from historical trial data. The questions one should address are as follows: What historical placebo control trials are available? Which collection of historical trials will be used to provide the estimate of the control effect? What kind of analysis will be used to provide this estimate? How reliable is the estimate of the historical control effect?

Although the control effect may be assessed by using data from historical randomized placebo-controlled studies, the current control effect may be different because of changes in patient population, standard care, medical practice, etc. One should design the current trial to be as similar as possible to the historical placebo control trials used in estimating the historical control effect. Important factors that may influence the trial outcome should be identified and may include region, gender, ethnicity, and patient characteristics. Standard of care differences may affect trial outcome. The current trial should also be as similar as possible to the historical trials with respect to these factors. Some adjustment of the historical control effect may be necessary by either replacing the control effect, $\log HR(P/C)$, in the hypotheses in Eq. 10 by $\theta \log HR(P/C)$, where $0 < \theta < 1$, or by adjusting using a mathematical model for population and study differences.

One should be clear about the primary objective of the active control trial. Should the objective be to demonstrate superiority of the new treatment to the control, non-inferiority to the control, or simply effectiveness of the new treatment? Each of these objectives requires a hypothesis with a different degree of tightness in the margin. For demonstration of non-inferiority, one needs to specify a fraction, ϕ_0 , of the control effect to be retained.

The specification of this fraction depends on the trial objective. If the active control is highly effective, then there may be no legitimate reason to consider a new treatment that is inferior to this control. In this case, one may require φ_0 to be closer to 1. However, if the new treatment offers some important benefits that are not available from the active control, such as a better toxicity profile, then φ_0 may be specified somewhat smaller. In the latter case, one may only conclude that the new treatment is effective, but not non-inferior to the control. The actual value of φ_0 may also depend not only on types of therapies, diseases, and endpoints but also on the distributional properties of the estimated control effect as well.

The conservatism of the two 95% CI-testing procedure does not take into account any of the concerns regarding the validity of the active control trial, nor the confidence one has regarding the estimate of the control effect. At the design stage of an active control trial, such concerns and lack of confidence may be reflected in the values specified for various design parameters. The design parameters include the fraction, φ_0 , of control effect to be retained and the fraction, θ , of the historical control effect to be reduced. The specification of these design parameters should also take into consideration how the control effect estimate is calculated and the resultant sample size requirement. Once φ_0 and θ have been specified, one can calculate the sample size required for testing the non-inferiority null hypothesis in Eq. 10.

SAMPLE SIZE DETERMINATION

The sample size required for testing the null hypothesis in Eq. 10 at a given significance level can be calculated as follows. The calculation depends on an appropriate assessment of the control effect. As mentioned before, using the statistic T to test the null hypothesis in Eq. 10 is conditionally equivalent to an asymmetric two CI-testing procedure, where the control effect is estimated by the lower limit of a $100(1 - \gamma_T)\%$ CI. The value of this γ_T is not known ahead of time. However, for time to event endpoint, this γ_T can be approximately calculated at design stage based on the pre-specified number of events at stopping time (the final analysis) and the significance level used in the test of the null hypothesis in Eq. 10.

In practice, the control effect sometimes cannot be assessed reliably either because there is insufficient historical data or the control effect may have changed over time. In such cases, the sample size calculation should be based on a conservatively estimated control effect, e.g., using the lower limit of the 95% or even 99% CI.

In oncology, we require, as stopping rules for time to event endpoints, pre-specified fixed number of events. A lower bound and approximation for the asymptotic standard error of $\log hr(T/C)$ for a 1:1 randomization is given by $2/\sqrt{n}$, where n denotes the pre-specified total number

Table 1 Number of Events Needed for 80% Power

HR(T/C)	Number of events	Cutoff
1	4801	1.0842
0.95	1505	1.0976
0.90	750	1.1044
0.85	446	1.1085
0.80	291	1.1114

of events at stopping. The use of this lower bound for the standard error at the design stage allows us to algebraically “equate” the use of a test statistic with a pre-specified two CI procedure.

Consider the case where $\alpha = 0.025$, $\beta = 0.2$, $\varphi_0 = 0.5$, $\theta = 1$, $\log hr(P/C) = 0.234$, and $SE[\log hr(P/C)] = 0.075$. Table 1 gives the number of events needed for various “alternatives” for $HR(T/C)$ and the corresponding non-inferiority cutoffs for the upper limit of the 95% CI for $\log HR(T/C)$. The cutoff is defined as $\delta = (1 - \varphi_0)\{\log hr(P/C) - c_{(1-\gamma)}SE[\log hr(P/C)]\}$, where $c_{(1-\gamma)}$ is the solution to the equation,

$$\begin{aligned} & 2/\sqrt{n} + c_{(1-\gamma)}(1 - \varphi_0)SE[\log hr(P/C)]/1.96 \\ & = \sqrt{\{4/n + (1 - \varphi_0)^2 SE^2[\log hr(P/C)]\}}, \end{aligned}$$

and the number of events needed at stopping is given by n where n solves

$$\begin{aligned} \frac{2\Phi^{-1}(1 - \beta)}{\sqrt{n}} &= (1 - \varphi_0)\log hr(P/C) - \log hr(T/C) \\ &+ \Phi^{-1}(\alpha)\sqrt{\{4/n + (1 - \varphi_0)^2 SE^2[\log hr(P/C)]\}}. \end{aligned}$$

CONCLUSION

In the design of a non-inferiority trial, if one believes that the historical active control effect is likely to be maintained in the current trial setting, then the hypothesis with a retention of certain fraction of the control effect could be a reasonable formulation, although as noted earlier a basic problem with this approach is inherent in the test statistic used to test the non-inferiority hypothesis that is partly based on data from past completed studies (if they exist). If there is reason to believe that the historical control effect has been reduced over time, then the fixed margin approach would be more appropriate, although it remains a challenge to provide a reasonable estimate of the current control effect. If the control effect in the current setting is greater than the estimate, then the margin would be conservative. On the other hand, if the control effect in the current setting is smaller than the estimate, then the margin would be liberal as illustrated by the earlier example. In either the fraction retention or the fixed margin

approach, there is a basic lack of a reference for assessing whether a given margin is too liberal or too restrictive. Further research in this regard would be needed in order to develop a more acceptable approach to the design of a non-inferiority study.

Active control non-inferiority trials should be considered on a case-by-case basis. We need to have assurance that the active control effect exists in the current study patient population. It is recommended that the best available treatment be used as the control. We need to assess whether the current effect size, if it exists, has diminished. We should have historical randomized placebo-controlled studies that can provide reliable estimate of the historical control effect. The current trial population should be as similar as possible to the populations studied in the historical control studies. A “non-inferiority” trial may be prone to bias toward “no difference.” Careful attention should be paid to the conduct and analysis of such a study. It should be well monitored and conducted. After the study is done and all efficacy and safety data collected, the evidence as to whether the new treatment is effective should be carefully examined. One should then assess this evidence by incorporating any other additional substantive benefits that the new treatment can provide to the patients that the control cannot. In view of the subjective nature of the margin determination, it is recommended that in the final efficacy assessment, the pre-specified non-inferiority margin should serve more as a guide than as an absolute cutpoint. The final conclusion of course should still follow the usual benefit/risk assessment.

DISCLAIMER

The views expressed in this article are those of the authors and not necessarily those of the U.S. Food and Drug Administration nor those of the individual companies affiliated with the co-authors.

REFERENCES

1. Leber, P. The placebo control in clinical trials—a view from the FDA. *Psychopharmacol. Bull.* **1986**, 22 (1), 30–32.
2. Leber, P. Hazards of inference: The active control investigation. *Epilepsia* **1989**, 30 (S1), s57–s63.
3. U.S. Food and Drug Administration. *International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), Guidance to Industry, E-10: Choice of Control Group and Related Issues in Clinical Trials* 1999—Federal Register 64, 185, September 24, 1999/Notices. 51767–51780.
4. World Medical Association. Declaration of Helsinki Ethical principles for medical research involving human subjects. *J. Am. Med. Assoc.* **2000**, 284 (23), 3043–3045.
5. Temple, R. Problems in interpreting active control equivalence trials. *Account. Res.* **1996**, 4, 267–275.
6. Guideline on the Choice of the Non-inferiority Margin. Committee for Medical Products for Human Use, European Medicines Agency, July 27, 2005.
7. Draft Guidance for Industry on Non-inferiority Clinical Trials, U.S. Food and Drug Administration, March 2010.
8. Hung, H.M.J.; Wang, S.-J.; Tsong, Y.; Lawrence, J.; O'Neill, R.T. Some fundamental issues with non-inferiority testing in active controlled trials. *Stat. Med.* **2003**, 22, 213–225.
9. Hung, H.M.J.; Wang, S.-J.; O'Neill, R.A. Issues with statistical risks for testing methods in NI trials without a placebo arm. *J. Biopharm. Stat.* **2007**, 17, 201–213.
10. Hung, H.M.J.; Wang, S.-J.; O'Neill, R.A. A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. *Biom. J.* **2009**, 51, 324–334.
11. Ng, T.H. Non-inferiority hypotheses and choice of non-inferiority margin. *Stat. Med.* **2008**, 27, 5392–5406.
12. Blackwelder, W. Proving the null hypothesis in clinical trials. *Control. Clin. Trials* **1982**, 3, 345–353.
13. White, H.D. Thrombolytic therapy and equivalence trials. *J. Am. Coll. Cardiol.* **1998**, 31, 494–496.
14. Hauck, W.W.; Anderson, S. Some issues in the design and analysis of equivalence trials. *Drug Inf. J.* **1999**, 33, 109–118.
15. Holmgren, E.B. Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained. *J. Biopharm. Stat.* **1999**, 9 (4), 651–659.
16. Koch, G.G.; Tangen, C.M. Nonparametric analysis of covariance and its role in non-inferiority clinical trials. *Drug Inf. J.* **1999**, 33, 1145–1159.
17. Rothmann, M.; Li, G.; Chen, G.; Chi, G.Y.H.; Temple, R.; Tsou, H.H. Design and analysis of non-inferiority mortality trials in oncology. *Stat. Med.* **2003**, 22, 239–264.
18. Chen, G.; Wang, Y.C.; Chi, G.Y.H. Hypotheses and type I error in active-control noninferiority trials. *J. Biopharm. Stat.* **2004**, 14, 301–313.
19. FDA. *Medical-Statistical Review for Xeloda, NDA 20-896*; FDA Division of Freedom of Information: Rockville, MD, 2001.
20. Wang, S.J.; Hung, H.M.J.; Tsong, Y. Utility and pitfalls of some statistical methods in active controlled trials. *Control. Clin. Trials* **2002**, 23, 15–28.
21. Wang, Y.C.; Chen, G.; Chi, G.Y.H. A ratio test in active control noninferiority trials with a time-to-event endpoint. *J. Biopharm. Stat.* **2006**, 16, 151–164.

Active Controlled Clinical Trials: Noninferiority Analysis

Sue-Jane Wang

Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

H. M. James Hung

Division of Biometrics I, Office of Biostatistics, U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.

Yi Tsong

Division of Biometrics VI, Office of Biostatistics, U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.

Abstract

This entry discusses the simulation in early phases of clinical trials, specifically through oncology dose-escalation trials to discuss some theoretical and practical issues in trial simulations.

Keywords: Active control studies; Noninferiority; Preservation.

INTRODUCTION

Noninferiority analysis for comparing a new experimental therapy with a selected active (or positive) control therapy is a difficult and often very controversial subject,^[1–17] particularly when the active controlled clinical trial does not include a placebo arm. In approximately late 1970s to early 1980s, the concept of noninferiority was generally defined using a one-sided hypothesis with a prespecified noninferiority margin (e.g., Ref. [18]). The aim of the noninferiority analysis is to demonstrate that the experimental therapy is not inferior to the active control treatment within the margin with respect to the efficacy endpoint(s) of main interest. Hypothesis testing or confidence interval approach was frequently employed. In late 1980s to mid-1990s, noninferiority was sometimes considered as two-sided therapeutic equivalence, mostly seen in cancer treatment trials, with a prespecified equivalence range. Indeed, for some time, it was not clear as to whether the real interest in a noninferiority clinical trial is that the new treatment is not much different in effect from the active control or that the new treatment is not much less effective than the active control. The terms equivalence and noninferiority were often used interchangeably for a little while in the mid-1990s. Then, some literature in 1997–2001—the points to consider of the Committee for Proprietary Medicinal Products (CPMP) document¹¹ from the European regulatory agency, and the discussions in the Food and Drug Administration (FDA)—made the distinction between equivalence

and noninferiority. Intuitively, noninferiority can be viewed as one-sided equivalence. Historically, growing awareness and regulatory attention to noninferiority and equivalence analysis appeared to begin around the 1980s.

Difficulty in defining a noninferiority margin has probably been overlooked for a long time. The noninferiority margin is tied to the clinically important difference that one hopes to rule out between the experimental and the control treatments. Hence the margin determination cannot be based solely on some kind of statistical criteria. It is generally a collective effort that consider many factors, such as, the effect of the active control, dosages of the active control agent and/or concomitant medication, background medications, etc. International Conference on Harmonization (ICH)-E9⁸ states that the margin is the largest difference that can be judged as being clinically acceptable and should be smaller than the differences observed in superiority trials of the active comparator. ICH-E10⁹ states that the margin chosen for a noninferiority trial cannot be greater than the smallest effect size that the active drug would be reliably expected to have compared with placebo in the setting of the planned trial. These are general principles. None of these documents discuss how to define a margin or how to analyze a noninferiority study.

OBJECTIVES OF NONINFERIORITY TRIAL

Two main objectives regarding the effectiveness of an experimental therapy are often considered in a noninferiority trial. One objective is to establish evidence for the efficacy of the experimental treatment, a regulatory requirement for

The views expressed in this article do not necessarily represent those of the U.S. Food and Drug Administration.

drug approval. The direct comparison of the experimental treatment with the active control together with historical evidence on the active control is often needed to lead to the assertion that the experimental treatment would have been more effective than placebo had the placebo been included in the noninferiority trial. Pursuance of this objective is the focus of many articles, e.g., Refs. [19,20]. The other objective is to assert that the experimental treatment is not much less effective than the control treatment by a noninferiority margin. This renders a direct assessment of the performance of the experimental treatment relative to the active control treatment, which is guarded by the noninferiority margin, e.g., Ref. [10]. The two objectives are interrelated. The former objective is usually a prerequisite consideration when pursuing the latter objective. Undoubtedly, the latter objective is a higher bar to achieve.

In many practices, when setting the noninferiority margin, one criterion often taken into consideration is that the experimental treatment retains at least some portion of the control effect. In effect, demonstration of the percent effect preservation may also be a goal for the noninferiority analysis. A well-known case example is retention of at least 50% effect for approving a new thrombolytic agent in noninferiority trials, which is considered by the Center of Biologics Evaluation and Research (CBER), FDA, as described in Ref. [10]. Other percentages of retention are certainly possible, depending on many factors such as diseases, drug safety profile, cost consideration, the confidence about the estimated control effect, etc. In fact, preservation of some fraction of the control effect is an important consideration to ensure that the efficacy of the experimental therapy relative to the unavailable concurrent placebo can be established with great confidence. In some cases, it may be sufficient to quantify how much risk that it would be the worst case is acceptable or to ensure that the amount of therapeutic effect lost by the experimental treatment is maximally allowable. However, the concept of “not inferior” cannot always be defined quantitatively only by a retention of some fraction of the control effect, see Ref. [21].

STATISTICAL METHODS OF NONINFERIORITY ANALYSIS

Literature provides two major classes of statistical methods for entertaining the objectives in noninferiority analysis. Use of confidence interval has been a common practice for decades, as found in such works as Fleming,^[2] CBER/FDA,^[10] and Blackwelder.^[18] First, one calculates a 95% confidence interval for the effect of the experimental treatment relative to the active control comparator from the active-controlled trial. If the entire confidence interval falls in the noninferiority range referenced by the noninferiority margin, then an argument for noninferiority defined on the basis of the prescribed margin can

be made. By asserting that the experimental treatment is not substantially less effective than the control, one can then indirectly infer that the experimental treatment is efficacious (relative to the putative placebo) or preserves the targeted percent of the control effect. As an example, in the case study of Fleming,^[2] mitoxantrone (MITX, an experimental treatment) was compared to doxorubicin (ADR, a widely considered standard of care) in treatment of advanced breast cancer on survival. Fleming pointed out that using the conventional 5% level of significance in the no treatment placebo controlled study and a reliable estimate of the active control effect on survival, in principle, survival with MITX relative to ADR can be represented by positioning ADR at 100% and no treatment at 80%. Survival on MITX was estimated to be 81% as long as that on ADR with 95% confidence interval of 58%–112% from an active control study. Because the lower bound of this confidence interval does not exceed 80%, the active control study failed to provide sufficient statistical evidence to assert that MITX gives greater survival than no treatment. The confidence interval analysis can be easily transformed to hypothesis testing with the same defined margin.^[18]

In the confidence interval approach, the noninferiority margin is the criterion for the determination of a noninferior experimental therapy as compared with a control therapy. Thus, the confidence interval method necessitates prespecification of the noninferiority margin. It is usually a very challenging task to quantitatively set the noninferiority margin for defining the concept of “not inferior” and at the same time indicating “not much less effective.” Consideration must include the efficacy of the experimental treatment, side effects, cost, convenience of administration, the effect of the active control including its size, the variability, and its consistency among the historical trials. For the traditional confidence interval method, once this margin is determined, it is treated as a fixed number without variability in the statistical inference for the active control trial.

The other class of statistical methods for noninferiority analysis may be collectively called synthesis method, such as those found in Refs. [19–26]. This type of methods implement a formal framework for cross trial statistical inference. Under such framework, the method synthesizes the estimated relative effect of the experimental treatment vs. the control from the active controlled trial and the estimated effect of the control treatment relative to placebo from historical evidence. The synthetic estimate results in a test statistic for formally testing the efficacy of the experimental treatment relative to a putative placebo and is used to estimate the fraction of the control effect that may have been preserved by the experimental treatment had the placebo been included in the active controlled trial. In the statistical hypothesis of percent effect preservation, the noninferiority margin is an explicit function of the control effect. If the control effect can be precisely

ascertained, a null percent of effect retention by the experimental treatment implies that it is more effective than the putative placebo and 100% preservation implies that the experimental treatment is more effective than the active control. In contrast to the confidence interval method, the synthesis method does not require the full pre-specification of the noninferiority margin in the hypothesis but needs only to set the level of percent for preservation. The statistical uncertainty in the effect of the active control is incorporated as a part of statistical inference with the synthesis method.

Simon^[27] introduced a Bayesian method that perhaps incorporates the concept of synthesization more naturally. Under this framework, the value of the expected placebo event rate is allowed to vary from past to present, from population to population, or from trial to trial. So are the effect of the active control and the effect of the new treatment relative to placebo. All these parameters are assumed to follow some statistical distributions identified by a number of hyperparameters that quantify central tendency and spread or scale. Using the data from historical trials and the active controlled trial and the given prior distributions, one can compute the posterior probability that the experimental treatment is effective relative to placebo or the probability that the experimental treatment retains at least some percent of the control effect. Statistical inference is then drawn on the basis of the posterior probability. It is worth noting that with noninformative priors on the placebo effect and the effect of the experimental treatment relative to the placebo, Simon's method renders essentially the same conclusion as the frequentist synthesis method gives.

Recently, Wang and Blume^[28] have considered a likelihood-based approach to define the noninferiority margin based on the support intervals from historical trials and to assess the strength of evidence for noninferiority under a pre-specified fixed margin.

CHALLENGES

Despite the fact that there has been considerable experience in noninferiority analysis, many difficulties and controversies remain. One of the difficult issues concerns choice of the active control. Usually, the selected active control for noninferiority trial must be effective that has been demonstrated in historical evidence. This is a part of the "assay sensitivity" issue discussed in great length in the ICH E-10 document.^[9] After an active control is selected, how to estimate the effect of the active control and which historical trials should be selected for estimation is also difficult. For the noninferiority analysis, improper handling of noncompliant patients might induce a bias that makes it easier to show no difference between the treatments. The analysis of the intent-to-treat population preserving randomization might not be conservative for noninferiority analysis

as generally thought in showing treatment differences in a superiority analysis. There are abundant resources in literature to enlighten many of such difficult issues.^[11-17]

Several other fundamental issues, however, have been rarely addressed in practical applications of noninferiority analysis, as stipulated in Ref. [29,30]. The main study objective is often not entirely clear in designing a noninferiority clinical trial. If the objective is showing that the experimental treatment is not considerably less effective than the active control, then the margin needs to be defined a priori and cross trial inference is not always necessary. The margin selection cannot be based purely on some statistical criteria and clinical consideration plays a vital role. Many approaches have been articulated, e.g., Refs. [31]. On the other hand, for showing that the experimental treatment is effective relative to a placebo or preserves some fraction of the control effect, cross trial inference cannot be avoided (see, e.g., Refs. [19-26,29,30-33]).

The value of the control effect is almost always unknown in both the active control trial population and the relevant historical trial populations. In the active control trial, the effect of the active control is not estimable. The historical trials provide an estimate for the control effect. This estimate may be applicable to the active control trial when the effect of control in the historical trial populations carries to the active control trial population, the so-called constancy condition. This condition implies that the active control trial and historical trial patient populations are similar, that trial design, background medications, etc. are similar, and that the control effect does not change over time or because of any other factors. Noninferiority analysis that requires cross trial statistical inference is extremely difficult because of the following reasons: (i) the constancy condition cannot be verified because of unavailable concurrent placebo in the noninferiority trial; (ii) trial-to-trial heterogeneity is mostly present in practice. In fact, when the constancy condition is in doubt, no straightforward statistical or clinical means can be used to properly adjust the bias and/or the variability, and no statistical method can render a conclusion that is defensible under the conventionally required criteria for statistical evidence in clinical trials. Placebo creep is often of concern if the control effect changes over time (see, e.g., Refs. [23,34]). It is therefore critical to assess the potential violation of the constancy assumption (Refs. [34]).

Even if the constancy condition holds, the fact that the effect of the active control is at best estimable from historical data presents a great challenge to the conventional framework of statistical inference. To conclude via noninferiority testing that the experimental treatment would have been effective compared to a placebo had the placebo been studied, cross trial inference is necessary. What is the probability of type I error pertaining to falsely making such a conclusion in the cross trial statistical inference? If the type I error probability cannot be defined, the *p*-value that is routinely used for quantification of statistical evidence

in placebo controlled clinical trials becomes meaningless. This issue also arises in testing for percent effect retention. Even with classical confidence interval approach, when the noninferiority margin involves the estimated control effect, the statistical uncertainty of the estimated margin in principle arguably needs to be incorporated in evaluation of the type I error of making a false assertion of noninferiority. This cross trial type I error probability has recently been considered by many researchers, e.g., Refs. [19–26,29,30,34]. Nevertheless, the relevance of this error probability would continue to be debated for the days to come.

If the constancy assumption is very much in doubt, the type I error probability, however it is defined, may be irrelevant. There is no way to know the worst difference in the control effect between the active controlled trial population and the historical trial populations without truly knowing the population similarities and whether the constancy assumption holds. The cross trial type I error probability may not be controllable in both confidence interval method and synthesis method. In this case, the noninferiority analysis must be avoided unless an extremely conservative margin can be defined so that the result of the analysis may be interpretable with a very small chance of drawing a false conclusion. However, as argued in Refs. [29,30], such a risk is beyond the traditional alpha error and is not statistically quantifiable.

CONCLUSION

Difficult and complex as it is with either limited or rich relevant historical information, a noninferiority active controlled trial design is not always informative for studying a new treatment. Interpretability of the results of a noninferiority trial depends on its study objective(s), how robust is the control effect estimated from the historical placebo controlled trials, whether this control effect is applicable to the targeted patient population, and many other factors. Cross-trial inference is almost unavoidable whether it is implicitly or explicitly made.

REFERENCES

1. Temple, R. Difficulties in Evaluating Positive Control Trials. *Proceedings of the Biopharmaceutical Section*; American Statistical Association: Alexandria, VA, 1983; 1–7.
2. Fleming, T.R. Treatment evaluation in active control studies. *Cancer Treat. Rep.* **1987**, *71*, 1061–1064.
3. Pledger, G.; Hall, D.B. Active Control Equivalence Studies: Do They Address the Efficacy Issue? *Statistical Issues in Drug Research and Development*; Marcel Dekker: New York, 1990; 226–238.
4. Jones, B.; Jarvis, P.; Lewis, J.A.; Ebbutt, A.F. Trials to assess equivalence: The importance of rigorous methods. *Br. Med. J.* **1996**, *313*, 36–39.
5. Temple, R. Problems in interpreting active control equivalence trials. *Account. Res.* **1996**, *4*, 267–275.
6. Ebbutt, A.F.; Frith, L. Practical issues in equivalence trials. *Stat. Med.* **1998**, *17*, 1691–1701.
7. Rohmel, J. Therapeutic equivalence investigations: Statistical considerations. *Stat. Med.* **1998**, *17*, 1703–1714.
8. *International Conference on Harmonisation: Statistical Principles for Clinical Trials (ICH E-9)*; Food and Drug Administration, DHHS, 1998.
9. *In Food and Drug Administration*, DHHS, International Conference on Harmonisation: Guidance on Choice of Control Group and Related Design and Conduct Issues in Clinical Trials (ICH E-10), 2000.
10. CBER/FDA Memorandum. Summary of CBER considerations on selected aspects of active controlled trial design and analysis for the evaluation of thrombolytics in acute MI, June 1999.
11. Committee for Proprietary Medicinal Products (CPMP) *Points to Consider on Switching Between Superiority and Non-inferiority*; 1999, Available at: <http://www.emea.eu.int/pdfs/human/ewp/048299en.pdf>
12. Siegel, J.P. Equivalence and non-inferiority trials. *Am. Heart J.* **2000**, *139*, S166–S170.
13. Fleming, T.R. Design and interpretation of equivalence trials. *Am. Heart J.* **2000**, *139*, S171–S176.
14. Temple, R.; Ellenberg, S.S. Placebo-controlled trials and active-control trials in the evaluation of new treatments—Part 1: Ethical and scientific issues. *Ann. Intern. Med.* **2000**, *133*, 455–463.
15. Ellenberg, S.S.; Temple, R. Placebo-controlled trials and active-control trials in the evaluation of new treatments—Part 2: Practical issues and specific cases. *Ann. Intern. Med.* **2000**, *133*, 464–470.
16. Wang, S.J.; Hung, H.M.J.; Tsong, Y.; Cui, L. Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. *Stat. Med.* **2001**, *20*, 1903–1912.
17. Chow, S.C.; Shao, J. *Statistics in Drug Research—Methodology and Recent Development*; Marcel Dekker, Inc.: New York, NY, 2002.
18. Blackwelder, W.C. Proving the null hypothesis in clinical trials. *Control. Clin. Trials* **1982**, *3*, 345–353.
19. Fisher, L.D.; Gent, M.; Bjeller, H.R. Active-control trials: How would a new agent compare with placebo? A method illustrated with clopidogrel, aspirin, and placebo. *Am. Heart J.* **2001**, *141*, 26–32.
20. Hasselblad, V.; Kong, D.F. Statistical methods for comparison to placebo in active-control trials. *Drug Inf. J.* **2001**, *35*, 435–449.
21. Hung, H.M.J.; Wang, S.J.; O'Neill, R. A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. *Biom. J.* **2005**, *47*, 28–36.
22. Holmgren, E.B. Establishing equivalence by showing that a prespecified percentage of the effect of the active control over placebo is maintained. *J. Biopharm. Stat.* **1999**, *9* (4), 651–659.
23. Wang, S.J.; Hung, H.M.J.; Tsong, Y. Utility and pitfall of some statistical methods in active controlled clinical trials. *Control. Clin. Trials* **2002**, *23* (1), 15–28.
24. Snappinn, S.M. Alternative for discounting in the analysis of non-inferiority trials. *J. Biopharm. Stat.* **2004**, *14*, 263–273.

25. Rothman, M.; Li, N.; Chen, G.; Chi, G.Y.H.; Temple, R. Design and analysis of non-inferiority mortality trials in oncology. *Stat. Med.* **2003**, *22*, 239–264.
26. Snapinn, S.; Jiang, Q. Controlling the type I error rate in non-inferiority trials. *Stat. Med.* **2008**, *27*, 371–381.
27. Simon, R. Bayesian design and analysis of active control clinical trials. *Biometrics* **1999**, *55*, 484–487.
28. Wang, S.J.; Blume, J. An evidential approach to non-inferiority clinical trials. *2008 Proceedings of the American Statistical Association, Biopharmaceutical Section [CD-ROM]*; Alexandria, VA: American Statistical Association.
29. Hung, H.M.J.; Wang, S.J.; Tsong, Y.; Lawrence, J.; O'Neill, R.T. Some fundamental issues with non-inferiority testing in active controlled trials. *Stat. Med.* **2003**, *22* (2), 215–225.
30. Hung, H.M.J.; Wang, S.J.; O'Neil, R.T. Issues with statistical risks for testing methods in non-inferiority trial without a placebo arm. *J. Biopharm. Stat.* **2007**, *17*, 201–213.
31. Temple, R. Active Control Studies—Regulatory Issues: The Past and the Present. Presented in the Annual Meeting of the Drug Information Association **2002**.
32. Wang, S.J.; Hung, H.M.J. TACT method for non-inferiority testing in active controlled trials. *Stat. Med.* **2003**, *22* (2), 227–238.
33. Julious, S.A.; Wang, S.J. How biased are indirect comparisons, particularly when comparisons are made over time in controlled trials? *Drug Inf. J.* **2009**, *42*, 625–633.
34. Hung, H.M.J.; Wang, S.J.; O'Neill, R.T. Challenges and regulatory experiences with non-inferiority trial design without placebo arm. *Biom. J.* **2009**, *51* (2), 324–334.

Adaptive Design Clinical Trials: Case Studies

Shiowjen Lee

*Center for Biologics Evaluation and Research, U.S. Food and Drug Administration,
Silver Spring, Maryland, U.S.A.*

Min Lin

*Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring,
Maryland, U.S.A.*

Abstract

For the past few decades, there has been considerable interest among pharmaceutical and other medical product developers in adaptive design clinical trials, in which knowledge learned from data of an ongoing trial affects design features or analysis of the study. Following the initial release of the FDA Draft Guidance “Adaptive Design Clinical Trials for Drugs and Biologics” in early 2010, there have been high expectations of an increase in regulatory submissions involving adaptive design features, particularly for trials that are intended to provide substantial evidence for product registration. More recently, the finalized FDA Guidance on Adaptive Designs for Medical Device Clinical Studies lays out the principles of conducting adaptive design clinical trials. The passed 21st Century Cures Act encourages a broader use of adaptive design in clinical studies. Despite all these, there remain some concerns and questions regarding statistical analyses and operational challenges in conducting adaptive design clinical trials. In this article, we provide some case studies reviewed at the FDA and general considerations for developing proposals for adaptive design clinical trials.

Keywords: Adaptive design; Adaptive seamless design; Critical path initiative; Food and Drug Administration; Operational bias; Study integrity; Type I error rate.

INTRODUCTION

To support regulatory standards for evidence in a medical product application, statistical inferences in larger scale trials are generally performed. One of the critical issues is whether results generated from the trials are interpretable with respect to the safety and effectiveness of the proposed medical product in the seeking indication. It is recognized that medical product development has become increasingly challenging, inefficient, and costly. The United States Food and Drug Administration’s (FDA) critical path initiative document^[1,2] had called for the need for more scientific efforts and efficient evaluation tools used in product development. Adaptive trial design was one of the topics focused on in the medical utility area. Additionally, the recently passed 21st Century Cures Act encourages a broader use of adaptive design in clinical studies, so a beneficial treatment can be made available sooner for patients in need.

Adaptive designs have been commonly used in clinical trials to advance medical research for many years.

An adaptive clinical trial design is one where data collected during the trial are monitored and analyzed in order to make modifications to the trial as it continues. For a clinical trial with an adaptive design, interim analyses are needed. Because of the usage of interim data for study modification, such designs have been thought by many to speed up the development process of investigational products and maximize the chance of trial success. Therefore, adaptive designs have been attractive to industry researchers who are looking to reduce costs and streamline the product development process. Additionally, there are potential ethical advantages of adaptive designs with respect to both efficacy and safety. For examples, adaptive design may help to protect patients from harm if interim results show a clear detrimental tendency of the investigational therapy, or adaptive design may lead to earlier availability of an effective treatment with exposure of fewer patients in the testing process.

Despite enthusiasm, the use of adaptive design in product development may or may not be appropriate depending upon a number of factors. Factors likely to determine a choice of an adaptive approach could include what is known at the planning stage, for example, objectives and target populations, accrual rate, choice of endpoints, characteristics of the investigational products, and others.

Note: The views expressed are those of the authors and do not represent the official position of the US Food and Drug Administration.

The potential advantages and disadvantages of an adaptive design over a traditional design at the study planning stage, especially for trials that are intended to provide substantial evidence for product registration, should be carefully assessed. Furthermore, all considerations should be evaluated whether simple and conventional adaptive designs, such as group sequential designs, will suffice to establish safety and efficacy of a product. On the contrary, complex adaptive designs may not be cost- or time-efficient as detailed design justification and possibly extensive simulation studies would usually be needed. This process can significantly delay the initiation of a trial. In general, an adaptive design with a way-too-early interim look may not be useful because there is too little information and it is subject to large variation to properly adapt the trial. Studies that stop too early for success may have inadequate safety data accrued, and both safety and efficacy are considerations for a premarket approval application. In addition, having sufficient sample size to assess secondary endpoints may need to be taken into consideration when planning an early study termination for efficacy, particularly if the secondary endpoint has less frequency (e.g., mortality).

Many adaptive design features have been proposed in the literature and/or have applied to clinical trials, including adaptive dose escalation, adaptive randomization, adaptive dose finding, and sample size re-estimation (SSR). However, modifications of an ongoing trial that is intended to provide substantial evidence for product registration, on the one hand, have been viewed by some as a contradiction to the confirmatory nature in the product development process. On the other hand, such modifications may be necessary for the well-being of patients. As a result, for many adaptive design clinical trials, there are outstanding statistical and operational issues that can affect the interpretability and validity of trial results. FDA Draft Guidance for drugs and biologics^[3]—“Adaptive Design Clinical Trials for Drugs and Biologics”—and final Guidance for medical devices^[4]—“Adaptive Designs for Medical Device Clinical Studies”—state explicitly that all planned adaptations should be prespecified at the design stage, and post hoc modification of study features is not considered to be an adaptive design. While both Guidance documents encouraged exploring the use of adaptive designs in early stage studies, their emphasis was on those studies that are intended to provide primary evidence for product registration. There are three major concerns associated with adaptive design clinical trials emphasized in the Guidance: (1) control of the study-wise type I error rate, (2) minimization of the impact of any adaptation-associated statistical or operational bias on the estimates of treatment effects, and (3) the interpretability of study results. These concerns are expected to be addressed during the design and study protocol stage when an adaptive design clinical trial is planned for a regulatory submission.

For the past few decades, adaptive design clinical trials have been proposed and conducted in regulatory

submissions. The proposed adaptive design features differ from one therapeutic area to another. FDA’s experiences in adaptive design have been in the review of proposals at the investigational new drug and investigational device exemption stages, as well as premarket applications. Some of the products have been approved based on results generated from adaptive design clinical trials. The following gives a brief summary of possible types of adaptive designs:

- **SSR:**
The sample size planned originally may not be sufficient. This could attribute to incorrect assumptions or variation greater than expected in the current ongoing trial.
- **Change of study objective:**
The superiority objective originally planned may be too ambitious after interim results are available, or vice versa, that the benefit of the investigational product based on interim results turns out to be greater than expected.
- **Change of the primary endpoint:**
The primary endpoint was chosen inappropriately at the planning stage. For example, the choice of components in a composite endpoint may be suboptimal.
- **Termination/modification of treatment arms:**
Interim results may suggest that some treatment arms are not efficacious or harmful and therefore should be dropped from the study, or treatment arms or dosing schedule is not optimal and should be modified.
- **Change of statistical methods:**
Interim results may suggest that more optimal statistical methods could be chosen for the primary analysis. For example, the distribution of some baseline covariates is different between groups, and therefore, analysis of covariance or propensity score method is more appropriate than the originally planned *t*-test.
- **Change of study population:**
Results from interim analyses may suggest the benefits of a subgroup, or may want to expand the inclusion criteria because of slow patient recruitment.
- **Adaptive seamless design:**
An adaptive seamless Phase 2/3 design in drugs and biologics is a two-stage design consisting of a learning stage and a confirmatory stage where data from both stages are combined for a final analysis. Such two-stage design clinical studies can apply to medical device submissions as well, whereas the terminology of Phase 2/3 is not used. Trials of this type are reviewed rigorously as Phase 3 trials or trials that are intended to provide substantial evidence for product registration in regulatory submissions. There are also operationally seamless but inferentially distinct Phase 2/3 studies, in which two traditional trials (Phase 2 and Phase 3) are conducted under a single study protocol but analyzed separately. For this type of trials, the control of type I error rate in the Phase 3 part is expected, which is the

same as Phase 3 trials intended to provide substantial evidence for regulatory approval of products. It should be noted that the terms “seamless” and “Phase 2/3” were not used in the FDA Draft Guidance for drugs and biologics as they have sometimes been adopted to describe various design features.

The layout of the article is as follows: Issues and considerations regarding the above referred design adaptations are discussed. Case studies are provided to illustrate the issues associated with study adaptations. The article ends with concluding remarks. The discussion in this article focuses on trials that are intended to provide substantial evidence for product registration.

ISSUES AND CONSIDERATIONS ON DESIGN ADAPTATION

The following provides issues and considerations for each possible type of study adaptation.

1. SSR:

SSR is a design adaptation that occurs most frequently in regulatory submissions just after group sequential design. In general, SSR based on blinded interim analyses of aggregate/overall data is considered not to be problematic, as these approaches have very limited potential to introduce bias or impair the validity and interpretability of study results. SSR based on unblinded knowledge of interim treatment effects, however, can raise issues related to type I error inflation and/or operational biases introduction, and has to be approached with greater caution. In planning any unblinded SSRs in regulatory submissions, clear analytical derivations or statistical justifications to demonstrate a control of type I error rate are expected, as well as strategies/plans to mitigate operational bias at the design stage. For example, reviewers at the Center for Biologics Evaluation and Research (CBER) in the FDA have agreed to multiple proposals using the conditional power approach.^[5] One caveat about SSRs is that the approaches can lead to changing the minimum effect size for which a statistically significant result will occur. Such a statistically significant result, however, may be no longer clinically meaningful. Therefore, decision rules with respect to sample size increase to ensure that statistically significant results are also clinically meaningful should be clearly prespecified. In addition, the timing of an SSR should be factored into the plan. An early SSR (e.g., 20% of information) may result in an unreliable new sample size because of limited accrued data that are subject to large variation, whereas a late SSR (e.g., 80% information) may be pointless as the initial planned accrual may already be completed by

this time. Simulations may be explored to determine the optimal timing for interim analyses for SSRs under different scenarios. Similar with any adaptive approaches, the timing and the rules for any SSRs should be clearly prespecified in the study protocol. For trials with multiple SSRs, it raises a serious concern that data generated from the study fluctuate and that the effect of the proposed product may not be understood to a degree of conducting trials that aim to provide substantial evidence for product registration. Studies aiming for exploratory purposes are usually performed to investigate unknowns about the product in this situation.

2. Termination or/and selection of treatment/dose arms

In many drugs and biologic product developments, there may be still some doubts about the optimal dose of the experimental product for Phase 3 even after a carefully conducted Phase 2 program. One of the reasons is that patients exposed to the product in the Phase 2 program may be limited, and therefore, questions about the estimates of treatment effect are still remained. As a result, several possible doses of the product may be investigated further. An adaptive design combined with multiple doses of the experimental product in trials is thought to offer the opportunity to select the optimal dose of the product and demonstrate the efficacy and safety of the selected dose at trial completion. There are considerations as stated next.

A statistical issue of this type of adaptive design is whether the type I error is controlled. In fact, simulation results have shown that the type I error rate can be inflated if the withdrawal of an arm is based on data of the ongoing trial^[6] and no adjustment has been made. Such an inflation of the type I error rate can be significant in the case of the redistribution of the unused significance level to the remaining arms for comparisons after the withdrawal of a treatment arm. There is ample literature investigating the control of type I error and power^[7] for dose selection. Recently, unconditional and conditional approaches have been proposed for the one-winner treatment selection.^[8–10] Another issue for this adaptation feature is the safety profile of the experimental doses tested in the study. The selection of the optimal dose in this type of trial is mostly focused on efficacy. As a result, a most efficacious dose may be selected with, however, an unacceptable safety profile for the intended indication. This adaptation feature could be carefully explored using simulations under different scenarios. Dose response studies are considered to be conducted prior to embarking larger scale trials for product registration if there are many unknowns about product dose.

3. Change of the primary endpoint

The primary endpoint is the variable that is most relevant to the product indication and best reflects the

disease progression and the effect of treatment. To establish the efficacy of an experimental treatment for a specified indication, study results would need to show the benefit of the experimental treatment to the control (can be placebo) statistically and clinically with respect to the primary efficacy endpoint that is prespecified in the study. For many applications, primary endpoints are well established within the regulatory agencies and clinical societies. For example, in a heart failure clinical trial, the primary efficacy endpoint is usually a composite endpoint that consists of several components, such as deaths, myocardial infarction (MI), and stroke. Death or MI alone is not adequate to capture the progression of heart failure. It is possible in a heart failure clinical trial that an experimental treatment does not show effect statistically for the specified composite endpoint. It, however, may show a statistically significant effect for a different composite endpoint that consists of different components. However, changing the primary endpoint may alter the intended fundamental clinical/statistical questions, which are the originally planned hypotheses. This could contradict the confirmatory nature of trials for product registration purposes. Changing the primary endpoint based on results from ongoing trials that are intended to provide primary evidence for product registration may not be justifiable.

4. Changes between superiority to non-inferiority

Many active control trials in regulatory submissions in some disease areas do not incorporate a placebo arm because of ethical reasons. The effectiveness of the experimental product in active control trials in the absence of the placebo arm can be established via two strategies—the superiority of the experimental product to the active control product or the non-inferiority that the experimental product is not much worse than the active control product, and therefore, it would have been superior to placebo if the placebo had been in the trials. For the later strategy, the effectiveness of the control group must have been established previously compared to placebo in a well-controlled clinical trial setting, as the non-inferiority margin would rely on such information. The non-inferiority margin in a non-inferiority trial would need to be prespecified in regulatory submissions.

There is ample statistical literature regarding the consequences of determination of the non-inferiority margin based on the data derived from the current ongoing trials.^[11–13] The change of the non-inferiority margin based on the data observed during the course of the trial is not justifiable, and the study results may not be interpretable. In contrast, if a trial is planned to demonstrate non-inferiority and this is achieved based on interim results, there may be a motivation

to complete the study enrollment to demonstrate the superiority of the experimental treatment to the active control in efficacy and to generate more safety data. This possibility, however, should be planned in advance and should be prespecified in the protocol.

5. Selection of patient subgroups based on interim results

In clinical trials, there will be some differential responses to the experimental treatment for subgroups or strata of patients. Although it is desirable that results derived from a clinical trial are generalized to a wider patient population, sometimes interest may focus on patient subgroups that are most beneficial from the experimental treatment. On the other hand, when a substantial difference in responses between subgroups (or strata) is apparent, one is tempted to drop the patient stratum that is minimally responsive to the treatment as this may increase study power. Another reason for dropping a stratum is that the experimental treatment shows detrimental effect for the stratum of patients.^[14]

Adaptive selection of patient subgroups based on interim results has been proposed in regulatory submissions. Some even proposed to use results of point estimates to drop a patient stratum midway through the trial. One immediate concern is that it is not clear how to quantify the operating characteristics of the study. Simulated type I error rate may no longer correspond to the proper hypotheses to be tested as the assumptions made at the beginning of the simulation (wider patient population) may no longer be considered for rejection after study adaptation or at the end of the trial (narrower patient population). When relatively little information is available about which groups of patients would benefit from investigational treatment, study adaptation of selecting patient subgroups should be explored in the early phases of investigational treatment evaluation. It is of a great concern if the purpose of such study adaptation is to salvage trials that would otherwise be compromised due to low study powers or heterogeneous response across patient strata.

6. Simulations

Trial simulations play an important role in adaptive design studies, especially in trials that intend to provide substantial evidence for product registration. The major design properties of type I error rate and power of the trial often cannot be evaluated analytically in this type of trials. Trial simulations may greatly facilitate the evaluation of the type I error rate and power, the average sample size required, and the timing of interim analysis under various realistic scenarios. Simulations have variously been used to support novel study designs in regulatory submissions. Depending on the degree of complexity of a proposed adaptive design, extensive simulation

studies could be necessary. However, trial simulations require a variety of assumptions that may be difficult to justify and may well turn out to be false in practice. In any case where a final study outcome is out of the range of simulated scenarios at the planning stage, interpretation of the trial results might be complicated, and this could affect the FDA review of a premarketing application.

CASE STUDIES

Three case studies with features of study design adaptations are presented in this section.

Case Study 1

Indacaterol 75 mcg once daily was approved by the FDA in 2011 for bronchodilator treatment of airflow obstruction in patients with chronic obstructive pulmonary disease (COPD).^[15,16] The premarket application was initially submitted in 2008 for FDA review, where data generated from a two-stage Phase 3 adaptive design trial compared four doses of Indacaterol (75, 150, 300, or 600 mcg) to placebo, Formoterol, or Tiotropium in patients with COPD were included. The study was a double-blind, parallel-group, placebo-controlled, and randomized study with approximately 2000 patients planned. The primary efficacy endpoint was trough forced expiratory volume (FEV1) at 24 h post-dose after 12 weeks of treatment. A selection of optimal Indacaterol doses was made through a prespecified interim analysis. The final analysis used Bonferroni alpha adjustment for the four initial Indacaterol dose groups. Trial simulations were used to evaluate the statistical properties of the two-stage adaptive trial design. The trial schema is shown in Fig. 1.

Sponsor’s interim analysis resulted in the selection of two Indacaterol doses (150 and 300 mcg) to continue into stage 2. The final study results of each selected dose

were statistically significantly superior to placebo with a *p*-value smaller than 0.001. However, there were safety concerns regarding the selected 150 and 300 mcg doses. Cardiovascular (CV) and cerebrovascular adverse events as well as possible asthma-related deaths were observed in the selected Indacaterol groups. After further review of efficacy curves, the doses selected that were carried forward to stage 2 of the adaptive design trial were not justifiable. As a result, the sponsor conducted additional trials for dose–response and product registration that led to the approval of Indacaterol 75 mcg dose that was not evaluated in stage 2 of the two-stage adaptive design trial.

It is worth noting that the initial intent of the two-stage adaptive design trial was to select doses and to demonstrate superiority over active control arms, Formoterol or Tiotropium. This may subsequently result in the selection of higher doses of Indacaterol to carry forward to stage 2, and later yield unacceptable benefit–risk analysis. This case study also illustrates the importance of careful dose–response assessments in a product development. By combining dose selection and confirmation of the selected dose in one study, there is a risk that the selected dose may provide an unacceptable safety profile though it was shown to be highly statistically significant compared to placebo with respect to efficacy. This example also illustrates that the “white space” between separate Phase 2 and 3 trials in a conventional program can be useful and important as it permits thoughtful considerations of data collected prior to embarking a large scale of Phase 3 trials.

Case Study 2

This case study concerns the change of the primary endpoint during the course of CAPRICORN clinical trial. The study was designed as multinational, double-blind, parallel-group, and randomized. The objective was to study the effects of Carvedilol on mortality and morbidity in patients with left ventricular dysfunction, with or

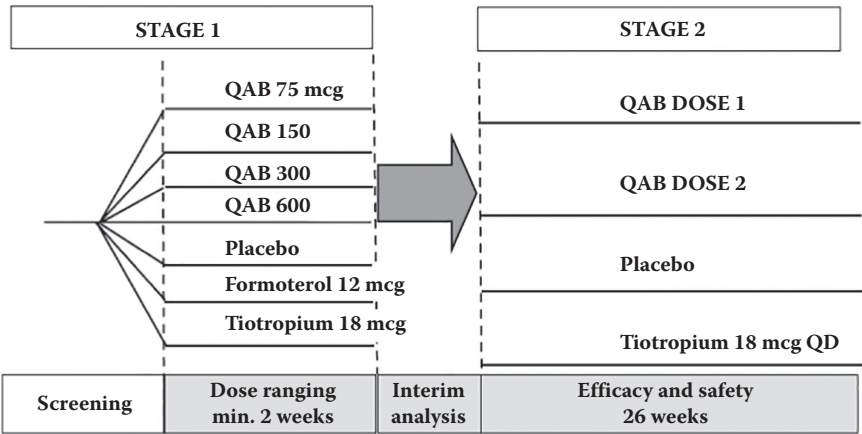


Fig. 1 Schema of two-stage adaptive trial design

without clinical evidence of heart failure, or post MI compared with placebo.^[17] A total of 975 and 984 patients were randomized to receive Carvedilol and placebo, respectively. Study results were presented at the Cardiovascular Advisory Committee meeting in 2003.^[18]

At the study design stage, the primary endpoint was specified as all-cause mortality, and the secondary endpoint was a composite of mortality and CV hospitalization. One interim analysis of all-cause mortality at 125 deaths was planned in the original protocol. On the basis of evidence from outside of the trial (MERIT-HF), Data Safety Monitoring Board (DSMB) recommended an unplanned change in the primary endpoint to enhance study power. Recommendation was made prior to having carried out any efficacy analyses of the trial. As a result, the composite endpoint of mortality and CV hospitalization, which was originally pre-specified as a secondary endpoint, was added to be another primary endpoint, in addition to all-cause mortality. The study would be declared as a success if Carvedilol was superior to placebo in either one of the primary endpoints. For testing purpose and to preserve a study-wise type I error rate, the revised protocol assigned the primary endpoint, a composite of all-cause mortality and CV hospitalization, with a significance level of $\alpha = 0.045$, and the primary endpoint of all-cause mortality with a significance level of $\alpha = 0.005$, where $\alpha = 0.001$ for the interim analysis of mortality at 125 deaths and $\alpha = 0.004$ for the final analysis.

Following the study completion, the trial results did not meet the superiority comparison statistically in both primary endpoints ($p = 0.30$ and 0.031 for composite of all-cause mortality and CV hospitalization and all-cause mortality, respectively, which are greater than the nominal significance levels of 0.045 and 0.004). On the other hand, the study would have shown a statistical superiority of Carvedilol to placebo if the changing of endpoints at mid-course of the trial had not been implemented. One would argue that Carvedilol showed a statistically significant effect on the all-cause mortality endpoint as the p -value of 0.031 is less than a significance level of 0.05 . However, with the change of endpoints during the course of the trial, the comparison of $p = 0.031$ for the endpoint of all-cause mortality to 0.05 is no longer valid.

The change of primary endpoint in this case study was based on external information. There exists risk as the patient populations observed, and clinical practice, may be factors attributed to outcomes and resulted in different treatment effect sizes. The risk of changing endpoints using accumulated data in the same and ongoing trial is not eliminated as the degree of reliability of study results at the interim could be uncertain and whether subjects enrolled in the study earlier adequately represent the target population.

Case Study 3

Seamless Phase 2/3 adaptive design can sometimes be useful in some trial settings that eliminate unpromising doses

early, while carrying forward to Phase 3 only those doses that demonstrate potential public health benefit. Seamless Phase 2/3 adaptive design was implemented in a vaccine trial of Merck's recently approved 9-valent human papillomavirus vaccine (HPV-9). The study was designed to select the optimal dose among several doses in Phase 2 part and carried forward the selected dose to Phase 3 part for efficacy and safety. In this trial, Phase 2 part randomized subjects in an equal allocation to one of three HPV-9 vaccine doses or the control group Gardasil in a three-dose regimen. The best HPV-9 vaccine dose was selected based on immunogenicity and safety data analyzed at the interim analysis. Following selection of the optimal HPV-9 vaccine dose in Phase 2, a Phase 3 efficacy study enrolling an additional 13,380 subjects was conducted. These additional subjects combined with those subjects enrolled in Phase 2 who received the selected HPV-9 vaccine dose or the control vaccine Gardasil were evaluated for efficacy and safety to provide the primary evidence for product registration. In this case study, although the "white space" between Phase 2 and 3 trials in a conventional program as well as the End-of-Phase 2 meeting with the FDA was eliminated, there were several characteristics of the trial setting and product that supported the use of seamless design for this trial. First, the HPV-9 vaccine is a higher valent, which is the second-generation version of Gardasil, the 4-valent HPV vaccine approved in the United States in 2006. Consequently, there were limited unknowns about the HPV-9 vaccine. Second, the interim results were based on immunogenicity data for the four original serotypes (6/11/16/18) that are common to both the HPV-9 vaccine and Gardasil. There were clear criteria for dose selection. Efficacy assessment in the Phase 3 part was based on clinical endpoints related to the five new serotypes in HPV-9. Since there was low correlation expected between the immune responses for the four original serotypes and the efficacy outcomes for the five new serotypes, type I error inflation due to dose selection was expected to be minimal. The study protocol has specified a conservative statistical approach to estimate vaccine efficacy to mitigate any minimal type I error inflation that might occur. Third, several committees were involved in the dissemination of interim results. A sponsor internal committee that did not involve in the study performed the dose selection at the interim based only on immunogenicity data for the four original serotypes. The selected dose was confirmed by an external Data Monitoring Committee (DMC) based on both immunogenicity and safety. Firewalls on how interim results were disseminated were in place to protect the trial integrity. Following the completion of the trial, the sponsor performed simulations and results showed the control of type I error rate empirically for the seamless design. The simulations results also appear to show low correlation between the immune response and clinical efficacy endpoints. The final result of relative vaccine efficacy of HPV-9 compared to the 4-valent Gardasil was estimated to be 96%–97%, depending on case definition.

The above HPV-9 trial may be considered a successful implementation of a seamless Phase 2/3 adaptive design as different characteristics of trial setting and product supported the design. Whether an End-of-Phase 2 meeting with FDA is needed in the product development, it would depend on numerous factors, and the proposed trials intended to provide substantial evidence for product registration would need to be rigorously evaluated. The sponsor should discuss with regulators the possibility of a seamless Phase 2/3 adaptive design trial in advance when they intend to plan one for product registration.

CONCLUDING REMARKS

Adaptive designs have become recognized in clinical trials and other areas of public health research. For a clinical trial with adaptive study design, the central issues of concern are the clinical interpretation of overall trial results, operational logistics, trial conduct, and whether the integrity of study results is maintained. Methods that simply gain control over the type I error rate are not adequate. Issues, such as the operational biases and characteristics due to interim looks of data, and pooling data before and after design adaptations, make the regulatory evaluation of outcomes generated from adaptive design trials more complex and challenge. Sponsors may propose adaptive design clinical trials when appropriate for their product development. One, however, should recognize that the best study design proposals may not be adaptive designs. Sometimes, a simpler design is most efficient. As indicated in a survey summary paper,^[19,20] the number of adaptive design submissions in CBER and CDRH did not continue to increase after the release of the FDA Draft Guidance for drugs and biologics. Our experience indicates that whether an adaptive feature can actually benefit a study may depend on a number of factors. For example, early-stage studies (e.g., Phase 1/2 or feasibility/pilot) with adaptive design elements are generally encouraged to explore. Trials that are intended to provide primary evidence for product registration would require more rigorous reviews and deliberations. We provide the following points for considerations when an adaptive design clinical trial is planned for product registration purposes:

1. What are the advantages of using an adaptive design instead of a conventional design (**why**)?
2. What adaptation features does the proposed design plan (**what**)?
3. How is the adaptive design conducted?
 - Are details related to timing and execution clearly specified in the protocol?
 - Is the study-wise type I error rate controlled? If simulations are used to demonstrate the control of type I error rate, are the included simulation studies adequate?

- Are strategies and logistics being planned to mitigate the operational bias? Are study decision rules, success criteria, and stopping rules clearly specified?

Finally, it should be emphasized that the added flexibility of adaptation is not without potential pitfalls. Sponsors are expected to address the issues laid out in the article when they propose an adaptive design clinical trial in regulatory submissions. Study adaptation should never be used as a way to salvage study results and should never be used to avoid careful planning through research and development of medical products during the development process.

REFERENCES

1. US FDA. Critical Path Initiative. Available at www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/default.htm (accessed May 23, 2017).
2. US FDA. Challenges and Opportunities Report—March 2004. Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products. Available at www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/ucm077262.htm (accessed May 23, 2017).
3. US FDA. Adaptive Design Clinical Trials for Drugs and Biologics Draft Guidance, 2010. Available at www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf (accessed May 23, 2017).
4. US FDA. Guidance for Industry and Food and Drug Administration Staff: Adaptive Designs for Medical Device Clinical Studies, 2016. Available at www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM446729.pdf (accessed May 23, 2017).
5. Chen, Y.; DeMets, D.; Lan, K. Increasing the sample size when the unblinded interim result is promising. *Stat. Med.* **2004**, *23* (7), 1023–1038.
6. Tsong, Y.; Hung, H.; Wang, S.; Cui, L.; Nuri, W. Dropping a treatment arm in clinical trial with multiple arms. In *Proceedings of the Biopharmaceutical Section*; American Statistical Association: Anaheim, CA, 1997; 58–63.
7. Li, G.; Zhu, J.; Ouyang, S.; Xie, J.; Deng, L.; Law, G. Adaptive designs for interim dose selection. *Stat. Biopharm. Res.* **2009**, *1* (4), 366–376.
8. Sampson, A.; Sill, M. Drop-the-losers design: Normal case. *Biomet. J.* **2005**, *47* (3), 257–268.
9. Stallard, N.; Todd, S. Sequential designs for phase III clinical trials incorporating treatment selection. *Stat. Med.* **2003**, *22* (5), 689–703.
10. Todd, S.; Stallard, N. A new clinical trial design combining phases 2 and 3: Sequential designs with treatment selection and a change of endpoint. *Drug Inform. J.* **2005**, *39* (2), 109–118.
11. Morikawa, T.; Yoshida, M. A useful testing strategy in phase III trials: Combined test of superiority and test of equivalence. *J. Biopharm. Stat.* **1995**, *5* (3), 297–306.

12. Wang, S.; Hung, H.; Tsong, Y.; Cui, L. Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. *Stat. Med.* **2001**, *20* (13), 1903–1912.
13. Huang, H.; Wang, S. Multiple testing of non-inferiority hypotheses in active controlled trials. *J. Biopharm. Stat.* **2004**, *14* (2), 327–335.
14. The Coronary Drug Project Research Group. Practical aspects of decision making in clinical trials: The coronary drug project as a case study. *Control. Clin. Trials.* **1981**, *1* (4), 363–376.
15. US FDA. Drugs@FDA: FDA Approved Drug Products. Available at www.accessdata.fda.gov/scripts/cder/daf/index.cfm?event=overview.process&ApplNo=022383 (accessed May 23, 2017).
16. US FDA. Briefing Information for the March 8, 2011 Meeting of the Pulmonary-Allergy Drugs Advisory Committee. Available at www.fda.gov/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/Pulmonary-AllergyDrugsAdvisoryCommittee/ucm245635.htm (accessed May 23, 2017).
17. US FDA. Drug Approval Package. Available at www.accessdata.fda.gov/drugsatfda_docs/nda/2003/20-297s009_Coreg.cfm (accessed May 23, 2017).
18. US FDA. Food and Drug Administration, Cardiovascular and Renal Drugs Advisory Committee, January 7, 2003, Briefing Information. Available at www.fda.gov/ohrms/dockets/ac/03/briefing/3920b2.htm (accessed May 23, 2017).
19. Lin, M.; Lee, S.; Zhen, B.; Scott, J.; Horne, A.; Solomon, G.; Russek-Cohen, E. CBER's experiences in adaptive design clinical trials. *Ther. Innov. Regul. Sci.* **2016**, *50* (2), 195–203.
20. Yang, X.; Thompson, L.; Chu, J.; Liu, S.; Lu, H.; Zhou, J.; Gomatam, S.; Tang, R.; Zhao, Y.; Ge, Y.; Gray, G. Adaptive design practice at the Center for Devices and Radiological Health (CDRH), January 2007 to May 2013. *Ther. Innov. Regul. Sci.* **2016**, *50* (6), 710–717.

Adaptive Design Methods in Clinical Trials

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Annpey Pong

Merck Research Laboratories, Rahway, New Jersey, U.S.A.

Abstract

The aim of this entry is not only to provide a comprehensive summarization of the issues that are commonly encountered when applying/implementing the adaptive design methods in clinical research but also to include recent development such as the role of the independent data safety monitoring board and sample size estimation/allocation, justification, and adjustment when implementing a much more complicated adaptive design in clinical trials.

Keywords: Adaptive Design; Flexible Design; Group Sequential Design; Adaptive design in clinical trials; Flexible SSRE design.

INTRODUCTION

In the past several decades, it has been recognized that increasing spending of biomedical research does not reflect an increase of the success rate of pharmaceutical/clinical research and development. The low success rate of pharmaceutical/clinical development could be due to following factors: (i) there is a diminished margin for improvement that escalates the level of difficulty in proving drug benefits, (ii) genomics and other new sciences have not yet reached their full potential, (iii) mergers and other business arrangements have decreased candidates, (iv) easy targets are the focus as chronic diseases are harder to study, (v) failure rates have not improved, and (vi) rapidly escalating costs and complexity decreases the willingness/ability to bring many candidates forward into the clinic.^[1] In early 2000, the U.S. Food and Drug Administration (FDA) kicked off the Critical Path Initiative to assist the sponsors in (i) identifying possible causes, (ii) providing resolutions, and (iii) increasing the efficiency and the probability of success in pharmaceutical research and development. In its 2004 Critical Path Report, the FDA presented its diagnosis of the scientific challenges underlying the medical product pipeline problems. Two years later, the FDA released a Critical Path Opportunities List that calls for advancing innovative trial designs by using prior experience or accumulated information in trial design. Many researchers interpret it as the encouragement of using innovative adaptive design methods in clinical trials, while some researchers believe it is the recommendation for the use of Bayesian approach for assessment of treatment effect in pharmaceutical/clinical development. The purpose of adaptive design methods in clinical trials is

to provide the flexibility to the investigator for identifying best (optimal) clinical benefit of the test treatment under study in a timely and efficient fashion without undermining the validity and integrity of the intended study.

The concept of adaptive design can be traced back to 1970s when adaptive randomization (play-the-winner) and a class of designs for sequential clinical trials were introduced.^[2] As a result, most adaptive design methods in clinical research and development are referred to as adaptive randomization,^[3–8] group sequential designs with the flexibility for stopping a trial early due to safety, futility and/or efficacy,^[9–13] and flexible sample size re-estimation at interim for achieving the desired statistical power by controlling the overall type I error rate at a pre-specified level of significance.^[14–16] The use of adaptive design methods for modifying the trial procedures and/or statistical methods of ongoing clinical trials based on accrued data has been practiced for years in pharmaceutical/clinical research and development. Adaptive design methods in clinical research are very attractive to pharmaceutical/clinical scientists for the following reasons. First, it reflects medical practice in real world. Second, it is ethical with respect to both efficacy and safety (toxicity) of the test treatment under investigation. Third, it is not only flexible but also efficient in the early phase of clinical development. However, one concern is whether the p-value or confidence interval regarding the treatment effect obtained after the modification is correct or reliable. In addition, another concern is that the use of adaptive design methods in a clinical trial may lead to a totally different trial that is unable to address scientific/medical questions that the trial is intended to answer.

In recent years, the potential use of adaptive design methods in clinical trials has attracted much attention.

For example, the Pharmaceutical Research and Manufacturers of America (PhRMA) and Biotechnology Industry Organization (BIO) have established adaptive design working groups and proposed/published white papers regarding strategies, methodologies, and implementations for regulatory consideration.^[17,18] However, there is no universal agreement in terms of definition, methodologies, applications, and implementations. In addition, many journals have also published special issues on adaptive design for evaluation the potential use of adaptive trial design methods in clinical research and development. These scientific journals include, but are not limited to, *Biometrics* (Vol. 62, No. 3), *Statistics in Medicine* (Vol. 25, No. 19), *Journal of Biopharmaceutical Statistics* (Vol. 15, No. 4 and Vol. 17, No. 6), *Biometrical Journal* (Vol. 48, No. 4), and *Pharmaceutical Statistics* (Vol. 5, No. 2). In addition, many professional conferences/meetings have devoted special sessions for discussion of the feasibility, applicability, efficiency, validity, and integrity of the potential use of the innovative adaptive design methods in clinical trials in the past several years. For example, the FDA/Industry Statistics Workshop has offered adaptive sessions and workshops from industrial, academic, and regulatory perspectives consecutively between 2006 and 2008. More details regarding the use of adaptive design methods in clinical trials can be found in the books by Chow and Chang^[19] and Chang^[18].

The purpose of this entry is not only to provide a comprehensive summarization of the issues that are commonly encountered when applying/implementing the adaptive design methods in clinical research but also to include recently development such as the role of the independent data safety monitoring board and sample size estimation/allocation, justification, and adjustment when implementing a much more complicated adaptive design in clinical trials. In the next section, commonly employed adaptations and the resultant adaptive designs are briefly described. Also included in this section are regulatory and statistical perspectives regarding the use of adaptive design methods in clinical trials. The impact of protocol amendments, challenges of by design adaptations, and obstacles of retrospective adaptations when applying adaptive design methods in clinical trials are described in the section “Impact, Challenges, and Obstacles”. Some trial examples and strategies for clinical development are discussed in the sections “Some Examples” and “Strategies for Clinical Development,” respectively. A brief concluding remark is given in the last section.

WHAT IS ADAPTIVE DESIGN?

In clinical trials, it is not uncommon to modify trial procedures and/or statistical methods during the conduct of clinical trials based on the review of accrued data at

interim. The purpose is not only to efficiently identify clinical benefits of the test treatment under investigation but also to increase the probability of success of the intended clinical trial. Trial procedures are referred to as the eligibility criteria, study dose, treatment duration, study endpoints, laboratory testing procedures, diagnostic procedures, criteria for evaluability, and assessment of clinical responses. Statistical methods include randomization scheme, study design selection, study objectives/hypotheses, sample size calculation, data monitoring and interim analysis, statistical analysis plan, and/or methods for data analysis. In this chapter, we will refer to the adaptations (changes or modifications) made to the trial and/or statistical procedures as the adaptive design methods. Thus, an adaptive design is defined as a design that allows adaptations to trial and/or statistical procedures of the trial after its initiation without undermining the validity and integrity of the trial.^[20] In one of their publications, with the emphasis of the feature of by design adaptations only (rather than ad hoc adaptations), the PhRMA Working Group on Adaptive Design refers to an adaptive design as a clinical trial design that uses accumulating data to decide on how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial.^[17] In many cases, an adaptive design is also known as a flexible design.^[21,22]

Adaptations

An adaptation is referred to as a modification or a change made to trial procedures and/or statistical methods during the conduct of a clinical trial. By definition, adaptations that are commonly employed in clinical trials can be classified into the categories of prospective adaptation, concurrent (or ad hoc) adaptation, and retrospective adaptation. Prospective adaptations include, but are not limited to, adaptive randomization; stopping a trial early due to safety, futility, or efficacy at interim analysis; dropping the losers (or inferior treatment groups); sample size re-estimation; and so on. Thus, prospective adaptations are usually referred to as by design adaptations as described in the PhRMA white paper.^[17] Concurrent adaptations are usually referred to as any ad hoc modifications or changes made as the trial continues. Concurrent adaptations include, but are not limited to, modifications in inclusion/exclusion criteria, evaluability criteria, dose/regimen and treatment duration, changes in hypotheses and/or study endpoints, and so on. Retrospective adaptations are usually referred to as modifications and/or changes made to statistical analysis plan prior to database lock or unblinding of treatment codes. In practice, prospective, ad hoc, and retrospective adaptations are implemented by study protocol, protocol amendments, and statistical analysis plan with regulatory reviewer's consensus, respectively.

Types of Adaptive Designs

Based on the adaptations employed, commonly considered adaptive designs in clinical trials include, but are not limited to (i) adaptive randomization design, (ii) group sequential design, (iii) N-adjustable (or flexible sample size re-estimation) design, (iv) drop-the-losers design, (v) adaptive dose finding design, (vi) biomarker-adaptive design, (vii) adaptive treatment-switching design, (viii) adaptive-hypothesis design, (ix) adaptive seamless (e.g., phase I/II or phase II/III) trial design, and (x) multiple adaptive design. These adaptive designs are briefly described below.

Adaptive Randomization Design

An adaptive randomization design is a design that allows modification of randomization schedules based on varied and/or unequal probabilities of treatment assignment in order to increase the probability of success. As a result, an adaptive randomization design is sometimes referred to as a play-the-winner design since it will increase the probability of success. Commonly applied adaptive randomization procedures include treatment-adaptive randomization,^[3,4] covariate-adaptive randomization, and response-adaptive randomization.^[6,8]

Although an adaptive randomization design could increase the probability of success, it may not be feasible for a large trial or a trial with relatively longer treatment duration because the randomization of a given subject depends on the response of the previous subject. A large trial or a trial with a relatively longer treatment duration utilizing adaptive randomization design will take a much longer time to complete. Besides, the randomization schedule may not be available prior to the conduct of the study. Moreover, statistical inference on treatment effect is often difficult to obtain due to the complexity of the randomization scheme. In practice, a statistical test is often difficult, if not impossible, to obtain due to complicated probability structure as the result of adaptive randomization, which has also limited the potential use of adaptive randomization design in practice.

Group Sequential Design

A group sequential design is a design that allows for prematurely stopping a trial due to safety, futility/efficacy, or both with options of additional adaptations based on results of interim analysis. Many researchers refer to a group sequential design as a typical adaptive design because some adaptations may be applied after the review of interim results of the study, such as stopping the trial early due to safety, efficacy, and/or futility. In practice, various stopping boundaries based on different boundary functions for controlling an overall type I error rate are available in the literature.^[6,9,10,19,23,24] In recent years,

the concept of two-stage adaptive design has led to the development of the adaptive group sequential design.^[11–14]

It should be noted that when additional adaptations such as adaptive randomization, dropping the losers, and/or adding additional treatment arms (in addition to the commonly considered adaptations such as stop the trial early due to safety, efficacy and/or futility and sample size re-estimation, and so on) are applied to a typical group sequential design after the review of the interim results, the resultant group sequential design is usually referred to as an adaptive group sequential design. In this case, the standard methods for the typical group sequential design may not be appropriate. In addition, it may not be able to control the overall type I error rate at the desired level of 5% if (i) there are additional adaptations (e.g., changes in hypotheses and/or study endpoints) and/or (ii) there is a shift in target patient population due to additional adaptations or protocol amendments.

Flexible Sample Size Re-Estimation (SSRE) Design

A flexible sample size re-estimation (or N-adjustable) design is referred to as an adaptive design that allows for sample size adjustment or re-estimation based on the observed data at interim. In clinical research, it can be verified that the selected sample size is a function of type I error rate, type II error rate (or power), treatment effect (or clinically meaningful difference), and variability associated with the response. A typical approach (power analysis) is to select a sample size that will achieve a desired power by fixing other parameters such as type I error rate, the clinically meaningful difference, and the variability associated with the response. In practice, it is impossible to select a sample size that controls all parameters. For a flexible SSRE design, sample size adjustment or re-estimation could be done in either a blinding or unblinding fashion based on the criteria of maintaining anticipated treatment effect-size, controlling variability within a tolerable limit, achieving a desired conditional power, and/or reaching certain degree of reproducibility probability.^[11,14–16,25–28] Thus, sample size in a flexible SSRE trial design is a random variable. Sample size re-estimation suffers from the same disadvantage as the original power analysis for sample size calculation prior to the conduct of the study because it is performed by treating estimates of the study parameters, which are obtained based on data observed at interim, as true values. Note that the criteria of maintain treatment effect size, controlling variability, and achieving conditional power are considered one-parameter problem, while the criterion of reaching reproducibility is a two-parameter problem.

According to an informal communication with the medical/statistical reviewers at the FDA, it is not a good clinical/statistical practice to start with a small number and then perform sample size re-estimation (adjustment) at interim by ignoring the clinically meaningful difference

that one wishes to detect for the intended clinical trial. It should be noted that the observed difference at interim based on a small number of subjects may not be of statistical significance (i.e., it may be observed by chance alone). In addition, there is variation associated with the observed difference which is an estimate of the true difference. Thus, standard methods for sample size re-estimation based on the observed difference with a limited number of subjects may be biased and misleading. To overcome these problems, in practice, a sensitivity analysis (with respect to variation associated with the observed results at interim) for sample size re-estimation design is recommended.

Drop-The-Losers Design

A drop-the-losers design is a design that allows dropping the inferior treatment groups as well as adding additional (promising) arms. A drop-the-losers design is useful in early phase of clinical development especially when there are uncertainties regarding the dose levels.^[29–32] The selection criteria (including the selection of initial dose, the increment of the dose, and the dose range) and decision rules play important roles for drop-the-losers designs. Dose groups that are dropped may contain valuable information regarding dose response of the treatment under study. Typically, drop-the-losers design is a two-stage design. At the end of the first stage, the inferior arms will be dropped based on some pre-specified criteria. The winners will then proceed to the next stage. In practice, the study is often powered for achieving a desired power at the end of the second stage (or at the end of the study). In other words, there may not be any statistical power for the analysis at the end of the first stage for dropping the losers (or picking up the winners).

In practice, it is not uncommon to drop the losers or pick up the winners based on so-called precision analysis.^[16] The precision approach is an approach based on the confidence level for achieving statistical significance. In other words, the decision will be made (i.e., to drop the losers) if the confidence level for observing a statistical significance (i.e., the observed difference is not by chance alone or it is reproducible with the pre-specified confidence level) is exceed a pre-specified confidence level. In addition, other criteria for dropping the losers such as predictive probability of success and probability of being the best dose (or treatment arm) are also commonly considered.^[69] Liu et al (2017) compared relative performances of these criteria in terms of the probability of correctly identifying the most promising dose (treatment arm) under a two-stage adaptive trial design.^[70]

Note that in a drop-the-losers design, a general principle is to drop the inferior treatment groups or add promising treatment arms but at the same time it is suggested that the control group be retained for a fair and reliable comparison at the end of the study. It should be noted that dose groups that are dropped may contain valuable information

regarding dose response of the treatment under study. In practice, it is also suggested that subjects who are assigned in the inferior dose groups should be switched to the better dose group for ethical consideration. Treatment switching in a drop-the-losers design could complicate statistical evaluation in the dose selection process. Note that some clinical scientists prefer the term pick-the-winners rather than drop-the-losers.

Adaptive Dose Finding Design

The purpose of an adaptive dose finding (e.g., escalation) design is multifold and includes (i) the identification whether there is a dose response, (ii) the determination of the minimum effective dose (MED) and/or the maximum tolerable dose (MTD), (iii) the characterization of dose response curve, and (iv) the study of dose ranging. The information obtained from an adaptive dose finding experiment is often used to determine the dose level for the next phase of clinical development.^[33–35] For adaptive dose finding design, the method of continual re-assessment method (CRM) in conjunction with Bayesian approach is usually considered.^[36–38] Mugno et al.^[39] introduced a non-parametric adaptive urn design approach for estimating a dose-response curve. More details regarding PhRMA's proposed statistical methods, the reader should consult with a special issue recently published at the Journal of Biopharmaceutical Statistics, Vol. 17, No. 6. Note that a typical approach for adaptive dose finding design focuses on dose response (severe toxicity and/or tolerability) curve. In practice, it is suggested that both safety including mild-to-moderate toxicity and efficacy be considered.

Note that according to the ICH E4 guideline on Dose-response Information to Support Drug Registration, there are several types of dose-finding (response) designs, which are (i) randomized parallel dose-response designs, (ii) crossover dose-response design, (iii) forced titration design (dose escalation design), and (iv) optimal titration design (placebo-controlled titration to endpoint). Some commonly asked questions for an adaptive dose finding design include, but are not limited to, (i) how to select the initial dose, (ii) how to select the dose range under study, (iii) how to achieve statistical significance with a desired power with a limit number of subjects, and (iv) what the selection criteria and decision rules should be if one would like to make a decision based on safety, tolerability, efficacy and/or pharmacokinetic information, and (v) what the probability is of achieving the optimal dose. In practice, a clinical trial simulation and/or sensitivity analysis is often recommended to evaluate/address the above questions.

Biomarker-Adaptive Design

A biomarker-adaptive design is a design that allows for adaptations based on the response of biomarkers such as genomic markers. An adaptive biomarker design involves

biomarker qualification and standard, optimal screening design, and model selection and validation. It should be noted that there is a gap between identifying biomarkers that associated with clinical outcomes and establishing a predictive model between relevant biomarkers and clinical outcomes in clinical development. For example, correlation between a biomarker and a true clinical endpoint makes a prognostic marker. However, correlation between a biomarker and a true clinical endpoint does not make a predictive biomarker. A prognostic biomarker informs the clinical outcomes, independent of treatment. They provide information about the natural course of the disease in individuals who have or have not received the treatment under study. Prognostic markers can be used to separate good- and poor-prognosis patients at the time of diagnosis. A predictive biomarker informs the treatment effect on the clinical endpoint.^[18]

A biomarker-adaptive design can be used to (i) select right patient population (e.g., enrichment process for selection of a better target patient population), (ii) identify nature course of disease, and (iii) detect disease earlier. In clinical research and development, biomarker-adaptive design, which has led to the research of targeted clinical trials, is not only the key to the success of precision medicine, but also help in developing personalized medicine.^[18,40,41]

Adaptive Treatment-Switching Design

An adaptive treatment-switching design is a design that allows the investigator to switch a patient's treatment from an initial assignment to an alternative treatment if there is evidence of lack of efficacy or safety of the initial treatment.^[42,43] In cancer clinical trials, estimation of survival is a challenge when treatment-switching has occurred in some patients. A high percentage of subjects who switched due to disease progression could lead to change in hypotheses to be tested. In this case, sample size adjustment for achieving a desired power is necessary.

Adaptive-Hypotheses Design

An adaptive-hypotheses design refers to a design that allows modification or change in hypotheses based on interim analysis results.^[44] Adaptive-hypotheses designs are often considered before database lock and/or prior to data unblinding, which are implemented by the development of statistical analysis plan (SAP). Typical examples include the switch from a superiority hypothesis to a non-inferiority hypothesis and the switch between the primary study endpoint and the secondary endpoints. The purpose for switching from a superiority hypothesis to a non-inferiority hypothesis is to increase the probability of the success of the clinical trial. A typical approach is to first establish non-inferiority and then test for superiority. In this way, we do not have to pay for statistical penalty due to the principle of closed testing procedure. The idea of

switching the primary study endpoints and the secondary endpoints is also to increase the probability of success of clinical development. In practice, it is not uncommon to observe positive results in secondary endpoint while fail to demonstrate clinical benefit for the primary endpoints. In this case, there is a strong desire to switch the primary endpoints and the secondary endpoints whenever it is scientifically, clinically, and regulatory justifiable.

It should be noted that, for the switch from a superiority hypothesis to a non-inferiority hypothesis, the selection of non-inferiority margin is critical and has an impact on sample size adjustment for achieving the desired power. According to the ICH guideline, the selected non-inferiority margin should be both clinical and statistical justifiable.^[45,46] Regarding the switch between the primary endpoint and the secondary endpoints, there has been a tremendous debate about controlling the overall type I error rate at the 5% level of significance. Thus, as an alternative, it has been recommended by many researchers considering switching from the primary endpoint to either a co-primary endpoint or a composite endpoint. However, the optimal allocation of the alpha spending function has raised another statistical/clinical/regulatory concern.

Seamless Adaptive Trial Design

A seamless adaptive trial design refers to a program that addresses within single trial objectives that are normally achieved through separate trials of clinical development. A seamless adaptive design is a trial design that would use data from patients enrolled before and after the adaptation in the final analysis.^[47–49] Commonly considered adaptive seamless trials in clinical development include an adaptive seamless phase I/II design in early clinical development and an adaptive seamless phase II/III trial design in late phase clinical development.

An adaptive seamless phase II/III design is a two-stage design consisting of a learning or exploratory stage (phase IIb) and a confirmatory stage (phase III). A typical approach is to power the study for the phase III confirmatory phase and obtain valuable information with certain assurance using confidence interval approach at the phase II learning stage. Its validity and efficiency, however, has been challenged.^[50–52] Statistical methods for a combined analysis under the situation that the study objectives (and/or endpoints) are similar but different at different stages are studied in the literature.^[71] One of the key assumptions is that there is a well-established relationship between study endpoints at different stages. In other words, study endpoint (e.g., biomarker, surrogate endpoint, or same clinical endpoint with shorter duration) at earlier stage is predictive of study endpoint (e.g., clinical endpoint) at later stage. More research regarding sample size estimation/allocation and statistical analysis for seamless adaptive designs with different study objectives and/or

study endpoints for various data types (e.g., continuous, binary, and time-to-event) is needed.

Most recently, on August 9, 2016, the FDA allows a sponsor to conduct a phase II/III/IV seamless adaptive clinical trial for NASH (Non-Alcoholic SteatoHepatitis) studies. This is very encouraging to the sponsors. However, statistical methods for multiple-stage seamless adaptive trial design of this kind is not fully developed. More research on this topic is necessary.

Multiple Adaptive Design

Finally, a multiple adaptive design is any combinations of the above adaptive designs. Commonly considered multiple adaptive designs include (i) the combination of adaptive group sequential design, drop-the-losers design, and adaptive seamless trial design and (ii) adaptive dose-escalation design with adaptive randomization.^[19] In practice, because statistical inference for a multiple-adaptation design is often difficult, it is suggested that a clinical trial simulation be conducted to evaluate the performance of the resultant multiple adaptive design at the planning stage.

When applying a multiple adaptive design, some frequently asked questions include (i) how to avoid/control potential operational biases which may be introduced due to various adaptations that apply to the trial, (ii) how to control the overall type I error rate at the 5%, (iii) how to determine the required sample size for achieving the study objectives with the desired power, and (iv) how to maintain the quality, validity, and integrity of the trial. The tradeoff between the flexibility/efficiency and scientific validity/integrity needs to be carefully evaluated before a multiple adaptive design is implemented in clinical trials.

Regulatory/Statistical Perspectives

From a regulatory point of view, the use of adaptive design methods based on accrued data in clinical trials may introduce operational bias such as selection bias, method of evaluation, early withdrawal, and modification of treatment. Consequently, it may not be able to preserve the overall type I error rate at the pre-specified level of significance. In addition, p-values may not be correct and the corresponding confidence intervals for the treatment effect may not be reliable. Moreover, it may result in a totally different trial that is unable to address the medical questions that original study intended to answer. Li^[53] also indicated that commonly seen adaptations that have an impact on the type I error rate include, but are not limited to, (i) sample size adjustments at interim, (ii) sample size allocations to treatments, (iii) deletions of, additions to, or changes to treatment arms, (iv) shifts in target patient population such as changes in inclusion/exclusion criteria, (v) changes in statistical test strategy, (vi) changes in study endpoints, and (vii) changes in study objectives such as the

switch from a superiority trial to a non-inferiority trial. As a result, it is difficult to interpret the clinically meaningful effect size for the treatments under study.^[54]

From a statistical point of view, major (or significant) adaptations to trial and/or statistical procedures could (i) introduce bias/variation to data collection, (ii) result in a shift in location and scale of the target patient population, and (iii) lead to inconsistency between hypotheses to be tested and the corresponding statistical tests. These concerns will not only have an impact on the accuracy and reliability of statistical inference drawn on the treatment effect but also present challenges to biostatisticians for development of appropriate statistical methodology for an unbiased and fair assessment of the treatment effect.

Note that although the flexibility of modifying study parameters is very attractive to clinical scientists, several regulatory questions/concerns arise. First, what level of modifications to the trial procedures and/or statistical procedures would be acceptable to the regulatory authorities? Second, what are the regulatory requirements and standards for review and approval process of clinical data obtained from adaptive clinical trials with different levels of modifications to trial procedures and/or statistical procedures of ongoing clinical trials? Third, has the clinical trial become a totally different clinical trial after the modifications to the trial procedures and/or statistical procedures for addressing the study objectives of the originally planned clinical trial? These concerns should be addressed by the regulatory authorities before the adaptive design methods can be widely accepted in clinical research and development.

IMPACT, CHALLENGES, AND OBSTACLES

Impact of Protocol Amendments

In practice, for a given clinical trial, it is not uncommon to have between three and five protocol amendments after the initiation of the clinical trial. One of the major impacts of many protocol amendments is that the target patient population may have been shifted during the process, which may have resulted in a totally different target patient population at the end of the trial. A typical example is the case when significant modifications are applied to inclusion/exclusion criteria of the study protocol. As a result, the resultant actual patient population following certain modifications to the trial procedures is a moving target patient population rather than a fixed target patient population. As indicated in Chow and Chang,^[19] the impact of protocol amendments on statistical inference due to shift in target patient population (moving target patient population) can be studied through a model that link the moving population means with some covariates.^[55] Chow and Shao^[55] derived statistical inference for the original target patient population for simple cases.

Challenges in By Design Adaptations

In clinical trials, commonly employed prospective (by design) adaptations include stopping the trial early due to safety, futility, and/or efficacy, sample size re-estimation (adaptive group sequential design), dropping the losers (adaptive dose finding design), and combining two separate trials into a single trial (adaptive seamless design). These designs are typical multiple-stage designs with different adaptations. In this section, major challenges in analysis and design are described. Recommendations and future development for resolution are provided whenever possible.

The major difference between a classic multiple-stage design and an adaptive multiple-stage design is that an adaptive design allows adaptations after the review of interim analysis results. These by-design adaptations include sample size adjustment (re-assessment or re-estimation), stopping the trials due to safety, efficacy/futility, or dropping the losers (picking the winners). Note that commonly considered adaptive group sequential design, adaptive dose finding design, and adaptive seamless trial design are special cases of multiple-stage designs with different adaptations. In this section, we will discuss major challenges in design (e.g., sample size calculation) and analysis (controlling type I error rate under moving target patient population) of an adaptive multiple-stage design with $K - 1$ interim analyses.

A multiple-stage adaptive group sequential design is very attractive to sponsors in clinical development. However, major (or significant) adaptations such as modification of doses and/or change in study endpoints may introduce bias/variation to data collection as the trial continues. To account for these (expected and/or unexpected) biases/variation, statistical tests are necessary adjusted to maintain the overall type I error, and the related sample size calculation formulas have to be modified for achieving the desired power. In addition, the impact on statistical inference is not negligible if the target patient population has been shifted due to major or significant adaptations and/or protocol amendments. This has presented challenges to biostatisticians in clinical research when applying a multiple-stage adaptive design. In practice, thus, it is worthy pursuing the following specific directions: (i) derive valid statistical test procedures for adaptive group sequential designs assuming model, which relates the data from different interim analyses, (ii) derive valid statistical test procedures for adaptive group sequential designs assuming the random-deviation model, (iii) derive valid Bayesian methods for adaptive group sequential designs, and (iv) derive sample size calculation formulas for various situations. Tsiatis and Mehta^[50] showed that there exists an optimal (i.e., uniformly more powerful) design for any class of sequential design with a specified error spending function. It should be noted that adaptive designs do not require in general a fixed error spending function. One of

major challenges for an adaptive group sequential design is that the overall type I error rate may be inflated when there is a shift in target patient population.^[56]

For adaptive dose-finding design, Chang and Chow's method can be improved by the following specific directions: (i) study the relative merits and disadvantage of their method under various adaptive methods, (ii) examine the performance of an alternative method by forming the utility first with different weights to the response levels and then modeling the utility, and (iii) derive sample size calculation formulas for various situations. An adaptive seamless phase II/III design is a two-stage design that consists of two phases namely a learning (or exploratory) phase and a confirmatory phase. One of the major challenges for designs of this kind is that different study endpoints are often considered at different stages for achieving different study objectives. In this case, the standard statistical methodology for assessment of treatment effect and for sample size calculation cannot be applied.

For a two-stage adaptive design, the aforementioned method by Chang^[57] can be applied. Chang's method, however like other stagewise combination methods, is valid under the assumption of constancy of the target patient populations, study objectives, and study endpoints at different stages.

Obstacles of Retrospective Adaptations

In practice, retrospective adaptations such as adaptive-hypotheses may encounter prior to database lock (or unblinding) and be implemented through the development of statistical analysis plan. To illustrate the impact of retrospective adaptations, we first consider the situation where switching hypotheses between a superiority hypothesis and a non-inferiority hypothesis. For a promising test drug, the sponsor would prefer an aggressive approach for planning a superiority study. The study is usually powered to compare the promising test drug with an active control agent.

However, the collected data may not support superiority. Instead of declaring the failure of the superiority trial, the sponsor may switch from testing superiority to testing the following non-inferiority hypotheses. The margin is carefully chosen to ensure that the treatment of the test drug is larger than the placebo effect and, thus, declaring non-inferiority to the active control agent means that the test drug is superior to the placebo effect. The switch from a superiority hypothesis to a non-inferiority hypothesis will certainly increase the probability of success of the trial because the study objective has been modified to establishing non-inferiority rather than showing superiority. This type of switching hypotheses is recommended provided that the impact of the switch on statistical issues and inference (e.g., appropriate statistical methods) on the assessment of treatment effect is well justified.

SOME EXAMPLES

In this section, we will present some examples for adaptive trial designs, which have been implemented in practice.^[58] These trial examples include (i) an adaptive dose escalation design for early phase cancer trials, (ii) a multiple-stage adaptive design for Non-Hodgkin's Lymphoma (NHL) trial, (iii) a phase IV drop-the-losers adaptive design for multiple myeloma trial, and (iv) a two-stage seamless phase I/II adaptive trial design for hepatitis C virus (HCV) trial, which are described below.

Example 1: Adaptive dose escalation design for early phase cancer trials

In a phase I cancer trial, suppose that the primary objective is to identify the maximum tolerated dose (MTD) for a new test drug. Based on the animal studies, it is estimated that the toxicity (dose limiting toxicity rate) is 1% for the starting dose 25 mg/m² (one-tenth of the lethal dose). The dose limiting toxicity (DLT) rate at MTD is defined as 0.25 and the MTD is estimated to be 150 mg/m². A typical approach is to consider so-called 3 + 3 traditional escalation rule (TER). On the other hand, an adaptive type approach is to apply the continual re-assessment method (CRM) in conjunction with a Bayesian approach. We can compare the 3 + 3 TER approach and the Bayesian CRM adaptive method through simulations. A logistic toxicity model is chosen for the simulations. The dose interval sequence is chosen to be the customized dose increment sequence (Increment factors = 2, 1.67, 1.33, 1.33, 1.33, 1.33, 1.33, 1.33). The evaluations of the escalation designs are based on the criteria of safety, accuracy, and efficiency. Simulation results for all three methods for three different MTD scenarios are summarized in Table 1.

In this example, as it can be seen that if the true MTD is 150 mg/m², the TER approach underestimates MTD (125 mg/m²), whereas the Bayesian CRM adaptive method also slightly underestimates the MTD (141 mg/m²). The average number of patients required is 19.4 and 15.5 for the TER approach and the Bayesian CRM adaptive method, respectively. From a safety perspective, the average number of DLTs is 2.9 and 2.5 per trial for the TER approach and the Bayesian CRM, respectively. As a result, the Bayesian CRM adaptive method is preferable. More details regarding Bayesian CRM adaptive method and adaptive dose finding can be found in Chang and Chow^[38].

Example 2: Multiple-stage adaptive design for NHL trial

A phase III two parallel group Non-Hodgkin's Lymphoma trial was designed with three analyses. The primary endpoint is progression-free survival (PFS), the secondary endpoints are (i) overall response rate (ORR) including complete and partial response and (ii) complete response rate (CRR). The estimated median PFS is 7.8 months and 10 months for the control and test groups, respectively. Assume a uniform enrollment with an accrual period of 9 months and a total study duration of 23 months. The estimated ORR is 16% for the control group and 45% for the test group. The classic design with a fixed sample size of 375 subjects per group will allow for detecting a three-month difference in median PFS with 82% power at a one-sided significance level of $\alpha = 0.025$. The first interim analysis will be conducted on the first 125 patients/group (or total $N_1 = 250$) based on ORR. The objective of the first IA is to modify the randomization. Specifically, if the difference in ORR (test-control), $\Delta_{\text{ORR}} > 0$, the enrollment will continue. If $\Delta_{\text{ORR}} \leq 0$, then the enrollment will stop. If the enrollment is terminated prematurely, there will be one final analysis for efficacy based on PFS and possible efficacy claimed on the secondary endpoints. If the enrollment continues, there will be a second interim analysis based on PFS. The second interim analysis will could lead to either claim efficacy or futility, or continue to the next stage with possible sample size re-estimation. For the final analysis of PFS, when the primary endpoint (PFS) is significant, the analyses for the secondary endpoints will be performed for the potential claim on the secondary endpoints. During the interim analyses, the patient enrollment will not stop.

Example 3: Drop-the-losers adaptive design for multiple myeloma trial

For phase IV study, the adaptive design can also work well. For example, a year after an oncology drug came on the market, physicians are using the drug with different combinations for treating patients with multiple myeloma (MM). However, there is a strong desire to know which combination is the best for the patient populations. Many physicians have their own experiences, but no one has convincing data. Therefore, the sponsor is planned a trial to investigate the optimal combination for the drug. In this scenario, we can use much smaller sample size than we would in phase III trials because the issue is focused on the type I control as the drug is approved. The issue is as follows: given a minimum

Table 1 Summary of Simulation Results for the Designs

Method	True MTD	Mean predicted MTD	Mean number of patients	Mean number of DLTs
3 + 3 TER	100	86.7	14.9	2.8
CRM	100	99.2	13.4	2.8
3 + 3 TER	150	125	19.4	2.9
CRM	150	141	15.5	2.5
3 + 3 TER	200	169	22.4	2.8
CRM	200	186	16.8	2.2

Source: From Ref. [58].

Table 2 Simulation Results of Phase IV Drop-the-Losers Design

Interim sample size per group	Final sample size per group	Probability of identifying the optimal arm (%)
25	50	80.3
50	100	91.3

Note: Response rate for the five arms: 0.4, 0.45, 0.45, 0.5, and 0.6.
Source: From Ref. [58].

clinically meaningful difference (e.g., two weeks in survival), what is the probability of the trial will be able to identify the optimal combination? Again this can be done through simulations, the strategy is to start with about five combinations, drop inferior arms, and calculate the probability of selecting the best arm under different combinations of sample size at the interim and final stages. In this adaptive design, we will drop two arms based on the observed response rate, that is, the two arms with lowest observed response rates will be dropped at interim analysis and the rest three arms will be carried forward to the second stage. The characteristics of the design are presented in Table 2 for different sample sizes.

Given the response rate 0.4, 0.45, 0.45, 0.5, and 0.6 for the five arms and 91% power, if a traditional design is used, there will be nine multiple comparisons (adjusted $\alpha = 0.0055$ using Bonferroni method). The required sample size for the traditional design is 209 per group or a total of 1,045 subjects comparing a total of 500 subjects for the adaptive trial (250 at the first stage and 150 at the second stage). In cases where the null hypothesis is true (the response rates are the same for all the arms), it doesn't matter too much which arm is chosen as optimal.

Example 4: Two-stage seamless Phase I/II adaptive design for HCV trial

A pharmaceutical company is interested in conducting a clinical trial utilizing a two-stage seamless adaptive design for evaluation of safety, tolerability, and efficacy of a test treatment as compared to a standard care treatment for treating subjects with hepatitis C virus (HCV) genotype 1 infection. The proposed adaptive trial design consists of two stages of dose selection and efficacy confirmation. The primary efficacy endpoint is the incidence of sustained virologic response (SVR), defined as an undetectable HCV RNA level (< 10 IU/mL) at 24 weeks after treatment is complete (Study Week 72). This assessment will be made when the last subject in Stage 2 has completed the Study Week 72 evaluation. In both Stages 1 and 2, consider the incidence of the following primary efficacy variables of (i) rapid virologic response (RVR), that is, undetectable HCV RNA level at Study Week 4; (ii) early virologic response (EVR), that is, ≥ 2 -log10 reduction in HCV RNA level at Study Week 12 compared with the baseline level; (iii) end-of-treatment response (EOT), that is, undetectable HCV RNA level at Study Week 48; and (iv) SVR, that is, undetectable HCV RNA level at Study Week 72 (24 weeks after treatment is complete).

Stage 1 is a four-arm randomized evaluation of three dose levels of continuous subcutaneous (SC) delivery of the test treatment compared with PegIntron (standard care) given as once weekly SC injections. All subjects will receive oral weight-based

ribavirin. After all Stage 1 subjects have completed Study Week 12, an interim analysis will be performed. The interim analysis will provide information to enable selection of an active dose of the test treatment based on safety/tolerability, outcomes, and early indications of efficacy to proceed to testing for non-inferiority compared with standard of care in Stage 2. Depending upon individual response data for safety and efficacy, Stage 1 subjects will continue with their randomization assignments for the full planned 48 weeks of therapy, with a final follow-up evaluation at Study Week 72. Stage 2 will be a non-inferiority comparison of a selected dose and the same PegIntron active-control regimen used in Stage 1, both again given with oral ribavirin, for up to 48 weeks of therapy, with a final follow-up evaluation at Study Week 72. A second interim analysis of all available safety/tolerability, outcomes, and efficacy data from Stage 1 and Stage 2 will be performed when all Stage 2 subjects have completed Study Week 12. Depending upon individual response data for safety and efficacy, Stage 2 subjects will receive the full planned 48 weeks of treatment, with final follow-up at Study Week 72 (see Fig. 1).^[71]

For power analysis for sample size estimation, a total study enrollment of 388 subjects will be required (120 subjects or 30 subjects per arm for Stage 1 and 268 subjects or 134 subjects per arm for Stage 2) to account for a probable dropout rate of 15% as well as 2 planned interim analyses using the O'Brien-Fleming method. Stage 1 will enroll a total of 120 subjects divided equally among the four treatment arms to gather sufficient data to select an active dose of the test treatment to proceed to testing in Stage 2. Stage 2 will enroll an additional cohort of 268 subjects divided equally between its two treatment arms to provide a sufficient number of subjects to evaluate non-inferiority of continuous interferon delivery to standard-of-care interferon therapy, for an overall total of 306 subjects enrolled in Stage 1 and Stage 2 combined. Therefore, accounting for 15% attrition in each Stage, we expect that 164 subjects from the selected dose arms (30 from Stage 1 plus 134 additional subjects enrolled in Stage 2) and 164 subjects from the PegIntron active-control arms (30 from Stage 1 plus an additional 134 subjects enrolled in Stage 2) will need to be enrolled to meet the study objective of achieving an 80% power for establishing non-inferiority (with a non-inferiority margin of 15%) at the overall type I error rate of 5%.

Note that above two-stage seamless trial design is a combination of group sequential design, drop-the-losers design and seamless adaptive phase I/II design utilizing precision analysis (i.e., confidence interval approach) for decision-making on dose selection at the first stage. If we apply adaptive randomization for the second stage, the study design would be even more complicated. From regulatory point of view, it is important to ensure that (i) no operational biases are introduced during the conduct of the trial and (ii) the overall type I error rate is well controlled at the 5% level.

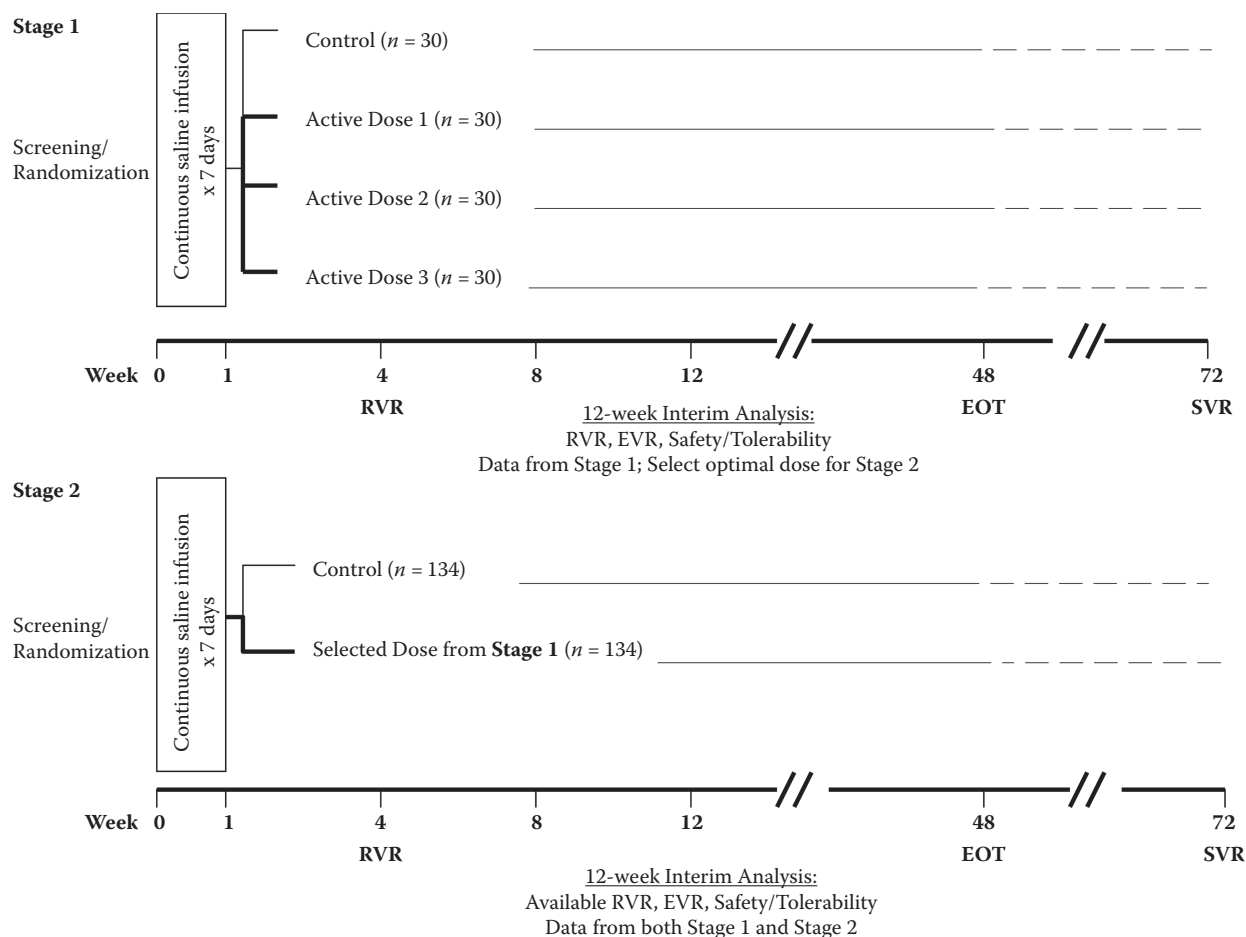


Fig. 1 Diagram of study design

Example 5: Targeted clinical trials/biomarker-adaptive trial design

Biomarker-adaptive trial design is often considered for identifying those patients most likely to respond to the test treatment under study through some validated diagnostic tests of biomarkers such as genomic markers. The process is referred to as the enrichment process in clinical trials, which has led to the concept of targeted clinical trials.^[59] As indicated by many researchers,^[60–64] the disease targets at the molecular level can be identified after completion of the Human Genome Project (HGP). As a result, the importance of diagnostic tests for identification of molecular targets increases as more targeted clinical trials will be conducted for the individualized treatment of patients (personalized medicine). For example, based on the risk of distant recurrence determined by a 21-gene Oncotype DXf0d2 breast cancer assay, patients with a recurrence score of 11–25 in the TAILORx (Trial Assigning Individualized Options for Treatment) trial sponsored by the United States National Cancer Institute (NCI) are randomly assigned to receive either adjuvant chemotherapy and hormonal therapy or adjuvant hormonal therapy alone.^[65]

Despite of different technical platforms employed in the diagnostic devices for molecular targets used in the trial, the assay falls in the category of the in vitro diagnostic multivariate index assays (IVDMIA)^[66] based on the selected differentially

expressed genes for detection of the patients with the molecular targets. In addition, to reduce the variation, the IVDMIA do not usually use all genes during the development stage. Therefore, identification of the differentially expressed genes between different groups of patients is the key to the accuracy and reliability of the devices for molecular targets. Once the differentially expressed are identified, the next task is to search an optimal representation or algorithm that provides the best discrimination ability between the patients with molecular targets and those without the targets. The current validation procedure for diagnostic device is for the assay based on one analyte. However, the IVDMIA are in fact the parallel assays based on the intensities of multiple analytes. As a result, the current approach to assay validation for one analyte may not be appropriate and is inadequate for validation of IVDMIA.

With respect to the enrichment design for the targeted clinical trials, patients with positive diagnosis for the molecular targets are randomized to receive the test drug or the control. However, because no IVDMIA can provide the perfectly correct diagnosis, some patients with positive diagnosis may not actually have the molecular targets. Consequently, the treatment effect of the test drug for the patients with targets is underestimated. On the other hand, estimation of the treatment effect based on the data from the targeted clinical trials needs to take into consideration the variability associated with the estimates of accuracy of the

IVDMIA such as positive predictive value and false positive rate obtained from the clinical effectiveness trials of the IVDMIA.

In general, there are three classes of targeted clinical trials. The first type is to evaluate the efficacy and safety of targeted treatment for the patients with molecular targets; Herceptin® clinical trials belong to this class. The second type of targeted clinical trials is to select the best treatment regimen for the patients based on the results of some tests for prognosis of clinical outcomes. The last type of the targeted clinical trials is to investigate the correlation of the treatment effect with variations of the molecular targets. Because the objectives of different targeted clinical trials vary, therefore the FDA Drug-Diagnostic Co-Development Concept Paper^[67] proposed three different designs to meet different objectives of targeted clinical trials. These three designs are given in Figs. 2–4.

Design A is the enrichment design,^[68] in which the only patients tested positively for identification of molecular targets are randomized either to receive the test drug or the concurrent control. The enrichment design is usually employed when there is a high degree of certainty that the drug response occurs only in the patients tested positively for the molecular targets and the mechanism of pathological pathways is clearly understood. Most of the Herceptin® phase III clinical trials used the enrichment design. However, as pointed out in the FDA Concept Paper, the description of test sensitivity and specificity will not be possible using this type of design without drug and placebo data in the patients tested negative for the molecular targets. Design B is a stratified randomized design and stratification factor is the results of the test for the molecular targets. In other words, the patients are stratified into two groups depending upon whether the diagnostic test is either positive or negative. Then a separate randomization is independently performed within each group to receive the test drug or concurrent control. The information of the test results for the molecular targets in Design C is primarily

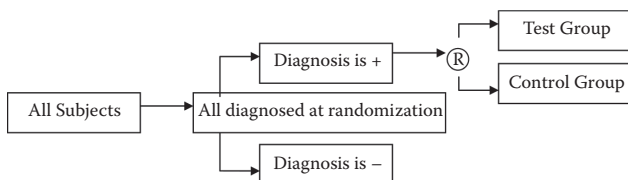


Fig. 2 Design A for targeted clinical trials

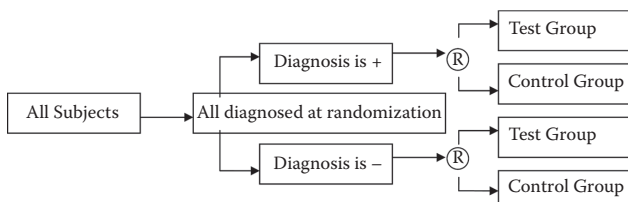


Fig. 3 Design B for targeted clinical trials

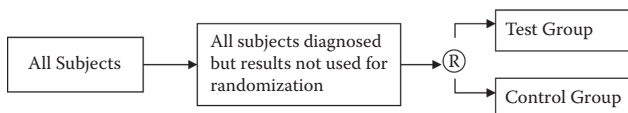


Fig. 4 Design C for targeted clinical trials

used as covariates and is not involved with randomization. Sometimes, only a part of the patients are tested for the molecular targets. Design C is useful when the association of the treatment effect of the drug with the results of the diagnostic test needs to be further explored.

STRATEGIES FOR CLINICAL DEVELOPMENT

Clinical development of a new drug product is a lengthy and costly process, which includes phases I to III clinical development (prior to regulatory review and approval) and phase IV clinical development (post-approval). For life-threatening diseases or diseases with unmet medical need (rare diseases), this lengthy clinical development process is not acceptable. Such cases call for the use of adaptive design methods in clinical trials in order to shorten the development process (speed) without compromising the safety and efficacy of the drug product under investigation (validity) by maximizing the power for identify best clinical benefit of the drug product under investigation with limited number of subjects (efficiency). As a result, many adaptive design methods are developed for archiving the ultimate goals of validity, efficiency, and speed in clinical development. In practice, however, many other important factors beyond statistical components may have an impact on the development process. These factors include, but are not limited to, patient enrollment, treatment durations, and time required for regulatory review/approval.

Commonly considered strategies for the use of adaptive design methods in clinical development include, but are not limited to, adaptive dose finding, adaptive seamless phase I/II in early clinical development, and adaptive seamless phase II/III in late phase clinical development. These strategies help not only to shorten the development time in a more efficient way but also to increase the probability of success in clinical development. As an example, let us consider the development of a new drug product for a rare disease of multiple myeloma (MM) with unmet medical need. A traditional approach is to conduct a typical phase I study for identifying the maximum tolerated dose (MTD), which is often considered as the optimal dose for the late phase of clinical development. In practice, there are several options that we can run this study with several doses or dose schedules. For example, this trial can be run with parallel dose groups or sequentially with the test drug as a single or add-on agent. The traditional approach will not be able to provide the entire spectrum of the dose response of the test drug. On the other hand, if we consider an adaptive approach, it allows us to use more options at the beginning and drop some of the arms (options) if they are either too toxic or ineffective (activity too low or requires very high dose that makes the treatment very costly—biologic product may be very costly). In addition, because the treatment arms are often correlated, the information such as dose limiting toxicity (DLT) from different

arms can be synchronized using methods such as Bayesian hierarchical model.^[38]

Similar approaches can be used for phase II trials, phase I/II, or phase II/III seamless design. In this approach, the interim endpoint may be a marker such as response rate or time to disease progression. Some of the inferior arms can be dropped at interim. If some of the arms are promising, we can apply different randomization scheme to assign more patients to the superior arms (play-the-winner). In this case, not only will the total cost not increase, but additionally we will increase the chance of success because the increase of sample size for the superior treatment arms is most promising. In addition, the schedules for the trial are similar because the total number of patients remains unchanged. This adaptive design strategy can be applied to phase III trials, but with much less arms. For phase IV studies, the adaptive design can also be applied. However, the situation is different because the drug has been approved during phase IV clinical development. The focus of phase IV clinical development will be the safety rather than efficacy.

CONCLUDING REMARKS

As indicated earlier, although the use of adaptive design methods in clinical trials is motivated by their flexibility and efficiency, many researchers are not convinced and still challenge their validity and integrity.^[50] As a result, many discussions concern the flexibility, efficiency, validity, and integrity of these methods. Li^[53] suggested a couple of principles to be used when one implements an adaptive design in a clinical trial: (i) adaptation should not alter trial conduct and (ii) type I errors should be preserved. Beyond these principles, some basic considerations such as dose/dose regimen, study endpoints, treatment duration, and logistics should be carefully evaluated for feasibility.^[54] To maintain the validity and integrity of an adaptive design with complicated adaptations, it is strongly suggested that an independent data monitoring committee (IDMC) be established. In practice, IDMCs have been widely used in group sequential design with adaptations of stopping a trial early and sample size re-estimation. The role and responsibility of an IDMC for a clinical trial using adaptive design should be clearly defined. An IDMC usually conveys very limited information to investigators or sponsors about treatment effects, procedural conventions, and statistical methods with recommendations in order to maintain the validity and integrity of the study.

When applying adaptive design methods in clinical trials, the feasibility of certain adaptations such as changes in study endpoints/hypotheses should be carefully evaluated to prevent from any possible misuse and abuse of the adaptive design methods. For a complicated multiple adaptive design, an independent data monitoring committee should be established to ensure the integrity of

the study. Clinical trial simulation provide a solution, not the solution, for a complicated multiple adaptive design. In practice, deciding how to validate the assumed predictive model for clinical trial simulation is a major challenge to both investigators and biostatisticians.

In February, 2010, FDA published adrafft guidance on Adaptive Design Clinical Trials for Drug and Biologics. FDA has expressed its intention to slow down the escalating momentum behind adaptive clinical trial designs to allow time to craft better working definitions of the new models and build a better base before moving forward. Efforts are primarily directed to current issues such as the impact of protocol amendments, challenges in by design adaptations, and obstacles of retrospective adaptations as described in the previous sections. Following the draft guidance, there are several successful regulatory submissions utilizing adaptive trial designs since 2010.^[69] As a result, the draft guidance is currently being revised and under review within the FDA.

In summary, from the clinical point of view, adaptive design methods reflect real clinical practice in clinical development. Adaptive design methods are very attractive due to their flexibility and are very useful especially in early clinical development. From the statistical point of view, the use of adaptive methods in clinical trials makes current good statistics practice even more complicated. The validity of the use of adaptive design methods is not well established and fully understood. The impact of statistical inference on treatment effect should be carefully evaluated under the framework of moving target patient population as the result of protocol amendments. In practice, regulatory agencies may not realize that the adaptive design methods for review and approval of regulatory submissions have been employed for years without any scientific basis. Guidelines regarding the use of adaptive design methods and *Good Adaptive Design Practices* must be developed so that appropriate statistical methods and statistical software packages can be developed accordingly.

REFERENCES

1. Woodcock, J. DFA introduction comments: clinical studies design and evaluation issues. *Clin. Trials* **2005**, 2, 273–275.
2. Wei, L.J. The adaptive biased-coin design for sequential experiments. *Ann. Stat.* **1978**, 9, 92–100.
3. Efron, B. Forcing a sequential experiment to be balanced. *Biometrika* **1971**, 58, 403–417.
4. Lachin, J.M. Statistical properties of randomization in clinical trials. *Control. Clin. Trials* **1988**, 9, 289–311.
5. Atkinson, A.C.; Donev, A.N. *Optimum Experimental Designs*; Oxford University Press: New York, 1992.
6. Rosenberger, W.F.; Stallard, N.; Ivanova, A.; Harper, C.N.; Ricks, M.L. Optimal adaptive designs for binary response trials. *Biometrics* **2001**, 57, 909–913.
7. Rosenberger, W.F.; Lachin, J. *Randomization in Clinical Trials*; John Wiley and Sons: New York.

8. Hardwick, J.P.; Stout, Q.F. Optimal few-stage designs. *J. Stat. Plann. Inference* **2002**, *104*, 121–145.
9. Lan, K.K.G.; DeMets, D.L. Group sequential procedures: calendar versus information time. *Stat. Med.* **1987**, *8*, 1191–1198.
10. Wang, S.K.; Tsiatis, A.A. Approximately optimal one-parameter boundaries for a sequential trials. *Biometrics* **1987**, *43*, 193–200.
11. Posch, M.; Bauer, P. Adaptive two stage designs and the conditional error function. *Biom. J.* **1999**, *41*, 689–696.
12. Lehmacher, W.; Wassmer, G. Adaptive sample size calculations in group sequential trials. *Biometrics* **1999**, *55*, 1286–1290.
13. Liu, Q.; Proschan, M.A.; Pledger, G.W. A unified theory of two-stage adaptive designs. *J. Am. Stat. Assoc.* **2002**, *97*, 1034–1041.
14. Cui, L.; Hung, H.M.J.; Wang, S.J. Modification of sample size in group sequential trials. *Biometrics* **1999**, *55*, 853–857.
15. Chung-Stein, C.; Anderson, K.; Gallo, P.; Collins, S. Sample size reestimation: a review and recommendations. *Drug Inf. J.* **2006**, *40*, 475–484.
16. Chow, S.C.; Shao, J.; Wang, H. *Sample Size Calculation in Clinical Research*, 2nd Ed; Chapman and Hall/CRC Press, Taylor & Francis: New York, 2007.
17. Gallo, P.; Chuang-Stein, C.; Dragalin, V.; Gaydos, B.; Krams, M.; Pinheiro, J. Adaptive design in clinical drug development—an executive summary of the PhRMA Working Group (with discussions). *J. Biopharm. Stat.* **2006**, *16* (3), 275–283.
18. Chang, M. *Adaptive Design Theory and Implementation Using SAS and R*; Chapman and Hall/CRC Press, Taylor and Francis: New York, 2007a.
19. Chow, S.C.; Chang, M. *Adaptive Design Methods in Clinical Trials*; Chapman and Hall/CRC Press, Taylor and Francis: New York, 2006.
20. Chow, S.C.; Chang, M.; Pong, A. Statistical consideration of adaptive methods in clinical development. *J. Biopharm. Stat.* **2005**, *15*, 575–591.
21. EMEA. Point to Consider on Methodological Issues in Confirmatory Clinical Trials with Flexible Design and Analysis Plan; The European Agency for the Evaluation of Medicinal Products Evaluation of Medicines for Human Use. CPMP/EWP/2459/02: London, 2002.
22. EMEA. Reflection paper on Methodological Issues in Confirmatory Clinical Trials with Flexible Design and Analysis Plan; The European Agency for the Evaluation of Medicinal Products Evaluation of Medicines for Human Use. CPMP/EWP/2459/02: London, 2006.
23. Jennison, C.; Turnbull, B.W. *Group Sequential Methods with Applications to Clinical Trials*; Chapman and Hall: New York, 2000.
24. Jennison, C.; Turnbull, B.W. Meta-analysis and adaptive group sequential design in the clinical development process. *J. Biopharm. Stat.* **2005**, *15*, 537–558.
25. Shih, W.J. Sample size re-estimation—a journey for a decade. *Stat. Med.* **2001**, *20*, 515–518.
26. Proschan, M.A.; Hunsberger, S.A. Designed extension of studies based on conditional power. *Biometrics* **1995**, *51*, 1315–1324.
27. Liu, Q.; Chi, G.Y.H. On sample size and inference for two-stage adaptive designs. *Biometrics* **2001**, *57*, 172–177.
28. Friede, T.; Kieser, M. Sample size recalculation for binary data in internal pilot study designs. *Pharm. Stat.* **2004**, *3*, 269–279.
29. Bauer, P.; Kieser, M. Combining different phases in development of medical treatments within a single trial. *Stat. Med.* **1999**, *18*, 1833–1848.
30. Brannath, W.; Koenig, F.; Bauer, P. Improved repeated confidence bounds in trials with a maximal goal. *Biom. J.* **2003**, *45*, 311–324.
31. Sampson, A.R.; Sill, M.W. Drop-the-loser design: normal case (with discussions). *Biom. J.* **2005**, *47*, 257–281.
32. Posch, M.; Konig, F.; Brannath, W.; Dunger-Baldauf, C.; Bauer, P. Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Stat. Med.* **2005**, *24*, 3697–3714.
33. Bauer, P.; Rohmel, J. An adaptive method for establishing a dose-response relationship. *Stat. Med.* **1995**, *14*, 1595–1607.
34. Whitehead, J. Bayesian decision procedures with application to dose-finding studies. *Int. J. Pharm. Med.* **1997**, *11*, 201–208.
35. Zhang, W.; Sargent, D.J.; Mandrekar, S. An adaptive dose-finding design incorporating both toxicity and efficacy. *Stat. Med.* **2006**, *25*, 2365–2383.
36. O’Quigley, J.; Pepe, M.; Fisher, L. Continual reassessment method: A practical design for phase I clinical trial in cancer. *Biometrics* **1990**, *46*, 33–48.
37. O’Quigley, J.; Shen, L. Continual reassessment method: A likelihood approach. *Biometrics* **1996**, *52*, 673–684.
38. Chang, M.; Chow, S.C. A hybrid Bayesian adaptive design for dose response trials. *J. Biopharm. Stat.* **2005**, *15*, 667–691.
39. Mugno, R.; Zhus, W.; Rosenberger, W.F. Adaptive urn designs for estimating several percentiles of a dose response curve. *Stat. Med.* **2004**, *23*, 2137–2150.
40. Charkravarty, A. Burzykowski, M.; Molenberghs, G.; Buyse, T. Regulatory aspects in using surrogate markers in clinical trials. In *The Evaluation of Surrogate Endpoint*; Eds.; Springer: New York, 2005.
41. Wang, S.J.; O’Neill, R.T.; Hung, H.M.J. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharm. Stat.* **2007**, *6*, 227–244.
42. Shao, J.; Chang, M.; Chow, S.C. Statistical inference for cancer trials with treatment switching. *Stat. Med.* **2005**, *24*, 1783–1790.
43. Branson, M.; Whitehead, W. Estimating a treatment effect in survival studies in which patients switch treatment. *Stat. Med.* **2002**, *21*, 2449–2463.
44. Hommel, G. Adaptive modifications of hypotheses after an interim analysis. *Biom. J.* **2001**, *43*, 581–589.
45. ICH. International Conference on Harmonization Guideline E10: Guidance on Choice of Control Group and Related Design and Conduct Issues in Clinical Trials; The United States Food and Drug Administration: Rockville, Maryland D, July. 2000.
46. Chow, S.C.; Shao, J. On margin and statistical test for noninferiority in active control trials. *Stat. Med.* **2006**, *25*, 1101–1113.

47. Kelly, P.J.; Stallard, N.; Todd, S. An adaptive group sequential design for Phase II/III clinical trials that select a single treatment from several. *J. Biopharm. Stat.* **2005**, *15*, 641–658.
48. Maca, J.; Bhattacharya, S.; Dragalin, V.; Gallo, P.; Krams, M. Adaptive seamless Phase II/III designs—background, operational aspects, and examples. *Drug Inf. J.* **2006**, *40*, 463–474.
49. Chow, S.C.; Tu, Y.H. On two-stage seamless adaptive design in clinical trials. *J. Formosan Med. Assoc.* **2008**, *107*, S51–S59.
50. Tsiatis, A.A.; Mehta, C. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* **2003**, *90*, 367–378.
51. Chow, S.C.; Lu, Q.; Tse, S.K. Statistical analysis for two-stage adaptive design with different study endpoints. *J. Biopharm. Stat.* **2007**, *17*, 1163–1176.
52. Chow, S.C.; Tu, Y.H. On two-stage seamless adaptive design in clinical trials. *J. Formosan Med. Assoc.* **2008**, *107*, 12, S52–S60.
53. Li, N. Adaptive trial design—FDA statistical reviewer's view. the CRT2006 Workshop with the FDA, Arlington, VA, April 4, 2006.
54. Quinlan, J.A.; Gallo, P.; Krams, M. Implementing adaptive designs: logistical and operational consideration. *Drug Inf. J.* **2006**, *40*, 437–444.
55. Chow, S.C.; Shao, J. Inference for clinical trials with some protocol amendments. *J. Biopharm. Stat.* **2005**, *15*, 659–666.
56. Feng, H.; Shao, J.; Chow, S.C. Group sequential test for clinical trials with moving patient population. *J. Biopharm. Stat.* **2007**, *17*, 1227–1238.
57. Chang, M. Adaptive design method based on sum of p-values. *Statistics in Medicine*, **2007b**, *26*, 2772–2784.
58. Chow, S.C.; Chang, M. Adaptive design methods in clinical trials—a review. *Orphanet J. Rare Dis.* **2008**, *3*, 1–13.
59. Liu, J.P.; Chow, S.C. Statistical issues on the diagnostics multivariate index assay for targeted clinical trials. *J. Biopharm. Stat.* **2008**, *18*, 167–182.
60. Simon, R.; Maitournam, A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin. Cancer Res.* **2004**, *10*, 6759–6763.
61. Maitournam, A.; Simon, R. On the efficiency of targeted clinical trials. *Stat. Med.* **2005**, *24*, 329–339.
62. Varmus, H. The new era in cancer research. *Science* **2006**, *312*, 1162–1165.
63. Dalton, W.S.; Friend, S.H. Cancer biomarkers—an invitation to the table. *Science* **2006**, *312*, 1165–1168.
64. Casciano, D.A.; Woodcock, J. Empowering microarrays in the regulatory setting. *Nat. Biotechnol* **2006**, *24*, 1–1103.
65. Sprarano, J. 2006.
66. FDA. *Draft Guidance on In Vitro Diagnostic Multivariate Index Assays*; The United States Food and Drug Administration: Rockville, MD, 2006.
67. FDA. *Draft concept paper on Drug-Diagnostic Co-Development*; The United States Food and Drug Administration: Rockville, MD, 2005.
68. Chow, S.C.; Liu, J.P. *Design and Analysis of Clinical Trials*, 2nd Ed.; John Wiley & Sons: New York, NY, 2004.
69. Lee, S.J.; Lin, M. Adaptive Trial Design – Case Studies. Short course presented at 2016 Duke-Industry Statistics Symposium, Durham, NC, September 14, 2016.
70. Liu, S.; Zheng, J.; Chow, S.C. On dose selection criteria for a two-stage seamless adaptive trial design. Submitted. 2017.
71. Chow, S.C. and Lin, M. Analysis of two-stage adaptive seamless trial design. *Pharmaceutica Analytica Acta*, **2015**, *6*(3). <http://dx.doi.org/10.4172/2153-2435.1000341>.

Adaptive Survival Trials

William F. Rosenberger

Department of Statistics, George Mason University, Fairfax, Virginia, U.S.A.

Lanju Zhang

Department of Biostatistics, MedImmune LLC, Gaithersburg, Maryland, U.S.A.

Abstract

This entry briefly examines clinical trials with a time-to-event primary outcome, or active survival trials, by summarizing two approaches: the treatment effect mapping approach and the optimal allocation approach.

Keywords: Censoring; Doubly biased-coin design; Optimal allocation proportion; Outcome; Treatment effect mapping.

INTRODUCTION

A survival trial is a clinical trial with a time-to-event primary outcome, such as time to death or cure. Typically, survival trials involve a limited recruitment period with staggered entry and are complicated by censoring because of losses to follow-up and other reasons, including administrative censoring. The primary outcome of a survival trial is usually analyzed using the log-rank test, or a member of the log-rank family of tests, including the Gehan test or the Peto-Peto Prentice Wilcoxon test.^[1]

Most survival trials comparing two treatments have traditionally employed equal sample sizes in each treatment arm. However, it has long been known that equal allocation does not always yield the most powerful test in survival analyses.^[2,3] Formulas that give the optimal allocation to maximize power have been derived.^[4] Equal allocation also results in half of patients being assigned to the inferior treatment arm, if one treatment is superior. To mitigate this problem, several authors have proposed adaptive randomization be used in the context of survival trials.^[5,6] In adaptive randomization, allocation probabilities are skewed away from 0.5 to favor the treatment that is performing “better” thus far in the trial. One method that has been proposed is to adapt the probabilities as a function of the current treatment effect in the trial. These probabilities can be adapted before each patient is randomized,^[5] or as groups of patients are randomized.^[4,6] Another method is to define an optimal proportion of patients to be allocated to one treatment, usually dependent on the some unknown parameters, and use a randomization procedure to target the optimal proportion. We refer to survival trials employing adaptive randomization as *adaptive survival trials*. The idea of using accruing data to assign the more favorable treatment in a survival setting can be traced to Ref. [7].

TREATMENT EFFECT MAPPING APPROACH

In the treatment effect mapping approach, one can use any treatment effect measure $S \in \Omega$ along with a function $g: \Omega \rightarrow [0,1]$, continuous, such that $g(0) = 1/2$, $g(x) > 1/2$ if $x > 0$, and $g(x) < 1/2$ if $x < 0$. Let $S(\tau-)$ be the computed treatment effect measure up to time τ . Let $X(\tau) = 1$ if a patient is randomized to treatment A at time τ , and 0 if the patient is randomized to treatment B. Let $\mathcal{F}_{\tau-}$ capture all the information on censoring and events up to time τ . Then a patient to be entered at time τ would be assigned to A with probability $E(X(\tau)|\mathcal{F}_{\tau-}) = g(S(\tau-))$,

As an example, one could use the Mantel formulation of the unnormalized log-rank statistic as S .^[5] Let $0 < \tau_1 < \dots < \tau_K$ be the ordered event times for all patients in the trial, where τ_i is the time from recruitment to event. At each event time τ_i , a 2×2 table is constructed with $\delta_i = 1$ if the event occurred on treatment A, 0 otherwise. Let n_{ji} be the number of patients in the risk set on treatment, $j = A, B$, immediately prior to the event time, and let $N_i = n_{Ai} + n_{Bi}$. Also define n_j , $j = A, B$ to be the total number randomized up to time τ , with $N = n_A + n_B$. Let $K(\tau-)$ be the number of events occurring up to τ . Then, we can let $S(\tau-) = \sum_{i=1}^{K(\tau-)} (\delta_i - n_{Ai}/N_i) / \max(n_A, n_B) \sum_{i=1}^{K(\tau-)} 1/(N - i)$. This formulation restricts $S \in [-1,1]$, where $S = 0$ means no treatment effect, so an appropriate function g would be $g(x) = 0.5(1 + x)$. Yao and Wei^[6] used the Gehan statistic as the treatment effect measure, and proposed the function

$$\begin{aligned} g(x) &= 0.5 + xr, \text{ if } xr \in [-0.4, 0.4] \\ &= 0.1, \text{ if } xr < -0.4 \\ &= 0.9, \text{ if } xr > 0.4 \end{aligned}$$

where r is a constant reflecting the degree to which one wishes to adapt the trial.

OPTIMAL ALLOCATION APPROACH

In the optimal allocation approach, one first defines an optimal allocation proportion based on an optimization problem, which usually establishes an objective function as a metric for ethical considerations and a constraint to preserve power of the test post-trial. This formulation addresses the tradeoff between ethics and efficiency.^[8] In this approach, we must use a parametric model to describe the survival/censoring distributions.

For example,^[9] consider a survival trial where patients randomized to treatment $j = A, B$ have a survival time T_j following an exponential distribution with parameter θ_j . The patients also are subject to an independent right censoring scheme. Let (t_{ij}, δ_{ij}) , $i = 1 \dots n_j$ be a random sample, where t_{ij} is a survival time and $\delta_{ij} = 1$ if the i th patient assigned to treatment j is not censored, and a censoring time and $\delta_{ij} = 0$ if the participant is censored. To test $H_0 : \theta_A = \theta_B$, we use the following statistic:

$$Z = \frac{\hat{\theta}_A - \hat{\theta}_B}{\sqrt{(\hat{\theta}_A^2/r_A + \hat{\theta}_B^2/r_B)}} \sim N(0,1),$$

where $r_j = \sum_{i=1}^{n_j} \delta_{ij}$ is the total deaths from treatment j and $\hat{\theta}_j$ is the maximum likelihood estimator of θ_j . To derive the optimal allocation in this case, assume that $\varepsilon_j = E(\delta_{ij})$ is the same for all $i = 1 \dots n_j$. Then $E(r_j) = n_j E(\delta_{Aj}) = n_j \varepsilon_j$. Minimizing the total expected hazard, we have the following optimization problem,

$$\begin{cases} \min_{n_A} n_A \theta_A^{-1} + (n - n_A) \theta_B^{-1} \\ s.t. \frac{\theta_A^2}{n_A \varepsilon_A} + \frac{\theta_B^2}{(n - n_A) \varepsilon_B} \equiv C_0 \end{cases}$$

where $n = n_A + n_B$ and C_0 is a constant. The solution is

$$\rho = \frac{n_A}{n} = \frac{\sqrt{\theta_A^3 \varepsilon_B}}{\sqrt{\theta_A^3 \varepsilon_B} + \sqrt{\theta_B^3 \varepsilon_A}}$$

One can also minimize other metrics for ethics to obtain other allocation proportions. Other models and different censoring schemes are considered in Ref. [9].

Once an optimal allocation proportion is defined, one can use a randomization procedure, such as the doubly adaptive biased-coin design,^[10] to randomize patients to the treatments such that the final proportion of patients allocated to treatment A would be very close to the target proportion ρ . During this process, the unknown parameters in ρ are replaced with their estimators based on available data. For example, the $(k + 1)$ th patient can be randomized to treatment A with the probability $p = h(N_{Ak}/k, \hat{\rho}_k)$ ^[10] where N_{Ak} is the number of patients randomized to treatment

A from the first k patients, $\hat{\rho}_k$ the estimate of ρ based on the available data from the first k patients and

$$h(x, y) = \begin{cases} 1 & \text{if } x = 0 \\ \frac{y \left(\frac{y}{x} \right)^\gamma}{y \left(\frac{y}{x} \right)^\gamma + (1-y) \left(\frac{1-y}{1-x} \right)^\gamma}, & \text{if } 0 < x < 1, \\ 0 & \text{if } x = 1 \end{cases}$$

where $\gamma \geq 0$ is a constant.

Extensive simulation studies have shown that these approaches are quite powerful.^[4-6,9] In fact, Refs. [5,6,9] report no significant losses in power and great savings in terms of expected treatment failures. Such gains in terms of individual patient experience should be attractive to both physicians and patients, and could influence recruitment and retention of patients positively.^[11] One must be wary, however, of *accrual bias*,^[12] where patients, understanding the principles behind the adaptive survival trial, will elect to enroll in the trial as late as possible, to obtain the best chance of being assigned to the better treatment.

ADAPTIVE SURVIVAL TRIALS IN PRACTICE

Planning an adaptive survival trial can be more time-consuming than planning a traditional survival trial with equal allocation. For the latter, straightforward formulas to compute requisite sample sizes and power of the log-rank test under parametric assumptions on the survival distribution are well known.^[13] Because adaptive randomization induces extra variability in the design, these formulas are not applicable. Careful simulation of power and sample size can be carried out using techniques similar to those found in Refs. [5,9]. In those papers, the authors programmed a priority queue that keeps track of patients' entry and response times. One must assume a staggered entry distribution, a delayed response distribution (that could depend on treatment assignment, response, or entry time) and a survival distribution. Programs are available from the authors upon request.

Many survival trials involve limited recruitment periods and long-term follow-up, so that patient responses are not available during randomization. Adaptive survival trials cannot be used in this context. However, the benefits of adaptive survival trials may make it attractive to extend recruitment over a longer period, so that some information may accrue to benefit future patients.

REFERENCES

1. Kalbfleisch, J.D.; Prentice, R.L. *The Statistical Analysis of Failure Time Data*; John Wiley: New York, 1980.

2. Sposto, R.; Krailo, M.D. Use of unequal allocation in survival trials. *Stat. Med.* **1987**, *7*, 629–637.
3. Hsieh, F.Y. Comparing sample size formulas for trials with unbalanced allocation using the logrank test. *Stat. Med.* **1992**, *11*, 1091–1098.
4. Hallstrom, A.; Brooks, M.M.; Peckova, M. Logrank, play the winner, power and ethics. *Stat. Med.* **1996**, *15*, 2135–2142.
5. Rosenberger, W.F.; Seshaiyer, P. Adaptive survival trials. *J. Biopharm. Stat.* **1997**, *7*, 617–624.
6. Yao, Q.; Wei, L.J. Play the winner for phase I/II clinical trials. *Stat. Med.* **1996**, *15*, 2413–2423.
7. Flehinger, B.J.; Louis, T.A. Sequential treatment allocation in clinical trials. *Biometrika* **1971**, *58*, 419–426.
8. Rosenberger, W.F.; Sverdlov, O. Handling covariates in the design of clinical trials. *Stat. Med.* **2008**, *23*, 404–419.
9. Zhang, L.; Rosenberger, W.F. Response-Adaptive Randomization for clinical trials with survival outcomes: the parametric approach. *J. R. Stat. Soc. C* **2007**, *53*, 153–165.
10. Hu, F.; Zhang, L.X. Asymptotic properties of doubly adaptive biased coin design for multi-treatment clinical trials. *Ann. Stat.* **2004**, *32*, 268–301.
11. Tamura, R.N.; Faries, D.E.; Andersen, J.S.; Heiligenstein, J.H. A case study of an adaptive clinical trial in the treatment of out-patients with depressive disorder. *J. Am. Stat. Assoc.* **1994**, *89*, 768–776.
12. Rosenberger, W.F. New directions in adaptive designs. *Stat. Sci.* **1996**, *11*, 137–149.
13. Lachin, J.M.; Foulkes, M.A. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics* **1986**, *42*, 507–519.

Adverse Event Reporting

Cynthia Pennise

Baxter Healthcare Corporation, New Providence, New Jersey, U.S.A.

Abstract

In order for a chemical intended for use as a medicinal product to be sold for use in humans, research must be carried out to prove that the benefits of the drug outweigh the risks associated with the drug. This entry discusses the importance of adverse event reporting, which is a crucial element in showing and understanding the side effects of a particular drug.

INTRODUCTION

In order for a chemical intended for use as a medicinal product to be sold for use in humans, research must be carried out to prove that the benefits of the drug outweigh the risks associated with the drug. Clinical studies of planned tests of the drug in people are performed following the protocols with predetermined statistical analyses planned for the safety and efficacy of the drug. These studies are conducted in a known number of patients. The results of these studies can be statistically analyzed to show that the results can be reproduced, and that the drug is useful to treat the condition and is safe in the intended population. Adverse events terminology is coded in order to group similar events and to categorize them into similar medical concepts within broad concepts for data retrieval. It is only by knowing the number of patients treated and by analyzing the adverse events reported that the safety profile of a drug can be drawn. By analyzing adverse events that occur within the test drug treatment and comparing it to a placebo-control or a known historical control, one can definitely show what side effects a drug causes.

Once a drug is marketed, the number of patients who are exposed is unknown, as is the indication that the drug was used to treat and the dose used by each patient. Unless further clinical trials are conducted for a different indication or subset of patients, information gathered from the post-marketing safety reporting cannot be utilized to request a regulatory authority to change the labeling. When a new drug is marketed, the treated population is not restricted by protocol, and there is an exponential increase in the number of patients exposed (which can only be estimated from sales volumes). New adverse events are reported that were not seen in the trials conducted by the pharmaceutical company. Postmarketing safety reporting of adverse event information is performed voluntarily. No statistical analyses can be performed on postmarketing data because of the number of patients treated being unknown as well

as the true number of occurrences of each adverse event. New adverse events can be added in the safety section of the package labeling from postmarketed use; however, no incidence for these new adverse events can be listed. This is why great care is taken in the design of clinical protocols so that appropriate claims for the efficacy and safety of a drug can be documented.

BACKGROUND OF THE DRUG APPROVAL PROCESS

The drug approval process is a long one, incorporating up to 15 years of research from the drug's inception at the chemist's bench to marketing. This process includes formulation testing, several stages of preclinical testing, and clinical testing that generally involves phases I through III clinical trials prior to marketing. If the drug survives preclinical testing, phase I studies are conducted in small populations, usually in 20–25 normal volunteers, to determine initial human safety. No formal statistical safety analysis is usually performed on these small studies because of the fact that the number of patients, or n , is small. If there appears to be no clinical safety issues after reviewing the results of the blood and urine analyses, preliminary efficacy studies are performed and are called phase II trials. These trials study the drug in the intended patient population with the target disease to be treated. The trials are small, usually 50–100 patients, and are usually single-blind studies, generally with no control arm. Although safety is looked at from a clinical perspective, it is usually not analyzed statistically. The main focus of these studies is to see if the drug appears to show some efficacy to justify additional clinical research. If efficacy appears to show a positive trend, phase III trials utilizing a placebo-control, a historical control, or a comparative control are conducted. Depending on the disease and the intended duration of human treatment (acute versus chronic use), these trials

may range from 100–200 patients treated for several days to several months, or up to several hundreds of patients to 1000 or more patients treated up to several years. Two separate controlled clinical trials are usually required for regulatory approval in order to ensure that the results can be statistically reproduced. Recently, the Food and Drug Administration (FDA) has allowed some drugs intended for critical use to be evaluated for approval utilizing a single, controlled trial. Although there are many aspects regarding the evaluation of clinical trials, the focus of this entry is on the adverse event reporting for the safety data analysis using the Medical Dictionary for Regulatory Activities (MedDRA) coding system.

CODING DICTIONARIES

The Need to Code

In order to prepare an analysis of safety data and provide any meaningful results of the safety that was seen in the trial, the safety terms, as they were reported, must be combined into similar terminology, or coded, to allow the true occurrence of similar events to be analyzed. For example, if the reported terminology for adverse events from a study includes stomach ache, stomachache, stomach pain, and pain in stomach, if analyzed as reported, there is one occurrence of each of the four terms. However, when a coding dictionary is utilized, all of the terms would be coded (depending on the dictionary used), for example, to “abdominal pain, upper” with four occurrences of this one adverse event. Combining terminology via coding dictionaries allows the safety results to be analyzed statistically and more truly reflect what is occurring in the patient population being treated with the drug product.

Common Dictionaries Utilized

There are many adverse event coding dictionaries that have been created over the years throughout the world. Some of the common ones include MedDRA,^[1] Costart,^[2] Whoart,^[3] and ICD-9.^[4] Many pharmaceutical companies have also created their own versions of adverse event coding dictionaries, an example of which is HARTS.^[5] One major problem with the existence of multiple coding dictionaries is that in addition to each pharmaceutical company choosing a dictionary to use, the regulatory authorities around the world are also using different dictionaries. This leads to many problems in companies who have to use multiple dictionaries for foreign versus domestic regulatory submissions and in regulatory authorities who have to recode the data into their dictionary of choice. This problem was the one that added to the need for the creation of the International Conference for Harmonization (ICH).^[6] The goal of this group, which represents the regulatory bodies and research-based industry in the European Union, Japan,

and the United States was to find a way to harmonize all the types of regulatory submissions for global acceptance and approval. One of the resulting guidances that arose from the formation of this group was the need for a common coding dictionary to be utilized around the world by countries in these regions. The resulting dictionary was called MedDRA. This dictionary was designed to provide a standardized coding dictionary to be utilized by industry and regulators around the world. The dictionary was made up of terminology already contained in many of the existing dictionaries, although its structure allowed greater specificity for the coding of terminology.

Medical Dictionary for Regulatory Activities

Instead of being divided into the common body systems used by previous dictionaries, MedDRA categorizes terminology into 26 System Organ Classes (SOCs). The division of terminology into these SOCs allows for the coding of medical history as well as adverse events arising from the use of both drugs and biological products. The dictionary is not currently used to code Medical Device Incidents or to code medications. The dictionary is hierarchical in that it provides vertical linking of terminology. The five-level hierarchy of the dictionary with the number of terms included in each level for version 4.1 is shown in Fig. 1.

The LLT is the place where the reported term is located. Both the British and American English spellings are used at the LLT level. Each LLT is linked to only one PT. Preferred terms represent a single medical concept and can consist of several LLTs. The dictionary utilizes the British spelling of words at the PT level and higher. Terminology utilized in statistical tables is usually represented by the verbatim term as reported and the coded term, in this case, at the PT level.

The dictionary allows for multiaxial coding, which means that certain terms at the PT level that can logically be coded to more than one SOC can be represented under more than one SOC. However, within each SOC, each PT

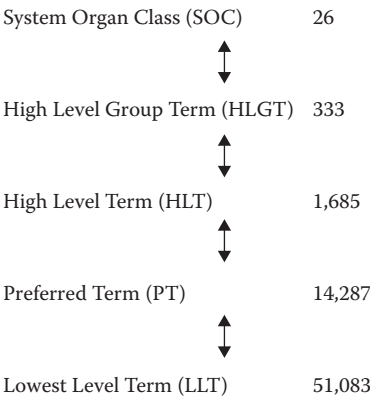


Fig. 1 MedDRA hierarchy

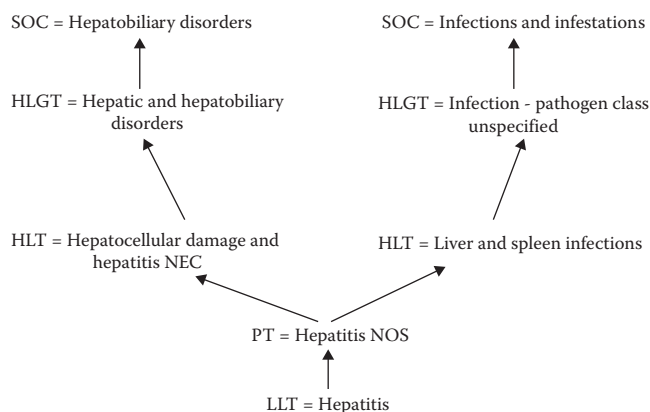


Fig. 2 Multi-axial pathway for one PT to two different SOC

is represented by only one HLT and one HLGT. Each PT has a primary pathway that must be used for the initial analysis by companies. Secondary analyses can be performed utilizing these secondary pathways. An example of the multi-axial pathway for a term is shown in Fig. 2.

The SOC are not mutually exclusive. The MedDRA dictionary also contains Special Search Categories (SSC) that contain clinical concepts that cross SOC. These SSCs allow linkage of PTs that are not necessarily equivalent or related in the hierarchy of MedDRA. These SSCs generally group PTs that are relevant to a medical issue, usually a disease or a syndrome. Preferred terms can exist in multiple SSCs. Examples of SSCs are listed in Fig. 3.

Dictionary Versions

The MedDRA dictionary is constantly updated for coding. Updates are needed because of the companies submitting requests for additions of terminology to the company managing the dictionary. Requests for changes in mapping of terms, as well as updates for dictionary consistency, also result in the need to update the dictionary. The current MedDRA Maintenance and Support Services Organization is TRW, Inc. The original MedDRA dictionary was updated four times per year; current updates are made twice per year. Major updates are indicated by an increase in the first number of the version (e.g., 4.0), with minor updates of a version indicated by an increase of the decimal (e.g., 4.1). Various pharmaceutical companies do coding of the clinical studies in many ways. One method is to initiate coding of a study in the current version of MedDRA that is being utilized at the company and to remain with the same version throughout the study. At the end of the study, the entire study can be recoded in the current version of MedDRA. An alternative way to code a clinical study is to continually update the versions as they are published. Whichever method of coding is utilized, it is important to document the version of MedDRA utilized for the coding of each study.

ANAPHYLAXIS
ARREST (CARDIAC)
BLOOD DYSCRASIS/BONE MARROW DEPRESSION
CARDIAC ISCHAEMIA
HAEMORRHAGE
HYPERSENSITIVITY REACTIONS
OEDEMA
PAIN
PREMALIGNANT LESIONS
SECONDARY IMMUNOCOMPROMISED STATE
THROMBOSIS
UPPER GI BLEEDING/PERFORATION
VASCULITIS

Fig. 3 Special Search Categories

STATISTICAL ANALYSES FOR ADVERSE EVENT DATA

Individual Study Report

Each individual study must be statistically analyzed to show its contribution toward the approval of a drug for both efficacy and safety.^[7] Depending on the size and the objectives of the study, safety data may be analyzed with descriptive statistics (frequency, percentage, mean, and standard deviation) for most cases, or with rigorous statistical methods that provide *p*-values for comparison between different treatment groups. Note that safety data usually consist of clinical laboratory data, ECG, physical examination, and adverse events reporting. Adverse event data are one of the most, if not the most, important piece of safety data that are closely watched and reviewed by regulatory agencies.

When safety data are analyzed, it is a common practice to include all patients who received at least one dose of the study drug(s) so that the investigation of the safety profile may be executed to the greatest possible extent. As far as adverse events are concerned, in a basic summary statistics, they are presented in tables showing the incidence of each adverse event in the comparative treatment groups. Incidence is represented by showing the number of occurrences of each adverse event with the percentage of patients who experienced the adverse event. When a patient experiences a particular adverse event on more than one occasion, the patient is only counted once for that adverse event. The same is true if a patient experiences a particular adverse event that recurs and that varies in intensity (mild, moderate, or severe). The patient is counted once for that adverse event, the occurrence with the greatest severity.

Different adverse events are counted separately for the same patient. In addition to the presentation of occurrence of adverse events by treatment group, an overall cumulative tabulation of each adverse event that occurred in all treatment groups is presented. Another critical piece of the summary is the count and the percent of patients in each treatment group who experienced any adverse events. This provides a simple and condensed overall safety picture of the study drug in relation to the control/comparator drug. Presentation of the adverse events in the summary table can be performed through sorting by SOC, or by the frequency (or percent) of each SOC in decreasing order of the active treatment group. The latter presentation shows a clear trend of the occurrences of the SOCs for the test drug. Within each SOC in the summary table is the listing of the PTs that are primarily coded to that specific SOC, in alphabetical order or by decreasing order of occurrence and percent. When a large-scale multicenter trial is involved, in addition to a pooled summary of all the study centers, it may be relevant to summarize the adverse events by each individual study site. An example of a typical side-by-side presentation of the safety data is shown in Table 1.

According to the ICH guidelines, the presentation of individual patient events by drug, adverse event, severity, and relationship to treatment can be presented in Table 2.

It may be pertinent to statistically compare some SOCs that showed the highest occurrences of adverse events between and among treatment groups for a large-scale clinical trial where these SOCs are of clinical relevance for the drug product. Because there are 26 SOCs, and some SOCs may have zero count of occurrences, it would not provide any clinically meaningful information to compare every SOC. The following statistical methods are commonly used to compare adverse events within each SOC between treatment groups:

1. Chi-square test for a 2×2 summary table: this is employed in cases of comparing two treatment groups, and the number of patients with adverse events (AE) and no AE for each treatment group should be equal to or exceed 5 (preferably over 10).
2. Fisher-exact test for a 2×2 summary table: this is employed in cases of comparing two treatment groups regardless of the magnitude of the number of patients with and without AE.
3. Mantel–Haenszel test for $2 \times 2 \times q$ summary tables: this is employed in cases of comparing two treatment groups while at the same time controlling another explanatory variable (such as study sites in a multicenter study, ages grouped by at least two classes, gender, and severity of underlying disease) as the stratum, where q refers to the number of levels for the stratum. This test is statistically valid when the combined row counts are sufficiently large (for example, greater than 30).
4. Extended Mantel–Haenszel test for $2 \times r \times q$ summary tables: this is employed in cases of comparing more than two treatment groups ($r > 2$) while at the same time controlling another explanatory variable as the stratum as in Case 3 above. The case when $q = 1$ may represent a single-center clinical trial with more than two treatment groups. A realistic and conservative guideline for the modeling validity is to choose one or more cutpoints for the column for each stratum, say j ($j \geq 1$). Add the 1st through j th columns together and $(j + 1)$ th through the last columns. Then the sample size is adequate if all the cells in each new reduced 2×2 table of the q tables are 5 or more. Note that the cutpoints across all strata are not necessary to be the same.^[8]

If there is more than one treatment period in a study per protocol, the adverse events that occurred during each period should be presented separately. Also, depending on the patient population and the disease being studied, subset analyses of adverse event data can be performed. These can be performed by specific demographics, or SSC, for example.

The incidence of adverse events is evaluated for clinical significance in addition to the statistical analysis that was performed. A high incidence of a particular adverse event can often be explained by the concurrent disease state of the patients being treated. Other times there are known side effects of a particular drug class. One of the primary reasons

Table 1 Typical Side-by-Side Presentation of Safety Data

		Incidence of adverse events: number (%)		
System organ class	MedDRA PT	Treatment A (n = 25)	Treatment B (n = 25)	Total (n = 50)
Cardiac disorders	Asystole	1 (4%)	5 (20%)	6 (12%)
	Myocardial infarction	0 (0%)	2 (8%)	2 (4%)
Hepatobiliary disorders	Hepatitis NOS	5 (20%)	2 (8%)	7 (14%)
	Jaundice	10 (40%)	0 (0%)	10 (20%)
Vascular disorders	Hypertension	1 (4%)	6 (24%)	7 (14%)
	Hypotension	2 (8%)	12 (48%)	14 (28%)
All	All	13 (52%)	20 (80%)	33 (66%)

Table 2 Incidence of Adverse Events (Primary Organ Class and Preferred Term) by Severity and Relationship^a

		Treatment group: AAAA											
Primary system organ class	Preferred term	Severity = mild			Severity = moderate			Severity = severe			Overall		
		Not Related	Not related	Total	Not Related	Not related	Total	Not Related	Not related	Total	Not Related	Not related	Total
Gastrointestinal disorders	Nausea	5	4	8	1	2	3	0	1	1	6	6	11
		N11 ^b	N11	N11	N19	N20	N19		N20	N20	N11	N11	N11
		N12	N16	N12		N21	N20				N12	N16	N12
		N13	N17	N13			N21				N13	N17	N13
		N14	N18	N14							N14	N18	N14
		N15		N15							N15	N20	N15
				N16							N19	N21	N19
				N17									N16
				N18									N17
													N18
													N20
													N21
	Vomiting NOS	3	2	4	2	3	3	0	0	0	4	5	6
		N22	N23	N22	N24	N24	N24				N22	N23	N22
		N23	N25	N23	N26	N26	N26				N23	N24	N23
		N24		N24		N27	N27				N24	N25	N24
				N25							N26	N26	N26
												N27	N25
												N27	

Related includes adverse events that are possibly, probably, and highly probably related to the study drug.

Not related includes adverse events that are not related or remotely related to the study drug.

^aPatients are counted only once in each cell.

^bNxx represents patient ID number.

that new drugs are developed is to be able to market a drug that is effective in treating the condition and is either the first of its kind or is better than the products already available. One of the ways this superiority can be documented is by using a comparative drug as a positive control in a study and being able to show a better safety profile, or that the innovator is faster in onset or lasts longer in duration.

Integrated Summary of Safety

When filing a New Drug Application (NDA), in addition to reporting a summary of each individual study where the safety and the efficacy are analyzed for that particular study, a combined safety analysis, the Integrated Summary of Safety (ISS), must be written.^[9] The safety data from each controlled clinical trial are combined into a single data set and are analyzed to see if the safety results from the individual studies are changed when the data from all studies are analyzed together. In addition to comparing the treatment groups to one another as a whole, if there is sufficient power, subset analyses can be performed. These subset analyses depend on the patient population treated. Data can be separated and analyzed by age, sex,

race, or disease state, for example. The results of these subset analyses can provide valuable information that will eventually be included in the indications, dosing, precautions, or adverse event sections of the package insert or the Core Data Sheet. Depending on the results of the subset analysis, some adverse events may be separately described to list their incidence, for example, in pediatric or geriatric populations compared with the overall incidence for that adverse event in the total population that was treated. Safety analyses of the different disease states have provided safety results that show that the drug cannot be administered to certain patient populations (e.g., renal-impaired or hepatic-impaired patients). Within the ISS, it is sometimes useful to determine whether there is a trend when several dosing levels of the test drug are involved.

Adverse Events as Secondary Endpoint

In addition to proving that a drug is effective, for certain disease states, there are certain known side effects, or sequelae, that occur in the usual course of the treatment. If these sequelae are known in advance of protocol design, the study can be designed to show that the study drug

is “safer” than other products available by resulting in a lower incidence of the known side effects. The protocol should be written to include detailed statistical approach to this secondary analysis.

CONCLUSION

The safety of a drug product is important for its approval for marketing. Careful planning of the clinical trials for study design is important. Analysis of adverse events can be performed as a basic necessity to show the safety of a drug product. In some cases, adverse event data can also be analyzed as a safety endpoint in order to provide data for labeling claims. This is one of the many reasons to be sure that your statisticians are well informed about the disease and the patients being treated when the protocol is written to ensure sufficient power for labeling claims.

REFERENCES

1. MSSO, *Medical Dictionary for Regulatory Activities Terminology, MedDRA Introductory Guide v4.1, MSSO-DI-6003-4.1.0*; TRW Inc., 2001.
2. Department of Health and Human Services. *Costart: Coding Symbols for Thesaurus of Adverse Reaction Terms, Fifth Edition, 1995, USA Food and Drug Administration, PB95-269023, FDA/CDER-95/24*; U.S. Department of Commerce, National Technical Information Service: Springfield, VA, 1995.
3. World Health Organization Collaborating Centre for International Drug Monitoring, WHO Adverse Reaction Terminology (WHOART), 1998, Uppsala, Sweden, <http://www.who.int/home-page/> (accessed in December 2001).
4. *ICD-9-CM: International Classification of Diseases*, 9th Revision; Clinical Modification, *PB2001-500083*, U.S. Department of Commerce, National Technical Information Service: Springfield, VA, 2001.
5. *Hoeschst Adverse Reaction Terminology System (HARTS)*; 1992, Aventis Pharma.
6. *ICH: International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use*; <http://www.ifpma.org/ich2.html> (accessed in December 2001).
7. Food and Drug Administration Code of Federal Regulations, 21 CFR 314.50 (d) (5) (ii).
8. Stokes, M.; Davis, C.; Koch, G.; Eds. *Categorical Data Analysis Using The SAS System*, 2nd Ed.; SAS Institute Inc.: Cary, NC; 2000.
9. Food and Drug Administration Code of Federal Regulations, 21 CFR 314.50 (d) (5) (vi).

Alpha Spending Function

Jihao Zhou

Department of Biostatistics, Allergan, Inc., Irvine, California, U.S.A.

Glen Andrews

Department of Biostatistics, Pfizer, Inc., San Diego, California, U.S.A.

Abstract

In this entry, we review the background for the alpha spending function, give an overview of classic group sequential methods, and focus on the alpha spending functions.

Keywords: Alpha spending function; Clinical trial; Data monitoring; Group sequential analysis; Interim analysis; Type I error.

INTRODUCTION

Clinical trials have become the major standard over the past three decades for evaluating new therapies and medical interventions. In the majority of studies, accruing data are monitored and analyzed at intervals by the sponsor or an independent data monitoring committee. Any repeated significance tests will inflate the type I error probability and lead to false claims of a treatment benefit even when one does not exist. Group sequential methods were introduced in the late 1970s in order to control the type I error; however, they require predetermination of the number of interim analyses and equal increments of information between looks. A more flexible approach was introduced by Lan and DeMets,^[1] which eliminated the constraints of the group sequential methods by introducing a spending function for alpha. In this entry, we review the background for the alpha spending function, give an overview of classic group sequential methods, and focus on the alpha spending functions.

BACKGROUND

Interim Analysis

In the ICH Guideline E9 (Statistical Principles for Clinical Trials), an interim analysis is defined as any analysis intended to compare treatment arms with respect to efficacy or safety at any time prior to the formal completion of a trial.^[2] Generally, the use of an interim analysis can aid in decision making on: (a) early stopping of a clinical trial for overwhelming efficacy, (b) early stopping for futility, and (c) modification of certain aspects of the trial design mid-course such as, sample size re-estimation in order to have sufficient power. In conducting interim analyses, we carry out repeated significance tests, which

inflate the overall significance level, and thus increase the chances of a false positive claim.^[3] Several group sequential methods^[4–7] have been developed for use to preserve the overall significance level alpha.

STANDARD GROUP SEQUENTIAL METHODS

Although not originally described as a group sequential procedure, a method described by Haybittle^[6] in 1971 and later supported by Peto et al.^[7] has often been used. For each analysis, Haybittle proposed a very conservative analysis critical value for all interim analyses (e.g., ± 3.0 or ± 3.5) such that the type I error increases almost negligibly. This implies for all practical purposes, one could use the usual 5% critical value of ± 1.96 at the last scheduled analysis should the trial continue that far. More formally:

$$\begin{aligned} C(k, \alpha) &= C_{H-P}, & k = 1, \dots, K-1; \\ C(k, \alpha) &= Z_{1-\alpha/2}, & k = K. \end{aligned}$$

In the above equations, the subscript H-P indicates Haybittle–Peto, K is the total number of analyses, and α is the overall significance level for a study. Obviously, a disadvantage is the somewhat ad hoc choice for the interim testing levels. Another disadvantage is the uncertainty regarding the overall probability of type I error.

Pocock^[4] more formally introduced the term “group sequential” with his work in 1977, modifying proposals by McPherson and Armitage. His method uses a constant boundary in order to make a decision about whether or not to reject the null hypothesis and to control the overall significance level alpha. Here:

$$C(k, \alpha) = C_p(K, \alpha), \quad k = 1, \dots, K.$$

His proposal uses up the overall significant level equally at each time point that an analysis is undertaken.

O'Brien and Fleming^[5] proposed an alternative method in 1979, which uses a monotone-decreasing boundary, against which the test statistic is compared to decide whether or not to reject the null hypothesis. In this case:

$$C(k, \alpha) = C_{O-F}(K, \alpha) \sqrt{(K/k)}, \quad k = 1, \dots, K.$$

It starts with a very stringent nominal significant level, making it very difficult to stop the trial early on, but ending with a nominal significant level very close to the overall significance level.

Fig. 1 shows the comparison of the three group sequential methods using $\alpha = 0.05$ with a two-sided standard normal test statistic Z and four interim analyses. The three methods have very different stopping boundaries. The O'Brien–Fleming method is unlikely to lead to stopping in the early stages. Later on, however this procedure leads to a great chance of stopping prior to the end of the study compared to the other two methods. Both the Peto and O'Brien–Fleming methods avoid the awkward situation of accepting the null hypothesis when the statistic at the end of the trial is much larger than the conventional critical value (i.e., 1.96 for a conventional two-sided significance level). It is not obvious what criteria should be used to select from among the various rules; however, of the three methods described, the O'Brien–Fleming more closely mimics the behavior of many data monitoring committees, who are usually conservative at the beginning stage of the trial when they will require a greater level of evidence to persuade them to stop a study early.

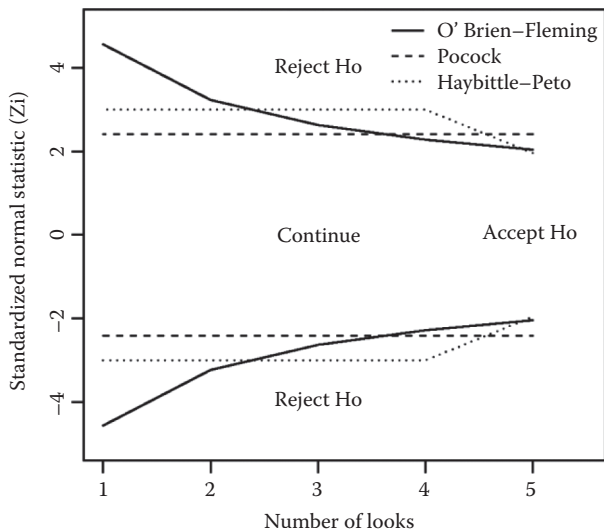


Fig. 1 Comparison of three group sequential methods using $\alpha = 0.05$ with a two-sided standard normal test statistic Z

ALPHA SPENDING FUNCTIONS

Origin of Alpha Spending Function

The practical problems in implementing the standard sequential methods are that first they require a prior specification of the number of interim analyses needed for a clinical trial, and secondly they require that the interim analyses must be equally spaced so that the amount of information which accrues for each interim analysis is uniform. These may actually cause difficulty and inconvenience in conducting a clinical trial. For example, the DSMB (Data and Safety Monitoring Board), which is usually responsible for conducting the interim analysis, may not be able to have all its members available at a specific time point when additional data has been accrued. Lan and DeMets proposed the alpha spending function approach to overcome those problems by extending the group sequential concept to a very flexible method that controls the overall alpha level, while allowing for the number and exact timing of the interim analyses to remain unspecified *a priori*.^[1] If a spending function is selected in advance of the study, starting according to a desired boundary, then interim analyses can be conducted as needed. Even determining the frequency of analyses based on emerging trends negligibly affects the false-positive error rate, especially for conservative boundaries.^[28]

The spending function of Lan and DeMets was developed in the context of designs that do not allow early stopping with rejection of the alternative hypothesis. Pampallona, Tsiatis, and Kim^[14] proposed an adaptation, whereby the total probability of the type II error (β) can be incorporated after successive looks. For this reason the alpha spending function has also been referred to as the use function,^[8] spending function,^[11,12] and the error spending function.^[13] Here we will use the term alpha spending function.

Concept of Alpha Spending Function

Slud and Wei^[15] discussed the idea of distributing the overall significance level alpha (α) to each interim analysis by using the increments of α , say, $\alpha_1, \dots, \alpha_K$, where $\alpha_1 + \dots + \alpha_K = \alpha$, and K is predefined as the total number of analyses in the study. However, no particular function was suggested or proposed to distribute the α , although they recommend that the sequence $\alpha_1, \dots, \alpha_K$ be increasing. Lan and DeMets^[1] first suggested the alpha spending function concept. The function $\alpha^*(t)$ describes the rate at which the total significance level alpha is to be spent or used as a continuous function of information fraction t . In the context of a clinical trial, t indicates how much of the trial has been completed in terms of the accumulated information, and thus indicates how much of the allowable type I error should be allocated.^[10] The $\alpha^*(t)$ generates the cumulative

significance level “used up” after the current analysis and relates the sequence of nominal significance levels through a Brownian motion process.^[1]

Derivation of the Original Spending Function

The original boundary suggested by Lan and DeMets was proposed because of its similarity to the O’Brien–Fleming boundary. The derivation of that spending function is presented here to help the reader understand the process. This is described for a one-sided test, but it is easy to generalize to a two-sided test by replacing α by $\alpha/2$.

Let $\{B(t), 0 \leq t \leq 1\}$ be a standard Brownian motion process, and let τ be the first exit time across the boundary. Here let’s consider a horizontal boundary $b(t) = z_{\alpha/2}$, so if we assign $\alpha^*(t) = \Pr(\tau \leq t)$, then it follows that:

1. $\alpha^*(0) = 0$;
2. For $0 < t \leq 1$,

$$\begin{aligned}\alpha^*(t) &= \Pr(\tau \leq t) \\ &= \Pr(|B(t)| > b(t)) \\ &= \Pr(|B(t)| > Z_{\alpha/2}) \quad (\text{because } b(t) = Z_{\alpha/2}) \\ &= \Pr(|B(t)/\sqrt{t}| > Z_{\alpha/2}/\sqrt{t}). \\ &= \Pr(|Z| > Z_{\alpha/2}/\sqrt{t}) \quad (\text{because } B(t) \sim N(0, t), \\ &\quad \text{then } B(t)/\sqrt{t} \sim N(0, 1)) \\ &= 2\left(1 - \Phi\left(Z_{\alpha/2}/\sqrt{t}\right)\right) \\ &\quad (\text{where } \Phi \text{ is the standard normal function}) \\ &= 2 - 2\Phi\left(Z_{\alpha/2}/\sqrt{t}\right).\end{aligned}$$

So, $\alpha^*(t)$ is strictly increasing in t ; and $\alpha^*(1) = \alpha$, which is equal to the prespecified significance level.

Assume that $B(t)$ be observed only at discrete time points $t_i, i = 1, \dots, K$, with $0 < t_1 < \dots < t_K = 1$. Assign an accumulated boundary crossing probability $\alpha^*(t_1)$ to the point t_1 , by defining b_1 to satisfy

$$\begin{aligned}\Pr\{B(t_1) > b_1\} \\ &= \Pr\{\tau \in [0, t_1]\} \\ &= \alpha^*(t_1)\end{aligned}$$

For $i = 2, \dots, K$,

$$\begin{aligned}\Pr\{B(t_j) < b_j, j = 1, \dots, i-1; B(t_i) > b_i\} \\ &= \Pr\{\tau \in (t_{i-1}, t_i)\} \\ &= \alpha^*(t_i) - \alpha^*(t_{i-1})\end{aligned}$$

Once we allocate the alpha to the spending function, then the next step is that we need to calculate the nominal boundary $(b_1, \dots, b_K)$ against which the observed test statistic can be compared to conduct the statistical hypothesis testing.

For $i = 1$,

$$\begin{aligned}\Pr\{B(t_1) < b_1\} \\ &= 1 - \Pr\{B(t_1) > b_1\} \\ &= 1 - \alpha^*(t_1) \\ &= 1 - \left(2 - 2\Phi\left(Z_{\alpha/2}/\sqrt{t_1}\right)\right) \\ &\quad (\text{plug in the function we derived above}) \\ &= 2\Phi\left(Z_{\alpha/2}/\sqrt{t_1}\right) - 1.\end{aligned}$$

From standard Brownian motion, we know that $B(t_1)/\sqrt{t_1} \sim N(0, 1)$, then

$$b_1 = \sqrt{t_1} \Phi^{-1}\left\{2\Phi\left(Z_{\alpha/2}/\sqrt{t_1}\right) - 1\right\}.$$

For $i = 2, \dots, K$, evaluation of b_i can be done by using numerical integration,^[3,11,13] because from the Brownian motion process, we know that the $B(t_1), \dots, B(t_K)$ jointly follow a multivariate normal distribution.^[16,17] The algorithm for the numerical integration can be found in references.^[3,11,13]

Note the above calculation process of the boundary depends only on the functional form $\alpha^*(t)$ and information $t_1, \dots, t_i$, therefore we do not need to know K (the total number of analyses) or $t_{i+1}, \dots, t_K$ (the future time points). Now we are in a position to use this derived function and the corresponding evaluated boundary values to conduct an interim analysis without the necessity of either knowing K (the total number of analyses) or the equal spacing of information t_i ; indeed the points $t_1, \dots, t_i$ need not be equally spaced.

TYPES OF ERROR SPENDING FUNCTIONS

The above error spending function was the first one derived from the standard Brownian motion process as we have shown, and is only one of many functions that have been proposed or developed since then. Other approaches defined in the original paper by Lan and DeMets^[1] were:

1. Pocock-type

$$\alpha_2^*(t) = \alpha \ln[1 + (e-1)t], \quad 0 < t \leq 1$$

It can approximate the standard group sequential method Pocock boundary.

2. Uniform function

$$\alpha_3^*(t) = t, \quad 0 < t \leq 1$$

It spends alpha uniformly over the information fraction t .

Additional functions introduced subsequent to this include:

3. ρ [Kim and DeMets^[8]]

$$\alpha_4^*(t) = \alpha t^\rho, \quad 0 < t \leq 1$$

Special mention was made of $\rho = 2$ and $\rho = 3/2$, which reduce the likelihood of the embarrassing situation of having a significant result at the k th analysis and a non-significant results at the $(k + 1)$ th analysis.

4. Segmental function [Li and Geller^[18]]

$$\alpha_5^*(t) = 0.8\alpha t | \{0 \leq t \leq 0.75\} \\ + (1.6\alpha t - 0.6\alpha) | \{0.75 < t \leq 1\}$$

They noted that by using the strictly convex functions, such as that by Kim and DeMets,^[8] the analysis times will be severely restricted, especially near the end of a trial, where frequent monitoring is often desirable.^[19] They proposed the above spending function to have the property of the uniform use of 60% of the alpha during the accrual of 75% of the information and 40% of alpha reserved for the last 25% information.

5. One-parameter Gamma Family [Hwang, Shi, DeCan^[9]]

$$\alpha_6^*(t) = \alpha(1 - e^{-\gamma t}) / (1 - e^{-\gamma}), \quad \gamma \neq 0$$

$$0 < t \leq 1,$$

$$\alpha_6^*(t) = \alpha t, \quad \gamma = 0, \quad 0 < t \leq 1.$$

It includes members similar in shape to the spending functions of Pocock ($\gamma = 1$) and O'Brien–Fleming ($\gamma = -4$). Negative values of γ yield convex spending functions that increase in conservatism as γ decreases, while positive values of γ produce convex spending functions that increase in aggressiveness as γ increases.

6. Other types include the unifying family of Kittleson and Emerson^[20] for the mean statistic and the Wang–Tsatis delta family,^[21] which includes standard group sequential methods such as Pocock's and O'Brien–Fleming's methods as special cases.

Generally speaking, the Pocock-type and the O'Brien–Fleming-type are two extreme cases. The former is an aggressive method in terms of spending alpha heavily at the early stage of the trial, while the latter is a conservative method because it spends a very little at the beginning but reserves the majority of alpha for later stage of the trial.

CHOICE OF ALPHA SPENDING FUNCTION $\alpha^*(t)$

Because there are many types of alpha-spending functions, we provide some guidance below:

The function should be an increasing function $\alpha^*(t)$ with $\alpha^*(0) = 0$, and $\alpha^*(1) = \alpha$ to construct the boundary

and to enable the control of the overall significance level for a study.^[11] A strictly convex $\alpha^*(t)$ is recommended in general.^[8,10,15,19] Using steadily increasing sequence of $\alpha^*(t_k) - \alpha^*(t_{k-1})$ so that the resulting boundary values $b_k / \sqrt{t_k}$ decrease. Using steadily decreasing boundary values will be unlikely to result in an embarrassing possibility of non-significant result at t_{k+1} when a decision to stop the trial with significant results at t_k , was deferred.^[15] A situation like this may result in the DSMB deciding to ignore the boundary being crossed and is called “overruling.”^[22]

When short-term effects are of interest in a clinical trial with planned interim analyses, then early stopping would be desirable, so $\alpha^*(t)$ with a smaller expected stopping time (e.g., Pocock-type α spending function) has a good chance of an early rejection of the null hypothesis if the alternative hypothesis is true.^[10] On the other hand, when long-term effects are of interest, use more convex $\alpha^*(t)$ that would generate boundaries similar to O'Brien and Fleming boundaries. By doing so, we will use very little alpha at the early interims and save the majority of the alpha for near the end of the trial.

INFORMATION FRACTION t

Once we've made a choice about the alpha spending function, then the next important step is to decide how to define operationally the information fraction t before the trial starts. Obviously, even if we use the same alpha spending function, different definitions of t would lead to different boundary values. Here we give some guidance. In general, the information fraction t should be defined as the amount of information provided by the accumulated subjects up to the current analysis, divided by the amount of total information planned or expected for the study.^[10] For comparing means or proportions, t is approximated by the observed sample size divided by the expected maximum sample size (n/N). For survival data, t may be approximated by the number of observed events (for example, deaths in a clinical oncology trial) divided by the total expected number of events (e/E); t may also be approximated by the time up to current analysis, divided by the total planned duration for the study (t/T). The former is called maximum information design and the latter maximum duration trial. For repeated measurements data, a definition of t in the context of the linear mixed model was provided by Lee and DeMets.^[23]

COMPUTING SOFTWARE

Because computation for using the alpha spending function normally requires complex numerical integration, several software packages have been developed and are available for sequential analyses. The software EaSt (version 3.1 or higher) is commercially available from Cytel Corporation, Cambridge, Massachusetts. LANDEM

is a Fortran-based program publicly available from the Department of Biostatistics, the University of Wisconsin, Madison, Wisconsin, and its detailed usage with examples has been published.^[11] R is free software widely used in the statistical circle around the world and it has a function called the `ldBands()` function which can carry out alpha spending function calculations.^[24] PEST is commercially available from the MPS unit, the University of Reading, United Kingdom, featuring the triangular sequential test.^[25] The S + Seq Add-on package for S-plus is commercially available from the Insightful Corporation, Seattle, Washington.^[13]

APPLICATION OF THE ALPHA SPENDING FUNCTION

Design of an Interim Trial Using Alpha Spending Function

The alpha spending function has been applied in clinical trials with planned interim analyses to aid decision making on early stopping in a number of different scenarios:

1. Stop early for efficacy only, i.e., to reject H_0 on a one-sided test.
2. Stop early for safety only, i.e., to reject H_0 on a one-sided test.
3. Stop early for futility only, i.e., to reject H_1 on a one-sided test.
4. Stop early for efficacy or safety, i.e., to reject H_0 on a two-sided test.
5. Stop early for efficacy or futility, i.e., to reject H_0 or reject H_1 on a one-sided test.
6. Stop early for efficacy, safety, or futility, i.e., to reject H_0 or reject H_1 on a two-sided test.
7. Stop early for futility only, i.e., to reject H_1 on a two-sided test.

In the above, scenarios 1 and 2 are essentially the same in terms of the methodology except that the statistic in 1 is related to efficacy, e.g., the reduction in viral load in an AIDs trial, while the statistic in 2 is related to safety, such as detecting an increase in the number of adverse events for a new compound compared to an approved drug in the same class.

Scenario 4 is a two-sided test, which is actually often used in clinical trial practice.

In scenario 3, this use is usually referred to as beta spending function instead of alpha spending function because the lower futility stopping boundary (critical values) is chosen to maintain the nominal type II error rate,^[14] i.e., the spending function is chosen to control the rate of spending type-II error. Monitoring for futility is usually combined with a test for benefit as in scenario 5, and this approach is becoming more popular in clinical trial practice. Scenario 6 is usually called a two-sided test with an inner wedge,^[13] which is used to accept H_0 . Scenario 7 is sometimes called “only an inner wedge,” which is rarely used in practice.^[13] The choice between a one-sided and a two-sided test is based on the parameter of interest. In some cases, a one-sided alternative is appropriate because departures from H_0 in the other direction are either implausible or impossible. In other cases, the parameter of interest may lie on either side of H_0 , but the alternative is chosen in the one direction in which departures from H_0 is of clinical interest. Obviously, a one-sided test tends to reduce the expected sample size.^[13]

Here we present examples for the commonly used scenarios 4 and 5. The other scenarios are relatively easy extensions of 4 and 5, so they are not illustrated here.

For scenario 4, let's design a hypothetical phase III, randomized, double-blind, active-controlled, clinical oncology trial to compare the efficacy of the new investigational drug monotherapy against that of the standard therapy to treat a specific cancer in adult patients. Three interim analyses are to be planned for this trial. The primary efficacy endpoint for this trial is progression free survival (PFS); and from literature, median PFS for standard therapy is 4 months and we expect that the median PFS for the new drug will be improved by 50% to 6 months. Patients are to be randomly assigned to a new investigational drug and to a standard therapy in a 2:1 fashion to allow more patients to receive the promising new therapy. We will use the O'Brien–Fleming-type spending function to design this trial. Table 1 shows the details we need given a two-sided, $\alpha = 0.05$ and 90% power, illustrated graphically in Fig. 2. The computation was carried out using the software EaSt 3.1.

For scenario 5, we will use the example presented by Pampallona, Tsiatis, and Kim,^[14] which was a typical, randomized, two-arm trial comparing a control to an

Table 1 Design of an Interim Trial Using O'Brien–Fleming-Type Spending Function

Look	Number of events	Information fraction (t)	Hazard ratio	P -value	Lower boundary (Z)	Upper boundary (Z)
1	70	0.25	0.34	0.0001	−4.33	4.33
2	140	0.50	0.59	0.003	−2.96	2.96
3	211	0.75	0.71	0.018	−2.36	2.36
4	281	1.00	0.78	0.044	−2.01	2.01

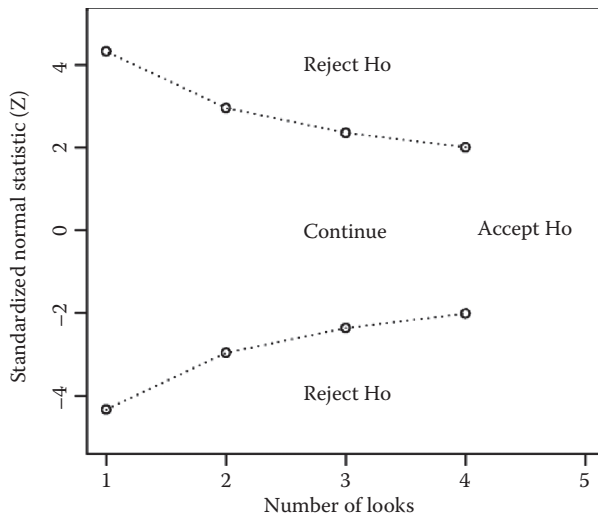


Fig. 2 Design of an interim analysis using O'Brien-Fleming-type function with $\alpha = 0.05$, 90% power and a two-sided test

experimental arm on the basis of a one-sided normally distributed statistic (Z). It's assumed that under the alternative hypothesis, the effect size (mean difference divided by standard deviation) equals 0.25. Three interim analyses (a total of 4 looks) were planned for this study. Patients were to be randomly assigned to the two groups in a 1:1 fashion. The O'Brien-Fleming-type spending function was used to design this trial. Table 2 shows the details we need given a one-sided $\alpha = 0.05$ and 90% power—this again is illustrated graphically in Fig. 3. As shown here, the use of a futility boundary allows us to stop the study early when the treatment is not working, which leads to a smaller expected sample size under the null hypothesis.

Analysis of an Interim Trial Using the Alpha Spending Function

For illustrative purposes, we will describe the approach to the Beta-Blocker Heart Attack Trial (BHAT). BHAT was a randomized double-blind, multicenter trial to evaluate the effect of propranolol in reducing mortality in patients who had recently suffered a myocardial infarction.^[26,27] The O'Brien and Fleming-type α spending function was used. The statistical analysis of time to event data was carried out using a log-rank statistic. For this study, the calculated

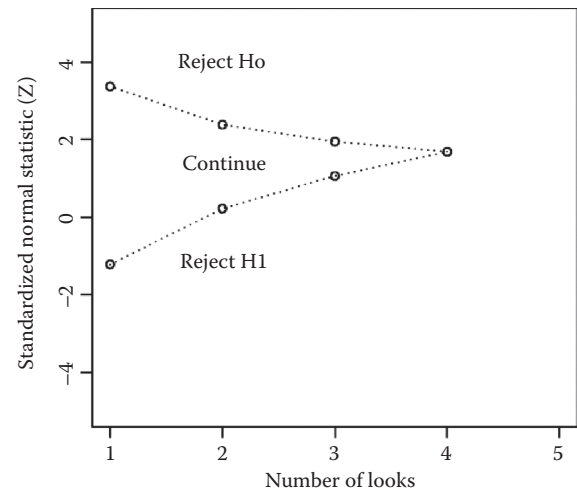


Fig. 3 Design of an interim analysis for simultaneous efficacy and futility using O'Brien-Fleming-type function with $\alpha = 0.05$, 90% power and a one-sided test

log-rank statistic for each interim analysis is shown in the last column of both Tables 3 and 4. Two scenarios of designs are shown here.

1. As a maximum information trial: if the study was planned with total amount of information $D = 400$ deaths, the information fraction at each interim analysis would be the number of deaths up to that point divided by D (d/D). After the information fraction is obtained, then a boundary value (Z) can be computed using the O'Brien and Fleming-type α spending function (see Table 3).
2. As a maximum duration trial: if the study was planned with duration 48 months, the information fraction for each analysis would be the time up to current analysis divided by the $T(t/T)$. After the information fraction is determined, then a boundary value (Z) can be computed using the O'Brien-Fleming-type α spending function (see Table 4).

Fig. 4 shows the interim analysis for the BHAT trial using the O'Brien-Fleming-type function. Obviously, the two boundaries are very different, although the stopping point for this example happens to coincide at the sixth interim analysis point, where the observed log-rank statistic

Table 2 Design of an Interim Trial Using O'Brien-Fleming-Type Spending Function for Simultaneous Efficacy/Futility Purpose

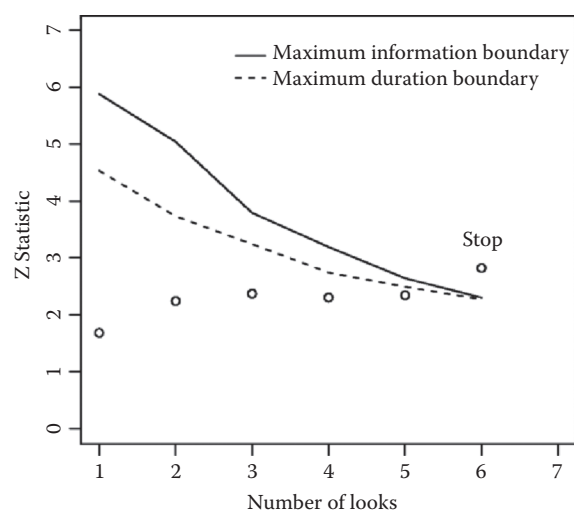
Look	Number of cases	Information fraction (t)	Lower stopping boundary	Upper stopping boundary
1	150	0.25	-1.220	3.372
2	300	0.50	0.220	2.384
3	450	0.75	1.063	1.947
4	600	1.00	1.686	1.686

Table 3 Interim Analysis for Beta-Blocker Heart Attack Trial (BHAT) as a Maximum Information Trial ($D = 400$)

Planned look	Number of deaths (d)	Information fraction (d/D)	Boundary value (Z)	Log-rank statistic
1	56	0.14	5.88	1.68
2	77	0.19	5.04	2.24
3	126	0.32	3.79	2.37
4	177	0.44	3.19	2.30
5	247	0.62	2.64	2.34
6	318	0.80	2.30	2.82

Table 4 Interim Analysis for Beta-Blocker Heart Attack Trial (BHAT) as a Maximum Duration Trial ($T = 48$)

Planned look	Calendar time in months (t)	Information fraction (t/T)	Boundary value (Z)	Log-rank statistic
1	11	0.23	4.53	1.68
2	16	0.33	3.73	2.24
3	21	0.43	3.24	2.37
4	28	0.58	2.74	2.30
5	34	0.70	2.49	2.34
6	40	0.83	2.27	2.82

**Fig. 4** Interim analysis for the Beta-Blocker Heart Attack Trial (BHAT) trial using the O'Brien–Fleming-type function

($Z = 2.82$) exceeded both boundaries at this interim analysis time point. However, we should bear in mind that generally that may not be the case. Therefore, it is critical to pre-specify how to define the information fraction t as well as the predetermined spending functional form.

Some Issues in Using the Alpha-Spending Functions

With the application of alpha-spending functions into clinical trials, questions involving interim analysis have been arising, such as how to deal with overruling,^[22] changing monitoring frequency,^[28] and multiple primary endpoints in clinical trials.^[29] For instance, because the

statistical decision rule is only one of the factors for consideration in making a final decision about whether to stop or continue a trial based on a certain time point interim analysis, the DSMB may decide to continue a trial even when the statistical boundary has been crossed. This is called overruling. In the case of overruling, we may buy back the previously spent probability to be respent or redistributed at future looks.^[22] Sometimes, the frequency has to be increased due to emerging trends. In this case, the probability of type I error will no longer be preserved exactly, but simulation results show that the changing frequency of interim analyses has a negligible inflation of the type I error^[28] for the common spending functions such as O'Brien–Fleming type. Recently, a global alpha spending function has been proposed for a trial with multiple primary endpoints.^[29] Asymmetric boundaries are also common and allow for stopping rules for both efficacy and safety, and to allow for early stopping for futility.^[14] Other issues include how to provide valid parameter estimates and confidence intervals at the end of a study, allowing for any “shrinkage” to the estimates that may have occurred because of the interim analyses. The detailed expositions of these issues are beyond the scope of this entry. Interested readers may refer to the respective reference for details.

CONCLUSION

The alpha spending function is a practical tool to control the rate at which the overall type I error alpha is to be spent or used as a continuous function of information fraction. It is very flexible. The total number of interim analyses need

not be prespecified, and the information fraction need not be equally spaced. The users can also specify their own spending function. Implementation is made easy because of the availability of computing software.

ACKNOWLEDGMENTS

We would like to thank Parke-Davis Pharmaceutical Research Division of Warner-Lambert Company for initial support of my work on this field. We want to express appreciation to Eric Yan, Nai-tee Ting, and Randy Allred for valuable suggestions in the preparation of this manuscript. Our continued motivation in this field would not be possible without collaboration with our Pfizer clinical colleagues on applications of the alpha-spending functions in clinical trials practice. R software was used for making the in-text graphs.

ARTICLES OF FURTHER INTEREST

Clinical Trials; Data Monitoring Committees (DMC); Group Sequential Methods; Group Sequential Tests and Variance Heterogeneity in Clinical Trials; Interim Analysis; International Conference on Harmonization (ICH); Statistical Principles for Clinical Trials; Statistical Significance.

REFERENCES

- Lan, K.K.G.; DeMets, D.L. Discrete sequential boundaries for clinical trials. *Biometrika* **1983**, *70* (3), 659–663.
- ICH E9. *Statistical Principles for Clinical Trials*. <http://www.ich.org> (accessed 08/05/2005), 1998.
- Armitage, P.; McPherson, C.K.; Rowe, B.C. Repeated significance tests on accumulating data. *J. R. Stat. Soc. A* **1975**, *132*, 235–244.
- Pocock, S.J. Group sequential methods in the design and analysis of clinical trials. *Biometrika* **1977**, *64* (2), 191–199.
- O'Brien, P.C.; Fleming, T.R. A multiple testing procedure for clinical trials. *Biometrics* **1979**, *35*, 549–556.
- Haybittle, J.L. Repeated assessment of results in clinical trials of cancer treatment. *Br. J. Radiol.* **1971**, *44*, 793–797.
- Peto, R.; Pike, M.C.; Armitage, P.; Design and analysis of randomized clinical trials requiring prolonged observations of each patient. I. Introduction and design. *Br. J. Cancer* **1976**, *34*, 585–612.
- Kim, K.; DeMets, D.L. Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* **1987**, *74*, 149–154.
- Hwang, I.K.; Shih, W.J.; De Cani, J.S. Group sequential designs using a family of type I error probability spending function. *Stat. Med.* **1990**, *9*, 1439–1445.
- DeMets, D.L.; Lan, K.K.G. Interim analysis: the alpha spending function approach. *Stat. Med.* **1994**, *13*, 1341–1352.
- Reboussin, D.M.; DeMets, D.L.; Kim, K.-M.; Lan, K.K. Computations for group sequential boundaries using the Lan–DeMets spending function method. *Control. Clin. Trials* **2000**, *21*, 190–207.
- Whitehead, J. *The Design and Analysis of Sequential Clinical Trials*, 2nd Ed.; Wiley: Chichester, 1997.
- Jennison, C.; Turnbull, B.W. *Group Sequential Methods with Applications to Clinical Trials*, Chapman and Hall: Boca Raton, FL, 2000.
- Pampallona, S.; Tsiatis, A.A.; Kim, K. *Spending functions for type I and type II error probabilities of group sequential test Technical Report*. Department of Biostatistics, Harvard School of Public Health: Boston, MA, 1995.
- Slud, E.; Wei, L.J. Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J. Am. Stat. Assoc.* **1982**, *77* (12), 862–868.
- Ross, S.M. *Stochastic Processes*, Wiley: New York, 1996.
- Tuckwell, H.C. *Elementary Applications of Probability Theory*, Chapman and Hall: London, 1995.
- Li, Z.; Geller, N.L. On the choice of times for data analysis in group sequential clinical trials. *Biometrics* **1991**, *47*, 745–750.
- Betensky, R.A. Construction of a continuous stopping boundary from an alpha spending function. *Biometrics* **1998**, *54*(3), 1061–1071.
- Kittelson, J.M.; Emerson, S.S. A unifying family of group sequential test designs. *Biometrics* **1999**, *55*(3), 874–882.
- Wang, S.K.; Tsiatis, A.A. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **1987**, *43*(1), 193–199.
- Lan, K.K.G.; Lachin, J.M.; Bautista, O. Over-ruling a sequential boundary—a stopping rule versus a guideline. *Stat. Med.* **2003**, *22*, 3347–3355.
- Lee, J.W.; DeMets, D.L. Sequential comparison of change with repeated measurement data. *J. Am. Stat. Assoc.* **1991**, *86*, 757–762.
- <http://lib.stat.cmu.edu/S/Harrell/help/Hmisc/html/lldBands.html> (accessed 09/29/2006).
- PEST (Planning and Evaluation of Sequential Trials). *The MPS Research Unit*; The University of Reading: Reading, 1995.
- Beta-Blocker Heart Attack Trial Research Group. A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. *JAMA* **1982**, *247*, 1707–1714.
- DeMets, D.L.; Hardy, R.; Friedman, L.M.; Lan, K.K.G. Statistical aspects of early termination in the Beta-Blocker Heart Attack Trial. *Control. Clin. Trials* **1984**, *5*, 362–372.
- Lan, K.K.G.; DeMets, D.L. Changing frequency of interim analyses in sequential monitoring. *Biometrics* **1989**, *45*, 1017–1020.
- Kosorok, M.R.; Shi, Y.; DeMets, D.L. Design and analysis of group sequential clinical trials with multiple primary endpoints. *Biometrics* **2004**, *60*, 134–145.
- Pocock, S.J.; Hughes, M.D. Practical problems in interim analyses, with particular regard to estimation. *Control. Clin. Trials* **1989**, *10*, 209–221.

Ames Test

Wherly P. Hoffman

Department of Biometrics, Elan Pharmaceuticals Inc., South San Francisco, California, U.S.A.

Michael L. Garriott

Eli Lilly and Company, Indianapolis, Indiana, U.S.A.

Abstract

Two- and threefold rules and various parametric and nonparametric methods have been applied to the Ames test data. This article will review these analysis methods and discuss the design, statistical analyses, and interpretation of results of the Ames test.

Keywords: In vitro assay; In vivo assay; Safety profile.

INTRODUCTION

Genetic toxicology tests are among the early studies conducted to assess the safety profile of a compound. A battery of tests, each assessing a different genetic endpoint, is typically performed to thoroughly evaluate a given compound because no single test is capable of detecting all mutagens. The tests are designed to determine whether the compound can interact with DNA and lead to the production of gene mutations or chromosomal breakage. Some tests can also detect compounds that interact with components of the mitotic spindle apparatus. The detection of mutagens is important because they may also be either teratogens or carcinogens. Of a battery of short-term genetic assays, the *Salmonella typhimurium*/microsome test developed by Ames et al.^[1,2] is the most commonly used genotoxicity test. Two- and threefold rules and various parametric and non-parametric methods have been applied to the Ames test data. This article will review these analysis methods and discuss the design, statistical analyses, and interpretation of results of the Ames test. “Background” includes information pertaining to the relevance of the Ames test to the toxicology safety profile. The “Design” section describes the basic design of the test. The section “Analysis Methods for Ames Test Data” provides a review of some analysis methods employed for the evaluation of the mutagenicity of a compound. In the “Discussion” section, we discuss the methods reviewed and some important issues in the evaluation of mutagenicity based on Ames test results, and in the last section we draw concluding remarks to summarize this article.

Notes: Adapted from Hoffman, W.P. In *Design and Analysis of Animal Studies in Pharmaceutical Development*; Chow, S.C.; Liu, J.P.; Eds.; Marcel Dekker: New York, 1998, 357–372.

BACKGROUND

This section presents a brief summary of why the Ames test was developed and how this assay is conducted. The procedure for conducting the Ames test is described without too much detail. This section is not intended to provide enough information to conduct the assay; it is intended to provide readers with a general understanding of how the responses are obtained and thus leads naturally into the following sections on the design and statistical methods employed for analysis of the test data.

WHY THE AMES TEST

When developing a compound for commercial purposes, one has to establish its safety profile. *In vivo* animal and *in vitro* cell culture tests are designed to serve this purpose. In general, the testing includes acute toxicity tests in rodents, genetic toxicity tests, developmental toxicity tests, general animal toxicity tests from 30 days up to 1 year in duration, and lifetime carcinogenicity tests in rodents. Of these general tests, which are needed to demonstrate the safety of a compound and to ensure its registration worldwide, the most costly and time-consuming test is the carcinogenesis bioassay in rodents. Assays for genotoxicity were initially developed in an attempt to predict in a shorter time frame the eventual outcome of the bioassay. The basic premise revolved around the somatic theory of carcinogenesis, which suggested that mutational events were causative factors in the development of cancers. Additionally, there was evidence that mutagenic events were associated with embryonic mortality, birth defects, and genetic disease. Therefore, the use of genotoxicity tests seemed to have broad

application in the overall safety assessment. Although bacterial mutation tests had been used since the 1950s, they were not efficient in the detection of compounds acting by various mutagenic mechanisms. Efforts by Bruce Ames and his colleagues at the University of California, Berkeley, to develop a more efficient bacterial screening system culminated in 1971 with the test that now bears his name. Despite the test's utility, it was realized that bacteria are not mammalian cells and that the organization of the genetic material was different so that other short-term mutagenicity tests would be required to ensure the detection of the majority of rodent carcinogens.^[3] Although over 200 mutagenicity tests have been developed, few have been well validated, and current regulatory requirements for product registration include only a battery consisting of two to four tests of the seven most thoroughly validated standard tests.^[4-6] Of these required tests, the Ames test is the most commonly used and for this reason alone merits discussion.

THE AMES TEST

The Ames bacterial mutation test is an in vitro assay in that it uses bacterial cells grown on agar contained in petri plates. Typically, the *Salmonella typhimurium* bacterial tester strains TA1535, TA1537, TA98, and TA100 and an *Escherichia coli* tester strain WP2uvrA are used in the Ames test. Although WP2uvrA was added to the bacterial test system after the Ames test was developed, it is handled the same as the four *S. typhimurium* strains in all aspects of the conduct and analysis of the experiment. Therefore, in this article, Ames test will refer to the assay of the four *S. typhimurium* strains and the *E. coli* strain. For each strain of the bacteria, a mixture of the proper amounts of three ingredients is first prepared in a tube. The three ingredients are (i) from 10–100 million bacterial cells from a particular strain; (ii) a compound, which can be a test article, negative control, or positive control; and (iii) minimal agar, which includes trace amounts of histidine, tryptophan, and biotin and may or may not contain S9 mix. See Fig. 1. The mixture is then poured into a plate containing a layer of solidified minimal agar for the growth of revertant colonies. Some compounds are not directly mutagenic in vivo but are metabolized to a compound that is. Therefore, all in vitro genotoxicity tests, including the Ames test, are conducted so that cells are exposed to the test article both in the presence and in the absence of an exogenous metabolic activation system. This metabolic activation system consists of a postmitochondrial supernatant (S9) of homogenized liver tissue from rats. This cell-free extract contains a variety of drug-metabolizing enzymes. The S9 is supplemented with cofactors, salts, and a buffer system collectively known as S9 mix.

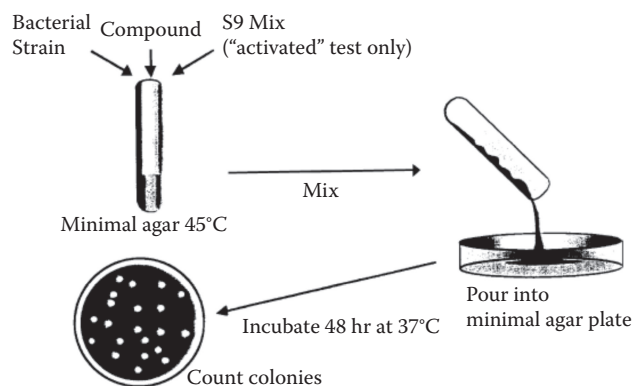


Fig. 1 Description of the conduct of the Ames test

Histidine and tryptophan are amino acids essential for the growth of the tester strains of *S. typhimurium* and *E. coli*, respectively. The Ames tester strains contain mutations in the genes that allow normal bacteria (prototrophs) to manufacture their own histidine or tryptophan. The tester strains are dependent upon the inclusion of the amino acids into the minimal agar to support their growth. These mutant bacteria are called auxotrophs. The Ames test is often referred to as a “reverse” mutation assay because what is measured is the reverse mutation rate from histidine/tryptophan dependency (auxotrophy) to histidine/tryptophan independence (prototrophy). Bacterial cells that have undergone the reverse mutational event are called “revertants.” Trace amounts of histidine and tryptophan are provided at the start of the experiment to sustain the bacterial cells at the beginning and thus to allow the possible expression of reverse mutation events later. After a 48-hour incubation period, the revertant cells will form visible colonies, which are counted. Because reverse mutation may occur without treatment with a mutagen, a negative control is always included in the assay to account for the spontaneous reverse mutation rate. If a compound is not a mutagen, then one would not expect a large increase in the number of revertant colonies compared to the number of spontaneous revertant colonies in the negative control. Therefore, the number of revertant colonies is the response evaluated to determine the mutagenic effect of a test article. A significant increase in revertant colonies on the compound-treated plates is an indication of a positive response for bacterial mutation.

DESIGN

The Ames test procedure is rather strictly defined for regulatory purposes.^[7-9] The assay is typically conducted both with and without metabolic activation for each of the five bacterial strains. For each bacterial strain, five

or more concentration levels of a compound and proper concentration levels for the negative and positive controls are determined by scientists. A positive control is required for the validity of the assay but is not included in the evaluation of the mutagenicity of a compound. A negative control is needed to account for spontaneous reverse mutation in the evaluation of the mutagenicity of the test article. Historical control data are useful for further evaluation of positive effects associated with a compound. Each dose level is tested in triplicate, and each of the triplicate plates is prepared independently. Therefore, if five concentrations are selected for each bacterial strain, then a total of 50 sets of triplicate data from the compound (five concentrations, two types of activation, and five strains of bacteria) and 10 sets of triplicate data from each of the negative control and positive control are the results of the assay.

Because the purpose of the Ames test is to evaluate the mutagenic effects of a compound, a suitable selection of the concentration levels of a compound for testing is essential. Logarithmically spaced concentration levels are determined by scientists in an attempt to capture an increasing trend in the response. However, if the concentration of the compound is too high, then one would expect a decrease in the number of revertant colonies as a result of toxicity. For determination of the concentration levels for the Ames test, a compound is initially assayed over a wide range of concentration levels up to 5,000 µg/plate for each bacterial strain. Based on the preliminary results, five to seven concentration levels are then selected for a definitive assessment. The highest concentration may be very close to the toxic level. In the absence of toxicity, a top concentration of 5,000 µg/plate is used in the definitive assessment of mutation induction.

ANALYSIS METHODS FOR AMES TEST DATA

Proper analysis methods should account for various sources of variability in the response. Therefore, one needs to identify different sources of variability in the assay data before considering the analysis. Based on the variability and some assumptions, we discuss various statistical and nonstatistical methods for the analysis of revertant colonies obtained from the Ames test. Edler^[10] and Mahon et al.^[11] provide a good overview of statistical methods for the Ames *Salmonella* test. For a more recent review of statistical methods, see Lin.^[12] Here we will begin the discussion with some basic information on data in the next section. Then selected methods will be discussed to bring about different approaches for the analysis of Ames test data. The methods discussed in the subsequent sections are modified two- and threefold rules, non-parametric methods, and parametric methods.

UNDERSTANDING DATA

What is the distribution of the number of revertants from the Ames test? Before suggesting an answer, let us first understand the data and try to identify various sources of variability in it. For each replicate plate, a separate minimal agar mixture of the selected strain of bacterial cells, the test article, and a solution of histidine, tryptophan, and biotin is made in a tube. Therefore, the numbers of revertants from the triplicate plates are independent. Because auxotrophs need to grow in a histidine/tryptophan-sufficient environment, trace amounts of these amino acids are supplied at the beginning of the test. For each strain of bacteria, variability in the number of revertants can be a result of the number of bacterial cells that are plated at the beginning of the test, the amount of histidine/tryptophan provided initially, the concentration of the test article, the contents of the minimal agar, the amount of S9 mix, the duration of the incubation time, or other factors. In a laboratory, attempts are made to control all factors so that changes in the revertant counts would reflect only the effect of the test article through the concentration levels.

To demonstrate that a compound is not mutagenic, one has to test it at sufficiently high concentrations to rule out doubts that the bacterial cells may not have been challenged enough. Pilot studies are usually conducted first to select proper concentration levels for each strain of bacteria. The final selection often includes concentration levels that are very close to the toxicity threshold of the bacterial cells. Consequently, one may observe a decrease in the number of revertants when some of the concentration levels of a compound are higher than the toxicity threshold of the bacterial cells.

Because the number of bacterial cells plated initially is large (about 10^7) and the probability of a cell undergoing a reverse mutation from auxotrophy to prototrophy is small, the resulting revertant counts of such rare events may have a Poisson distribution. Does the number of revertants X follow a Poisson distribution $p(X = x, \lambda)$ where

$$p(X = x, \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \text{for } x = 0, 1, 2, \dots \quad (1)$$

and λ is the mean rate of reverse mutation? One can examine this either graphically or by performing a statistical test. To get a good assessment of the distribution assumption, both approaches require more data than are available from one assay in general. When a sufficient amount of data is available, one can check this Poisson assumption by plotting the sample variance against the sample mean on a logarithmic scale. A linear relationship with slope near one is an indication of a Poisson distribution. Or one can perform Fisher's dispersion test^[13] for a sample of n counts, $x_1, x_2, \dots, x_n$, which compares the test statistic

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\bar{x}} \quad (2)$$

to a chi-square distribution with $n - 1$ degrees of freedom, where $\bar{x}$ is the average of the n revertant counts. Stead et al.^[14] reported no overdispersion in their experience, whereas Vollmar^[15] and Margolin et al.^[16] observed overdispersion. Because each laboratory may have its unique features in conducting the assay and controlling the assay environment, an assessment of overdispersion should be performed based on its own historical data, and reevaluation should be performed periodically.

If the rate of reverse mutation λ from replicate to replicate is not a constant, then the excess of variability in replicates over the mean would suggest that the revertant counts are a random sample from a mixture of Poisson distributions. For ease of further references in later sections, we will follow the parameterization of Margolin et al.^[16] If λ is a random variable with a gamma distribution, then the revertant counts have a negative binomial distribution

$$p(X = x | \mu, c) = (x + c^{-1} - 1) \left(\frac{\mu}{\mu + c^{-1}} \right)^x \times \left(\frac{c^{-1}}{\mu + c^{-1}} \right)^{c^{-1}} \quad (3)$$

for $x = 0, 1, 2, \dots$

where $c \geq 0$ and $\mu > 0$. The mean of this distribution is μ , and the variance is $\mu(1 + c\mu)$. The extra variability in the Poisson counts is reflected by the dispersion parameter c . Large values of c indicate inadequacy of modeling without accounting for the extra variability in the sampling.

MODIFIED TWO- AND THREEFOLD RULES

The two- and threefold rules declare a given test article to have induced a positive response if a concentration-related increase in the number of revertant colonies is at least twice the control count for strains TA98, TA100, and WP2uvrA or three times the control count for strains TA1535 and TA1537.^[17] The comparison is based on the average number of revertants from the replicates for the control and each concentration level. The treated averages are compared to the control average. In the event that there is a decrease in the number of revertants owing to toxicity, the evaluation criteria are applied only to the results at the non-toxic level.

NON-PARAMETRIC METHODS

Vollmar^[15] has examined the results from the European Collaborative Ames-Test Study 1977–1978, and the plots

of log-range versus log-mean indicated that the number of revertants did not follow a Poisson distribution. Overdispersion was evident between laboratories as well as within each laboratory. It was recommended that non-parametric methods be applied to Ames test results. Two rank-based tests were considered: the Jonckheere test^[18] for monotonic concentration effects and the Kruskal-Wallis test^[19] for other effects, including a downturn at the toxic levels. Given 713 Ames test results, the Jonckheere test was superior to the Kruskal-Wallis test and was recommended as the routine statistical method for qualitative evaluations. Vollmar pointed out that, in general, three to five replicate samples were necessary for achieving a statistical significance level of 0.05.

Wahrendorf et al.^[20] have proposed a non-parametric approach to the analysis of Ames test data. Assume that there are K concentration levels with n_k replicates at the k th level and the total number of replicates is N . In their approach, the concentration levels were arranged in K increasing levels, with $K = 1$ for the control. For a given j , $1 \leq j \leq K$, the revertant counts from all K concentration levels are assigned to one of the two categories: 1 is the “below” category for levels 1 to j ; 2 is the “above” category for levels $j + 1$ to K . The revertant counts are ranked from low to high. If there are no concentration-related effects, then the sum of the ranks, L_j , from the “below” category should be the product of the midrank, $(N + 1)/2$, and the number of observations in the “below” category. Define L as the sum of the “below” categories corresponding to j for $1 \leq j \leq K - 1$. If a compound does induce a positive trend in the number of revertants, then the ranks in the “above” category would be relatively higher than those in the “below” category. Therefore, positive trends are associated with smaller values of L , and a test statistic for positive trends is L . Wahrendorf et al. have provided critical values for L based on Monte Carlo simulations for standard designs in their paper. Incorporating the expectation, $E_0(L)$, and variance, $\text{Var}_0(L)$, of L under the null hypothesis of no concentration-related effects, the resulting standardized test T for the alternative hypothesis of an increasing trend is a one-sided test:

$$T = \frac{E_0(L) - L}{\sqrt{\text{Var}_0(L)}} \quad (4)$$

where $E_0(L)$ and $\text{Var}_0(L)$ are defined^[20] by the total sample N and s_j , the number of observations in the “below” category for a given j as follows:

$$E_0(L) = \frac{N+1}{2} \sum_{j=1}^{K-1} s_j \quad (5)$$

$$\text{Var}_0(L) = \frac{N+1}{2} \left[\sum_{j=1}^{K-1} s_j (N - s_j) + 2 \sum_{j=1}^{K-2} \sum_{k=j+1}^{K-1} s_j (N - s_k) \right] \quad (6)$$

The T statistic is approximated by a normal distribution. In addition to the test proposed in the foregoing, Wahrendorf et al. also provided a calculation for the probability $\hat{q}_j$ that a revertant count from the “below” category is smaller than that from the “above” category for a given j . Under the null hypothesis of no concentration-related effects, $\hat{q}_j$ is expected to have a value of 0.5. This probability gives insight to the change of trends in the revertant counts and is helpful for further understanding of the type and strength of the trends.

PARAMETRIC MODELING

Parametric methods include the following: (i) regression on transformed data, assumed to be normally distributed;^[21] (ii) non-linear regression on data assumed to be Poisson distributed, either with or without extra variation, using a full likelihood or quasi-likelihood approach; and (iii) biologically based models incorporating extra-Poisson variation.^[15,22,23]

LINEAR REGRESSION ON LOGARITHMICALLY TRANSFORMED REVERTANT COUNTS

Chu et al.^[21] at the In Vitro Program of the National Cancer Institute (NCI)/National Toxicology Program (NTP) evaluated the statistical methods for Ames test data based on results from 2,362 tests performed in four laboratories on 17 test articles. In defining the screening criteria for adequate data, they defined “toxic dose level” as “any dose level which was greater than the dose eliciting the highest average response and which had every response less than the lowest single response in the highest average response dose level.” Evidence of toxicity is a decrease in the number of revertants after reaching the toxic level. Therefore, when a downturn is present in the revertant counts and the peak mean revertant count occurs at concentration level c_i , then the toxic levels are those concentration levels, c_j , for $j > i$, that have all revertant counts smaller than the lowest revertant count at concentration level c_i .

Once the toxic concentration levels are identified, evaluation of the mutagenic effects will be based on data obtained in the non-toxic concentration range. One approach to the statistical evaluation of the mutagenic effects is regressing the logarithmically transformed revertant count on the logarithmically transformed concentration level.^[21] Because the control concentration level is 0 and the log transformation of 0 does not exist, the recommendation was to increase all concentration levels by 1. The validity of this regression approach lies in the appropriateness of the normality assumption resulting from the log transformation on the revertant counts. Mutagenicity is established by a statistically significant positive slope of the regression line. Chu et al.

required a test to have at least five concentration levels, including a negative control, and each with a minimum of two replicates. When concentration levels are approximately equally spaced, instead of adding 1 to all concentration levels to allow for the log transformation on the 0 concentration level, an alternative was recommended by Margolin et al.^[24] The concentration level of the negative control is selected so that all concentration levels are equally spaced on the logarithmic scale.

NON-LINEAR REGRESSION ON REVERTANT COUNTS

Although rare events are often modeled by Poisson models, and it seems reasonable to consider the small number of revertants in about 10 million bacterial cells a Poisson random variable, there are statistical concerns in the assumption of simple Poisson distribution. Evidence of extra-Poisson variation was observed by some, but not all. Because replicates are required at each concentration level, accumulating historical control data for examination of the adequacy of the Poisson assumption is essential. Under the Poisson assumption, the revertant counts can be modeled by a generalized linear model^[25] with a Poisson distribution and logarithmic link. Positive trends in the treated groups can be tested in a sequential fashion by a trend test^[26] if identifying the concentration level associated with no observable effect is of interest. For laboratories that exhibit extra-Poisson variation in the historical control data, the generalized linear model can be adjusted using a quasi-likelihood method for the extra-Poisson variation. The extra-Poisson variation is accounted for by including an extra factor in the variance of the model. The extra factor can be estimated by the square root of Pearson's chi-square divided by the degrees of freedom. This analysis can be carried out using PROC GENMOD in SAS.^[27]

If the mean rate for reverse mutation, λ , in Eq. 1 is a random variable with gamma distribution, then the posterior distribution of the revertant counts is a negative binomial distribution as in Eq. 3. Thus, the extra-Poisson variation can be incorporated in the modeling and evaluation of the mutagenicity of a compound using a full likelihood approach.

BIOLOGICALLY BASED MODELS

Non-monotonicity observed as a decrease in the revertant counts is not uncommon in practice. Bacterial cells subject to toxicity are inhibited from full expression of mutagenicity. Once a definition of a toxic concentration level is established, then one can eliminate the toxic concentration levels and model the mutagenic effects of a compound using the remaining data. An alternative to this two-step modeling approach is to model both mutagenicity and toxicity of a

compound simultaneously. Margolin et al.^[16] have developed biologically based models for modeling the mutagenicity and toxicity while accounting for the extra variation in the Poisson distribution in the revertant counts from the Ames test. The choice of their models depends on the number of generations the bacterial cells have for their histidine/tryptophan supply, the number of generations the compound is mutagenic, and the number of generations the compound is toxic. Two basic assumptions of the biologically based models are that mutagenic and toxic effects are independent, and durations of bacterial cells, being mutagenic and toxic, are not affected by the concentration of the compound. The mean, μ , of the negative binomial distribution, is related to the number of bacterial cells exposed to the test article, N_0 , and the probability of resulting a revertant colony when exposed to the compound at concentration level d , $P(d)$, as

$$\mu = N_0 P(d) \quad (7)$$

We will discuss the two simplest cases, when the bacterial cells have enough histidine/tryptophan supply for just one generation and the compound is mutagenic for one generation. The first case is when the compound is toxic for one generation. The second case is when it is toxic for infinitely many generations. If a compound is toxic only for one generation, then $P(d)$ is

$$P(d) = (1 - e^{-(\alpha + \beta d)})e^{-\gamma d} \quad (8)$$

where α and β are positive and γ is non-negative. However, if it has long-lasting toxicity, then $P(d)$ is

$$P(d) = (1 - e^{-(\alpha + \beta d)})[2 - e^{-\gamma d}]_+ \quad (9)$$

where $[y]_+ = \max(y, 0)$. The spontaneous reverse mutation rate, mutagenicity, and toxicity are modeled through parameters α , β , and γ . The evaluation of the Ames data is based on Models 8 and 9 in combination with 7 and the negative binomial distribution in Eq. 3 through parameters α , β , γ , and c . If there is no toxicity, then $\gamma = 0$ and Models 8 and 9 are identical. Mutagenicity of a compound is evaluated by testing the null hypothesis of $H_0: \beta = 0$ against the alternative of $H_a: \beta > 0$. Margolin et al.^[16] have approximated the distribution of the ratio of the maximum likelihood estimate of β , $\hat{\beta}$, and its standard error, $SE(\hat{\beta})$, by a standard normal distribution. The $SE(\hat{\beta})$ was obtained based on the Fisher information matrix evaluated at the maximum likelihood estimates of α , β , γ , and c . However, further research by Margolin et al.^[22] has identified problems with the two models. Inflated type I error rates were reported for both models. Their proposal to contain the type I error rates was performing a pretest on $\gamma = 0$ to identify the presence of toxicity. The pretest of $H_0: \gamma = 0$ against $H_a: \gamma > 0$ is a likelihood ratio test:

$$\lambda_\gamma = 2(L(\alpha, \beta, \gamma; c) - L(\alpha, \beta, 0; c)) \quad (10)$$

where $L(\alpha, \beta, \gamma; c)$ is the maximum of the log-likelihood of the data. If the toxicity is not supported by the data, then the mutagenicity parameter β will be tested by the likelihood ratio test in^[10] constrained on $\gamma = 0$:

$$\lambda_\beta = 2(L(\alpha, \beta, 0; c) - L(\alpha, 0, 0; c)) \quad (11)$$

Otherwise, the mutagenicity parameter β will be tested by

$$\lambda_\beta = 2(L(\alpha, \beta, \gamma; c) - L(\epsilon, 0, \gamma; c)) \quad (12)$$

Histidine/tryptophan is necessary for the growth of auxotrophic cells. In the biologically based models described in the foregoing, amino acid diffusion was not considered and each cell will have a local supply of the essential amino acid. However, if the amino acids do diffuse through the plate agar, then the revertant counts should be a function of the amount of histidine/tryptophan as well. Krewski et al.^[23] have proposed generalizing the biologically based models by Margolin et al.^[16] to allow for diffusion of the amino acids in an agar plate.

DISCUSSION

The original and modified two- or threefold rules are simple to use and have been compared to other methods in Refs [21,28]. Chu et al.^[21] have reported favorable performance of the modified twofold rule based on results from 2,362 tests performed in four laboratories on 17 test articles. Definitions of the false positive and false negative are as follows:

- False positive = negative tests determined by consensus of microbiologists that is concluded positive using the decision rule.
- False negative = positive tests determined by consensus of microbiologists that is concluded negative using the decision rule.

Among the methods that gave higher false-positive conclusions than false-negative conclusions are the modified twofold rule, which gave 4.1% false-positive and 1.8% false-negative results, and the positive linear trend test described in the section “Linear Regression on Logarithmically Transformed Revertant Counts,” which gave 20.0% false-positives and 0.4% false-negative results. If the modified twofold rule is supplemented with a modified threefold rule, one would expect the proportion of false positives to decrease further.

For situations where there is no convincing evidence for the assumptions of Poisson distribution, overdispersion in the Poisson distribution, and the normality distribution

of the transformed data, the non-parametric methods discussed in the section “Non-parametric Methods”—namely, the Jonckheere test and the method proposed by Wahrendorf et al.—seem plausible. The latter also provide a descriptive quantity for assessment of the types of increasing trends and the possibility of a downturn due to toxicity. An example of revertant counts of strain TA100 in triplicate of (66, 82, 64), (66, 73, 87), (89, 84, 86), (76, 83, 87), (87, 103, 91), and (89, 98, 82) corresponding to the concentration levels of 0, 10, 30, 50, 100, and 300 μg per plate, respectively, was taken from Stead et al.^[14] The non-parametric method by Wahrendorf concluded a highly significant mutagenic effect with p value = 0.0028 based on normal approximation and p value = 0.0016 based on critical values established by Monte Carlo simulations. If the modified twofold rule were applied to the data, no mutagenic effects would be concluded.

Parametric methods with the assumptions of Poisson distribution with or without extra variation and negative binomial distribution of the revertant counts utilize statistical tools in modeling the counts for mutagenicity, toxicity, and dispersion, whereas the modified twofold rule calls only for simple averages of the revertant counts. Although the modified twofold rule demonstrated good agreement with the consensus of the microbiologists in the NCI/NTP study,^[21] it was considered too conservative by statisticians in general. It is clear that if the statistical methods are adopted, then there will be a great deal of statistically significant findings than using the modified twofold and threefold rules. To address the differences in the mutagenicity of compounds using statistical evaluation and biological evaluation, several issues merit further discussion here:

- i. The use of historical control data.
- ii. The toxicity threshold of a bacterial cell to a test article.
- iii. Extra-Poisson variation in the revertant counts.
- iv. The amount of histidine/tryptophan provided.
- v. The relationship between statistical significance and biological significance.

Clearly, these issues are all interrelated and, therefore, the discussion may not follow the order listed here. Because, in practice, most laboratories conduct Ames tests in triplicate, it is difficult to identify if extra-Poisson variation is present in the revertant counts within each experiment. A historical database for negative control data should be a must for examining the assumption of excessive variation in the revertant counts. In addition, a historical database can provide a base range for the evaluation of compounds that may yield statistical significance that is not biologically important. Because there may be variations among various laboratories, it is best for each laboratory to establish its own database for the historical negative control.

The toxicity threshold of a bacterial cell to a test article may be determined using the rule described by Chu

et al. or by 50–90% lethality. In other words, the number of revertants in the treated plates has to be reasonably low to indicate a toxic effect. Statistical methods can then be applied without modeling the toxic effects. However, if the rules are not accepted by the regulatory agencies, statistical modeling may be the best way to determine the presence of the toxic effects.

Biologically based models for different combinations of mutagenicity and toxicity may lead to different outcomes. How one determines the number of generations that the trace amino acid supply will last, the number of generations the compound will be mutagenic, and the number of generations the compound will be toxic is unclear.

CONCLUSION

At this point, no consensus has been reached on statistical methods for analyzing and interpreting the Ames data. This is reflected both in the existing ICH guidelines for genotoxicity testing,^[5,29,30] and in the revised guideline currently in development,^[31] neither of which mentions any statistical methods for the evaluation of the test results. Considering the small set of numbers that the Ames test provides for each bacterial strain, statistical methods designed to evaluate the mutagenic effects should reflect the biological significance to be of any help to microbiologists. Statistical methods that find biologically small increases to be statistically significant are not desirable and may slow the development of a compound unnecessarily. Any statistical methods recommended for the routine data analysis for the Ames test should be validated by a collection of compounds with known mutagenicity. A historical database is a must in the evaluation of the mutagenicity of compounds and should be established in each laboratory. When a historical database is not available, non-parametric approaches may be a better choice than parametric methods. Statistical methods should support microbiologists' findings when the effects, mutagenic or non-mutagenic, are clear, and should guide scientists when the effects are not clear. Because the modified two- and threefold rules seem to agree well with the consensus of microbiologists,^[21] further work in this area is needed to compare the proposed statistical methods with the modified two- and threefold rules based on known compounds.

REFERENCES

1. Ames, B.N.; Lee, F.D.; Durston, W.E. An improved bacterial test system for the detection and classification of mutagens and carcinogens. *Proc. Natl. Acad. Sci. U.S.A.* **1973**, *70*, 782–786.
2. Ames, B.N.; McCann, J.; Yamasaki, E. Methods for detecting carcinogens and mutagens with the *Salmonella*/mammalian microsome mutagenicity test. *Mutat. Res.* **1975**, *31*, 347–364.

3. Casciano, D.A. Introduction: Historical Perspectives of Genetic Toxicology. *Genetic Toxicology*; CRC Press: Boca Raton, FL, 1991, 1–12.
4. Auletta, A.E.; Dearfield, K.L.; Cimino, M.C. Mutagenicity test schemes and guidelines: U.S. EPA Office of Pollution Prevention and Toxics and Office of Pesticide Programs. *Environ. Mol. Mutagen* **1993**, *21*, 38–45.
5. Muller, L.; Kikuchi, Y.; Probst, G.; Schechtman, L.; Shimada, H.; Sofuni, T.; Tweats, D. ICH-harmonised guidances on genotoxicity testing of pharmaceuticals: Evolution, reasoning and impact. *Mutat. Res.* **1999**, *436*, 195–225.
6. Cimino, M.C. New OECD Genetic Toxicology Guidelines and Interpretation of Results. *Genetic Toxicology and Cancer Risk Assessment*; Marcel Dekker: New York, 2001; 223–248.
7. Claxton, L.D.; Allen, J.; Auletta, A.; Mortelmans, K.; Nestmann, E.; Zeiger, E. Guide for the Salmonella typhimurium/mammalian microsome tests for bacterial mutagenicity. *Mutat. Res.* **1987**, *189*, 83–91.
8. Gatehouse, D.; Haworth, S.; Cebula, T.; Gocke, E.; Kier, L.; Matsushima, T.; Melcion, C.; Nohmi, T.; Ohta, T.; Venitt, S.; Zeiger, E. Recommendations for the performance of bacterial mutation assays. *Mutat. Res.* **1994**, *312*, 217–233.
9. Organisation for Economic Cooperation and Development (OECD). Bacterial Reverse Mutation Test, No. 471. *OECD Guidelines for Testing of Chemicals*; OECD: Paris, 1997.
10. Edler, L. Statistical methods for short-term tests in genetic toxicology: The first fifteen years. *Mutat. Res.* **1992**, *277*, 11–33.
11. Mahon, G.A.T.; Middleton, B.; Robinson, W.D.; Green, M.H.L.; Mitchell, I.; Tweats, D.J. Analysis of Data from Microbial Count Assays. *Statistical Evaluation of Mutagenicity Test Data*; Cambridge University Press: Cambridge, 1989, 26–65.
12. Lin, K.K. Statistical review and evaluation of in vitro mutagenicity study data. *Drug Inf. J.* **1997**, *31*, 335–344.
13. Fisher, R.A. The significance of deviations from expectation in a poisson series. *Biometrics* **1950**, *6*, 17–24.
14. Stead, A.G.; Hasselbald, V.; Greason, J.P.; Claxton, L. Modelling the Ames test. *Mutat. Res.* **1981**, *85*, 13–27.
15. Vollmar, J. Statistical Problems in the Ames Test. *Progress in Mutation Research*; Elsevier/North-Holland Biomedical Press: Amsterdam, 1981, 2, 179–186.
16. Margolin, B.H.; Kaplan, N.; Zeiger, E. Statistical analysis of the Ames Salmonella/microsome test. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 3779–3783.
17. Kier, L.; Brusick, D.J.; Auletta, A.E.; Von Halle, E.S.; Brown, M.M.; Simmon, V.F.; Dunkel, V.; McCann, J.; Mortelmans, K.; Prival, M.; Rao, T.K.; Ray, V. The Salmonella typhimurium/mammalian microsome assay. A report of the U.S. Environmental Protection Agency Gene-Tox Program. *Mutat. Res.* **1986**, *168*, 69–240.
18. Jonckheere, A.R. A distribution-free k-sample test against ordered alternative. *Biometrika* **1954**, *41*, 133–145.
19. Kruskal, W.H.; Wallis, W.A. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621.
20. Wahrendorf, J.; Mahon, G.A.T.; Schumacher, M. A non-parametric approach to the statistical analysis of mutagenicity data. *Mutat. Res.* **1985**, *147*, 5–13.
21. Chu, K.C.; Patel, K.M.; Lin, A.H.; Tarone, R.E.; Linhart, M.S.; Dunkel, V.C. Evaluating statistical analyses and reproducibility of microbial mutagenicity assays. *Mutat. Res.* **1981**, *85*, 119–132.
22. Margolin, B.H.; Kim, B.S.; Risko, K.J. The Ames Salmonella/microsome mutagenicity assay: Issues of inference and validation. *J. Am. Stat. Assoc.* **1989**, *84*, 651–661.
23. Krewski, D.; Leroux, B.G.; Bleuer, S.R.; Broekhoven, L.H. Modeling the Ames Salmonella/microsome assay. *Biometrics* **1993**, *49*, 499–510.
24. Margolin, B.H.; Resnick, M.A.; Rimpo, J.Y.; Archer, P.; Galloway, S.M.; Bloom, A.D.; Zeiger, E. Statistical analysis for in vitro cytogenetic assays using Chinese hamster ovary cells. *Environ. Mutagen.* **1986**, *8*, 183–204.
25. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd Ed.; Chapman and Hall: London 1989, 193–214.
26. Tukey, J.W.; Ciminera, J.L.; Heyse, J.F. Testing the statistical certainty of a response to increasing doses of a drug. *Biometrics* **1985**, *41*, 295–301.
27. SAS Institute Inc. *The GENMOD procedure*. SAS/STAT User's Guide, version 8; SAS Institute: Cary, NC, 1999; 2, 1363–1464.
28. Margolin, B.H. Statistical studies in genetic toxicology: A perspective from the U.S. national toxicology program. *Environ. Health Perspect.* **1985**, *63*, 187–194.
29. International Conference on Harmonisation; Guidance on specific aspects of regulatory genotoxicity tests for pharmaceuticals. *Fed. Regist.* **1996**, *61*, 18198–18202.
30. International Conference on Harmonisation. Guidance on genotoxicity: A standard battery for genotoxicity testing of pharmaceuticals. *Fed. Regist.* **1997**, *62*, 62472–62475.
31. International Conference on Harmonisation; Draft Guidance on S2(R1) Genotoxicity Testing and Data Interpretation for Pharmaceuticals Intended for Human Use. *Fed. Regist.* **2008**, *73*, 16024–16025.

Analysis of $2 \times K$ Tables

Shiva Gautam

Harvard Medical School, Boston, Massachusetts, U.S.A.

Abstract

This entry presents some existing methods of analyzing data in $2 \times K$ nominal and ordinal tables, and then discusses some recently developed methods for $2 \times K$ ordinal table as an extension to Pearson chi-square test.

INTRODUCTION

Data in $2 \times K$ contingency tables are encountered quite frequently in biomedical, epidemiological, social, and behavioral studies. The variable representing two rows is often called the row variable, whereas the variable representing K columns is called column variable. (Representation of a data set either in a $2 \times K$ or a $K \times 2$ table is just a matter of convenience.) Depending on the research design, either of the column and row variables may be outcome (response) variables or only one of them may be an outcome variable. More specifically, an observation may have been simultaneously categorized into one of the two rows and into one of the K column categories, or an observation may have been first drawn from a given classification of one of the variables (row or column) and then classified into one of the categories of the other variable (column or row). For example, without taking into account the pros and cons of study designs, consider a possible study to evaluate the association between smoking and lung cancer. The investigator may choose a design in which he/she first selects two groups of people according to whether they have or have no cancer. Then each subject is classified into one of the smoking history categories (e.g., nonsmoker, light smoker, heavy smoker, etc.). Similarly, the investigator may first select people according to smoking status, and then classify each subject from each smoking group according to whether he/she has or has no lung cancer. Finally, the investigator may select a fixed number of subjects and then simultaneously classify them into one of the two lung cancer categories and into one of the several smoking categories. In many situations, the same computational procedures can be used while analyzing data regardless of the study design.

In the analysis of $2 \times K$ nominal table it is important to distinguish between a nominal table and an ordinal table. The data from the lung cancer and smoking study alluded above give rise to an ordinal $2 \times K$ table as the column of the tables (e.g., nonsmoker, light smoker, heavy smoker, etc.) follow an ordering (increasing) or a hierarchy in the

sense that any one category will either be at a higher level or at a lower level than any of the other remaining categories. Sometimes such an ordering among categories is also called simple ordering. A $2 \times K$ table with no ordering, in the sense that any category is neither at a higher nor at a lower level than any one of the other remaining category, is called nominal category. For example, a table showing low-birthweight babies (low birthweight = yes or no) and ethnicity (Asians, Blacks, Hispanics, Whites, etc.) is an example of a $2 \times K$ nominal table.

This paper presents some existing methods of analyzing data in $2 \times K$ nominal and ordinal tables, and then discusses some recently developed methods for $2 \times K$ ordinal table as an extension to Pearson chi-square test.

ANALYSIS OF $2 \times K$ NOMINAL TABLES

Let n_{ij} denote the number of observations in the i th row ($i = 1, 2$) and j th column ($j = 1, 2, \dots, K$) as displayed in Table 1. Also, let $n_{i+} = \sum_{j=1}^K n_{ij}$, $n_{+j} = \sum_{i=1}^2 n_{ij}$, and $n = \sum_{i=1}^2 \sum_{j=1}^K n_{ij} = \sum_{i=1}^2 n_{i+} = \sum_{j=1}^K n_{+j}$. The columns of the table, for the time being, are assumed to be nominal.

Pearson's Chi-Square Test

Perhaps the most popular method for analyzing $2 \times K$ nominal tables is the Pearson's chi-square procedure introduced by Karl Pearson in 1900. The Pearson chi-square test statistic is defined as

$$X^2 = \sum_{i=1}^2 \sum_{j=1}^K (n_{ij} - \hat{\theta}_{ij})^2, \quad (1)$$

where $\hat{\theta}_{ij} = \frac{n_{i+} n_{+j}}{n}$.

The statistic X^2 is asymptotically distributed as the chi-square variate with $(K - 1)$ degrees of freedom. A large value of X^2 provides evidence against the null hypothesis.

Table 1 A $2 \times K$ Contingency Table

Row	Column					Total
	1	2	...	...	K	
1	n_{11}	n_{12}	...	...	n_{1k}	n_{1+}
2	n_{21}	n_{22}	...	...	n_{2k}	n_{2+}
Total	n_{+1}	n_{+2}	...	...	n_{+k}	n

The null hypothesis is often stated as there is no association between the row and column variables. Depending on the research question, the null hypothesis could be that the distribution of proportions in each row (two populations) is the same or the column proportions (K -populations) are the same. As usual, the decision to reject (or not to reject) the null hypothesis is based on the p -value. Agresti^[1] is an excellent source on chi-square analysis of two-way nominal categorical tables.

Likelihood Ratio Chi-Square

Likelihood ratio chi-square statistic is also used to make inference from a nominal contingency table. It is defined as

$$G^2 = -2 \sum_{i=1}^2 \sum_{j=1}^k n_{ij} \log(n_{ij}/\hat{\theta}_{ij}), \quad (2)$$

where $\hat{\theta}_{ij} = \frac{n_{i+} n_{+j}}{n}$.

For large n , G^2 also has chi-square distribution with $(K-1)$ degrees of freedom. Hence both X^2 and G^2 analyses of a given data set in a $2 \times K$ nominal table will generally yield similar results for large n .

Logistic Regression

When the two rows of a $2 \times K$ table represent response, logistic regression may be used to analyze the data by modeling the probability of response (e.g., present vs. absent). Let p = probability of response in the first row. Define dummy variable $X_2, X_3, \dots, X_k$ such that $X_j = 1$ if the observation is from the j th category ($j = 2, 3, \dots, k$), and $X_j = 0$, otherwise. The logistic regression model can be represented as

$$\text{logit}(p) = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k \quad (3)$$

where $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$.

Note that $\hat{\beta}_1$ is log odds of responding in row 1 from column 1 (reference column) or equivalently when $X_2, X_3, \dots, X_k$ equal to 0 and X_1 equals to 1. In other words, $\hat{\beta}_1 = \log(n_{11}/n_{21})$. From the above equation $\hat{\beta}_j$ is the excess of log odds responding in row 1 due to the j th column

than the response due to the first column. In other words, $\exp(\hat{\beta}_j)$ represents odds ratio (odds of response due to the j th column compared to odds of response due to the first column).

[Note: Suppose π_1 denotes the probability of lung cancer for smoker, and π_2 denotes the chance of lung cancer for a nonsmoker, then the odds in favor of lung cancer for a smoker is given by $\pi_1/(1-\pi_1)$, and the odds in favor of lung cancer for a nonsmoker by $\pi_2/(1-\pi_2)$. The ratio $\pi_1(1-\pi_2)/\pi_2(1-\pi_1)$ is referred to as the odds ratio which is often used as a direct/indirect measure of the relative risk of a disease (cancer) with an exposure (smoking) relative to the same disease without the exposure. Sample odds ratios are calculated using the observed proportions instead of probabilities.]

One of the advantages of using logistic regression is that it quantifies the magnitude of association. Furthermore, the effect of any additional variables can be adjusted in the model. For example, while evaluating the association of low birthweight (yes, no) and race (Blacks, Hispanics, Whites, Others), investigators may want to adjust for the effect of a covariate (e.g., weight of mothers). Because of these and some other desirable properties and ease of interpretation of coefficients, the logistic regression procedure^[2,3] is widely used to model binary response data from biomedical and other studies.

Maximal Correlation and Pearson's Chi-Square

Consider Table 1 as a sample from a bivariate distribution U (row) and V (column) variable. Let $U = 1$ if an observation is classified into the first row, $U = 0$ if an observation is classified into the second row. Let $V = s_j$ if an observation is classified into the j th category ($j = 1, 2, \dots, k$), where s_j is a real number. The value s_j taken by the variable V corresponding to the j th column of Table 1 will be referred to as a "score" hereafter. Let $r^2\{s_1, s_2, \dots, s_k\}$ denote the square of the Pearson's correlation between U and V for a given set of scores $\{s_1, s_2, \dots, s_k\}$. Let $r_{\max}^2$ denote the maximum of $r^2\{s_1, s_2, \dots, s_k\}$ over all possible sets of scores. Then it can be shown^[4,5] that,

$$nr_{\max}^2 = X^2 \quad (4)$$

where X^2 is the Pearson's chi-square statistics.

It can be shown that the maximal score for $r_{\max}^2$ is given by the set of scores $\{n_{11}/n_{+1}, n_{12}/n_{+2}, \dots, n_{1k}/n_{+k}\}$ or any set of scores obtained from a linear transformation of these scores.^[5] If $r_{\max} = \sqrt{r_{\max}^2}$, then it can be shown that $r_{\max}^2 = (r_{\max})^2 = (r_{\min})^2$, where $r_{\min} = -r_{\max}$. Thus $r_{\max}$ is the maximum possible correlation between the row and column variable, and is always nonnegative.

As mentioned earlier X^2 is simply a significance test and does not provide information on the magnitude of association between row and column variables. As evident from

Eq. 4, a large value of chi-square may result if there is a large sample size even when the association is poor. However, the maximum possible correlation $r_{\max} = \sqrt{\frac{X^2}{n}}$ can be used as a measure of association as it meets several criteria outlined by Goodman and Kruskal for a measure of association.^[6] As $\sqrt{\frac{X^2}{n}}$ is also equal to Cramer's V and the ϕ coefficient, the maximal correlation gives a new meaning to these quantities which are incorporated into the output of some statistical packages. Note that $0 \leq r_{\max} \leq 1$ and $r_{\max} = 1$ if and only if total observations in each column are contributed by only one of the two rows.^[5]

Regression and Pearson's Chi-Square

Consider U and V as row and column variables of Table 1. Furthermore, define $K - 1$ dummy variables $X_2, X_3, \dots, X_k$ corresponding to the second, third, ..., and k th category, respectively (these are the same variables defined earlier in the context of logistic regression). Consider the following predicted line from the linear regression U on $X_2, X_3, \dots, X_k$

$$\hat{U} = \hat{\alpha}_1 + \hat{\alpha}_2 X_2 + \hat{\alpha}_3 X_3 + \dots + \hat{\alpha}_k X_k \quad (5)$$

If R denotes the multiple correlation coefficient then it can be shown that,^[5]

$$nR^2 = X^2 \quad (6)$$

It follows from Eqs. 4 and 6 that $r_{\max}^2 = R^2$. It may be desirable to collapse the columns of a $2 \times K$ table without losing much information.^[7] Gautam and Kimeldorf showed that two columns of a $2 \times K$ table can be collapsed into one if the regression coefficients are equal.^[5] Similarly, if a regression coefficient equals zero then the corresponding column can be collapsed to the column representing the intercept. Therefore, one can test the hypothesis that only a given subset of categories is responsible for the association between the variables. A post hoc analysis may also be performed to find out the reduced table obtained by collapsing categories. This is analogous to testing of the hypothesis

for a subset of regression parameters in multiple linear regression setting.

Exact Tests

Inferences drawn from the above-described analyses of $2 \times K$ tables are based on a large sample theory. When the entries n_{ij} of Table 1 are small then the p -value is obtained directly. For example, exact tests such as Fisher's exact test for 2×2 tables can be extended to a $2 \times K$ table. An exact test for a $2 \times K$ table is often performed by generating all possible tables or randomly generating some large number of tables (e.g., 10,000 tables) with the same marginal totals as the observed table assuming the null hypothesis is true. The p -value is the proportion of tables yielding values of a statistic (e.g., X^2) that are equal to or larger than the value of the same statistic obtained from the observed table.

An Example

Consider Table 2 from Helmes and Fekken.^[8] The table is also reproduced in Agresti.^[1] The table classifies psychiatric patients by their diagnosis and whether the treatment prescribed drugs.

The Pearson chi-square statistic from Table 2 is $X^2 = 84.180$ ($p < 0.0001$, $df = 4$). This suggests an association between the diagnosis and whether or not a patient's treatment prescribed drugs. The above table shows that a schizophrenic patient is most likely to be treated by drugs followed by patient diagnosed as active disorder and personality disorder, respectively. A patient with neurosis has an almost 50% chance of being prescribed drugs, whereas a patient classified as having special symptoms is not likely to be treated by a drug. The Pearson chi-square test rejects the hypothesis that these proportions in the population are the same. In other words, there seems to be an association between the diagnosis and whether or not the treatment prescribed a drug. The last row shows the odds ratio of being prescribed with a drug for a given diagnosis compared to odds of being prescribed with a drug if a patient is diagnosed as schizophrenic.

Table 2 Diagnosis of Patients and Whether Their Treatment Prescribed Drugs

Treatment	Diagnosis					Total
	Schizophrenia	Active disorder	Neurosis	Personality disorder	Special symptoms	
Drugs	105	12	18	47	0	182
No drugs	8	2	19	52	13	94
Total	113	14	37	69	13	276
% Drugs	92.92	85.71	48.65	68.12	0	
Odds ratio	1.0	0.46	0.07	0.07	0	

Source: Ref. [8].

Logistic Regression

Let

$$\begin{aligned} X_2 &= \begin{cases} 1, & \text{if Active Disorder} \\ 0, & \text{otherwise} \end{cases} \\ X_3 &= \begin{cases} 1, & \text{if Neurosis} \\ 0, & \text{otherwise} \end{cases} \\ X_4 &= \begin{cases} 1, & \text{if Personality Disorder} \\ 0, & \text{otherwise} \end{cases} \\ X_5 &= \begin{cases} 1, & \text{if Other Symptoms} \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

Table 3 shows the results from the logistic regression model presented in Eq. 3.

A quick inspection of Table 3 reveals that the entries in the last column are odds ratios given in the last row of Table 2 except for the entry corresponding to “constant.” The entry corresponding to constant is simply the odds of being prescribed drug in the schizophrenia category ($105/8 = 13.215$). Logistic regression analysis shows that at least one odds ratio is different from one, thus indicating an association between row and column variable. Some caution must be taken while analyzing data in tables with zero entries. A small number is often added to such entries. As an alternate, exact logistic regression analysis can be performed.

A large value of chi-square is indicative of an association between treatment prescribing drug and diagnosis, but it does not indicate whether the statistical significance is a reflection of a strong association or an artifact of weaker association and a large sample size. As discussed above, the maximal correlation could shed further light into this association. From the relationship between the Pearson chi-

square and $r_{\max}$, we have $r_{\max} = \sqrt{\frac{X^2}{n}} = \sqrt{\frac{84.180}{276}} = 0.5523$.

This shows that the observed significant association is not solely due to a large sample size. A natural question that may arise is whether some of the categories can be combined without losing much information.^[7] An investigator may be further interested to determine whether this

association is mostly due to only a few selected categories of the table. Gautam and Kimeldorf^[5] used multiple regression procedure to compute the contribution of a category to the total association noting that $r_{\max}$ is equal to the multiple correlation coefficient R obtained from the regression of row variable U ($U = 1$ if first row and $U = 0$ if second row) on dummy variables X_2, X_3, X_4 , and X_5 defined earlier. They showed that if the 2×3 table is obtained by collapsing columns 1 and 2 together, columns 3 and 4 collapsing together and leaving column 5 the way it is, then the maximal correlation (Pearson chi-square) from this reduced table is 0.5511 (chi-square = 83.824) which is about 99.8% of the maximal correlation (chi-square) from the original 2×5 table.

If the table is reduced to the 2×2 table by collapsing the last four columns into one column, then the maximal correlation from this 2×2 table is 0.4740 (chi-square = 62.011). The reduction in the correlation is 0.0783, and the corresponding reduction in chi-square is 22.179 (p -value < 0.001). Furthermore, Gautam and Kimeldorf argue that this significance is due to the sample size rather than an indication of the weakening of the association due to collapsing of the categories.^[5] The two columns of this table classify patients into schizophrenia and nonschizophrenia groups, and the 2×2 table still retains about 86% of the information provided by the original 2×5 table. Hence the association between row and column variables in the original table is basically the association between diagnosis of schizophrenia (yes or no) and whether the treatment prescribed (yes or no) a drug. In terms of odds ratios, the odds of being on prescription drugs for a person diagnosed with schizophrenia are 14.66 times the odds of being on prescription drugs for a person diagnosed with a nonschizophrenia (95% CI : 6.71–32.04) category.

ANALYSIS OF $2 \times K$ ORDERED TABLES

Pearson's chi-square procedures and other tests developed for analyzing data in $2 \times K$ nominal tables do not incorporate the information on ordering among the columns of the table. These tests are not directed toward any specific alternate hypothesis. In analyzing data in a $2 \times K$ ordered table, investigators will obviously want to use as much information as possible provided by the data and also often want to determine whether the null hypothesis can be rejected against a specific alternate hypothesis (e.g., increasing response with the columns). A test that utilizes ordering information will have increased power compared to a test for nominal tables.^[1]

Methods for analyzing data in $2 \times K$ ordered tables may be broadly classified into two groups, namely, methods that assign and that do not assign numerical scores to the ordered categories, respectively. Methods that do assign numerical scores to the ordered categories may further be

Table 3 Results from Logistic Regression on Data in Table 2

Variable	β	SE(β)	p -value	Exp(β)
X_2	− 0.783	0.847	0.356	0.457
X_3	− 2.629	0.493	0.000	0.072
X_4	− 2.676	0.418	0.000	0.069
X_5	− 23.777	11,147.524	0.998	0.000
Constant	2.575	0.367	0.00	13.125

divided into two subgroups. In the first subgroup of methods, scores are chosen a priori and the analysis is carried out using these scores, whereas in the second subgroup, scores are extracted from the data and thus are functions of the observed data.

Methods Without Scores

One of the widely used methods for analyzing data in $2 \times K$ ordered tables with the two rows representing two populations is the Mann–Whitney or equivalently Wilcoxon rank sum procedure.^[9] Dykstra et al.^[10] proposed a likelihood ratio statistic for testing whether one population with ordered outcome is larger than the other in the sense likelihood ratio ordering. The Likelihood Proportional Odds Model is another method used to analyze $2 \times K$ ordered data without assigning scores to ordinal categories.^[11]

Methods with Scores

Several methods of data analysis for $2 \times K$ ordered tables assign order-preserving scores. The Cochran–Armitage–Mantel trend test is widely used to evaluate the trend in proportion.^[12–14] This trend test requires assignment of order-preserving scores. Another widely used method that utilizes order-preserving scores is logistic regression. The scores are chosen a priori by the investigator, and often, equally spaced scores (e.g., 1, 2, 3, ..., k) are assigned to the ordered categories. In some situations, the ordered categories are defined by actual intervals or actual quantity (e.g., dose level) in which case the scores may be chosen as mid values of the interval or the numerical numbers used to define the categories. The trend test is simply a test of significance and does not provide the magnitude of the association. The slope from the logistic regression is a function of the odds ratio [$\exp(\text{slope}) = \text{odds ratio}$].^[2,3] It is worth noting that the test statistic for the trend test and logistic regression are equivalent for a given set of scores. Hence the p -value from the trend test under the null hypothesis of no trend and the p -value for the logistic regression under the null hypothesis unit odds ratio are the same. When the two rows represent two independent samples, one can compare the row means with a given set of scores using Student's t -test. Again, the p -value from the t -test will also be equal to the p -value from the trend test or the logistic regression if one were to use these methods. Finally, the test of zero correlation (Pearson) between row and column variables would also have yielded the same p -value.

Let the row variable be denoted by U and column variable V with its values as category scores. If $r = \text{corr}(U, V)$ is the correlation between U and V for a given set of order-preserving scores, then the test statistic for the trend test and logistic regression is equal to nr^2 , and the test statistic for the t -test is $t = \frac{nr}{\sqrt{1-r^2}}$ (where n is the number of observations). Therefore, if linear regression instead of

logistic regression is used, then one would obtain the same value of the test statistic and the same p -value while testing the null hypothesis of zero slope. Although the research questions being asked and the design generating a $2 \times K$ ordered table may be different, a common computational procedure may be employed to obtain the statistical significance. Graubard and Korn listed several equivalent test statistics for a given set of scores.^[15]

Choice of Scores

As discussed, several methods (e.g., trend test) of analyzing data in $2 \times K$ ordered tables use order-preserving scores. Even the Mann–Whitney (or Wilcoxon rank sum) test which apparently is a non-score-based method is equivalent to a score-based method. The t -test with mid-rank as category scores is equivalent to the Wilcoxon rank sum test. Cochran noted that any set of scores gives a valid test.^[12] However, different sets of scores may yield different results.^[15]

Iso-Chi-Square Test

Gautam et al. proposed the Iso-chi-square test which is an extension of Pearson's chi-square test.^[16] This test also addresses the issue of arbitrariness of the scores assigned to the columns of $2 \times K$ ordered tables. It was shown earlier that the Pearson chi-square test statistic is equal to $nr_{\max}^2$ where the maximum was taken over all possible column scores (with no restriction on the scores). The Iso-chi-square statistic is given by the same expression but the maximum is taken over all possible order-preserving scores. There is a closed form of solution for the maximal scores and it is given by isotonic regression. However, it is not necessary to extract the maximal scores. The Iso-chi-square is equal to the Pearson chi-square obtained from either the original $2 \times K$ table or from a reduced table obtained by collapsing certain adjacent categories. The null distribution of the Iso-chi-square therefore is given by a mixture of chi-square distributions with 1 to $K - 1$ degrees of freedom. Exact p -values can be obtained by generating tables with given marginal totals. Gautam et al. listed 5% and 1% cut-off values for several values of K .^[16]

Iso-chi-square also addresses the issue of arbitrary assignment of scores to the ordered categories as it reports the maximal value of the test statistics. In cases where there is no clear indication of what scores are to be used, Agresti^[1] suggests using sensitivity analysis by choosing a few sets of scores. Iso-chi-square actually assigns all possible sets of order-preserving scores. It is obvious that Iso-chi-square is also related to the maximal t -statistic and the maximal trend statistic. Iso-chi-square which utilizes all possible scores is equivalent to a method that does not utilize order-preserving scores proposed by Dykstra et al.^[10] and thus links traditional statistical methods with

Table 4 Maternal Alcohol Consumption and Congenital Sex Organ Malformation of the Children

Malformation	Alcohol consumption (average number of drinks per day)				
	0	< 1	1–2	3–5	≥ 6
Absent	17,066	14,464	788	126	37
Present	48	38	5	1	1

the correspondence analysis or dual scaling.^[17] It was also pointed out in an earlier section that the Pearson chi-square test statistic is equivalent to the maximal correlation obtained without order restriction on the scores.

An Example

Consider Table 4, which classifies maternal drinking and congenital sex organ malformation of babies.^[15]

If the two sample Wilcoxon rank sum test is used then the p -value = 0.56 which is also the p -value from the trend test with midranks as category scores. If equally spaced scores {1, 2, 3, 4, 5} are used then the p -value = 0.20 (from the trend test, t -test, logistic regression, linear regression, and correlation analysis). In an example such as this perhaps the mid-values of the interval represent the underlying continuous measure. Graubard and Korn used scores of 0, 1.5, 4.0, and 7 (somewhat arbitrary) which yield a p -value equal to 0.01.^[15] Iso-chi-square analysis for this data set yields a p -value of 0.02. These are exact p -values.

Stochastic Ordering

Stochastic ordering, in the context of a $2 \times K$ ordered table, is defined as having the cumulative distribution function (CDF) of one of the rows not crossing the distribution function of the other. In terms of the entries of Table 1, $F_{1j} = \sum_{i=1}^j n_{1i}/n_{1+}$ and $F_{2j} = \sum_{i=1}^j n_{2i}/n_{2+}$. If $F_{2j} \leq F_{1j}$ for all $j = 1, 2, \dots, K$, then row 2 is stochastically larger than row 1 (in the observed sample data). It is interesting to note that row 2 is stochastically larger than row 1 if and only if all order-preserving scores yield a larger (or equal) mean for row 2 than the corresponding mean for row 1. Kimeldorf et al. computed the minimum and maximum t -statistics, and argued that if the minimum t -statistic is positive and significant (or the maximum t -statistic is negative and significant), then any set of order-preserving scores will produce a significant result.^[18] This could be used as an evidence of stochastic ordering in a given $2 \times K$ table. Berger et al. propose the convex hull test for ordered categorical data.^[19]

CONCLUSION

In this article some existing methods of analyzing data in $2 \times K$ ($K > 2$) contingency tables are discussed. Pearson's

chi-square test statistic which is widely used to analyze nominal data is shown to be related to maximal correlation. Using maximal correlation an investigator may determine if only a few categories contribute to the observed association. This relationship between the chi-square and the maximal correlation may also shed light on whether the large value of the chi-square test statistic is only due to a large sample size.

The paper also discusses methods of analysis of $2 \times K$ ordered table. Some of these methods use order-preserving scores and others that do not use such scores. Several of the methods that utilize scores are equivalent to each other. As these methods are directed toward a particular alternative hypothesis they have more power in general than the methods that do not utilize such scores. Also, these methods are computationally simple. However, the scores chosen are often arbitrary. In many situations the columns may provide some indication (e.g., interval, actual dose of a drug, etc.), where it makes sense to use certain scores. But in a situation where the columns are defined as "low," "medium," and "high," it may be difficult to come up with a set of score. In such situations, the Iso-chi-square method may be useful. Iso-chi-square may be considered as a natural extension of the Pearson chi-square to the $2 \times K$ ordered table in the sense that if this procedure is applied to $2 \times K$ nominal tables, the test statistic is the Pearson's chi-square test statistic. Also, Iso-chi-square may be considered as a link between methods that do and do not utilize order-preserving score.

All the $2 \times K$ tables discussed here are assumed to have simple ordering. There may be other types of tables where the ordering between two categories is not simple. For example, parental drinking or smoking may be classified as "neither parent," "mother only," "father only," and "both parents." The level of the first or the last category has a distinct hierarchy compared with any other categories. However, such a hierarchy between the second, the first, and the third is not defined. Similarly, some $2 \times K$ tables may have mixed categories (both nominal and ordinal categories) or may have open-ended categories.^[20,21] The method of Iso-chi-square may be extended in such situations. Gautam presented a case where the choice of an open-ended category may influence the statistical conclusion in the context of the trend test and argued that such an analysis may be misleading.^[22]

ACKNOWLEDGMENTS

The author would like to express his sincere thanks to Roger Davis ScD for his valuable comments. This research was supported in part by grant RR 01032 to the Beth Israel Deaconess Medical Center General Clinical Research Center from the National Institutes of Health.

REFERENCES

1. Agresti, A. *Categorical Data Analysis*, 2nd Ed.; Wiley: New York, 2002.
2. Hosmer, D.W.; Lemeshow, S. *Applied Logistic Regression*, 2nd Ed.; Wiley: New York, 2000.
3. Collett, D. *Modeling Binary Data*; Chapman & Hall: London, 1991.
4. Haberman, S.J. Test for independence in two-way contingency tables based on canonical correlation and linear-by-linear interaction. *Ann. Stat.* **1981**, *9*, 1178–1186.
5. Gautam, S.; Kimeldorf, G. Some results on the maximal correlation in $2 \times K$ contingency tables. *Am. Stat.* **1999**, *53* (4), 336–341.
6. Goodman, L.A.; Kruskal, W.H. Measures of association for cross-classifications. *J. Am. Stat. Assoc.* **1954**, *49*, 732–764.
7. Bishop, Y.M.M.; Fienberg, S.E.; Holland, P.W. *Discrete Multivariate Analysis: Theory and Practice*; The MIT Press: Cambridge, MA, 1995.
8. Helmes, E.; Fekken, G.C. Effects of psychotropic drugs and psychiatric illness on vocational aptitude and interest assessment. *J. Clin. Psychol.* **1986**, *42*, 569–576.
9. Conover, W.J. *Practical Nonparametric Statistics*, 2nd Ed.; Wiley: New York, 1980.
10. Dykstra, R.; Kocher, S.; Robertson, T. Inference for likelihood ratio ordering in the two sample problem. *J. Am. Stat. Assoc.* **1995**, *90*, 1034–1040.
11. McCullugh, P. Regression models for ordinal data. *J. R. Stat. Soc., Ser. B.* **1980**, *42*, 109–142.
12. Cochran, W.G. Some methods of strengthening the common chi-square test. *Biometrics* **1954**, *10*, 417–451.
13. Armitage, P. Tests for linear trend in proportion and frequency. *Biometrics* **1955**, *11*, 375–386.
14. Mantel, N. Chi-square test with one degree of freedom: Extension of the Mantel–Haenszel procedure. *J. Am. Stat. Assoc.* **1963**, *58*, 690–700.
15. Graubard, B.I.; Korn, E.L. Choice of column scores for testing independence in ordered $2 \times K$ contingency tables. *Biometrics* **1987**, *43*, 471–476.
16. Gautam, S.; Singh, H.; Sampson, A. Iso-chi-square testing in $2 \times K$ ordered tables. *Can. J. Stat.* **2002**, *29*, 609–629.
17. Nishisato, S. *Analysis of Categorical Data: Dual Scaling and Its Applications*; University of Toronto Press: Toronto, 1980.
18. Kimeldorf, G.; Sampson, A.; Whitaker, L. Min max scoring for two sample ordinal data. *J. Am. Stat. Soc.* **1992**, *87*, 241–247.
19. Berger, V.W.; Permutt, T.; Ivanova, A. The convex hull test for ordered categorical data. *Biometrics* **1998**, *54*, 1541–1550.
20. Gautam, S. Test for linear trend in $2 \times K$ ordered tables with open ended categories. *Biometrics* **1997**, *53*, 1163–1169.
21. Gautam, S. Analysis of mixed categorical data in $2 \times K$ contingency tables. *Stat. Med.* **2002**, *21*, 1471–1484.
22. Gautam, S.; Ashikaga, T. Assessing the effect of open-ended category on the trend in $2 \times K$ ordered tables. *J. Data Sci.* **2003**, *1*, 167–183.

Analysis of Clustered Binary Data

Valerie Durkalski

Division of Biostatistics and Epidemiology, Medical University of South Carolina, Charleston, South Carolina, U.S.A.

Abstract

When a subject contributes more than one binary response (paired or unpaired), it results in clustered data in which the subject is the cluster and each response within a cluster is a unit. The focus of this entry is on various methods of analysis in terms of clustered binary data. Specifically, this entry deals with the ICC, pooled estimators, method of moments estimators, and GEE approaches.

Keywords: Binary outcomes; Chi-square test; Clustered data; GEE.

INTRODUCTION

Clinical trial designs often incorporate binary outcomes (i.e., proportion of responders; success probabilities) or change continuous outcomes into binary outcomes for interpretation purposes. When a subject contributes more than one binary response (paired or unpaired), it results in clustered data in which the subject is the cluster and each response within a cluster is a unit. In this scenario, responses are considered dependent within a cluster, and independent between clusters.

One can imagine several distinct contexts in which clustered binary data might arise. Examples include ophthalmology studies, in which the unit is each eye, but more generally the number of units may vary across clusters. This more general formulation then includes dental studies, in which the unit of analysis is each tooth, oncology studies, in which the units are infected nodes or lesions within a patient, and community intervention studies, in which the unit is the individual within the community. When dealing with two treatments, it would be possible for some clusters, and every unit in those clusters, to be assigned to one treatment group, and for the other clusters, and every unit in these other clusters, to be assigned to the other treatment group. For example, to examine the success of two teaching programs on the graduation rate in one school, one may randomize classrooms (cluster) to the treatment and have students be the unit of analysis. Fear of contamination would preclude one from randomizing students within the same classroom to different teaching conditions. This is also the context for repeated measures studies, in which the cluster is the patient and the units are the sampling times. Such a design allows for the separation of effects due to between-subject variability from effects due to within-subject variability.^[1] It is also possible that each cluster receives both treatments, but each unit within

each cluster receives one treatment or the other, such as in some ophthalmology studies, where one eye is given one treatment and the other is given the comparison treatment. A third context would be in which each unit receives both treatments, such as in matched-pair designs, where clusters (i.e., subjects) act as their own control and receive both procedures/interventions, or clusters are matched with each receiving one of the procedures/interventions.

At the unit level, the correlation between units within a cluster violates the independence assumption of many of the statistical methods for analyzing binary outcomes, such as Pearson's chi-square, McNemar's test, and logistic regression. This violation may decrease standard error estimates and the p -value associated with the test statistic, thereby yielding misleading statistical conclusions. The challenge that arises when analyzing clustered binary data, therefore, is handling the correlation among units within a cluster. While cluster summary approaches such as collapsing data within a cluster or focusing the design on cluster-level analyses are valid in some cases, they do not use all available data, and so more informative analysis techniques would be expected to yield more precise results. Unit-level analyses are the focus of this entry.

METHODS OF ANALYSIS

Methods for the analysis of clustered binary data (paired and unpaired) are well developed.^[2,3] Most research directly addresses important theoretical and applied issues regarding the effect of clustering and the consequential effect on the variance and on the statistical conclusions. Adjustments for correlated binary responses within a cluster have been incorporated into statistical tests using a correlated binomial model,^[4] Fisher's permutation test,^[5] binomial ratio estimators,^[6,7] pooled estimators,^[8] method of moments

estimators,^[9] and intracluster correlation (ICC).^[10] Another approach that is widely applied is marginal regression modeling using the general estimating equation (GEE) approach of Zeger and Liang.^[11] Although these comparison methods have been developed for clustered binary data, not all tests are consistent in terms of performance. Depending on the cluster sizes and the correlation structure within clusters, the tests performances may vary in terms of size and power. This entry focuses on the ICC, pooled estimators, method of moments estimators, and GEE approaches.

To review different analysis methods, three subscripts are defined in the case of two proportions, as follows. Let the response variable Y_{ijk} be dichotomous (1 = success, 0 = failure), where i (=1, 2) is the treatment or intervention, j (=1, 2, ..., n_k) is the unit within the cluster, and k (=1, 2, ..., K) is the cluster. Let $\hat{p}_{ik} = n_{ik}^{-1} \sum_{j=1}^{n_k} Y_{ijk}$ be the event rate (i.e., proportion of successes) in group i for cluster k ; K is the total number of clusters in the study population, and $N = \sum_{k=1}^K n_k$ is the total number of units across all the clusters for both treatments. The n_k units in the k th cluster are assumed to be fixed for all K clusters. With this framework, the data from each unit within each cluster can be summarized as a standard 2×2 contingency table, as both treatment and outcome vary, and are binary variables (assuming that there are only two treatments). The frequencies of the responses per group may be summed over all K clusters. This data display is presented in Table 1 for independent bivariate random variables and in Table 2 for dependent bivariate variables (matched-pair data).

HYPOTHESIS TESTING

When dealing with bivariate random variables, one for each of two treatment conditions, it is often of interest to test the hypothesis of equality of the probabilities of a positive response across the treatment conditions. Doing so is fairly straightforward when subjects are allocated to treatment groups by random allocation, with either no restrictions at all on the randomization or with the only restriction being that the group totals are fixed. In either case (conditioning on the random group totals in the first case), the

design-based analysis is Fisher's exact test.^[12] When clustering is present, however, this would not be an appropriate analysis, because it would ignore the clustering.

To avoid underestimating the variance of the parameter in the presence of clustering, Donner and colleagues explore a model using the intracluster correlation coefficient (ICC), which adjusts the chi-square test for correlation within clustered data.^[13,14] The ICC, originally applied to interobserver agreement analyses as an alternative to Cohen's kappa, an index of agreement between two or more raters, is estimated using the mean square errors of analysis of variance (ANOVA) for mixed models, where the treatment is constant and the cluster is random. The estimated ICC represents the proportion of variation due to between-cluster differences. Donner presents a detailed explanation of how a consistent estimate of the ICC is calculated under the assumption of a constant ICC across clusters.^[5] The ICC adjustment is applied to the Pearson chi-square test,^[13] chi-square test for linear trend,^[14] Mantel-Haenszel test,^[15] and McNemar test.^[16] The test statistic is adjusted by dividing it by an inflation factor, $C_i = 1 + (m-1)\hat{\rho}$, where m is the average number of units per cluster for treatment group i and ρ is the estimated ICC for clustering,

$$\hat{\rho} = \frac{\text{BMS} - \text{WMS}}{\text{BMS} + (S_0 - 1)\text{WMS}}$$

In this equation, BMS is the mean squared error between subjects, WMS is the mean squared error within subjects, and S_0 is the adjusted mean cluster size. The ICC approach for adjustment of a chi-square test is an extension of a standard chi-square test, because when $\rho = 0$, the adjusted test reduces to the standard test with one degree of freedom. Therefore, if there is only one unit per cluster, the standard test to compare overall event rates can be adopted,

$$\chi^2 = \sum_{i=1}^2 \frac{n_i(p_i - \hat{p})^2}{C_i \hat{p}(1 - \hat{p})}$$

where $p_i = \sum_{k=1}^K p_{ik}$ and $\hat{p}$ is the overall event rate. In addition, the test can be extended to the comparison of more than two treatment groups.^[14] Although the ICC appears to be quite adaptable to adjusting a variety of chi-square tests, this method has limitations regarding the required size of the study population and the test assumptions. Donald and Donner^[13] suggest that the number of units per cluster should be greater than 10 and the ratio of the number of units to the number of observations per unit should be greater than two when applying the adjusted chi-square to Pearson's chi-square test for homogeneity of proportions. They also note that the method's assumption, that the correlation between responses within a cluster is equal, may not be appropriate when the probability of a correct response varies

Table 1 Contingency Table for Clustered Binary Data

	Response	
	Success	Failure
Group 1	$\sum_{k=1}^K \sum_{j=1}^{n_k} y_{1jk}$	$\sum_{k=1}^K \sum_{j=1}^{n_k} (n_1 - y_{1jk})$
Group 2	$\sum_{k=1}^K \sum_{j=1}^{n_k} y_{2jk}$	$\sum_{k=1}^K \sum_{j=1}^{n_k} (n_2 - y_{2jk})$
	$\sum_{k=1}^K \sum_{j=1}^{n_k} (y_{1jk} + y_{2jk})$	

Table 2 Contingency Table for Clustered Matched-Pair Binary Data

		Procedure 1		
		Success	Failure	
Procedure 2	Success	$\sum_{k=1}^K \sum_{j=1}^{n_k} a_{jk}$	$\sum_{k=1}^K \sum_{j=1}^{n_k} b_{jk}$	$\sum_{k=1}^K \sum_{j=1}^{n_k} (a_{jk} + b_{jk})$
	Failure	$\sum_{k=1}^K \sum_{j=1}^{n_k} c_k$	$\sum_{k=1}^K \sum_{j=1}^{n_k} d_k$	
		$\sum_{k=1}^K \sum_{j=1}^{n_k} (a_{jk} + c_{jk})$		$\sum_{k=1}^K \sum_{j=1}^K n_k = N$

The joint probabilities of success and failure between the two procedures are $p_{11} = N^{-1} \sum_{k=1}^K E[a_k]$, $p_{10} = N^{-1} \sum_{k=1}^K E[b_k]$, $p_{01} = N^{-1} \sum_{k=1}^K E[c_k]$, and

$$p_{00} = N^{-1} \sum_{k=1}^K E[d_k].$$

between units within a cluster or when the correlation is dependent upon the size of the cluster. However, Jung et al.^[17] illustrate through simulation studies that the adjustment method performs well even if the assumption of a common intracluster correlation is not met, as long as the intraclusters are only moderately different. Nonetheless, certain data situations where this assumption may not hold include lesion detection, where the number of lesions per subject can vary widely and the detection of a lesion can depend on several factors, including its size and location.

To avoid the limitation inherent in the aforementioned analysis, other techniques would need to be implemented. For example, Rao and Scott^[6] have developed a method that avoids having to calculate an intracluster correlation and thus avoids the assumption of a constant correlation between units within cluster. Based on a sampling technique, an inflation factor is defined that accounts for clustered observations. This inflation factor is equal to the variance estimate of the ratio of positive responses,

$$\text{var}(\hat{p}_i) = \frac{K_i}{K_i - 1} \left(\frac{\sum_{k=1}^K (y_{ik} - n_{ik} \hat{p}_i)^2}{\left(\sum_{k=1}^K n_{ik} \right)^2} \right)$$

divided by the estimated binomial variance of the proportion of positive responses, $n_i^{-1} \hat{p}_i (1 - \hat{p}_i)$. Rao and Scott refer to the inflation factor as the “design effect” and apply it to the standard chi-square test with $I - 1$ degrees of freedom,

$$\tilde{\chi}^2 = \sum_{i=1}^I \frac{(\tilde{y}_i \tilde{p})^2}{[\tilde{n}_i \tilde{p}(1 - \tilde{p})]}$$

The number of positive responses (y_i), number of units (n_i) and event rate (p) in the chi-square equation are each adjusted by the design effect and applied to the one degree of freedom chi-square statistic. The design effect approach is best utilized when the mean cluster size differs between comparison groups. The limiting factor of this *sampling*

approach is that a relatively large number of clusters is suggested in order to obtain a consistent ratio estimator.

Following a similar approach, Obuchowski^[8] has developed a test for clustered matched-pair data that pools the probability of a success across all clusters and all tests, where p in the above equation is replaced with $\bar{p}$, which is the average of the proportions of success for each procedure: $\bar{p} = (\hat{p}_i + \hat{p}_i') / 2$. The homogeneity hypothesis and the numerator of the test statistic remain the same as in the McNemar test, except that the overall sample proportion of positive units detected by procedure i ($\hat{p}_i$) is estimated by counting the number of positive responses by procedure i summed over all K clusters, e.g., $\sum (a_k + b_k)$ for Procedure 1 and $\sum (a_k + c_k)$ for Procedure 2. The test is asymptotically distributed as a chi-square with one degree of freedom under the null hypothesis,

$$\chi_0^2 = \frac{(\hat{p}_i - p_i')^2}{\text{var}(p_i - p_i') \bar{p}}$$

where

$$\text{var}(\hat{p}_i - \hat{p}_i')_{\bar{p}} = \text{var}(\hat{p}_i)_{\bar{p}} - 2 \text{cov}(\hat{p}_i, \hat{p}_i')_{\bar{p}}.$$

Obuchowski compares this method to the ICC adjustment. Under specific conditions (i.e., a cluster size ≤ 5 , three different correlation structures, and various sample sizes), the proposed method performs similar to the ICC in terms of maintaining a type I error rate below 0.05. The main difference between the two procedures is that the sampling approach does not take into account within-cluster correlation.

An alternative approach proposed by Durkalski and colleagues utilizes a method of moments approach for the analysis of clustered matched-pair data.^[9] The proposed test statistic is

$$\chi_V^2 = \frac{\left[\sum_{k=1}^K (1/n_k)(b_k - c_k) \right]^2}{\sum_{k=1}^K [(b_k - c_k)/n_k]^2}$$

because under the null hypothesis, the estimated variance of the difference in success rates is

$$\text{var}(\hat{p}_{10} - \hat{p}_{01}) = \frac{1}{K^2} \sum_{k=1}^K \left[\frac{(b_k - c_k)}{n_k} \right]^2$$

This estimate of the variance is consistent according to large sample theory. Based on simulations, the method of moments approach performs as well in terms of size and power as the McNemar test adjusted by the ICC and Obuchowski's approach for clustered matched-pair data. Moreover, this method has been extended to account for non-inferiority study designs.^[18]

GENERAL ESTIMATING EQUATIONS

Although adjusting the chi-square test for the equality of probabilities is popular, more complex analyses that incorporate covariate effects may also be of interest. These analyses require sophisticated models that account for the correlation within clusters. A popular method due to the availability of statistical software is the application of generalized linear models.^[19] Based on this model, marginal regression modeling using GEEs was developed as a semiparametric approach to fitting logistic regression models for binary clustered data.^[11] The logistic regression model using the GEE estimating function obtains similar results as the adjusted chi-square tests previously discussed in this entry when covariates are not present. This model is defined as $\log \text{it } \text{pr}(\gamma_{ijk} = 1) = \beta_0 + \beta_1 x_{ijk}$ where x_{ijk} identifies the i th treatment variable from the j th unit of the k th cluster (in the case of two treatments) and can be run using PROC GENMOD in SAS.^[20] The p -value is generated from a comparison of the log odds ratio to the estimated variance. Because clustering affects the standard errors of the parameter estimates rather than the parameter estimates themselves, the parameter estimates can be obtained by running a regression analysis on each response. The logit link expresses the linear relationship between a cluster's responses and the corresponding covariates. The correlation within a cluster is accounted for in the variance-covariance matrix. To estimate the variance of the response, a specified "working" correlation matrix that defines the association among units within a cluster substitutes for the true, often unknown, correlation matrix. Although the working correlation matrix may be misspecified, which could possibly result in a decrease of efficiency, the GEE method can produce consistent estimates.^[11] Furthermore, the decreased efficiency may be minimal if the number of clusters is large.^[21] Prentice^[22] extends Zeger and Liang's model to incorporate the modeling of correlations within a cluster in order to improve upon the efficiency of the GEE estimates of the regression parameters. Although the GEE method is more complex

than the straightforward comparison of event rates, the attractiveness of this approach is that it easily incorporates adjustments for covariates when needed. Methods for binary regression using clustered data deserve an entry of their own and are not discussed further.

EXAMPLE

To illustrate the clinical application of commonly used methods for analyzing clustered binary data and the importance of accounting for the clustered nature of the data, Donner and Klar^[23] consider data collected from schools randomly allocated to one of two interventions. The outcome of interest is the proportion of children who use smokeless tobacco after two years of follow up. The performance of various methods for testing the equality of clustered binary data is observed, including the test statistics and p -values of the unadjusted Pearson chi-square, the ICC inflation factor, the design effect inflation factor, and the GEE approach. The test statistics and p -values illustrate that statistical conclusions can be false when clustering is ignored (using the standard chi-square test), while all methods that adjust for clustering give relatively the same conclusions with some variability in the actual test values and significance levels.

For matched-pair data, a data set containing clustered matched-pair data that appeared in Obuchowski's paper^[8] is assessed with the unadjusted McNemar test, the ICC adjustment, pooled estimators, and the method of moments approach. A trial in diagnostic methods for hyperparathyroidism is designed to compare the sensitivity and specificity of two techniques, positron emission tomography (PET) and a single photon emission CT (SPECT) scan. The data consist of 21 patients whose glands were examined for the presence of hyperparathyroidism (Table 3). Of the 21 patients, a total of 72 glands were evaluated by both diagnostic tools. The specificity of the two scanners is considered to be of interest and, therefore, only the glands that are confirmed negative by surgery (considered the gold standard) are evaluated. Of the 72 glands among the 21 patients, a total of 51 glands were confirmed negative. The estimated intraclass correlation is equal to 0.46. Obuchowski's, Donner's, and Durkalski's test results are chi-square values of 2.86 ($p = 0.091$), 3.66 ($p = 0.056$), and 2.32 ($p = 0.128$), respectively. The unadjusted McNemar test statistic is 4.5 ($p = 0.034$).

All the tests that account for clustering fail to reject the null hypothesis of no difference between the two procedures (at the 0.05 level), whereas the McNemar test (which does not account for the clustering) concludes that a statistically significant difference does exist (at the 0.05 level) in the specificity of PET and SPECT. This is not surprising, because as seen in Donner and Klar's comparisons, accounting for clustering means assigning less weight to multiple observations from the same cluster. Without doing

Table 3 Hyperthyroid Data

K	n_k	y_{ik}	y'_{ik}	a_k	b_k	c_k	d_k
1	3	0	2	0	0	2	1
2	3	2	3	2	0	1	0
3	3	3	3	3	0	0	0
4	1	1	1	1	0	0	0
5	3	2	3	2	0	1	0
6	4	4	4	4	0	0	0
7	3	3	3	3	0	0	0
8	2	2	2	2	0	0	0
9	2	2	1	1	1	0	0
10	1	1	1	1	0	0	0
11	3	2	2	2	0	0	1
12	2	2	2	2	0	0	0
13	3	3	3	3	0	0	0
14	2	2	2	2	0	0	0
15	2	0	2	0	0	2	0
16	3	2	2	2	0	0	1
17	3	2	2	2	0	0	1
18	3	2	3	2	0	1	0
19	2	2	2	2	0	0	0
20	1	1	1	1	0	0	0
21	2	2	2	2	0	0	0

$$K = 21, N = \sum_{k=1}^K n_k = 51, y_{ik} = \sum_{j=1}^{n_k} y_{ijk} = 40, y'_{ik} = \sum_{j=1}^{n_k} y'_{ijk} = 46,$$

$$a = \sum_{k=1}^K \sum_{j=1}^{n_k} a_{jk} = 39,$$

$$b = \sum_{k=1}^K \sum_{j=1}^{n_k} b_{jk} = 1,$$

$$c = \sum_{k=1}^K \sum_{j=1}^{n_k} c_{jk} = 7,$$

$$d = \sum_{k=1}^K \sum_{j=1}^{n_k} d_{jk} = 4. \text{ (From Ref. [8].)}$$

this, these multiple observations would have more influence than they perhaps should, and this could easily lead to pseudopower, or a false rejection of a true null hypothesis.

CONCLUSIONS

Clustered binary data are prevalent in medical research. Ophthalmology studies, dental research, oncology studies, and community research programs often involve multiple layers, which violate the basic assumption of independence for common statistical methods. Methods for analyzing clustered binary data are available and continue to be developed and enhanced. All methods account for the within-cluster correlation; yet, it is the approach to doing so that makes them different. The appropriate strategy is dependent on the research question being explored and the data structure in terms of cluster size and cluster covariates.

Due to the availability of current statistical software, the majority of methods discussed in this entry are simple to implement in practice. Therefore, the collapsing of the data to a single-level model should be avoided. Ignoring the cluster effect or collapsing of data has been shown throughout the literature to create biased estimates, which can lead to false statistical conclusions (as was the case in our examples). If a simple test of the equality of event rates is being performed and a relatively small difference in mean cluster size between comparison groups is present, then the ICC approach is convenient to implement and can be extended to a number of data scenarios, including stratified or matched-pair data.

For more complex study designs, such as those that incorporate covariates or multiple cluster levels (perhaps classrooms within school districts or designs that have a combination of clustered and longitudinal data), hierarchical modeling approaches are available and continue to be explored.^[1,3] It is worthwhile to mention that during the planning stages of a study that involves clustered binary data, the statistician needs to consider inflating the sample size to offset the loss of information that occurs due to the clustering. Considerations for sample size estimation for clustered binary data are available in the literature.^[24,25]

ACKNOWLEDGMENT

The author would like to thank Dr. Vance Berger for his constructive comments, which have added to this entry.

REFERENCES

1. Neuhaus, J. Assessing change with longitudinal and clustered binary data. *Annu. Rev. Public Health* **2001**, 22, 115–128.
2. Ashby, M.; Neuhaus, J.M.; Hauck, W.W.; Bacchetti, P.; Heilbron, D.C.; Jewell, N.P.; Segal, M.R.; Fusaro, R.E. An annotated bibliography of method for analyzing correlated categorical data. *Stat. Med.* **1992**, 11, 67–99.
3. Pendergast, J.F.; Gange, S.J.; Newton, M.A.; Lindstrom, M.J.; Palta, M.; Fisher, M.R. A survey of methods for analyzing clustered binary response data. *Int. Stat. Rev.* **1996**, 64 (1), 89–118.
4. Hujoel, P.P.; Moulton, L.H.; Loesche, W.J. Estimation of sensitivity and specificity of site-specific diagnostic tests. *J. Periodontal Res.* **1990**, 25, 193–196.
5. Donner, A. Statistical methodology for paired cluster designs. *Am. J. Epidemiol.* **1987**, 126 (5), 972–979.
6. Rao, J.N.K.; Scott, A.J. A simple method for the analysis of clustered binary data. *Biometrics* **1992**, 48, 577–585.
7. Lee, E.W.; Dubin, N. Estimation and sample size considerations for clustered binary responses. *Stat. Med.* **1994**, 13, 1241–1252.
8. Obuchowski, N.A. On the comparison of correlated proportions for clustered data. *Stat. Med.* **1998**, 17, 1495–1507.
9. Durkalski, V.; Palesch, Y.; Lipsitz, S.; Rust, P. The analysis of clustered matched-pair data. *Stat. Med.* **2003**, 22, 2417–2428.

10. Donner, A. The analysis of intraclass correlation in multiple samples. *Ann. Hum. Genet.* **1985**, *49*, 75–82.
11. Zeger, S.L.; Liang, K. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **1986**, *42*, 121–130.
12. Berger, V.W. Pros and cons of permutation tests in clinical trials. *Stat. Med.* **2000**, *19*, 1319–1328.
13. Donner, A.; Donald, A. The statistical analysis of multiple binary measurements. *J. Clin. Epidemiol.* **1988**, *41* (9), 899–905.
14. Donner, A.; Banting, D. Adjustment of frequently used chi-square procedures for the effect of site-to-site dependencies in the analysis of dental data. *J. Dent. Res.* **1989**, *68* (9), 1350–1354.
15. Donald, A.; Donner, A. Adjustments to the Mantel-Haenszel chi-square statistic and odds ratio variance estimator when the data are clustered. *Stat. Med.* **1987**, *6*, 491–499.
16. Donner, A.; Eliasziw, M. Application of matched pair procedures to site-specific data in periodontal research. *J. Clin. Periodontol.* **1991**, *18*, 755–759.
17. Jung, S.; Ahn, C.; Donner, A. Evaluation of an adjusted chi-square statistic as applied to observational studies involving clustered binary data. *Stat. Med.* **2001**, *20*, 2149–2161.
18. Durkalski, V.; Palesch, Y.; Lipsitz, S.; Rust, P. The analysis of clustered matched-pair data under a non-inferiority study design. *Stat. Med.* **2003**, *22*, 279–290.
19. Nelder, J.A.; Wedderburn, R.W.M. Generalized linear models. *J. R. Stat. Soc. A* **1972**, *135*, 370–384.
20. SAS Statistical Software V8 or V9, Cary, North Carolina.
21. Ahn, C. Statistical methods for the estimation of sensitivity and specificity of site-specific diagnostic tests. *J. Periodont. Res.* **1997**, *32*, 351–354.
22. Prentice, R.L. Correlated binary regression with covariates specific to each binary observation. *Biometrics* **1988**, *44*, 1033–1048.
23. Donner, A.; Klar, N. Analysis of binary outcomes. *Cluster Randomization Trials*; Arnold, Hodder Headline Group: London, 2000, 86–95.
24. Lee, E.W.; Dubin, N. Estimation and sample size considerations for clustered binary responses. *Stat. Med.* **1994**, *13*, 1241–1252.
25. Jung, S.H.; Kang, S.H.; Ahn, C. Sample size calculations for clustered binary data. *Stat. Med.* **2001**, *20*, 1971–1982.

Analysis of Clustered Categorical Data

Kalyan Das

Department of Statistics, University of Calcutta, Kolkata, India

Abstract

This entry discusses models in the bivariate setup. It also discusses the limitations considering the normality assumption of the subject specific effects and proposes some nonparametric distribution instead.

INTRODUCTION

Social, behavioral, and epidemiological studies are often concerned with multiple discrete indicators that are utilized as responses in lieu of an obvious single measure for an outcome of interest. Sometimes these multiple indicators corresponding to a subject are each rated with natural ordering. For example, in order to study the efficacy of a new drug, patients who received a certain dose of that drug are observed at weekly intervals. Their responses are recorded in terms of multiple indicators and an obvious ordering in terms of patient's improvement may be seen. The ordered responses of the patient (within a cluster) are likely to be positively correlated. Repeated measurements instead of different time points may refer to matched sets. In statewide studies of cancer screening in primary care practices, interest is in the physician and patient factors that predict the extent of breast cancer screening that a woman receives. The outcome for each woman takes a value of 0, 1, or 2 for the number of types of screening procedures (mammogram and clinical breast examination) observed over successive periods in the past year. The statistical analysis should consider that the level of one woman's cancer screening may be correlated to that of another patient in the same medical practice. In a randomized clinical trial on survey for duodenal ulcer, the postoperative pain level (none/slight/moderate/severe) and medication requirements (never/seldom/occasionally/regularly) of each patient are observed together over days. The objective may be to study the association between these two as related to different types of operation for duodenal ulcer (VP—vagotomy and drainage procedure, VA—vagotomy and distal antrectomy, VA—vagotomy and hemigastrectomy, RA—resection alone). The prime concern, therefore, is to find a suitable model to describe the relationship between the paired ordinal responses and the available covariates. In oral hygiene studies, plaque formed on six well-defined teeth of each subject may be measured as per Quigley-Hein index for plaque scoring. The primary objective may be to study the role of factors—such as smoking

habits, food habits, and efficacy of toothbrushes—that predict the extent of formation of plaque. Once again, those six teeth in an individual's mouth are all intradependent. Prior to the advent of increasingly powerful computing, statistical analysis was mostly limited to large sample χ^2 methods based on contingency tables. In recent years, there has been growing interest in exact inference. As a result, estimation and hypothesis testing for tables with small samples can be made efficiently by using exact methods with the help of advanced computing techniques. Although the non-model-based Mantel-Haenszel's method is quite useful when the basic requirements of sample size are met, variations of this method that adjusts for continuous covariates are yet to be studied extensively. This review primarily focuses on the second approach. Due to enormity of the subject, we do not attempt an exhaustive review of the literature but often refer to other books or articles that document many contributions to this field.

In many biological studies, categorical responses are observed from highly stratified random samples with a large number of strata (clusters) each consisting of a few observations. Our primary objective is to analyze data under this kind of setting. Marginal models analyze effects averaged over all clusters at particular levels of predictors. Alternatively, one can proceed through cluster specific (conditional) models using random effects that describe effects at cluster level.

MODELS FOR CLUSTERED CATEGORICAL DATA

Approaches based on logistic regression models were developed four decades ago to describe the relationship between a categorical response variable and a set of explanatory variables that may include both categorical and continuous variables. We emphasize here unconditional logistic regression mainly for its application in life data analysis. Models for ordinal responses have received much attention and since 1960, there have been several proposals (Smell,^[1]

Bock and Jones,^[2] McCullagh,^[3] Goodman.^[4] The proportional odds model^[3] specifies a cumulative logit link to relate p covariates to (a r category) ordinal response with corresponding multinomial probabilities $\pi_1, \dots, \pi_r$ such that $\sum_{j=1}^r \pi_j = 1$. The model is then

$$\text{logit } P(y_t \leq j | x_t) = \alpha_j - x_t' \beta_j, j = 1, \dots, r \quad (1)$$

where y_t denotes an individual's t th response, $t = 1, \dots, T$, which can take one of r possible categories, $j = 1, \dots, r$. x_t consists of the values of the explanatory variables corresponding to the t th observation in a cluster. The parameters $\{\alpha_j\}$ called cut points are usually nuisance parameters whereas the parameters $\{\beta_j\}$ have their usual interpretations.

Proportional Odds Mixed Models

Ordinal outcomes are observed frequently in many field of investigations such as small area surveys, longitudinal studies, disease mapping, and others. Typically in those situations, data observed across time or locations form some sort of clustering and therefore become correlated within themselves. This causes additional variability that needs to be taken into account in order to draw appropriate inference from the data. To do that, it is natural to introduce an unobserved random component in the linear predictor of the model defined in Eq. 1. But a major difficulty in these models is that a full likelihood analysis often involves intractable numerical integration with respect to mixing distribution of random effects. Usually, two classes of models are considered: marginal models or population average models, where effects are averaged over clusters at particular levels of predictors; or subject specific (cluster specific) models, which are conditional models that describe effects at each cluster level. The basic purpose of this review is to highlight the recent methodological developments on subject specific models. Essentially, one can extend the proportional odds model to a subject specific model (see Tutz and Hennevoogt^[5]) as

$$\text{logit } P(y_{it} \leq j | x_{it}, u_i) = \alpha_j - x_{it}' \beta + z_{it}' u_i, \quad (2)$$

$i = 1, \dots, n, t = 1, \dots, T, j = 1, \dots, c$

where $(y_{i1}, y_{i2}, \dots, y_{iT})$ denote T repeated responses in the i th cluster (subject), u_i denotes the subject (cluster) specific random effect with mean 0 and variance Σ . The parameters $\{\alpha_j\}$ called cut points are usually nuisance parameters. Specifically, this model implies that odds ratios for describing effects of explanatory variables as the response variable are same for each possible ways of collapsing a “c” category response to a binary variable (McCullagh^[3]). In the binary set up the model in Eq. 2 reduces to

$$\text{logit } P(y_{it} = 1 | x_{it}, u_i) = x_{it}' \beta + z_{it}' u_i.$$

Latent Mixed Models

A clustered ordinal model can be developed through threshold concept. In such models, conventionally one assumes that there is an unobserved latent variable (y^*) that is related to the actual response though the introduction of a threshold variable. In fact for the ordinal model a series of threshold values $\alpha_1, \dots, \alpha_{J-1}$ where J equals the number of ordered categories $\alpha_0 = -\infty, \alpha_J = \infty$. Here a response occurs in category j ($y_{it} = j$) if the latent response process y^* exceeds the threshold value α_{j-1} but does not exceed the threshold value α_j . Thus for the categorical variable “ y ” obtained by chopping y^* into categories (with cut point $\{\alpha_j\}$) the proportional odds model holds for predictor with effects proportional to those in the continuous model. One can therefore, write the model

$$\begin{aligned} P(y_{it} = j | \beta, \alpha) &= \Phi \left[\left(\frac{\gamma_j - x_{it}' \beta - u_i}{\sigma} \right) \right] - \Phi \left[\left(\frac{\gamma_{j-1} - x_{it}' \beta - u_i}{\sigma} \right) \right] \\ \text{or, } &= \bar{\Psi} \left[\left(\frac{\gamma_j - x_{it}' \beta - u_i}{\sigma} \right) \right] - \bar{\Psi} \left[\left(\frac{\gamma_{j-1} - x_{it}' \beta - u_i}{\sigma} \right) \right] \end{aligned} \quad (3)$$

where $\Phi(\cdot)$ and $\bar{\Psi}(\cdot)$ represent the cumulative standard normal distribution function and the logistic function, respectively. For convenience, we take $\gamma_1 = 0$ and $\sigma = 1$ ($\sigma = \pi^2/3$ for logistic model).

Note that alternatively we can consider a log linear model (Goldstein^[6]) for clustered categorical data that can include multiple random effects but such a model corresponds to an ordinal regression model only when the number of categories is 2 (see McCullagh^[3])

Multilevel Models

Model in Eq. 3 can be viewed as a two-level model where i denotes the level-2 units (cluster in the clustered data context or subjects in the longitudinal data context) and t denotes the level-1 units (subjects in the clustered data context or repeated observations in the longitudinal data context). Thus the random effect regression model for the latent response y_{it}^* is

$$y_{it}^* = x_{it}' \beta + z_{it}' u_i + \varepsilon_{it} \quad (4)$$

where the conjugate vector x_{it} and random effects design vector z_{it} are associated with the t th level-1 unit nested within level-2 unit i and ε_{it} are residuals. Hedeker and Gibbons^[7] and Liu and Hedeker^[8] considered multi-level representation of ordinal response model in a more extended way.

Multivariate Clustered Ordinal Model

Marginal Models

We discuss a multivariate ordinal categorical model accounting for the heterogeneity. This kind of models has received much attention lately. Here emphasis is on the efficient estimation of the effect of covariates on the marginal distribution of ordinal responses. Such a model without any random term has been proposed earlier (Molenberghs and Lessaffre^[9], Dale^[10]). In view of its elegant statistical properties such a model appears to be a good candidate for modeling categorical dependent variables. This model is quite flexible as the marginal and association structures can be modeled easily. Furthermore, here the odds ratio is easier to interpret than a multivariate probit model. The details of the model construction have been addressed in Molenbergh and Lessaffre.^[9] In fact, in particular, for the bivariate model we may consider the following model for subject i as

$$\begin{aligned}\text{logit } P(y_{1it} \leq j_1 | x_{1t}) &= \alpha_{1j_1} - x'_{1it} \beta_1 + z'_{1it} u_{1i} \\ \text{logit } P(y_{2it} \leq j_2 | x_{2t}) &= \alpha_{1j_2} - x'_{2it} \beta_2 + z'_{2it} u_{2i}\end{aligned}$$

and

$$\log \psi(j_1, j_2, x) = x'_{3it} \beta_3 + u_{3i}$$

where

$$\psi(j_1, j_2) = \frac{P(y_{1it} \leq j_1, y_{2it} \leq j_2 | x) P(y_{1it} \leq j_1, y_{2it} \leq j_2 | x)}{P(y_{1it} > j_1, y_{2it} \leq j_2 | x) P(y_{1it} > j_1, y_{2it} > j_2 | x)}$$

is called (global) cross ratio and u_{ki} , $k = 1, 2, 3$ are the subject specific random effects. In particular when $j_1, j_2 = 0, 1$, the model reduces to

$$\text{logit } P(y_{1it} = 1 | x_{1t}) = x'_{1it} \beta_1 + u_{1it} \quad (5)$$

$$\text{logit } P(y_{2it} = 1 | x_{2t}) = x'_{2it} \beta_2 + u_{2it} \quad (6)$$

$$\text{and } \log \psi_{it}(x) = x'_{3it} \beta_3 + u_{3it} \quad (7)$$

where $\psi(x)$ given by $\log \left(\frac{p_{00}(x)p_{11}(x)}{p_{01}(x)p_{10}(x)} \right)$ is linear in covariances and

$$p_{j_1 j_2}(x) = P(y_{1it} = j_1, y_{2it} = j_2), \quad j_1, j_2 = 0, 1.$$

The model in Eqs. 5 and 6 combined with Eq. 7 accounts for three types of associations. Firstly the association between two responses of the t th unit in the i th cluster is modeled directly through odds ratio. Secondly, the association between responses of two different units in the

same cluster occurs through the components of the vector $u_i = (u_{1i}, u_{2i}, u_{3i})'$. In the context of repeated measurements or longitudinal data. This is basically the longitudinal association. Finally the cross-sectional association between one response in one unit and other response in another unit is accommodated through u_{3i} considered in Eq. 7.

Latent Mixed Models

The primary difficulty in extending the univariate ordinal category set up is to write a flexible statistical model. In a multivariate ordinal model, responses are observed from several characteristics of a subject. Such a mixture model accommodates correlation and extravariation in the data through the involvement of a random term. Unfortunately a full likelihood analysis in such a mixture model is often hampered by the need for numerical integration. Alternatively, one can proceed through extending the idea of a latent variable-based approach to multivariate set up where a latent vector variable can be modeled. To illustrate such a model, we consider an oral hygiene study where the interest may be to see the cleaning efficacy of different toothbrushes commercially available in the market. The objective may also be to study the plaque-removing efficiency of different toothbrushes as related to factors like smoking habits, food habits, etc. Plaque formation is an important causative factor in dental caries and periodontal disease. Reduction of plaque before and after brushing is measured according to the Tureskey et al. modification, and for every subject, measurements are recorded for selected teeth which form a cluster. Further repeated measurements on each tooth of a subject may be observed across time or on application of different treatments. One can then think of a polytomous random vector $y_{ij} = (y_{ij1}, \dots, y_{ijk}, \dots, y_{ijq})$ where y_{ijk} denotes the k th categorical response of the i th subject when j th treatment is applied. Accordingly, a latent random vector $y_{ij}^* = (y_{ij1}^*, \dots, y_{ijk}^*, \dots, y_{ijq}^*)$ can be seen to act in the following way

$$\begin{aligned}y_{ij1} &= m_1 & \text{if } \alpha_{1,m_1-1} \leq y_{ij1}^* \leq \alpha_{1,m_1} \\ y_{ijq} &= m_q & \text{if } \alpha_{q,m_q-1} \leq y_{ijq}^* \leq \alpha_{q,m_q}\end{aligned} \quad (8)$$

where $1 \leq m_k \leq r_k$, $k = 1, \dots, q$ indicating that there are r_k categories for the k th variable y_{ijk} . We take $\alpha_{k,0} = -\infty$ and $\alpha_{k,r_k} = \infty$. As in the univariate setup, the latent vector is subject to a random effect model.

$$y_{ijk}^* = x'_{ijk} \beta + Z'_{ijk} u_i + \varepsilon_{ijk} \quad (9)$$

where u_i is a q -dimensional vector of random effects that take into account the cluster (subject) specific variation or longitudinal variation.

This model is basically a three-level model where i ($i = 1, \dots, n$) denote the level-3 subject, j denote the level-2 treatment for subject i and k denote the level-1 response for the i th subject in the j th occasion.

In the section “Methods for Clustered Categorical Data,” we will discuss methodologies that may be applied to draw inference from this bivariate dichotomous clustered (longitudinal) model.

METHODS FOR CLUSTERED CATEGORICAL DATA

We briefly describe here several methods for ordinal data that exhibit extra heterogeneity due to repeated measurements.

Likelihood Based Approach

A large class of model-based procedures arises from defining a model for the variables with clusters and making inferences based on maximum likelihood (ML) methods. Here one models the joint distribution by using random effects for clusters. The models have conditional interpretations with cluster (subject) specific effects. The repeated responses are typically assumed to be independent, given the random effect, but variability in the random effects induces a marginal nonnegative association between pair of responses after averaging over the random effects. Let y_i be the vector of pattern ordinal responses from subject i for all of the T occasions, then assuming independence of the responses conditional on the random effects, the conditional likelihood of any pattern y_i given u_i can be written from (2.2) as

$$L(y_i | u_i) = \prod_{t=1}^T \prod_{j=1}^C \left\{ P(y_{it} \leq j | x_{it}) - P(y_{it} \leq j-1 | x_{it}) \right\}^{d_{ij}} \quad (10)$$

where

$$d_{ij} = \begin{cases} 1 & \text{if } y_{it} = j \\ 0 & \text{otherwise.} \end{cases}$$

In fact Eq. 10 leads to the marginal likelihood as

$$L(y | \alpha, \beta) = \prod_{i=1}^n \int L(y_i | u_i) g(u_i) du_i.$$

One can either proceed through MCEM or can use Fisher scoring algorithm to obtain ML estimates. But the difficulty arises in the integration over the random effects. In order to tackle the numerical integration, required for obtaining the unconditional likelihood, several approximation inference procedures have been proposed. These include Laplace approximations of the integrated likelihood (Lin and Pierce,^[11] Solomon and Cox,^[12] Lin and Breslow^[13]), and penalized quasi likelihood (Breslow and Clayton,^[14] Schall^[15]).

In case more than one random component appears in the model, ML model fitting can be challenging. An early approach (Harville and Mee^[16]) used best linear unbiased prediction of parameters of an underlying continuous model. Because the random effects are unobserved, to obtain the likelihood function we construct the usual product of multinomials that would apply if they were known and then integrate out the random effects. Except in rare cases (see Crouchley,^[17] Tenhave^[18]) this integral does not have closed form and it is thus necessary to use some approximation for the likelihood function. We can then maximize the approximated likelihood using a variety of standard methods. For simple models like random intercept models, straightforward Gauss-Hermite quadrature is usually adequate (Hedeker and Gibbons^[7,19]). An adaptive version of Gauss-Hermite quadrature (e.g., Lin and Pierce,^[20] Pinheiro and Bates^[21]) uses the same weights and nodes for the finite sum as Gauss-Hermite quadrature, but to increase efficiency, it centers the nodes with respect to the mode of the function being integrated and scales them according to the estimated curvature at the node. For models with higher dimensional integrals, more feasible methods use Monte Carlo (MC) methods which use the randomly samples nodes to approximate integrals. Booth and Hobert^[22] proposed an automated MCEM algorithm that assesses the MC error in the current parameter estimates and increases the number of nodes if the error exceeds the change in the estimates from the previous iteration.

Bayesian Approach

There have been several studies on the Bayesian estimation of cell probabilities in contingency tables (Good,^[23] Fienberg and Holland,^[24,25] Agresti and Chung,^[26] Some Bayesian (Evans et al.,^[27] Kateri et al.^[28]) later stressed estimation of association parameters. In this review, we emphasize the Bayesian approach for analyzing ordinal response data. One of the most important contributions is by Johnson.^[29] In the cumulative link models, they considered beta prior distributions for the cumulative probabilities at several values of the explanatory variables. The hybrid Metropolis-Hasting's/Gibbs sampler they used for fitting takes into account the ordering restriction on the intercept $\{\alpha_j\}$ parameters without any difficulty.

Chipman and Hamada^[30] carried out Bayesian analysis for a cumulative probit model. Bayesian ordinal models with latent mixed regression have been addressed earlier by Albert and Chib,^[31] Johnson.^[32] Other literature on Bayesian analysis with ordinal responses include Bradlow and Zaslarsky,^[33] Biswas and Das,^[34] Cowles et al.,^[35] Ishwaran and Zarepour,^[36] Ishwaran and Gatsonis,^[37] Tan et al.,^[38] Rossi et al.^[39]

Bayesian clustered ordinal data analysis has received attention for the last few years. Chib and Grenberg^[40] considered a multivariate normal latent random vector to define

the categories of the discrete variables repeatedly observed over time or over clusters. Recently O'Brien and Dunson^[41] formulated a multivariate logistic distribution in a very elegant manner where the correlation parameters are directly involved. Webb and Forster^[42] proposed a model where the conditional posterior of distributions are standard and simulations are easy. Chen and Shao^[43] employed a scale mixture of multivariate normal links and established condition for propriety under noninformative priors. Tan et al.^[38] generalized the methodology in Qu and Tan^[44] to propose Bayesian hierarchical proportional odds model with a complex random effect structure for three levels data, allowing each cluster to have a different number of observations. They used a MCMC fitting method that avoids numerical integration to estimate the fixed effects and the correlation parameters.

GEE Approach

GEE provides a non-likelihood based approach for fitting models for repeated or clustered ordinal data. This method seems to be useful as often in ML fitting of the marginal model though it is sometimes quite difficult to maximize the log-likelihood function. The GEE methodology originally specified for marginal GLM, extends to cumulative logit models^[45] and cumulative probit models^[46] for repeated ordinal responses. Though computationally GEE approach is much simpler than the full likelihood based approach, one naturally loses information to some extent by just relying on the first two moments. In fact, it really loses significant amount of information unless the sample size is extremely large.

ANALYSIS OF CLUSTERED MULTIVARIATE ORDINAL MODELS

Analysis of Dale's Model for Binary Set Up

In the following two subsections "Classical Likelihood Based Analysis" and "Bayesian Analysis," we would consider analysis of Dale's likelihood base correlated binary model under the classical and Bayesian framework.

Classical Likelihood Based Analysis

Following the previous notations for correlated binary models in Eqs. 5–7 under the assumption that $u_i \sim N_{3T}(0, \Sigma)$ for all i , we write the likelihood as

$$\ell(\psi/\text{data}) \propto \prod_i \int \ell_{ic}(\psi) du_i \quad (11)$$

where the parameter vector ψ consists of $\beta = (\beta'_1, \beta'_2, \beta'_3)'$ and those nuisance parameters that appear in Σ . Further the complete data likelihood corresponding to subject i is

$$\ell_{ic}(\Psi/\text{data}) = \prod_{t=1}^T \prod_{r=0}^1 \prod_{m=0}^1 p_{rmit}^{\tilde{y}_{rmit}} \quad (12)$$

where

$$p_{1it} = \frac{a_{it} - \sqrt{a_{it}^2 + c_{it}}}{2(\xi_{it} - 1)} \quad \text{if } \xi_{it} \neq 1$$

$$= p_{1it} p_{2it} \quad \text{if } \xi_{it} = 1$$

$$a_{it} = 1 + (p_{1it} + p_{2it})(\xi_{it} - 1), c_{it} = -4\xi_{it}(\xi_{it} - 1)p_{1it}p_{2it}$$

and

$$z_{rmit} = y_{1it}^1 y_{2it}^m (1 - y_{1it})^{1-r} (1 - y_{2it})^{1-m}, \quad r, m = 0, 1.$$

In view of the difficulty in evaluating the integrals, one can proceed through Monte Carlo EM (MCEM) to estimate the parameters. Note that Tenhave and Morabia^[47] considered a particular setup where they introduced standard deviations as a coefficient corresponding to the random components u_1 , u_2 , and u_3 in Eqs. 5–7 and assumed b_{ik} 's ($k = 1, 2, 3$) to be standard normal deviate vector.

The model in Eqs. 5–7 is more general as the assumption of dispersion matrix helps to study the longitudinal dependence (see Naskar and Das^[48] for more discussion). Numerical studies in this regard have been carried out in the section "Illustrative Example." Normality assumption of the random effects, though a bit convenient mathematically, may not always hold. In fact such an assumption fails to accommodate multimodality, skewness, and heavy-tailed features of the distribution of random effects. A very general class of mixing could be considered by introducing nonparametric Dirichlet process for unknown distribution of the random effects. The resulting semiparametric correlated binary model, though computer intensive, is capable of providing more accurate inference when the heterogeneity in the data exhibits multimodality/skewness/heavy-tailed features.

Bayesian Analysis

For the correlated binary set up in Eqs. 5–7, a Bayesian semiparametric model can be formed by presenting the hierarchical priors for p -dimensional fixed effects β . Writing $\mathbf{d} = (d_1, \dots, d_p)'$ and $\sigma = (\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2)'$, we choose the priors.

$$\beta|\mathbf{d}, \sigma \sim N_p(\mathbf{d}, \text{diag}(\sigma)) \quad (13)$$

$$d_r|\sigma_b^2 \stackrel{iid}{\sim} N(0, \sigma_{ob}^2), \quad r = 1, 2, \dots, p$$

$$\sigma_r^2|s_1, s_2 \stackrel{iid}{\sim} \text{Gamma}(s_1, s_2), \quad r = 1, 2, \dots, p$$

where $\text{diag}(\sigma) = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$.

Further we assume

$$\mathbf{u}_i \stackrel{iid}{\sim} G, G \sim DP(\alpha G_0) \quad (14)$$

The unknown distribution G of $\mathbf{u}_i$ is assumed to have a Dirichlet process prior where the base measure G_0 can be thought of as a guess of the distribution of random effects and α as a measure of the strength of this belief.

The full Bayesian model in the present context is completed by the prior assumptions on α and G_0 . It is assumed that

$$\alpha | \gamma_1, \gamma_2 \sim \text{Gamma}(\gamma_1, \gamma_2), G_0 | \Sigma \sim N_{3T} \left(0, \Sigma \right)$$

where the baseline covariance matrix Σ is

$$\begin{aligned} \Sigma &= \text{Block diag} \left(\Sigma_1, \Sigma_2, \Sigma_3 \right), \\ \Sigma_k &= \sigma_{ok}^2 \left[(1 - \rho_k) I_T + \rho_k \mathbf{1}_T \mathbf{1}_T' \right], \quad k = 1, 2, 3. \\ \mathbf{1}_T &= (1, \dots, 1)' \end{aligned}$$

The full conditionals can be derived as follows

$$\beta | \mathbf{d}, \sigma, \mathbf{u}, \text{data} \propto \ell_c \times \text{Diag}(\sigma) \quad (15)$$

where $\ell_c = \prod_{i=1}^n \ell_{ic}$, ℓ_{ic} is given in Eq. 12

$$\mathbf{d} | \beta, \sigma \sim N_p(\Delta^* \text{Diag}(\sigma^{-2}) \beta, \Delta^*) \quad (16)$$

where $\Delta^* = (\text{Diag}(\sigma^{-2}) + \sigma_{ou}^{-2} I)^{-1}$, $\text{Diag}(\sigma^{-2}) = \text{diag}(\sigma_1^{-2}, \dots, \sigma_p^{-2})$

Further, for $\ell = 1, \dots, p$.

$$\sigma_\ell^{-2} | \beta, \mathbf{d}, \stackrel{iid}{\sim} \text{Gamma} \left(s_1 + \frac{1}{2}, s_2 + \frac{(\beta_\ell - d_\ell)^2}{2} \right) \quad (17)$$

and marginalizing over G , the predictive distribution of $\mathbf{b}$ is

$$\begin{aligned} \mathbf{u} | \beta, \alpha, \theta &\propto \prod_{i=1}^n \ell_{ic} \\ &\times \prod_{i=1}^n \left\{ \frac{\alpha}{\alpha + i - 1} g_0(\mathbf{u}_i | \theta + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \gamma_{ji} \delta_{uj}(\mathbf{u}_i)) \right\} \end{aligned} \quad (18)$$

where $g_0(\cdot | \theta)$ is the baseline density indexed by the parameter (vector) θ , γ_{ji} is the number of $\mathbf{u}_{j'}$, $j' = 1, \dots, i-1$

such that $\mathbf{u}_i = \mathbf{u}_{i'}$ and $\delta_x(\cdot)$ is a function concentrated at x . Carrying out a component-wise MH algorithm within Gibbs sampler, we can generate samples from Eqs. (4.7)–(4.10) and hence make necessary posterior inferences.

Analysis of Latent Mixed Model

In view of Eqs. 8 and 8, conditional on $\mathbf{u}_i$, we compute

$$\begin{aligned} P^{u_i} &= P(y_{i1} = m_1, \dots, y_{iq} = m_q | \mathbf{u}_i) \\ &= P^{u_i}(\alpha_1, m_1 - 1 \leq y_{i1}^* < \alpha_{1,m_1}, \dots, \alpha_{q,m_q-1} \\ &\leq y_{iq}^* < \alpha_{q,m_q}) \\ &= \prod_{k=1}^q P^{u_i}(\alpha_{k,m_k-1} \leq y_{ik}^* < \alpha_{k,m_k}) \\ &= \prod_{k=1}^q P_k^{u_i}. \end{aligned}$$

Thus conditionally given $\mathbf{u}_i$, the likelihood is

$$\begin{aligned} \ell^{u_i} &= \prod_{t=1}^T \sum_{m_q=1}^{r_1} \dots \sum_{m_q=1}^{r_q} 1_{m_1, \dots, m_q}^{it} P^{u_i} \\ &= \prod_{t=1}^T \left[\sum_{s=1}^q \sum_{m_s=1}^{r_s} 1_{m_1, \dots, m_q}^{it} \prod_{k=1}^q P^{u_i} \right] \\ &= \prod_{t=1}^T \left[\sum_{s=1}^q \sum_{m_s=1}^{r_s} 1_{m_1, \dots, m_q}^{it} 1_{m_1, \dots, m_q}^{*it} \prod_{k=1}^q \phi(x'_{ik} \beta + Z'_{ik} \mathbf{u}_i, 1) \right] \end{aligned}$$

where ϕ is the normal pdf and

$$\begin{aligned} 1_{m_1, \dots, m_q}^{it} &= 1 \text{ if } y_{i1} = m_1, \dots, y_{iq} = m_q \\ &= 0 \text{ otherwise} \\ 1_{m_1, \dots, m_q}^{*it} &= 1 \text{ if } y_{i1}^* = m_1, \dots, y_{iq}^* = m_q \\ &= \text{otherwise.} \end{aligned}$$

In fact the unconditional likelihood is

$$L = \prod_{i=1}^n \int_{\mathbf{u}_i} \ell^{u_i} g(\mathbf{u}_i) d\mathbf{u}_i$$

where $g(\mathbf{u}_i)$ is the pdf of $\mathbf{u}_i$.

Because the likelihood involves high dimensional integration, no analytically tractable form can be obtained. One way to obtain the maximum likelihood estimate (MLE) is to proceed through multidimensional Gauss-Hermite Quadrature for integration (Liu and Hedeker^[8]). But that also creates difficulty in case of high dimension integration. Alternatively, we can proceed by using MCEM technique. In the section “Illustrative Example” we provide the estimates based on this technique.

Bayesian Analysis

Assume that the random component $u_i \sim N_q(0, \Sigma)$, $i = 1, \dots, n$. For the model Eqs. 8 and 9, the following prior is then assumed

$$\begin{aligned} \pi\left(\beta, \sum\right) &\propto \pi(\beta) \pi\left(\sum\right) \\ \beta &\sim N_p\left(0, \sum_0\right) = \text{diag}(\sigma_0^2) \\ \sigma^2 &\sim \phi_k\left(\sigma \mid \sigma_0, T_0^{-1}\right), \sigma \in \wp \end{aligned}$$

where $\sigma = (\sigma_{12}, \dots, \sigma_{q-1,q})'$ are the distinct $q/2(q-1)$ elements of Σ . Note that in order to avoid non-identifiability without loss in generality, we assume Σ to be a correlation matrix. Further we choose the prior for σ to be a truncated $N_k(\sigma_0, T_0^{-1})$ where the truncated region $\wp \subset S = [-1, 1]^{q(q-1)/2}$. One can then apply MH algorithm within the Gibbs sampler to generate observations from the full conditionals (see Das and Chattopadhyay^[49]). A general Bayesian semiparametric analysis could be done where u_i assumed to have some unknown distribution G . The following priors are considered.

$$\begin{aligned} \beta &\sim N_p\left(0, \sum_0\right), \quad G \sim DP(\alpha G_0) \\ G &\sim N_q\left(0, \sum\right), \quad \sum \sim W_q(\gamma, R) \end{aligned}$$

where G refers to a random distribution generated by a Dirichlet process with base distribution G_0 and total mass parameter α (known), γ and R , respectively, the degrees of freedom and known covariance matrix associated with the inverted Wishart prior. A key feature is that $u_i s'$ are marginally samples from G_0 and with positive probability some of the $u_i s'$ are identical. This is because of discreteness of the random measure G . The full conditionals for such model can be derived as follows.

$$\begin{aligned} \left[\beta \mid \sum, u, y^*, y\right] &\sim N_\phi\left(\hat{\beta}, (x'x)^{-1}\right) \\ \left[\sum \mid \beta, u, y^*, y\right] &\sim w_q(v + v_0, (R^{-1} + uu')^{-1}) \\ \left[y_{ik}^* \mid \beta, \sum, u, y_{-ik}^*\right] &\sim N(x_{ik}'\beta + u_{ik}, 1) \quad \alpha_{m_k-1} \leq y_{ik}^* < \alpha_{m_k} \\ \left[u_i \mid \beta, \sum, u_{-i}, y, y^*\right] &\propto \left(\sum_{j \neq i} e^{-\frac{1}{2}(y_i^* - x_i\beta)' V_i (y_i^* - x_i\beta)} \times \delta u_j(u_i)\right) \\ &\quad + |Q_i|^{-1/2} \left|\sum\right|^{-1/2} \int e^{-1/2(y_i^* - x_i\beta)' (y_i^* - x_i\beta)} du_i \\ &\quad \times g(u_i \mid 0, \sum) f(y_i^* \mid u_i, \beta, y, y_{-j}^*) \end{aligned}$$

where v_0 is the number of distinct $u_i s'$, $Q_i = (\Sigma^{-1} + Z_i'Z_i)^{-1}$, $V_i = Z_i Q_i Z_i' - I$ and u_{-i} , denote the random effects for

subjects excluding subject i . Note that $\delta_s(\cdot)$ is a degenerate distribution with mass point s and c is a constant.

ILLUSTRATIVE EXAMPLE

Example 1: Tibolone Data

To examine the performance and compare classical approach (both with and without random effects) and Bayesian approach for the correlated binary models, we consider a data on 200 women who have undergone Hormone Replacement Therapy. The primary objective in this study was to see whether tibolone is effective for reducing physiological problems like hot flashes (HF) and vaginal dryness (VD). The effectiveness of tibolone is, in fact, neglected through these pair of responses, collected as “mild” or “severe” corresponding to those two symptoms. The responses on these symptoms were observed at pre-treatment stage and after every 30 doses in a period of three months. The age and number of doses (NOD) considered to be covariates play important role on the effectiveness of tibolone. Empirical study revealed (see Naskar and Das^[48]) the following correlated binary model (CBM) conditional on the random components u_{kit} , $k = 1, 2, 3$

$$\begin{aligned} \text{logit}(p_{kit}) &= \beta_{k0} + \beta_{k1}(\text{age})_{it} + \beta_{k2}(\text{NoI})_{it} \\ &\quad + \beta_{k3}(\text{NoD})^2 + u_{kit}, \quad k = 1, 2 \end{aligned}$$

$$\text{log}(\xi_{it}) = \beta_{30} + \beta_{31}(\text{age})_{it} + \beta_{32}(\text{NoD})_{it} + u_{3it}.$$

Table 1 gives the result obtained by likelihood based MCEM approach in the classical paradigm and also the full Bayes Gibbs sampling approach in the Bayesian paradigm. In both the classical and Bayesian set up, we have considered DP mixing of CBM to account for all possible types of heterogeneity. For the Bayesian approach the posterior mean and standard errors are computed by generating observations from the full conditionals. One can also compute the 95% confidence interval or the predictive interval for the parameters of interest.

Simulation Study

An extensive investigation has been carried out to study the finite sample performances of the ML and Bayes estimates in CBM. Assume a single covariate x_{10it} for the first marginal model, x_{20it} , x_{21it} second marginal model and x_{30it} for the log association model. The covariates x_{20it} , x_{21it} and x_{30it} are generated from a standard normal variate and the covariate effect x_{21it} is generated from $U[-1, 1]$. To simulate two-level bivariate binary data, we take for $i = 1, \dots, 50$, $t = 1, \dots, 4$.

$$u_i \sim G, G \mid G_0 \sim DP(\alpha G_0), G_0 = N_{12}\left(0, \sum\right) \quad (19)$$

Table 1 Analysis of Tibolone Treatment Data

Effect	H F			V D			H F $\times$ V D		
	CBM	CBMM	DPMCMBM	CBM	CBMM	DPMCMBM	CBM	CBMM	DPMCMBM
Intercept	-1.331 (0.146)	-1.368 (0.182)	-1.397 (0.151)	-1.276 (0.143)	-1.420 (0.179)	-1.777 (0.166)	1.540 (0.241)	1.719 (0.285)	1.893 (0.327)
Age	-0.128 (0.096)	-0.130 (0.104)	-0.134 (0.099)	0.016 (0.091)	0.032 (0.114)	0.048 (0.105)	0.272 (0.211)	0.431 (0.268)	0.572 (0.287)
NoD	-1.268 (0.197)	-1.289 (0.213)	-1.364 (0.200)	-0.699 (0.158)	-0.706 (0.171)	-0.843 (0.178)	0.739 (0.248)	0.763 (0.317)	0.832 (0.335)
NoD ²	-0.040 (0.157)	0.042 (0.212)	0.046 (0.216)	0.603 (0.198)	0.984 (0.237)	1.369 (0.230)			

Table 2 Simulation Results—Mean and Standard Deviation (SD) of the Estimates

Parameter	True value	Mean $\pm$ SD ^a	Mean $\pm$ SD ^b	Mean $\pm$ SD ^c
β_{10}	1.0	0.757 $\pm$ 0.160	0.849 $\pm$ 0.122	0.952 $\pm$ 0.093
β_{20}	0.5	0.383 $\pm$ 0.074	0.419 $\pm$ 0.086	0.480 $\pm$ 0.059
β_{21}	1.0	0.767 $\pm$ 0.133	0.855 $\pm$ 0.177	0.965 $\pm$ 0.096
β_{30}	1.0	0.563 $\pm$ 0.213	0.598 $\pm$ 0.280	0.912 $\pm$ 0.196
α	0.5	—	—	0.612 $\pm$ 0.571
β_{10}	1.0	0.742 $\pm$ 0.130	0.873 $\pm$ 0.098	0.900 $\pm$ 0.092
β_{20}	0.5	0.373 $\pm$ 0.085	0.435 $\pm$ 0.085	0.482 $\pm$ 0.056
β_{21}	1.0	0.807 $\pm$ 0.092	0.867 $\pm$ 0.089	0.951 $\pm$ 0.081
β_{30}	1.0	0.601 $\pm$ 0.235	0.606 $\pm$ 0.235	0.885 $\pm$ 0.219
α	1.0	—	—	0.151 $\pm$ 0.769
β_{10}	1.0	0.737 $\pm$ 0.111	0.892 $\pm$ 0.094	0.910 $\pm$ 0.093
β_{20}	0.5	0.361 $\pm$ 0.082	0.4511 $\pm$ 0.092	0.470 $\pm$ 0.062
β_{21}	1.0	0.840 $\pm$ 0.103	0.875 $\pm$ 0.110	0.916 $\pm$ 0.086
β_{30}	1.0	0.570 $\pm$ 0.191	0.619 $\pm$ 0.191	0.891 $\pm$ 0.170
α	2.0	—	—	0.220 $\pm$ 0.707

^aValues obtained under ML method^bValues obtained under Bayesian approach with normality of random effect^cValues obtained under DP mixed Bayesian model

where

$$\Sigma = \text{block diag}(\Sigma_1, \Sigma_2, \Sigma_3), \Sigma_k = 2[.5I_4 + .51_41_4'] \\ k = 1, 2, 3.$$

We also take $\alpha = 1.0$. Then the bivariate binary data set is formed by first generating u_i 's from Eq. 19 according to Polya urn-based algorithm (Blackwell and McQueen^[50]) and then following models Eqs. 5–7 with $\beta_{10} = \beta_{30} = \beta_{21} = 1.0$ and $\beta_{20} = 0.5$. ML estimates obtained under this simulated data set by using MCEM are given in Table 2. In the Bayesian context we consider the prior hyperparameters $\sigma_0 = 10$, $s_1 = s_2 = 0.001$. To ensure regularity on the influence of small as well as large values of α we choose hyperparameters $\gamma_1 = \gamma_2 = 2.0$ in the gamma prior for α . Table 2 lists the estimates and their standard errors obtained by the likelihood approach and the Bayesian approach under DP mixing and simple normal mixing.

Oral Hygiene Data

In order to compare the cleaning efficacy of different designs of toothbrushes, a survey was conducted of 220 students and staff members from hospitals in and around Kolkata. A detailed history for each sample individual recorded a week prior to the beginning of the study to collect information like age, sex, occupation, food habit, smoking habit, and teeth-cleaning habit. The amount of plaque on the facial and lingual surfaces of 6 selected teeth for each individual was scored according to the Turesky et al.^[51] modification of the Quigley-Hein index. Each individual was given a total of six different designs of toothbrushes to be used during the four-month long clinical study (see Das and Chattopadhyay^[49] for details). Final data set essentially based on a cross over trial, provided the differences of pre- and post-brushing scores for 220 individuals corresponding to six different types of brushes. Table 3 compares the result

Table 3 Various Estimates of Bush Effects for Oral Hygiene Data

Effects	Estimated coefficients		
	ML	FB	DPB
Brush 1	0.423(0.962)	0.426(1.524)	0.541(1.404)
Brush 2	0.597(0.920)	0.672(0.640)	0.660(0.716)
Brush 3	1.179(0.883)	1.405(0.941)	1.208(0.842)
Brush 4	0.837(1.154)	0.805(1.063)	0.911(1.288)
Brush 5	0.109(0.643)	0.012(0.671)	0.142(0.761)
Brush 6	0.226(0.973)	0.368(1.152)	0.212(0.919)
Food habit	0.106(2.531)	0.112(3.432)	0.124(2.404)
Smoking habit	1.163(0.724)	1.166(0.550)	1.172(0.761)

ML—likelihood-based MCEM approach, FB—full Bayesian-Gibbs sampler, and DPB—Dirichlet process-based Bayesian approach.

obtained by likelihood based MCEM approach (where $u_i \sim N(0, 0.51 \pm 0.511')$ discussed in the previous section and by the full Bayes with hyperparameter approach ($\sum_0 = \sigma_0^2 I$, $\sigma_0 = 10$, and $T_0 = I$) and the DP mixed Hierarchical Bayes approach (with $v = 10$, $R = 0.1I + 0.911'$) discussed in the section “Bayesian Analysis.” The details of the computations have been demonstrated in Das and Chattopadhyay.^[49] The study revealed that the plaque formation rate is higher for smokers and nonvegetarians. Also, the cleaning efficacy of a certain brush is likely to be the best among all brands of brushes.

DISCUSSION

In view of the computational advances, it is now possible to encompass a much broader area of techniques for the analysis of categorical data.^[52–54] Though chi-square and unconditional logistic regression methods still comprise the staple diet for analysis by the health science researchers, recently developed methods for ordinal data analysis are receiving wider use. Furthermore, complex data sets involving cluster samples with discrete and continuous multilevel covariates are being analyzed these days by using sophisticated tools leading to exact inference. This survey describes recent enhancements in the area of clustered categorical data analysis and discusses several applications of the recent methodology. One possible direction would be to analyze multivariate repeated measurements (or clustered) data is to proceed through the latent vector approach. However, the association among the variable of a subject is not truly exposed. Alternatively, one can proceed by developing a multivariate association model. In this survey we discuss this kind of models in the bivariate setup. We also have discussed the limitations considering the normality assumption of the subject specific effects and proposed some nonparametric distribution instead. The analysis of clustered categorical data using the Bayesian

approach specification of prior distribution still remains a problem.

As emphasized in our discussion, we believe there is a great need for further studies in clustered data with multivariate responses both from analysis and design viewpoints. The analysis of multilevel models needs to be developed further. In the regression setup, it may be of interest to look into the properties of regression parameters, especially when the nonparametric distribution of the random effects are considered.

Most methodologies discussed here have appeared in the last few years and have yet to be disseminated among the health science/social science researchers. We hope that computational advances will continue expanding the scope of exact methods and that complex data sets will more commonly be analyzed by statistical methods that incorporate various patterns of depending in the data.

REFERENCES

1. Snell, E.J. A scaling procedure for ordered categorical data. *Biometrics* **1964**, 20, 592–607.
2. Bock, R.D.; Jones, L.V. *The Measurement and Prediction of Judgement and Choice*; Holden Day: San Francisco, CA, 1968.
3. McCullagh, P. Regression models for ordinal data (with discussion). *J. R. Stat. Soc. B* **1980**, 42, 109–142.
4. Goodman, L.A. The analysis of dependence in cross classifications having ordered categories, using log-linear models for frequencies and loglinear models for odds. *Biometrics* **1983**, 39, 149–160.
5. Tutz, G.; Hennevogl, W. Random effects in ordinal regression model. *Comput. Stat. Data Anal.* **1996**, 22, 537–557.
6. Goldstein, H. Nonlinear multilevel models with an application to discrete response data. *Biometrika* **1991**, 78, 45–51.
7. Hedeker, D.; Gibbons, R.D. A random effects ordinal regression model for multilevel analysis. *Biometrics* **1994**, 50, 933–944.
8. Liu, L.C.; Hedeker, D. A mixed effects regression model for longitudinal multivariate ordinal data. *Biometrics* **2006**, 62, 261–268.
9. Molenberghs, G.; Lessaffre, E. Marginal modeling of correlated ordinal data using plackett distribution. *J. Am. Stat. Assoc.* **1994**, 89, 633–644.
10. Dale, J. Global cross ratio models for bivariate discrete ordered responses. *Biometrics* **1986**, 41, 909–917.
11. Lin, Q.; Pierce, D.A. Heterogeneity in Mantel-Haenszel type models. *Biometrika* **1993**, 80, 543–556.
12. Solomon, P.J.; Cox, D.R. Nonlinear components of variance models. *Biometrika* **1992**, 79, 1–11.
13. Lin, X.; Breslow, N.E. Bias correction generalized linear mixed models with multiple components of dispersion. *J. Am. Stat. Assoc.* **1996**, 91, 1007–1016.
14. Breslow, N.; Clayton, D.G. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* **1993**, 88, 9–25.

15. Schall, R. Estimation in generalized linear models with random effects. *Biometrika* **1991**, *40*, 917–927.
16. Harville, D.A.; Mee, R.W. A mixed-model procedure for analyzing ordered categorical data. *Biometrics* **1984**, *40*, 393–408.
17. Crouchley, R. A random effects model for ordered categorical data. *J. Am. Stat. Assoc.* **1995**, *90*, 489–498.
18. Tarhame, T.R. A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics* **1996**, *52*, 472–491.
19. Hedeker, D.; Gibbons, R.D. MIXOR: A computer program for mixed-effects ordinal regression analysis. *Comp. Biomd.* **1996**, *49*, 157–176.
20. Liu, Q.; Pierce, D.A. A note on Gauss-Hermite quadrature. *Biometrika* **1994**, *81*, 624–629.
21. Pinhero, J.C.; Bates, D.M. Approximations to the loglikelihood functions in the nonlinear mixed effects model. *J. Comp. Graph. Stat.* **1995**, *4*, 12–35.
22. Booth, J.G.; Hobert, J.P. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. B* **1999**, *61*, 265–285.
23. Good, I.J. The estimation of probabilities: an essay on modern Bayesian methods; MIT Press: Cambridge, MA, 1965.
24. Fienberg, S.E.; Holland, P.W. On the choice of flattening constants for estimating multinomial probabilities. *J. Mul. Anal.* **1972**, *2*, 127–134.
25. Fienberg, S.E.; Holland, P.W. Simultaneous estimation of multinomial cell probabilities. *J. Am. Stat. Assoc.* **1973**, *68*, 683–690.
26. Agresti, A.; Chuang, C. Model based Bayesian methods for estimating cell proportions in cross classification tables having ordered categories. *Comp. Stat. Data Anal.* **1989**, *7*, 245–258.
27. Evans, M.; Gilula, Z.; Guttman, I. Computational issues in the Bayesian analysis of categorical data: Loglinear and Goodman's RC model. *Stat. Sin.* **1993**, *3*, 391–406.
28. Kateri, M.; Nicolaou, A.; Ntzoufras, I. Bayesian inference for the RC(m) association model. *J. Comp. Graph. Stat.* **2005**, *14*, 116–138.
29. Johnson, B.M. On the admissible estimators for certain fixed sample binomial problems. *Ann. Math. Stat.* **1971**, *42*, 1579–1587.
30. Chipman, H.A.; Hamada, M.S. Comments on "Follow-up designs to resolve confounding in multifactor experiments." *Techometrics* **1996**, *38*, 317–320.
31. Albert, J.H.; Chib, S. Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.* **1993**, *88*, 669–679.
32. Johnson, V.E. On Bayesian analysis of multirater ordinal data: an application to automated essay grading. *J. Am. Stat. Assoc.* **1996**, *91*, 42–51.
33. Bradley, E.T.; Zaslavsky, A.M. A hierarchical latent variable model for ordinal data from a customer satisfaction survey with "no answer" responses. *J. Am. Stat. Assoc.* **1999**, *94*, 45–52.
34. Biswas, A.; Das, K. A Bayesian analysis of bivariate ordinal data. Wisconsin epidemiologic study of diabetic retinopathy revisited. *Stat. Med.* **2002**, *21*, 549–559.
35. Cowles, M.K.; Carlin, B.P.; Connett, J.E. Bayesian tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness. *J. Am. Stat. Assoc.* **1996**, *91*, 86–98.
36. Ishwaran, H.; Zarepour, M. Markov chain Monte Carlo in approximate dirichlet and beta two parameter process hierarchical models. *Biometrika* **2000**, *87*, 371–390.
37. Ishwaran, H.; Gatsonis, C.A. A general form of hierarchical ordinal regression models with application to correlated ROC analysis. *Can. J. Stat.* **2000**, *28*, 731–750.
38. Tan, M.; Qu, Y.S.; Mascha, E.; Schubert, A. A Bayesian hierarchical model for multilevel repeated ordinal data: analysis of oral practice examinations in a large anaesthesiology training program. *Stat. Med.* **1999**, *18*, 1983–1992.
39. Rossi, P.E.; Gilula, Z.; Allenby, G.M. Overcoming scale usage heterogeneity: a Bayesian hierarchical approach. *J. Am. Stat. Assoc.* **2001**, *90*, 20–31.
40. Chib, S.; Greenberg, E. Analysis of multivariate probit models. *Biometrika* **1998**, *85*, 347–361.
41. O'Brien, S.M.; Dunson, D.B. Bayesian multivariate logistic regression. *Biometrics* **2004**, *60*, 739–746.
42. Webb, E.L.; Forster, J.J. Bayesian model determination for multivariate ordinal and binary data. Tech. Report 2004.
43. Chen, M.H.; Shao, Q.M. Properties of prior and posterior distributions for multivariate categorical response data models. *J. Mul. Anal.* **1999**, *71*, 277–296.
44. Qu, Y.; Tan, M. Analysis of clustered ordinal data with subclusters via a Bayesian hierarchical model. *Commun. Stat. Theory Meth.* **1998**, *27*, 1461–1475.
45. Lipsitz, S.R.; Kim, K.; Zhao, L. Analysis of repeated categorical data using generalized estimating equations. *Stat. Med.* **1994**, *13*, 1149–1163.
46. Toledano, A.V.; Gatsonis, C. Generalized estimating equations for ordinal categorical data: arbitrary pattern of missing responses and missingness in a key covariate. *Biometrics* **1999**, *55*, 488–496.
47. Ten Have, T.R.; Morabia, A. Mixed effects models with bivariate and univariate association parameters for longitudinal bivariate binary response data. *Biometrics* **1999**, *55*, 85–93.
48. Naskar, M.; Das, K.; Ibrahim, J.G. A semiparametric mixture model for analyzing clustered competing risks data. *Biometrics* **2005**, *61*, 729–737.
49. Das, K.; Chattopadhyay, A.K. An analysis of clustered categorical data - application in dental health. *Stat. Med.* **2004**, *23*, 2895–2910.
50. Blackwell, D.; Mac Queen, J. Ferguson distribution via polya win schemes. *Ann. Stat.* **1973**, *1*, 353–355.
51. Turesky, S.; Gilmore, J.D.; Glickman, I.R. Reduced plaque formation by the chlormethyl analogue of Vit.C. *J. Periodontol.* **1970**, *41*, 41–43.
52. Agresti, A. *Categorical Data Analysis*, 2nd Ed; Wiley: Hoboken, NJ, 2002.
53. Naskar, M.; Das, K. Semiparametric analysis of two level bivariate binary data. *Biometrics* **2006**, *62* (4), 1004–1013.
54. Tutz, G.; Hennevoogl, W. Random effects in ordinal regression models. *Comp. Stat. Data Anal.* **1996**, *11*, 275–295.

Analysis of Heritability

Brent D. Burch

U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Ian R. Harris

Southern Methodist University, Dallas, Texas, U.S.A.

Abstract

The heritability of a measurable characteristic is the most commonly used statistic for expressing the degree to which genetic material is transferred from parent to progeny. This entry reviews methods of analyzing heritability such as statistical modeling and point and interval estimation.

INTRODUCTION

Statistical methods applied to the field of quantitative genetics provide insight concerning the role that nature and nurture play in the development of life forms. Physical traits of organisms depend on genetic as well as environmental influences and quantifying these effects are major concerns in the study of plants, animals, and human beings. The heritability of a measurable characteristic is the most commonly used statistic for expressing the degree to which genetic material is transferred from parent to progeny. Heritability is defined in terms of the variation in the trait under study and relies on the partitioning of this variation into components that are attributable to genetic and environmental sources. Variance components in mixed linear models are used to measure how different sources contribute to the variation of a characteristic.

Statistical models serve as the basis for analyzing heritability. The analysis of heritability is accomplished by measuring the trait of interest, and other important variables, on a collection of subjects. In addition, the relationships among the subjects are recorded and used in the statistical model. These relationships provide information about the pedigree structure or lineage of the subjects. Point estimators, confidence intervals, and hypothesis tests are some of the statistical inferential procedures employed to analyze heritability.

HERITABILITY

The phenotype of an individual is a quantitative characteristic or trait that is the result of combined influences of genetic and nongenetic effects. In equation form,

$$P = G + E \quad (1)$$

where P is the phenotypic value, G is the genotypic value which represents the effect of the set of genes that contribute to the phenotypic value, and E is the environmental (nongenetic) effect on the phenotypic value. The genotypic value (G) is further partitioned into distinct genetic effects. Specifically,

$$G = A + D + I \quad (2)$$

where A is the additive genetic effect, D is the dominance genetic effect, and I is the epistatic or interaction genetic effect. The additive genetic effect (A) represents that part of G which is transferable from one generation to the next. In some applications, A is referred to as the breeding value of the individual because it is the primary factor in determining the resemblance between relatives. The dominance genetic effect (D) is a result of the dominance of alleles at a locus. In other words, D represents the interaction of alleles within a locus or within-locus interactions. The epistatic genetic effect (I) accounts for genotypic values that depend on more than one locus. Genes are said to be epistatic if the effects of the different loci are not independent. The overall genotypic value for an individual is determined by the sum of the effects of the individual loci.

Heritability is defined in terms of the variances associated with the genetic influences on the phenotypic value. If $\text{Var}(G)$ and $\text{Var}(P)$ are the variances of the genotypic and phenotypic values, respectively, then heritability in the broad sense is defined as

$$H^2 = \frac{\text{Var}(G)}{\text{Var}(P)} \quad (3)$$

H^2 is the proportion of the total phenotypic variance attributable to genetic effects. H^2 is sometimes referred to as the coefficient of genetic determination. The more

commonly used form of heritability is known as heritability in the narrow sense. If $\text{Var}(A)$ is the variance of the additive genetic effect, then heritability in the narrow sense is defined as

$$h^2 = \frac{\text{Var}(A)}{\text{Var}(P)} \quad (4)$$

and is the proportion of the total phenotypic variance resulting from additive genetic effects. The coefficient h^2 measures the importance of transmissible genetic material because it quantifies to what extent the variation in the phenotype of a progeny is determined by the genes transmitted from its parents. See Refs. [1–4] for a thorough discussion of heritability and quantitative genetics.

In the analysis of h^2 , the primary source of variation under study is that which results from additive genetic effects. The other sources of variation (environmental, dominance, epistatic) play a secondary role and are often combined into a single term. In this manner

$$\begin{aligned} P &= A + (D + I + E) \\ &= A + \text{Other} \end{aligned} \quad (5)$$

so that

$$\text{Var}(P) = \text{Var}(A) + \text{Var}(\text{Other}) \quad (6)$$

with the understanding that additive genetic effects and the other effects combined are treated as uncorrelated. In most applications, h^2 is written in terms of variance components. Variance components correspond to the sources that produce variation in the phenotypic value. If

$$\text{Var}(A) = \sigma_1^2 \quad (7)$$

and

$$\text{Var}(\text{Other}) = \sigma_2^2 \quad (8)$$

it follows that heritability (in the narrow sense) is a function of σ_1^2 and σ_2^2 , say, $h^2 = f(\sigma_1^2, \sigma_2^2)$, where

$$h^2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \quad (9)$$

This form of h^2 enables the investigator to utilize statistical models to analyze heritability. See Refs. [5–7] for further details concerning statistical models involving genetic variances.

STATISTICAL MODELING OF HERITABILITY

The statistical models used to analyze heritability contain the effects that influence the quantitative characteristic

under study. Effects have different values of interest called levels. If all levels of an effect are included in a model, or only those particular levels of an effect included in the model are of interest, the effect is called a fixed effect. Gender (Male, Female) and smoking status (Yes, No) are examples of fixed effects. If the levels of an effect included in the model may be considered as a random sample of all possible levels, the effect is called a random effect. The additive genetic effect is considered random if the subjects in the study are a random sample from a population of subjects. In this scenario, the focus of attention is not on the specific levels of the effect but on the underlying distribution of the random effect. In particular, the variance of the random effect, for example, the variance of the additive genetic effect, is used to measure the contribution of the effect on the overall variation of the characteristic under study.

Mixed linear models, models that include a linear combination of fixed and random effects, are used in the analysis of heritability. Using matrix notation,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \quad (10)$$

where $\mathbf{Y}$ is the vector of measured characteristics, $\boldsymbol{\beta}$ is the vector of parameters associated with the fixed effects, $\mathbf{u}$ is the vector of unobservable additive genetic effects, and $\boldsymbol{\varepsilon}$ is the vector of unobservable random variables corresponding to “other” effects as given in Eq. 5. For a particular individual, Y_i is the phenotypic value, $x_i'\boldsymbol{\beta}$ is the mean phenotypic value for members of the population associated with x_i where x_i is the i th row of $\mathbf{X}$, u_i is the additive genetic effect, and ε_i represents the combined effect of sources of variation in the phenotypic value not accounted for by the additive genetic effect.

For the purpose of statistical inference, it is assumed that $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ are independent and multivariate normally distributed. Specifically, $\mathbf{u} \sim N(0, \sigma_1^2 \mathbf{R})$ and $\boldsymbol{\varepsilon} \sim N(0, \sigma_2^2 \mathbf{I})$ where the matrix $\mathbf{R}$ is referred to as the relationship matrix and $\mathbf{I}$ is the identity matrix. It follows that $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ where

$$\mathbf{V} = \sigma_2^2 \mathbf{I} + \sigma_1^2 \mathbf{Z}\mathbf{R}\mathbf{Z}' \quad (11)$$

The matrix $\mathbf{R}$ describes the degree to which the subjects in the study are genetically related.

Consider a simple example of the modeling process. Suppose the phenotypic values of six subjects have been recorded. The gender (M,F) and the relationships among the six subjects are displayed in Fig. 1. For instance, subject 1 is a male having a son (subject 3), a daughter (subject 4), and a grandson (subject 6). Note that the parents of subjects 1 and 2 have not been recorded, and subjects 3, 4, 5, and 6 have only one identifiable parent each. Based on Fig. 1, one can also see that subjects 1 and 2 are unrelated and thus the offspring of subject 1 will be unrelated to the offspring of subject 2.

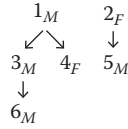


Fig. 1 Pedigree structure

Let Y_i be the phenotypic value of the i th subject, $i = 1, \dots, 6$. Then $\mathbf{Y}' = [Y_1, Y_2, Y_3, Y_4, Y_5, Y_6]$ is the vector of responses for the subjects. The mixed linear model for this example may be expressed as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 & 0 \\ & & 1 & 0 & 0 & 0 \\ & & & 1 & 0 & 0 \\ & & & & 1 & 0 \\ & & & & & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{bmatrix} \quad (12)$$

The influence of the fixed effect gender on the responses is taken into account by the 2×1 vector β , where the first element of β corresponds to males and the second element of β corresponds to females. The influence of additive genetic effects on the responses is taken into account by the vector 6×1 vector u . The additive genetic effect is a random effect in this example.

From Fig. 1, subjects 1, 3, 4, and 6 are related and subjects 2 and 5 are related. The relationship matrix $\mathbf{R}$ for the example is

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 & 0 & 1/4 \\ & 1 & 0 & 0 & 1/2 & 0 \\ & & 1 & 1/4 & 0 & 1/2 \\ & & & 1 & 0 & 1/8 \\ & & & & 1 & 0 \\ & & & & & 1 \end{bmatrix} \quad (13)$$

Because subject 3 receives one-half of its genetic material from subject 1 (the other half coming from an unidentified source), the genetic covariance between subjects 1 and 3, denoted by $\text{Cov}(u_1, u_3)$, is $1/2\sigma_1^2$. Similarly, because subjects 3 and 4 are half-sibs, that is, they have only one parent in common, $\text{Cov}(u_3, u_4) = 1/4\sigma_1^2$. Note, for instance, that $\text{Cov}(u_1, u_5) = 0$ as subjects 1 and 5 do not share any

common genetic material. A general procedure used to construct the relationship matrix is given in Ref. [8].

Heritability of the characteristic under study is based on the variance components in the statistical model. It is assumed that $\sigma_1^2 \geq 0$ and $\sigma_2^2 > 0$. This setup allows for the possibility that the genetic influence on the variability in the phenotypic value may be negligible. It follows, from Eq. 9, that $0 \leq h^2 < 1$. In the following three sections, the analysis of heritability is considered by applying the statistical inferential procedures of point estimation, interval estimation, and hypothesis testing.

POINT ESTIMATION

The point estimation techniques used for h^2 are essentially the same as those used to estimate the individual variance components. That is, if $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are the estimates of σ_1^2 and σ_2^2 , respectively, then $\hat{h}^2 = f(\hat{\sigma}_1^2, \hat{\sigma}_2^2)$, where

$$\hat{h}^2 = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2} \quad (14)$$

is the estimate of h^2 . In this section, we discuss estimators based on analysis of variance (ANOVA), maximum likelihood (ML), and restricted maximum likelihood (REML) procedures. Because ML estimators are translation invariant, the properties of the ML estimators of h^2 carry over from the properties of the ML estimators of σ_1^2 and σ_2^2 . In addition to the estimators considered here, alternative estimators of h^2 may be of interest.

Analysis of Variance Estimation

The ANOVA estimator of h^2 may be obtained from the ANOVA table associated with the model. The procedure involves equating the sums of squares to their expected values and solving for the estimates of the resulting variance components. In general, if $\mathbf{B}_i$ is a symmetric matrix having the same dimension as the length of the $\mathbf{Y}$ vector, then the expected value of $\mathbf{Y}'\mathbf{B}_i\mathbf{Y}$ is

$$E(\mathbf{Y}'\mathbf{B}_i\mathbf{Y}) = \text{tr}(\mathbf{B}_i\mathbf{V}) + \beta' \mathbf{X}'\mathbf{B}_i\mathbf{X}\beta \quad (15)$$

$\text{tr}(\mathbf{B}_i\mathbf{V})$ is the trace of the matrix $\mathbf{B}_i\mathbf{V}$ and is a linear combination of the variance components. The expectation will be free of the fixed effects β if $\beta' \mathbf{X}'\mathbf{B}_i\mathbf{X}\beta$ is zero. There exists a $\mathbf{B}_i$ such that

$$E(\mathbf{Y}'\mathbf{B}_i\mathbf{Y}) = r_i\sigma_2^2 + r_i\lambda_i\sigma_1^2 \quad (16)$$

where $0 \leq \lambda_1 < \dots < \lambda_d$ are the distinct eigenvalues, each having multiplicity r_i , of a matrix related to the $\mathbf{B}_i$'s. See Ref. [9] for more details. Under normal theory the quadratic forms, $\mathbf{Y}'\mathbf{B}_i\mathbf{Y}$, more simply written as Q_i , are

independently distributed as $(\sigma_2^2 + \lambda_i \sigma_1^2) \chi^2$ variates with degrees of freedom r_i .

In many applications $\lambda_1 = 0$, which implies that replications exist in the experiment. The ANOVA table associated with this scenario is given in Table 1.

For those applications in which $\lambda_1 > 0$, there is not a simple estimator of σ_2^2 . In either case, the ANOVA estimator of h^2 takes the form of

$$\hat{h}^2 = \frac{\sum_{i=1}^k r_i \sum_{i=k+1}^d Q_i - \sum_{i=k+1}^d r_i \sum_{i=1}^k Q_i}{\sum_{i=k+1}^d r_i (\lambda_i - 1) \sum_{i=1}^k Q_i - \sum_{i=1}^k r_i (\lambda_i - 1) \sum_{i=k+1}^d Q_i} \quad (17)$$

where the choice of k represents the partitioning of $Q_1, \dots, Q_d$ into two distinct groups and lies at the discretion of the investigator. Only for the case when $d = 2$ is the ANOVA estimator of h^2 unique. This occurs when the model is based on balanced data. Note that $0 \leq h^2 < 1$ whereas $\hat{h}^2$ may be negative.

Maximum Likelihood and Restricted Maximum Likelihood Estimation

The ML and REML procedures estimate variance components, and hence heritability, by determining the values of the variance components that maximize a likelihood function based on the response vector $\mathbf{Y}$. ML methods require the investigator to specify the distribution of the responses and, in many applications, it is assumed that $\mathbf{Y}$ is multivariate normally distributed. The ML estimation procedure uses a likelihood function that is based on the original responses and produces estimates of β in addition to estimates of σ_1^2 and σ_2^2 . The logarithm of the likelihood function, which is the function that is usually maximized to obtain the ML estimators, is often denoted by $L(\beta, \sigma_1^2, \sigma_2^2 | \mathbf{Y})$. The ML estimate of heritability may be obtained by first estimating the variance components and then using Eq. 14.

The REML estimation procedure uses a likelihood function that is based on a transformation of the responses. The distribution of the transformed responses, also referred to as error contrasts, does not depend on the fixed effects part of the model in Eq. 10. The logarithm of the likelihood function for REML estimation can be written as $L(\sigma_1^2, \sigma_2^2 | \mathbf{Q})$ where $\mathbf{Q}$ is the vector $(Q_1, \dots, Q_d)$. The elements of $\mathbf{Q}$ represent the squares of the error contrasts.

Table 1 ANOVA Table

Source	df	SS	E(SS)
Model (free of β)	$\sum_{i=2}^d r_i$	$\sum_{i=2}^d Q_i$	$\sigma_2^2 \sum_{i=2}^d r_i + \sigma_1^2 \sum_{i=2}^d r_i \lambda_i$
Error	r_1	Q_1	$r_1 \sigma_2^2$

It follows that the REML procedure is used to estimate the variance components only. By applying Eq. 14, one obtains the REML estimate of heritability. See the *Mixed Effects Model* entry for more information concerning ML procedures in mixed linear models.

By construction, both the ML methods result in estimates of heritability that are in the parameter space. That is, $0 \leq \hat{h}_{\text{ML}}^2 < 1$ and $0 \leq \hat{h}_{\text{REML}}^2 < 1$. Whereas the REML estimation procedure accounts for the degrees of freedom that are associated with estimating fixed effects, the ML estimation procedure does not. For all but the simplest of models, closed-form expressions for the ML and REML estimators do not exist and iterative procedures are required to compute the estimates. It can be shown that the REML estimator of h^2 is

$$\hat{h}_{\text{REML}}^2 = \frac{\sum_{i=1}^d a_i(h^2, \lambda_i, r_i) Q_i}{\sum_{i=1}^d b_i(h^2, \lambda_i, r_i) Q_i} \quad (18)$$

where a_i and b_i are functions of h^2 , λ_i , and r_i that serve as coefficients of Q_i . In essence, the REML estimator of h^2 is a ratio of linear combinations of the quadratic forms $Q_1, \dots, Q_d$ whose coefficients depend in part on the unknown parameter. The value of $\hat{h}_{\text{REML}}^2$ must be obtained iteratively by selecting a starting value and relying on the convergence of the procedure. See Refs. [7,10,11] for general information about ML estimation procedures.

INTERVAL ESTIMATION

Confidence intervals for variance components and for h^2 in particular are not typically symmetric. That is, the point estimate of h^2 does not necessarily lie at the center of the interval. Intervals are obtained by inverting a quantity that is a function of the responses and h^2 whose distribution is independent of h^2 . Specifically, if $\mathbf{Y}$ is multivariate normally distributed, then the elements of $\mathbf{Q}$ are chi-squared distributed and are independent as discussed in the subsection "Analysis of Variance Estimation." Because the ratio of independently chi-squared variates give rise to F -distributed quantities, it follows that a $100(1 - \alpha)\%$ confidence interval for h^2 can be computed by recognizing that

$$P \left(F_{\alpha_1} \leq \frac{\sum_{i=k+1}^d \frac{Q_i}{1 + h^2(\lambda_i - 1)} / \sum_{i=k+1}^d r_i}{\sum_{i=1}^k \frac{Q_i}{1 + h^2(\lambda_i - 1)} / \sum_{i=1}^k r_i} \leq F_{\alpha_2} \right) = 1 - \alpha \quad (19)$$

where $\alpha_1 + \alpha_2 = \alpha$ and F_{α_1} , F_{α_2} represent the α_1 and α_2 percentiles of the F -distribution having numerator and

denominator degrees of freedom equal to $\sum_{i=k+1}^d r_i$ and $\sum_{i=1}^k r_i$, respectively. By rewriting the joint inequality in the probability statement in Eq. 19 so that h^2 alone appears in the center, one can obtain a confidence interval for h^2 . If $d > 2$, this can only be accomplished using numerical techniques as no closed-form expression for the endpoints of the confidence interval exists. As previous discussions suggested, the choice of k is up to the investigator. Additional details are presented in Ref. [12].

HYPOTHESIS TESTS

Hypothesis tests for h^2 are useful if the investigator wants to determine if the sample information is consistent with a specific value of heritability under study. The hypothesis test developed in Ref. [9] is equivalent to the confidence interval given in the section “Interval Estimation.” To conduct the hypothesis test $H_0: h^2 = h_o^2$ vs. the two-sided hypothesis $H_1: h^2 \neq h_o^2$, define the test statistic to be

$$F^* = \frac{\sum_{i=k+1}^d \frac{Q_i}{1+h_o^2(\lambda_i-1)}}{\sum_{i=1}^k \frac{Q_i}{1+h_o^2(\lambda_i-1)}} \bigg/ \frac{\sum_{i=k+1}^d r_i}{\sum_{i=1}^k r_i} \quad (20)$$

Under H_0 , F^* is F -distributed with numerator and denominator degrees of freedom equal to $\sum_{i=k+1}^d r_i$ and $\sum_{i=1}^k r_i$, respectively. For a prespecified value of α , the type I error probability, reject H_0 if $F^* < F_{\alpha/2}$ or $F^* > F_{\alpha/2}$ and do not reject H_0 if $F_{\alpha/2} \leq F^* \leq F_{\alpha/2}$. One-sided tests may also be conducted using a similar procedure.

The locally most powerful (LMP) test has the property of being more powerful than any other test for values of h^2 that are close to h_o^2 . By definition, the LMP test is a test whose power function has maximum slope at $h^2 = h_o^2$. The test statistic for the LMP test is

$$T_{\text{LMP}} = \frac{(1-h_o^2) \sum_{i=1}^d \frac{\lambda_i Q_i}{(1+h_o^2(\lambda_i-1))^2}}{\sum_{i=1}^d \frac{Q_i}{1+h_o^2(\lambda_i-1)}} \quad (21)$$

Recall that $Q_i \sim (\sigma_e^2 + \lambda_i \sigma_1^2) \chi^2(r_i)$ are independent for $i = 1, \dots, d$. Under H_0 , the distribution of the test statistic is

$$\frac{(1-h_o^2) \sum_{i=1}^d \frac{\lambda_i \chi^2(r_i)}{1+h_o^2(\lambda_i-1)}}{\sum_{i=1}^d \chi^2(r_i)} \quad (22)$$

where the chi-squared random variables are independent. This testing procedure is discussed in some detail

in Ref. [13]. Neyman–Pearson tests and alternative testing procedures are given in Ref. [14].

DISCUSSION

In the preceding analyses, the investigator may be faced with the task of splitting the information contained in the quadratic forms $Q_1, \dots, Q_d$ into two separate pieces, namely $Q_1, \dots, Q_k$ and $Q_{k+1}, \dots, Q_d$. This becomes apparent by examining the ANOVA estimator of h^2 in Eq. 17, the confidence interval for h^2 as determined by Eq. 19, and the test statistic for h^2 in Eq. 20. The choice of k is simplified if $\lambda_1 = 0$. In this case, it is recommended that $k = 1$ so that Q_1 is separated from the remaining quadratic forms. See Table 1 for an example of this application. The choice of $k = 1$ corresponds to the Wald method in Ref. [15] for the unbalanced one-way random effects model. See Ref. [12] for guidance on how to select the value of k when $\lambda_1 > 0$.

The ML procedures are popular methods used to estimate variance components. Many statistical software packages will compute ML estimates of σ_1^2 , σ_2^2 and hence, h^2 , with very little effort required on the part of the investigator. These methods are based on iterative procedures that require initial estimates and employ well-designed algorithms to arrive at the solutions. While these procedures perform admirably in a wide variety of applications, there is no guarantee that the methods will always converge to a solution. Closed-form approximations to the REML estimator of h^2 are discussed in Ref. [16].

CONCLUSION

Other statistical procedures may be used to analyze heritability. Bootstrap and jackknife approaches may be useful. An overview of bootstrap estimates and bootstrap confidence intervals for variance components may be found in Ref. [17]. The analysis of heritability using Bayesian procedures may also be performed. In Ref. [7], the authors note that, because the REML estimation of variance components is based on a specific marginal likelihood function, it is equivalent to Bayesian estimation using noninformative priors.

REFERENCES

1. Falconer, D.S. *Introduction to Quantitative Genetics*, 3rd Ed.; Longman Scientific & Technical: Essex, England, 1989.
2. Fraser, G.R.; Mayo, O. *Textbook of Human Genetics*; Blackwell Scientific Publications: Oxford, 1975.
3. Narain, P. *Statistical Genetics*; Wiley: New York, 1990.
4. Wright, S. *Evolution and the Genetics of Populations*; University of Chicago Press: Chicago, IL, 1969, Vol. 2.
5. Fisher, R.A. *The Genetical Theory of Natural Selection*; Dover: New York, 1958.

6. Kempthorne, O. *An Introduction to Genetic Statistics*; Iowa State University Press: Ames, IA, 1969.
7. Searle, S.R.; Casella, G.; McCulloch, C.E. *Variance Components*; Wiley: New York, 1992.
8. Henderson, C.R. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*. **1976**, *32*, 69–83.
9. LaMotte, L.R.; McWhorter, A. Jr., An exact test for the presence of random walk coefficients in a linear regression model. *J. Am. Stat. Assoc.* **1978**, *73*, 816–820.
10. Hartley, H.O.; Rao, J.N.K. Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*. **1967**, *54*, 93–108.
11. Harville, D.A. Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* **1977**, *72*, 320–338.
12. Burch, B.D.; Iyer, H.K. Exact confidence intervals for a variance ratio (or heritability) in a mixed linear model. *Biometrics*. **1997**, *53*, 1318–1333.
13. Westfall, P.H. Power comparisons for invariant variance ratio tests in mixed ANOVA models. *Ann. Stat.* **1989**, *17*, 318–326.
14. Lin, T.-H.; Harville, D.A. Some alternatives to Wald's confidence interval and test. *J. Am. Stat. Assoc.* **1991**, *86*, 179–187.
15. Wald, A. A note on the analysis of variance with unequal class frequencies. *Ann. Math. Stat.* **1940**, *11*, 96–100.
16. Burch, B.D.; Harris, I.R. Closed-form approximations to the REML estimator of a variance ratio (or heritability) in a mixed linear model. *Biometrics*. **2001**, *57*, 1148–1156.
17. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman & Hall/CRC: Boca Raton, FL, 1993.

Analysis of Variance (ANOVA)

Maria Overbeck-Larisch and Werner Sanns

University of Applied Sciences, Darmstadt, Germany

Abstract

The classical idea of the analysis of variance was to find out if there exists an influence of one or more categorical variables over a normally distributed random variable. This entry addresses this classical use of ANOVA as well as the modern approach which includes quantitative analysis by providing several sample analyses.

INTRODUCTION

The analysis of variance (ANOVA) is a statistical method, which originally was developed by R.A. Fisher (1890–1962) for problems in biological studies.^[2] Since that time, the ANOVA has been used in many different areas and there exist a comprehensive literature concerning the theory and practice of ANOVA.^[4,5,7–10] The classical idea of the ANOVA was to find out if there exists an influence of one or more categorical variables (factor variables) over a normally distributed random variable. This classical task of ANOVA can be regarded as a qualitative analysis. The modern approach to ANOVA, however, includes a quantitative analysis, too, via the general linear model (GLM) (i.e., it allows to measure the strength of the influence of each factor variable).

Practicing the ANOVA requires an appropriate software system. SAS® is one of these systems; it is a well-known tool in biopharmaceutical statistics. Therefore we include in this contribution the basic commands of the SAS procedures ANOVA and GLM (Ref. [6]).

MATHEMATICAL PROBLEM

Assume k random variables to be independent and normally distributed with the same variance σ^2 but with eventually different mean values (expected values) $\mu_1, \mu_2, \dots, \mu_k$. The objective of the ANOVA is to test the hypothesis H_0 of all mean values being equal against the alternative H_1 , which states that there are at least two different mean values. Thus the ANOVA is a statistical test procedure which, for $k = 2$, corresponds to the situation of the two-sample t -test. If the hypothesis holds, the k random variables have identical normal distributions. Otherwise, they have different mean values.

EXAMPLE 1 OF TOTAL PROTEIN IN SERUM

In toxicological studies, many parameters such as the number of red and white blood corpuscles (erythrocytes

and leucocytes) or the Na concentration in urine play an important role, and it is of great interest to know the statistical distribution of these parameters. Some of them such as the parameter TPRO (Total **PRO**tein in serum) can be assumed to be normally distributed. The following table contains the values of the parameter TPRO, where the measurements come from three control groups (not treated with the test compound) consisting of dogs, mice, and rats.

	Sex = f						Sex = m					
Species = Dog	5.7	5.4	5.8	5.5	5.1	5.5	5.5	5.4	5.4	5.0	5.4	5.9
Species = Mouse	5.9	5.4	5.7	5.4	5.3	5.9	6.1	5.4	6.1	5.8	5.6	6.3
Species = Rat	6.7	6.7	6.6	6.6	6.6	6.4	6.9	6.9	7.0	6.5	6.4	6.5

For further computations, we save the data set as a SAS file named TPRO with the three variables Species, Sex, and Value. In submitting the following SAS command lines, the file is created and printed on the screen in the output window:

```
data TPRO;

input Species$ Sex$ Value;

cards;

Dog f 5.7

Dog f 5.4

...

Dog m 5.9

Mouse f 5.9

...

Rat m 6.5

run;

proc print; run;
```

By means of a one-way ANOVA, we shall show that the values of the parameter TPRO differ significantly from species to species. One-way ANOVA means that the influence of only one factor, for instance the factor species, is taken into account.

But if one also takes into account the fact that the first six animals of each species are female whereas the last six animals are male, one may wonder if the factor Sex significantly influences the value of the parameter TPRO, too. By means of a two-way ANOVA, we will answer this question.

MODEL ASSUMPTIONS FOR THE ONE-WAY ANALYSIS OF VARIANCE

In the general situation of a one-way ANOVA, there are n observations in k different samples (because of the k levels of one categorical factor variable) of eventually different sample sizes $n_1, n_2, \dots, n_k$:

Sample 1	$y_{11}, y_{12}, \dots, y_{1n_1}$
Sample 2	$y_{21}, y_{22}, \dots, y_{2n_2}$
...	...
...	...
Sample k	$y_{k1}, y_{k2}, \dots, y_{kn_k}$

The observations y_{ij} , $1 \leq i \leq k$, and $1 \leq j \leq n_i$ come from random variables:

$$Y_{ij} = \mu_i + E_{ij} = \mu + \alpha_i + E_{ij}$$

where the E_{ij} variables are independent $N(0, \sigma^2)$ -distributed random variables. $\mu := (1/n) \sum_{i=1}^k 1\mu_i$ denotes the overall mean, and the parameters $\alpha_i : \mu_i - \mu$ denote the deviations of the particular mean values from μ . Because of this definition, the parameters α_i satisfy the condition:

$$\sum_{i=1}^k \alpha_i = 0$$

which later on could be used for recoding the parameters within the frame of the linear model. In terms of the parameter α_i , we have to test the hypothesis H_0 of all α_i being zero against the alternative H_1 that at least one $\alpha_i \neq 0$.

NOTATIONS

Some of the following notations are well known from other statistical methods:

$$n = \sum_{i=1}^k n_i$$

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad \text{and} \quad s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$$\text{for } 1 \leq i \leq k$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} = \frac{1}{n} \sum_{i=1}^k n_i \bar{y}_i \quad \text{and} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

$$\text{sst} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = (n-1)s^2$$

$$\text{ss1} = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$$

$$\text{ss2} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^k (n_i - 1) s_i^2$$

The name *analysis of variance* comes from the equation:

$$\text{sst} = \text{ss1} + \text{ss2}$$

which is valid as the following computation shows:

$$\begin{aligned} \text{sst} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} \left((y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y}) \right)^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} \left((y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{y})^2 + 2(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) \right) \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + 2 \sum_{i=1}^k (\bar{y}_i - \bar{y}) \underbrace{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)}_{=0} \\ &\quad + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 = \text{ss2} + \text{ss1} \end{aligned}$$

Interpretation: The total sum of square (sst) (i.e., the sum over all squared differences between the total mean and the particular observation) can be split into two sums ss1 and ss2. The first one is a result of the differences *between* the samples, and the second one is a result of the differences *within* the samples. So the above equation can be written in the form:

$$\text{sst} = \text{ss1} + \text{ss2} = \text{ss}_{\text{between}} + \text{ss}_{\text{within}}$$

It seems to be reasonable to reject the hypothesis that the k normal distributions are identical if the quotient $\text{ss}_{\text{between}} / \text{ss}_{\text{within}}$ is large. The question how large the quotient has to be in order to reject the hypothesis can be answered by the following theorem.

THEOREM

If the model assumptions are fulfilled and if the hypothesis of equal means is true, we can state the following:

$$\frac{sst}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2 \sim \chi_{n-1}^2$$

$$\frac{ss1}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y})^2 \sim \chi_{k-1}^2$$

$$\frac{ss2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \sim \chi_{n-k}^2$$

$$V := \frac{ss1 / (k-1)}{ss2 / (n-k)} \sim F_{k-1, n-k}$$

Under H_0 , the random variable V assumes small values. Thus H_0 can be rejected if the realized value v of the test variable V exceeds a critical value v^* . For a given significance level α , the critical value v^* is determined in such a way that:

$$P(V > v^* | H_0) = \alpha$$

Let F be the cumulative distribution function of the $F_{k-1, n-k}$ distribution. Then the critical value v^* is the solution of the equation:

$$1 - F(v^*) = \alpha$$

or

$$v^* = F^{-1}(1 - \alpha)$$

respectively. This means that v^* is the $(1 - \alpha)$ percentile of the $F_{k-1, n-k}$ distribution.

Therefore we have the following procedure for a one-way ANOVA:

1st step: Choose an appropriate significance level α .

2nd step: Compute the $(1 - \alpha)$ percentile of the $F_{k-1, n-k}$ distribution.

3rd step: Compute $\bar{y}_1, \dots, \bar{y}_k, \bar{y}, s_1^2, \dots, s_k^2, ss1, ss2$, and $v = [ss1/(k-1)]/[ss2/(n-k)]$.

4th step: Reject H_0 if $v > v^*$.

REMARK

It is obvious that the above theorem and the resulting procedure for a one-way ANOVA strongly depend on the assumption of normally distributed random variable Y_{ij} . If this condition is not satisfied, the test variables V is not appropriate and a distribution-free test like the test of Kruskal–Wallis should be chosen.

EXAMPLE 2 OF TOTAL PROTEIN IN SERUM

As the TPRO value can be assumed to be normally distributed, we follow the given procedure for a one-way ANOVA ($k = 3, n = 36$):

1st step: If the aim is to show that the values of the parameter TPRO differ significantly from species to species, we choose a small significance level. We choose the standard level of 5%.

2nd step: We have to compute the 95% percentile of the $F_{2,33}$ distribution. This can be performed, for instance, by the following SAS program:

```
data temp;
vstar=Finv(0.95,2,33); run;
proc print; run;
```

We get $v^* = 3.28492$.

3rd step: Submitting the SAS commands:

```
proc means data=TPRO; run;
proc means data=TPRO; by Species; run;
```

the following results can be achieved:

$\bar{y} = 5.9527778$	$\bar{y}_1 = 5.4666667$	$\bar{y}_2 = 5.7416667$	$\bar{y}_3 = 6.6500000$
$s = 0.5744494$	$s_1 = 0.2570226$	$s_2 = 0.3287949$	$s_3 = 0.1977142$

So we have:

$$ss1 = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 = 12 \sum_{i=1}^3 (\bar{y}_i - \bar{y})^2 = 9.20388889$$

$$ss2 = \sum_{i=1}^k (n_i - 1) s_i^2 = 11 \sum_{i=1}^3 s_i^2 = 2.34583333$$

$$v = \frac{ss1 / (k-1)}{ss2 / (n-k)} = \frac{9.20388889 / 2}{2.34583333 / 33} = 64.738$$

4th step: As $v > v^*$, the hypothesis H_0 can be rejected at the 5% significance level. The means of the three species differ significantly.

These and a lot of more computational results can be achieved by putting the SAS procedure ANOVA into action. The SAS program:

```
proc anova data=TPRO;
class Species;
model Value=Species;
run;
```

yields, among other results, the ANOVA table at the base of this page (we added our symbols in parentheses).

There are two more or less remarkable differences between our handmade calculation and the SAS output. The first one concerns the notation: The sum $ss1 = ss_{\text{between}}$ describes that part of the total sum of squares that corresponds to the model, whereas the sum $ss2 = ss_{\text{within}}$ is the remaining error sum of squares. For a better understanding of the SAS output, denote $ss1$ by ss_{model} and denote $ss2$ by ss_{error} and write the equation of a one-way ANOVA in the form:

$$sst = ss_{\text{model}} + ss_{\text{error}}$$

The second difference is more essential: SAS does not compute the 95% percentile v^* of the $F_{2,33}$ distribution, but in the column with the title “ $Pr > F$ ” it produces the probability $1 - F(v)$ or an upper bound. Here this upper bound is 0.0001. Because of the equivalence:

$$v > v^* \quad F(v) > F(v^*) \quad F(v) > 1 - \alpha \quad 1 - F(v) < \alpha$$

the hypothesis can be rejected as long as the value $1 - F(v)$ is less than the chosen significance level α . As already mentioned, SAS produces some more computational results that are not printed above. As a very important number, for instance, we get:

$$R^2 = ss1 / sst = ss_{\text{between}} / ss_{\text{total}} = ss_{\text{model}} / ss_{\text{total}}$$

which tells us how much of the total sum of squares can be explained by the model. In the example of TPRO, $R^2 = 0.796893$ means that about 80% of the variation in the data set can be explained by the fact that the animals belong to different species.

MODEL ASSUMPTIONS FOR THE TWO-WAY ANALYSIS OF VARIANCE

Within the framework of a two-way ANOVA or a higher-way ANOVA, we simultaneously explore if there is an influence of two or more categorical variables (factor variables) over a normally distributed random variable. For the sake of simplicity, let us restrict ourselves to a situation with only two factor variables. Regarding our TPRO data set, we want to find out simultaneously if the Species or the Sex has an effect on the value of the parameter TPRO. The two factor variables divide the observations in r rows and p columns with $r \times p$ cells (samples). Assume, furthermore, that there are m observations in each cell: So for the number n of observations, we have $n = rpm$. (In case of $m = 1$, we have the situation of a simple block experiment.)

Let us assume the observations y_{ijl} , $1 \leq i \leq r$, $1 \leq j \leq p$, and $1 \leq l \leq m$ to come from random variables:

$$Y_{ijl} = \mu + \alpha_i + \beta_j + E_{ijl}$$

with independent $N(0, \sigma^2)$ -distributed random variables E_{ijl} . α_i denotes the effect of the i th row and β_j denotes the effect of the j th column. As in the model of a one-way ANOVA, we assume two reparameterization conditions to be valid:

$$\sum_{i=1}^r \alpha_i = 0 \quad \text{and} \quad \sum_{j=1}^p \beta_j = 0$$

So we have to test the hypothesis $H_0: \alpha_1 = \dots = \alpha_r = \beta_1 = \dots = \beta_p = 0$ against the alternative H_1 that at least one of these $r + p$ parameters does not have a value of zero.

Class Level Information					
Class		Levels	Values		
Species		3	Dog	Mouse	Rat
Number of observations				36	
The ANOVA Procedure					
Dependent Variable: Value					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2 ($k-1$)	9.20388889 (ss1)	4.60194444 ($ss1/(k-1)$)	64.74 (v)	<.0001
Error	33 ($n-k$)	2.34583333 (ss2)	0.07108586 ($ss2/(n-k)$)		
Corrected Total	35 ($n-1$)	11.54972222 (sst)			

NOTATIONS

$$\bar{y} = \frac{1}{rpm} \sum_{i=1}^r \sum_{j=1}^p \sum_{l=1}^m y_{ijl} \quad sst = \sum_{i=1}^r \sum_{j=1}^p \sum_{l=1}^m (y_{ijl} - \bar{y})^2$$

$$\bar{y}_{i..} = \frac{1}{pm} \sum_{j=1}^p \sum_{l=1}^m y_{ijl} \quad ss1 = \sum_{i=1}^r pm(\bar{y}_{i..} - \bar{y})^2$$

for $1 \leq i \leq r$

$$\bar{y}_{.j.} = \frac{1}{rm} \sum_{i=1}^r \sum_{l=1}^m y_{ijl} \quad ss2 = \sum_{j=1}^p rm(\bar{y}_{.j.} - \bar{y})^2$$

for $1 \leq j \leq p$

$$ss3 = \sum_{i=1}^r \sum_{j=1}^p \sum_{l=1}^m (y_{ijl} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y})^2$$

As in the case of a one-way ANOVA, the basic idea is to split the total sum of squares:

$$sst = ss1 + ss2 + ss3$$

Here $ss1$ is a result of the differences between the rows, $ss2$ is a result of the differences between the columns, and $ss3$ is the remaining error sum of squares. If the model assumptions are fulfilled and if the hypothesis is true, we can state: The random variables $ss1/\sigma^2$, $ss2/\sigma^2$, and $ss3/\sigma^2$ are χ^2 -distributed. The number of degrees of freedom of $ss1/\sigma^2$ and $ss2/\sigma^2$ is $r-1$ and $p-1$, respectively. The number of degrees of freedom of $ss3/\sigma^2$ is $n-1-(r-1)-(p-1) = n+1-r-p$ with $n = rpm$. Furthermore, we have:

$$V_1 : \frac{ss1/(r-1)}{ss3/(n+1-r-p)} \sim F_{(r-1), (n+1-r-p)}$$

$$V_2 : \frac{ss2/(p-1)}{ss3/(n+1-r-p)} \sim F_{(p-1), (n+1-r-p)}$$

The hypothesis H_0 can be rejected if the realization v_1 of the test variable V_1 is greater than the $(1-\alpha)$ percentile v_1^* of the F -ratio distribution with $(r-1), (n+1-r-p)$ degrees of freedom, or if the realized value v_2 of the test variable V_2 is greater than the $(1-\alpha)$ percentile v_2^* of the F -ratio distribution with $(p-1), (n+1-r-p)$ degrees of freedom. This rejection rule is equivalent to the demands:

$$P(V_1 > v_1 | H_0) = 1 - F_1(v_1) < \alpha \quad \text{or}$$

$$P(V_2 > v_2 | H_0) = 1 - F_2(v_2) < \alpha$$

Here the functions F_1 and F_2 are the cumulative distribution functions of the F -ratio distributions with $(r-1), (n+1-r-p)$ and $(p-1), (n+1-r-p)$ degrees of freedom, respectively.

	Column			
	1	2	...	P
1	$y_{111}, y_{112}, \dots, y_{11m}$	$y_{121}, \dots, y_{12m}$	...	$y_{1p1}, \dots, y_{1pm}$
Row 2	$y_{211}, y_{212}, \dots, y_{21m}$	$y_{221}, \dots, y_{22m}$	...	$y_{2p1}, \dots, y_{2pm}$
3	$y_{311}, y_{312}, \dots, y_{31m}$	$y_{321}, \dots, y_{32m}$	...	$y_{3p1}, \dots, y_{3pm}$
...	...	...	...	...
r	$y_{r11}, y_{r12}, \dots, y_{r1m}$	$y_{r21}, \dots, y_{r2m}$	...	$y_{rp1}, \dots, y_{rpm}$

EXAMPLE 3 OF TOTAL PROTEIN IN SERUM

Submitting the SAS program:

```
proc anova data=TPRO;
    class Species Sex;
    model Value=Species Sex;
run;
```

one gets the ANOVA table at the base of this page.

As $1 - F_1(v_1) = 1 - F_1(65.58) < 0.0001 < \alpha = 5\%$, the hypothesis of no differences between the Species can be rejected. But as $1 - F_2(v_2) = 1 - F_2(1.43) = 0.2407$ is much greater than the significance level of $\alpha = 5\%$, the hypothesis of no differences between the Sex cannot be rejected.

Note that the R^2 is nearly the same as in the model with the one factor variable, Species.

If the number m of observations in the single cells is greater than one, the model of a two-way ANOVA allows one to include interactions. The corresponding model equations are:

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijl}$$

where the following equations are assumed to be true:

$$\sum_{i=1}^r \alpha_i = 0, \quad \sum_{j=1}^p \beta_j = 0$$

$$\sum_{i=1}^r (\alpha\beta)_{ij} = 0 \text{ for all } 1 \leq i \leq r;$$

$$\sum_{i=1}^r (\alpha\beta)_{ij} = 0 \text{ for all } 1 \leq j \leq p$$

Without discussing this model in detail, we submit the following SAS program:

```
proc anova data=TPRO;
    class Species Sex;
    model Value=Species Sex Species*Sex;
run;
```

The SAS output contains an additional line for the interactions between the categorical variables Species and Sex. As the computed probability $1 - F(v_3) = 0.2776$ is much greater than the assumed significance level of 5%, we conclude that there are no significant interactions.

LINEAR MODEL OF ANALYSIS OF VARIANCE

To begin with, assume the situation of a one-way ANOVA with $k = 3$ samples (such as in the example of TPRO, where the samples correspond to different species):

Class Level Information		
Class	Levels	Values
Species	3	Dog Mouse Rat
Sex	2	f m
Number of observations		36

The ANOVA Procedure					
Dependent Variable: Value					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	9.30416667	3.10138889	44.20	<.0001
Error	32	2.24555556	0.07017361		
Corrected Total	35	11.54972222			
	R-Square	Coeff Var	Root MSE	Value Mean	
	0.805575	4.450074	0.264903	5.952778	
Source	DF	Anova SS	Mean Square	F Value	Pr > F
Species	2	9.20388889	4.60194444	65.58	<.0001
Sex	1	0.10027778	0.10027778	1.43	0.2407

Sample 1	$y_{11}, y_{12}, \dots, y_{1n_1}$
Sample 2	$y_{21}, y_{22}, \dots, y_{2n_2}$
Sample 3	$y_{31}, y_{32}, \dots, y_{3n_3}$

The n model equation is:

$$Y_{ij} = \mu_i + E_{ij} = \mu + \alpha_i + E_{ij} \quad \text{for } i = 1, 2, 3 \text{ and } 1 \leq j \leq n_i$$

Apply the following transformation to the observed value y_{ij} and to the corresponding random variables Y_{ij} and E_{ij} :

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ \dots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \dots \\ y_{2n_2} \\ y_{31} \\ y_{32} \\ \dots \\ y_{3n_3} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_{n_1} \\ y_{n_1+1} \\ y_{n_1+2} \\ \dots \\ y_{n_1+n_2} \\ y_{n_1+n_2+1} \\ y_{n_1+n_2+2} \\ \dots \\ y_n \end{pmatrix}$$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_{n_1} \\ Y_{n_1+1} \\ Y_{n_1+2} \\ \dots \\ Y_{n_1+n_2} \\ Y_{n_1+n_2+1} \\ Y_{n_1+n_2+2} \\ \dots \\ Y_n \end{pmatrix} \quad \mathbf{E} = \begin{pmatrix} E_1 \\ E_2 \\ \dots \\ E_{n_1} \\ E_{n_1+1} \\ E_{n_1+2} \\ \dots \\ E_{n_1+n_2} \\ E_{n_1+n_2+1} \\ E_{n_1+n_2+2} \\ \dots \\ E_n \end{pmatrix}$$

Now the n model equations can be rewritten in the form:

$$Y_i = \mu x_{i0} + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + E_i$$

where

$$x_{i0} = 1 \quad \text{for all } 1 \leq i \leq n$$

$$x_{i1} = \begin{cases} 1 & \text{if } 1 \leq i \leq n_1 \\ 0 & \text{else} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } n_1 + 1 \leq i \leq n_1 + n_2 \\ 0 & \text{else} \end{cases}$$

$$x_{i3} = \begin{cases} 1 & \text{if } n_1 + n_2 + 1 \leq i \leq n \\ 0 & \text{else} \end{cases}$$

Then we have:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_{n_1} \\ Y_{n_1+1} \\ Y_{n_1+2} \\ \dots \\ Y_{n_1+n_2} \\ Y_{n_1+n_2+1} \\ Y_{n_1+n_2+2} \\ \dots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ \dots & \dots & \dots & \dots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$$

$$+ \mathbf{E} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{E}; \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$$

The equation:

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{E}; \quad \mathbf{X} = (x_{ij})_{1 \leq i \leq n, 0 \leq j \leq 3}$$

is similar to a linear intercept regression model with $k = 3$ factor variables (called regressors). But there is the following important difference concerning the elements of the design matrix $\mathbf{X}$: When $\mathbf{X}$ is the design matrix of a linear intercept regression model, for $j \geq 1$, the element x_{ij} is the i th adjusted value for the j th regressor. In the present linear model of a one-way ANOVA, the element x_{ij} of the design matrix $\mathbf{X}$ is an element of the set $\{0,1\}$. For $j \geq 1$, the value x_{ij} only denotes which sample the i th observation belongs to.

For generalizing the one-way model, assume that the number of samples is an arbitrary number $k \geq 2$. In the general case, the design matrix $\mathbf{X}$ has n rows and $k + 1$ columns. All elements of the first column have a value of 1. In the second column, we have n_1 leading elements with a value of 1, followed by $n - n_1$ zeros. In the 3rd column, we have n_1 leading zeros, followed by n_2 elements with a value of 1 and further $n - n_1 - n_2$ zeros. The last column begins with $n - n_k$ zeros, followed by n_k ones. So we have:

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} & \dots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{0}_{n_2} & \dots & \mathbf{0}_{n_2} \\ \mathbf{1}_{n_3} & \mathbf{0}_{n_3} & \mathbf{0}_{n_3} & \mathbf{1}_{n_3} & \mathbf{0}_{n_3} & \dots & \mathbf{0}_{n_3} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \mathbf{1}_{n_k} & \mathbf{0}_{n_k} & \mathbf{0}_{n_k} & \mathbf{0}_{n_k} & \mathbf{0}_{n_k} & \dots & \mathbf{1}_{n_k} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \dots \\ \alpha_k \end{pmatrix}$$

Here $\mathbf{1}_{n_j}$ denotes the n_j -dimensional column vector with all its components having a value of 1 and $\mathbf{0}_{n_j}$ is the n_j -dimensional zero vector.

EXAMPLE 4 OF TOTAL PROTEIN IN SERUM

Neglecting the influence of the factor variable Sex, we have a one-way ANOVA with the linear model equation:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \dots \\ Y_{12} \\ Y_{13} \\ \dots \\ Y_{24} \\ Y_{25} \\ \dots \\ Y_{36} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{12} & \mathbf{1}_{12} & \mathbf{0}_{12} & \mathbf{0}_{12} \\ \mathbf{1}_{12} & \mathbf{0}_{12} & \mathbf{1}_{12} & \mathbf{0}_{12} \\ \mathbf{1}_{12} & \mathbf{0}_{12} & \mathbf{0}_{12} & \mathbf{1}_{12} \end{pmatrix} \cdot \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$$

$$+ \mathbf{E} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{E}; \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$$

The equations for the linear model of a two-way ANOVA with $r \times p$ samples are:

$$Y_i = \mu x_{i0} + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \dots + \alpha_r x_{ir} + \beta_1 x_{i(r+1)} + \dots + \beta_p x_{i(r+p)} + E_i$$

Here for all $1 \leq i \leq n$, the coefficient x_{i0} of the parameter μ has a value of 1. For $1 \leq j \leq r$, the coefficient x_{ij} of the

parameter α_j has a value of 1, if the observation with number i belongs to the j th category of the first factor variable; else x_{ij} has a value of 0. For $1 \leq j \leq p$, the coefficient $x_{i(r+j)}$ of the parameter β_j has a value of 1, if the observation with number i belongs to the j th category of the second factor variable. Otherwise, $x_{i(r+j)}$ has a value of 0.

EXAMPLE 5 OF TOTAL PROTEIN IN SERUM

Taking into account the fact that the first six observations for each species are female whereas the last six observations are male animals, we write the model in the following way ($r = 3, p = 2$):

$$Y_i = \mu x_{i0} + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \beta_1 x_{i4} + \beta_2 x_{i5} + E_i$$

Here the numbers x_{i0} , x_{i1} , x_{i2} , and x_{i3} are defined as above in the linear model of one-way ANOVA, whereas the additional numbers x_{i4} and x_{i5} characterize the sex of the animals. Formally, they are defined in the following way:

$$x_{i4} = \begin{cases} 1 & \text{if the } i\text{th observation comes from a female} \\ 0 & \text{animal else} \end{cases}$$

$$x_{i5} = \begin{cases} 1 & \text{if the } i\text{th observation comes from a male animal} \\ 0 & \text{else} \end{cases}$$

So we get the linear model for the two-way ANOVA:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \dots \\ Y_{12} \\ Y_{13} \\ \dots \\ Y_{24} \\ Y_{25} \\ \dots \\ Y_{36} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_6 & \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{0}_6 & \mathbf{1}_6 & \mathbf{0}_6 \\ \mathbf{1}_6 & \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{0}_6 & \mathbf{0}_6 & \mathbf{1}_6 \\ \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{1}_6 & \mathbf{0}_6 \\ \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{0}_6 & \mathbf{1}_6 \\ \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{0}_6 & \mathbf{1}_6 & \mathbf{1}_6 & \mathbf{0}_6 \\ \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{0}_6 & \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{1}_6 \end{pmatrix} \cdot \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

$$+ \mathbf{E} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{E}; \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

The aim of the classical ANOVA is to examine if all sample effects are 0. But as soon as one has formulated the

linear model $\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{E}$ of the ANOVA, one can do more than this *qualitative analysis* and ask the same questions as in regression analysis. Finding estimators for the parameters and testing the hypotheses on them is a *quantitative analysis*.

In regression analysis, the parameter vector $\boldsymbol{\beta}$ of the linear model $\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{E}$ is estimated by the least squares method (LSM). This means that the least squares estimator (LSE) $\mathbf{b}$ has to minimize the function:

$$f(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b})$$

Simple computations show that the function f becomes minimal if and only if $\mathbf{b}$ is the solution of the so-called normal equations:

$$(\mathbf{X}^T \mathbf{X}) \mathbf{b} = \mathbf{X}^T \mathbf{y}$$

If the matrix $\mathbf{X}$ has a full rank, the matrix $\mathbf{X}^T \mathbf{X}$ is regular and the unique solution of the system of normal equations is:

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

It is important to emphasize that the LSE $\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ is an unbiased estimator of the unknown parameter vector $\boldsymbol{\beta}$ of the linear model $\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{E}$ for every distribution of the random vector $\mathbf{E}$. Thus contrary to the classical approach to ANOVA, where we assume from the very beginning that $\mathbf{E}$ follows a normal distribution law, this assumption is not necessary for estimating the parameter vector $\boldsymbol{\beta}$ of the linear model by means of the LSM.

But there arises another problem: The design matrix $\mathbf{X}$ of the linear model for the one-way ANOVA does not have a full rank. If the factor variable divides the observations in k samples, the matrix $\mathbf{X}$ has $1 + k$ columns. The first column with elements $x_{i0} = 1, 1 \leq i \leq n$ is the sum of the following k columns. The problem is intensified if $\mathbf{X}$ is the design matrix of a linear model for two-way ANOVA. If the two factor variables divide the observations in $r \times p$ samples, the matrix $\mathbf{X}$ has $1 + r + p$ columns. The first column of $\mathbf{X}$ with elements $x_{i0} = 1, 1 \leq i \leq n$ is the sum of the following r columns and it is the sum of the last p columns as well. As a consequence, the matrix $\mathbf{X}^T \mathbf{X}$ is not regular and does not have an inverse.

EXAMPLE 6 OF TOTAL PROTEIN IN SERUM

Neglecting the influence of the factor variable Sex, the design matrix $\mathbf{X}$ of the linear model of this one-way ANOVA satisfies the equation:

$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} 36 & 12 & 12 & 12 \\ 12 & 12 & 0 & 0 \\ 12 & 0 & 12 & 0 \\ 12 & 0 & 0 & 12 \end{pmatrix}$$

This matrix is not regular: The first column is the sum of the last three columns. If the factor Sex, too, is taken into account, the design matrix $\mathbf{X}$ of the linear model of this two-way ANOVA satisfies the equation:

$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} 36 & 12 & 12 & 12 & 18 & 18 \\ 12 & 12 & 0 & 0 & 6 & 6 \\ 12 & 0 & 12 & 0 & 6 & 6 \\ 12 & 0 & 0 & 12 & 6 & 6 \\ 18 & 6 & 6 & 6 & 18 & 0 \\ 18 & 6 & 6 & 6 & 0 & 18 \end{pmatrix}$$

This matrix is not regular: The first column is the sum of the following three columns and it is the sum of the last two columns as well.

There are two methods for getting LSEs in this situation. The SAS procedure GLM uses a generalized inverse, which is defined in the following way: The (m,n) matrix $\mathbf{A}^-$ is a generalized inverse of the (n,m) matrix $\mathbf{A}$ if the equation:

$$\mathbf{A} \cdot \mathbf{A}^- \cdot \mathbf{A} = \mathbf{A}$$

is satisfied. If $\mathbf{A}$ is a regular quadratic matrix, it has exactly one generalized inverse and this is the inverse matrix $\mathbf{A}^{-1}$; else the generalized inverse of the matrix $\mathbf{A}$ is not unique. If the equation system $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$, where $\mathbf{A}$ is an (n,m) matrix and $\mathbf{b}$ is an n -dimensional vector, is solvable, the set of all solutions is:

$$\{\mathbf{x} \mid \mathbf{x} = \mathbf{A}^- \cdot \mathbf{b} + (\mathbf{A}^- \cdot \mathbf{A} - \mathbf{I}_m) \cdot \mathbf{z}, \mathbf{z} \in \mathbb{R}^m\}$$

So one special solution of the linear equation system $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$ is the vector $\mathbf{x} = \mathbf{A}^- \cdot \mathbf{b}$.

Applying this statement to the system $(\mathbf{X}^T \mathbf{X}) \cdot \mathbf{b} = \mathbf{X}^T \mathbf{y}$ of normal equations, we have the following: If $\mathbf{G}$ is a generalized inverse of the matrix $\mathbf{X}^T \mathbf{X}$, the vector $\mathbf{b} = \mathbf{G} \mathbf{X}^T \mathbf{y}$ is only a special solution, which, moreover, is not an unbiased estimator for the unknown parameter vector $\boldsymbol{\beta}$.

EXAMPLE 7 OF TOTAL PROTEIN IN SERUM

The appropriate SAS procedure is GLM. In submitting the commands:

```
proc glm data=TPRO;

class Species;

model Value=Species;

run;
```

the SAS generates the same output as when using the SAS procedure ANOVA for the one-way case. But extending

the model command line by the option solution makes SAS compute, in addition, a special solution $\mathbf{b}$ of the system of normal equations by means of a special generalized inverse $\mathbf{G}$ of the matrix $\mathbf{X}^T \mathbf{X}$. Extending the model command line once more by the two options xpx and i, the matrix $\mathbf{X}^T \mathbf{X}$ and the generalized inverse $\mathbf{G}$ are printed in the output window, too. Thus when submitting for the one-way ANOVA the SAS commands:

```
proc glm data=TPRO;

class Species;

model Value=Species/solution xpx i;

run;
```

the SAS computes:

$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} 36 & 12 & 12 & 12 \\ 12 & 12 & 0 & 0 \\ 12 & 0 & 12 & 0 \\ 12 & 0 & 0 & 12 \end{pmatrix}$$

$$\mathbf{G} = \begin{pmatrix} 0.083333333 & -0.083333333 & -0.083333333 & 0 \\ -0.083333333 & 0.166666667 & 0.083333333 & 0 \\ -0.083333333 & 0.083333333 & 0.166666667 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{b} = \begin{pmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{pmatrix} = \begin{pmatrix} 6.650000000 \\ -1.183333333 \\ -0.908333333 \\ 0.000000000 \end{pmatrix}$$

The output contains the note that the estimates for the intercept μ and for the three effects $\alpha_1, \alpha_2, \alpha_3$ of the three species Dog, Mouse, and Rat are biased and not unique.

The second method for solving the problem of a less-than-full-rank design matrix $\mathbf{X}$ requires a recoding of the parameters to be estimated, which implies a reduction of its number and another interpretation. In the case of a one-way ANOVA with k samples, we make use of the fact that for every i , the number x_{i0} equals 1 and the sum $x_{i1} + x_{i2} + \dots + x_{ik}$ has a value of 1, too. So the model can be rewritten in the following form:

$$\begin{aligned} Y_i &= \mu x_{i0} + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \dots + \alpha_{k-1} x_{i(k-1)} + \alpha_k x_{ik} + E_i \\ &= \mu x_{i0} + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \dots + \alpha_{k-1} x_{i(k-1)} + \alpha_k x_{ik} \\ &\quad + E_i + \alpha_k x_{i0} - \alpha_k (x_{i1} + x_{i2} + \dots + x_{ik}) \\ &= (\mu + \alpha_k) x_{i0} + (\alpha_1 - \alpha_k) x_{i1} + (\alpha_2 - \alpha_k) x_{i2} + \dots \\ &\quad + (\alpha_{k-1} - \alpha_k) x_{i(k-1)} + (\alpha_k - \alpha_k) x_{ik} + E_i \\ &= (\mu + \alpha_k) x_{i0} + (\alpha_1 - \alpha_k) x_{i1} + (\alpha_2 - \alpha_k) x_{i2} + \dots \\ &\quad + (\alpha_{k-1} - \alpha_k) x_{i(k-1)} + E_i \end{aligned}$$

The design matrix $\mathbf{X}$ of this linear model has k columns instead of $k + 1$ and the parameter vector β has k components instead of $k + 1$:

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} & \mathbf{0}_{n_1} & \dots & \mathbf{0}_{n_1} \\ \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \mathbf{0}_{n_2} & \mathbf{0}_{n_2} & \dots & \mathbf{0}_{n_2} \\ \mathbf{1}_{n_3} & \mathbf{0}_{n_3} & \mathbf{0}_{n_3} & \mathbf{1}_{n_3} & \mathbf{0}_{n_3} & \dots & \mathbf{0}_{n_3} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \mathbf{1}_{n_{k-1}} & \mathbf{0}_{n_{k-1}} & \mathbf{0}_{n_{k-1}} & \mathbf{0}_{n_{k-1}} & \mathbf{0}_{n_{k-1}} & \dots & \mathbf{1}_{n_{k-1}} \\ \mathbf{1}_{n_k} & \mathbf{0}_{n_k} & \mathbf{0}_{n_k} & \mathbf{0}_{n_k} & \mathbf{0}_{n_k} & \dots & \mathbf{0}_{n_k} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \mu + \alpha_k \\ \alpha_1 - \alpha_k \\ \alpha_2 - \alpha_k \\ \dots \\ \alpha_{k-1} - \alpha_k \end{pmatrix}$$

Obviously, the new design matrix $\mathbf{X}$ has a full rank k . The first component of the parameter vector β is the expectation of the last sample. The further components of the parameter vector β denote differential effects; they measure the effect of the j th sample minus the effect of the last sample.

EXAMPLE 8 OF TOTAL PROTEIN IN SERUM

The new design matrix $\mathbf{X}$ and the new parameter vector β for the one-way model are:

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{12} & \mathbf{1}_{12} & \mathbf{0}_{12} \\ \mathbf{1}_{12} & \mathbf{0}_{12} & \mathbf{1}_{12} \\ \mathbf{1}_{12} & \mathbf{0}_{12} & \mathbf{0}_{12} \end{pmatrix} \quad \beta = \begin{pmatrix} \mu + \alpha_3 \\ \alpha_1 - \alpha_3 \\ \alpha_2 - \alpha_3 \end{pmatrix}$$

This matrix $\mathbf{X}$ has full rank, so the matrix $\mathbf{X}^T\mathbf{X}$ is regular. The following computations can be performed by hand or by means of a computer algebra system like Mathematica®:

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 36 & 12 & 12 \\ 12 & 12 & 0 \\ 12 & 0 & 12 \end{pmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{12} & \frac{1}{12} & \frac{-1}{12} \\ \frac{-1}{12} & \frac{1}{6} & \frac{1}{12} \\ \frac{-1}{12} & \frac{1}{12} & \frac{1}{6} \end{pmatrix}$$

The normal equations are solved uniquely by the estimation vector:

$$\mathbf{b} = \begin{pmatrix} 6.6500 \\ -1.18333 \\ -0.883333 \end{pmatrix}$$

Comparing these results with the SAS output of the procedure GLM, it is obvious that the only meaningful difference is the missing part of the 4th component. The estimates computed by SAS by means of a generalized inverse are unbiased estimators for the new set of parameters: The number 6.6500 estimates the expectation $\mu + \alpha_3$ of the TPRO value in the species Dog. For the species Mouse, the expectation of the TPRO value is $\mu + \alpha_1 = (\mu + \alpha_3) + (\alpha_1 - \alpha_3)$. It can be estimated by $6.6500 + (-1.18333) = 5.4667$. For the species Rat, the expectation of the TPRO value is $\mu + \alpha_2 = (\mu + \alpha_3) + (\alpha_2 - \alpha_3)$. It can be estimated by $6.6500 + (-0.883333) = 5.7667$.

Obviously, these estimators are identical with the averages $\bar{y}_1$, $\bar{y}_2$, and $\bar{y}_3$ of the three samples corresponding to the single species which were computed by the SAS procedure MEANS, and the question arises if this linear model is worth the effort. Moreover, it has to be admitted that for a one-way ANOVA, it would be possible to avoid a less-than-full-rank design matrix $\mathbf{X}$ in choosing a no-intercept model where the design matrix $\mathbf{X}$ arises from the original design matrix by canceling the first column. For the no-intercept linear model, the design matrix and the parameter vector are given by:

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{12} & \mathbf{0}_{12} & \mathbf{0}_{12} \\ \mathbf{0}_{12} & \mathbf{1}_{12} & \mathbf{0}_{12} \\ \mathbf{0}_{12} & \mathbf{0}_{12} & \mathbf{1}_{12} \end{pmatrix} \quad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$$

The justification for leaving out the possibility of a no-intercept model is the fact that this method does not work in the case of a two-way ANOVA, as the sum of the last p columns again gives a column where all elements have a value of 1. So we have to go on with the method of recoding the $1 + r + p$ parameters. For the sake of simplicity, we demonstrate the following method for our example.

EXAMPLE 9 OF TOTAL PROTEIN IN SERUM

We make use of the fact that for every i , the following equations are true:

$$x_{i0} = 1, \quad x_{i1} + x_{i2} + x_{i3} = 1, \quad x_{i4} + x_{i5} = 1$$

So the model can be rewritten in the following form:

$$\begin{aligned} Y_i &= \mu x_{i0} + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \beta_1 x_{i4} + \beta_2 x_{i5} + E_i \\ &= \mu x_{i0} + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \beta_1 x_{i4} + \beta_2 x_{i5} + E_i \\ &\quad + (\alpha_3 + \beta_2) x_{i0} - \alpha_3 (x_{i1} + x_{i2} + x_{i3}) - \beta_2 (x_{i4} + x_{i5}) \\ &= (\mu + \alpha_3 + \beta_2) x_{i0} + (\alpha_1 - \alpha_3) x_{i1} + (\alpha_2 - \alpha_3) x_{i2} \\ &\quad + (\beta_1 - \beta_2) x_{i4} + E_i \end{aligned}$$

The design matrix $\mathbf{X}$ of this linear model has four columns instead of six and the parameter vector $\boldsymbol{\beta}$ has four components instead of six:

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_6 & \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{1}_6 \\ \mathbf{1}_6 & \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{0}_6 \\ \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{1}_6 & \mathbf{1}_6 \\ \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{1}_6 & \mathbf{0}_6 \\ \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{0}_6 & \mathbf{1}_6 \\ \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{0}_6 & \mathbf{0}_6 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu + \alpha_3 + \beta_2 \\ \alpha_1 - \alpha_3 \\ \alpha_2 - \alpha_3 \\ \beta_1 - \beta_2 \end{pmatrix}$$

Obviously, the new design matrix $\mathbf{X}$ has a full rank of 4. The first component of the parameter vector $\boldsymbol{\beta}$ is the expectation of the last sample (i.e., of male rats). The next two components of the parameter vector $\boldsymbol{\beta}$ again denote differential effects; they measure the effect of the species Dog minus the effect of the species Rat and the effect of the species Mouse minus the effect of the species Rat. The last component of $\boldsymbol{\beta}$ measures the differential effect of Sex: female minus male. Submitting the SAS commands:

```
proc glm data=TPRO;
class Species Sex;
model value=Species Sex/solution;
run;
```

yields the estimation:

$$\mathbf{b} = \begin{pmatrix} 6.70277778 \\ -1.18333333 \\ -0.90833333 \\ -0.10555556 \end{pmatrix}$$

with the note that the estimates are biased and not unique. Indeed, the output does not allow to compute unbiased estimators for the original parameters μ , α_1 , α_2 , α_3 , β_1 , and β_2 . But with the correct interpretation of the output, the expectation of the random variable TPRO for each of the six samples defined by Species and Sex can be estimated unbiasedly:

Dog, female	$6.70277778 - 1.18333333 - 0.10555556 \approx 5.41389$
Dog, male	$6.70277778 - 1.18333333 \approx 5.51944$
Mouse, female	$6.70277778 - 0.90833333 - 0.10555556 \approx 5.68889$
Mouse, male	$6.70277778 - 0.90833333 \approx 5.79444$
Rat, female	$6.70277778 - 0.10555556 \approx 6.59722$
Rat, male	$6.70277778 \approx 6.70278$

At the end of this contribution, we make some remarks concerning additional possibilities of the linear model. The

further discussion of these possibilities is beyond the scope of this article.

REMARK 1

Neither the method of generalized inverses nor the described method of recoding the parameters allows to compute unbiased estimators for the parameters μ , α_1 , α_2 , ..., α_k or μ , α_1 , ..., α_r , β_1 , ..., β_p of the original linear model of the one-way or the two-way ANOVA, respectively. But there are linear functions of these parameters for which unbiased estimates do exist. In the one-way case, all sums $\mu + \alpha_i$, $1 \leq i \leq k$ for instance can be estimated. In the two-way case, all sums $\mu + \alpha_i + \beta_j$, $1 \leq i \leq r$, $1 \leq j \leq p$ can be estimated.

REMARK 2

It was emphasized that for the task of estimating the parameters of a full-rank linear model, special distribution assumptions are not necessary. So up to now, in our example, the fact that the TPRO value can be assumed to be normally distributed was not taken advantage of. But with this normal distribution in mind, one can test the hypotheses on the parameters. SAS, for instance, automatically tests the hypothesis that this component has a value of 0 using a Student-distributed test variable for each component of the parameter vector. The output shows: The estimators for the intercept and for the differential effects “Dog minus Rat” and “Mouse minus Rat” differ significantly from 0, whereas the last component estimating the differential effect “Female minus Male” does not.

Within the frame of the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$ of ANOVA, it is possible to test more general hypotheses of the form $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$. Here $\mathbf{C}$ is a matrix with an appropriate number of rows and columns. In the two-way model for the TPRO value for instance, the matrix $\mathbf{C}$ could have one single row: With $\mathbf{C} = (0 \ 1 \ 0 \ 0)$, the hypothesis $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ means that $\alpha_1 - \alpha_3 = 0$ (i.e., there is no difference between the species Dog and the species Rat). With $\mathbf{C} = (0 \ 1 \ 0 \ -10)$, the hypothesis $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ means that $\alpha_1 - \alpha_3 = 10(\beta_1 - \beta_2)$ (i.e., the difference between the effects of the species Dog and species Rat is 10 times as great as the difference between the effects Female and Male).

REMARK 3

In the linear model of a two-way or higher-way ANOVA, it is possible to include interactions. In the example of TPRO, there are six interaction parameters. Including them in the model increases the number of columns of the design matrix by six, but the rank is only increased by two. So one has to reduce the number of interaction terms from six to two, and it is rather difficult to explain the meaning of the recoded parameters.

REMARK 4

It should be mentioned that there is another recoding method for solving the problem of a less-than-full-rank design matrix in the linear model of ANOVA. Making use of the reparameterization conditions, the number of parameters for each factor variable can be reduced by 1. The resulting design matrix $\mathbf{X}$ has elements with values 0, +1, or -1. In the example of TPRO for instance, we have the two conditions $\alpha_1 + \alpha_2 + \alpha_3 = 0$ and $\beta_1 + \beta_2 = 0$. These two equations yield:

$$\alpha_3 = -\alpha_1 - \alpha_2 \quad \text{and} \quad \beta_2 = -\beta_1$$

So the original model equations can be rewritten in the following form:

$$\begin{aligned} Y_i &= \mu x_{i0} + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \beta_1 x_{i4} + \beta_2 x_{i5} + E_i \\ &= \mu x_{i0} + \alpha_1 x_{i1} + \alpha_2 x_{i2} + (-\alpha_1 - \alpha_2) x_{i3} + \beta_1 x_{i4} \\ &\quad + (-\beta_1) x_{i5} + E_i \\ &= \mu x_{i0} + \alpha_1 (x_{i1} - x_{i3}) + \alpha_2 (x_{i2} - x_{i3}) \\ &\quad + \beta_1 (x_{i4} - x_{i5}) + E_i \end{aligned}$$

Again the design matrix $\mathbf{X}$ of this new linear model with only four columns instead of six has a full rank and the parameter vector β has four components instead of six:

$$\mathbf{X} = \begin{pmatrix} 1_6 & 1_6 & 0_6 & 1_6 \\ 1_6 & 1_6 & 0_6 & -1_6 \\ 1_6 & 0_6 & 1_6 & 1_6 \\ 1_6 & 0_6 & 1_6 & -1_6 \\ 1_6 & -1_6 & -1_6 & 1_6 \\ 1_6 & -1_6 & -1_6 & -1_6 \end{pmatrix} \quad \beta = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \end{pmatrix}$$

The advantage of this reparameterization method is that the meaning of the parameters is left unchanged. This method is not supported by the SAS procedure GLM, but it is used in the procedure CATMOD for the linear model in categorical data analysis, where the response variable is categorical as well.^[3,6]

REMARK 5

In many animal studies, each animal is measured repeatedly through time. In the ideal case, the number of time

points is identical for each animal. Performing a one-way or a higher-way ANOVA in treating “time point” as a factor variable violates the primary assumption of independence.

There exist different methods for solving this problem. If the influence of time is negligible, the simplest method is to choose one value per animal. For example, one can choose the first observation, or one can choose one observed value at random. If (and only if) the set of time points is identical for each animal, a time-by-time ANOVA can be performed. It consists of a number of separate analyses, one for each time of observation.

The repeated-measures ANOVA, however, includes the influence of time. It requires a modification of the underlying linear model, adding further parameters for the interactions between treatments and time points. By the repeated statement, the SAS procedures ANOVA and GLM offer the opportunity of performing a repeated-measures ANOVA.

The repeated-measures ANOVA and further GLMs for the analysis of longitudinal data like marginal models, random effects models, and transition models are discussed in Ref. [1].

REFERENCES

1. Diggle, P.J.; Liang, K.-Y.; Zeger, S.L. *Analysis of Longitudinal Data*; Clarendon Press: Oxford, 1994.
2. Fisher, R.A. *The Design of Experiments*; Oliver and Boyd: Edinburgh, 1947.
3. Forthofer, R.N.; Lehnen, R.G. *Public Program Analysis: A New Categorical Data Approach*; Lifetime Learning Publ: Belmont, CA, 1981.
4. Hocking, R.R. *Methods and Applications of Linear Models: Regression and the Analysis of Variance*; Wiley-Interscience: New York; 1996.
5. Lindman, H.R. *Analysis of Variance in Experimental Design*; Springer-Verlag: New York, 1992.
6. Littell, R.C.; Freund, R.J.; Spector, Ph C. *SAS System for Linear Models, 3rd ed.*; SAS Institute Inc.: Cary, NC, 1991.
7. Roberts, M.J.; Russo, R. *Student's Guide to Analysis of Variance*; Routledge: London; 1999.
8. Sahai, H.; Ageel, M. *The Analysis of Variance: Fixed, Random and Mixed Models*; Springer: New York; 2000.
9. Scheffé, H. *The Analysis of Variance*; John Wiley & Sons: New York, 1959.
10. Searle, S.R. *Linear Models*; John Wiley and Sons: New York, 1971.

Analytical Similarity Assessment

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Fuyu Song

Center for Food and Drug Inspection, China Food and Drug Administration, Beijing, China

He Bai

Department of Drug & Cosmetics Registration, China Food and Drug Administration, Beijing, China

Abstract

For assessment of biosimilarity for biosimilar products, the United States (US) Food and Drug Administration (FDA) proposed a stepwise approach for providing the totality of the evidence of similarity between a proposed biosimilar product and a US-licensed (reference) product. The stepwise approach starts with assessment of critical quality attributes (CQAs) that are relevant to clinical outcomes in structural and functional characterization in the manufacturing process of the proposed biosimilar product. The FDA suggests that these CQAs be identified and classified into three tiers depending on their criticality or risk ranking. To assist the sponsors, the FDA also suggests some statistical approaches for the assessment of analytical similarity for CQAs from different tiers, namely equivalence test for Tier 1, quality range approach for Tier 2, and descriptive raw data and graphical comparison for Tier 3. In this article, challenging issues to the FDA's recommended approaches are discussed followed by alternative methods for the assessment of analytical similarity (mainly for CQAs from Tier 1).

Keywords: Equivalence test; Fixed SD approach; Quality range approach; Stepwise approach; Tiered approach; Totality of the evidence.

INTRODUCTION

In recent years, the assessment of biosimilarity for biosimilar products has received much attention by scientists, researchers, and reviewers from the pharmaceutical industry (biosimilar sponsors), academia, and regulatory agencies such as the United States Food and Drug Administration (FDA) and the China Food and Drug Administration (CFDA).^[6] As indicated in its recent guidance, the FDA recommends a stepwise approach for obtaining the totality of the evidence for demonstrating biosimilarity between a proposed biosimilar product and an innovative (reference) biological product.^[1,9] The stepwise approach starts with analytical studies for functional and structural characterization of critical quality attributes (CQAs) that are relevant to clinical outcomes at various stages of the manufacturing process followed by animal studies for assessment of toxicity, clinical pharmacology pharmacokinetic (PK) or pharmacodynamic (PD) studies, and clinical studies for assessment of immunogenicity, safety/tolerability, and efficacy.^[2,3,5] For the analytical studies, the FDA suggests that CQAs be identified and classified into three tiers according to their criticality or risk ranking based on the mechanism of action (MOA) or PK using appropriate

statistical models or methods. CQAs that are most relevant to clinical outcomes are classified as Tier 1, whereas CQAs that are less (mild-to-moderate) or least relevant to clinical outcomes are classified as Tier 2 and Tier 3, respectively.

The FDA proposes equivalence tests for CQAs in Tier 1, quality range approach for CQAs in Tier 2, and raw data or graphical presentation for CQAs in Tier 3 to assist sponsors in analytical similarity assessment for obtaining the totality of the evidence for demonstrating similarity between the proposed biosimilar product and the reference product.^[7,12] As indicated by the FDA, the equivalence test for Tier 1 CQAs is more statistically rigorous than that of the quality range approach for Tier 2 CQAs, which is in turn more rigorous than that of raw data or graphical presentation for Tier 3 CQAs.^[12] In practice, however, for a given CQA, there is no guarantee that passing a Tier 1 test will pass a Tier 2 test and vice versa. This has raised a number of debatable (controversial) issues in analytical similarity assessment.^[2,3,5] These issues include, but are not limited to, (1) fundamental similarity assumption, (2) primary assumptions for the tiered approach, (3) statistical properties of the FDA's recommended Tier 1 equivalence test, (4) criticism of the fixed approach for margin/range selection, (5) inconsistencies between tiered approaches,

(6) sample size requirement, (7) heterogeneity within lots and across lots within and between test product and reference product, (8) interpretation of the FDA's current thinking on scientific input, (9) relationship between similarity limit and variability, and (10) a proposed unified tiered approach.

Analytical similarity assessment is a complicated problem that involves characterization of bioactivity and protein content whose relationship with clinical outcomes may not be fully explored and understood. This uncertainty has made analytical similarity assessment even more complicated. The purpose of this article is to give a brief summary of these controversial issues in analytical similarity assessment rather than provide solutions.

FUNDAMENTAL SIMILARITY ASSUMPTION

For small-molecule drug products, bioequivalence studies are necessarily conducted for regulatory review and approval of small-molecule generic drug products.^[8] This is because it constitutes legal basis (from the *Hatch-Waxman* Act) under the *Fundamental Bioequivalence Assumption*, which states that

If two drug products are shown to be bioequivalent, it is assumed that they will reach the same therapeutic effect or they are therapeutically equivalent.

Under the Fundamental Bioequivalence Assumption, bioavailability (defined as the rate and extent of drug absorbed into the bloodstream and become available) serves as a surrogate endpoint for clinical outcomes (safety and efficacy). Thus, under the Fundamental Bioequivalence Assumption, an *approved* generic drug product can serve as a substitute for the innovative (brand-name) drug product. Although this Fundamental Bioequivalence Assumption constitutes legal basis, it has been challenged by many researchers. In practice, there are following four possible scenarios:

1. Drug absorption profiles are similar, and they are therapeutic equivalent.
2. Drug absorption profiles are not similar, but they are therapeutic equivalent.
3. Drug absorption profiles are similar, but they are not therapeutic equivalent.
4. Drug absorption profiles are not similar, and they are not therapeutic equivalent.

The Fundamental Bioequivalence Assumption is considered scenario 1. Scenario 1 works if the drug absorption (in terms of the rate and extent of absorption) is predictive of clinical outcome. In this case, PK responses such as AUC (area under the blood or plasma concentration–time curve for measurement of the extent of drug absorption) and C_{max} (maximum concentration for measurement of the rate of

drug absorption) serve as surrogate endpoints for clinical endpoints for the assessment of efficacy and safety of the test product under investigation. Scenario 2 is the case in which generic companies argue for generic approval of their drug products especially when their products fail to meet regulatory requirement for bioequivalence. In this case, it is doubtful that there is a relationship between PK responses and clinical endpoints. The innovator companies usually argue with the regulatory agency for generic approval with scenario 3. However, more studies are necessarily conducted in order to verify scenario 3. There are no arguments with respect to scenario 4.

It should be noted that a generic drug contains *identical* active ingredient(s) as the brand-name drug. Thus, it is reasonable to assume that the generic (test) drug products and the brand-name (reference) drug products have identical means, i.e., $\mu_T = \mu_R$. In addition, bioequivalence testing focuses on the mean difference (i.e., $\mu_T - \mu_R$) or the ratio of means (i.e., μ_T / μ_R) and ignores heterogeneity in variability of the test and reference products (i.e., $\sigma_T \neq \sigma_R$). As a result, a generic drug product may fail the bioequivalence testing when σ_R is relatively large (say > 30%) even when $\mu_T = \mu_R$. Unlike small-molecule drug products, biosimilar products are large-molecule drug products which are made of living cells or living organisms. Chow et al.^[4] indicated that there are fundamental differences between small-molecule drug products and biosimilar products. For example, biosimilar products are often very sensitive to environmental factors during the manufacturing process. A small change and variation may translate to a huge change in clinical outcomes. Consequently, biosimilar products are expected to have much larger variability compared to that of generic drug products. In this case, statistical methods for similarity assessment following the concept of bioequivalence testing (i.e., focusing on the mean difference or the ratio of means but ignoring variability) for assessing the biosimilarity of biosimilar products may not be appropriate. [Table 1](#) provides a comparison between bioequivalence test for generic drug products and biosimilarity test for biosimilar products.

Following a similar idea to bioequivalence assessment for small-molecule drug products, Scientific Statistical Advisory Board on Biosimilars^[11] proposed the following Fundamental Similarity Assumption:

When a follow-on biologic product is claimed to be biosimilar to an innovator product in some well-defined study endpoints, it is assumed that they will reach similar therapeutic effect or they are therapeutically equivalent.

Although the aforementioned proposed Fundamental Similarity Assumption does not constitute legal basis, the FDA seems to adopt the assumption for analytical similarity assumption without verifying the validity of the assumption. In other words, the FDA assumes that analytical similarity in terms of CQAs identified at various stages

Table 1 Comparison of Bioequivalence and Biosimilarity Testing

Characteristics	Bioequivalence		Biosimilarity	
	In vitro BE testing	In vivo BE testing	Analytical	PK/clinical
Fundamental assumption	√	√	×	×
Log data	×	√	×	×
Primary focus	Mean	Mean	Mean	Mean
Variability	<10%	20%–30%	<10%	40%–50%
Criterion	(90%, 111%)	(80%, 125%)	$EAC = \pm 1.5\sigma_R$	SABE?
Analysis	Profile/non-profile	Hypothesis/CI	Hypothesis/CI	Hypothesis/CI Biosimilarity index

such as functional and structural characterization of the manufacturing process is predictive of clinical outcomes.

For biosimilars, it can be expected that $\mu_T \neq \mu_R$ and $\sigma_T \neq \sigma_R$, i.e., $\mu_T = \mu_R + \Delta$ and $\sigma_T = C\sigma_R$, where Δ is the true mean difference and C is an inflation factor for the variability.

PRIMARY ASSUMPTIONS FOR THE TIERED APPROACH

Tier 1 Equivalence Test

For CQAs in Tier 1, the FDA recommends that an equivalence test be performed for assessing analytical similarity. As indicated by the FDA, for a given CQA, we may test for equivalence by the following interval (null) hypothesis:

$$H_0: \mu_T - \mu_R \leq -\delta \text{ or } \mu_T - \mu_R \geq \delta,$$

where $\delta > 0$ is the equivalence limit (or similarity margin), and μ_T and μ_R are the mean responses of the test (the proposed biosimilar) product and reference product lots, respectively. Analytical equivalence (similarity) is concluded if the null hypothesis of nonequivalence (*dis-similarity*) is rejected. Note that inequivalence can be defined as when the confidence interval falls entirely outside the equivalence limits. Similarly to the confidence interval approach for bioequivalence testing under the raw data model, analytical similarity would be accepted for a quality attribute if the $(1 - 2\alpha)100\%$ two-sided confidence interval of the mean difference is within $(-\delta, \delta)$. The FDA further recommended that the equivalence acceptance criterion (EAC), $\delta = EAC = 1.5\sigma_R$, where σ_R is the variability of the reference product being used based on extensive simulation studies and internal scientific input. Chow^[3] provided statistical justification for the selection of $c = 1.5$ in EAC following the idea of scaled average bioequivalence (SABE) criterion for highly variable drug products proposed by the FDA.^[10]

For the establishment of EAC, the FDA made the following assumptions: First, the FDA assumes that the true

difference in means is proportional to σ_R , i.e., $\mu_T - \mu_R$ is proportional to σ_R . Second, the FDA adopts the similarity limit as $EAC = 1.5\sigma_R$ and recommends that σ_R be estimated by the sample standard deviation (SD) of test values from reference lots (one test value from each lot). Third, in the interest of achieving a desired power of the similarity test, the FDA further recommends that an appropriate sample size be selected by evaluating the power under the alternative hypothesis at $\mu_T - \mu_R = \frac{1}{8}\sigma_R$. The assumption that $\mu_T - \mu_R$ is proportional to σ_R , the selection of $c = 1.5$, and the allowed mean shift of $\mu_T - \mu_R = \frac{1}{8}\sigma_R$ have generated tremendous discussion among the FDA, biosimilar sponsors, and academia, and they are debatable.

To provide a better understanding of the debatable issue and the FDA's proposal, we would like to point out the following, which may be helpful to resolve the debatable issues: (1) unlike the traditional bioequivalence test, the FDA's intention is to take variability into account by considering the effect size adjusted for variability, i.e.,

$$\text{Effect size} = \text{eff} = \frac{\mu_T - \mu_R}{\sigma_R} = \frac{\frac{1}{8}\sigma_R}{\sigma_R} = \frac{1}{8} = 0.125,$$

which is halfway between 1 and 1.25 (unity to the upper equivalence limit of 125%); (2) the EAC for effect size adjusted for variability becomes fixed, i.e., $EAC = c = 1.5$; and (3) if the true difference falls on the halfway between 1 and 1.25, the worst possible observed difference could fall on 1.25 (this may happen if the worst possible reference lot is selected for comparison). In this case, the original EAC (bioequivalence testing for generic drug products with $\mu_T = \mu_R$) is necessarily shifted by 0.25. Thus, the upper limit is shifted from 1.25 to $c = 1.25 + 0.25 = 1.5$.

Tier 2 Quality Range Approach

For CQAs in Tier 2, the FDA suggests that analytical similarity be performed based on the concept of quality ranges, i.e., $\pm x\sigma_R$, where σ_R is the SD of the reference product and x is a constant which should be

appropriately justified. Thus, the quality range of the reference product for a specific quality attribute is defined as $(\hat{\mu}_R - x\hat{\sigma}_R, \hat{\mu}_R + x\hat{\sigma}_R)$. Analytical similarity would be accepted for the quality attribute if a sufficiently large percentage of test lot values falls within the quality range. Under normality assumption, if $x = 1.645$, we would expect 90% of the test results from reference lots to lie within the quality range. If x is chosen to be 1.96, we would expect that about 95% test results of reference lots will fall within the quality range. Thus, the selection of x could have an impact on the width of the quality range and consequently the percentage of test lot values that will fall within the quality range.

At the 2015 Duke Industry Statistics Symposium held in Duke Campus on October 22–23, one of the FDA speakers indicated that x should be selected between 2 and 3 to guarantee that the majority of test values of the test lots will fall within the quality range established based on the test values of the reference lots. Under the normality assumption, in practice, we would expect that there are about 95% of data would fall within $\pm 2SD$ (i.e., $x = 2$) of the mean, and about 99.7% of data would fall within $\pm 3SD$ (i.e., $x = 3$) of the mean. Under the normality assumption, the FDA recommended the quality range approach; it is considered a reasonable approach only under the assumption that $\mu_T \approx \mu_R$ and $\sigma_T \approx \sigma_R$. In this case, it can be expected that the majority of test values obtained from the test lots will fall within $\pm xSD$ of the range established based on the test values of the reference lots.

In practice, however, the assumption that $\mu_T \approx \mu_R$ and $\sigma_T \approx \sigma_R$ are usually not true due to the nature of biosimilar products. Thus, one of the major criticisms of the quality range approach is that it ignores the fact that there are differences in population mean and population SD between the proposed biosimilar product and the reference product, i.e., $\mu_T \neq \mu_R$ and $\sigma_T \neq \sigma_R$. In practice, it is recognized that biosimilarity between a proposed biosimilar product and a reference product could be established even under the assumption that $\mu_T \neq \mu_R$ and $\sigma_T \neq \sigma_R$. Thus, under the assumption that $\mu_T \approx \mu_R$ and $\sigma_T \approx \sigma_R$, the quality range approach for analytical similarity assessment for CQAs from Tier 2 is considered more stringent compared to equivalence testing for CQAs from Tier 1 (most relevant to clinical outcomes), regardless of the fact that they are mild-to-moderate relevant to clinical outcomes. This is because that equivalence testing allows a possible mean shift of $\sigma_R/8$, whereas the quality range approach does not. In practice, there are several possible scenarios that include the cases where (1) $\mu_T \approx \mu_R$ or there is a significant mean shift (either a shift to the right or a shift to the left), and (2) $\sigma_T \approx \sigma_R$, $\sigma_T > \sigma_R$, or $\sigma_T < \sigma_R$.

Thus, one of the most controversial issues for the quality range approach for CQAs in Tier 2 is that the approach does not reflect the real practice that $\mu_T \neq \mu_R$ and $\sigma_T \neq \sigma_R$. As a result, the test results are somewhat misleading and not reliable.

Tier 3 Raw Data and Graphical Comparison

For CQAs in Tier 3 with the lowest risk ranking, the FDA recommends an approach that uses raw data/graphical comparisons. The examination of similarity for CQAs in Tier 3 by no means is less stringent, which is acceptable because they have least impact on clinical outcomes in the sense that a notable dis-similarity will not affect clinical outcomes.

Evaluation based on raw data and graphical presentation is not only somewhat subjective but also biased. Tier 1 equivalence tests and Tier 2 quality range similarity tests are supposed to be more rigorous than Tier 3 raw data and graphical comparison. That is, passing the Tier 1 equivalence test and Tier 2 quality range similarity test will pass the Tier 3 raw data and graphical comparison test. In practice, however, there is no guarantee that a given CQA that passes a Tier 1 equivalence test or Tier 2 quality range similarity test will pass a Tier 3 raw data and graphical comparison test and vice versa. Since CQAs in Tier 3 are considered least relevant to clinical outcomes, it is necessary that all Tier 3 CQAs pass the test. If not, it is of interest to know about what percentage of CQAs need to pass in order to pass a Tier 3 test.

STATISTICAL PROPERTIES OF THE FDA'S TIER 1 EQUIVALENCE TEST

For equivalence test for CQAs in Tier 1, the FDA recommends testing one sample from each reference lot for obtaining an estimate of σ_R . Wang and Chow^[13] evaluated statistical properties of the FDA's recommended method. Without loss of generality, suppose multiple test samples from each reference lot are available. Let x_{Rij} be the test value of the j th test sample from the i th reference lot, $i = 1, \dots, k, j = 1, \dots, n_i$, and it follows a normal distribution with mean μ_i and variance σ_i^2 , where μ_i and σ_i^2 are also random variables. The expectations of μ_i and σ_i^2 are μ and σ^2 , and the variances are σ_μ^2 and σ_σ^2 . Then, the variance of x_{Rij} is given by

$$\begin{aligned}\sigma_R^2 &= \text{Var}(x_{Rij}) \\ &= \text{Var}\left(E(x_{Rij} | \mu_i, \sigma_i^2)\right) + E\left(\text{Var}(x_{Rij} | \mu_i, \sigma_i^2)\right) \\ &= \text{Var}(\mu_i) + E(\sigma_i^2) = \sigma_\mu^2 + \sigma^2.\end{aligned}$$

Define $\bar{x}_{Ri} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{Rij}$, $\bar{x}_R = \frac{1}{k} \sum_{i=1}^k \bar{x}_{Ri}$. Then,

$$\begin{aligned}E(\bar{x}_{Ri}) &= \frac{1}{n_i} \sum_{j=1}^{n_i} E(x_{Rij} | \mu_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} E(\mu_i) = \mu \\ \text{Var}(\bar{x}_{Ri}) &= \text{Var}\left(E(\bar{x}_{Ri} | \mu_i, \sigma_i^2)\right) + E\left(\text{Var}(\bar{x}_{Ri} | \mu_i, \sigma_i^2)\right) \\ &= \text{Var}(\mu_i) + \frac{1}{n_i} E(\sigma_i^2) = \sigma_\mu^2 + \frac{1}{n_i} \sigma^2,\end{aligned}$$

where

$$E(\bar{x}_{R..}) = \frac{1}{k} \sum_{i=1}^k E(\bar{x}_{Ri.}) = \mu$$

$$Var(\bar{x}_{R..}) = \frac{1}{k^2} \sum_{i=1}^k Var(\bar{x}_{Ri.}) = \frac{1}{k} \sigma_{\mu}^2 + \frac{\sigma^2}{k^2} \left(\sum_{i=1}^k n_i \right)$$

If we assume that $n_1 = \dots = n_k = n$, then we have

$$\hat{\sigma}_R^2 = \frac{1}{nk-1} \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{Rij} - \bar{x}_{R..})^2$$

$$E(\hat{\sigma}_R^2) = \frac{1}{nk-1} \sum_{i=1}^k \sum_{j=1}^{n_i} E \left((x_{Rij} - \mu)^2 + (\bar{x}_{R..} - \mu)^2 - 2(x_{Rij} - \mu)(\bar{x}_{R..} - \mu) \right)$$

$$= \frac{1}{nk-1} \sum_{i=1}^k \sum_{j=1}^{n_i} (Var(x_{Rij}) - Var(\bar{x}_{R..}))$$

$$= \frac{1}{nk-1} (n(k-1)\sigma_{\mu}^2 + (nk-1)\sigma_{\mu}^2) = \sigma_{\mu}^2 + \sigma^2 - \frac{n-1}{nk-1} \sigma_{\mu}^2.$$

Thus, it can be seen that the FDA's approach is an unbiased estimate of σ_R . However, when $n > 1$, the FDA's recommended approach with multiple test samples per lot is a biased estimate of σ_R . In the interest of having an unbiased estimate with multiple test samples per lot, Wang and Chow^[13] proposed an alternative approach to correct the biasedness. As indicated by Wang and Chow,^[13] multiple test samples per lot provide valuable information regarding the heterogeneity across lots, which is useful especially when extreme lots (i.e., lots with extremely low or high variability) are selected for equivalence test.

ISSUES OF FIXED MARGIN/RANGE IN THE TIERED APPROACH

In the tiered approach, the FDA seems to prefer a *fixed margin approach* for EAC by treating $s = \hat{\sigma}_R$ as the true σ_R (i.e., $1.5\sigma_R$) for Tier 1 CQAs. The fixed margin approach is also referred to as a *fixed SD approach* for similarity quality range by treating the estimate of SD (which is obtained based on the test values of reference lots) as the true σ_R (i.e., xSD) for Tier 2 CQAs. The fixed margin or SD is in fact a statistic, which is a random variable rather than a fixed constant. In other words, it may vary depending upon the selected reference lots for tiered testing. The fixed SD approach is a conditional approach rather than an unconditional approach. In practice, if reference lots with *less* variability are selected for tiered testing, the proposed biosimilar product is most likely to fail the test. As

a result, the scientific validity of the fixed SD approach is questionable.

One of the major criticisms of the fixed approach for margin/range selection in tiered analysis is that the fixed approach does not take into consideration the variability of the estimate of the SD. Thus, it is considered *bad luck* to the biosimilar sponsors if reference lots with less variability are selected for a Tier 1 equivalence test. Another criticism is that the reference product cannot pass a Tier 1 equivalence test itself if we divide all of the reference lots into two groups: one group with less variability and the other group with large variability. In this case, the group with large variability may not pass a Tier 1 equivalence test with EAC established based on the test values from the reference lots with less variability. It would be a concern that the reference product cannot pass a Tier 1 equivalence test when comparing to itself.

INCONSISTENT RESULTS BETWEEN TIERED APPROACHES

As indicated in Tsong,^[12] a Tier 1 equivalence test is considered more rigorous than a Tier 2 quality range approach, which is in turn more rigorous than a Tier 3 raw data and graphical comparison. The primary assumptions for these tiered approaches, however, are different. Thus, under different assumptions, there is no guarantee that passing a Tier 1 equivalence test will pass a Tier 2 quality range approach and vice versa although these tests are conducted based on the same data set collected from the test and reference lots under study. In practice, it is of interest to evaluate inconsistencies regarding the passages between a Tier 1 equivalence test and Tier 2 quality range approach.

For a given CQA, the inconsistencies between a Tier 1 equivalence test and a Tier 2 quality range approach can be assessed by means of clinical trial simulation as follows: Let p_{ij} be the probability of passing the i th tier test given that the CQA has passed the j th tier test. Thus, we have the following 2×2 contingency table for comparison between a Tier 1 equivalence test and Tier 2 quality range approach (Table 2).

Let $\mu_T = \mu_R + \Delta$ and $\sigma_T = C\sigma_R$. It is then suggested that the inconsistencies between a Tier 1 equivalence test and Tier 2 quality range approach be evaluated at various combinations of (1) $\Delta = 0$ (no mean shift), $\frac{1}{8}\sigma_R$ (FDA recommended mean shift allowed), and $\frac{1}{4}\sigma_R$ (the worst possible

Table 2 Probabilities of Inconsistencies

Tier 1 equivalence test	Tier 2 quality range approach	
	Pass	Fail
Pass	p_{11}	p_{12}
Fail	p_{21}	p_{22}

scenario); and (2) $C = 0.8$ (deflation), 1.0, and 1.2 (inflation) to provide a complete picture of the relative performance of a Tier 1 equivalence test and Tier 2 quality range approach.

SAMPLE SIZE REQUIREMENT

One of the most commonly asked questions for analytical similarity assessment is probably that of how many reference lots are required for establishing an acceptable EAC to achieve a desired power. For a given EAC, formulas for sample size calculation under different study designs are available in the work of Chow, Shao, and Wang (2008). In general, sample size (the number of reference lots, k) required is a function of (1) overall type I error rate (α), (2) type II error rate (β) or power ($1 - \beta$), (3) clinically or scientifically meaningful difference (i.e., $\mu_T - \mu_R$), and (4) the variability associated with the reference product (i.e., σ_R) assuming that $\sigma_T = \sigma_R$. Thus, we have

$$k = f(\alpha, \beta, \mu_T - \mu_R, \sigma).$$

In practice, we select an appropriate k for achieving a desired power of $1 - \beta$ for detecting a clinically meaningful difference of $\mu_T - \mu_R$ at a prespecified level of significance α , assuming that the true variability is σ . If α , $\mu_T - \mu_R$, and σ are fixed, the above equation becomes $k = f(\beta)$. We can then select an appropriate k for achieving the desired power. The FDA's recommendation attempts to control all parameters at the desired levels (e.g., $\alpha = 0.05$ and $1 - \beta = 0.8$) by *knowing* that $\mu_T - \mu_R$ and σ are varying. In practice, it is often difficult, if not impossible, to control (or find a balance point among) α (type I error rate), $1 - \beta$ (power), $\mu_T - \mu_R = \Delta$ (clinically meaningful difference), and σ (variability in observing the response) *at the same time*. For example, controlling α at a prespecified level of significance may be at the risk of decreasing power with a selected sample size.

HETEROGENEITY WITHIN/BETWEEN LOTS WITHIN/BETWEEN TEST AND REFERENCE PRODUCTS

Suppose there are n_R and n_T lots for analytical similarity assessment. For a given reference (test) lot, assume that the test value follows a distribution with mean μ_{Ri} (μ_{Ti}) and variance σ_{Ri}^2 (σ_{Ti}^2). FDA's recommended approach assumes that $\mu_{Ri} = \mu_{Rj}$ and $\sigma_{Ri}^2 = \sigma_{Rj}^2$ for $i \neq j, i, j = 1, \dots, n_R$ and $\mu_{Ti} = \mu_{Tj}$ and $\sigma_{Ti}^2 = \sigma_{Tj}^2$ for $i \neq j, i, j = 1, \dots, n_T$ for equivalence test in Tier 1 and quality range approach in Tier 2. Now let σ_R^2 and σ_T^2 be the variabilities associated with the reference product and the test product, respectively. Thus, we have

$$\sigma_R^2 = \sigma_{WR}^2 + \sigma_{BR}^2 \text{ and } \sigma_T^2 = \sigma_{WT}^2 + \sigma_{BT}^2,$$

where $\sigma_{WR}^2, \sigma_{BR}^2$ and $\sigma_{WT}^2, \sigma_{BT}^2$ are the within-lot variability and the between-lot (lot-to-lot) variability for the reference product and the test product, respectively. In practice, it is very likely that $\sigma_R^2 \neq \sigma_T^2$ and often $\sigma_{WR}^2 \neq \sigma_{WT}^2$ and $\sigma_{BR}^2 \neq \sigma_{BT}^2$ even when $\sigma_R^2 \approx \sigma_T^2$. This has posted a major challenge to the FDA's proposed approaches for the assessment of analytical similarity for CQAs from both Tier 1 and Tier 2, especially when there is only one test sample from each lot from the reference product and the test product. FDA's proposal ignores within-lot variability, i.e., when $\sigma_{WR}^2 = 0$ or $\sigma_R^2 = \sigma_{BR}^2$. In other words, sample variance based on $x_i, i = 1, \dots, n_R$ from the reference product may underestimate the true σ_R^2 and consequently may not provide a fair and reliable assessment of analytical similarity for a given quality attribute.

In practice, it is well recognized that $\mu_{Ri} \neq \mu_{Rj}$ and $\sigma_{Ri}^2 \neq \sigma_{Rj}^2$ for $i \neq j$, where μ_{Ri} and σ_{Ri}^2 are the mean and variance of the i th lot of the reference product. A similar argument is applied to the proposed biosimilar (test) product. As a result, the selection of reference lots for the estimation of σ_R is critical for the proposed approach. The selection of reference lots has an impact on the estimation of σ_R and consequently on the EAC. Assuming that $n_R > n_T$, the FDA suggested using the remaining $n_R - n_T$ lots to establish EAC to avoid selection bias. It sounds a reasonable approach if $n_R \gg n_T$. In practice, however, there might be a few lots available. Alternatively, it is suggested that all of the n_R lots be used to establish EAC.

The assumption that $\mu_{Ri} = \mu_{Rj}$ and $\sigma_{Ri}^2 = \sigma_{Rj}^2$ for $i \neq j; i, j = 1, \dots, n_R$ and $\mu_{Ti} = \mu_{Tj}$ and $\sigma_{Ti}^2 = \sigma_{Tj}^2$ for $i \neq j; i, j = 1, \dots, n_T$ is a strong assumption that does not reflect real practice. Since biosimilar products are made of living cells and/or living organisms, it is expected that

$$\mu_{Ri} \neq \mu_{Rj}, \sigma_{Ri}^2 \neq \sigma_{Rj}^2 \text{ for } i \neq j; i, j = 1, \dots, n_R$$

$$\mu_{Ti} \neq \mu_{Tj}, \sigma_{Ti}^2 \neq \sigma_{Tj}^2 \text{ for } i \neq j; i, j = 1, \dots, n_T.$$

Heterogeneity within lots and across lots between the test product and the reference product has posted the following controversial issues in analytical similarity assessment. First, suppose two extreme reference lots, one lot with the smallest within-lot variability and the other lot with the largest within-lot variability, are randomly selected for analytical similarity assessment. In this case, chances are that the reference product (the two selected extreme lots) may not even pass the equivalence test itself. Thus, analytical similarity between the test product and the reference product is not comprehensive. The other controversial issue is that if reference lots selected for establishment of EAC are extreme lots with the smallest variability, the established EAC could be too narrow to penalize good test products.

THE FDA'S CURRENT THINKING ON SCIENTIFIC INPUT

In his recent presentation, Tsong^[12] indicated that FDA's current thinking for establishment of EAC is to consider $1.5\sigma_R + \Delta$, where Δ is a regulatory allowance depending upon *scientific input*. From a statistical perspective, we may interpret the scientific input as scientific justification for accounting for the worst possible reference lot (i.e., a reference lot with extremely large variability) in establishment of EAC. Thus, $1.5\sigma_R + \Delta$, can be rewritten as $1.5\sigma'_R$, where $\sigma'_R = \sigma_R + \varepsilon$. Thus, we have

$$\text{EAC} = \pm 1.5(\sigma_R + \varepsilon).$$

The FDA's original proposal is to estimate σ'_R using sample SD (s) of the test values obtained from the reference lots, assuming that there is only one single test value per lot. Although Wang and Chow^[13] showed that s is an unbiased estimate of σ_R , it underestimates σ'_R because it does not take the variability associated with s into consideration. Thus, Chow^[3] suggested using the 95% upper confidence bound to estimate σ'_R , i.e.,

$$\hat{\sigma}'_R = \sqrt{\frac{n_R - 1}{\chi^2_{\alpha/2, n_R - 1}}} s.$$

This leads to

$$\varepsilon = \left(\sqrt{\frac{n_R - 1}{\chi^2_{\alpha/2, n_R - 1}}} - 1 \right) s.$$

One of the controversial issues for the establishment of EAC is whether the margin should be fixed. The FDA seems to recommend using the estimate of σ_R as the true σ_R without taking into consideration the variability associated with the observed sample variance. The variability associated with the observed sample variance depends upon the sample size (i.e., the number of reference lots) used for analytical similarity assessment. As a result, the result of an equivalence test may not be *reproducible*.

RELATIONSHIP BETWEEN SIMILARITY LIMIT AND VARIABILITY

As it can be seen from Table 1, similarity (equivalence) limit for assessment of similarity (equivalence) depends upon the variability associated with the drug product. For example, if the variability is less than 10%, (90%, 111%), similarity limit is recommended, while (80%, 125%) similarity limit is used for drug product with variability between 20% and 30%. For drug products with large variability (or highly variable drug products) such as biosimilar products, it is suggested that a scaled similarity limit

adjusted for variability be considered. As an alternative to the scaled similarity limit and in the interest of one-size-fits-all criterion, some researchers suggest (70%, 143%) to be considered. The selection of similarity limits based on the associated variability is somewhat arbitrary without scientific/statistical justification. In what follows, we attempt to describe the relationship between similarity limit and variability for achieving a desired probability of claiming similarity.

Let (δ_L, δ_U) be the similarity limits for the evaluation of similarity between a test product (T) and a reference product (R). Current regulation indicates that we can claim similarity if the 90% confidence interval of difference in mean (i.e., $\mu_T - \mu_R$) falls entirely within the lower and upper similarity limits. Let (L, U) be the 90% confidence interval for $\mu_T - \mu_R$. Define

$$p = P\{(L, U) \subset (\delta_L, \delta_U) | \sigma_R^2, \sigma_T^2\},$$

where σ_R^2 and σ_T^2 are the variabilities associated with the reference product and the test product, respectively, and p is a desired probability of claiming similarity. Thus, for a given set of p , σ_R^2 , and σ_T^2 , appropriate (δ_L, δ_U) can be determined.

A UNIFIED TIERED APPROACH

For biosimilar products, their population means and population variances are expected to be different. The relationship between a proposed biosimilar (test) product and an innovative (reference) product can be described as $\mu_T = \mu_R + \Delta$ and $\sigma_T = C\sigma_R$, where Δ is a measure of a possible shift in population mean and C is an inflation factor. There is a significant shift in mean (e.g., $\Delta \gg 0$) or notable heterogeneity between the proposed biosimilar product and the reference product (e.g., either $C \gg 1$ or $C \ll 1$). In this case, the validity of equivalence test for Tier 1 CQAs and the quality range approach for Tier 2 CQAs are questionable because the equivalence test and quality range approach are unable to handle a major shift in population mean and a significant change in population variance.

To overcome the problems of major mean shift and significant change in variability, alternatively, Chow et al.^[5] consider the equivalence test and quality range approach applied to standardized test values (or *effect size adjusted for SD*) rather than untransformed raw data. For simplicity and without loss of generality, consider the quality range approach for Tier 2 CQAs.

Define $eff_R = \mu_R / \sigma_R$ and $eff_T = \mu_T / \sigma_T$ and let $y_i, i = 1, \dots, n_R$ be the test values of the reference lots, where n_R is the number of reference lots considered for the test. Also, let $z_i, i = 1, \dots, n_T$ be the test values of the reference lots, where n_T is the number of test lots considered for the test. The FDA's quality range approach is applied

on $\{y_i, i = 1, \dots, n_R\}$ to establish the *quality range* with an appropriate selection of x . Instead, we suggested that the quality range approach be applied to $\{y_i / \hat{s}_R, i = 1, \dots, n_R\}$, where $\hat{s}_R$ is the sample SD of the test values of the reference lots. The quality range can then be established with the following adjustment on the selected x for achieving $eff_R \approx eff_T$. For simplicity and illustration purposes, consider that case that $\sigma_R \approx \sigma_T$ (i.e., $C \approx 1$). In this case, under the assumption that $eff_R \approx eff_T$, we have

$$\frac{\mu_R}{\sigma_R} \approx \frac{\mu_T}{\sigma_T} = \frac{\mu_R + \Delta}{\sigma_T}.$$

This leads to

$$\sigma_T = \sigma_R \left(1 + \frac{\Delta}{\mu_R} \right) \approx \sigma_R,$$

where $1 + \frac{\Delta}{\mu_R}$ is referred to as the adjustment factor for selection of x . Note that the left-hand side of the above equation is for test lots, and the right-hand side is referring to the reference lots. Various selection of x and adjusted x when there is a shift in mean are given in Table 3.

As it can be seen from the table, under the assumption that $eff_R \approx eff_T$, if there is a 20% shift in mean, i.e., $\frac{\Delta}{\mu_R} = 0.2$, the selection of $x = 2.5$ is equivalent to the selection of $x = 3$ without a shift in population mean.

A similar idea can be applied to the case in which there is a shift in scale parameter (i.e., $C \neq 1$). It should also be noted that the above proposal is similar to the justification (based on the percentage of coefficient of variation) as described in your questions. It should be noted that for biosimilar products, the assumption that $\mu_T \approx \mu_R$ and $\sigma_T \approx \sigma_R$ is usually not true. In practice, it is reasonable to assume that $eff_R \approx eff_T$. The above proposal accounts for possible shift in population mean and heterogeneity in variability.

Table 3 Adjustment on Selection of x When There Is a Shift in Mean

Δ/μ_R	x	$x(\text{adj})$
0.1	2	2.2
	2.5	2.75
	3	3.3
0.2	2	2.4
	2.5	3
	3	3.6

CONCLUDING REMARKS

The concept of stepwise approach recommended by the FDA is well taken. The purpose is to obtain the totality of the evidence in order for demonstration of biosimilarity between a proposed biosimilar product and an innovative biological product. The totality of the evidence consists of evidence from analytical studies for characterization of the molecule, animal studies for toxicity, PK and PD for pharmacological activities, and clinical studies for safety/tolerability, immunogenicity, and efficacy. The FDA, however, does not indicate whether these evidences should be obtained sequentially or simultaneously. It is a concern that a sequential approach may kill good products early purely by chance alone.

The stepwise approach starts with structural and functional characterization of CQAs that may be relevant to clinical outcomes. The FDA's recommended tiered approach is to serve the purpose. The recommended tiered approach depends upon the classification of identified CQAs based on their criticality (or risk ranking) relevant to clinical outcomes. The assessment of criticality, however, is somewhat subjective and often lack of scientific/statistical justification.

The FDA recommended that the tiered approach have raised a number of scientific and/or controversial issues. These controversial issues are related to difference in population means and heterogeneity within and across lots within and between the test product and the reference product. In practice, the primary assumption that $\mu_T \approx \mu_R$ and $\sigma_T \approx \sigma_R$ is usually not true. In this case, it is reasonable to assume that $eff_R = \frac{\mu_R}{\sigma_R} \approx \frac{\mu_T}{\sigma_T} = eff_T$ so that assessment of similarity between the proposed biosimilar product and the reference product is possible.

As indicated by Tsong,^[12] the Tier 1 equivalence test is supposed to be more rigorous than the Tier 2 quality range approach. Thus, we would expect that passing a Tier 1 test will pass a Tier 2 test. In practice, however, there is no guarantee that a given CQA that passes a Tier 1 test will pass a Tier 2 test and vice versa. This may be due to difference in primary assumptions made for Tier 1 equivalence test and Tier 2 quality range approach. Since there may be a large number of CQAs in both Tier 1 and Tier 2, "Does FDA require all CQAs at either Tier pass the corresponding test in order to claim the totality of the evidence?" is probably the most commonly asked question. If not, are there any rules to follow? These questions, however, remain unanswered.

Note that recently FDA circulated a draft guidance on analytical similarity assessment for public comments in September 2017. [14] In the draft guidance, most of the issues described in this chapter remained unsolved. Specific guidance on these issues are necessarily developed.

REFERENCES

1. Chow, S.C. *Biosimilars: Design and Analysis of Follow-on Biologics*; Chapman and Hall/CRC Press; Taylor & Francis: New York, 2013.
2. Chow, S.C. On assessment of analytical similarity in biosimilar studies. *Drug Des.: Open Access* **2014**, *3*, e124. doi: 10.4172/2169-0138.1000e124.
3. Chow, S.C. Challenging issues in assessing analytical similarity in biosimilar studies. *Biosimilars* **2015**, *5*, 33–39.
4. Chow, S.C.; Endrenyi, L.; Lachenbruch, P.A.; Yang, L.Y.; Chi, E. Scientific factors for assessing biosimilarity and drug interchangeability of follow-on biologics. *Biosimilars* **2011**, *1*, 13–26.
5. Chow, S.C.; Song, F.Y.; Bai, H. Similarity assessment in analytical biosimilar studies. *AAPS J.* **2016**, *18* (3), 670–677.
6. Chow, S.C.; Song, F.Y.; Endrenyi, L. A note on Chinese draft guidance on biosimilar products. *Chin. J. Pharm. Anal.* **2015**, *35* (5), 762–767.
7. Christl, L. Overview of regulatory pathway and FDA's guidance for development and approval of biosimilar products in US. Presented at FDA ODAC meeting, Silver Spring, MD, January 7, 2015.
8. FDA. *Guidance on Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Considerations*. Center for Drug Evaluation and Research, the US Food and Drug Administration, Rockville, MD, 2003.
9. FDA. *Guidance on Scientific Considerations in Demonstrating Biosimilarity to a Reference Product*. The United States Food and Drug Administration, Silver Spring, MD, 2015.
10. Haidar, S.H.; Davit, B.; Chen, M.L.; Conner, D.; Lee, L.; Li, Q.H.; Lionberger, R.; Makhoul, F.; Patel, D.; Schuirmann, D.J.; Yu, L.X. Bioequivalence approaches for highly variable drugs and drug products. *Pharm. Res.* **2008**, *25*, 237–241.
11. SSAB. A communication package submitted to the FDA. Scientific Statistical Advisory Board on Biosimilars (SSAB) sponsored by Amgen, Thousand Oaks, CA, 2010.
12. Tsong, Y. Analytical Similarity Assessment. Presented at Duke-Industry Statistics Symposium, Durham, NC, October 22–23, 2015.
13. Wang, T.R.; Chow, S.C. On establishment of equivalence acceptance criterion in analytical similarity assessment. *J. Biopharm. Stat.* **2017**, *27*, 206–212.
14. FDA. *Guidance for Industry Statistical Approaches to Evaluate Analytical Similarity*. Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER). The United States Food and Drug Administration, Silver Spring, MD, September, 2017.

ANCOVA Approach: Premarketing Shelf Life Determination with Multiple Factors

Yi Tsong, Chi-wan Chen, Wen Jen Chen, Roswitha E. Kelly,
Daphne T. Lin, and Karl K. Lin

Office of Biostatistics/Office of Pharmacoepidemiology and Biostatistical Sciences,
Center of Drug Evaluation and Research, U.S. Food and Drug Administration,
Silver Spring, Maryland, U.S.A.

Abstract

The stability of a drug substance or drug product is its capacity to remain within the established acceptance criteria to ensure its identity, strength, quality, and purity for a specified period of time.

In this entry, the discussion focuses on the shelf life determination based on ANCOVA modeling.

INTRODUCTION

The stability of a drug substance or drug product is the capacity of the drug substance or drug product to remain within the established acceptance criteria to ensure its identity, strength, quality, and purity for a specified period of time. Regulatory agencies require that adequate testing be performed by an applicant to demonstrate the stability of the drug substance or product in support of a marketing registration application. The purpose of stability testing is to provide evidence on how the quality of a drug substance or drug product varies with time under the influence of a variety of environmental factors, such as temperature, humidity, and light. This assessment is used to establish the recommended storage condition and a shelf life for the drug product. Shelf life (also referred to as expiration dating period) is the period of time during which a drug product is expected to remain within the approved acceptance criteria, provided the drug product has been stored under the conditions defined on the container label.

For the sampling of stability data, a sufficient number of container units from each of at least three batches are put on long-term storage in a controlled environment (e.g., 25°C/60% RH), and samples are taken from the chamber at predetermined time points, namely at 0, 3, 6, 9, 12, 18, 24, 36, 48, etc. months, and tested for appropriate physical, chemical, biological, and microbiological attributes. The tested data of all time points collected before the drug is marketed will be used to support the marketing application of a drug product. A minimum of three batches is a compromise between cost and statistical requirements for estimating between-batch variability. Data analysis can determine

whether data from the batches could be combined, which usually lead to a longer shelf life estimate.

Many products are supplied in several strengths and different container closure systems, e.g., bottle, blister pack, and various container sizes and fills. In the United States, prior to 1990, the shelf life of a pharmaceutical product was determined separately for the individual manufacturing and marketing factors.^[1–3] It was realized that multifactorial designs in which the effects of batch, strength, container, etc. on the stability of the dosage form are investigated may be more efficient. If the assumption of equal error variances for various combinations of factors holds, there is a great potential advantage of the multifactorial over separate regression models. It lies in the improvement in the precision of the variance estimate by using all data. It lies also in the size of the estimation error through the potential pooling of data across design factors. In an effort to save cost and resources, several complicated multifactorial designs, most notably bracketing and matrixing designs, have been proposed in the early 1990s.^[4–8] However, their applicability to a given case needs to be cautiously evaluated.

The regulatory requirements of the submission of stability studies of pharmaceuticals at New Drug Application (NDA) or premarketing stage are described in the recently updated guidelines issued by the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, or ICH. For example, *ICH Q1A(R) Stability Testing of New Drug Substances and Products*^[9] defines the core stability data requirements that are sufficient to support the registration of a new drug application in the tripartite regions of the European Union, Japan, and the United States. It recommends that at least three primary batches of the drug substance or product be tested for stability at prescribed storage conditions and time points. *ICH Q1D Bracketing*

This paper represents the authors' professional opinion but it does not represent an official position of the U.S. Food and Drug Administration.

and *Matrixing Designs for Stability Testing of New Drug Substances and Products*^[10] provides guidance on reduced designs for stability studies. It outlines the circumstances under which a bracketing or matrixing design can be used. *ICH Q1E Stability Data Evaluation*^[11] describes the principles of stability data evaluation and various approaches to statistical analysis of stability data when establishing a retest period for the drug substance or a shelf life for the drug product.

In this entry, the discussion focuses on the shelf life determination based on ANCOVA modeling.

ANCOVA MODELING AND POOLING TESTS

The linear change of the drug potency, or any other appropriate attribute, of an individual batch is often represented by

$$Y(t) = \alpha + \beta t + \varepsilon_i \quad (1)$$

where $Y(t)$ denotes the observed mean value of a batch at month t , α and β are the intercept and slope of the batch, respectively, ε_i is the model error term that follows a $N(0, \sigma^2)$ distribution. The true shelf life T of the batch is the date when the expected regression line intersects with either the lower acceptance criterion S_L or the upper acceptance criterion S_U . With the data collected, the shelf life is estimated by the earlier date (i.e., time point) between the two dates when either one of the two 95% confidence limits (upper and lower) of the expected regression line separately intersects the acceptance criteria, S_L and S_U .

Linear regressions of multiple batches of the same drug product can be represented by the ANCOVA model,

$$Y_i(t) = \alpha_0 + \alpha_i + (\beta_0 + \beta_i)t + \varepsilon_i t \quad (2)$$

where α_0 and β_0 are the common intercept and slope of all batches, respectively, α_i and β_i are the deviations of the individual intercept and slope of the i th batch from α_0 and β_0 , respectively.

Often the batches may share the same regression line or have the same changing rate (slope). In order to determine whether the batches share the same slope or intercept, pooling tests of slope and intercept are to be used. They are performed in a hierarchical order such that the poolability of the slope is tested before the intercept. To determine whether the batches share the same slope, one tests

$$H_{0\beta}: \beta_i = 0 \quad \text{vs.} \quad H_{a\beta}: \beta_i \neq 0 \quad (3)$$

If $H_{0\beta}$ is not rejected, one concludes that all batches share the same slope. In this case, one may proceed to test

$$H_{0\alpha}: \alpha_i = 0 \quad \text{vs.} \quad H_{a\alpha}: \alpha_i \neq 0 \quad (4)$$

Similarly, if $H_{0\alpha}$ is not rejected, one concludes that all batches share the same intercept as well.

Depending on the results of testing $H_{0\beta}$ and $H_{0\alpha}$, one can determine whether the batches should be pooled in estimating the shelf life using a common slope only or a common slope and intercept.

Conventionally, the modeling of simple stability studies with multiple batches is carried out with hierarchical pooling tests of slope and intercept^[1–3,13–16] based on the type I sums of squares when using the SAS PROC ANOVA or GLM.^[17] The two tests are performed in a single ANCOVA analysis of the model (2) using F -tests for hypotheses (3) and (4). With type I sums of squares, hypotheses (3) are tested using $R(\beta_i | \alpha_0, \beta_0, \alpha_i)$, i.e., the sum of squares of β_i adjusted for α_0 , β_0 , and α_i . Hypotheses (4) are tested with $R(\alpha_i | \alpha_0, \beta_0)$, i.e., the sum of squares of α_i adjusted for α_0 and β_0 . Once a model is determined, either testing or estimation of shelf life can be carried out using this final model.

The objective of the NDA stability study is to demonstrate that the sponsor can produce the same product consistently. This means that the same change over time occurs from batch to batch. It is expected that the sponsor can release batches within tightly controlled limits. The pooling tests are structured in a hierarchical order such that the slope test is carried out before the intercept test. The ANCOVA model (2) may be used to represent a common slope–common intercept (i.e., $\beta_i = 0$ and $\alpha_i = 0$) model, common slope–individual intercept (i.e., $\beta_i = 0$ and $\alpha_i \neq 0$) model, or an individual slope–individual intercept (i.e., $\beta_i \neq 0$ and $\alpha_i \neq 0$) model. In practice, the individual slope–common intercept model is not of interest and cannot be determined by hierarchical tests based on type I sums of squares in a single SAS PROC GLM run.

When evaluating the stability data in support of a single proposed shelf life T_0 for all batches of the drug product, the shortest estimated shelf life T^* of any batch in (2) will be compared with T_0 . T_0 is granted if $T^* > T_0$. T_0 may be extrapolated 6–12 months beyond the last observation date, when at least 12 months' data are available.

When extending the same ANCOVA–type I sums of square application to a multiple-factor design, a prespecified hierarchical ordering of the pooling tests^[12,18] is required. For example, when applying the two-step ANCOVA model, i.e., testing first slopes then intercepts to a study designed with two factors (e.g., strength and container size) and multiple batches, the ANCOVA model with all factors and interactions can be represented by

$$\begin{aligned} Y_{spbt} = & \alpha + \alpha_s + \alpha_p + \alpha_b + \alpha_{sb} + \alpha_{sp} + \alpha_{pb} \\ & + \alpha_{spb} + \beta t + \beta_s t + \beta_p t + \beta_b t + \beta_{sb} t \\ & + \beta_{sp} t + \beta_{pb} t + \beta_{spb} t + \varepsilon_{spbt} \end{aligned} \quad (5)$$

where subscript $s = 1, \dots, S$ represents the level of strength, $p = 1, \dots, P$ represents the level of container size, and $b = 1, \dots, B$ represents the batch number. The α s represent

the intercept coefficients, the β s represent the slope coefficients of the regression lines, ε_{spbt} represents the iid random error term that follows a standard normal distribution with mean zero and variance σ^2 .

A hierarchical ordering of null hypotheses of the pooling tests may be given as follows:

1. $H_0: \beta_{spb} = 0$, strength-by-container size-by-batch interaction for slope.
2. $H_0: \alpha_{spb} = 0$, strength-by-container size-by-batch interaction for intercept.
3. $H_0: \beta_{sb} = 0$, strength-by-batch interaction for slope.
4. $H_0: \alpha_{sb} = 0$, strength-by-batch interaction for intercept.
5. $H_0: \beta_{pb} = 0$, container size-by-batch interaction for slope.
6. $H_0: \alpha_{pb} = 0$, container size-by-batch interaction for intercept.
7. $H_0: \beta_{sp} = 0$, strength-by-container size interaction for slope.
8. $H_0: \alpha_{sp} = 0$, strength-by-container size interaction for intercept.
9. $H_0: \beta_b = 0$, batch slope.
10. $H_0: \alpha_b = 0$, batch intercept.
11. $H_0: \beta_s = 0$, strength slope.
12. $H_0: \alpha_s = 0$, strength intercept.
13. $H_0: \beta_p = 0$, container size slope.
14. $H_0: \alpha_p = 0$, container size intercept.

When using type I Sum of Squares (SS), the F -test for each factor or interaction is based on the sums of squares adjusted for the factors and interactions following it in the ordering. The process of the pooling test stops whenever a significant F -value is encountered. For example, if $H_0: \alpha_{pb} = 0$ is rejected. The final model is

$$Y_{spbt} = \alpha + \alpha_s + \alpha_p + \alpha_b + \alpha_{sp} + \alpha_{pb} + \beta t + \beta_s t + \beta_p t + \beta_b t + \beta_{sp} t + \beta_{pb} t + \varepsilon_{spbt} \quad (6)$$

even if both β_{sp} and α_{sp} are not significantly different from 0.

Hierarchical ordering of the pooling test involving batch, strength, and container size is not unique. There are six possible ordering arrangements to test for β_{sb} , β_{sp} , β_{pb} and six arrangements to test for β_b , β_p , and β_s leading to as many as 36 different test orderings for analysis. Because of the “sudden death” nature of the one-step model selection, any of the alternative arrangements may lead to a different shelf life for the product.

Alternatively, Tsong, Chen, and Chen^[18] proposed to use a more flexible stepwise modeling procedures with type III sums of squares under some proper restrictions on the order of the pooling tests. This approach will accommodate the possibility of eliminating any term within the limitation of exchangeable ordering.

To illustrate, consider the following model,

$$Y_{spbt} = \alpha + \alpha_s + \alpha_p + \alpha_b + \alpha_{sb} + \alpha_{sp} + \alpha_{pb} + \beta t + \beta_s t + \beta_p t + \beta_b t + \beta_{sb} t + \beta_{sp} t + \beta_{pb} t + \varepsilon_{spbt}$$

after eliminating the nonsignificant three-way interaction terms after pooling tests.

The F -statistics for testing $\beta_{sb} = 0$, $\beta_{sp} = 0$, and $\beta_{pb} = 0$ are calculated using type III SS, $R(\beta_{sb} | \alpha, \alpha_s, \alpha_p, \alpha_b, \alpha_{sb}, \alpha_{sp}, \alpha_{pb}, \beta, \beta_s, \beta_p, \beta_b, \beta_{sp}, \beta_{pb})$, $R(\beta_{sp} | \alpha, \alpha_s, \alpha_p, \alpha_b, \alpha_{sb}, \alpha_{sp}, \alpha_{pb}, \beta, \beta_s, \beta_p, \beta_b, \beta_{sb}, \beta_{pb})$, and $R(\beta_{pb} | \alpha, \alpha_s, \alpha_p, \alpha_b, \alpha_{sb}, \alpha_{sp}, \alpha_{pb}, \beta, \beta_s, \beta_p, \beta_b, \beta_{sp}, \beta_{sb})$, respectively. The term with the largest nonsignificant p -value (say β_{sb}) is to be eliminated from the model. Then, in the reduced model

$$Y_{spbt} = \alpha + \alpha_s + \alpha_p + \alpha_b + \alpha_{sb} + \alpha_{sp} + \alpha_{pb} + \beta t + \beta_s t + \beta_p t + \beta_b t + \beta_{sp} t + \beta_{pb} t + \varepsilon_{spbt}$$

α_{sb} , the corresponding intercept term, is tested for poolability. It is eliminated from the model if the F -value using $R(\alpha_{sb} | \alpha, \alpha_s, \alpha_p, \alpha_b, \alpha_{sp}, \alpha_{pb}, \beta, \beta_s, \beta_p, \beta_b, \beta_{sp}, \beta_{pb})$ for testing $H_0: \alpha_{sb} = 0$ is not significant. In this case, the model is further reduced to

$$Y_{spbt} = \alpha + \alpha_s + \alpha_p + \alpha_b + \alpha_{sp} + \alpha_{pb} + \beta t + \beta_s t + \beta_p t + \beta_b t + \beta_{sp} t + \beta_{pb} t + \varepsilon_{spbt}$$

Otherwise, if α_{sb} is not equal to zero, it will be retained in the model. Based upon the above further reduced model, the next term to be eliminated is either β_{sp} or β_{pb} depending on which has the larger nonsignificant p -value calculated by $R(\beta_{sp} | \alpha, \alpha_s, \alpha_p, \alpha_b, \alpha_{sp}, \alpha_{pb}, \beta, \beta_s, \beta_p, \beta_b, \beta_{pb})$ and $R(\beta_{pb} | \alpha, \alpha_s, \alpha_p, \alpha_b, \alpha_{sp}, \alpha_{pb}, \beta, \beta_s, \beta_p, \beta_b, \beta_{sb})$, respectively. Depending on the testing results of β_{sp} and β_{pb} , the model may be even further reduced to one of the following two:

$$Y_{spbt} = \alpha + \alpha_s + \alpha_p + \alpha_b + \alpha_{sp} + \alpha_{pb} + \beta t + \beta_s t + \beta_p t + \beta_b t + \beta_{pb} t + \varepsilon_{spbt}$$

$$Y_{spbt} = \alpha + \alpha_s + \alpha_p + \alpha_b + \alpha_{sp} + \alpha_{pb} + \beta t + \beta_s t + \beta_p t + \beta_b t + \beta_{sp} t + \varepsilon_{spbt}$$

The stepwise modeling process stops when no further term can be eliminated.

There are two restrictions to the testing orders.

1. Test higher-level interaction terms before lower-level interactions or main factors. However, the lower-level interactions or main design factors can be tested even with significant higher-level interaction terms, as long as there are no common design factors involved.
2. For each factor or interaction, test the equality of slope before testing the equality of the corresponding intercept. In addition, no equality of intercept is tested if the equality of the corresponding slope has been rejected.

At the premarketing stage when the sample size (i.e., number of observation times) is small, the major concern with the pooling test is its power to reject the null hypothesis.^[1–3,13–16] In order to reduce the rate of false pooling, a significance level of 0.25 was recommended for testing for pooling batches in simple stability designs.^[13–16] Fairweather, Lin, and Kelly^[19] proposed to use large significance level such as $\alpha = 0.25$ for all testing that involves the batch term and $\alpha = 0.05$ for any other pooling test. These significance levels are currently recommended in *ICH Guidance*.^[10] Note that Fairweather, Lin, and Kelly^[19] used the error rate for the *F*-tests based on type I sums of squares instead of the *F*-tests based on type III sums of squares as proposed by Tsong, Chen, and Chen.^[18]

DISCUSSION AND CONCLUSIONS

The ANCOVA approach has been the conventional and standard tool to analyze stability study data. With appropriate significance levels set for the intercept and slope pooling tests, decision on pooling data of different batches or across design factors can be made for the determination of the shelf lives. The extension of the conventional modeling with *F*-test based on type I sums of squares for pooling to multifactorial designs, including both complete and fractional, has some unwanted limitations. The alternative approach with ANCOVA stepwise modeling using *F*-tests based on type III sums of squares for pooling tests may eliminate these difficulties. Alternative approaches, such as pooling by equivalence assessment in order to replace pooling tests of low power, are topics of great interest to nonclinical statisticians.^[20] Random effect models for shelf life prediction has also been proposed in the literature.^[21–25] The supporting argument of such an approach is that the shelf life is determined for future batches rather than for the few batches studied. However, in practice, the batches at the NDA stage are not a random sample and their number is often too small to permit a variance estimate with precision.

ACKNOWLEDGMENT

This manuscript was prepared with the support of Regulatory Science Research Grant RSR02-015 of the Center of Drug Evaluation and Research, U.S. FDA. The authors wish to thank the FDA CDER Office of Biostatistics Stability Working Group for the support and discussion on the development of this manuscript.

REFERENCES

1. FDA. *In Guidelines for Submitting Documentation for the Stability of Human Drugs and Biologics*; U.S. Food and Drug Administration, Center for Drugs and Biologics: Rockville, MD, 1987.
2. Lin, K.K.; Lin, T.D.; Kelly, R.E. Stability of drugs. *In Statistics in the Pharmaceutical Industry*, 2nd Ed.; Buncher, C.R.; Tsay, J.Y., Eds.; Marcel Dekker: New York, 1993, 419–444.
3. Chow, S.C.; Liu, J.P. *Statistical Design and Analysis in Pharmaceutical Sciences*; Marcel Dekker: New York, 1995.
4. Helboe, P. New designs for stability testing programs: matrix or factorial designs. Authorities viewpoint on the predictive value of such studies. *Drug Inf. J.* **1992**, *26*, 629–634.
5. Nordbrock, E. Statistical comparison of stability study designs. *J. Biopharm. Stat.* **1992**, *2*, 91–113.
6. Lin, T.D. Applicability of Matrix and Bracket Approach to Stability Study Design. Proceedings of the Association Biopharmaceutical Section of American Statistical Association, 1994, 142–147.
7. Lin, T.-Y.D.; Chen, C.-W. Overview of stability designs. *J. Biopharm. Stat.* **2003**, *133*, 337–354.
8. Chen, C.-W. *U.S. FDA's Perspective of Matrixing and Bracketing*. Proceedings from EFPIA Symposium: Advanced Topics in Pharmaceutical Stability Testing Building on the ICH Stability Guideline; EFPIA: Brussels, 1996.
9. *Guidance for Industry: ICH Q1A(R) Stability Testing of New Drug Substances and Products*, Food and Drug Administration, Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research; August, 2001 (www.fda.gov/cder/guidance/4282fnl.pdf).
10. *Draft ICH Consensus Guideline Q1E Stability Data Evaluation*, Food and Drug Administration, Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research; August, 2001 (www.fda.gov/cder/guidance/4983dft.pdf).
11. *Guidance for Industry: ICH Q1D Bracketing and Matrixing Designs for Stability Testing of New Drug Substances and Products*, Food and Drug Administration, Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research; January, 2003 (www.fda.gov/cder/guidance/4985fnl.pdf).
12. Tsong, Y.; Chen, C.-W.; Chen, W.J.; Kelly, R.; Lin, K.K.; Lin, T.D. Stability of pharmaceutical products. *In Statistics in the Pharmaceutical Industry*, 3rd Ed.; Buncher, C.R.; Tsay, J.Y., Eds.; Marcel Dekker: New York, 2003.
13. Bancroft, T.A. On biases in estimation due to the use of preliminary tests of significance. *Ann. Math. Stat.* **1944**, *15*, 190–204.
14. Larson, H.J.; Bancroft, T.A. Sequential model building for prediction in regression analysis, I. *Ann. Math. Stat.* **1963**, *34*, 231–242.
15. Bancroft, T.A. Analysis and inference for incompletely specified models involving the use of preliminary tests of significance. *Biometrics*. **1964**, *203*, 427–442.
16. Johnson, J.P.; Bancroft, T.A.; Han, C.P. A pooling methodology for regressions in prediction. *Biometrics*. **1977**, *33*, 57–67.
17. *SAS/STAT User's Guide, Version 8*; SAS Institute Inc.: Cary, NC, 1999; vol. 2.
18. Tsong, Y.; Chen, W.-J.; Chen, C.-W. ANCOVA approach for shelf life analysis of stability study of multiple factor designs. *J. Biopharm. Stat.* **2003**, *13* (3), 375–394.

19. Fairweather, W.R.; Lin, T.D.; Kelly, R. Regulatory, design, and analysis aspects of complex stability studies. *J. Pharm. Sci.* **1995**, *84* (11), 1322–1326.
20. Tsong, Y.; Chen, W.-J.; Chen, C.-W. Shelf life determination based on equivalence assessment. *J. Biopharm. Stat.* **2003**, *13* (3), 431–450.
21. Chen, J.J.; Hwang, J.-S.; Tsong, Y. Estimation of the shelf life of drugs with mixed effects models. *J. Biopharm. Stat.* **1995**, *5* (1), 131–140.
22. Chen, J.; Ahn, H.; Tsong, Y. Shelf life estimation for multi-factor stability studies. *Drug Inf. J.* **1997**, *31* (2), 573–587.
23. Chow, S.-C.; Shao, J. Estimating drug shelf-life with random effect batches. *Biometrics*. **1991**, *47*, 1071–1079.
24. Shao, J.; Chen, L. Prediction bounds for random shelf-lives. *Stat. Med.* **1997**, *16*, 1167–1173.
25. Shao, J.; Chow, S.C. Statistical inference in stability analysis. *Biometrics*. **1994**, *50*, 753–763.

Assay Development

Timothy L. Schofield

Merck Research Laboratories, West Point, Pennsylvania, U.S.A.

Abstract

Assays are utilized in the pharmaceutical industry to characterize drugs and biological products. Assay development is coordinated with pharmaceutical development to furnish the tools necessary to guide the research process. This entry will discuss the varieties of technologies and statistical methodologies used in the development of an assay.

INTRODUCTION

Assays are utilized in the pharmaceutical industry to characterize drugs and biological products. Assay development is coordinated with pharmaceutical development to furnish the tools necessary to guide the research process. Measures of content, potency, and purity are combined to warrant the safety and effectiveness of marketed products. Thus special care is taken to develop assays that are sensitive to changes in the important properties of pharmaceutical products.

Statistics plays a primary role in the processing of assay results and in the development of experimental strategies to optimize an analytical method. Most of these strategies employ basic statistical moieties, and they are thus accessible to the analytical researcher and biostatistician alike. Their simplicity does not absolve the user, however, from careful attention to their implementation and interpretation.

This entry will discuss the varieties of technologies and statistical methodologies used in the development of an assay. A review of standard calibration methodologies will illustrate the synergy between the scientific and statistical properties of an assay. Sound statistical experimental design can be utilized to explore the effect of operational factors on assay response and to establish the levels of the factors that yield reliable results.

TERMINOLOGY

Several features common to all assays performed in pharmaceutical research, development, and quality control are associated with specific terminology. These terms will be defined here and illustrated with specific cases.

The *analyte* is the substance being measured. This may be the active drug in the product or that compound in clinical samples obtained during clinical development. The analyte can also be a known impurity or a degradate of the active chemical species. Biotechnological and biological

products are subject to the introduction of adventitious agents, such as retroviruses associated with the biological system used to produce the product. In vaccine studies, products of the biological response of the vaccinee, such as antibodies or responder cells, are measured to establish successful immunization with the immunogen.

The *sample matrix* is the environment in which the analyte exists. For a drug product, this will be the inactive excipients that comprise the formulation, while for a biological it will be these as well as byproducts of the process, such as cellular debris, surfactants, and other process components. The clinical sample is the matrix for specimens tested from pharmacokinetic studies as well as from immunogenicity studies.

Various terms are used to describe the units of measure of an assay. Most assays of a drug substance measure the *content* or *gravimetric mass* of the compound, which is reported in units of mass (e.g., micrograms) or mass per volume of sample (e.g., micrograms per milliliter). The *potency* of a drug represents the biological activity of the compound, and it is usually reported as the amount of the analyte that induces a specific response (e.g., the amount of a vaccine that is effective in inducing antibody response in half of immunized animals is called the 50% effective dose, or ED_{50}). Many potency assays involve a titration of a drug or biological (i.e., a series of doses of the analyte), yielding a *titer* for the sample. Finally, many assays measure the physical properties of a drug or biological, such as pH, polydispersity, and solubility.

The term *assay* is used synonymously to represent the particular analytical method, an individual run of the procedure, or the composite result obtained for a sample in multiple runs of the procedure. Where not otherwise specified in this and related entries, the term *assay* will refer to the analytical method being discussed.

In this entry, a calibration *standard* represents the material in an assay, which is utilized to construct the “ruler” against which test samples are measured, while an assay *control* is a material that is measured alongside test

samples and that is used to monitor the performance of the procedure.

ASSAY TECHNOLOGIES

A variety of technologies form the foundation of analytical method development. A basic understanding of the mechanism and operation of these technologies is useful to their successful implementation.

Separation techniques, such as high-performance liquid chromatography (HPLC), are among the most widely employed assay technologies in the pharmaceutical industry. The basis of HPLC is the separation of molecular species, which are forced by a mobile phase through a resin-packed column (solid phase) under fixed pressure. The separation can be used on different properties of the analyte, including size, ionization, hydrophobicity, and affinity to a ligand. The output from an HPLC assay is a chromatographic trace, displaying spiked peaks with areas that are directly proportional to the amount of each molecular species in the sample. High-performance liquid chromatography is used either to detect or to quantify the amount of an analyte in a product or clinical sample. Quantitation of the analyte is usually accomplished using standard calibration techniques, in correspondence to a reference preparation.

Immunoassays employ the principle of antibody–antigen binding as a means to quantify an analyte. There are a variety of technologies associated with immunoassay, but common to all of these is a *detector*. Among the most common detectors are radioactive isotopes (associated with radioimmunoassay, or RIA), which discharge radioactive emissions in counts per minute (CPM), and enzyme substrates (associated with enzyme immunoassay, or EIA), which yield a change in the optical density (OD) of the solution at a specific wavelength of light. In either case, the level of detector signal is directly proportional to the amount of analyte in the sample. As in HPLC, quantitation of the analyte is accomplished using standard calibration techniques in correspondence to a reference preparation.

Many pharmacologically active compounds yield a response in vivo, in a biological milieu. Activity can be tested in a specified tissue, in cell culture, or in animals, utilizing biological assay techniques. A typical *bioassay* in animals is comprised of groups of animals, with each group receiving a graded dose of the test compound. The number of animals eliciting a particular response in each group is noted, and an endpoint is calculated from the dose series using one of a variety of quantal calculation methods.

A number of emerging technologies have been introduced into the analytical portfolio of many drugs and vaccines. In most cases these technologies fit into the traditional analytical paradigm. Thus fluorescence and refractory index can now easily be measured in a sample, which serve as the detector in a traditional calibration

assay design. Powerful methods for exploring the structure of a drug molecule or biological derivative, such as nuclear magnetic resonance (NMR) and polymerase chain reaction (PCR), bring with them new challenges for assay development and validation.

CALIBRATION

The most common assay design used to quantify the amount or potency of analyte in a drug or clinical sample is a standard curve, or calibration, design. A typical run of a calibration assay is comprised of a series of known amounts of a calibration, or reference, standard, run alongside one or several levels of a test preparation. The responses from the standard series are statistically fit using one of a variety of mathematical equations; this fit is then used to predict the amount of the analyte in the test sample. A typical calibration assay is illustrated in Fig. 1.

Linear Calibration

One of several strategies can be employed in selecting an appropriate calibration equation for an assay. During development, data from several runs of the standard preparation can be used to identify the “linear” range of the standard curve, in order to utilize a linear equation to fit the standard curve. This range is usually determined graphically, and can be verified statistically using common goodness-of-fit techniques. Some of the factors related to the implementation of the linear standard curve design include (1) the range of concentrations used in the standard curve; (2) the number of standard curve concentrations; and (3) the concentration scaling—i.e., arithmetic (e.g., 10, 20, 30, 40, etc.) or geometric (e.g., 10, 20, 40, 80, etc.).

These factors will, for the most part, be selected on the basis of the intended use and specific characteristics of the assay. The range of the standard curve should embrace the expected range of concentrations in test samples, with the exception that high-concentration samples can usually be manipulated (either by dilution or by loading

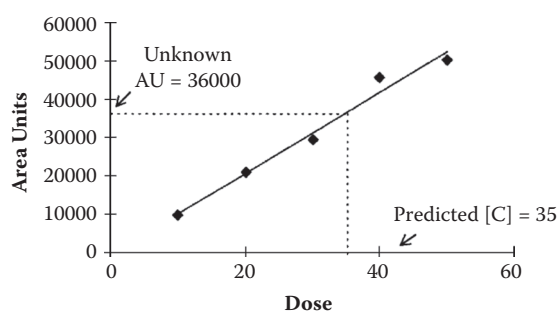


Fig. 1 Calibration using a linear fit to a standard curve; inverse regression is employed to predict the concentration corresponding to the response in an unknown

a different amount of the test specimen) to achieve a level that falls within the standard curve range. The number of standard curve concentrations can be selected on the basis of the impact on the margin of error of measurement of the analyte concentration in the test sample. The concentration scaling can be selected on the basis of the behavior in the variability of the instrument readout (e.g., CPM, OD); if the variability of the readout is well behaved (i.e., there is no association between variability and response level), then an arithmetic scaling of standard curve concentrations can be used. If the variability of the instrument readout increases with increased response, a geometric scaling of response may be preferred. In this case, the standard curve equation should be fit using a log-log transformation of the data, thereby achieving both equal spacing among the standard curve concentrations and homogeneity of variability across the (log) responses. A linear-linear fit and a log-log fit to a standard curve with geometric scaling are illustrated in Fig. 2.

Once a linear equation has been fit to the standard curve data, and prior to predicting the concentrations in unknowns, the statistical “goodness of fit” of the linear equation can be assessed using common regression diagnostics. Customarily, a restriction is placed upon the coefficient of determination (R^2) associated with the regression fit, and an assay run that fails to meet this restriction is judged invalid and is usually repeated. This criterion is frequently and inappropriately used to select the linear range of the standard curve during development of an assay as well as during its routine use. The dangers of relying exclusively on R^2 as a measure of goodness of fit are illustrated in an article by Anscombe.^[1] Several conditions yielding satisfactory R^2 values, such as nonlinear standard curve kinetics, a regression outlier, or heterogeneity of variability across the range of the standard curve, might still require remedial action. These conditions might more aptly be addressed through graphical diagnostic techniques, such as residual plots. Fig. 3 illustrates residual patterns associated with a properly behaved regression fit, along with several commonly

occurring alternatives. Residual plots can, in turn, be supplemented with more sophisticated diagnostic criteria, such as statistical tests of patterned residuals, of curvature, or of regression outliers.^[2]

Another graphical device used to diagnose the linearity of the standard curve is a “response factor” plot.^[3] The “response factor” is calculated as the ratio of the instrument readout to the nominal concentration of each standard curve point, and it is plotted against concentration. When the regression intercept is zero, the “response factor” corresponds to the slope of the standard curve, and it should be constant across levels. If a trend is observed, this indicates either a nonlinear calibration curve or a nonzero intercept. For more complex calibration functions, or when a nonzero intercept can be tolerated, the “response factor” plot of the standard can be compared with that of a concentration series of a test sample; similarly shaped “response factor” profiles indicate an assay that is “linear,” in the sense that measurements at each level of the test sample will be directly proportional to the test sample concentrations (see “Linearity” in the entry *Assay Validation*).

Having established goodness of fit of the standard curve, parameter estimates for the intercept (a) and the slope (b) can be utilized to predict the concentration of analyte in an unknown sample:

$$\hat{X} = \frac{Y_{\text{unknown}} - a}{b}$$

where Y_{unknown} represents the response in the unknown (note that this pertains to the case of a “linear-linear” fit to the standard curve; prediction from a “log-log” fit of the standard curve can be developed similarly). The reliability of the estimate can be ascertained through the prediction interval on the result:^[4]

$$\frac{\hat{X} \pm \frac{t_{n-2} \cdot \hat{\sigma}}{b} \sqrt{(1 + 1/n) \cdot (1 - g) + \hat{X}^2 / SXX}}{1 - g}$$

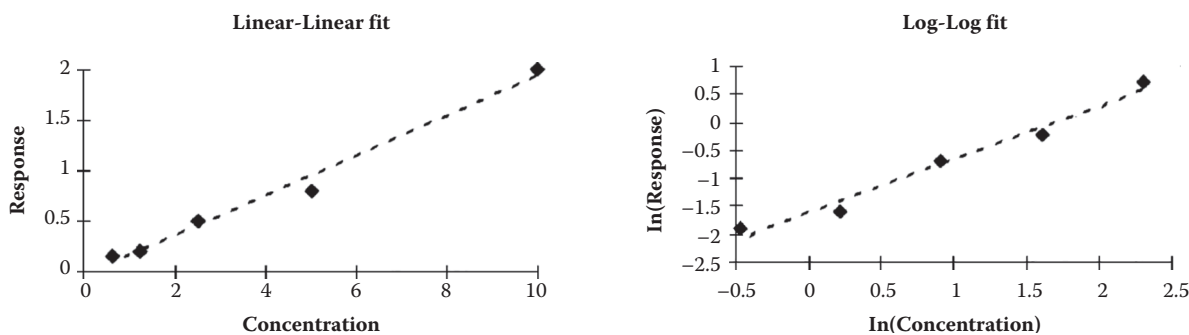


Fig. 2 A linear-linear fit to a geometrically scaled standard curve alongside a log-log fit, illustrating equal spacing between log concentrations

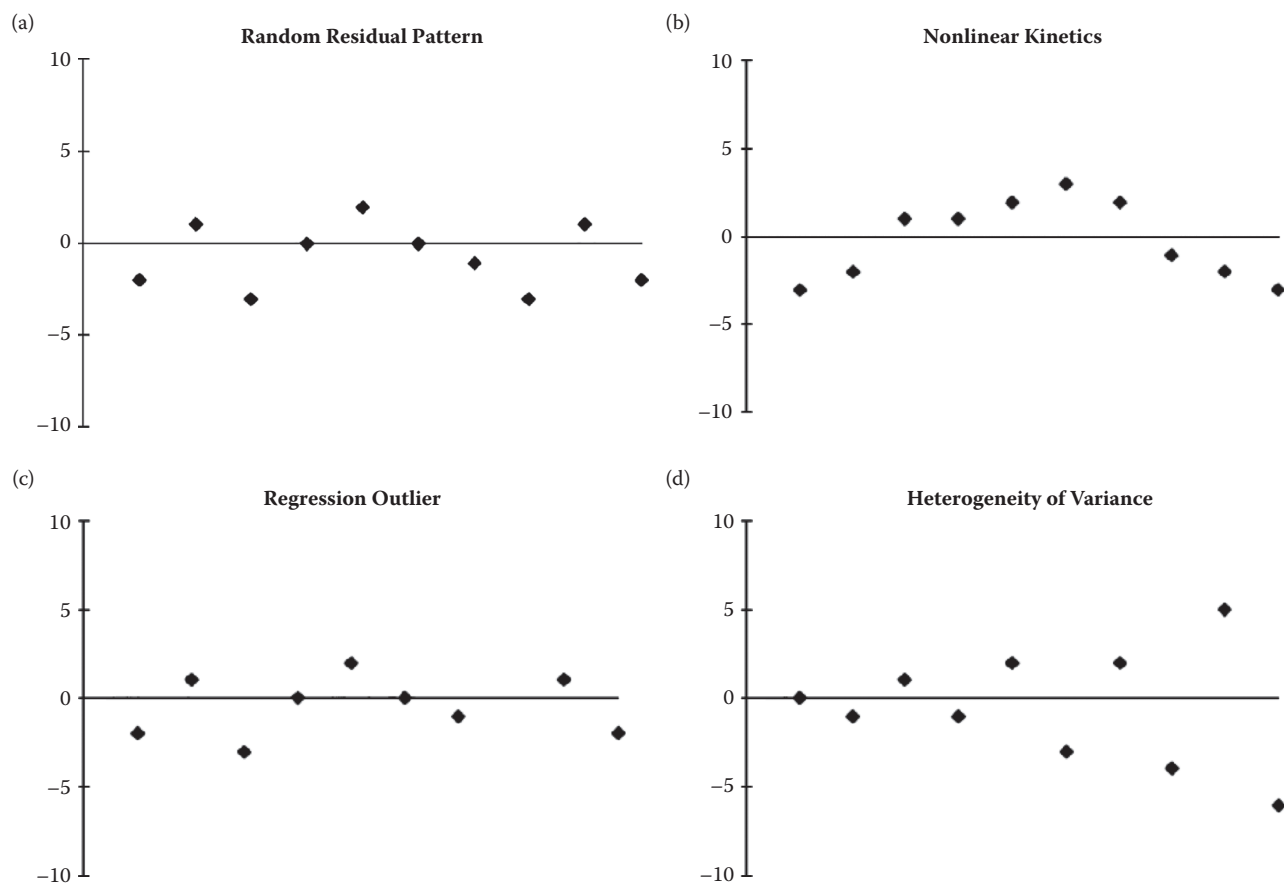


Fig. 3 Residual plots for standard curve fits exhibiting various behaviors: (a) normal behavior with random variability, (b) nonlinear kinetics, (c) a regression outlier, and (d) heterogeneity of variance

where

$$g = \frac{t_{n-2}^2}{t_b^2} = \frac{t_{n-2}^2 \cdot \hat{\sigma}^2}{b^2 \cdot SXX}$$

t_{n-2} is the appropriate percentile from the t -distribution with $n - 2$ degrees of freedom, $\hat{\sigma}$ is the standard error of the regression, t_b is the statistic associated with a test of significance of the regression slope, and SXX is the sum of squared deviations in the concentrations. In most cases, g will be small, and can be treated as if it were equal to “0” in the calculation of the standard error of estimate of the predicted concentration. In this case, the standard error can be utilized in conjunction with the predicted concentration to calculate the relative standard deviation (RSD) of the result:

$$\text{RSD} = 100 \cdot \frac{\frac{\hat{\sigma}}{b} \cdot \sqrt{1 + 1/n + \hat{X}^2/SXX}}{\hat{X}} \%$$

In this way, the variability of each prediction can be determined, and results that meet a prespecified restriction (e.g., $\leq 20\%$ RSD) can be reported.

Nonlinear Calibration

Many analytical methodologies, primarily immunoassays, generate nonlinear standard curves. As indicated earlier, the linear range of the standard curve can be identified, and linear calibration can be used to fit the standard curve and interpolate sample results. This portion of the standard curve is only approximately linear, however, and it might be more suitable to fit the entire curve using one of a variety of concentration kinetics equations. In fact, basic immunochemistry principles predict that the standard curve can be modeled using the four-parameter logistic regression equation (Fig. 4). Letting Y be the instrument readout associated with an individual standard curve concentration X , the four-parameter logistic equation is as follows:

$$Y = D + \frac{A - D}{1 + \left(\frac{X}{C}\right)^B}$$

In this equation, the parameter D represents the minimum response (usually the background), while A represents the level in response associated with the saturation of the

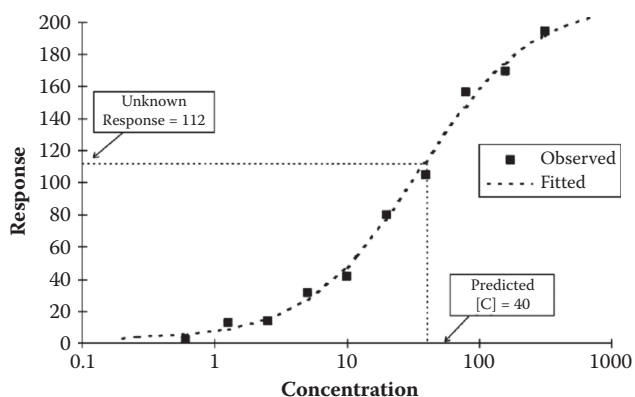


Fig. 4 Illustration of a standard curve fit using four-parameter logistic regression; inverse prediction is utilized to estimate the concentration corresponding to the response in an unknown

“binder.”^[5] The parameters B and C are related to the slope and intercept from the linear regression of logistic transformed response on log concentration. In fact, C corresponds specifically to EC_{50} , the concentration of the standard predicted to yield 50% activity in the assay. This parameter can be used to monitor the stability of the standard.

The nonlinear equation can be fit using conventional nonlinear curve-fitting methodology. Similar regression diagnostics as employed with linear calibration can be utilized in conjunction with nonlinear regression (e.g., residuals plots, regression outliers), while the interpolated concentration for an unknown sample can be supplemented by its standard error, yielding an assessment of the quantifiable range for an individual assay run:

$$\hat{X} = C \cdot \left[\frac{A - Y}{Y - D} \right]^{1/B}$$

The standard error of a predicted value can be established via estimates of the variances and covariances of the parameter estimates, and measurement variability, using a first-order Taylor series approximation:^[6]

$$\text{var}(\hat{X}) = \sum \left(\frac{\partial \hat{X}}{\partial P} \right)^2 \cdot \text{var}(P) + \sum \sum \left(\frac{\partial \hat{X}}{\partial P} \right) \left(\frac{\partial \hat{X}}{\partial Q} \right) \cdot \text{cov}(P, Q) + \left(\frac{\partial \hat{X}}{\partial Y} \right)^2 \cdot \hat{\sigma}^2$$

where $\hat{X}$ is the predicted concentration, P and Q are parameters of the nonlinear model, and $\text{var}(P)$ and $\text{cov}(P, Q)$ represent the approximate variances and covariances among the parameters.

The design of an assay utilizing a nonlinear standard curve will differ from that of a linear calibration system in several ways. Typically more standard concentrations will be necessary to support the estimation of additional parameters in the nonlinear model. Immunoassay design

usually requires that standard curve points be extended to measure the asymptotes of the curve (A and D in the four-parameter logistic model). A common concern is that this implies that test samples can be read from these “flat” portions of the standard curve; however, the reliable portion of the standard curve can be calculated for each run using methods discussed previously for calculating the standard error of a predicted concentration from the nonlinear curve fit.

Other Calibration Conventions

One means to ameliorate the effects of increasing variability with increasing response is to transform the data. Commonly used transformations are the log and square root. Another widely used solution is to implement a weighted regression, where the weight assigned to each standard curve concentration is a function of the predicted response. This method is called *iteratively reweighted least squares*;^[7] it can easily be implemented with computation software such as EXCEL®.

The weighting functions of the predicted response (y) that are commonly used are (1) weight = $1/y$, when the response is believed to be generated by a Poisson process (e.g., when the response is a count, such as in RIAs; note that this is analogous to performing a square root transformation of the instrument readout); and (2) weight = $1/y^2$, when it can be established that the %RSD is a constant (note that this is analogous to performing a log transformation of the instrument readout).

An adaptation of the standard calibration method is a parallel-line design. In the parallel-line design, a test sample is diluted in conjunction with the standard preparation, forming simultaneous dose–response curves. Statistical conformance to the assumptions of the parallel-line model is tested prior to the measurement of the *relative potency* (RP) of the test preparation to the standard. The assumptions tested are specific to the mathematical model utilized to fit the simultaneous dose–response profiles. When a simple linear model is used to fit the data, tests of linearity and parallelism of the dose–response curves usually precede potency estimation.^[8] The measured relative potency corresponds to the horizontal distance between parallel response profiles, at any level of response; thus, when the response profiles are parallel, this horizontal distance is the same at all levels of response. When the response profiles are not parallel, the estimated relative potency (horizontal distance) is not unique and can vary depending upon the level of response used to predict the relative potency (see Fig. 5).

The four-parameter logistic regression equation is frequently used to model response in a parallel-line immunoassay. The goodness of fit of various parameterizations of the simultaneous four-parameter logistic equation can be tested, preliminary to estimating the relative potency.^[5]

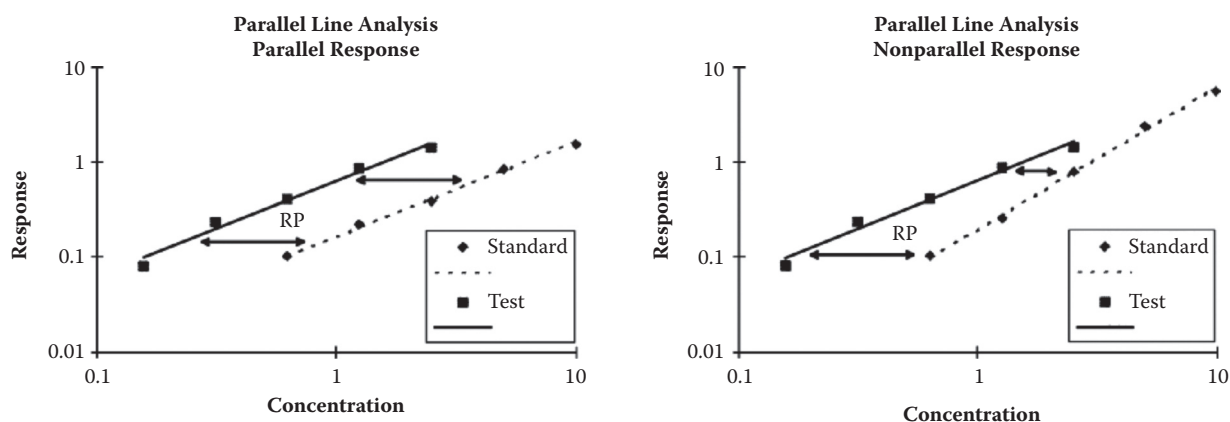


Fig. 5 The relative potency represents the horizontal distance between dose–response profiles. The relative potency is unique when the profiles are parallel, but it is different for different levels of response when the profiles are not parallel

In either case, when the assumptions are met, the relative potency of the test preparation to the standard can be estimated, along with an appropriate confidence interval:

$$M = \log(\text{RP}) = \bar{x}_s - \bar{x}_t - \frac{\bar{y}_s - \bar{y}_t}{b}$$

and

$$(\bar{x}_s - \bar{x}_t) + \frac{\left[M - \bar{x}_s + \bar{x}_t \pm \frac{t_{df}}{b} \left\{ \hat{\sigma}^2 (1-g)(1/n_s + 1/n_t) + (M - \bar{x}_s + \bar{x}_t)^2 / \sum SXX \right\}^{1/2} \right]}{1-g}$$

where

$$g = \frac{t_{df}^2}{t_b^2} = \frac{t_{df}^2 \cdot \hat{\sigma}^2}{b^2 \sum SXX}$$

x represents the dose, y is the response, b is the pooled slope from the standard and test preparations, t_{df} is the appropriate percentile from the t -distribution with $n_s + n_t - 2$ degrees of freedom, $\hat{\sigma}$ is the standard error of the regression, t_b is the statistic associated with a test of significance of the pooled regression slope, and SXX is the sum of squared deviations in the concentrations, which is summed between standard and test preparations.

The relative potency is unitless and represents the “shift” in concentration of the dose–response curve for the test sample relative to the standard sample. A relative potency of test to standard equal to 1.00 represents “equipotence” of the two samples. The reported potency of the test preparation can be converted to an absolute scale by multiplying the known potency of the standard by the relative potency.

Design components of a parallel-line assay are the same as those for a calibration assay, with the additional

requirements implicit in the reliable estimation of a relative potency. Thus in addition to the choice of the dose range and the number of doses of the test and standard preparations, these should be selected so as to achieve approximately equal average responses across doses for the two preparations.

ASSAY OPTIMIZATION

Assay development commences with scientifically reasonable levels of factors that might affect the assay (factors such as pressure, pH, temperature of incubation, time of incubation, and levels of certain key reagents) and then proceeds with empirical studies that seek to optimize the levels of these factors. The goal of the optimization is to achieve satisfactory performance in one or several key attributes of the assay. Examples of performance attributes include background reactivity in an immunoassay, linearity in an HPLC, and repeatability and precision of the method (see the entry *Assay Validation*).

Key to the successful optimization of a method is a strategic plan for studying the effects of the assay factor(s). This can be accomplished using tools such as factorial and fractional-factorial experimental designs.^[9] Consider an example in which an experiment is conducted to minimize the background in an immunoassay. The factors that are thought to influence background reactivity (Y in optical density, OD) are incubation temperature ($X1$, ranging from 30°C to 35°C), incubation time ($X2$, ranging from 6 to 8 hr), and dilution of conjugate ($X3$, ranging from

Table 1 A 2³ Factorial Experiment Studying the Effects of Temperature, Time, and Dilution of Conjugate on Background Level in an EIA

Run	Temperature (X1)	Time (X2)	Dilution (X3)	OD
1	30°C	6 hr	1:1000	0.29
2	30°C	6 hr	1:4000	0.16
3	30°C	8 hr	1:1000	0.34
4	30°C	8 hr	1:4000	0.22
5	35°C	6 hr	1:1000	0.30
6	35°C	6 hr	1:4000	0.19
7	35°C	8 hr	1:1000	0.35
8	35°C	8 hr	1:4000	0.20

Runs are usually performed in randomized order.

1:1000 to 1:4000). A 2³ factorial design is used to study simultaneously the effects of these three factors in Table 1. By inspection, it appears that conjugate dilution (X3) has an effect on background reactivity, with the 1:1000 dilution yielding a higher background than the 1:4000 dilution. The effect of dilution of conjugate can be estimated from experimental results:

$$X3 = \frac{0.29 + 0.34 + 0.30 + 0.35}{4} - \frac{0.16 + 0.22 + 0.19 + 0.20}{4} = 0.32 - 0.19 = 0.13$$

The interaction between factors can also be estimated; for example, the interaction of temperature with dilution of conjugate (X1* X3) is:

$$X1 * X3 = \frac{0.29 + 0.34 + 0.19 + 0.20}{4} - \frac{0.16 + 0.22 + 0.30 + 0.35}{4} = 0.255 - 0.2575 = -0.0025$$

The effects of these factors and their interactions can be plotted in order of largest to smallest absolute effect in a Pareto plot,^[9] as shown in Fig. 6. This plot shows that dilution of conjugate (X3) asserts the greatest effect on background OD, while incubation time (X2) has the next largest effect. Temperature of incubation (X1) and the two- and three-factor interactions appear to have little affect.

Another aspect of assay optimization is the replication strategy used to test a sample. This is discussed in the entry *Assay Validation*.

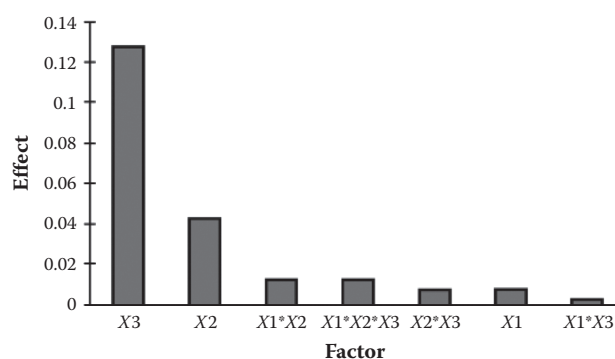


Fig. 6 Pareto plot showing absolute effects of temperature (X1), time (X2), and dilution of conjugate (X3) on background reactivity in an immunoassay

BIOLOGICAL ASSAYS AND ASSAYS OF ANALYTES IN COMPLEX MATRICES

Many biologicals and some drugs are tested using biological assays (bioassays), which measure the biological activity of the analyte. This is typically true of vaccines, which induce an antibody response in animals as well as in man. In a bioassay for a vaccine, a fixed number of mice (or some other animal species that responds to the vaccine; vaccines are also tested ex vivo, in tissue cells that demonstrate a prescribed response, usually cytopathicity due to infection with a vaccine virus) are injected, each with one of a graded series of doses of the vaccine. After a prescribed period of time, the mice are bled and the presence or absence of antibody is determined in each animal. The data become a series of response rates across doses. Assays of this sort are often called “quantal response” assays. A typical quantal response curve is illustrated in Fig. 7. The reported potency is usually either the dose predicted to yield responses in 50% of immunized animals (usually called either the ED₅₀ in an in vivo animal model or TCID₅₀ when tissue culture infection is the analytical

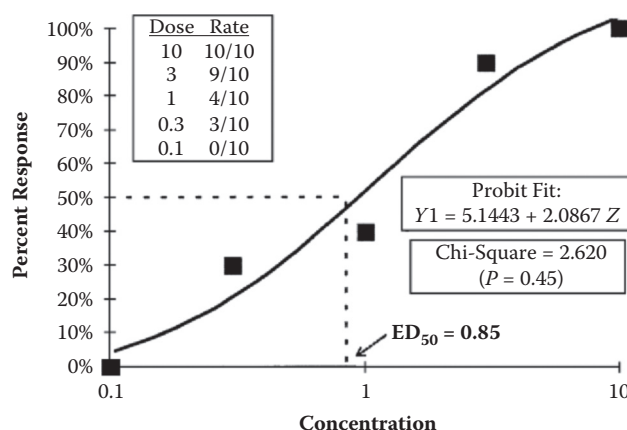


Fig. 7 Quantal response curve fit using probit analysis; the estimated ED₅₀ represents the level of the test preparation that is predicted to yield 50% response in animals

endpoint) or the relative potency of the test preparation to a standard (this will be mathematically equal to the ratio of ED_{50} 's).

There are several ways to calculate these endpoints, among them methods due to Reed and Muench^[10] and to Spearman and Kärber, as well as probit and logit analyses. The Reed and Muench method is easy to table and was popular before computers were enlisted to perform the complex calculations associated with other methods of analysis. It does not provide, however, any confirmation of the reliability of the measured potency. The Spearman and Kärber method furnishes an estimate of the standard error of the estimate; the probit method offers this along with a test of goodness of fit of the probit model.

Several design factors affect the statistical properties of an in vivo assay. The number of doses, the dose range, and the number of animals or tissue culture wells per dose affect the variability of the measured potency. Care should be taken in defining the cutoff when an immunoassay is used to score positive responders. A cutoff that yields a high "false-negative" rate in the immunoassay will result in a potency that is negatively biased. A parallel-quantal-response design can be used to ameliorate the affects of the operating characteristics of the immunoassay. As with parallel-line design in immunoassay, this requires that the quantal response curves be "parallel," and it results in the measurement of a relative potency of the test to the standard.

CONCLUSION

Assays are used to characterize the content and properties of pharmaceutical products and vaccines. Careful development of these methods helps ensure the consistency, safety, and effectiveness of products for human and animal

use. Statistical considerations during the development of an assay include the analytical method design, the processing algorithm, and optimization strategies such as multifactor design of experiments. Due consideration should also be given to the variability of the analytical measurement, in assessing its value as a descriptor of the analytical property of interest. Development is complete when all important processing parameters have been identified and acknowledged in a standard operating procedure for the method. Only when development is complete should assay validation commence, and the method be put into routine use in conjunction with adequate assay quality control.

REFERENCES

1. Anscombe, F. Assay development. *Am. Stat.* **1973**, 27, 17–21.
2. Barnett, V.; Lewis, T. *Outliers in Statistical Data*; Wiley: New York, 1978, 252–256.
3. Riley, C.M., Rosanske, T.W., Eds.; *Development and Validation of Analytical Methods*; Elsevier Science Inc.: Tarrytown, NY, 1996, 263–270.
4. Snedecor, G.; Cochran, W. *Statistical Methods*, 6th; Iowa State University Press: Ames, IA, 1967,; 159–160.
5. De Lean, A.; Munson, P.; Rodbard, D. *Am. J. Physiol.* **1978**, 235, 97–102.
6. Morgan, B. *Analysis of Quantal Response Data*; Chapman & Hall: New York, 1992, 370, 371.
7. Myers, R. *Classical and Modern Regression with Applications*; Duxbury Press: Boston, MA, 1986; 204–211.
8. Finney, D. *Statistical Method in Biological Assay*, 3rd Ed.; Macmillan: New York, 1978, 69–99.
9. Haaland, P. *Experimental Design in Biotechnology*; Marcel Dekker: New York, 1989, 5–7, 45.
10. Reed, L.; Muench, H. *Am. J. Hyg.* **1938**, 27, 493–497.
11. Morgan, B. *Analysis of Quantal Response Data*; Chapman & Hall: New York, 1992, 306–310.

Assay Validation

Timothy L. Schofield

Merck Research Laboratories, West Point, Pennsylvania, U.S.A.

Abstract

Upon completion of the development of an assay, and prior to implementation, a well-conceived assay validation is carried out to demonstrate that the procedure is fit for use. This entry will introduce the validation parameters used to describe the characteristics of an analytical method.

INTRODUCTION

Upon completion of the development of an assay, and prior to implementation, a well-conceived assay validation is carried out to demonstrate that the procedure is *fit for use*. Regulatory requirements specify the assay characteristics that are subject to validation and suggest experimental strategies to study these properties. These requirements, when united with sound statistical experimental design, furnish an effective and efficient validation plan. That plan should also address the form of statistical processing of the experimental results and thereby ensure reliable and meaningful measures of the properties of the method.

This entry will introduce the validation parameters used to describe the characteristics of an analytical method. Performance metrics associated with those parameters will be defined, with illustrations of their utility in application. Realistic and relevant acceptance criteria are proposed, as a basis for evaluating a procedure. An example of a validation protocol will illustrate the design of a comprehensive exploration, in conjunction with the statistical methods used to extract the relevant information.

TERMINOLOGY

Most of the terminology related to assay validation can be found in the entry on *Assay Development*. Some terms that are unique to this topic are *relative standard deviation*, *spike recovery*, and *dilution effect*.

The relative standard deviation (RSD) is a measure of the proportional variability of an assay and is usually expressed as a percentage of the measured value:

$$\text{RSD} = 100 \times \frac{\hat{\sigma}_{\text{assay}}}{\hat{x}} \%$$

Note that the RSD is the same as the coefficient of variation in statistics. Other conventions for calculating the RSD, which are appropriate for measurements that are

lognormally distributed, utilize $\hat{\sigma}_{\log \hat{x}}$, the log standard deviation or the fold variability:

$$\begin{aligned} \text{RSD} &= 100 \hat{\sigma}_{\log \hat{x}} \% \\ \text{or RSD} &= 100(\text{FV} - 1)\% = 100(e^{\hat{\sigma}_{\log \hat{x}}} - 1)\% \end{aligned}$$

Note that the log standard deviation provides an approximation to the RSD; it should be restricted to levels below 20% RSD. An important feature of the RSD is that it is calculated from the predicted measurements for a sample, such as concentrations, and not from an intermediate, such as the instrument readout. These are the same only in the case of linear calibration.

Spike recovery is the process of measuring an analyte that has been distributed as a known amount into the sample matrix, with results usually reported as percent recovery or percent bias. Dilution effect is a measure of the proportional bias measured across a dilution series of a validation sample; it is usually reported as percent bias per dilution increment (usually percent bias per twofold dilution).

REGULATORY REQUIREMENTS

The U.S. Pharmacopeia specifies that the “Validation of an analytical method is the process by which it is established, in laboratory studies, that the performance characteristics of the method meet the requirements for the intended analytical applications.”^[1] Key to this definition is the treatment of a validation study as an *experiment*, with the goal of that experiment being to demonstrate that the assay is fit for use. As an experiment, a validation study should be suitably designed to meet its goal, while the acceptance criteria should yield opportunities to judge the assay as unsuitable as well as suitable, possibly requiring further development. This does not suggest that it is the role of the validation study to optimize the assay; this should be completed during the assay development. In fact, the properties

of the assay are probably well understood by the time the validation experiment is undertaken. The validation experiment should serve as a “snapshot” of the long-term routine performance of the assay, thereby documenting its characteristics so that we can make reliable decisions during drug development and control.

The International Conference on Harmonization (ICH) has devoted two topics of its guidelines to analytical method validation. The first topic, entitled “Guideline on Validation of Analytical Procedures: Definitions and Terminology,”^[12] offers a list of validation parameters that should be considered when implementing a validation experiment and guidance toward the selection of those parameters. The second topic, entitled “Validation of Analytical Procedures: Methodology,”^[13] provides direction on the design and analysis of results from a validation experiment.

ASSAY VALIDATION PARAMETERS AND VALIDATION DESIGN

It is convenient for the purposes of planning a validation experiment to dichotomize assay validation parameters into two categories: parameters related to the accuracy of the analytical method and parameters related to variability. In this way, the analyst will choose the test sample set that will be carried forward into the validation, in accord with accuracy parameters, and devise a replication plan for the validation to satisfy the estimation of the variability parameters.

The validation parameters related to the accuracy of an analytical method are accuracy, linearity, and specificity. The *accuracy* of an analytical method “expresses the closeness of agreement between the value which is accepted either as a conventional value or an accepted reference value, and the value found.” Accuracy is usually established by spiking known quantities of analyte (see the entry *Assay Development* for definitions of terms) into the sample matrix and demonstrating that these can be completely recovered. Conventional practice is to spike using five levels of the analyte: 50%, 75%, 100%, 125%, and 150% of the declared content of analyte in the drug (note that this is more stringent than the ICH, which requires “a minimum of nine determinations over a minimum of three concentration levels”). The experimenter is not restricted to these levels but should plan to spike through a region that embraces the range of expected measurements from samples that will be tested during development and manufacture. Accuracy can also be determined relative to a “referee” method; in the case of complex mixtures, such as combination vaccines, accuracy can be judged relative to a monovalent control.

The *linearity* of an analytical procedure is “its ability (within a given range) to obtain test results that are directly proportional to the concentration (amount) of analyte in the sample.” This parameter is often confused with the

“graphical linearity” of the standard curve; in fact, the guidelines encourage the experimenter to evaluate linearity “by visual inspection of a plot of signals as a function of concentration or content.” The principle of analytical linearity, however, should not be confused with an abstract property, such as the “straightness of the standard curve,” but rather should be based upon the attribute that graded concentrations of the analyte yield measurements that are in proper proportion. Thus, for example, if a sample is tested in twofold dilution, the measured concentrations in those samples should yield a twofold series. In this way, linearity can be established from the spiking experiment outlined earlier, while a dilution series is usually employed when the assessment of linearity cannot be achieved through spiking (i.e., when the purified analyte does not exist, such as in vaccines).

Specificity is “the ability to assess unequivocally the analyte in the presence of components which may be present.” Thus specificity speaks to the variety of sample matrices containing the analyte, including process intermediates and stability samples. The ICH guidelines provide a comprehensive description of the means to establish specificity in identity tests and in analytical methods for impurities and content. When the samples are available, specificity can be established through testing of the analyte-free matrix.

The assay validation parameters related to variability are precision, repeatability, ruggedness, limit of detection, and limit of quantitation. The *precision* of an assay “expresses the closeness of agreement (degree of scatter) between a series of measurements obtained from multiple sampling of the same homogeneous sample under the prescribed conditions.” Precision is frequently called “interrun” variability, where a run of an assay represents the independent preparation of assay reagents, tests samples, and a standard curve. The *repeatability* of an assay “expresses the precision under the same operating conditions over a short interval of time,” and is frequently called “intrarun” variability. Repeatability and precision can be studied simultaneously, using the levels of a sample required to establish accuracy and linearity, in a consolidated validation study design such as that depicted in Table 1. Here precision is associated with the multiple experimental runs, while repeatability is associated with the run-by-level interaction. In many cases, the experimenter may wish to test true within-run replicates rather than employ this subtle statistical artifact. The runs in this design can be strategically

Table 1 Strategic Validation Design Employing Five Levels of the Analyte Tested in k Independent Runs of the Assay

Run	Level				
	1	2	3	4	5
1	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_{n1}	y_{n2}	y_{n3}	y_{n4}	y_{n5}

allocated to *ruggedness* parameters, such as laboratories, operators, and reagent lots, using a combination of nested and factorial experimental design strategies. Thus *operators* might be nested within *laboratory*, while *reagent lot* can be crossed with *operator*.

The *robustness* of an analytical method “is a measure of its capacity to remain unaffected by small, but deliberate, variations in method parameters,” and might be more suitably established during the development of the assay (see the entry on *Assay Development*). Note that while ruggedness parameters represent uncontrollable factors affecting the analytical method, robustness parameters can be varied and should be controlled when it has been observed that they have an effect on assay measurement.

The *limit of detection* (LOD) and *limit of quantitation* (LOQ) of an assay can be obtained from replication of the standard curve during the implementation of the consolidated validation experiment. The LOD of an analytical procedure “is the lowest amount of analyte in a sample that can be detected but not necessarily quantitated as an exact value,” while the LOQ “is the lowest amount of analyte in a sample that can be quantitatively determined with suitable precision and accuracy.” The limit of quantitation need not be restricted to a lower bound on the assay, but might be extended to an upper bound of a standard curve, where the fit becomes flat (for example, when using a four-parameter logistic regression equation to fit the standard curve).

Finally, the composite of the ranges identified with acceptable accuracy and precision is called the *range* of the assay.

CHOICE OF VALIDATION PARAMETERS

The choice of parameters that will be explored during an assay validation is determined by the intended use as well as by the practical nature of the analytical method. For example, a biochemical assay using a standard curve to establish drug content might require the exploration of all of these parameters if the assay is to be used to determine low as well as high levels of the analyte (such as an assay used to determine drug level in clinical samples or an assay that will be used to measure the content of an unstable analyte). If the assay is used, however, to determine content in a stable preparation, the LOD and LOQ need not be established. On the other hand, an assay for an impurity requires adequate sensitivity to detect and/or quantify the analyte; thus the LOD and LOQ become the primary focus of the validation of an impurity assay.

Many of these assay validation parameters have limited meaning in the context of biological assay and various other potency assays. There is no means to explore the accuracy and linearity of assays in animals or tissue culture, where the scale of the assay is defined by the assay; thus validation experiments for this sort of analytical method are usually restricted to a study of the specificity and ruggedness

of the procedure. As discussed previously, the accuracy of a potency assay may be limited to establishing linearity with dilution when the purified analyte is unavailable to conduct a true spike-recovery experiment.

VALIDATION METRICS AND ACCEPTANCE CRITERIA

As previously defined, the goal of the validation experiment is to establish that the analytical method is fit for use. Thus for an impurity assay, for example, the goal of the validation experiment should be to show that the sensitivity of the analytical procedure (as measured by the LOD for a limit assay and by the LOQ for a quantitative assay) is adequate to detect a meaningful level of the impurity. Assays for content and potency are usually used to establish that a drug conforms to specifications that have been determined either through clinical trials with the drug or from process/product performance. It is important to point out that the process capability limits should include the manufacturing distribution of the product characteristic under study, the change in that characteristic because of instability under recommended storage conditions, as well as the effects on the measurement of the characteristic because of the performance attributes of the assay. In the end, the analytical method must be capable of reliably discriminating satisfactory from unsatisfactory product against these limits. The portion of the process range that is a result of measurement, including measurement bias as well as measurement variability, serves as the foundation for setting acceptance criteria for assay validation parameters.

Validation parameters related to accuracy are rated on the basis of recovery or bias (bias = 100 – recovery, usually expressed as a percentage), while validation parameters related to random variability are appraised on the basis of variability (usually expressed as % RSD). These can be combined with the process variability to establish the “process capability” of the measured characteristic:^[4]

$$\begin{aligned} \text{CP} &= \frac{\text{Specification range}}{6 \cdot \text{Product range}} \\ &= \frac{\text{Specification range}}{6\sqrt{\sigma_{\text{product}}^2 + \sigma_{\text{assay}}^2 + \text{Bias}^2}} \end{aligned}$$

Process capability, in turn, is related to the percentage of measurements that are likely to fall outside of the specifications. Thus acceptance criteria on the amount of bias and analytical variability can be established based upon the knowledge of the product variability and the desired process capability.

When the limits have not been established for a particular product characteristic, such as during the early development of a drug or biologic, acceptance criteria for an assay attribute might be specified on the basis of a typical

expectation for the analytical method as well as on the nature of the measurement in the particular sample. Thus for example, high-performance liquid chromatography for the active compound in the final formulation of a drug might be typically capable of yielding measurements equal to or less than 10% RSD, while an immunoassay used to measure antigen content in a vaccine might only be capable of achieving up to 20% RSD. Measurement of a residual or an impurity by either method, on the other hand, may only achieve up to 50% RSD, owing to the variable nature of measurement at low analyte concentrations.

VALIDATION DATA ANALYSIS

Analysis of Parameters Related to Accuracy

Analytical accuracy and linearity can be parameterized as a simple linear model, assuming either absolute or relative bias. In one model, if we let μ represent the known analyte content and x is the measured amount, then the accuracy can be expressed as $x = \mu$ at a single concentration, or as $x = \alpha + \beta\mu$, where $\alpha = 0$ and $\beta = 1$, across a series of concentrations. This linear equation can be rewritten as

$$x = \mu + [\alpha + (\beta - 1)\mu].$$

In this parameterization, α is related to the accuracy of the analytical method and represents the constant bias in an assay measurement (usually reported in the units of measurement of the assay), while $(\beta - 1)$ is related to the linearity of the analytical method and represents the proportional bias in an assay measurement (usually reported in the units of measurement in the assay per unit increase in that measurement, e.g., 0.02 μg per microgram increase in content). Data from a spiking experiment can be utilized to estimate α and $\beta - 1$, and these (with statistical confidence limits if the validation experiment has been designed to establish statistical conformance) can be compared with the acceptance criteria established for these parameters.

To illustrate, consider the example in Table 2. A validation study has been performed in which samples have been prepared with five levels of an analyte ([C] in mg), and their

Table 2 Example of Results from a Validation Experiment in Which Five Levels of an Analyte Were Tested in Six Runs of the Assay

[C]	1	2	3	4	5	6	Avg.
3	3.3	3.4	3.2	3.1	3.4	3.3	3.3
4	4.2	4.2	4.2	4.4	4.0	4.3	4.2
5	5.0	4.9	4.9	5.0	4.8	4.9	4.9
6	5.8	5.8	5.6	5.9	5.9	5.7	5.8
7	6.4	6.7	6.7	6.8	6.6	6.4	6.6

content is determined in six runs of an assay. These data are depicted graphically, along with the results from their analysis, in Fig. 1. A linear fit to the data yields the following estimates of intercept (note that the data have been centered in order to obtain a “centered” intercept) and slope:

$$\hat{\alpha}' = -0.04 \text{ } (-0.08, 0.01), \quad \hat{\beta} = 0.81 \text{ } (0.70, 0.85)$$

The slope can be utilized to estimate the bias per unit increase in drug concentration:

$$\hat{\beta} - 1 = -0.19 \text{ } (-0.30, -0.15)$$

This proportional bias is significant, owing to the fact that the confidence interval excludes 0. It is more desirable, however, to react to the magnitude of the bias and to judge that there is an unacceptable nonlinearity when the bias is in excess of a prespecified acceptance limit. Thus there is a 0.2-mg decrease per unit increase in concentration (as much as -0.3 mg per unit decrease in concentration based on the confidence interval). If the range in the concentration of the drug is typically 4–6mg (i.e., a 2-mg range), this would predict that the bias as a result of “nonlinearity” is 0.4 mg (0.6 mg in the confidence interval). This can be judged to be of consequence or not in testing product; for example, if the specification limits on the analyte are 5 ± 1 $\mu\text{g}/\text{dose}$, much of this range will be consumed by the proportional bias, potentially resulting in an undesirable failure rate in the measurement of that analyte.

When the purified analyte is unavailable for spiking, and the levels of the analyte in the validation have been attained through dilution, then the data generated from that series can be alternatively evaluated using the model $x = \alpha\mu^\beta$ (note that μ can represent either the expected concentration upon dilution or the actual dilution). Taking the logs, the data from the dilution series is used to fit the linear model,

$$\log(x) = \log(\alpha) + \beta \log(\mu)$$

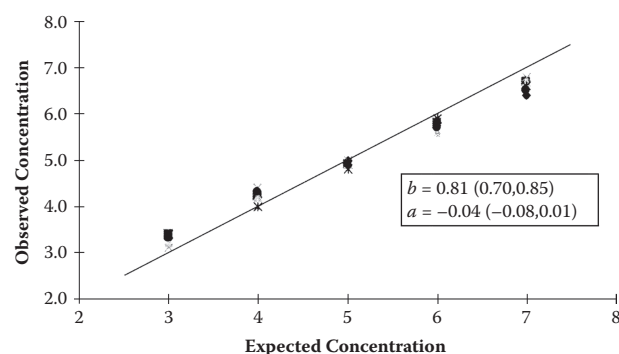


Fig. 1 Validation results from an assay demonstrating proportional bias, with low-concentration samples yielding high measurements and high-concentration samples yielding low measurements

The intercept of this model has no practical meaning, while the proportional bias (sometimes called “dilution effect”) can be estimated as $2^{\beta-1} - 1$ if μ is in units of concentration and as $2^{\beta+1} - 1$ if μ is in units of dilution (note that the units of “dilution effect” are percent bias per dilution increment; the dilution increment used here is twofold).

To illustrate, consider the example in Table 3. A dilution series of a sample is performed in each of the five runs of an assay. These data are depicted graphically, along with the results from their analysis, in Fig. 2. A linear fit of log-titer to log-dilution yields an estimated slope equal to $\hat{\beta} = -1.01$ ($-1.04, -0.98$). The corresponding dilution effect is equal to $2^{-1.01+1} - 1 = -0.007$ (i.e., a 0.7% decrease per twofold dilution).

The dilution effect can also be calculated as a function of the estimated slopes for the standard (slope as a function of dilution rather than concentration) and the test sample:

$$\text{Dilution effect} = 2^{1-\hat{\beta}_T/\hat{\beta}_S}$$

with the corresponding standard error a function of the standard error of a ratio:

$$\text{SE}(\hat{\beta}_T/\hat{\beta}_S) = \sqrt{\frac{1}{\hat{\beta}_S^2} \cdot \left[\text{VAR}(\hat{\beta}_T) + (\hat{\beta}_T/\hat{\beta}_S)^2 \cdot \text{VAR}(\hat{\beta}_S) \right]}$$

This has the advantage over the earlier calculation of being a more reliable estimate of the underlying variability in the measurements, coming from the contribution to the estimate of variability from the data for the standard.

Table 3 Example of Results from a Validation Experiment in Which Five Twofold Dilutions of an Analyte were Tested in Five Runs of the Assay

Dilution	1	2	3	4	5	Titer
1	75	90	93	79	72	81
2	43	45	46	35	40	83
4	20	23	22	18	21	83
8	11	10	12	10	10	85
16	4	5	6	4	5	78

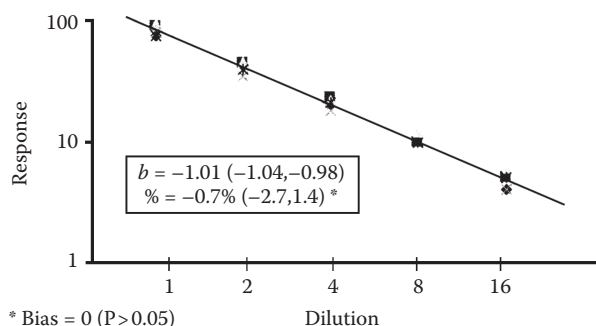


Fig. 2 Results from a validation series demonstrating satisfactory linearity, i.e., approximately a twofold decrease in response with twofold dilution

A special case of accuracy comes from a comparison of an experimental analytical method with a validated “referee” method. Paired measurements in the two assays can be made on a panel of samples, and these can be utilized to establish the “linearity” and the accuracy of the experimental method relative to the “referee” method. The paired measurements are a sample from a multivariate population, and the “linearity” of the experimental method to the “referee” method can be estimated using principal component analysis.^[5] A “concordance slope” can be estimated from the elements of the first characteristic vector of the sample covariance matrix (a_{21}, a_{11}):

$$\hat{\beta}_C = \frac{a_{21}}{a_{11}}$$

The standard error of the “concordance slope” is obtained from these along with the other elements of the characteristic value evaluation (a_{i1} is the i th element of the first characteristic vector and l_i is the corresponding characteristic roots):

$$\text{SE}(\hat{\beta}_C) = \sqrt{\left[\frac{V(a_{21})}{a_{21}^2} + \frac{V(a_{11})}{a_{11}^2} - 2 \frac{\text{cov}(a_{21}, a_{11})}{a_{21}a_{11}} \right] \hat{\beta}_C^2}$$

where

$$V(a_{i1}) = \left[\frac{l_i l_j}{n \cdot (l_j - l_i)^2} \right] a_{i2}^2$$

and

$$\text{cov}(a_{21}, a_{11}) = \left[\frac{l_i l_j}{n \cdot (l_j - l_i)^2} \right] a_{12} a_{22}$$

The “concordance slope” can be used, along with the standard error of that estimate, to measure the discordance of the experimental analytical method relative to the “referee” method and its corresponding confidence interval:

$$\text{Discordance} = 2^{\hat{\beta}_C - 1}$$

This is similar to the “dilution effect” discussed earlier and represents the percentage difference in reported potency per twofold increase in result as measured in the “referee” method.

When the two assays are scaled alike, the accuracy of experimental method relative to the “referee” method can be assessed through a simple paired variate analysis. When the percent difference is expected to be constant across the two methods, the data are analyzed on a log scale. An estimate of the simple paired difference in the logs can be used, in conjunction with the standard error of the paired difference, to estimate the percentage difference in

measurements obtained from the experimental procedure relative to the “referee” method.^[6]

$$\bar{d} \pm t_{n-1} s_d / \sqrt{n}$$

$$\text{Percent difference} = 100(e^{\bar{d}} - 1)\%$$

Note that as usual, the significance of the percentage difference should be judged primarily on the basis of practical considerations rather than statistical significance.

The specificity of an analytical procedure can be assessed through an evaluation of the results obtained from a “placebo” sample (i.e., a sample containing all constituents except the analyte of interest) relative to background. A statistical evaluation might be composed of either a comparison of instrument measurements obtained for this “placebo,” with background measurements (i.e., through background-corrected “placebo” measurements), or from measurements made on the “placebo” alone. In either case, an increase over the detection level of the assay would indicate that some constituent of the sample, in addition to analyte, is contributing to the measurement in the sample. This should be assessed by noting that the upper bound on a confidence interval falls below some prespecified level in the assay.

Analysis of Parameters Related to Variability

Estimates of repeatability (intrarun, or within-run, variability) and precision (interrun, or between-run, variability) can be established by performing a variance component analysis on the results obtained from the data generated to assess accuracy and linearity.^[7] Other sources of variability (e.g., laboratories, operators, and reagent lots) can be incorporated into the pattern of runs performed during the validation to estimate the effects of these sources of variability. The expected mean squares are shown in Table 4. From this, $\hat{\sigma}_{\text{withinrun}}^2$ = mean square error (MSE) and $\hat{\sigma}_{\text{betweenrun}}^2$ = (mean square for runs (MSR)–MSE)/5 (note that the MSE is a composite of pure error and the interaction of run and level; pure error could be separated from the interaction by performing replicates at each level; however, the within-run variability will still be reported as the composite of pure error and the level-by-run interaction).

Table 4 Analysis of Variance Table Showing Expected Mean Squares as Functions of Intrarun and Interrun Components of Variability

Effect	df	MS	Estimated mean square	F
Level	4	Mean square for levels (MSL)	$\hat{\sigma}_w^2 + Q(L)$	
Run	5	MSR	$\hat{\sigma}_w^2 + 5\hat{\sigma}_B^2$	MSR/MSE
Error	20	MSE	$\hat{\sigma}_w^2$	

The total variability associated with testing a sample in the assay can be calculated from these variance component estimates as:

$$\hat{\sigma}_{\text{Total}}^2 = \frac{\hat{\sigma}_B^2}{r} + \frac{\hat{\sigma}_w^2}{nr}$$

Note that this is the variance of $\bar{y}$, where r represents the number of independent runs performed on the sample and n represents the number of replicates within a run.

A confidence interval can be established on the total variability as follows:^[8]

$$\begin{aligned} \hat{\sigma}_{\text{Total}}^2 &= \frac{\hat{\sigma}_B^2}{r} + \frac{\hat{\sigma}_w^2}{nr} \\ &= \left(\frac{1}{Jr} \right) \cdot \text{MSR} + \left(\frac{n-J}{Jnr} \right) \cdot \text{MSE} \\ &= c_1 \cdot \text{MSR} + c_2 \cdot \text{MSE} \\ &= \sum_q c_q S_q \end{aligned}$$

where J represents the number of replicates of a sample tested in each run of the assay validation experiment (here $J = 5$ for the number of levels of the sample). The $1 - 2\alpha$ confidence interval on $\hat{\sigma}_{\text{Total}}^2$ becomes:

$$\left[\hat{\sigma}_{\text{Total}}^2 - \sqrt{\sum_q G_q^2 c_q^2 S_q^4}, \hat{\sigma}_{\text{Total}}^2 + \sqrt{\sum_q H_q^2 c_q^2 S_q^4} \right]$$

where

$$G_q = 1 - \frac{1}{F_{\alpha, df_q, \infty}}, H_q = \frac{1}{F_{1-\alpha, df_q, \infty}} - 1, F_{\alpha, df_q, \infty}, F_{1-\alpha, df_q, \infty}$$

represent the percentiles of the F -distribution.

The LOD and the LOQ are calculated using the standard deviation of the responses ($\hat{\sigma}$) and the slope of the calibration curve ($\hat{\beta}$) as $\text{LOD} = 3.3 \hat{\sigma}/\hat{\beta}$ (note, the 3.3 is derived from the 95th percentile of the standard normal distribution, 1.64; here, $2 \times 1.64 = 3.3$, thereby limiting both the false-positive and false-negative error rates to be equal to or less than 5%) and $\text{LOQ} = 10 \hat{\sigma}/\hat{\beta}$ (note, the 10 corresponds to the restriction of no more than 10% RSD in the measurement variability; $\frac{\text{RSD}}{100} = \frac{\hat{\sigma}}{\bar{x}} < 0.10$ gives $\bar{x}$; thus if the restriction was changed to “not more than 20% RSD,” then the LOQ could be calculated as $\text{LOQ} = 5 \hat{\sigma}/\hat{\beta}$).

These estimates do not account for the calibration curve, and they can be improved upon by acknowledging the variability in prediction from the curve.^[9] For a linear fit, the LOD is depicted graphically in Fig. 3. The LOQ can

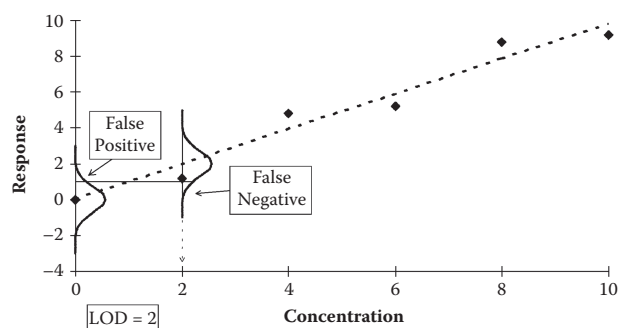


Fig. 3 Determination of the LOD using the linear fit of the standard curve plus restrictions on the proportions of false positives and false negatives

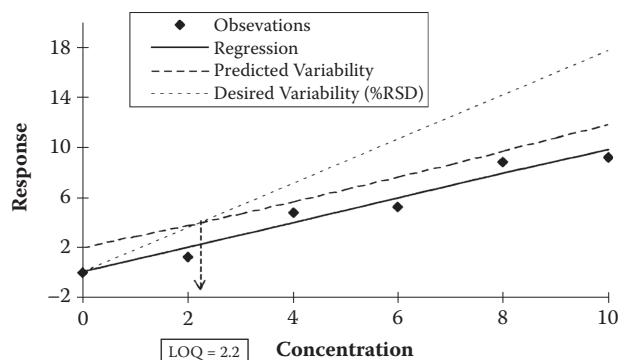


Fig. 4 Determination of the LOQ using the linear fit of the standard curve and a restriction on the variability (% RSD) in the assay

be calculated as prescribed in the ICH guideline, solving for x_Q in the following equation:

$$LOQ = 10 \left\{ \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_Q - \bar{x})^2}{SXX}} \right\}$$

The limit of quantitation is depicted graphically in [Fig. 4](#).

The LOQ is more complicated to calculate from a non-linear calibration function, such as the four-parameter logistic regression equation, and involves the first-order Taylor series expansion of the solution equation for x_Q ^[10] (see the entry on *Assay Development*). The calculation of the LOQ for this equation is illustrated in [Fig. 5](#).

FOLLOW-UP

The information from the analytical method validation can be utilized to establish a testing format for the samples in the assay. In particular, variance component estimates of the within-run and between-run variabilities can be used to predict the total variability associated with a variety of testing formats. For example, consider the case where component estimates have been determined to be equal to

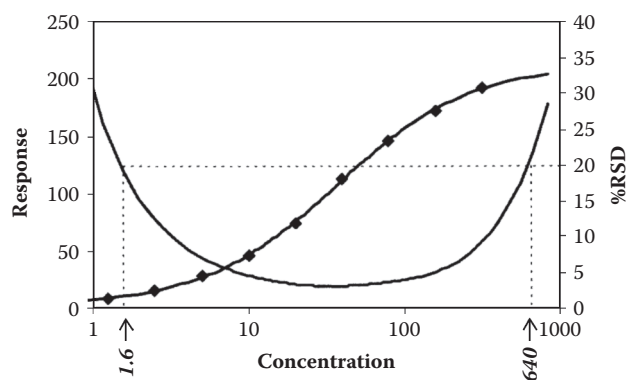


Fig. 5 Quantifiable range determined as limits within which assay variability (% RSD) is less than or equal to 20%

Table 5 Predicted Variability (% RSD) for a Given Number of Runs (r) and a Given Number of Replicates within Each Run (n)

r/n	1	2	3	∞
1	6%	5%	5%	5.0%
2	4%	4%	4%	3.5%
3	3%	3%	3%	2.9%

$\hat{\sigma}_B = 0.05$ (5%) and $\hat{\sigma}_W = 0.03$ (3%). Letting r represent the number of independent runs that will be performed on a sample, and n be the number of replicates within a run, the variabilities are listed in [Table 5](#). Suppose that, because of the process capability, the analytical variability must be restricted to $RSD \leq 3\%$. As can be observed from the table, this cannot be accomplished using a single run of the assay; it must therefore be achieved through a combination of multiple runs with multiple replicates within each run.

ANALYTICAL METHOD QUALITY CONTROL

Assay validation does not end with the formal validation experiment but is continuously assessed through assay quality control. Control measures of variability and bias are utilized to bridge to the assay validation and to warrant the continued reliability of the procedure.

Extra variability in replicates can be detected utilizing standard statistical process control metrics;^[4] detection and prespecified action upon detection can help maintain the variability characteristics of an assay. For example, when the measurement of an analyte is obtained from multiple runs of an assay, those results can be examined by noting the range (either “log(max) – log(min)” or “max – min,” depending upon whether the measurements are scaled proportionally or absolutely). A bound on the range can be derived:

$$\text{Range} \leq \hat{\sigma}_{\text{dimension}} (d_2 + kd_3)$$

where $\hat{\sigma}_{\text{dimension}}$ is the variability in the dimension of replication (here, run to run), d_2 and d_3 are the factors used to scale the standard deviation to a range, and k is a factor representing the desired degree of control (e.g., $k = 2$ represents approximately 95% confidence). Having detected “extra variability,” additional measurements can be obtained, and a result for the sample can be determined from the remainder of the measurements made after the maximum and minimum results have been eliminated from the calculation. When “extra variability” has been identified in instrument replicates of a standard concentration or a dilution of a test sample, that concentration of the standard or dilution of the test sample might be eliminated from subsequent calculations.

Measurements on control sample(s) help identify shifts in scale and therefore bias in a run of an assay. Information in multiple control samples can typically be reduced to one or two meaningful scores, such as an average and/or a slope. These can help isolate problems and provide information regarding the cause of a deviation in a run of an assay. Suppose, for example, that three controls of various levels are used to monitor shifts in sensitivity in the assay. The average can be used to detect absolute shifts, because of some underlying source of constant variability, such as inappropriate preparation of standards or the introduction of significant background. The slope can be used to detect proportional shifts as a result of some underlying source of proportional variability.

Suitability criteria on the performance of column peaks can detect degeneration of a column, while characteristics of the standard curve, such as the slope or EC_{50} (EC_{50} corresponds to C in a four-parameter logistic regression) help predict an assay with atypical or substandard performance.

Care must be taken, however, to avoid either meaningless or redundant controls. The increased false-positive rate inherent in the multiplicity of a control strategy must be considered, and a carefully contrived quality control scheme should acknowledge this.

CONCLUSION

A carefully conceived assay validation should follow assay development, and precede routine implementation of the method for product characterization. The purpose of the validation is to demonstrate that the assay is fundamentally reliable, and possesses operating characteristics that make it *suitable for use*. The statistical estimates collected during the validation can also be used to design a test plan, that provides the maximum reliability for the minimum effort in the laboratory. Most importantly, validation of the assay does not end with the documentation of the validation experiment, but should continue with assay

quality control. The proper choice of control samples and control parameters helps assure the continued validity of the procedure during routine use in the laboratory.

ACKNOWLEDGEMENT

Minor updates have been made by the Editor-in-Chief in order to reflect developments since publication of the *Encyclopedia of Biopharmaceutical Statistics, Third Edition*.

REFERENCES

1. USP/NF, United States Pharmacopodia 35 – National Formulary 30, United States Pharmacopeia Convention, Rockville, MD.
2. Federal Register. *International Conference on Harmonization, Guideline on Validation of Analytical Procedures: Definitions and Terminology*. March 1, 1995; 11259–11262.
3. *International Conference on Harmonization, Guideline on Validation of Analytical Procedures: Methodology*. November 6, 1996.
4. Montgomery, D. *Introduction to Statistical Quality Control*, 2nd; Wiley: New York, 1991, 203–206, 208–210.
5. Morrison, D. *Multivariate Statistical Methods*; McGraw-Hill: New York, 1967, 221–258.
6. Snedecor, G.; Cochran, W. *Statistical Methods*, 6th; Iowa State Univ. Press: Ames, IA, 1967, 91–100.
7. Winer, B. *Statistical Principles in Experimental Design*, 2nd; McGraw-Hill: New York, 1971, 244–251.
8. Burdick, R.; Graybill, F. *Confidence Intervals on Variance Components*; Marcel Dekker: New York, 1992, 28–39.
9. Oppenheimer, L.; Capizzi, T.; Weppelman, R.; Mehta, H. Determining the lowest limit of reliable assay measurement. *Anal. Chem* **1983**, 55, 638–643.
10. Morgan, B. *Analysis of Quantal Response Data*; Chapman & Hall: New York, 1992, 370–371.
11. Massart, D.L.; Dijkstra, A.; Kaufman, L. *Evaluation and Optimization of Laboratory Methods and Analytical Procedures*; Elsevier: New York, 1978.
12. Currie, L.A. Limits for qualitative detection and quantitative determination: Application to radiochemistry. *Anal. Chem* **1968**, 40, 586–593.
13. Cardone, M.J. Detection and determination of error in analytical methodology. Part I. The method verification program. *J. Assoc. Off. Anal. Chem* **1983**, 66, 1257–1281.
14. Cardone, M.J. Detection and determination of error in analytical methodology. Part II. Correction for corrigible systematic error in the course of real sample analysis. *J. Assoc. Off. Anal. Chem* **1983**, 66, 1283–1294.
15. Miller, J.C.; Miller, J.N. *Statistics for Analytical Chemistry*, 2nd; Wiley: New York, 1988.
16. Caulcutt, R.; Boddy, R. *Statistics for Analytical Chemists*; Chapman & Hall: New York, 1983.

Assay Validation: Evaluation of Linearity

Jen-pei Liu

Division of Biometry, Department of Agronomy, and Institute of Epidemiology, National Taiwan University, Taipei, Taiwan; and Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan

Eric Hsieh

Division of Biometry, Department of Agronomy, National Taiwan University, Taipei, Taiwan

Abstract

In validation of quantitative analytical laboratory procedures, one of the most important characteristics of the accuracy is the linearity. This entry introduces experimental design and statistical testing procedures for testing linearity based on specific criteria.

Keywords: Allowable limit; General pivotal quantities; Linearity; Quantitative analytical laboratory methods.

INTRODUCTION

In validation of quantitative analytical laboratory procedures, one of the most important characteristics of the accuracy is the linearity. The ICH Q2A guideline^[1] defines the linearity of an analytical method as its ability (within a given range) to obtain the test results, which are directly proportional to the concentration (amount) of the analyte in the test sample. The objective for evaluation of linearity is to validate existence of a mathematically verified straight-line relationship between the observed values and the true concentrations or activities of the analyte. Linearity represents the simplest mathematical relationship and it also permits simple and easy interpolations of results for clinical practitioners. The approved Clinical Laboratory Standard Institute (CLSI) guideline EP6-A^[2] recommends that at least five solutions of different concentration levels across the anticipated range be included in an experiment for evaluation of linearity. At each concentration level, two to four replicates should be run. According to the evaluation criterion suggested by EP6-A,^[2] if the difference between the best-fit non-linear polynomial curve and simple linear regression equation at each concentration is smaller than some pre-defined allowable bias δ_0 , the linearity then can be claimed. For instance, in Fig. 1, it shows that the linearity is claimed because the magnitude of deviation from linear regression at all concentrations for the best-fit model are less than δ_0 , while the linearity cannot be claimed in Fig. 2 since the magnitude of deviation from linear regression at concentration S2, S3 and S4 for the best-fit model are greater than δ_0 .

The evaluation criterion of EP6-A is a disaggregate criterion since it requires the difference between the best-fit non-linear polynomial curve and simple linear regression equation must be smaller than pre-defined allowable limit at each concentration. On the other hand, Hsieh et al.^[3] propose the sum of squares of deviation from linearity (SSDL) as an unscaled aggregate criterion for assessment of linearity. The linearity of an assay method can be claimed at α significance level if observed SSDL is smaller than the upper $100(1 - \alpha)\%$ generalized confidence limit of SSDL. In addition, two scaled aggregate criteria including the linearity average deviation from linearity (ADL) and the coefficient of variation of the deviation from linearity (CVDL) are also proposed by Kroll et al.^[4] and Liu et al.,^[6] respectively. The ADL is scaled by the mean concentration while CVDL is scaled by the square of the variability of the best-fit model. The disaggregate criterion is in general more conservative than the aggregate criterion. On the other hand, compared with the unscaled aggregate criterion, the scaled aggregate criterion, which is the measures expressed in the relative magnitude of bias based on the unitless scaled deviations from linearity, may provide a better and more meaningful interpretation for the accuracy of analytical procedures.

In this entry, we will introduce the experimental design suggested by EP6-A guideline^[2] for assessment of linearity firstly. The evaluation criteria with the corresponding statistical hypotheses will then be provided. In the latter part of the entry, we will review several statistical testing procedures for assessing of linearity based on the evaluation

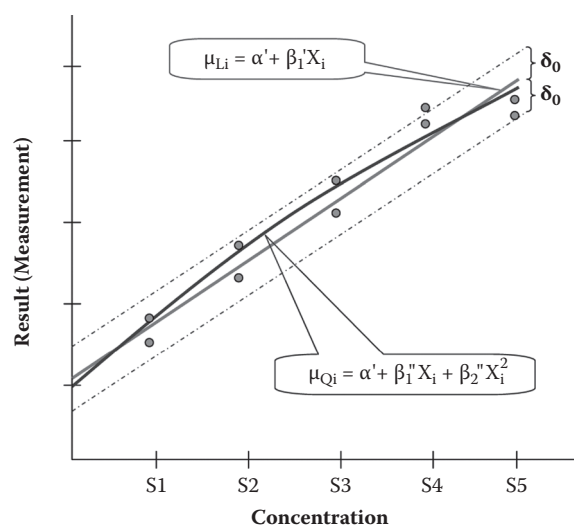


Fig. 1 Acceptance of linearity by CLSI EP6-A guideline

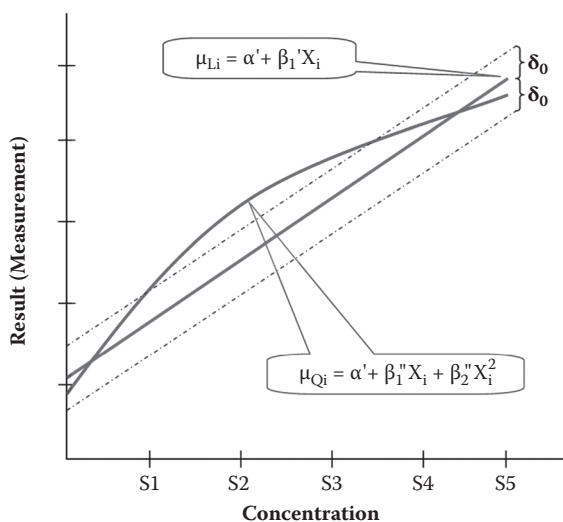


Fig. 2 Failure of linearity by CLSI EP6-A guideline

criteria introduced above. The simulation results for comparing the empirical size and powers of each method will be discussed. Two numerical examples will be used to discuss the applications of the introduced methods. Concluding remarks will be given at the end.

EXPERIMENT DESIGN

The approved CLSI guideline EP6-A^[2] recommends that the experiment for linearity assessment should be conducted at least five solutions of different concentrations run at least in duplicates. Let Y_{ij} be the test result of replicate j at concentration X_i , where $j = 1, \dots, j$; $i = 1, \dots, L$. The approved CLSI guideline EP6-A considers the following linear, quadratic, and cubic models fitting the data obtained from the experiment:

Linear (First order):

$$\mu_{Li} = \alpha' + \beta_1' X_i$$

Quadratic (Second-order polynomial):

$$\mu_{Qi} = \alpha'' + \beta_1'' X_i + \beta_2'' X_i^2 \quad (1)$$

or

Cubic (Third-order polynomial):

$$\mu_{Ci} = \alpha''' + \beta_1''' X_i + \beta_2''' X_i^2 + \beta_3''' X_i^3$$

where μ_{Li} , μ_{Qi} , and μ_{Ci} are the predicted mean of the corresponding models and α' , α'' , α''' ; β_1' , β_1'' , β_1''' ; β_2'' , β_2''' , and β_3''' , are the intercepts, regression coefficients for the corresponding models in Eq. 1. The best-fit model is the model such that lack-of-fit is not statistically significant and the repeatability meets the manufacturer's claim. In addition, for the purpose of illustration, the random error is assumed to be approximately constant rather proportional in the range of concentrations considered in the experiment.

The spirit of the EP6-A guideline^[2] is to determine the concentrations(s) where an assay method is not linear and the extent of the non-linearity at that level. As addressed in the guideline "The guideline emphasizes the necessity that each user establishes his or her requirements for linearity, or the allowable error due to nonlinearity. It also places less importance on global tests such as Lack-of-Fit test for linearity across the tested range. Global tests merely indicate that statistically significant non-linearity exists; they do not show where that non-linearity is, nor do they show the magnitude of the error." Therefore, even if the best-fit model is a non-linear model, it does not necessarily imply that the assay cannot be concluded linear. Based on the above concept, the guideline proposes the following rule in instead of Global test for assessing linearity: if the best-fit model is the linear model over the some range of concentrations employed in the experiment, then the assay method can be concluded to be linear over the some range of concentrations. However, if the best-fit model is not linear, the linearity of the analytical procedure can still be claimed if the magnitude of deviations from the linearity at each concentration is within some pre-specified allowable limit of δ_0 as showed in Fig. 1.

EVALUATION CRITERIA

EP6-A's Criterion

The EP6-A defines the best-fit model is the one such that the lack-of-fit of the model is not statistically significant and repeatability meets the manufacturer's claim. The deviation from linearity at concentration i is defined as the difference between the predicted mean of the best-fit model and that of the linear model at concentration i .

The linearity of an analytical procedure can be claimed if the one of following is conditions are satisfied:

- If the best-fit model is the linear model over the same range of concentrations employed in the experiment.
- If the best-fit model is not linear, the magnitude of deviations from the linearity at each concentration is within some pre-specified allowable limit.

Let μ_{pi} denote the predicted mean of the best-fit model, where μ_{pi} can be either μ_{Qi} or μ_{Ci} . The deviation from linearity at each concentration level is defined as the difference in the predicted means between the best-fit nonlinear model and linear model $\mu_{pi} - \mu_{Li}$. Let δ_0 be some pre-specified allowable margin. The linearity of an assay can be concluded by EP6-A's criterion if

$$|\mu_{pi} - \mu_{Li}| < \delta_0, \quad i = 1, \dots, L. \quad (2)$$

The corresponding statistical hypothesis can be expressed as

$$H_{0i} : |\mu_{pi} - \mu_{Li}| \geq \delta_0 \text{ vs. } H_{ai} : |\mu_{pi} - \mu_{Li}| < \delta_0, \quad \text{for all } i = 1, \dots, L. \quad (3)$$

Since the hypothesis in Eq. (3) requires all differences in the predicted means between the best-fitting and linear models be within the pre-specified allowable limit, it is a disaggregate criterion.

Sum of Squares of Deviation from Linearity (SSDL)

Considering the evaluation proposed by EP6-A guideline^[2] in Eq. 2, a natural summary measure for evaluation of linearity is the sum of squares of deviation (SSDL) from linearity:^[3]

$$\tau = \text{SSDL} = \sum_{i=1}^L (\mu_{pi} - \mu_{Li})^2. \quad (4)$$

It follows that the hypotheses for proving the assay linearity can be formulated based on SSDL as follows:

$$H_0 : \sum_{i=1}^L (\mu_{pi} - \mu_{Li})^2 \geq L\delta_0^2 \text{ vs. } H_a : \sum_{i=1}^L (\mu_{pi} - \mu_{Li})^2 < L\delta_0^2 \quad (5)$$

or equivalently

$$H_0 : \sum_{i=1}^L (\mu_{pi} - \mu_{Li})^2 / L \geq \delta_0^2 \text{ vs. } H_a : \sum_{i=1}^L (\mu_{pi} - \mu_{Li})^2 / L < \delta_0^2$$

The SSDL is an aggregate criterion since it is the sum of squares of the differences in the predicted means between the best-fit and linear models

Average Deviation from Linearity (ADL)

The SSDL is the sum of squares of unscaled deviations from linearity and its unit is the square of the unit of concentration of analyte. Sometimes, the measures expressed in the relative magnitude of bias based on the unitless scaled deviations from linearity may provide a better and more meaningful interpretation for the accuracy of analytical procedures. The first type of the scaled deviations is the deviations scaled by μ , the mean concentration for all solutions of the assay:

$$(\mu_{pi} - \mu_{Li}) / \mu$$

Kroll et al.^[4] considers the average deviation from linearity (ADL) for assessment of linearity. The ADL is defined as

$$\theta = \text{ADL} = \frac{\sqrt{\sum_{i=1}^L (\mu_{pi} - \mu_{Li})^2 / L}}{\mu}, \quad (6)$$

where μ is the population mean concentration for all solutions of the assay.

The ADL is a scaled deviation defined as the square root of sum of squares of the difference in predicted means between the best-fit and linear models. The hypothesis for evaluation of linearity based on ADL proposed by Hsieh and Liu^[5] is given as

$$H_0 : \theta \leq \theta_0 \text{ vs. } H_a : \theta > \theta_0, \quad (7)$$

where θ_0 is the maximum allowable average deviation from linearity. Because ADL is a function of standardized sum of squares of the differences in the predicted means between the best-fit and linear models, it is also an aggregate criterion.

Coefficient of Variation of the Deviations from Linearity (CVDL)

The second type of the scaled deviations is the deviations scaled by σ , the variability or repeatability of the best-fit model:

$$(\mu_{pi} - \mu_{Li}) / \sigma$$

The CVDL^[6] for assessment of linearity is the square root of the average sum of squares of the scaled deviations by σ :

$$\eta = \text{CVDL} = \frac{\sqrt{\sum_{i=1}^L (\mu_{pi} - \mu_{Li})^2 / L}}{\sigma}.$$

The hypotheses for evaluation of linearity is given for CVDL as

$$H_0: \eta \geq \eta_0 \text{ vs. } H_a: \eta < \eta_0. \quad (8)$$

where η_0 is the allowable limit of CVDL. The CVDL is also an aggregate criterion. Furthermore, the relationship among τ , θ , and η can be given as

$$\tau = L(\theta\mu)^2 = L(\eta\sigma)^2 \quad (9)$$

STATISTICAL TESTING PROCEDURES

Two One-sided Tests (TOST) Procedures

Let $\hat{Y}_{Pi}$ and $\hat{Y}_{Li}$ denote the least squared (LS) estimators of the predicted mean of the best-fit and linear models, respectively, where

$$\hat{Y}_{Li} = a' + b'_2 X_i, \text{ and}$$

$$\hat{Y}_{Pi} = \begin{cases} a'' + b''_1 X_i + b''_2 X_i^2, & \text{if the best-fit model is quadratic,} \\ a''' + b'''_1 X_i + b'''_2 X_i^2 + b'''_3 X_i^3, & \text{if the best-fit model is cubic,} \end{cases}$$

and a' , a'' , a''' ; b'_1 , b'_2 , b'_3 ; b''_1 , b''_2 , b''_3 ; b'''_1 , b'''_2 , b'''_3 are the LS estimators of the intercepts, regression coefficients for the corresponding models in Eq. (1).

The estimation method suggested by the EP6-A guideline^[2] conclude the linearity of the proposed analytical method concluded if

$$|\hat{Y}_{Pi} - \hat{Y}_{Li}| < \delta_0, \quad \text{for } i = 1, \dots, L.$$

However, the estimation method may inflate the type I error rate in assessment of linearity because it completely ignores the variability and distribution associated with the estimators. The linearity of an assay is claimed at the α significance level if the $(1-2\alpha)100\%$ confidence interval for $\alpha_p - \alpha_L$ given by

$$(\hat{Y}_{Pi} - \hat{Y}_{Li}) \pm t_{\alpha, LJ-d-1} \hat{\sigma}_{di}, \quad (10)$$

is within $(-\delta_0, \delta_0)$, for $i=1, \dots, I''$.

Estimators of SSDL, ADL, and CVDL

Define

$\mathbf{Y}$ = the $LJ \times 1$ vector of observations,

$\mathbf{X}_L = (\mathbf{1}, \mathbf{X})$.

$\mathbf{X}_P = \begin{cases} (\mathbf{1}, \mathbf{X}, \mathbf{X}_2), & \text{if the best-fit model is quadratic,} \\ (\mathbf{1}, \mathbf{X}, \mathbf{X}_2, \mathbf{X}_3), & \text{if the best-fit model is cubic, and} \end{cases}$

$\Delta = \mu_P - \mu_L$,

where $\mu_P = LJ \times 1$ predicted mean vector of best-fit polynomial model, μ_L = the $LJ \times 1$ predicted mean vector of linear model, and $\mathbf{1}$ is $LJ \times 1$ vector of 1s $\mathbf{X} = (\mathbf{X}_i)$, $\mathbf{X}_2 = (\mathbf{X}_i^2)$, and $\mathbf{X}_3 = (\mathbf{X}_i^3)$.

Then a least squares unbiased estimator for the vector of the deviations from linearity Δ , is given as

$$\hat{\Delta} = \mathbf{H}\mathbf{Y},$$

where $\mathbf{H} = (\mathbf{H}_P - \mathbf{H}_L)$, $\mathbf{H}_P$ and $\mathbf{H}_L$ be the projection matrices corresponding to the column spaces spanned by the design matrices of the best-fit and linear models, respectively, that is, $\mathbf{H}_P = \mathbf{X}_P(\mathbf{X}'_P \mathbf{X}_P)^{-1} \mathbf{X}'_P$ and $\mathbf{H}_L = \mathbf{X}_L(\mathbf{X}'_L \mathbf{X}_L)^{-1} \mathbf{X}'_L$.

It follows that $\hat{\Delta}$ is distributed as a multivariate normal distribution with mean vector Δ and covariance matrix $\mathbf{H}\mathbf{H}'\sigma^2$.

An estimator of SSD, ADL, and CVDL can be expressed by the least squares unbiased estimator for Δ as follows:

$$\hat{\tau} = (\hat{\Delta}'\hat{\Delta})/J.$$

$$\hat{\theta} = \frac{\sqrt{\sum_{i=1}^L (\hat{Y}_{Pi} - \hat{Y}_{Li})^2 / L}}{\bar{Y}} = \frac{\sqrt{\hat{\tau}/L}}{(\bar{Y})/LJ}, \quad (11)$$

$$\hat{\eta} = \frac{\sqrt{\sum_{i=1}^L (\hat{Y}_{Pi} - \hat{Y}_{Li})^2 / L}}{s} = \sqrt{\frac{\hat{\tau}/L}{[\mathbf{Y}'(\mathbf{I} - \mathbf{H}_P)\mathbf{Y}]/(LJ - d - 1)}},$$

Where $\hat{Y}$ is the observed mean concentration for all solutions of the assay, s is the square root of the residual mean square obtained from the best-fit model with degrees of freedom of $LJ-d-1$, and d is the degrees of freedom for regression of the best-fit model, respectively.

Generalized Pivotal Quantity Approach

Generalized Pivotal Quantity for SSDL, ADL, and CVDL

Suppose that U is a random variable whose distribution depends on a vector of unknown parameters $\xi = (\theta, \lambda)$, where θ is a parameter of interest and λ is a vector of nuisance parameter. Let $\mathbf{U}$ be a random sample from U and $\mathbf{u}$ be the observed value of $\mathbf{U}$. Also let $\mathbf{R} = \mathbf{R}(\mathbf{U}; \mathbf{u}, \xi)$ be a function of $\mathbf{U}$, $\mathbf{u}$, and ξ . The random quantity $\mathbf{R}$ is said to be a GPQ if satisfies the following two conditions:

- $\mathbf{R}$ has a probability distribution that is free of unknown parameter.
- The observed pivotal of R , say $r = \mathbf{R}(\mathbf{U}; \mathbf{u}, \xi)$, does not depend upon nuisance parameters λ . In other words, r is only a function of $\mathbf{u}$ and θ .

In particular, the GPQ is called the fiducial generalized pivotal quantity (FGPQ) when the observed quantity $r = \theta$. It has been shown that the generalized confidence interval (GCI) based on FGPQ are proven to have asymptotically correct frequent coverage probability.^[7] Therefore, an upper $100(1 - \alpha)$ th percentile GCI for θ is given by $R_{1-\alpha}$, where $R_{1-\alpha}$ are the $100(1 - \alpha)$ th percentile of the distribution of $\mathbf{R}$. The percentiles of $\mathbf{R}$ can be estimated using Monte Carlo algorithms.

A GPQ for SSDL^[3] is given as

$$R_\tau = \frac{1}{J} (\mathbf{R}_\Delta)' (\mathbf{R}_\Delta), \quad (12)$$

where $\mathbf{R}_\Delta = \mathbf{H}\mathbf{y} - [\mathbf{R}_{\sigma^2} \mathbf{H}\mathbf{H}']^{1/2} \mathbf{Z}$, a pivotal quantity for Δ , $R_{\sigma^2} = \frac{(\mathbf{L}\mathbf{J} - d - 1)S^2}{V}$, a pivotal quantity for σ^2 ,

$$\mathbf{Z} = \sum^{-1/2} [\hat{\Delta} - \Delta], V = \frac{(\mathbf{L}\mathbf{J} - d - L)S^2}{\sigma^2}, \sum = \sigma^2 \mathbf{H}\mathbf{H}',$$

where matrix $\Lambda^{1/2}$ denotes the positive definite square root of a positive definite matrix Λ and $\Lambda^{-1/2} = (\Lambda^{1/2})^{-1}$.

In addition, $\hat{\Lambda}$ can be expressed by the projection matrices as

$$\begin{aligned} \hat{\Lambda} &= \mathbf{H}\mathbf{Y} \\ &= (\mathbf{H}_p - \mathbf{H}_L)\mathbf{Y} \\ &= [(\mathbf{I} - \mathbf{H}_L) - (\mathbf{I} - \mathbf{H}_p)]\mathbf{Y}. \end{aligned}$$

Because $\bar{\mathbf{Y}} = (\mathbf{I}'\mathbf{Y})/LJ$ and $\mathbf{I}'[(\mathbf{I} - \mathbf{H}_L) - (\mathbf{I} - \mathbf{H}_p)] = 0$, $\bar{\mathbf{Y}}$ and $\hat{\Delta}$ are also independent. It is straightforward to verify that the estimators $\hat{\Delta}$, S^2 and $\mathbf{Y}$ are associated with pivotal quantities $\mathbf{Z}$, V , and Z_μ , which are independent with the following distributions:

$$\begin{aligned} \mathbf{Z} &= \sum^{-1/2} [\hat{\Delta} - \Delta] \sim \mathbf{N}_{LJ}(\mathbf{0}, \mathbf{I}), \\ V &= \frac{(\mathbf{L}\mathbf{J} - d - 1)S^2}{\sigma^2} \sim \chi_{LJ-d-1}^2, \end{aligned}$$

and

$$Z_\mu = \frac{\bar{\mathbf{Y}} - \mu}{\sqrt{\sigma^2/LJ}} \sim N(0, 1), \quad (13)$$

where $\mathbf{N}_{LJ}(\mathbf{0}, \mathbf{I})$, χ_{LJ-d-1}^2 and $N(0, 1)$ denote the multivariate standard normal distribution with $LJ \times 1$ random vector, the chi-square random variable with $LJ - d - 1$ degrees and the univariate standard normal distribution, respectively.

It follows that a GPQ for μ is given as

$$\begin{aligned} R_\mu &= \bar{\mathbf{y}} - \frac{1}{\sqrt{LJ}} \sqrt{R_{\sigma^2} Z_\mu} \\ &= \bar{\mathbf{y}} - \frac{1}{\sqrt{LJ}} \sqrt{\frac{S^2 \sigma^2}{S^2}} \frac{(\bar{\mathbf{Y}} - \mu)}{\sqrt{\sigma^2/LJ}}, \end{aligned} \quad (14)$$

where R_{σ^2} is the generalized pivotal quantity for σ^2 defined in Eq. (12).

From Eqs. (12–14), the distributions of $\mathbf{R}_\Delta$, R_{σ^2} , and R_μ are free of their respective nuisance parameters. In addition, it is readily to verify when $\hat{\Delta}$, S^2 and $\bar{\mathbf{Y}}$ are substituted by their observed values, then $(\mathbf{R}_\Delta, R_{\sigma^2}, \text{ and } R_\mu) = (\Delta, \sigma^2, \mu)$. Hence, they fulfill the requirement (a) and (b) of GPQ's given above.

As a result, generalized pivotal quantities for ADL and CVDL^[6] can be obtained respectively as

$$R_\theta = \frac{\sqrt{R_\tau/L}}{R_\mu}, \quad (15)$$

and

$$R_\eta = \frac{\sqrt{R_\tau/L}}{R_{\sigma^2}}, \quad (16)$$

where R_τ , R_μ , and R_{σ^2} are defined in Eqs. 12 and 14.

Generalized Pivotal Quantity for SSDL, ADL, and CVDL

An upper $100(1 - \alpha)$ th percentile GCI for SSDL, ADL, and CVDL can be obtained from the following Monte-Carlo algorithm:

- Step 1:** Choose a large simulation sample size, say $K = 10,000$. For k equal to 1 through K , carry out the following two steps.
- Step 2:** Independently generate $LJ \times 1$ standard normal random vector $\mathbf{Z}$, univariate standard normal variable Z_μ , and central chi-square random variables V with degree of freedom $LJ - d - 1$.
- Step 3:** For the realized values of $\mathbf{Y}$ and S^2 , compute R_τ , R_θ , and R_η as defined in Eqs. (12), (15), and (16), respectively.

The required upper $100(1 - \alpha)$ th percentiles of the distribution of GPQ for SSDL (or ADL or CVDL), which is also the upper $100(1 - \alpha)$ th generalized confidence limit for $\theta(\eta)$, is then estimated by the $100(1 - \alpha)$ th sample percentiles of the collection of $K = 10,000$ realizations $R_{\tau,1}$ (or $R_{\theta,1}$ or $R_{\eta,1}$), $R_{\tau,2}$ (or $R_{\theta,2}$ or $R_{\eta,2}$),, $R_{\tau,10000}$ (or $R_{\theta,10000}$ or $R_{\eta,10000}$).

Statistical Testing Procedure

The upper $100(1 - \alpha)\%$ generalized confidence limits for SSDL or ADL or CVDL based on GPQ can be used to test their respective statistical hypotheses in Eqs. (5) or (7) or (8) for linearity. The linearity of an analytical method is concluded at the α significance level if the upper $100(1 - \alpha)\%$ generalized confidence limit for SSDL or ADL or CVDL is less than $L\delta_0$ or θ_0 or η_0 .

SIMULATION STUDY

The simulation studies are performed to compare the empirical sizes between disaggregate-criterion based estimation method and TOST procedures, and which among three aggregate-criterion based GPQ-SSDL, GPQ-ADL, and GPQ-CVDL methods under the various combination of parameters of number of replicates, number of concentration level and standard deviation of normal random error of the best-fit model. For each of combinations, 5,000 random samples are generated for comparisons of estimation and TOST methods, whereas 10,000 random samples are generated for comparison of three GPQ methods. Therefore, for the 5% nominal significance level, a simulation study with 5,000 random samples implies that 95% of the empirical sizes evaluated at the equivalence limits will be within 0.04396 and 0.05604 if the proposed methods can adequately control the size at the nominal level of 0.05. On the other hand, the size at the nominal level of 0.05 can be adequately controlled if the empirical sizes are within 0.0457 and 0.0543 for ten thousand (10,000) random samples. As presented in Table 1, only 8.33% of the empirical sizes (1/12) of TOST method are not included in (0.04395, 0.05604), while all of the empirical sizes of the estimation method are between 0.4930 and 0.5066. Recall that the estimation method suggested in the approved CLSI guideline EP6-A^[2] ignores the variation of the estimates of $\mu_{pi} - \mu_{ci}$. Under the normal assumption, the size should be equal to 0.5 as confirmed by the empirical sizes of the simulation. With respect to the comparison of GPQ-SSDL, GPQ-ADL, and GPQ-CVDL methods, all of empirical sizes of the GPQ methods based on SSDL, ADL, and CVDL presented in Table 2 are within the range between 0.0457 and 0.0543. The simulation results reveal that the

Table 1 Results of Empirical Sizes (Uncorrected Kroll vs. Corrected Kroll and Estimation Method vs. TOST)

Type of comparisons	No. of sol.	No. of rep.	SD	EP6-A	TOST
Estimation vs. TOST.	5	2	0.1	0.4984	0.0522
			0.2	0.5050	0.0564
		3	0.1	0.4930	0.0440
			0.2	0.5050	0.0486
		4	0.1	0.5048	0.0490
			0.2	0.5066	0.0560
	7	2	0.1	0.4972	0.0512
			0.2	0.5024	0.0484
		3	0.1	0.4978	0.0478
			0.2	0.4946	0.0504
		4	0.1	0.5044	0.0504
			0.2	0.5066	0.0494

Sol.: Solution; Rep.: Replication

Table 2 Empirical Sizes (GPQ-based SSDL vs. GPQ-based ADL vs. GPQ-based CVDL Methods)

No. of solutions	No. of replicates	Standard deviation	GPQ-based SSDL	GPQ-based ADL	GPQ-based CVDL
5	2	0.1	0.0462	0.0467	0.0540
		0.2	0.0523	0.0517	0.0467
	3	0.1	0.0499	0.0502	0.0513
		0.2	0.0522	0.0517	0.0511
	4	0.1	0.0498	0.0505	0.0489
		0.2	0.0504	0.0508	0.0509
7	2	0.1	0.0504	0.0501	0.0490
		0.2	0.0495	0.0494	0.0504
	3	0.1	0.0505	0.0509	0.0473
		0.2	0.0495	0.0498	0.0504
	4	0.1	0.0498	0.0498	0.0529
		0.2	0.0498	0.0510	0.0452

GPQ-based methods for SSDL, ADL, and CVDL can adequately control the size at the nominal level. Fig. 3 presents the empirical powers of the three aggregate-criterion based method. The figure demonstrates that the GPQ-ADL procedure is uniformly more powerful than the GPQ-SSDL and GPQ-CVDL methods.

NUMERICAL EXAMPLES

Estimation Method Vs. TOST Method

We consider an experiment for evaluation of the linearity of a new analytical procedure for determination of β -HCG (β -Human Chorionic Gonadotropic, mIU/mL). The design consists of five dilutions with two replicates at each dilution of concentrations. Table 3 presents a set of hypothetical measurements under the design described above. For the purpose of the illustration, the allowable limit of $|\mu_{pi} - \mu_{Li}|$ is set as 0.4 for the estimation method suggested in the approved CLSI guideline EP6-A, and for the TOST procedure. The cubic model is the best-fit model recommended by the approved CLSI guideline EP6-A for the data in Table 3.

Table 4 presents the results of linearity of the two methods. With respect to the estimation method, the observed differences in the predicted means between the cubic and linear regression models at all dilutions are within the allowable margin of ± 0.4 . As a result, the linearity is claimed by the estimation method. On the other hand, the results of the TOST procedure show that the 95% confidence intervals for $\mu_{pi} - \mu_{ci}$ at the first two dilutions are not contained within $(-0.4, 0.4)$. With respect to hypotheses in Eq. 10, the analytical method cannot be concluded linear at the 5% significance level. Because the estimation

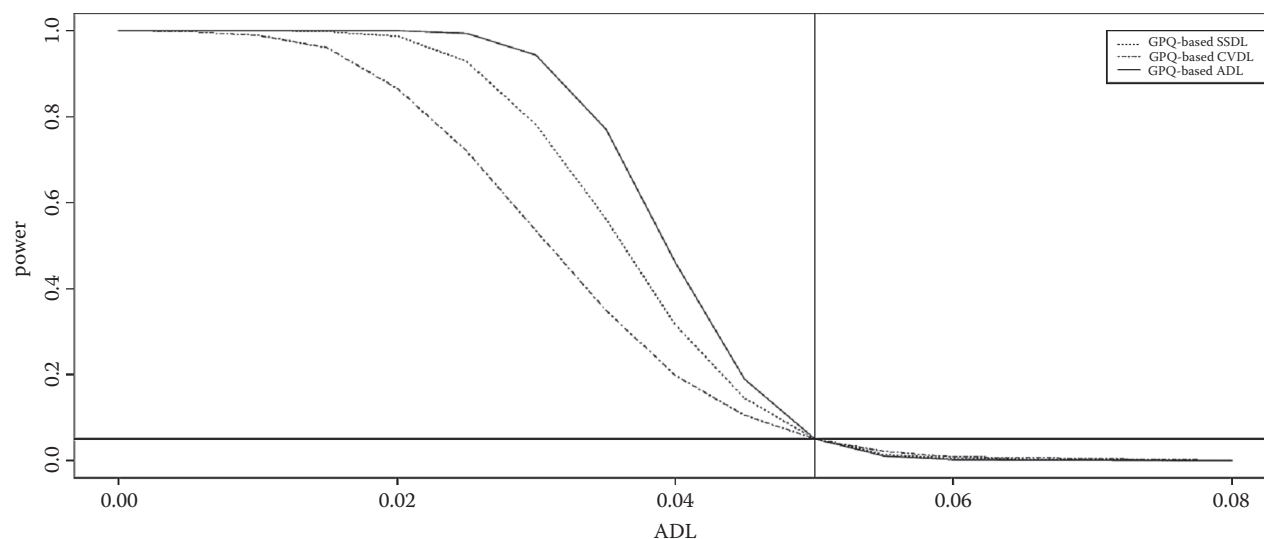


Fig. 3 The empirical powers when standard deviation of normal random error is 0.1, number of solutions is 5, and number of replicates is 3 (GPQ-based SSDL vs. GPQ-based CVDL vs. GPQ-based ADL methods)

Table 3 Measurement of β -HCG

Dilution	Replicate 1	Replicate 2
1	1.00	0.99
2	1.60	1.59
3	2.50	2.60
4	4.36	4.39
5	5.10	5.00

method completely ignores the variability in the observed differences in the predicted means, its conclusion is made without any statement of the probability of type I error. However, in fact, as demonstrated by the simulation, the probability of type I error of the estimation method far exceeds its nominal significance level.

GPQ-SSDL Method versus GPQ-ADL Method versus GPQ-CVDL Method

The first five concentrations given in Example 2 of CLSI guideline EP6-A^[2] illustrate the proposed testing procedures in evaluation of linearity of an analytical procedure.

Following EP6-A,^[2] the criterion of $|\mu_{p_i} - \mu_{L_i}|$ for linearity is set as 0.2 mg/dL for all five concentrations. In this example, the allowable margin of percent bound for ADL is set as 0.05 for all methods based on ADL as suggested by Kroll et al.^[4] On the other hand, the allowable limit of the GPQ-based SSDL is set as 0.2 which is calculated by square of 0.2 mg/dL multiplying five concentrations. We also assume that the allowable repeatability set by the manufacturer is 0.2. Therefore, the allowable margin of the GPQ-based CVDL is 1, which is equal to the square root of the value computed by $\tau = L(\eta\sigma)^2$. The quadratic model is the best-fit model among the three models recommended by the approved CLSI guideline EP6-A^[2] for the data in Table 5.

The results of the three GPQ-based procedures are provided in Table 6. The 95% upper confidence limit for the ADL computed by the GPQ method is 0.0218 which is smaller than the allowable upper limit of 0.05. Hence, the linearity of the analytical procedure can be concluded at the 5% significance level by the GPQ-based ADL procedure. On the other hand, the 95% upper confidence limits for SSDL of the GPQ-based SSDL and CVDL methods are 0.2471 and 1.9125, respectively. Both 95% upper

Table 4 Results of the Linearity by Four Different Methods of β -HCG

Dil (i)	TOST		Estimation method			
	90% C.I. of $\mu_{p_i} - \mu_{L_i}$	Result		$\hat{Y}_{p_i} - \hat{Y}_{L_i}$	Result	
		Ind.	Overall		Ind.	Overall
1	(0.064, 0.529)	NL	NL	0.296	L	L
2	(-0.524, -0.223)	NL		-0.374	L	
3	(-0.318, 0.027)	L		-0.146	L	
4	(0.078, 0.379)	L		0.228	L	
5	(-0.237, 0.228)	L		-0.005	L	

Table 5 Measurement of Calcium

Dilution	Replicate 1	Replicate 2
1	4.7	4.6
2	7.8	7.6
3	10.4	10.2
4	13.0	13.1
5	15.5	15.3

Source: The approved CLSI guideline. (From Ref. [2]).

Table 6 Results of the Linearity Evaluation by Three Different Methods

Method	Sample statistic/critical value or allowable bound		Conclusion
GPQ-based SSDL	Upper 95% C.L.	0.2471	Nonlinear
	Allowable upper bound	0.2	
GPQ-based CVDL	Upper 95% C.L.	1.9125	Nonlinear
	Allowable upper bound	1	
GPQ-based ADL	Upper 95% C.L.	0.0218	Linear
	Allowable upper bound	0.05	

95% C.L.: Upper 95% Confidence limit.

confidence limits are larger than their respective allowable upper limits of 0.2 and 1. As a result, both methods can not conclude the linearity of the analytical procedure at the 5% significance level. The results presented above show the different conclusions between the GPQ-based methods. As shown in simulation results, all three GPQ-based methods can control the size at the nominal size of 0.05, the GPQ-based ADL method is uniformly more powerful than the other two GPQ-based methods. This might be one the reasons why the linearity can be claimed by the GPQ-based ADL method.

CONCLUDING REMARKS

One of the most important characteristics for evaluation of accuracy and precision in assay validation is linearity. Even though the best-fit model is not linear, linearity of the analytical procedure can still be claimed if the difference in the predicted means between the best-fit and linear models is smaller than some pre-specified allowable limit at all concentrations employed in the validation experiment. As a result, the deviation from linearity is the fundamental unit for assessment of bias for evaluation of linearity.

In this entry, we introduced the TOST procedure for the disaggregate criterion as well as the GPQ-based

procedures for the difference aggregate criteria. All of the proposed procedures show the good performance in controlling the size and provide sufficient power for assessment of linearity in assay validation. In addition, the evaluation procedure based on the disaggregate criterion is more conservative than which based on the aggregate criterion because it requires that the difference in predicted mean between the best-fit model and linear models for all solutions be within the pre-specified limit, while the aggregate criterion only requires the magnitude of sum of deviations from linearity be controlled within a aggregate limit. The choice of the disaggregate-criterion based procedure and aggregate-criterion based procedure may depend on how accuracy the assay method is required. In addition, although the GPQ-based ADL procedure is recommended to be used for assessment of linearity in assay validation because it is the uniformly more powerful than the other two GPQ-based methods. However, one may consider using CVDL as the criterion for assay validation if he or she would like to evaluate accuracy and reliability simultaneously.

ACKNOWLEDGMENT

The authors would like to thank the editor of EBS, Dr. Shein-Chung Chow, for inviting us to contribute this entry.

REFERENCES

1. ICH Expert Working Group. International Conference on Harmonization Tripartite Guideline Q2A: Test on Validation of Analytical Procedures, 1995.
2. Tholen, D.W.; Kroll, M.; Astles, J.R.; Caffo, A.L.; Happe, T.M.; Krouwer, J.; Lasky, F. *EP6A: Evaluation of the Linearity of Quantitative Measurement Procedures: A Statistical Approach; Approved Guideline*; Clinical Laboratory Standard Institute: Wayne, PA, 2003.
3. Hsieh, E.; Hsiao, C.F.; Liu, J.P. Statistical methods for evaluating the linearity in assay validation. *J. Chemometrics* **2009**, *25*, 56–83.
4. Kroll, M.H.; Præstgaard, J.; Michaliszyn, E.; Styer, P.E. Evaluation of the extent of Nonlinearity in reportable range studies. *Arch. Pathol. Lab. Med.* **2000**, *124*, 1331–1338.
5. Hsieh, E.; Liu, J.P. On statistical evaluation of linearity in assay validation. *J. Biopharm. Stat.* **2008**, *18*, 677–690.
6. Hsieh, E.; Hsiao, C.F.; Liu, J.P. Deviations from linearity in statistical evaluation of linearity in assay validation. *J. Chemometrics* **2009**, *25*, 489–494.
7. Hanning, J.; Iyer, H.K.; Patterson, P. Fiducial generalized confidence intervals. *J. Am. Stat. Assoc.* **2006**, *101*, 254–269.

Bayesian Designs: Phase II Oncology Clinical Trials

Sze Huey Tan

*Division of Clinical Trials & Epidemiological Sciences, National Cancer Centre, Singapore;
Centre of Quantitative Biology and Medicine, Duke-NUS Graduate Medical School,
Singapore*

David Machin

Children's Cancer and Leukaemia Group, University of Leicester, Leicester, U.K.

Say Beng Tan

*Singapore Clinical Research Institute, Centre of Quantitative Biology and Medicine,
Duke-NUS Graduate Medical School, Singapore*

Abstract

In this entry, we discuss two variants of a Bayesian two-stage design for clinical trials. Through a specific example, the entry also discusses how these designs can be used in practice.

Keywords: Bayesian; Dual threshold design; Oncology; Phase II clinical trial; Sample size; Single threshold design.

INTRODUCTION

Many different designs have been proposed and used for phase II oncology clinical trials. The majority of these are based on frequentist statistical approaches. Bayesian methods provide an alternative to frequentist approaches, allowing for the incorporation of relevant prior information and the presentation of the trial results in a manner which some feel is more intuitive and helpful. In this article, we discuss two variants of a Bayesian two-stage design for such clinical trials.^[1] The designs determine the overall sample size on the basis of achieving a minimum posterior probability of exceeding the pre-specified target response rate should the trial data really suggest a positive outcome.

We give sample size tables and cutoff response rates for decisions for these two-stage designs. We also discuss how the designs can be used in practice, using an example of an actual phase II trial involving the treatment of metastatic nasopharyngeal cancer.^[2] We also discuss adaptations to these designs that have been proposed by Mayo and Gajewski^[3,4] and Sambucini.^[5]

PHASE II CLINICAL TRIAL DESIGNS

The process of testing a new therapy for the treatment of a disease involves many stages. In oncology, the typical

process involves a new drug first undergoing extensive development and testing in the laboratory. Once this is done, and the drug is felt to be ready for testing in human subjects, a phase I trial is conducted. The purpose of a phase I trial is to study the drug's toxicity in humans and to identify the "best" dose to be used (in diseases like cancer, this is typically the highest dose which does not result in excessive toxicity). Following this, a phase II trial using this "best" dose is then conducted. The goal of such a trial is to assess the effect of the drug on the disease in question to determine if it should be further tested in a large-scale randomized phase III trial, to determine its efficacy. If such a decision is made, a phase III trial is then conducted to directly compare the new drug with other compounds being used to treat the condition. Phase II clinical trials thus play an important role in the development and testing of a new therapy. Here, by "new," we refer not just to totally new compounds but also to established compounds that are to be tested within a new multi-drug combination therapy for the disease in question.

Phase II clinical trial designs can be single-stage or multi-stage, single-arm or randomized multi-arm. Single-stage designs include the Fleming^[6] and A'Hern^[7] designs. As for two-stage single-arm designs, those that are commonly used include that of Gehan^[8] and the Simon optimal and minimax designs.^[9] Other designs include those by Sposto and Gaynon,^[10] which take into account patient heterogeneity in the design, and an adaptive two-stage design by Lin and Shih,^[11] which allows investigators who are uncertain about the target outcome proportion to specify two-target outcome proportions. The decision whether to proceed

Notes: Some of this article has been adapted from: Tan, S.B.; Machin, D. Bayesian two-stage designs for phase II clinical trials. *Stat. Med.* 2002, 21 (14), 1991–2012.

to Stage 2 of the trial and the sample size for stage two will depend on the effect size experienced at stage one. Trial designs for randomized multi-arm designs have also been developed and utilized, including the randomized selection design proposed by Simon et al.^[12] for binary endpoints and by Liu et al.^[13] for time-to-event endpoints.

There has also been increasing interest in the use of time-to-event endpoints, such as progression-free survival, in the design of phase II oncology clinical trials. In part, this is driven by the large number of regimens that have shown promising activity in single-arm phase II trials (that used tumor response as the primary endpoint of interest) but fail to demonstrate superiority in subsequent phase III trials. This has raised interest and discussion in the use of randomized phase II trial designs with the standard treatment as the control arm.^[14–17] Despite that, there remains a clear role for single-arm phase II designs with tumor response rates as the primary outcome of interest. In particular, as highlighted by Cannistra,^[16] they remain appropriate in situations where the behavior of historical controls has remained stable over time, or the likelihood of response to standard treatment options is low, or the desired treatment effect size is large, or the drug's mechanism of action is expected to be cytotoxic as opposed to cytostatic (permitting the use of response as an endpoint).

It should be noted that all these designs, and many others like them, are based on frequentist statistical approaches. These usually require the specification of various inputs, including a fixed target response rate for the new treatment as well as limits on the type I and type II errors.

The main goal of a phase II trial is not to obtain a precise estimate of the effect (response rate) of the new drug, but rather to accept or reject the drug for further testing in a phase III trial. Nevertheless, the analysis of the trial results typically includes the obtaining of an estimate of the true response proportion, along with the associated 95% confidence interval (CI). Such an analysis does not always answer the question of principal interest to the investigator. For example, even if the point estimate of the response rate exceeds the pre-specified target value, much lower rates may nevertheless be consistent with the CI reported. Thus the investigator might rather wish to know the probability that the true response proportion exceeds the pre-specified target value. Only if this probability is “high” is a subsequent phase III trial justified. With a Bayesian approach, we can obtain the posterior probability distribution of the true response proportion. From this we can then compute the probability that the true response proportion falls within any pre-specified region of interest, including the region above the target value. A Bayesian design also allows for the formal incorporation of relevant information from other sources of evidence in the monitoring and analysis of the phase II trial.

Several Bayesian designs have been proposed for phase II clinical trials, see for example, the methods proposed by Thall,^[18–23] Heitjan,^[24] Sylvester,^[25–27] Mayo and Gajewski,^[3,4] Sambucini,^[5] and Huang et al.^[28] Review

articles on the use of Bayesian statistical methods for clinical trials in general have been written by Berry^[29] and Zohar et al.^[30] In particular, Berry^[29] describes their use in the design and analysis of trials and emphasizes their value for the whole drug development process. Zohar et al.^[30] focuses on design and conduct issues specific to phase II single-arm trials with binary outcomes.

To be specific, Thall and Simon propose a design that, in its simplest form,^[18] involves the continual accrual of patients until the new drug is shown with high posterior probability to be either promising or not promising or until a pre-determined maximum sample size is reached. Their design requires the specification of an informative clinical prior^[31] for the response rate of the “standard” drug and a non-informative clinical prior for the response rate of the new drug. Here and throughout, by “standard” we mean the drug that has been found to be the best to date. Thall and Wathen propose a similar design that also takes account of between patient heterogeneity.^[23]

In contrast, Heitjan's design^[24] does not require the elicitation of the investigators' clinical prior beliefs for the new drug. Instead it requires the specification of hypothetical skeptical and enthusiastic priors.^[31] These represent the prior beliefs of a skeptic (or an enthusiast as appropriate) with regards to the response rate of the drug being tested. The design then works by requiring that the evidence from the data for stopping the trial be strong enough so as to convince either the skeptic (in the case of concluding efficacy) or the enthusiast (in the case of concluding insufficient efficacy).

Both Thall and Simon's designs, as well as Heitjan's, explicitly make use of probability distributions for both the response proportions of the standard drug as well as the new drug. This is unlike what happens in frequentist designs in which the response rate of the standard drug is only taken into account implicitly (via the value of the pre-specified fixed target threshold). Mehta and Cain^[32] make use of a fixed-target threshold in their Bayesian approach for the early stopping of pilot studies. However they do not consider the problem of sample size determination as such, but instead address the issue of the early termination of trials with planned sample sizes that have already been determined using other methods.

Huang et al.^[28] uses short-term response information to adaptively randomize subjects to treatment arms. The design is suitable for trials for which the primary endpoint is to evaluate survival outcomes, in particular of progression-free survival. They model the relationship between the short-term response and long-term survival using a Bayesian model, first using prior clinical information and then dynamically updating the model using accruing data from the trial. The trial will terminate once the pre-specified threshold probability or the maximum number of patients is reached, and hence there is no explicit sample size calculation at the design stage.

Here we discuss two variants of a Bayesian design for phase II clinical trials. As with the designs of Thall and

Simon, and of Heijttian, but in contrast to many of the other Bayesian designs proposed,^[25–27, 33–35] these designs do not require the specification of a loss or utility function. However, unlike them, they allow for the specification of a prior distribution for the response rate of the new drug and not one for the standard drug. This is intended to bring the designs closer to the more familiar frequentist designs and hence facilitate their adoption into clinical research practice.

The Tan-Machin designs we describe are two-stage designs. In such designs n patients are recruited to Stage 1 and (possibly) a further $N - n$ to Stage 2. However, we recognized when proposing these new designs that there are practical constraints, arising from experience with other designs, which need to be considered. These suggest that designs with Stage 1 of fewer than 5 patients are unlikely to be ever adopted in practice, and often realistic bounds to the lower and upper limits to the total study size, N , are set. In what follows, we have set these to 10 and 90 patients respectively.

SINGLE THRESHOLD DESIGN

We consider the first variant of Tan-Machin design, which is referred to as the Single Threshold Design (STD). Let R_U be the target response rate for any new drug and π_{prior} be the anticipated response rate for the drug being tested. Furthermore, let λ_1 and λ_2 (with $0 < \lambda_1 < \lambda_2 < 1$) be the minimum desired threshold probabilities, at the interim stage and at the end of the trial respectively, that the true response rate, π , exceeds R_U . We note that λ_1 and λ_2 replace the frequentist type I and type II error probabilities.

The design then determines the sample sizes for the trial based on the following principle. Suppose that the trial was to be conducted and that x_1 and x_2 were the resulting data obtained from Stage 1 and Stage 2 respectively. (Note that as no actual data has yet been observed from the trial, x_1 and x_2 at this point represent the hypothetical data that we assume would arise from the trial.) Now, suppose also that the (hypothetical) response proportion underlying x_1 and x_2 is just larger than the pre-specified R_U , that is, $R_U + \varepsilon_U$, where ε_U is some pre-selected small value (say 0.05). We then want the smallest overall sample size, N , that will enable the posterior probability at the end of the trial, $\Pr(\pi > R_U | x_1, x_2)$, to be at least λ_2 . At the same time, we also want the Stage 1 sample size n to be large enough so that the posterior probability at the end of Stage 1, $\Pr(\pi > R_U | x_1)$ is at least λ_1 . We want the smallest such value for n .

Once the sample sizes have been determined and the trial started, the n patients are recruited into Stage 1 of the trial. At the end of Stage 1, the posterior probability, $\Pr(\pi > R_U | x_1)$, is then evaluated. Note that x_1 now represents the actual data from Stage 1 of the trial, and not the hypothetical data used in the sample size determination. If $\Pr(\pi > R_U | x_1)$ is less than λ_1 , the trial terminates on the grounds that there is insufficient potential that the drug

will be found efficacious enough to be recommended for testing in a phase III trial. In addition, if the number of responses in Stage 1 are so low to the extent that, even if all the patients to be recruited in Stage 2 were to be responders, the final probability of θ exceeding R_U will still not exceed λ_2 , the trial would also terminate.

If neither of the above events occur, a further $N - n$ patients would be recruited into Stage 2 of the trial. The final posterior distribution, $\Pr(\pi > R_U | x_1, x_2)$ would then be evaluated. If this is found to be less than λ_2 , the trial would conclude that the drug is insufficiently efficacious to justify proceeding for phase III testing. Otherwise, if $\Pr(\pi > R_U | x_1, x_2)$ is greater than or equal to λ_2 , progression to a phase III trial would be recommended.

Prior Distribution

Assume that there is relatively little information available regarding the efficacy of the drug in treating the particular cancer in question. In such a situation, it would be appropriate to utilize a vague prior distribution in the trial design. In particular, we make use of a vague Beta distribution, which takes values between 0 and 1. (The case of having informative prior distributions is discussed by Mayo and Gajewski.^[3,4])

Following the form of the optimistic prior suggested by Heijttian,^[24] except that we center our prior on π_{prior} rather than R_U , we construct a vague Beta (α, β) prior distribution with parameters: $\alpha = \pi_{\text{prior}} + 1$, and $\beta = (1 - \pi_{\text{prior}}) + 1$. This represents weak prior information with mode at π_{prior} . Such a prior distribution is sufficiently vague to allow for the possibility that π may take any value in the range $0 < \pi < 1$, although the most likely value is π_{prior} .

Computational Algorithm and Sample Sizes

Before the start of the trial, the investigator is asked to provide, for the type of patients under investigation and in the knowledge of currently available drugs, the values of R_U , π_{prior} , λ_1 and λ_2 . The investigator may also provide the value of ε_U or alternatively, use the default value of $\varepsilon_U = 0.05$.

The algorithm then sets up a vague Beta prior distribution based on the elicited value of π_{prior} . Following this, a search is made for the smallest $N \geq 10$ that satisfies $\Pr(\pi > R_U | x_1, x_2) > \lambda_2$, where here, x_1 and x_2 are hypothetical data with a total sample size of N and a response proportion of $R_U + \varepsilon_U$.

To carry out this search, the algorithm starts with $N = 9$ and determines whether the condition $\Pr(\pi > R_U | x_1, x_2) > \lambda_2$ is satisfied. If so, it proceeds to test whether the condition $\Pr(\pi > R_U | x_1) > \lambda_1$ is met for $n = N - 5 = 4$. If this condition is also met, the algorithm terminates and concludes that the optimal N is less than 10.

Assume that the algorithm does not terminate at this point. It then considers $N = 10$ and calculates the corresponding value of $\Pr(\pi > R_U | x_1, x_2)$. If this exceeds λ_2 , the

algorithm proceeds to search for the optimal n (see next paragraph). Otherwise, it moves on to consider $N + 1$ and as before evaluates whether $\Pr(\pi > R_U | x_1, x_2)$ exceeds λ_2 . This process continues until a suitable N is found or until the maximum pre-specified value of N ($= 90$) is reached. If no suitable N is eventually found, the algorithm terminates with the conclusion that the recommended sample size exceeds $N = 90$.

Now, supposing that a suitable N in the range $10 \leq N \leq 90$ has been found. A further search is then made to identify the smallest n between 5 and $N - 5$ that satisfies $\Pr(\pi > R_U | x_1) > \lambda_1$. This is done in the same way as for the search over N , except that we now want to satisfy the condition that $\Pr(\pi > R_U | x_1) > \lambda_1$ instead. If no such n is found in the range $5 \leq n \leq N - 5$, the algorithm rejects this value of N and moves on to $N + 1$ in the process of searching for the next smallest value of N which satisfies $\Pr(\pi > R_U | x_1, x_2) > \lambda_2$.

The algorithm will eventually arrive at one of 4 possible outcomes: (a) it finds and outputs the recommended n and

N values, in the desired ranges, (b) it concludes that $N < 10$, (c) it concludes that $N \geq 10$ but $n < 5$, (d) it concludes that either $N \geq 90$ or $n \geq 85$.

Sample sizes and cutoff values for a range of R_U , π_{prior} , λ_1 and λ_2 values have been tabulated. A selection of these (with $\varepsilon_U = 0.05$) are given in Table 1. For more extensive tables, see Tan and Machin.^[1]

DUAL THRESHOLD DESIGN

The Dual Threshold Design (DTD) is identical to the STD except that the Stage 1 sample size is determined, not on the basis of the probability of exceeding R_U , but on the probability that π will be less than the “no further interest” proportion, R_L . This represents the response rate below which the investigator would have no further interest in the new drug. R_L thus functions as a lower threshold on the response rate, as opposed to the upper threshold represented by R_U . The rationale behind

Table 1 Number of Patients Required for the Single Threshold Design with $\varepsilon_U = 0.05$

π_{prior} R_U	0.1		0.3		0.5		0.7		0.9	
	Stage 1	Overall	Stage 1	Overall	Stage 1	Overall	Stage 1	Overall	Stage 1	Overall
0.2	<i>a</i>		<i>a</i>		<i>a</i>		<i>a</i>		<i>a</i>	
	<i>b</i>		<i>b</i>		<i>b</i>		<i>a</i>		<i>a</i>	
	1/8	9/38	<i>b</i>		<i>b</i>		<i>a</i>		<i>a</i>	
0.3	1/5	8/24	<i>b</i>		<i>a</i>		<i>a</i>		<i>a</i>	
	1/5	21/61	<i>b</i>		<i>b</i>		<i>b</i>		<i>b</i>	
	8/24	21/61	5/15	18/53	<i>b</i>		<i>b</i>		<i>b</i>	
0.4	6/14	15/35	3/7	12/28	<i>b</i>		<i>a</i>		<i>a</i>	
	6/14	35/78	3/7	31/70	<i>b</i>		<i>b</i>		<i>b</i>	
	15/35	35/78	12/28	31/70	8/20	27/62	3/8	23/53	<i>b</i>	
0.5	10/20	24/44	8/15	20/37	4/9	16/30	<i>b</i>		<i>b</i>	
	10/20	48/88	8/15	44/81	4/9	40/73	<i>b</i>		<i>b</i>	
	24/44	48/88	20/37	44/81	16/30	40/73	12/22	35/65	6/12	31/57
0.6	16/26	32/50	13/21	28/44	9/15	24/37	5/9	19/30	<i>b</i>	
	<i>c</i>		13/21	55/86	9/15	50/78	5/9	46/71	<i>b</i>	
	<i>c</i>		28/44	55/86	24/37	50/78	19/30	46/71	14/22	40/63
0.7	23/31	39/53	19/26	35/47	15/21	30/41	11/15	25/34	7/10	20/27
	<i>c</i>		19/26	62/84	15/21	57/77	11/15	52/70	7/10	47/63
	<i>c</i>		35/47	62/84	30/41	57/77	25/34	52/70	20/27	47/63
0.8	29/35	45/53	25/30	39/47	21/25	34/41	16/20	30/36	12/15	24/29
	29/35	70/83	25/30	65/77	21/25	59/70	16/20	53/63	12/15	47/56
	45/53	70/83	39/47	65/77	34/41	59/70	30/36	53/63	24/29	47/56
0.9	34/36	46/49	30/32	41/44	25/27	36/38	21/23	31/33	17/18	25/27
	34/36	63/67	30/32	57/61	25/27	52/55	21/23	46/49	17/18	40/43
	46/49	63/67	41/44	57/61	36/38	52/55	31/33	46/49	25/27	40/43

For each value of R_U , the first, second, and third rows correspond to $(\lambda_1, \lambda_2) = (0.6, 0.7)$, $(0.6, 0.8)$, and $(0.7, 0.8)$, respectively. The drug is rejected if the response rate is less than or equal to the fractions above. Where no fractions are given, the interpretation is as follows: “*a*”—Overall Sample Size < 10 ; “*b*”—Overall Sample Size ≥ 10 but Stage 1 Sample Size < 5 ; “*c*”—Overall Sample Size > 90 or Stage 1 Sample Size > 85 .

this aspect of the DTD is that we want our Stage 1 sample size to be large enough so that, if the trial data really does suggest a response rate that is below R_L , we want the posterior probability of π being below R_L , to be at least λ_1 . We want to find the smallest Stage 1 sample size that satisfies this criterion.

To implement the design, n patients are recruited into Stage 1 of the trial. At the end of Stage 1, the posterior probability, $\Pr(\pi < R_L | x_1)$, is evaluated. If this is found to be greater than λ_1 , the trial terminates on the grounds that the response rate of the drug is likely to have a high probability of being below the “no further interest” response proportion, R_L . Also, as for the STD, if the number of responses in Stage 1 are so low to the extent that it is not possible for final probability of θ exceeding R_U to meet the λ_2 threshold, the trial will also terminate at this point.

Otherwise, the trial proceeds to recruit a further $N - n$ patients. The final posterior distribution, $\Pr(\pi > R_U | x_1, x_2)$ is then evaluated. As in the STD, the decision as to whether to recommend further testing of the drug in a phase III trial will then depend on whether $\Pr(\pi > R_U | x_1, x_2)$ is greater than or less than λ_2 .

Prior Distribution, Computational Algorithm, and Sample Sizes

The same vague Beta prior distribution used for the STD is again used for the DTD. As for the computational algorithm, this works as follows. Before the start of the trial, values of R_U , π_{prior} , λ_1 and λ_2 need to be elicited as in the STD. (However, note that we no longer require the condition that $\lambda_1 < \lambda_2$ as in the STD.) In addition, the investigator now also needs to provide a value for R_L . Values for ε_U and ε_L ($0 < \varepsilon_L < R_L$) may also be specified or the default values of $\varepsilon_U = \varepsilon_L = 0.05$ used.

The algorithm then searches for the optimal value of N as in the STD. For each selected N , it then searches for the smallest n in the range $5 \leq n \leq N - 5$ such that, if (hypothetically) the trial data were to have a response proportion of $R_L - \varepsilon_L$, the value of $\Pr(\pi < R_L | x_1)$ would be strictly greater than λ_1 .

Sample sizes for selected values of R_L , R_U , π_{prior} , λ_1 and λ_2 using the DTD with $\varepsilon_L = \varepsilon_U = 0.05$ have been tabulated. A selection of these are given in Table 2, for the case when $R_U - R_L = 0.2$ (see Tan and Machin^[1] for a wider range of values).

Table 2 Number of Patients Required for the Dual Threshold Design With $\varepsilon_L = \varepsilon_U = 0.05$ ($R_U - R_L = 0.2$)

π_{prior}		0.1		0.3		0.5		0.7		0.9	
R_L	R_U	Stage 1	Overall	Stage 1	Overall	Stage 1	Overall	Stage 1	Overall	Stage 1	Overall
0.1	0.3	2/18	8/24	4/23	9/28	5/27	10/32	6/32	11/37	8/36	13/41
		0/18	21/61	1/23	18/53	1/27	15/44	7/32	12/37	9/36	14/41
		1/27	21/61	1/33	18/53	9/38	15/44	11/44	16/49	13/49	18/54
0.2	0.4	2/15	15/35	4/20	12/28	8/25	13/30	10/30	15/35	12/35	17/40
		2/15	35/78	3/20	31/70	3/25	27/62	4/30	23/53	10/35	19/44
		4/29	35/78	5/36	31/70	6/41	27/62	17/47	23/53	20/53	25/58
0.3	0.5	2/10	24/44	3/15	20/37	7/21	16/30	11/26	16/31	14/31	19/36
		2/10	48/88	3/15	44/81	5/21	40/73	6/26	35/65	7/31	31/57
		6/27	48/88	8/34	44/81	10/41	40/73	17/47	35/65	26/53	31/58
0.4	0.6	<i>b</i>		3/9	28/44	5/15	24/37	10/21	19/30	14/26	19/31
		<i>c</i>		3/9	55/86	5/15	50/78	7/21	46/71	9/26	40/63
		<i>c</i>		10/30	55/86	12/37	50/78	19/44	46/71	27/50	40/63
0.5	0.7	<i>b</i>		<i>b</i>		4/9	30/41	6/15	25/34	13/20	20/27
		<i>c</i>		<i>b</i>		4/9	57/77	6/15	52/70	9/20	47/63
		<i>c</i>		9/22	62/84	13/30	57/77	19/37	52/70	28/44	47/63
0.6	0.8	<i>b</i>		<i>b</i>		<i>b</i>		3/7	30/36	9/14	24/29
		<i>b</i>		<i>b</i>		<i>b</i>		3/7	53/63	7/14	47/56
		<i>b</i>		4/8	65/77	11/20	59/70	18/28	53/63	26/35	47/56
0.7	0.9	<i>b</i>		<i>b</i>		<i>b</i>		<i>b</i>		3/5	25/27
		<i>b</i>		<i>b</i>		<i>b</i>		<i>b</i>		3/5	40/43
		<i>b</i>		<i>b</i>		<i>b</i>		12/15	46/49	21/24	40/43

For each value of R_U , the first, second and third rows correspond to $(\lambda_1, \lambda_2) = (0.6, 0.7)$, $(0.6, 0.8)$, and $(0.7, 0.8)$, respectively. The drug is rejected if the response rate is less than or equal to the fractions above. Where no fractions are given, the interpretation is as follows: “*b*”—Overall Sample Size ≥ 10 but Stage 1 Sample Size < 5 ; “*c*”—Overall Sample Size > 90 or Stage 1 Sample Size > 85 .

PROBABILITIES OF EARLY TERMINATION AND EXPECTED SAMPLE SIZES

We now obtain the probabilities of early termination (PET) and the expected sample sizes (EN) of STD and DTD, given that the true response proportion, π , is $(R_L - 0.05)$. Here we note that as we are in a Bayesian framework, π does not actually have a fixed value but instead has a probability distribution. Nevertheless, for the purpose of making these computations, we assume that π has a single “true value,” $(R_L - 0.05)$. We can also view this as a probability distribution with zero variance and mean $(R_L - 0.05)$.

For the DTD, the PET is then given by $\Pr [\Pr (\pi < R_L | x_1) > \lambda_1]$, where x_1 is the data from a binomial distribution with parameters n and $(R_L - 0.05)$. For the STD, the PET is given by $\Pr [\Pr (\pi > R_U | x_1) < \lambda_1]$, where x_1 is also from a binomial distribution but with parameters n and $(R_U - 0.05)$. As for the expected sample size, this is given by $EN = n + (1 - PET)(N - n)$.

Tables 3 and 4 give the PET and EN values corresponding to the STD and DTD, respectively, with $R_U - R_L = 0.2$, $\varepsilon_L = \varepsilon_U = 0.05$ and $(\lambda_1, \lambda_2) = (0.6, 0.7)$, $(0.6, 0.8)$ and $(0.7, 0.8)$. The PET values for the STD (Table 3) are mostly above 0.7 (many above 0.9), suggesting that the design is likely to recommend stopping at the end of Stage 1, if π is really equal to $R_U - 0.05$. For the DTD (Table 4), the PET values tend to be lower, generally around 0.5. We note however, that for both the STD and DTD, the PET values depend on what we set the true π to be. Taking smaller true π values will increase the PET. As for the EN, these range from 11.1–53.9 for the STD and from 21.2–60.2 for the DTD.

We have also investigated other properties of the design, such as the variation of the sample sizes with π_{prior} , λ_1 , λ_2 and ε_U . Comparisons between Tan-Machin and Simon's optimal and minimax designs^[9] have also been made. The interested reader is referred to Tan and Machin^[1] for a detailed discussion of all these.

Table 3 Probability of Early Termination (PET) and Expected Sample Sizes (EN) when $\pi = R_U - 0.05$, for the Single Threshold Design with $\varepsilon_U = 0.05$

π_{prior}	0.1		0.3		0.5		0.7		0.9	
R_U	PET	EN	PET	EN	PET	EN	PET	EN	PET	EN
0.2	*	*	*	*	*	*	*	*	*	*
	*	*	*	*	*	*	*	*	*	*
	0.66	18.1	*	*	*	*	*	*	*	*
0.3	0.64	11.9	*	*	*	*	*	*	*	*
	0.64	25.3	*	*	*	*	*	*	*	*
	0.88	28.4	0.86	20.5	*	*	*	*	*	*
0.4	0.81	17.9	0.81	11.1	*	*	*	*	*	*
	0.81	26.0	0.81	19.2	*	*	*	*	*	*
	0.87	40.5	0.86	34.1	0.77	29.8	0.70	21.6	*	*
0.5	0.75	25.9	0.82	19.0	0.62	17.0	*	*	*	*
	0.75	36.9	0.82	27.0	0.62	33.4	*	*	*	*
	0.92	47.5	0.90	41.5	0.87	35.7	0.86	27.9	0.74	23.7
0.6	0.81	30.6	0.81	25.5	0.74	20.7	0.63	16.7	*	*
	*	*	0.81	33.6	0.74	31.4	0.63	31.6	*	*
	*	*	0.90	48.1	0.91	40.5	0.87	35.4	0.85	38.3
0.7	0.90	33.2	0.86	39.0	0.80	24.9	0.83	18.2	0.73	14.5
	*	*	0.86	34.2	0.80	31.9	0.83	24.3	0.73	24.1
	*	*	0.94	49.1	0.90	44.7	0.89	37.9	0.89	30.8
0.8	0.90	36.8	0.90	31.7	0.91	26.4	0.77	23.7	0.77	18.3
	0.90	39.8	0.90	34.7	0.91	29.0	0.77	29.9	0.77	24.6
	0.97	53.9	0.93	49.1	0.91	43.5	0.92	38.2	0.88	32.2
0.9	0.98	36.2	0.97	32.4	0.93	27.8	0.88	24.2	0.95	18.5
	0.98	36.6	0.97	33.0	0.93	29.0	0.88	26.0	0.95	19.3
	0.98	49.3	0.97	44.4	0.99	38.2	0.97	33.5	0.93	28.2

For each value of R_U , the first, second, and third rows correspond $(\lambda_1, \lambda_2) = (0.6, 0.7)$, $(0.6, 0.8)$, and $(0.7, 0.8)$, respectively. The label * means that there is no recommended sample size in the desired range.

Table 4 Probability of Early Termination (PET) and Expected Sample Sizes (EN) When $\pi = R_L - 0.05$, for the Dual Threshold Design with $\varepsilon_L = \varepsilon_U = 0.05$

π_{prior}		0.1		0.3		0.5		0.7		0.9	
R_L	R_U	PET	EN	PET	EN	PET	EN	PET	EN	PET	EN
0.10	0.30	0.39	21.7	0.68	24.6	0.61	29.0	0.52	34.4	0.46	38.7
		0.39	44.2	0.68	32.7	0.61	33.7	0.52	34.4	0.46	38.7
		0.62	39.9	0.51	42.9	0.43	41.4	0.62	45.9	0.55	51.2
0.20	0.40	0.60	23.1	0.65	22.8	0.47	27.6	0.53	32.3	0.58	37.1
		0.60	40.4	0.65	37.4	0.47	44.6	0.53	40.8	0.58	38.8
		0.55	51.1	0.54	51.7	0.58	49.7	0.59	49.5	0.45	55.8
0.30	0.50	0.52	26.3	0.47	26.6	0.57	24.9	0.51	28.4	0.46	33.7
		0.52	47.5	0.47	49.8	0.57	43.4	0.51	45.0	0.46	45.0
		0.47	59.3	0.52	56.5	0.55	55.5	0.48	56.3	0.54	55.3
0.40	0.60	*	*	0.61	22.6	0.56	24.7	0.54	25.1	0.57	28.1
		*	*	0.61	38.8	0.56	42.8	0.54	44.0	0.57	41.9
		*	*	0.51	57.4	0.44	60.2	0.52	56.9	0.51	56.3
0.50	0.70	*	*	*	*	0.62	21.3	0.45	25.4	0.59	22.8
		*	*	*	*	0.62	35.1	0.45	45.2	0.59	37.5
		*	*	0.44	57.0	0.50	53.3	0.48	54.1	0.47	54.2
0.60	0.80	*	*	*	*	*	*	0.40	24.4	0.46	22.1
		*	*	*	*	*	*	0.40	40.6	0.46	36.8
		*	*	0.52	41.2	0.59	40.7	0.51	45.1	0.54	44.8
0.70	0.90	*	*	*	*	*	*	*	*	0.57	14.4
		*	*	*	*	*	*	*	*	0.57	21.2
		*	*	*	*	*	*	0.43	34.2	0.48	33.9

For each value of R_L and R_U , the first, second, and third rows correspond $(\lambda_1, \lambda_2) = (0.6, 0.7)$, $(0.6, 0.8)$, and $(0.7, 0.8)$, respectively. The label * means that there is no recommended sample size in the desired range.

PHASE II TRIAL USING A TRIPLET COMBINATION: PACLITAXEL, CARBOPLATIN, AND GEMCITABINE IN METASTATIC NASOPHARYNGEAL CARCINOMA

A phase II trial on the activity of a triplet combination of paclitaxel, carboplatin, and gemcitabine in chemotherapy-naïve patients with metastatic nasopharyngeal carcinoma was conducted at the National Cancer Centre in Singapore, with patient enrollment between July 2001 and April 2003.^[2]

This trial was designed to test the hypothesis of no further interest in the regimen if the response rate (complete responses and partial responses) was as low as 60% against the alternative hypothesis of an overall response rate of 80% or greater. The Bayesian DTD design with parameters $\{R_L, R_U, \pi_{\text{prior}}, \lambda_1, \lambda_2\} = \{0.6, 0.8, 0.8, 0.65, 0.7\}$ was used giving rise to a required Stage 1 sample size of $n = 19$ and a possible total sample size of $N = 32$ patients.

At the end of Stage 1, the data collected from the first $n = 19$ patients were used to update the prior distribution, based on the value of $\pi_{\text{prior}} (= 0.8)$ specified for the trial.

The probability that π was less than $R_L = 0.6$ was evaluated and under the DTD, the trial was only to proceed to Stage 2 if $\Pr(\pi < R_L | x_1)$ is less than λ_1 .

Eighteen responses (out of the 19 patients) were reported at the end of Stage 1. Based on this data, it was found that $l_1 = \Pr(\pi < R_L | x_1) = 0.0005 < \lambda_1 (= 0.65)$. Moreover it remained possible for the trial to eventually conclude “success” because should all the 13 Stage 2 patients be responders, we would have $\Pr(\pi > R_U | x_1, x_2) = 0.9931$, which would exceed λ_2 . Hence the Stage 1 criteria of the DTD to proceed to the next stage was met, and the trial proceeded to Stage 2.

At the end of Stage 2, a further 7 responses were reported, resulting in a total of 25 responses (out of the total 32 patients). This is used to further update the probability distribution of π to give a final posterior probability of $u_2 = \Pr(\pi > R_U | x_1, x_2) = 0.37$. Since $u_2 = 0.37 < \lambda_2 = 0.7$, this suggests that the triplet combination is not quite as promising as was hoped before the start of the trial. We note, however, that while $u_2 = \Pr(\pi > R_U | x_1, x_2)$ is only 0.37, $l_2 = \Pr(\pi < R_L | x_1, x_2)$ is still smaller than 0.65 at 0.0167. Thus, we clearly also cannot conclude “no further interest” in the triplet combination.

Now assume, for the purpose of illustration, that the STD was instead used to design the trial, with a hypothesis to test for an overall response rate of 80% or greater, with similar prior expected response rate, Stage 1 and 2 threshold probabilities of 0.8, 0.65 and 0.7, respectively. The STD design with parameters $\{R_U, \pi_{\text{prior}}, \lambda_1, \lambda_2\} = \{0.8, 0.8, 0.65, 0.7\}$ would require a Stage 1 sample size of $n = 24$ and a total sample size of $N = 32$ patients. Assuming that the study was conducted using the STD, at the end of Stage 1, the data collected from the first 24 patients would be used to update the prior distribution, based on the value of π_{prior} ($= 0.8$) specified. The probability that π is greater than $R_U = 0.8$ would then be evaluated.

The actual data from the trial showed that there were 22 responses out of the first 24 eligible patients. Based on this data, it was found that $u_1 = \Pr(\pi > R_U | x_1) = 0.8965$. Since $u_1 = 0.8965 > \lambda_1 = 0.65$, then the response rates at Stage 1 would have also satisfied the Stage 1 criteria of the STD to allow the trial to proceed to Stage 2.

At the end of Stage 2, based on the result of 25 responses (out of the total 32 patients), the updated probability distribution of π would give a final posterior probability of $u_2 = \Pr(\pi > R_U | x_1, x_2) = 0.37 < \lambda_2 (= 0.7)$ and conclude that the triplet combination is not quite as promising as anticipated at the start of the trial. Note that this is the same calculation as for the DTD.

It can be seen that for this particular trial, the initial very encouraging results seen in patients recruited in the first stage of the trial were generally not reproduced in patients recruited into the second stage.

ADAPTATIONS OF THE DESIGNS

A number of authors have proposed Bayesian phase II designs that were adaptations to the original design proposed by Tan and Machin. Mayo and Gajewski^[3] had in their first adaptation to the design extended the sample size calculations to include informative prior distributions. The informative prior distributions were defined based on several strategies. These strategies allow researchers with a pessimistic or optimistic prior to be involved in the sample size decision-making. Four approaches to constructing the informative prior were discussed in the article. Sample size was then determined using the same principles as in Tan-Machin of calculating a posterior probability that the true response rate will exceed a pre-specified target. However, unlike the Tan-Machin design, the sample size calculations were for single-stage phase II trials and hence did not have analogous single threshold and dual threshold variants.

In Gajewski and Mayo's second adaptation,^[4] they incorporated the use of a mixture of informative priors. The mixture would be defined based on several sources of information, and could also be constructed from a combination of pessimistic and optimistic opinions from

clinicians. Similar to their earlier adaptation, the sample size calculation is for single-stage phase II trials.

Sambucini^[5] introduced a "predictive version" of the STD design. The proposed method is motivated by the uncertainty surrounding the observations to be (but not yet) made and wanting to take that into account in the sample size calculation. It utilizes a design prior with mode greater than the target value to compute the prior predictive distribution of the data and a non-informative analysis prior to obtain the posterior probabilities of the treatment efficacy. The two-stage sample sizes are determined based on the criteria that the probabilities, computed with respect to the prior predictive distributions of the data, of getting the posterior probability greater than or equal to λ_1 and λ_2 (the STD thresholds) is greater than or equal to two additional thresholds, γ_1 and γ_2 respectively. As such, the design requires a larger number of design parameters to be specified than the original STD and differs from the STD in that the optimal second stage sample size and cutoff level depends on the first-stage optimal sample size and cutoff level.

DISCUSSION

We believe that Bayesian methods provide a good alternative to frequentist approaches for phase II clinical trials and have much to offer to the design, monitoring, and analysis of such trials. They allow for the incorporation of relevant prior information and the presentation of the trial results in a manner that seems more intuitive and helpful than what is obtained from a frequentist analysis. However, one of the main reasons why Bayesian methods are not in common use is because of their unfamiliarity to investigators and the perceived difficulty of implementation. However, the Tan-Machin designs are relatively simple compared to many of the other Bayesian phase II designs in the literature.

These designs are presented in a way that should be reasonably familiar to anyone who has had experience working with frequentist designs such as the Simon two-stage designs. As with Simon's designs, the investigator needs to specify for the STD appropriate values for R_U and (for the DTD) R_L . They also need to specify λ_1 and λ_2 values (as opposed to type I and II errors). Finally they need to specify a value for π_{prior} , their prior anticipated response rate for the new drug. This is a single value that will "automatically" form the basis of a vague prior. Thus there is no need to elicit a full prior distribution (although it is straightforward to adapt the design for use with, for example, an elicited informative Beta prior). Once all these values have been specified, the design establishes Stage 1 and overall sample sizes for the trial. Other variables, such as ε_U and ε_L can also be specified if desired, but it is usually sufficient to use the default values for these.

The choice of λ_1 and λ_2 values will clearly have a strong impact on the recommended sample sizes (in the same way that the choice of type I and type II error values will affect frequentist designs). In Tables 1 and 2, we make use of $(\lambda_1, \lambda_2) = (0.6, 0.7)$, $(0.6, 0.8)$ and $(0.7, 0.8)$ values. These were chosen on the basis that they gave rise to sample sizes which generally covered a similar range to those recommended by the commonly used Simon designs. Choosing smaller threshold values would result in smaller sample sizes. However, this could potentially be at the cost of having a design that has an unreasonably low probability of detecting a true effect (that is, the distribution of π is mostly above R_U or is below R_L) when the data does actually suggest this. On the other hand, using threshold values that are too high could result in sample sizes that are too large to be of practical value. There is thus a tradeoff between wanting a small sample size for practical reasons, while at the same time, ensuring that the corresponding λ_1 or λ_2 value is not too low.

The STD and DTD designs have been prospectively used in actual clinical trials. In particular, the STD has been used in the design of a phase II study to evaluate the use of Interferon- α for recurrent WHO grade 1 intracranial meningiomas^[36] and the DTD has been used in the design of a phase II clinical trial to evaluate the use of a triplet combination of paclitaxel, carboplatin, and gemcitabine in metastatic nasopharyngeal carcinoma conducted at the National Cancer Centre in Singapore.^[2]

We have utilized the latter trial^[2] to illustrate the use of the DTD. We have also applied the STD in that example by removing the no further interest threshold, R_L , which is only applicable in the DTD setting, to demonstrate the different objectives that the two variants of the design serve. Both the STD and DTD recommended that the trial should carry on to the second stage. However, the STD is concerned with ensuring, at the interim stage, that the final response rate of the drug has a reasonable probability of exceeding the target R_U at the end of the trial. On the other hand, the DTD is concerned at the interim stage in making sure that the final response rate is not lower than the no further interest threshold, R_L . The final results of the trial (designed using DTD) concluded that while there was insufficient evidence that the response rate exceeded R_U , it was also clear that the rate was not below R_L . Thus the DTD have correctly recommended that the trial moved on to the next stage. The STD however, did not recommend that the trial be terminated after Stage 1, and this was largely due to the initial encouraging results seen in first stage that were not reflected in the second stage.

Alongside that for other commonly used phase II designs, interactive and user-friendly sample size calculation software^[37,38] has been developed to implement the Tan-Machin designs. Adaptations of the basic designs to address two primary (binary) endpoints, and an endpoint that is ordinal, is currently in progress. The first arises because both efficacy and toxicity need to be considered

simultaneously when a new cancer therapy is being developed. The latter arises in situations concerning patients with retinoblastoma in which one or both eyes may be involved, when response could be in neither, one, or both organs.

REFERENCES

1. Tan, S.B.; Machin, D. Bayesian two-stage designs for phase II clinical trials. *Stat. Med.* **2002**, *21* (14), 1991–2012.
2. Leong, S.S.; Wee, J.; Tay, M.H.; Toh, C.K.; Tan, S.B.; Thng, C.H., et al. Paclitaxel, carboplatin, and gemcitabine in metastatic nasopharyngeal carcinoma: a phase II trial using a triplet combination. *Cancer* **2005**, *103* (3), 569–575.
3. Mayo, M.S.; Gajewski, B.J. Bayesian sample size calculations in phase II clinical trials using informative conjugate priors. *Control Clin. Trials* **2004**, *25* (2), 157–167.
4. Gajewski, B.J.; Mayo, M.S. Bayesian sample size calculations in phase II clinical trials using a mixture of informative priors. *Stat. Med.* **2006**, *25* (15), 2554–2566.
5. Sambucini, V. A Bayesian predictive two-stage design for phase II clinical trials. *Stat. Med.* **2008**, *27* (8), 1199–1224.
6. Fleming, T.R. One-sample multiple testing procedure for phase II clinical trials. *Biometrics* **1982**, *38* (1), 143–151.
7. A'Hern, R.P. Sample size tables for exact single-stage phase II designs. *Stat. Med.* **2001**, *20* (6), 859–866.
8. Gehan, E.A. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *J. Chronic Dis.* **1961**, *13*, 346–353.
9. Simon, R. Optimal two-stage designs for phase II clinical trials. *Control Clin. Trials* **1989**, *10* (1), 1–10.
10. Sposto, R.; Gaynon, P.S. An adjustment for patient heterogeneity in the design of two-stage phase II trials. *Stat. Med.* **2009**, *28* (20), 2566–2579.
11. Lin, Y.; Shih, W.J. Adaptive two-stage designs for single-arm phase IIA cancer clinical trials. *Biometrics* **2004**, *60* (2), 482–490.
12. Simon, R.; Wittes, R.E.; Ellenberg, S.S. Randomized phase II clinical trials. *Cancer Treat. Rep.* **1985**, *69* (12), 1375–1381.
13. Liu, P.Y.; Dahlberg, S.; Crowley, J. Selection designs for pilot studies based on survival. *Biometrics* **1993**, *49* (2), 391–398.
14. Rubinstein, L.V.; Korn, E.L.; Freidlin, B.; Hunsberger, S.; Ivy, S.P.; Smith, M.A. Design issues of randomized phase II trials and a proposal for phase II screening trials. *J. Clin. Oncol.* **2005**, *23* (28), 7199–7206.
15. Wieand, H.S. Randomized phase II trials: what does randomization gain? *J. Clin. Oncol.* **2005**, *23* (9), 1794–1795.
16. Cannistra, S.A. Phase II trials in journal of clinical oncology. *J. Clin. Oncol.* **2009**, *27* (19), 3073–3076.
17. Ratain, M.J.; Sargent, D.J. Optimising the design of phase II oncology trials: the importance of randomisation. *Eur. J. Cancer* **2009**, *45* (2), 275–280.
18. Thall, P.F.; Simon, R. Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics* **1994**, *50* (2), 337–349.
19. Thall, P.F.; Simon, R. A Bayesian approach to establishing sample size and monitoring criteria for phase II clinical trials. *Control. Clin. Trials* **1994**, *15* (6), 463–481.

20. Thall, P.F.; Simon, R.M.; Estey, E.H. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Stat. Med.* **1995**, *14* (4), 357–379.
21. Thall, P.F.; Simon, R.M.; Estey, E.H. New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. *J. Clin. Oncol.* **1996**, *14* (1), 296–303.
22. Thall, P.F.; Sung, H.G. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Stat. Med.* **1998**, *17* (14), 1563–1580.
23. Thall, P.F.; Wathen, J.K. Bayesian designs to account for patient heterogeneity in phase II clinical trials. *Curr. Opin. Oncol.* **2008**, *20* (4), 407–411.
24. Heitjan, D.F. Bayesian interim analysis of phase II cancer clinical trials. *Stat. Med.* **1997**, *16* (16), 1791–1802.
25. Staquet, M.; Sylvester, R. A decision theory approach to phase II clinical trials. *Biomedicine* **1977**, *26* (4), 262–266.
26. Sylvester, R.J. A Bayesian approach to the design of phase II clinical trials. *Biometrics* **1988**, *44* (3), 823–836.
27. Sylvester, R.J.; Staquet, M.J. Design of phase II clinical trials in cancer using decision theory. *Cancer Treat. Rep.* **1980**, *64* (2–3), 519–524.
28. Huang, X.; Ning, J.; Li, Y.; Estey, E.; Issa, J.P.; Berry, D.A. Using short-term response information to facilitate adaptive randomization for survival clinical trials. *Stat. Med.* **2009**, *28* (12), 1680–1689.
29. Berry, D.A. Bayesian clinical trials. *Nat. Rev. Drug Discov.* **2006**, *5* (1), 27–36.
30. Zohar, S.; Teramukai, S.; Zhou, Y. Bayesian design and conduct of phase II single-arm clinical trials with binary outcomes: a tutorial. *Contemp. Clin. Trials* **2008**, *29* (4), 608–616.
31. Spiegelhalter, D.J.; Freedman, L.S.; Parmar, M.K.B. Bayesian approaches to randomized trials. *J. R. Stat. Soc. Ser. A* **1994**, *157* (3), 357–387.
32. Mehta, C.R.; Cain, K.C. Charts for the early stopping of pilot studies. *J. Clin. Oncol.* **1984**, *2* (6), 676–682.
33. Brunier, H.C.; Whitehead, J. Sample sizes for phase II clinical trials derived from Bayesian decision theory. *Stat. Med.* **1994**, *13* (23–24), 2493–2502.
34. Stallard, N. Sample size determination for phase II clinical trials based on Bayesian decision theory. *Biometrics* **1998**, *54* (1), 279–294.
35. Stallard, N.; Thall, P.F.; Whitehead, J. Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics* **1999**, *55* (3), 971–977.
36. Chamberlain, M.C.; Glantz, M.J. Interferon-alpha for recurrent World Health Organization grade 1 intracranial meningiomas. *Cancer* **2008**, *113* (8), 2146–2151.
37. Tan, S.B.; Tan, S.H.; Machin, D. *Early Phase Clinical Trials (EPCT) Software*; UKCCSG Data Centre: Leicester, 2005.
38. Machin, D.; Campbell, M.J.; Tan, S.-B.; Tan, S.-H. *Sample Size Tables for Clinical Studies*, 3rd Ed.; Wiley-Blackwell: Chichester, 2009.

Bayesian Estimate: Concordance Correlation Coefficient

Dai Feng, Richard Baumgartner, and Vladimir Svetnik

Biometrics Research, MRL, Rahway, New Jersey, U.S.A.

Abstract

Concordance correlation coefficient (CCC) is one of the most frequently used and reported scaled indices of agreement. There have been several frequentist methods proposed to estimate the CCC. Thus, several questions pertaining to the Bayesian estimate of the CCC naturally arose. In this entry, we aimed at answering the following questions: Is there any added value of a Bayesian approach when compared to frequentist methods? How is a Bayesian estimate of the CCC executed? Are there any potential limitations of Bayesian approaches and what are the possible solutions? Most of the Bayesian methods discussed in the entry were implemented in a package *agRee* available from the Comprehensive R Archive Network.

Keywords: Concordance correlation coefficient; Bayesian; MCMC; Accommodation of covariates; Missing values and multiple replications; Model diagnostics and comparison.

INTRODUCTION

The need for assessment of agreement between observers is ubiquitous in a wide variety of substantive fields.^[1] Note that observer is a generic term, and in the drug development environment, it could refer to measurement method, test, assay, device, etc. There are, in general, three approaches for measuring agreement—descriptive tools, unscaled agreement indices, and scaled agreement indices.^[1] Among them, a widely used scaled index is the concordance correlation coefficient (CCC).

The CCC originally proposed by Lin^[2] measures the variation of linear relationship between readings from two observers from the 45° line through the origin. For pairs of samples (y_{i1}, y_{i2}), from $i = 1, 2, \dots, n$ subjects, the expectation of the squared difference $E[(Y_{i1} - Y_{i2})^2]$ measures the degree of concordance between two observers. Furthermore, the value of $E[(Y_{i1} - Y_{i2})^2]$ is transferred to make it scaled between -1 and 1. The larger the CCC value, the better the agreement. Hereafter, an uppercase Y (with subscript) is used to denote a random variable (or vector) and a lowercase y (with subscript) is used to denote the corresponding sample of Y .

The concept of CCC was extended to multiple observers for data without or with replications^[3–6] and to data with repeated measures from two or multiple fixed or random observers.^[7–11] These extensions usually ensue from using different analysis of variance models. Inference is typically based on the normality assumption or performed using semiparametric generalized estimating equations or non-parametric approaches.^[1,6]

Given several frequentist methods having been proposed to estimate the CCC, a natural question of a need for another Bayesian approach and its utility arose. What would be the additional benefits of the Bayesian methods and how could the Bayesian estimation be carried out? Are there potential limitations and challenges with the Bayesian methods and what are the possible solutions? This entry is aimed at answering these questions. Section “Motivation for Bayesian Estimate of the CCC” presents the motivation for the Bayesian estimate of the CCC. Section “How to Execute the Bayesian Estimate of the CCC” describes the details of Bayesian estimation. Section “Potential Limitations and Future Research for the Bayesian Estimate of the CCC” focuses on the potential issues with the Bayesian approaches and discusses possible solutions and future research topics.

MOTIVATION FOR BAYESIAN ESTIMATE OF THE CCC

A series of Bayesian methods for the estimation of the CCC were proposed and investigated.^[12–15] A motivating example in the study by Feng, Baumgartner, and Svetnik^[15] was from insomnia drug development. The gold standard for obtaining sleep measurements is through a polysomnography (PSG) study whereby a subject sleeps in a specially equipped lab, in which the brain's electrical activity, along with multiple other signals, is measured with a number of sensors attached to the subject's head, face, and body. Collected signals, or waveforms, are then

analyzed and each 30-second epoch of a subject's sleep (typically 8-hour) is attributed to (or scored as) one of five states (four for sleep and one corresponding to wake^[16]). Once scoring is completed, various summary measures of sleep are calculated.

The PSG scoring by visual analysis, though considered a gold standard for PSG data processing, is usually quite tedious and takes huge amount of time. There have been efforts made to develop semiautomated or automated procedures for PSG scoring, which would provide objective outcomes and greatly reduce variability, time, and cost of the sleep clinical trials. Thus, evaluation of agreement between the gold standard manual PSG scoring and these procedures is of paramount importance in the development of sleep drugs.

A widely used continuous measurement of sleep quality is total sleep time (TST), the total amount of time in minutes spent at different sleep stages. In a clinical study detailed in Svetnik et al.,^[17] PSG recordings from 82 subjects were collected during two nights of sleep, one under placebo, and one under drug (zolpidem, 10 mg) treatment. For each recording, different methods were used to provide the TST scores. The data from two observers of the drug arm, the semiautomatic procedure WideMedPartial versus manual, are plotted in Fig. 1.

The sample skewness of the TST from manual and WideMedPartial was -2.50 and -2.70 , respectively. The sample kurtosis of the TST from manual and WideMedPartial was 8.47 and 9.11 , respectively. These data are obviously skewed to the left and heavy tailed. To our knowledge, there was no discussion before Feng, Baumgartner, and Svetnik^[15] that focused on this type of data in the literature of the CCC (investigation of the right-skewed data was presented in Carrasco et al.^[18]). In Feng, Baumgartner, and Svetnik,^[15] the issue was addressed

using multivariate skew normal (MVSN) and multivariate skew t (MVST) distributions to accommodate various type of data structure—symmetric or skewed to the left/right and light or heavy tailed.

Another important question that relates to the estimation of CCC is the performance of non-parametric methods such as bootstrap.^[19] Can they provide accurate estimate, for example, when the data is left skewed and heavy tailed? A simulation study in Feng, Baumgartner, and Svetnik^[15] showed that, in different scenarios for bivariate skew t distributions (BVSTs), the Bayesian approach tended to provide more accurate coverage probabilities. For completeness, the results of Feng, Baumgartner, and Svetnik^[15] study on the Bayesian and bootstrap methods, respectively, are presented in Tables 1 and 2. The discrepancy between the actual and the nominal coverage from using the U-statistics as in King and Chinchilli^[20] was even larger and the corresponding results are not shown here. When data are from bivariate t -distributions, Feng, Baumgartner, and Svetnik^[13] conducted a thorough investigation on the difference between the bootstrap distribution and sample distribution and demonstrated that the bootstrap estimates tended to be biased downward. The details on Bayesian estimation (including the treatment of BVST) are presented in section “How to Execute the Bayesian Estimate of the CCC.”

Can data be transformed and then treated as normal and the CCC be estimated? The coverage probabilities for the CCC after taking the Box–Cox transformation^[21] of the left-skewed and heavy-tailed data and treating them as normally distributed (estimated by variance components method proposed by Carrasco and Jover^[22]) are shown in Table 3. The results were inferior to that from the Bayesian approach.

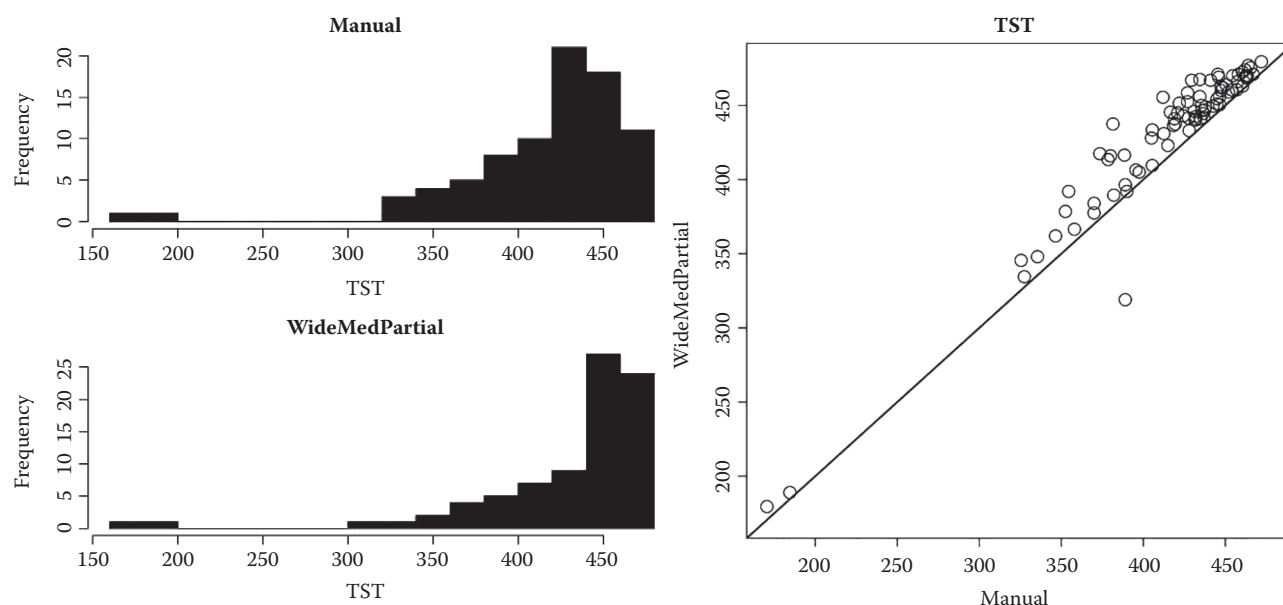


Fig. 1 Histograms and scatter plot of measurements of the TST in minutes from manual and WideMedPartial

Table 1 Coverage probabilities of the 95% credible intervals for the CCC for data simulated from BVSTs and estimated by the Bayesian method

n	δ	$\mu_1 = 450, \mu_2 = 450, \sigma_1 = 45, \sigma_2 = 50$						$\mu_1 = 450, \mu_2 = 400, \sigma_1 = 45, \sigma_2 = 45$						$\mu_1 = 450, \mu_2 = 400, \sigma_1 = 45, \sigma_2 = 50$					
		$\rho_{12} = 0.99$	$\rho_{12} = 0.9$	$\rho_{12} = 0.7$	$\rho_{12} = 0.5$	$\rho_{12} = 0.99$	$\rho_{12} = 0.9$	$\rho_{12} = 0.99$	$\rho_{12} = 0.9$	$\rho_{12} = 0.7$	$\rho_{12} = 0.5$	$\rho_{12} = 0.99$	$\rho_{12} = 0.9$	$\rho_{12} = 0.99$	$\rho_{12} = 0.9$	$\rho_{12} = 0.7$	$\rho_{12} = 0.5$	$\rho_{12} = 0.99$	$\rho_{12} = 0.9$
50	-60	0.95	0.94	0.93	0.92	0.93	0.94	0.93	0.95	0.93	0.92	0.94	0.94	0.94	0.94	0.93	0.92	0.94	0.94
	-30	0.95	0.96	0.94	0.92	0.93	0.94	0.93	0.95	0.94	0.93	0.95	0.94	0.93	0.95	0.94	0.94	0.94	0.94
	0	0.94	0.94	0.92	0.92	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
	30	0.96	0.96	0.94	0.92	0.92	0.94	0.92	0.94	0.94	0.94	0.94	0.94	0.93	0.94	0.94	0.93	0.94	0.93
80	60	0.95	0.95	0.94	0.92	0.94	0.94	0.94	0.94	0.92	0.92	0.94	0.94	0.93	0.94	0.92	0.92	0.94	0.92
	-60	0.94	0.96	0.94	0.92	0.94	0.94	0.94	0.94	0.92	0.90	0.94	0.94	0.94	0.94	0.93	0.91	0.94	0.91
	-30	0.94	0.97	0.94	0.94	0.93	0.94	0.93	0.94	0.93	0.93	0.94	0.94	0.92	0.94	0.94	0.93	0.94	0.93
	0	0.93	0.94	0.93	0.93	0.94	0.94	0.94	0.94	0.95	0.94	0.94	0.95	0.94	0.95	0.95	0.94	0.94	0.94
100	30	0.94	0.97	0.95	0.93	0.94	0.95	0.94	0.95	0.94	0.94	0.95	0.95	0.94	0.95	0.94	0.93	0.94	0.93
	60	0.95	0.96	0.95	0.92	0.93	0.94	0.93	0.95	0.93	0.91	0.94	0.94	0.94	0.94	0.93	0.91	0.94	0.91
	-60	0.94	0.96	0.95	0.93	0.94	0.94	0.94	0.94	0.92	0.91	0.94	0.94	0.94	0.95	0.93	0.91	0.94	0.91
	-30	0.93	0.96	0.94	0.94	0.93	0.94	0.93	0.94	0.93	0.93	0.93	0.94	0.92	0.94	0.94	0.94	0.94	0.94
	0	0.95	0.94	0.95	0.95	0.94	0.95	0.94	0.95	0.95	0.95	0.94	0.95	0.94	0.94	0.95	0.95	0.95	0.95
	30	0.94	0.97	0.94	0.94	0.93	0.94	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
	60	0.94	0.96	0.95	0.92	0.93	0.94	0.93	0.94	0.92	0.91	0.94	0.94	0.94	0.93	0.93	0.94	0.93	0.90

Note: The different simulation scenarios had the parameter $\nu = 5$ and $\mu_1, \mu_2, \sigma_1, \sigma_2$ with different values and various combinations of $\rho_{12}, \delta (= \delta_1 = \delta_2)$ values and sample sizes.

Table 2 Coverage probabilities of the 95% confidence intervals for the CCC for data simulated from BVSTs and estimated by the bootstrap method with percentile interval

n	δ	$\mu_1 = 450, \mu_2 = 450, \sigma_1 = 45, \sigma_2 = 50$				$\mu_1 = 450, \mu_2 = 400, \sigma_1 = 45, \sigma_2 = 45$				$\mu_1 = 450, \mu_2 = 400, \sigma_1 = 45, \sigma_2 = 50$			
		$\rho_{12} = 0.99$	$\rho_{12} = 0.9$	$\rho_{12} = 0.7$	$\rho_{12} = 0.5$	$\rho_{12} = 0.99$	$\rho_{12} = 0.9$	$\rho_{12} = 0.7$	$\rho_{12} = 0.5$	$\rho_{12} = 0.99$	$\rho_{12} = 0.9$	$\rho_{12} = 0.7$	$\rho_{12} = 0.5$
50	-60	0.89	0.91	0.92	0.92	0.92	0.93	0.92	0.90	0.93	0.93	0.92	0.91
	-30	0.84	0.90	0.92	0.92	0.91	0.90	0.91	0.91	0.90	0.91	0.91	0.91
	0	0.93	0.93	0.93	0.93	0.83	0.87	0.90	0.91	0.84	0.88	0.90	0.91
	30	0.85	0.91	0.93	0.93	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.92
80	60	0.90	0.92	0.93	0.93	0.93	0.93	0.91	0.90	0.92	0.93	0.92	0.91
	-60	0.85	0.89	0.92	0.92	0.93	0.93	0.92	0.90	0.93	0.94	0.92	0.91
	-30	0.76	0.88	0.93	0.92	0.93	0.92	0.92	0.91	0.93	0.92	0.91	0.91
	0	0.92	0.93	0.93	0.93	0.87	0.89	0.90	0.92	0.87	0.90	0.91	0.92
100	30	0.77	0.90	0.94	0.93	0.94	0.93	0.92	0.91	0.93	0.93	0.91	0.92
	60	0.86	0.91	0.94	0.94	0.93	0.94	0.93	0.91	0.93	0.94	0.93	0.91
	-60	0.82	0.87	0.93	0.92	0.93	0.93	0.93	0.91	0.93	0.93	0.92	0.91
	-30	0.70	0.87	0.92	0.93	0.92	0.91	0.91	0.92	0.92	0.92	0.91	0.92
	0	0.93	0.93	0.93	0.93	0.88	0.89	0.91	0.92	0.88	0.89	0.90	0.92
	30	0.72	0.89	0.93	0.93	0.93	0.92	0.92	0.91	0.92	0.92	0.92	0.92
	60	0.83	0.90	0.94	0.94	0.93	0.94	0.92	0.90	0.93	0.93	0.92	0.91

Note: The different simulation scenarios had the parameter $\nu = 5$ and $\mu_1, \mu_2, \sigma_1, \sigma_2$ with different values and various combinations of $\rho_{12}, \delta (= \delta_1 = \delta_2)$ values and sample sizes.

Table 3 Coverage probabilities of the 95% confidence intervals for the CCC for data simulated from BVSTs and estimated by taking the Box-Cox transformation first

n	δ	$\mu_1 = 450, \mu_2 = 450, \sigma_1 = 45, \sigma_2 = 50$						$\mu_1 = 450, \mu_2 = 400, \sigma_1 = 45, \sigma_2 = 45$						$\mu_1 = 450, \mu_2 = 400, \sigma_1 = 45, \sigma_2 = 50$					
		$\rho_{12} = 0.99$	$\rho_{12} = 0.9$	$\rho_{12} = 0.7$	$\rho_{12} = 0.5$	$\rho_{12} = 0.99$	$\rho_{12} = 0.9$	$\rho_{12} = 0.99$	$\rho_{12} = 0.9$	$\rho_{12} = 0.7$	$\rho_{12} = 0.5$	$\rho_{12} = 0.99$	$\rho_{12} = 0.9$	$\rho_{12} = 0.99$	$\rho_{12} = 0.9$	$\rho_{12} = 0.7$	$\rho_{12} = 0.5$	$\rho_{12} = 0.99$	$\rho_{12} = 0.9$
50	-60	0.85	0.89	0.90	0.92	0.90	0.91	0.90	0.91	0.92	0.91	0.90	0.92	0.90	0.92	0.92	0.91	0.90	0.92
	-30	0.80	0.88	0.90	0.89	0.86	0.88	0.86	0.88	0.88	0.89	0.86	0.89	0.86	0.89	0.88	0.89	0.86	0.88
	0	0.91	0.88	0.87	0.87	0.78	0.83	0.78	0.83	0.86	0.88	0.78	0.83	0.78	0.83	0.87	0.87	0.78	0.87
	30	0.79	0.88	0.89	0.90	0.87	0.88	0.87	0.88	0.88	0.89	0.87	0.88	0.87	0.88	0.89	0.89	0.87	0.89
80	60	0.84	0.88	0.91	0.91	0.90	0.91	0.90	0.91	0.91	0.90	0.90	0.91	0.89	0.90	0.91	0.91	0.89	0.91
	-60	0.78	0.84	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.87	0.88	0.90	0.88	0.90	0.89	0.88	0.88	0.88
	-30	0.70	0.83	0.87	0.87	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.87	0.87	0.87	0.86	0.87
	0	0.92	0.86	0.86	0.87	0.78	0.83	0.78	0.83	0.85	0.86	0.79	0.83	0.79	0.83	0.86	0.87	0.79	0.86
100	30	0.69	0.85	0.89	0.89	0.87	0.88	0.87	0.88	0.88	0.87	0.87	0.88	0.87	0.88	0.87	0.88	0.87	0.88
	60	0.77	0.86	0.91	0.91	0.90	0.91	0.90	0.91	0.90	0.89	0.90	0.91	0.90	0.90	0.90	0.89	0.90	0.89
	-60	0.75	0.83	0.90	0.90	0.89	0.90	0.89	0.90	0.90	0.87	0.89	0.90	0.89	0.90	0.89	0.88	0.89	0.88
	-30	0.62	0.82	0.87	0.87	0.86	0.87	0.86	0.87	0.87	0.86	0.86	0.87	0.87	0.87	0.87	0.87	0.87	0.87
	0	0.91	0.86	0.85	0.85	0.79	0.82	0.79	0.82	0.85	0.86	0.79	0.82	0.79	0.82	0.86	0.87	0.79	0.86
	30	0.62	0.83	0.88	0.88	0.87	0.87	0.87	0.87	0.87	0.88	0.87	0.87	0.87	0.88	0.88	0.87	0.87	0.87
	60	0.73	0.84	0.91	0.91	0.88	0.90	0.88	0.90	0.89	0.88	0.88	0.90	0.88	0.90	0.90	0.88	0.88	0.88

Note: The different simulation scenarios had the parameter $\nu = 5$ and $\mu_1, \mu_2, \sigma_1, \sigma_2$ with different values and various combinations of $\rho_{12}, \delta (= \delta_1 = \delta_2)$ values and sample sizes.

In general, the Bayesian method has advantages such as obeying the likelihood principle and having a natural interpretation of confidence intervals (known as “credible sets” or “credible intervals”). The estimate of CCC can benefit from using the Bayesian approach specifically due to its capability of providing inferential quantities of any functions of parameters directly for a variety of parametric and non-parametric models without using the plug-in method and asymptotic approximation.

More detailed comparison results between the Bayesian and frequentist methods can be found in the studies by Feng, Baumgartner, and Svetnik.^[12–15] The Bayesian approaches provided very compelling results even from a frequentist point of view such as accurate coverage probabilities. In addition, the Bayesian methods could avoid less ideal situations such as increasing number of observers, subjects, or decreasing percentage of missing data leading to worse coverage. Furthermore, practically relevant issues such as accommodation of covariates and missing data and model diagnostics and comparison, were addressed coherently within the Bayesian framework. Section “How to Execute the Bayesian Estimate of the CCC” presents the details of implementing the Bayesian approach to estimate the CCC.

HOW TO EXECUTE THE BAYESIAN ESTIMATE OF THE CCC

In this section, the Bayesian estimation for a basic setup is discussed first and then the extension to address other practically relevant issues is discussed.

Estimation for a Basic Setup

A typical setup in an agreement study is that for each subject i , we have samples $\mathbf{y}_i = (y_{i1}, \dots, y_{id})$ from d observers.

When there are d observers, the CCC is defined as:

$$CCC = \frac{2 \sum_{j=1}^{d-1} \sum_{j'=j+1}^d \sigma_{jj'}}{(d-1) \sum_{j=1}^d \sigma_j^2 + \sum_{j=1}^{d-1} \sum_{j'=j+1}^d (\mu_j - \mu_{j'})^2} \quad (1)$$

where σ_j^2 and μ_j are the variance and mean of the measurements made by observer j , respectively, and $\sigma_{jj'}$ is the covariance between measurements from observers j and j' ($j, j' = 1, \dots, d \geq 2$).

In principle, it is straightforward to conduct a Bayesian estimate of the CCC. First, given certain assumptions, a likelihood function is formulated and the corresponding priors are specified. The posterior distributions of parameters in the likelihood can then be estimated using, e.g., the Markov chain Monte Carlo (MCMC) approach. Finally, the posterior distribution of function of parameters, which is the CCC in this case, and the corresponding point estimate and credible interval can be obtained.

To formulate a likelihood and facilitate the MCMC and improve mixing, data augmentation is a very effective tool.^[23] Besides the original vector of parameters $\boldsymbol{\theta}$, some latent variables $\boldsymbol{\theta}_{\text{aug}}$ are created. With the introduction of $\boldsymbol{\theta}_{\text{aug}}$,

the likelihood $f(\mathbf{y}_i | \boldsymbol{\theta})$ equals $\int f_1(\mathbf{y}_i | \boldsymbol{\theta}_{\text{aug}}, \boldsymbol{\theta}) f_2(\boldsymbol{\theta}_{\text{aug}} | \boldsymbol{\theta}) d\boldsymbol{\theta}_{\text{aug}}$.

A Markov chain $\{(\boldsymbol{\theta}^t, \boldsymbol{\theta}_{\text{aug}}^t), t \geq 1\}$ is formed and $\boldsymbol{\theta}^t, \boldsymbol{\theta}_{\text{aug}}^t$ are drawn iteratively from the respective conditional distributions.

Feng, Baumgartner, Svetnik,^[15] proposed a general Bayesian framework to accommodate various types of data structure—symmetric or skewed to the left/right and light or heavy tailed. The proposal was based on the MVS/N/MVST distributions discussed in the study by Sahu, Dey, and Branco.^[24] The density of $\mathbf{y}_i$ having a MVST distribution with parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}$, and ν is as follows:

$$\begin{aligned} f(\mathbf{y}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \nu) &= 2^d t_{d, \nu}(\mathbf{y}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma} + \mathbf{D}^2) \times T_{d, \nu+d} \\ &\times \left[\left(\frac{\nu + q(\mathbf{y}_{i*})}{\nu + d} \right)^{-\frac{1}{2}} \right. \\ &\times (\mathbf{I} - \mathbf{D}(\boldsymbol{\Sigma} + \mathbf{D}^2)^{-1} \mathbf{D})^{-\frac{1}{2}} \\ &\times \mathbf{D}(\boldsymbol{\Sigma} + \mathbf{D}^2)^{-1} \mathbf{y}_{i*} \left. \right] \quad (2) \end{aligned}$$

where $\mathbf{y}_{i*} = \mathbf{y}_i - \boldsymbol{\mu}$; $q(\mathbf{y}_{i*}) = \mathbf{y}_{i*}^T (\boldsymbol{\Sigma} + \mathbf{D}^2)^{-1} \mathbf{y}_{i*}$; $t_{d, \nu}(\mathbf{y}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma} + \mathbf{D}^2)$ is the density of multivariate t -distribution of a d dimension vector $\mathbf{y}_i$ with location vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$, and degrees of freedom ν , respectively; $T_{d, \nu+d}(\cdot)$ denotes the cumulative distribution function (cdf) of a multivariate t -distribution for a d dimensional vector with location vector $\mathbf{0}$, scale matrix $\mathbf{I}$ (an identity matrix), and degrees of freedom $\nu + d$.

Instead of directly assigning prior distributions on parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \nu)$ based on the density function (2), Feng, Baumgartner, and Svetnik^[15] adopted the formulation in the study by Sahu, Dey, and Branco^[24] as follows:

$$\begin{aligned} \mathbf{Y}_i | \mathbf{Z}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \omega_i &\sim \text{MVN}(\boldsymbol{\mu} + \mathbf{D}\mathbf{Z}_i, \frac{1}{\omega_i} \boldsymbol{\Sigma}) \\ \mathbf{Z}_i &\sim \text{MVN}(\mathbf{0}, \mathbf{I}) \mathbf{I}(\mathbf{Z} > \mathbf{0}) \\ \omega_i | \nu &\sim \Gamma(\nu/2, \nu/2) \end{aligned} \quad (3)$$

where $\text{MVN}(\cdot)$ denotes a multivariate normal distribution, $\Gamma(\cdot)$ denotes a gamma distribution with corresponding shape and rate parameter, and $\mathbf{I}$ is an indicator function. Note that with the introduction of latent variables $\boldsymbol{\theta}_{\text{aug}} = (\mathbf{Z}_i, \omega_i)$, most of the parameters in $(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{aug}})$ have conjugate priors. The data augmentation was adopted in other proposals that used the MCMC to estimate the CCC^[13,14] as well.

Extension to Address Other Practically Relevant Issues

Practically relevant issues, including enabling incorporation of confounding covariates, handling missing data and

multiple replications (from each observer on the same subject), accessing goodness of fit of model, and conducting model comparison, can be addressed within the Bayesian framework.

For the example in Section “Motivation for Bayesian Estimate of the CCC” that studies the agreement between semiautomatic and manual procedure to measure the TST (shown in Fig. 1), there seems to be a shift of TST values depending on gender—females tended to have higher values than males (see Fig. 2 for details). How to address the case in which gender is a potential confounding covariate?

To accommodate covariates, such as gender in the previous example, the covariate effects can be added into the conditional mean of $\mathbf{Y}_i$. For example, if $\mathbf{Y}_i$ follows a MVST distribution, then considering the covariate effects, the conditional distribution of $\mathbf{Y}_i$ is as follows:

$$\mathbf{Y}_i | \mathbf{Z}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\omega}_i \sim \text{MVN}(\boldsymbol{\mu} + \mathbf{X}_i \boldsymbol{\beta} + \mathbf{DZ}_i, \frac{1}{\boldsymbol{\omega}_i} \boldsymbol{\Sigma}) \quad (4)$$

where $\mathbf{X}_i$ is the design matrix for subject i and $\boldsymbol{\beta}$ is the vector of regression coefficients.

When there are multiple replications, let Y_{ijk} be the k^{th} reading for subject i by observer j and consider a model as follows:

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} \quad (5)$$

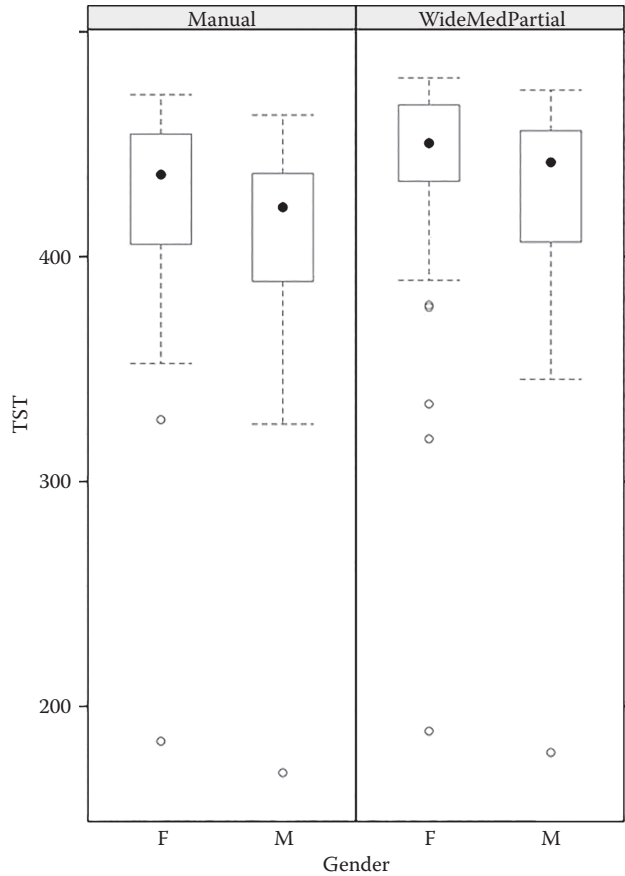


Fig. 2 Boxplots of TST by gender from two different methods: WideMedPartial and manual

In this case, besides the original CCC, the inter-CCC among observers ($\text{CCC}_{\text{inter}}$) and intra-CCC for observer j ($\text{CCC}_{j,\text{intra}}$) could be defined as follows:

$$\text{CCC}_{\text{inter}} = 1 - \frac{\sum_{j=1}^{d-1} \sum_{j'=j+1}^d E(\mu_{ij} - \mu_{ij'})^2}{\sum_{j=1}^{d-1} \sum_{j'=j+1}^d E_1(\mu_{ij} - \mu_{ij'})^2} \quad (6)$$

$$\text{CCC}_{j,\text{intra}} = 1 - \frac{\sum_{k=1}^{k-1} \sum_{k'=k+1}^K E(Y_{ijk} - Y_{ijk'})^2}{\sum_{k=1}^{k-1} \sum_{k'=k+1}^K E_1(Y_{ijk} - Y_{ijk'})^2} \quad (7)$$

where E_1 is the conditional expectation given independence of two corresponding random variables (refer the study by Barnhart, Haber, and Lin^[1] and references therein for more details). With the introduction of latent variables, the key to solving the expectations in Eqs. 6 and 7 is computing expectation by conditioning ($E(A) = E(E(A|B))$) (e.g., refer the study by Feng, Baumgartner, and Svetnik^[13]).

To address missing observations, assuming the mechanism is ignorable—missing completely at random or missing at random, missing data could be treated as additional unknown parameters to be estimated along with the other parameters of interest and are simulated in each MCMC iteration. The missing data require no change to the basic model structure (as the data were complete) and only need to be represented in the data file (as “NA” when running Bayesian inference using Gibbs sampling (BUGS)^[25] or JAGS^[26]).

To assess the goodness of fit of posited Bayesian model to data, a widely used method is the posterior predictive approach. A reference distribution (with corresponding tail-area probabilities) for any test statistic^[27] can be constructed. For example, refer the study by Feng, Baumgartner, and Svetnik,^[14] in which not only the classical χ^2 but an order statistic as discrepancy measures was also used for goodness-of-fit assessment.

To select a model from several candidates, one method is to use the deviance information criterion (DIC) as in the study by Spiegelhalter et al.^[28] DIC is a generalization of the Akaike information criterion (AIC) and is also related to Bayesian information criterion (BIC). Models with smaller DIC should be preferred to models with larger DIC. In the study by Feng, Baumgartner, and Svetnik,^[15] a model based on the MVST assumption fitted the data illustrated in Fig. 1 better than several other parametric models considered.

POTENTIAL LIMITATIONS AND FUTURE RESEARCH OF THE BAYESIAN ESTIMATE OF THE CCC

Among the proposed Bayesian estimate of the CCC, a parametric distributional assumption underpins each model. The less appropriate the assumption, the worse the results. For example, in the study by Feng, Baumgartner, and Svetnik,^[14] when the variation of the data was dominated by random error, the bootstrap outperformed

the Bayesian approach. The goodness-of-fit of each model can be checked and based on the DIC, different distributional assumptions can be compared to each other. However, there might be other parametric models beyond consideration that fits the data even better. A more accurate estimate of the CCC may be obtained through some non-parametric Bayesian approaches. This warrants further investigation given especially that non-parametric Bayes would require a relatively large sample size that is uncommon, e.g., in proof-of-concept agreement studies in early drug development.

Vague conjugate priors were typically used for the Bayesian estimate of the CCC. The Bayesian paradigm allows for a straightforward adoption of informative priors in cases when there is available prior knowledge or information related to the CCC. Using conjugate priors, the point estimates of parameters are estimated from the historical data and the hyperparameters based on these estimates and the strength of belief are specified. In an agreement study, a typical setup is that a gold standard already exists and the question is whether measurements from a “new” approach can be used interchangeably. Therefore, there is often accumulated knowledge about the gold standard but limited information on the “new” approach.

As an alternative to the simple vague prior, an “objective prior” could be a better choice in terms of both appealing statistical concept and optimal property (such as probability matching property).^[29] The performance of using Jeffreys prior in the Bayesian estimate of the CCC for normally distributed data was investigated in the study by Feng, et al.^[12] Further studies on the topic are required especially in small sample size cases.

Compared to the frequentist approach, the Bayesian estimate of the CCC may demand a higher computational cost. This limitation, however, is becoming less important. First, the iterative simulation based on conditional distributions is not always necessary to conduct a Bayesian analysis. For example, when there are only two raters, under normal assumption and using the Jeffreys prior, it is straightforward to obtain independent samples from the posteriors (without using Gibbs sampling or other computationally intensive sampling methods) and hence are fast to compute. Second, a frequentist approach may require more time than originally thought. For example, although Efron and Tibshirani suggested that 1000 resampling is enough for bootstrap confidence intervals construction,^[19] 10,000 resamples were argued to be necessary for more accuracy.^[30] Finally, with the constant advancement of computing tools and availability of powerful Bayesian sampling engines such as BUGS and JAGS, the computation time is often not a concern for the Bayesian estimate of the CCC, given the relatively simple model adopted on small size of data. Most of the proposed Bayesian methods for the CCC estimation were implemented in a R^[31] package “agRee.”^[32]

The package calls internally JAGS or BUGS to run the MCMC, when necessary.

The Bayesian estimate of the CCC has been applied to continuous data. As a topic for further research, Bayesian approach could be developed to handle categorical and ordinal data based on the Bayesian generalized linear mixed models.

REFERENCES

1. Barnhart, H.X.; Haber, M.J.; Lin, L.I. An overview on assessing agreement with continuous measurements. *J. Biopharm. Stat.* **2007**, *17* (4), 529–569.
2. Lin, L. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **1989**, *45* (1), 255–268.
3. Barnhart, H.X.; Haber, M.; Song, J. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics* **2002**, *58* (4), 1020–1027.
4. Lin, L.; Hedayat, A.; Sinha, B.; Yang, M. Statistical methods in assessing agreement: Models, issues, and tools. *J. Am. Stat. Assoc.* **2002**, *97* (457), 257–270.
5. Barnhart, H.X.; Song, J.; Haber, M.J. Assessing intra, inter and total agreement with replicated readings. *Stat. Med.* **2005**, *24* (9), 1371–1384.
6. Lin, L.; Hedayat, A.; Wu, W. A unified approach for assessing agreement for continuous and categorical data. *J. Biopharm. Stat.* **2007**, *17* (4), 629–652.
7. Chinchilli, V.M.; Martel, J.K.; Kumanyika, S.; Lloyd, T. A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics* **1996**, *52* (1), 341–353.
8. Quiroz, J. Assessment of equivalence using a concordance correlation coefficient in a repeated measurements design. *J. Biopharm. Stat.* **2005**, *15* (6), 913–928.
9. King, T.S.; Chinchilli, V.M.; Carrasco, J.L. A repeated measures concordance correlation coefficient. *Stat. Med.* **2007**, *26* (16), 3095.
10. Carrasco, J.L.; King, T.S.; Chinchilli, V.M. The concordance correlation coefficient for repeated measures estimated by variance components. *J. Biopharm. Stat.* **2009**, *19* (1), 90–105.
11. Chen, C.-C.; Barnhart, H.X. Assessing agreement with repeated measures for random observers. *Stat. Med.* **2011**, *30* (30), 3546–3559.
12. Feng, D.; Svetnik, V.; Coimbra, A.; Baumgartner, R. A comparison of confidence interval methods for the concordance correlation coefficient and intraclass correlation coefficient with small number of raters. *J. Biopharm. Stat.* **2014**, *24* (2), 272–293.
13. Feng, D.; Baumgartner, R.; Svetnik, V. A robust Bayesian estimate of the concordance correlation coefficient. *J. Biopharm. Stat.* **2015**, *25* (3), 490–507.
14. Feng, D.; Baumgartner, R.; Svetnik, V. A Bayesian estimate of the concordance correlation coefficient with skewed data. *Pharm. Stat.* **2015**, *14* (4), 350–358.
15. Feng, D.; Baumgartner, R.; Svetnik, V. A Bayesian framework for estimating the concordance correlation coefficient. 2016.
16. Rechtschaffen, A.; Kales, A. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep*

- Stages of Human Subjects*; Public Health Service, US Government Printing Office: Washington, DC, 1968.
17. Svetnik, V.; Ma, J.; Soper, K.A.; Doran, S.; Renger, J.J.; Deacon, S.; Koblan, K.S. Evaluation of automated and semi-automated scoring of polysomnographic recordings from a clinical trial using zolpidem in the treatment of insomnia. *Sleep* **2007**, *30* (11), 1562–1574.
 18. Carrasco, J.L.; Jover, L.; King, T.S.; Chinchilli, V.M. Comparison of concordance correlation coefficient estimating approaches with skewed data. *J. Biopharm. Stat.* **2007**, *17* (4), 673–684.
 19. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; CRC Press: Boca Raton, FL, 1994.
 20. King, T.S.; Chinchilli, V.M. A generalized concordance correlation coefficient for continuous and categorical data. *Stat. Med.* **2001**, *20* (14), 2131–2147.
 21. Box, G.E.; Cox, D.R. An analysis of transformations. *J. R. Stat. Soc. Ser. B.* **1964**, *26*, 211–252.
 22. Carrasco, J.L.; Jover, L. Estimating the generalized concordance correlation coefficient through variance components. *Biometrics* **2003**, *59*, 849–858.
 23. Van Dyk, D.A.; Meng, X-L. The art of data augmentation. *J. Comput. Graph. Stat.* **2001**, *10* (1), 1–50.
 24. Sahu, S.K.; Dey, D.K.; Branco, M.D. A new class of multivariate skew distributions with applications to Bayesian regression models. *Can. J. Stat.* **2003**, *31* (2), 129–150.
 25. Lunn, D.J.; Thomas, A.; Best, N.; Spiegelhalter, D. Winbugs-a Bayesian modelling framework: concepts, structure, and extensibility. *Stat. Comput.* **2000**, *10* (4), 325–337.
 26. Plummer, M. Jags: A program for analysis of Bayesian graphical models using gibbs sampling. *Proceedings of the 3rd international workshop on distributed statistical computing*, Vol. 124; Technische Universit at Wien Wien: Austria, 2003, 125 pp.
 27. Gelman, A.; Meng, X-L.; Stern, H. Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sinica* **1996**, *6*, 733–760.
 28. Spiegelhalter, D.J.; Best, N.G.; Carlin, B.P.; Van Der Linde, A. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B.* **2002**, *64* (4), 583–639.
 29. Ghosh, M. Objective priors: An introduction for frequentists. *Stat. Sci.* **2011**, *26*, 187–202.
 30. Hesterberg, T.C. What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *Am. Stat.* **2015**, *69*, 371–386.
 31. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2010. Available at: <http://www.R-project.org/> ISBN 3-900051-07-0 (accessed July 2016).
 32. Feng, D. agRee: Various Methods for Measuring Agreement, 2016, Available at: <https://CRAN.R-project.org/package=agRee> version 0.5-0 (accessed July 2016).

Bayesian Methods: Meta-Analysis

Elizabeth Stojanovski

*School of Mathematics and Physical Sciences, University of Newcastle, Callaghan,
New South Wales, Australia*

Kerrie L. Mengersen

*School of Mathematical Sciences, Queensland University of Technology, Brisbane,
Queensland, Australia*

Abstract

This entry discusses some applications of Bayesian approaches in meta-analysis, the steps involved in a meta-analysis, computer implementation of Bayesian models for meta-analysis, extensions to the simple Bayesian model, and dealing with publication bias in the Bayesian framework.

INTRODUCTION

Meta-analysis refers to the collection and subsequent statistical analysis of results from numerous studies. The purpose of a meta-analysis is to arrive at an overall conclusion about an issue of interest based on the available data. Meta-analysis has a long history, but it has enjoyed a surge of interest and corresponding rapid methodological development in the past two decades. Philosophical and practical problems surround the combination of studies with discrepancies in aims, study designs and quality, sampling frames, populations, and reported information. Other concerns include dealing with studies of small sample size or with peculiar results, inclusion of predictors at study level, methods by which results are updated as new information is available, and accounting for the impact of other biases because of publication, selection, and so on. A Bayesian approach to meta-analysis has been advocated as a way to address many of these issues. A Bayesian model describes the structural relationship between data and unknown parameters, whereby both data and parameters are considered random variables with uncertainty. The full posterior distributions of parameters borrow strength from all studies and enable direct inference regarding probabilities about expectations and, perhaps more notably in a pharmacology context, comparisons such as the probability with which subjects receiving a particular medication are better on average recovery, symptom level, or survival, than those on an alternative medication. Moreover, by their construction, these distributions are not confined to “usual” (normal) representations or based on asymptotic assumptions, and give substantially more information than single point estimates.

The article begins with a description of some applications of Bayesian approaches in meta-analysis. The steps involved in a meta-analysis are discussed next, followed

by a description of Bayesian models for meta-analysis and selection of prior distributions. Computer implementation of Bayesian models for meta-analysis next, followed by extensions to the simple Bayesian model and dealing with publication bias in the Bayesian framework.

APPLICATIONS OF BAYESIAN METHODS IN META-ANALYSIS

Meta-analysis is among the most fundamental methods used in clinical research. Integration of results from multiple studies is linked with increased power and precision of estimated effects with the potential for clarification of contentious findings, and can be of especial benefit for studies of small sample size that tend to be linked with non-statistically significant effects because of low power. Results can also aid in constructing hypotheses for future tests in situations involving conflicting results, and provide knowledge in terms of how to better design future studies.

Meta-analysis has a long history, but it has enjoyed rapid development in the past two decades due in part to the rapid growth in research literature^[1] and expanding interest in this technique in a wide range of disciplines. Moreover, meta-analysis is an integral part of evidence-based practice, which is becoming a mainstay of many professions, particularly in medicine.^[2]

Meta-analysis is often applied to studies that involve clinical trials, due in part to government requirements via the pharmaceutical benefits advisory scheme. Meta-analysis is used to examine clinical trial information and to support clinical guidelines, to compare newly developed medications with alternative medications or placebos, to assess vaccine effectiveness, and to make decisions regarding health policies.^[3] The growing popularity of

meta-analysis has engendered commensurate discussion and critical review of the general appropriateness and validity of these methods and associated inferences.^[3–7]

A Bayesian approach to meta-analysis has been advocated as a way of addressing many of these issues.^[8] A landmark paper by DuMouchel and Harris^[9] introduced a Bayesian hierarchical model to synthesize data from five study types that assess the effects of exposure to nine environmental agents on lung cancer risk in both humans and animals. The combination of differing design types has also been demonstrated by many authors, including Dominici et al.^[10] for open and closed studies that deal with different treatment regimes for migraines, described further below; and Higgins and Whitehead^[11] demonstrate methods that can be applied to studies based on randomized control trial (RCTs) with differing interventions as the control, features that cannot typically be incorporated into a meta-analysis. The process of expressing uncertainty about parameters through prior distributions has been explored by Sutton and Abrams,^[12] and methods for quantitative updating of treatment effect assessments as data are collected have been proposed by Larose and Dey,^[1] among others. Through an example involving a meta-analysis of studies concerning breast cancer therapy, Berry et al.^[12] address problems of lacking exchangeability within and among studies, of multiple variables, and of acceptance of results from meta-analyses by practitioners.

An application considered by Sun and Suresh^[13] concerned combining results from several studies that assess efficacy of an antihistaminic drug for both seasonal allergic rhinitis, otherwise known as hay-fever, and perennial allergic rhinitis, which is an allergy unrelated to time of year. Contrary to a classical meta-analysis, which assumes exchangeability of treatment benefit among studies, exchangeability in this application is assumed only between studies with the same type of rhinitis, thus enabling the treatment benefit distribution to vary with differing rhinitis types. Prior information about disease differences with regard to treatment effect was quantified and included in the model. Covariates can also be included by generalizing the model. This may reduce heterogeneity by allowing more specific therapeutic recommendations. The study reported by Dominici et al.^[10] combined results from studies of headache treatment, which were presented in numerous ways, including as continuous effects for each treatment, as effectiveness differences for pairs of treatments, or as two-by-two contingency tables for dichotomous responses. The majority of studies were small clinical trials reporting conflicting results about treatment effectiveness. A Bayesian hierarchical random effects model for meta-analysis of grouped studies was proposed to enable an explanation for heterogeneity between studies and study types.

The logic that underlies any Bayesian approach is as follows: The data enter the model via the likelihood. A likelihood, by definition, is the probability, given model parameters, of observing the data of interest. Information

about parameters, independent of the data, is expressed through prior probability distributions that represent the uncertainty about these unknown variables. Priors can be based on a review of relevant literature, on previous phases of the same study or on the researcher's own subjective beliefs. Alternatively, they can be "vague" or "uninformative," representing little or no prior opinion and allowing all information to come from the data. These components are combined using Bayes' theorem, resulting in a joint posterior probability distribution for the unknown variables. The posterior is a revision of prior beliefs regarding parameters after having observed the data.^[14] All subsequent inferences about parameters are based on the posterior distribution.

The full posterior distributions of parameters "borrow strength" from all studies^[14] and enable direct probabilistic inference about expectations and, perhaps more notably in a pharmacology context, comparisons such as the probability with which subjects receiving a certain medication are better on average symptom level than those receiving an alternative medication. As stated in section "Introduction," these distributions give substantially more information than single point estimates. For example, a single measure cannot adequately act as a summary when a parameter's posterior distribution is very skewed. The posterior probability distribution can further be used as a prior when assessing new data in the Bayesian framework.^[14, 15] Predictions from the model can be made through posterior predictive distributions, again giving full probabilistic information about these forecasts.^[16]

STEPS IN A BAYESIAN META-ANALYSIS

Regardless of the paradigm adopted, any meta-analysis involves several steps. These include defining the research question, defining the outcome, defining the studies to collect, selecting the studies, obtaining required information from studies, developing the model, determining the computational method, interpreting results, and assessing sensitivity of the model. An additional step required under the Bayesian paradigm is the expression of various kinds of uncertainty in the model and the incorporation of these via the construction of prior distributions. These steps are now discussed in more detail.

The first step, prior to undertaking the meta-analysis, is to develop a research question. Part of this will include identification of outcomes. The outcomes for each study that form the data for the meta-analysis depend on both the required inferences from the meta-analysis and the type of data. For binary data, three common outcomes are risk difference, relative risk, and odds ratio. For continuous data, the mean with the associated standard error for each study is a potential outcome. The estimated effect size may be used if this is more commonly reported or if studies are of differing designs.

It is imperative in meta-analysis to identify both published and unpublished studies applicable to the relationship of interest.^[17,18] To obtain meaningful answers from the meta-analysis, however, criteria need to be set a priori about studies that will be considered for a meta-analysis. These criteria will depend on whether only refereed or published studies will be included (because of the perceived quality and validity of the reported results) or whether all studies will be included (thus reducing potential publication bias). The search for relevant studies should include database and Internet searches, scanning references of related manuscripts and reports, and reports from funding sources and pharmaceutical companies. If primarily published studies are included then it is imperative that the potential influence of unpublished studies is tested for. Although studies are not regarded as identical in that they are sampled from the same underlying population, they should be at least comparable^[19] on outcome and design or incompatibilities should be described in the model. For each study, the agreed outcome must be either reported or able to be calculated. Most commonly, this includes an effect size estimate as well as the corresponding standard error.

Formation of Bayesian models and computation methods are described below. Importantly, these steps should be followed by a sensitivity analysis of the model construction, priors, input data, and possible biases because of publication,^[16] selection, and so on.

BAYESIAN MODELS FOR META-ANALYSIS

Simple Hierarchical Model

The following notation will be used throughout this article. The study number will be denoted by i , and the number of studies by k , where $i = 1, \dots, k$. The true underlying effect in study i is denoted by θ_i , and the vector of true effects across all studies by $\theta = (\theta_1, \dots, \theta_k)'$. The observed effect from study i is denoted by Y_i , and the vector of observed effects by $Y = (Y_1, \dots, Y_k)'$. Variation in the i th study of the summary statistic (within-study variance) is denoted by σ_i^2 , and the corresponding precision by τ_i , where $\tau_i = 1/\sigma_i^2$. If all k studies are exchangeable, a fixed-effects model assumes all the θ_i to be equal, so that all studies are random draws from a single population. A random-effects model assumes that each θ_i is a sample from a population of effects with common mean μ . In other words, $E(\theta_i) = \mu$ with an associated variance ω^2 reflecting the between-study heterogeneity. The hyperparameters, μ and ω^2 , can also be given prior distributions to reflect the degree of uncertainty in their value. This simple random-effects model is depicted in Fig. 1. Note that Fig. 1 is similar to that produced by Liermann and Hilborn,^[20] but the notation differs accordingly.

Related meta-analysis models have been developed by various authors in the last 15 years, based on early work

Global parameters	μ, τ^2	$P(\mu), P(\omega^2)$
Study-specific parameters	$\theta_1 \theta_2 \dots \theta_k$	$P(\theta_i \mu, \omega^2)$
Data	$Y_1 Y_2 \dots Y_k$	$P(y_i \theta_i \sigma_i^2)$

Fig. 1 Simple hierarchical Bayesian model with three levels of random variables: global hyperparameters μ and τ^2 representing the overall mean effect and corresponding variance, respectively, study-specific parameters θ_i and σ_i^2 , and data Y_i . Bayesian analysis seeks to discover a joint posterior distribution of θ_i and μ , given the data

by DuMouchel and Harris,^[9] DuMouchel,^[21] Carlin,^[19] Biggerstaff, Tweedie, and Mengersen,^[5] Hasselblad and McCrory^[22] and references therein, Smith, Spiegelhalter, and Thomas,^[8] Tweedie et al.,^[7] and Larose and Dey,^[1] among others. It is evident, through these examples, that Bayesian meta-analysis is developing into an accepted approach in medical research.^[5,23,24]

As a more detailed example of the case that combines study-specific observed means and corresponding variances in a random-effects model, consider the fully Bayesian model proposed by DuMouchel^[21]

$$\begin{aligned} Y | \theta, \sigma^2 &\sim N(\theta, \sigma^2 C), & \sigma^{-2} &\sim \chi^2(df_\sigma)/(df_\sigma) \\ \theta | \mu, \tau^2 &\sim N(X\mu, \tau^2 V), & \tau^{-2} &\sim \chi^2(df_\tau)/(df_\tau) \\ \mu | \tau &\sim N(0, D \rightarrow \infty) \end{aligned}$$

where the $k \times k$ observed variance–covariance matrix C describes the within-study variability and the $k \times k$ prior variance–covariance matrix V is specified a priori or given a prior distribution such as a Wishart. For independent studies, C is taken as a diagonal matrix with individual variance estimates of Y_i on the diagonal. Uncertainty about matrices C and V is described by the parameters σ^2 and τ^2 , which are centered around unity (so that the expected values of the variances of Y and θ are C and V , respectively), given chi-squared priors with degrees of freedom df_σ and df_τ , so that larger degrees of freedom indicate greater confidence in these specified values. For example, the average number of exposed cases from each study could be used as a conservative estimate of df_σ , and the number of studies could indicate the magnitude of df_τ . The notation $D \rightarrow \infty$ indicates a very large prior variance on the overall mean, reflecting almost no prior information.

This model has been examined extensively by DuMouchel and coworkers.^[9,21,25,26] Other authors have compared it against alternatives and applied it to important public health problems such as the impact of passive smoking on health.^[7] Smith, Spiegelhalter, and Thomas^[8] also adopt this model, but their use of a more general Gamma distribution to represent variance parameters indicates that the variance representation is quite influential with respect to posterior distributions of the means and resultant inferences.

Prior Distributions

A feature that is exclusive to Bayesian modeling is the specification of prior beliefs as distributions. These specifications depend on the application. If the likelihood dominates the posterior, the posterior distribution is considered invariant over a wide range of priors. If not, priors must be developed with care and detailed in any publication.

An uninformative prior distribution, or reference prior, is designed to contain no information about a parameter (at least on one scale of interest) in that the produced posterior distribution equates to the likelihood. This approach is appealing if there is no prior information, beliefs, or expectations regarding parameters or studies, if it is necessary to present posterior distributions that reflect only objective sample information even in the presence of prior beliefs, or if it is very difficult to formulate an appropriate informative prior distribution. Jeffrey's prior is based on Fisher's information of the likelihood and is the most widely used method of obtaining potential reference priors.^[27] Improper uniform priors or proper but very diffuse priors are also often adopted as uninformative. It is important, though, to acknowledge the potential effects of these priors, including identifiability of the posterior, the weak contribution of the likelihood if the sample size is small or data are missing, the representation of the prior on other scales, and so on.^[14]

An informative prior incorporates information from various sources of evidence. These sources may include other studies or a meta-analysis of similar studies, pharmacological knowledge of a new medication, information regarding a treatment from previous data of similar medications, or the opinions of experts in their field. An informative prior is also important for obtaining a good representation of between-study heterogeneity if there are only a few studies in the meta-analysis. For example, skeptical priors formalize the belief that large treatment differences are unlikely, whereas an enthusiastic prior allows a small probability of no effect. Methodology and examples in which the prior distributions are based upon formally elicited subjective beliefs are relatively rare.^[28]

Semiparametric priors may also be considered, which may better accommodate asymmetric characteristics of uncertainty.^[29] Alternatively, instead of selecting a particular specification of a prior, a community of priors may be constructed that cover differing views. Such a community may include, for example, a reference prior, intended to add as little as possible to data, and a clinical prior, expressing reasonable opinions held by individuals.^[30]

Additional priors that may be considered include conjugate priors, which simplify calculations and produce easy to understand results as an analytic form of the mean and variance. Conjugate priors for μ and τ^2 are the normal and inverse gamma (IG) distributions, respectively. Non-informative priors may be approximated by a very flat normal distribution for μ and an IG distribution with near zero parameters for τ^2 .

An example of creating priors follows: Credibility of various evidence sources may be dubious and so the contribution of each study type to the overall population estimate is questioned. A reasonable assumption is that variability among RCTs is considerably less than that among non-RCTs. This information can be implemented by assuming each study type effect to have been drawn from differing population distributions that all share a common mean μ with different variances τ_i^2 . A higher τ_i^2 for a certain study type, for example, will relocate the estimated summary effect toward the other study types.

With respect to model (1) above, DuMouchel^[21] indicated that the specified prior distributions were selected on a convenience basis so that the posterior distribution of separate study effects is a mixture of multivariate Student-*t* distributions. A non-informative prior was used for the global population effect μ , because even with a small number of studies, the combined data are relatively informative about this parameter. Other potential prior distribution forms for the scale parameters include Pareto and log Cauchy distributions on σ^{-2} and $\tau^{-2[1]}$ or a uniform distribution on the standard deviation, σ .^[8]

COMPUTATION

In meta-analysis, there are often only a few parameters of particular interest. Inferences are based on the corresponding marginal posterior distributions for these parameters, after nuisance parameters are eliminated. For particular choices of the likelihood and priors, the marginal posterior distributions can be generated analytically by integrating the joint posterior density over remaining parameters.^[1] This comprises evaluating several high-dimensional integrals that do not permit analytical solutions for models with many parameters. Moreover, for non-standard distributions or where the normalizing constants are not available, analytic solutions are again not available.

In these circumstances, one alternative is to approximate the posterior. For example, DuMouchel^[21] suggested approximating the multivariate-*t* distributions arising from model (1) by multivariate normal distributions, which can then be described through the posterior mean and covariance matrices.^[1] As a second example, Morris and Normand^[31] proposed a Pearson approximation to the posterior distribution of the average of the overall variance term.

The posterior distribution can alternatively be obtained by simulation. The main computational method is Markov Chain Monte Carlo (MCMC).^[32] This method generates simulations from conditional posterior distributions, which, under mild assumptions, represent the joint posterior distribution.^[14] Common algorithms include Gibbs, Metropolis-Hastings, slice sampling, and perfect sampling.^[32]

MCMC methods are used to perform Bayesian inference in the computer software package BUGS (Bayesian inference Using the Gibbs sampling technique), which

is available at www.mrc-bsu.cam.ac.uk/bugs.^[33] Given a user-specified model in the form of likelihood and priors, BUGS constructs the necessary posterior conditional distributions, carries out MCMC sampling, provides statistical and graphical summaries of the posterior distributions, and reports some convergence diagnostics. Methods for assessing convergence of the MCMC simulation to the required distribution have been developed and discussed by many authors; see, for example, Spiegelhalter et al.,^[16] Robert and Casella,^[32] and Cowles and Carlin.^[34]

The WinBUGS manual gives worked examples of a number of meta-analyses relevant to the health and pharmacological profession. Congdon^[15] also uses WinBUGS to illustrate a wide range of Bayesian models and analysis, including meta-analysis.

Furthermore, DuMouchel^[35] developed an Splus function to implement a fully Bayesian meta-analysis which is publicly available. The empirical Bayes model can be also implemented in MLn,^[36] which requires the global mean to be normally distributed and uses bootstrapping to obtain standard errors for the model coefficients. Specialist routines are also available for particular approaches.

EXTENSIONS

Extending the Simple Hierarchical Model

The simple model described in Eq. (1) can be extended in various directions. For example, if studies are only exchangeable within subgroups, an additional hierarchy of group-specific distributions can be inserted between the study-specific and the global levels. For example, Wolpert and Mengersen^[37] describe a meta-analysis of log odds ratios, in which conditionally independent normal distributions are used to describe study-specific true log odds ratios that are centered at a country-group-specific level. The country-group means are then generated from a normal distribution that is centered at a global log odds ratio, whose distribution in turn is normally distributed and centered at zero, with a large variance. This indicates vague prior knowledge of the overall effect size. Informative priors derived from other study information are used for the other parameters.

Alternatively or in parallel, covariates can be added at any level to explain study or group differences. For example, fixed covariates are included in model (1) through the matrix X . If multiple outcomes are of interest, a multivariate analogue of the model can be easily developed;^[38] see also the separate article on Multivariate Meta-Analysis in this volume.

An empirical Bayesian model for meta-analysis is a more limited Bayesian approach that requires less information, because hyperparameters are specified rather than given full prior distributions and nuisance parameters are not integrated out as in a fully Bayesian model but instead

are eliminated by conditioning likelihoods.^[39] For example, an empirical Bayesian version of DuMouchel's^[21] model was proposed by Carlin^[19] in which C is taken to be a diagonal matrix with known study-specific variances, and the global variance τ^2 is replaced by its observed value. This approach is shown to have parallels with the popular frequentist random effects model described by Dersimonian and Laird.^[40]

Accounting for Bias

Publication bias, also referred to as the file drawer problem, arises when decisions about manuscripts that are accepted to publish are based on the statistical significance of results. Studies with statistically significant results are more likely to be published than are studies with non-significant or inconclusive results. Publication bias may also arise if theses of students are not published, or if studies are suppressed for a variety of reasons. Publication bias may be pervasive in scientific literature and can create potentially severe distortions in meta-analysis.^[17] Sutton, Abrams, and Jones,^[41] for example, reported an assessment of 48 meta-analyses in which about half demonstrated signs of potential publication bias. An investigation of publication bias by Tweedie et al.^[7] found that as much as 45% of the observed effect size of interest in this study could be because of publication bias. Tweedie et al.^[7] recommend that sensitivity of summary point estimate and confidence intervals to possible biases be acknowledged and accounted for when possible.

Several techniques have been presented to examine publication bias impact on results of a meta-analysis. These include graphical techniques to examine the presence of missing studies visually such as the common funnel plot^[6] and similar quantitative methods.^[42,43] Bayarri and Degroot^[44] use a Bayesian framework to examine the effect of different rules for selecting papers for publication, concluding that published significant results may support a hypothesis of no effect. Larose and Dey^[45] use weight functions to describe a variety of possible mechanism of publication bias, where the weight is in proportion to the probability of study results being published. The sensitivity of parameter estimates over a range of models can then be compared.

Various approaches have been proposed to estimate numbers of missing studies.^[18,44,45] Smith, Givens, and Tweedie^[18] adopt a Bayesian hierarchical model and data augmentation to examine the possibility of publication bias and MCMC to estimate the number of missing studies. The constraint of ascribing all publication bias to significance levels is removed by Givens, Smith, and Tweedie,^[17] who developed a multitier model that allow for other reasons for publication bias such as study quality or design, sample size, or population studied.

Adjustment for bias induced by study quality has also been provided in the context of Bayesian meta-analysis.

Studies will vary in quality because of different designs and different attempts to control for known biases, confounding, and misclassification. Wolpert and Mengersen^[37] describe a range of methods for combining data in a meta-analysis, including simple pooling, fixed and random effects, and exchangeable and partially exchangeable models. They then identify approaches for accounting for study quality, including threshold exclusion in which unfit studies are simply excluded from the analysis, likelihoods are weighted according to study quality, block mixtures are utilized in which studies are grouped according to quality and successively added to the analysis, and hierarchical models are used whereby separate estimation of treatment effects is performed for homogeneous studies. The authors propose an alternative approach that adjusts the likelihood directly for the specific quality issues and give a detailed illustration of the method by accounting for three types of misclassification in a hierarchical Bayesian model.

CONCLUSIONS

Bayesian approaches to meta-analysis are developing into accepted tools of analyses in medical literature. Bayesian approaches to meta-analysis offer more flexibility in terms of combining studies with differing aims, study characteristics, sampling frames, and reported information than the corresponding classical techniques. Different forms of uncertainty can be incorporated into the modeling process via the prior distributions, which can be based on various sources. The choice of prior distribution is dependent on the specific prior beliefs or expectations on the unknown quantities or parameters. Inferences are based on posterior distributions, which are commonly generated using MCMC methods. These methods can be implemented with the software package WinBUGS for which documentation is widely available. Furthermore, several methods have also been proposed for assessing and dealing with publication bias in the Bayesian framework, the choice of which is dependent on the particular source of bias.

REFERENCES

1. Larose, D.T.; Dey, D.K. Grouped random effects models for Bayesian meta-analysis. *Stat. Med.* **1997**, *16*, 1817–1829.
2. Sutton, A.J.; Abrams, K.R. Bayesian methods in meta-analysis and evidence synthesis. *Stat. Methods Med. Res.* **2001**, *10*, 277–303.
3. Stangl, D.K.; Berry, D.A. *Meta-analysis in Medicine and Health Policy*; Marcel Dekker: New York, 2000.
4. NRC Committee on Applied and Theoretical Statistics. In *Combining Information: Statistical Issues and Opportunities for Research*; National Academy Press: Washington, DC, 1992.
5. Biggerstaff, B.J.; Tweedie, R.L.; Mengersen, K.L. Passive smoking in the workplace: classical and Bayesian meta-analyses. *Int. Arch. Occup. Environ. Health.* **1994**, *66*, 269–277.
6. Mengersen, K.; Tweedie, R.L.; Biggerstaff, B. The impact of method choice in meta-analysis. *Aust. J. Stat.* **1995**, *37*, 19–44.
7. Tweedie, R.L.; Scott, D.J.; Biggerstaff, B.J.; Mengersen, K.L. Bayesian meta-analysis, with application to studies of ETS and lung cancer. *Lung Cancer.* **1996**, *14* (Suppl. 1), S171–S194.
8. Smith, T.C.; Spiegelhalter, D.J.; Thomas, A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Stat. Med.* **1995**, *14*, 2685–2699.
9. DuMouchel, W.H.; Harris, J.E. Bayesian methods for combining the results of cancer studies in humans and other species. *J. Am. Stat. Assoc.* **1983**, *78*, 293–308.
10. Dominici, F.; Parmigiani, G.; Wolpert, R.L.; Hasselblad, V. Meta-analysis of migraine headache treatments: combining information from heterogeneous designs. *J. Am. Stat. Assoc.* **1999**, *94*, 16–28.
11. Higgins, J.P.T.; Whitehead, A. Borrowing strength from external trials in a meta-analysis. *Stat. Med.* **1996**, *15*, 2733–2749.
12. Berry, D.A.; Thor, A.; Cirrincione, C.; Edgerton, S.; Muss, H.; Marks, J. Scientific inference and predictions: multiplicities and convincing stories: a case study in breast cancer therapy. In *Bayesian Statistics 5*; Bernardo, J.; Berger, J.O., Lindley, D.V., Eds.; Smith, A.F.M. Oxford University Press: Oxford, 1996, 45–67.
13. Sun, S.; Suresh, R. Evaluating data from closely related clinical trials. *Drug Inform. J.* **2001**, *35*, 1317–1327.
14. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*; Chapman and Hall: London, 1995.
15. Congdon, P. *Bayesian Statistical Modelling*; John Wiley & Sons: Chichester, 2001.
16. Spiegelhalter, D.J.; Best, N.G. Bayesian methods for evidence synthesis and complex cost-effectiveness models: an example in hip prostheses. *Stat. Med.* **2003**, *22*, 3687–3709.
17. Givens, G.; Smith, D.D.; Tweedie, R.L. Publication bias in meta-analysis: a Bayesian data augmentation approach to account for issues exemplified in the passive smoking debate. *Stat. Sci.* **1997**, *12*, 221–250.
18. Smith, D.D.; Givens, G.H.; Tweedie, R.L. *Adjustment for Publication Bias and Quality Bias in Bayesian Meta-Analysis. Meta-Analysis in Medicine and Health Policy*; Marcel Dekker: New York, 2000.
19. Carlin, J.B. Meta-analysis for 2×2 tables: a Bayesian approach. *Stat. Med.* **1992**, *11*, 141–158.
20. Liermann, M.; Hilborn, R. Depensation in fish stocks: a hierarchic Bayesian meta-analysis. *Can. J. Fisheries Aquatic Sci.* **1997**, *54*, 1976–1984.
21. DuMouchel, W. Bayesian meta analysis. In *Statistical Methodology in the Pharmaceutical Sciences*; Berry, D.A., Ed.; Marcel Dekker: New York, 1990; 509–529.
22. Hasselblad, V.; McCrory, D.C. Meta-analytic tools for medical decision-making—a practical guide. *Med. Decision Making.* **1995**, *15*, 81–96.
23. Berry, D.A.; Berry, S.M.; McKellar, J.; Pearson, T.A. Comparison of the dose-response relationships of 2 lipid-lowering agents: a Bayesian meta-analysis. *Am. Heart J.* **2003**, *145*, 1036–1045.

24. Berry, S.M. Meta-analysis versus large trials: resolving the controversies. In *Meta-analysis in Medicine and Health Policy*; Stangl, D.K., Berry, D.A., Eds.; Marcel Dekker: New York, 2000, 65–81.
25. DuMouchel, W.; Waternaux, C. “Discussion of “Hierarchical models for combining information and for meta-analyses. In *Bayesian Statistics 4*; Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M., Eds.; Clarendon Press: Oxford, 1992; 338–341.
26. DuMouchel, W.; Berry, D.A. Meta-analysis for dose-response models. *Stat. Med.* **1995**, *14*, 679–685.
27. Box, G.; Tiao, G. *Bayesian Inference in Statistical Analysis*; Wiley Classics Library Ed.; Wiley: Hoboken, NJ, 1992.
28. Cooper, N.; Abrams, K.; Sutton, A.; Turner, D.; Lambert, P. A Bayesian approach to Markov modelling in cost-effectiveness analyses: application to taxane use in advanced breast cancer. *J. Roy. Stat. Soc. (A)* **2003**, *166*, 389–405.
29. Burr, D.; Doss, H.; Cooke, G.E.; Goldschmidt-Clermont, P.J. A meta-analysis of studies on the association of the platelet P1A polymorphism of glycoprotein IIIa and risk of coronary heart disease. *Stat. Med.* **2003**, *22* (10), 1741–1760.
30. Spiegelhalter, D.J.; Parmar, M.K.B.; Freedman, L.S. *Bayesian Approaches to Randomised Trials*; Read before the Royal Statistical Society at a meeting organised by the Medical Section; Feb. 16, 1994.
31. Morris, C.N.; Normand, S.-L. Hierarchical models for combining information and for meta-analysis. In *Bayesian statistics 4, Proceedings of the Fourth Valencia International Meeting*; Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M., Eds.; Clarendon Press: Oxford, 1992; 321–344.
32. Robert, C.P.; Casella, G. *Monte Carlo Statistical Methods*, 2nd Ed.; Springer-Verlag: New York, 2004.
33. Spiegelhalter, D.J.; Thomas, A.; Best, N.G.; Gilks, W.R. *BUGS: Bayesian Inference using Gibbs Sampling, Version 0.5; (Version ii)*; MRC Biostatistics Unit: Cambridge, 1996.
34. Cowles, M.K.; Carlin, B.P. Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Am. Stat. Assoc.* **1996**, *91*, 883–904.
35. DuMouchel, W.H. Predictive cross-validation of Bayesian meta-analysis. In *Bayesian Statistics 5*; Oxford University Press: Oxford, 1996, 107–128.
36. Woodhouse, G., Eds; *A Guide to MLn for New Users*; University of London, Institute of Education: London, 1995.
37. Wolpert, R.L.; Mengersen, K. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. *Stat. Sci.* **2004**, *19*, 450–471.
38. Nam, I.; Mengersen, K.; Garthwaite, P. Multivariate meta-analysis. *Stat. Med.* **2003**, *22*, 2309–2333.
39. Liao, J.L. A hierarchical Bayesian model for combining multiple 2×2 tables using conditional likelihoods. *Biometrics* **1999**, *55*, 268–272.
40. Dersimonian, R.; Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials.* **1986**, *7*, 177–188.
41. Sutton, A.J.; Abrams, K.R.; Jones, D.R. An illustrated guide to the methods of meta-analysis. *J. Eval. Clin. Practice.* **2000**, *7*, 135–148.
42. Duval, S.; Tweedie, R.L. Trim and Fill: a simple funnel plot based methods of testing and adjusting for publication bias in meta-analysis. *Biometrics.* **2000**, *56*, 276–284.
43. Duval, S.; Tweedie, R.L. A nonparametric “Trim and Fill” method for assessing publication bias in meta-analysis. *J. Am. Stat. Assoc.* **2000**, *95*, 89–98.
44. Bayarri, M.J.; Degroot, M.H. Optimal reporting of predictions. *J. Am. Stat. Assoc.* **1991**, *84*, 214–222.
45. Larose, D.T.; Dey, D.K. Modelling publication bias using weighted distributions in a Bayesian framework. *Comput. Stat. Data Anal.* **1998**, *26*, 279–302.

Bayesian Methods: Stability Analysis

Jie Chen

Abbott Laboratories, Abbott Park, Illinois, U.S.A.

Jinglin Zhong

Division of Biometrics IV, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Lei Nie

Division of Biometrics VI, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

This entry aims to illustrate how a Bayesian approach to drug stability analysis naturally presents itself as a fundamentally legitimate way to provide a common ground in that it assumes some connection among batches, takes into account the batch-to-batch variation by defining appropriate prior distributions, and makes the use of prior knowledge on population parameters.

INTRODUCTION

Shelf-life or expiration dating period is one of the key components of biopharmaceutical products. The U.S. Food and Drug Administration (FDA) requires that the shelf-life, defined as the time interval during which the potency of a product is expected to remain within a pre-defined specification limit after manufacture, be indicated on the immediate container label. A formal stability study is usually conducted, as part of the new drug application (NDA) or investigational new drug (IND) to establish an appropriate shelf-life, or as post-marketing monitoring of product labeled shelf-life. Regulatory guidelines^[1,2] stipulate that at least three batches, preferably more, be tested to allow for some estimate of batch-to-batch variation and to provide the justification on which a single shelf-life for all batches is based.

Current practice for determination of shelf-life is to employ the linear fixed-effect model, or simply the analysis of covariance (ANCOVA) method to fit the stability data. The shelf-life is determined as the time point at which the lower one-sided confidence limit on the predicted mean intersects the pre-determined specification limit or minimum effective dose. To estimate the shelf-life for all batches, the ICH^[3] guideline suggests two-step statistical tests to determine whether the regression lines from different batches share a common slope and/or a common intercept given a common slope. Each of the tests uses a significance level of 0.25 to compensate for the expected low power due to the relatively small sample size. There have been some alternative offers in literature pertaining the poolability of batches, e.g., Ruberg and Stegeman,^[4] Ruberg and Hsu,^[5] Chen and Tsong,^[6] Tsong

et al.,^[7] and Liu et al.,^[8] where either a sequential hypothesis test or confidence interval approach is suggested. After the sequential tests, one ends up with one of the following decisions in terms of poolability: (i) individual slopes with individual intercepts, (ii) a common slope with individual intercepts, and (iii) a common slope with a common intercept. If decision (i) or (ii) is reached, one estimates individual shelf-lives by applying respective parameter estimates and pooled mean squared error, and chooses the shortest one as the shelf-life for all batches; if decision (iii) is achieved, one simply pools all the data together to obtain a common slope, a common intercept and a single shelf-life estimate for all batches.

Some authors have argued that the batches under study are sampled from all batches manufactured under similar circumstance, and thus it would be more appropriate to consider the batches as correlated and the batch effects as random. Chow and Shao^[9,10] are among the first who studied batch-to-batch variation in drug stability study and proposed a random-effect model for shelf-life estimation by assuming random slopes and random intercepts from batches. Some recent developments on random-effect models for drug stability studies include Shao and Chow^[11] and Chen et al.^[12] However, as pointed out in Chen et al.,^[13] the random-effect models may be more appropriate in some situation such as post-marketing stability studies where a large number of batches are tested; there are at least two drawbacks to this approach. First, a small number of batches tested in the IND and NDA stability studies may result in unreliable estimates of random-batch effect. Second, as a frequentist method, it does not utilize any prior information that might have been available on parameters. Following the ideas in Chen et al.,^[13] this entry aims to

illustrate how a Bayesian approach to drug stability analysis naturally presents itself as a fundamentally legitimate way to provide a common ground in that it assumes some connection among batches, takes into account the batch-to-batch variation by defining appropriate prior distributions, and makes the use of prior knowledge on population parameters. With available software packages such as WinBUGS and R, one can easily implement the computational algorithm of Bayesian methods in determining the poolability of batches and posterior shelf-life estimation of drug products.

PRELIMINARIES AND NOTATIONS

Suppose that k batches are taken in a stability study. Let y_{ij} be the observed value of the j th assay result (percent of label claim) from the i th batch, then the following model can be assumed:

$$y_{ij} = \alpha_i + \beta_i(x_{ij} - \bar{x}_i) + \varepsilon_{ij}, \quad i = 1, \dots, k, j = 1, \dots, n_i, \quad (1)$$

where α_i and β_i are, respectively, the unknown individual regression intercept and slope for batch i , x_{ij} is the known time point determined at the design stage, $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$, and ε_{ij} are iid $N(0, \sigma^2)$ corresponding to the random error of y_{ij} . Here the i th regression line is shifted towards $\bar{x}_i$; the “centering” of x_{ij} values about their mean leads to algebraic simplifications and better computational and statistical properties of the estimates.^[14] Eq. 1 is called the linear fixed-effect model if α_i and β_i are considered as fixed effects.

Equality of individual slopes in Eq. 1 can be formally checked by performing the null hypothesis test $H_\beta: \beta_1 = \dots = \beta_k$ against $\bar{H}_\beta$: at least one β_i is not equal to the others. If H_β is rejected, then the test stops and Eq. 1 with individual slopes and individual intercepts is assumed; if H_β is not rejected, then the assumption of a common intercept for the k batches is examined by testing the null hypothesis $H_\alpha: \alpha_1 = \dots = \alpha_k$ against $\bar{H}_\alpha$: at least one α_i is not equal to the rest, given a common slope β . If H_α is rejected, then Eq. 1 with β_i being replaced by a common slope β is taken; if H_α is not rejected, then Eq. 1 with α_i being replaced by a common intercept α and β_i by β is assumed. These two tests can easily be accomplished by ANCOVA. Given the decision on whether or not the batches are poolable, one calculates the corresponding intercept(s), slope(s) as well as the (pooled) mean squared error, and then estimate the shelf-life that applies to all batches according to the so-called minimum approach.^[3] That is, if the batches are not completely poolable (decisions (a) and (b)), the shelf-lives for individual batches are calculated and the shortest shelf life is chosen as the shelf-life for all batches; on the other hand, if the batches are completely poolable (decision (c)), then a single shelf life is calculated as the

shelf life for all batches. For technical details of current practice on the poolability tests and shelf-life estimation of frequentist methods, either fixed effect or random-effect model, the reader is referred to Chow and Liu^[15] and Chow.^[16]

A BAYESIAN HIERARCHICAL APPROACH

A Bayesian approach relies on the specification of hierarchical structure of Eq. 1 being the data model. Given the data and a set of prior distributions, one proceeds with poolability tests using Bayes factors (or other Bayes means for hypothesis test) and then shelf-life estimation using posterior distributions of relevant parameters.

The Hierarchical Model

This subsection describes how the prior distributions can easily be defined and what points should be taken into account in specifying the priors. Although there are many different ways to specify the hierarchical models, what we offer below represents our prior believes on the process parameters. In all cases, the prior specification should totally be up to one's prior knowledge on drug development steps, such as manufacturing release, drug potency change over time, and other parameters in the study.

It is fairly reasonable to assume that, conditional on some hyper-parameter λ , $\alpha = \{\alpha_1, \dots, \alpha_k\}$, $\beta = \{\beta_1, \dots, \beta_k\}$ and σ^2 are mutually independent. The independence between α and β is practically valid, especially under Eq. 1 in which the maximum likelihood estimates of α_i and β_i are uncorrelated.

Given the above assumption, the prescription for the prior densities of α_i and β_i follows the hierarchical structure as given in Berger,^[17] (section 4.6), and hence consists of two stages: a distribution for α_i and β_i , $i = 1, \dots, k$, given $\lambda = (\mu, \xi, \sigma^2)$ and a distribution for $\lambda = (\mu, \xi, \sigma^2)$, where $\mu = (\mu_\alpha, \mu_\beta)$ and $\xi = (\xi_\alpha, \xi_\beta)$. Specifically, for the first stage prior $\pi_1(\alpha_i, \beta_i | \lambda)$, we take

$$\pi_1(\alpha_i, \beta_i | \lambda) = \pi_{11}(\alpha_i | \mu_\alpha, \xi_\alpha, \sigma^2) \pi_{12}(\beta_i | \mu_\beta, \xi_\beta, \sigma^2), \quad i = 1, \dots, k, \quad (2)$$

with

$$\pi_{11}(\alpha_i | \mu_\alpha, \xi_\alpha, \sigma^2) = N(\mu_\alpha, \xi_\alpha \sigma^2)$$

and

$$\pi_{12}(\beta_i | \mu_\beta, \xi_\beta, \sigma^2) = N(\mu_\beta, \xi_\beta \sigma^2).$$

Let the second stage prior be

$$\pi_2(\mu, \xi, \sigma^2) = \pi_{211}(\mu_\alpha) \pi_{212}(\mu_\beta) \pi_{221}(\xi_\alpha) \pi_{222}(\xi_\beta) \pi_{223}(\sigma^2), \quad (3)$$

with

$$\begin{aligned}\pi_{211}(\mu_\alpha) &= 1, \\ \pi_{212}(\mu_\beta) &= U(\theta, 0), \\ \pi_{221}(\xi_\alpha) &= \pi_\alpha I(\xi_\alpha = 0) + (1 - \pi_\alpha) \pi_{221}^*(\xi_\alpha) I(\xi_\alpha > 0), \\ \pi_{222}(\xi_\beta) &= \pi_\beta I(\xi_\beta = 0) + (1 - \pi_\beta) \pi_{222}^*(\xi_\beta) I(\xi_\beta > 0), \\ \pi_{223}(\sigma^2) &= \sigma^{-2}\end{aligned}$$

where I is an indicator function of respective parameters, π_α and π_β are some subjectively chosen constants, and $\pi_{221}^*(\xi_\alpha)$ and $\pi_{222}^*(\xi_\beta)$ are arbitrary. Hence, the hyperparameter μ_α is given a non-informative prior and μ_β a uniform distribution in $(\theta, 0)$, a class of constrained non-informative priors. The constraint $\theta < 0$ forces the degradation rate to be negative *a priori*, indicating a prior believe on non-increasing trend of potency over time. In addition, σ^2 is also given a non-informative prior which is a limit of proper conditionally-conjugate inverse-gamma $IG(a, b)$ density with $a, b \rightarrow 0$, a commonly used class of priors for normal variance.

The prior parameters π_α and π_β can be interpreted as the prior probabilities of the null hypotheses $H_\alpha : \alpha_1 = \dots = \alpha_k$ and $H_\beta : \beta_1 = \dots = \beta_k$ being true, respectively, because π_α is the probability for $\sigma_\alpha^2 = \xi_\alpha \sigma^2 = 0$ (hence H_α holds) and π_β for $\sigma_\beta^2 = \xi_\beta \sigma^2 = 0$ (hence H_β holds). The most common use of these special forms of mixture prior distributions on σ_α^2 and σ_β^2 is to test the equality of parameters, such as normal means, from multiple populations. A natural choice for π_α and π_β would conventionally be non-informative $\pi_\alpha = \pi_\beta = 1/2$ unless otherwise some subjective prior belief on H_α or H_β is available.

Although $\pi_{221}^*(\xi_\alpha)$ and $\pi_{222}^*(\xi_\beta)$ could also be non-informative, they are usually chosen to be informative, especially in testing situation. The default prior $\pi_{221}^*(\xi_\alpha)$ that we adopt in this testing problem is an inverse gamma distribution $IG(1/2, 1/2)$. The advantage for this prior is that the resulting conditional distribution $\pi(\alpha | \sigma^2)$ of α , when $k = 1$, is Cauchy (μ_α, σ^2) , an analysis recommended by.^[18] With a similar argument, we take $\pi_{222}^*(\xi_\beta) = IG(1/2, 1/2)$.

Testing for Poolability of Batches

Testing the equality of slopes and equality of intercepts given a common slope can be approached in different ways in the Bayesian context, such as F test^[19] and model selection using pairwise Bayes factors between models.^[20] However, the specification of mixture priors for the variance parameters ξ_α and ξ_β allows us to pursue the poolability tests through Bayes factors for prior variances being equal to 0 versus that they are greater than 0.

Testing Equality of Slopes With Individual Intercepts

Suppose that given individual intercepts, we want to test $H_\beta : \beta_1 = \dots = \beta_k$ against $\bar{H}_\beta : \text{not all the } \beta_i \text{'s are equal}$. This

is a typical analysis of variance problem and is equivalent to testing the null hypothesis $H_{\xi_\beta} : \xi_\beta = 0$ against $\bar{H}_{\xi_\beta} : \xi_\beta \neq 0$. Following Berger^[17] (pp. 149–150) and Berger and Deely,^[21] we first define

$$L(\xi_\beta) = \int f(y | \alpha, \beta, \sigma^2) \pi_1(\alpha, \beta | \mu, \xi, \sigma^2) \times \pi_2(\mu, \xi, \sigma^2) d\alpha d\beta d\mu d\sigma^2 d\xi_\alpha \quad (4)$$

as the likelihood of ξ_β and

$$K_\beta(y) = \int L(\xi_\beta) \pi_{222}^*(\xi_\beta) d\xi_\beta \quad (5)$$

as the marginal p.d.f. of y with respect to $\pi_{222}^*(\xi_\beta)$. Then, the Bayes factor (B) for testing H_{ξ_β} (or H_β) against $\bar{H}_{\xi_\beta}$ (or $\bar{H}_\beta$) is given by

$$B_\beta = \frac{L(\xi_\beta = 0)}{K_\beta(y)} \quad (6)$$

The resulting Bayes factor B_β is compared with a constant c . If $B_\beta < c$, then H_β is rejected; otherwise H_β is retained. Although different authors^[22–24] have suggested using different constants for rejecting the null hypothesis, we will, adopt the principle of Berger et al.^[23] and use 1 for the constant c . This makes sense because a Bayes factor less than 1 would simply imply that the null model is less likely to be true as compared with the alternative model.

Testing Equality of Intercepts With a Common Slope

If H_{ξ_β} is not rejected, then the degradation rates for all batches are considered to be equal. Given this hypothesis being true, we now want to know whether the initial potencies at time 0 are equal for all batches by testing the null hypothesis $H_\alpha : \alpha_1 = \dots = \alpha_k$ against $\bar{H}_\alpha : \text{not all the } \alpha_i \text{'s are equal}$. Similar to testing H_β , this can be accomplished by testing the null hypothesis $H_{\xi_\alpha} : \xi_\alpha = 0$ against $\bar{H}_{\xi_\alpha} : \xi_\alpha \neq 0$, conditional on $\xi_\beta = 0$. The Bayes factor for testing H_{ξ_α} against $\bar{H}_{\xi_\alpha}$ is given by

$$B_\alpha = \frac{L(\xi_\beta = 0, \xi_\alpha = 0)}{K_\alpha(y)}, \quad (7)$$

where $L(\xi_\beta = 0, \xi_\alpha = 0)$ is the likelihood of $\xi = (\xi_\alpha, \xi_\beta)$ evaluated at $\xi_\alpha = 0$ and $\xi_\beta = 0$, and

$$K_\alpha(y) = \int L(\xi_\alpha, \xi_\beta = 0) \pi_{221}^*(\xi_\alpha) d\xi_\alpha$$

is the marginal p.d.f. of y with respect to $\pi_{221}^*(\xi_\alpha)$ conditional on $\xi_\beta = 0$. Again the decision on rejection or acceptance of H_{ξ_α} is made by comparing the resulting Bayes factor B_α with c ; if $B_\alpha < c$ then H_{ξ_α} is rejected; otherwise H_{ξ_α} is accepted.

Shelf-Life Estimation

Instead of using confidence limits as in frequentist methods, a Bayesian employs credible limits on the mean predicted values that are based on the posterior distributions of regression parameters to estimate drug shelf-life, denoted by y^B . In this subsection we present some necessary formulas for calculating y^B for various situations corresponding to the decisions of poolability of batches resulted from the hypothesis testing, as discussed in the previous subsection.

Suppose that the Bayesian test of $H_{\xi\beta}$ rejects the null hypothesis of equality of slopes among batches, then decision (a) is reached and shelf-lives are calculated separately for individual batches using the credible limits on individual regression lines. It is straightforward to see that conditional on (y, μ, ξ, σ^2) , α_i and $\beta_i, i = 1, \dots, k$, are independently and normally distributed with $\alpha_i \sim N(\hat{\alpha}_i^B, \sigma^2 v_{\alpha i})$ and $\beta_i \sim N(\hat{\beta}_i^B, \sigma^2 v_{\beta i})$, where $\hat{\alpha}_i^B$ and $\hat{\beta}_i^B$ are respective posterior estimates of α_i and β_i , and $v_{\alpha i}$ and $v_{\beta i}$ are posterior estimates of their variance factors. For any given x' , the $100(1-\gamma)\%$ one-sided lower credible limit for the predicted mean potency $p_i^B = \alpha_i^B + \beta_i^B(x' - \bar{x}_i)$ is given by

$$\ell_i^B(x') = \hat{p}_i^B - z_{1-\gamma} \sigma \sqrt{v_{\alpha i} + v_{\beta i}(x' - \bar{x}_i)^2}, \quad i = 1, \dots, k, \quad (8)$$

where $\hat{p}_i^B = \hat{\alpha}_i^B + \hat{\beta}_i^B(x' - \bar{x}_i)$, $z_{1-\gamma}$ is the $100(1-\gamma)\%$ th percentile of the standard normal distribution. Consequently, the shelf-life for all batches is given by

$$\psi^B = \inf\{x_i \geq 0 : l_i^B(x') \leq \eta\}, \quad (9)$$

the smallest x' that satisfies $l_i^B(x') = \eta$

Similarly, if the Bayesian test of $H_{\xi\beta}$ fails to reject the null hypothesis of equality of slopes and the test of $H_{\xi\alpha}$ concludes inequality of intercepts, then decision (b) is made and the credible limits for individual batches are obtained using the posterior distributions of the common slope β^B and individual intercepts α_i^B . Some algebra can show that conditional on (y, μ, ξ, σ^2) , α_i and β are independent and normally distributed with $\alpha_i \sim N(\hat{\alpha}_i^B, \sigma^2 v_{\alpha i}), i = 1, K, k$, and $\beta \sim N(\hat{\beta}^B, \sigma^2 v_{\beta})$, where $\hat{\beta}^B$ is the posterior estimate of the common slope β and v_{β} the posterior estimate of its variance factor. Hence, the $100(1-\gamma)\%$ one-sided lower

credible limit for the posterior mean predicted potency $p_{\alpha i}^B = \alpha_i + \beta(x' - \bar{x}_i)$ is obtained as

$$l_i^B(x') = \hat{p}_{\alpha i}^B - z_{1-\gamma} \sigma \sqrt{v_{\alpha i} + v_{\beta i}(x' - \bar{x}_i)^2}, \quad i = 1, K, k, \quad (10)$$

where $\hat{p}_i^B = \hat{\alpha}_i^B + \hat{\beta}^B(x' - \bar{x}_i)$, and $\hat{\alpha}_i^B$ and $v_{\alpha i}$ are as defined in (8). The shelf-life for all batches is determined using the minimum approach as defined in (9).

Finally, if both tests of $H_{\xi\beta}$ and $H_{\xi\alpha}$ fail to reject the respective null hypotheses of equality of slopes and equality of intercepts, then decision (c) is reached and the data of all batches are pooled to obtain the posterior estimates of the common slope β and the common intercept α . Conditional on (y, μ, ξ, σ^2) , α and β are independent and $\alpha \sim N(\hat{\alpha}^B, \sigma^2 v_{\alpha})$, where $\hat{\alpha}^B$ is the posterior estimate of the common intercept α and v_{α} the posterior estimate of its variance factor. The $100(1-\gamma)\%$ lower credible limit for the posterior mean predicted potency $p^B = \alpha^B + \beta^B(x' - \bar{x})$ is

$$l^B(x') = \hat{p}^B - z_{1-\gamma} \sigma \sqrt{v_{\alpha} + v_{\beta}(x' - \bar{x})^2}, \quad (11)$$

where $\hat{p}^B = \hat{\alpha}^B + \hat{\beta}^B(x' - \bar{x})$, and $\hat{\beta}^B$ and v_{β} are as given in (10). It should be pointed out that for simplicity we have assumed a known value of σ^2 and used normal approximation to obtain the credible limits. In reality, σ^2 must be estimated and applied to the above estimation, and the normal percentile z score should be replaced by the percentile of the t distribution with appropriate degrees of freedom.^[19] This may be more important for studies with a small sample size such as NDA studies.

Obviously there are no closed forms of the Bayes factors and the unconditional posterior distributions of parameters of interest. The Bayes factors and shelf-lives have to be therefore estimated through Markov chain Monte Carlo (MCMC) methods. Consequently, all computations can be carried out in R and WinBUGS. Readers who are interested in trying other software packages are referred to WinBUGS. Readers who are interested in trying other software packages are referred to Chen et al.^[13] for the complete conditional posterior distributions of all parameters involved in the MCMC under various combinations of the null hypotheses for the hierarchical setting discussed above.

Table 1 NDA Stability Study: Potency Assay Results (Percent of Claim)^[9]

Batch	Time (month)						
	0	3	6	9	12	18	24
1	101	97	96	95	95	94	92
2	102	100	99	100	102	102	100
3	100	104	96	98	97	96	99

AN EXAMPLE

We illustrate the proposed Bayesian approach through an example of an NDA stability study. The data set, taken from Chow and Shao,^[9] consists of three batches with each tested at time points 0, 3, 6, 9, 12, 18, 24 and 36 months (Table 1). We first applied the ANCOVA and the likelihood ratio test, as discussed in Chen et al.,^[13] to test the equality of slopes and obtained the p -values being 0.0561 and 0.0380, respectively, which exceed the ICH guideline's suggested cutoff p -value of 0.25. Hence, we employed the respective linear fixed-effect model with individual slopes and individual intercepts and the linear random-effect model with both coefficients being random to estimate the shelf-lives for all future batches. Such estimated shelf-lives for $\eta = 90$ are 28.96 months by the linear fixed-effect model and 25.77 months by the random-effect model.

We then applied the proposed Bayesian hierarchical approach to the example data set with the same hierarchical prior structure as prescribed in section "A Bayesian Hierarchical Approach" and $\theta = -0.3$, we obtained a Bayes factor $B_\beta = 0.0160$ for testing the null hypothesis of equal slopes which leads to the rejection of the null hypothesis $H_{\beta\beta}$. Accordingly, the shelf-lives for the three batches are calculated separately based on the posterior densities of (α_i, β_i) and the minimum of such estimated shelf-lives is 29.53 months.

CONCLUDING REMARKS

Bayesian hierarchical modeling approach has abroad application in a variety of fields; it might be even more appropriate in the biopharmaceutical industry where heavy regulations and frequent testing take place throughout product development. The rationale behind this is that some prior information can easily be derived from regulation parameters and many specifications on product release. For instance, a drug product can only be released to the market after manufacture if its release tests satisfy a set of pre-prescribed specification limits, especially the potency limits that may serve as the extreme points for the prior on $\mu\alpha$. Also, the potency specification limits may be used to derive the prior distribution of variance parameter for the regression intercept. The prior distribution for degradation parameter may be obtained from a collection of pilot study, manufacturing process, product formulation as well as storage conditions. In case of no other prior information available, non-informative priors can always be used.

In this research we used threshold value $c = 1$ for the Bayes factor. It would be interesting to see how the choice of this threshold value affect the power, conditional power, and estimated shelf-lives in particular, under various designs. Although we illustrated the

hierarchical model in the balanced design, it is applicable to most of the stability designs including bracketing design, matrix design, and some recently proposed stability designs.^[25] A more complex design with covariates such as package type and dosage strength can also be accommodated by including more parameters and their priors, in which case an attention should be given to the balance of the number of data points and the number of parameters including hyperparameters to ensure their identifiability.

REFERENCES

1. FDA. *Guideline for Submitting Documentation for Stability Studies on Human Drugs and Biologics*; Center for Drugs and Biologics: Rockville, MD, 1987.
2. ICH. *Stability Testing of New Drug Substances and Products*. Tripartite International Conference on Harmonization Guideline, 2003.
3. ICH. *Evaluation of Stability Data*. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, 2002.
4. Ruberg, S.J.; Stegeman, J.W. Pooling data for stability studies: Testing the equality of batch degradation slopes. *Biometrics* **1991**, *47*, 1059–1069.
5. Ruberg, S.J.; Hsu, J.C. Multiple comparison procedures for pooling batches in stability studies. *Technometrics* **1992**, *34* (4), 465–472.
6. Chen, W.J.; Tsong, Y. Significance levels for stability pooling test: A simulation study. *J. Biopharm. Stat.* **2003**, *13*, 355–374.
7. Tsong, Y.; Chen, W.J.; Lin, T.Y.; Chen, C.W. Shelf life determination based on equivalence assessment. *J. Biopharm. Stat.* **2003**, *13*, 431–449.
8. Liu, W.; Jamshidian, M.; Zhang, Y.; Bretz, F.; Han, X. Pooling batches in drug stability study by using constant-width simultaneous confidence bands. *Stat. Med.* **2007**, *26* (14), 2759–2771.
9. Chow, S.-C.; Shao, J. Test for batch-to-batch variation in stability analysis. *Stat. Med.* **1989**, *8*, 883–890.
10. Chow, S.-C.; Shao, J. Estimating drug shelf-life with random batches. *Biometrics* **1991**, *47*, 1071–1079.
11. Shao, J.; Chow, S.-C. Statistical inference in stability analysis. *Biometrics* **1994**, *50*, 753–763.
12. Chen, J.J.; Hwang, J.-S.; Tsong, Y. Estimation of the shelf-life of drug with mixed effects models. *J. Biopharm. Stat.* **1995**, *5*, 131–140.
13. Chen, J.; Zhong, J.; Nie, L. Bayesian hierarchical modeling of drug stability data. *Stat. Med.* **2008**, *27* (13), 2361–2380.
14. Morrison, D.F. *Applied Linear Statistical Methods*; Prentice-Hall, Inc.: Englewood Cliffs, NJ, 1983.
15. Chow, S.C.; Liu, J.P. *Statistical Design and Analysis in Pharmaceutical Science*; Marcel Dekker, Inc.: New York, 1995.
16. Chow, S.-C. *Statistical Design and Analysis of Stability Studies*; Chapman & Hall/CRC: New York, 2007.

17. Berger, J.O. *Statistical Decision Theory and Bayesian Analysis*, 2nd Ed.; Springer-Verlag: New York, 1985.
18. Jeffreys, H. *Theory of Probability*; Oxford University Press: New York, 1961.
19. Box, G.E.P.; Tiao, G.C. *Bayesian Inference in Statistical Analysis*; John Wiley & Sons, Inc.: New York, 1973.
20. Berger, J.O.; Pericchi, L.R. Objective Bayesian Methods for Model Selection: Introduction and Comparison. In *Model Selection*; Lahiri, P.; Ed.; Institute of Mathematical Statistics Lecture Notes-Monograph Series: Beachwood, OH, 2001.
21. Berger, J.O.; Deely, J. A Bayesian approach to ranking and selection of related means with alternatives to analysis of variance methodology. *J. Am. Stat. Assoc.* **1988**, *83*, 364–373.
22. Berger, J.O.; Brown, L.D.; Wolpert, R.L. A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *Ann. Stat.* **1994**, *22*, 1787–1807.
23. Berger, J.O.; Boukai, B.; Wang, Y. Unified frequentist and Bayesian testing of a precise hypothesis. *Stat. Sci.* **1997**, *12*, 133–160.
24. Kass, R.E.; Raftery, A.E. Bayes factor. *J. Am. Stat. Assoc.* **1995**, *90*, 773–795.
25. Hedayat, A.; Yan, X.; Lin, L. Optimal designs in stability studies. *J. Biopharm. Stat.* **2006**, *16* (1), 35–59.

Bayesian Model: Prior Effective Sample Size

Satoshi Morita

Yokohama City University Medical Center, Yokohama, Kanagawa, Japan

Peter F. Thall

The University of Texas, M.D. Anderson Cancer Center, Houston, Texas, U.S.A.

Abstract

This entry presents definition for the prior effective sample size (ESS) of a parametric prior distribution in a Bayesian model, and proposes methods for computing the prior ESS for a wide variety of Bayesian parametric models, and also discusses the implementation of this method in specific Bayesian biomedical data analysis and clinical trial design settings.

Keywords: Bayesian analysis; Bayesian clinical trial design; Parametric prior distribution; Prior effective sample size.

INTRODUCTION

Understanding the strength of a prior distribution relative to the likelihood is a fundamental issue when applying Bayesian methods. The processes of formulating a putatively non-informative prior or eliciting a prior from an area expert typically require one to make many arbitrary choices, including choosing particular distributional forms and specifying numerical hyperparameter values. In practice, these choices are often dictated by technical convenience. A common criticism of Bayesian analysis is that an inappropriately informative prior may unduly influence posterior inferences and decisions. These issues may be addressed directly by quantifying the prior information in terms of a number of hypothetical patients, that is, a prior effective sample size (ESS). Such a summary allows one to judge the relative contributions of the prior and the data to the final conclusions. A useful property of prior ESS is that it is readily interpretable by any scientifically literate reviewer without requiring expert mathematical training. This is important, for example, for consumers of clinical trial results.

For some distributions, computing the prior ESS is straightforward. An example is the beta prior, $Be(a, b)$ commonly used for the success probability of a binomial sampling model, which is well known to have $ESS = a + b$. However, there is no obvious answer for many widely used models, including such basic models as a normal linear regression model with conjugate priors. We present a definition for the prior ESS of a parametric prior distribution in a Bayesian model, and propose methods for computing the prior ESS for a wide variety of Bayesian parametric models.^[1] In addition, we discuss the implementation of this method in specific Bayesian biomedical data analysis and clinical trial design settings.

PRIOR EFFECTIVE SAMPLE SIZE

The intuitive motivation for our method is to mimic the rationale for why the ESS of a $beta(a, b)$ equals $a + b$. If a binomial random variable Y from a sample of size n has success probability θ following a $beta(a, b)$ prior, then θ follows a $beta(a + Y, b + n - Y)$ posterior. In terms of ESS, given a sample of size n , the prior sum $a + b$ becomes the posterior sum $a + b + n$. Thus, saying that a given $beta(a, b)$ prior has ESS $m = a + b$ requires the implicit reasoning that the $beta(a, b)$ may be identified with a $beta(c + Y, d + m - Y)$ posterior arising from a previous $beta(c, d)$ prior having a very small amount of information. A simple way to formalize this is to set $c + d = \varepsilon$ for an arbitrarily small value $\varepsilon > 0$ and solve for $m = a + b - (c + d) = a + b - \varepsilon$.

As a general Bayesian framework, let $f(Y | \theta)$ denote the sampling model for a random vector Y with parameter vector $\theta = (\theta_1, \dots, \theta_d)$. That is, f is a parametric probability distribution function (pdf) or probability mass function (pmf). Denote $Y_m = (Y_1, \dots, Y_m)$, with $Y_i \sim f(Y_i | \theta)$ an i.i.d. sample. The likelihood is $f_m(Y_m | \theta) = \prod_{i=1}^m f(Y_i | \theta)$. Let $p(\theta | \tilde{\theta})$ be the prior on θ with hyperparameters $\tilde{\theta}$. An ε -information prior, $q_0(\theta | \tilde{\theta}_0)$, is defined by requiring matching means, $E_{q_0}(\theta) = E_p(\theta)$, and correlations, $Corr_{q_0}(\theta_j, \theta_{j'}) = Corr_p(\theta_j, \theta_{j'})$, $j \neq j'$, while inflating the variances of the elements of θ , with the requirement that $Var_{q_0}(\theta_j)$ must exist for each $j = 1, \dots, d$. The idea is that $q_0(\theta | \tilde{\theta}_0)$ is a version of $p(\theta | \tilde{\theta})$ that has a very small amount of information. For example, when $p(\theta | \tilde{\theta})$ is a beta distribution, $Be(\tilde{\alpha}, \tilde{\beta})$, $q_0(\theta | \tilde{\theta}_0)$ is specified as $Be(\tilde{\alpha}/c, \tilde{\beta}/c)$ using an arbitrarily large value $c > 0$, so that the $ESS = (\tilde{\alpha} + \tilde{\beta})/c$ of q_0 is very small while

the mean is still $\tilde{\alpha}/(\tilde{\alpha} + \tilde{\beta})$. Given a sample $\mathbf{Y}_m$ and an ε -information prior $q_0(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}_0)$, we denote the posterior by $q_m(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_m) \propto q_0(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}_0) f_m(\mathbf{Y}_m|\boldsymbol{\theta})$.

To define the prior-to-posterior distance between $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ and $q_m(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_m)$, the basic idea is to find the sample size, m , that would be implied by normal approximations of the prior $p(\boldsymbol{\theta})$ and the posterior $q_m(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_m)$. The second derivatives of the log densities are used to define the distance in terms of the trace of the information matrix (minus 2nd derivative of the log density) of the prior $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ and the expected information matrix of the posterior $q_m(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_m)$, where the expectation is with respect to the marginal $f_m(\mathbf{Y}_m)$. The ESS of $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ with respect to the sampling distribution $f(\mathbf{Y}|\boldsymbol{\theta})$ is defined to be the interpolated value of m minimizing the distance. If interest is focused on a subvector $\boldsymbol{\theta}_r$ of $\boldsymbol{\theta}$, the ESS can be determined similarly in terms of the marginal prior $p(\boldsymbol{\theta}_r|\tilde{\boldsymbol{\theta}})$. The reader is referred to Morita, Thall and Müller (MTM)^[1] for details of these definitions.

COMPUTATIONAL METHODS

We briefly summarize computational methods for obtaining ESS, details of which are provided in MTM.^[1] First, the following algorithm is used for computing prior ESS(s). Let M be a positive integer chosen so that it is reasonable to assume that $m \leq M$.

Step A1. Specify $q_0(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}_0)$.

Step A2. For each $m = 0, \dots, M$, compute the prior-to-posterior distance

Step A3. The ESS is the interpolated value of m minimizing the distance.

Steps A2 and A3 may be repeated to compute the ESSs for subvectors of $\boldsymbol{\theta}$, when subvector-specific ESS values as well as the ESS for the whole vector $\boldsymbol{\theta}$ may be of interests. When the expected information matrix of the posterior $q_m(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_m)$ cannot be computed analytically, the simulation-based numerical approximation is used.

The following algorithm may be used to calibrate the hyperparameters of prior distributions of subvectors of $\boldsymbol{\theta}$ using ESSs. Assume that $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2)$ and $\dim(\tilde{\boldsymbol{\theta}}) = 2$.

Step B1. Specify a prior ESS value.

Step B2. Fix $\tilde{\boldsymbol{\theta}}_1$.

Step B3. Find $\tilde{\boldsymbol{\theta}}_2$ to achieve the specified ESS.

Step B4. Repeat Steps B2 and B3 for a range of $\tilde{\boldsymbol{\theta}}_1$. The values of $(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2)$ define a contour of equal ESS value.

Step 5. Repeat Steps B1 to B4 for a set of ESSs values.

When $p = \dim(\tilde{\boldsymbol{\theta}}) \geq 3$, we suggest two strategies to reduce the dimension of the plot to two coordinates of

$\tilde{\boldsymbol{\theta}}$. Introduce constraints, such as $\tilde{\boldsymbol{\theta}}_2 = \tilde{\boldsymbol{\theta}}_3$. Alternatively fix $(\tilde{\boldsymbol{\theta}}_3, \dots, \tilde{\boldsymbol{\theta}}_p)$ at some value and carry out steps B2 through B5 for $(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2)$.

APPLICATIONS TO DATA ANALYSIS AND STUDY DESIGN

Cross-tabulated Counts from Small Samples

Carlin^[2] analyzed small-sample contingency table data from an experiment carried out to examine the effects of mechanical ventilator devices on lung damage in rabbits. In the experiment, the lungs of newborn rabbits were altered to simulate lung defects seen in human infants with underdeveloped lungs due to premature birth. The aim was to learn about the joint effects of different frequency and amplitude settings of the ventilators on lung damage. Six groups of six to eight animals each were compared using a factorial design with three frequency values crossed with two amplitudes. For amplitude indexed by $g = 1, 2$ for 20 and 60 respectively and frequency indexed by $h = 1, 2, 3$ for 5, 10, and 15 Hz, let $Y_{g,h}$ denote the number of animals with lung damage out of $n_{g,h}$ studied and let $\pi_{g,h}$ denote the probability of lung damage. The data are shown in Table 1, where each cell reports the empirical frequency $Y_{g,h}/n_{g,h}$. Carlin^[2] assumed the following logistic regression model

$$\pi_{g,h}(\boldsymbol{\theta}) = \text{logit}^{-1}(\mu + \alpha_g + \beta_h + \gamma_h I_{(g=2)}),$$

where $I_{(g=2)}$ indicates $(g = 2)$. The parameters α_g and β_h are main effects for amplitude and frequency while γ_h is the interaction effect at frequency h and amplitude $g = 1$. Hence, the model parameter is $\boldsymbol{\theta} = (\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \gamma_3)$.

Assuming a binomial model, $Y_{g,h} | \boldsymbol{\theta} \sim \text{Bin}(n_{g,h}, \pi_{g,h}(\boldsymbol{\theta}))$, the likelihood is

$$f(\mathbf{Y}_m | \boldsymbol{\theta}) \propto \prod_{g=1}^2 \prod_{h=1}^3 \pi_{g,h}(\boldsymbol{\theta})^{Y_{g,h}} \{1 - \pi_{g,h}(\boldsymbol{\theta})\}^{n_{g,h} - Y_{g,h}}$$

Carlin^[2] assumed independent normal priors for $\mu, \{\alpha_g\}, \{\beta_h\}$, and $\{\gamma_h\}$,

Table 1 The Effects of Mechanical Ventilator Devices on Lung Damage in Rabbits (Carlin, 2002)^[2]; Each Cell Contains the Number of animals with Lung Damage Over the Number Studied at the (Amplitude, Frequency) Combination

Amplitude	Frequency (Hz)		
	5	10	15
20	1/6	0/6	0/6
60	4/6	0/6	4/8

$$\mu \sim N(0, 1000^2), \alpha_g \sim N(0, \tilde{\sigma}_\alpha^2), \beta_h \sim N(0, \tilde{\sigma}_\beta^2), \gamma_h \sim N(0, \tilde{\sigma}_\gamma^2). \quad (1)$$

We use prior ESS to investigate sensitivity of inferences to hyperparameter values by considering seven alternative choices that cover a range of reasonably non-informative settings. The seven hyperparameter choices (N1 to N7) are shown in Table 2.

For each $\tilde{\theta}$, we compute an overall ESS for $p(\theta | \tilde{\theta})$, and also $ESS_\mu, ESS_\alpha, ESS_\beta$, and ESS_γ , for the marginal priors on the subvectors $\mu, \alpha = (\alpha_1, \alpha_2), \beta = (\beta_1, \beta_2, \beta_3)$, and $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ of θ . For convenience, below we will refer to the elements of θ as $\theta_1, \dots, \theta_9$. Let $D_{p,j}(\theta)$ and $D_{q,j}(m, \theta, Y_m)$ denote the j th diagonal elements of the information matrices of the prior $p(\theta | \tilde{\theta})$ and the posterior $q_m(\theta | \tilde{\theta}_0, Y_m)$, respectively. We specify an ε -information prior by inflating the variances in (1) by multiplying the variance hyperparameters by $c = 10,000$. It is straightforward to derive analytically that $D_{p,j}(\theta) = \tilde{\sigma}_\theta^{-2}$ and $D_{q,j}(m, \theta, Y_m) = \sum_{g,h} m_{g,h} \pi_{g,h} (1 - \pi_{g,h})$, for each $j = 1, \dots, 9$. Each $D_{q,j}$ depends on the numbers of animals in the cells but not on Y_m , which greatly simplifies averaging $D_{q,j}(m, \theta, Y_m)$ over the future data Y . The ESS values are summarized in Table 2.

The prior N7 with $\tilde{\sigma}_\alpha = \tilde{\sigma}_\beta = \tilde{\sigma}_\gamma = 0.5$ has $ESS = 36.6$, so that it has impact roughly equal to that of the data (sample size $n = 36$) on any posterior inference. Thus, based on its ESS, the prior N7 may be criticized as being overly informative. The priors N4 and N7 both assume very high prior information for the interaction effects. Both priors have $ESS_\gamma = 96.1$. Because the prior means of the interactions are 0, these priors have the effect of shrinking the posterior estimates of the interactions excessively toward 0. This illustrates the importance of evaluating not only the overall ESS, but also the ESS values of subvectors of θ that have particular interpretations in the model. Such ESS computations help readers to interpret the reported posterior results, such as the estimates between the frequency groups displayed in Carlin (Fig. 1).^[2] In this example, since the sample size is 38, the ESS_γ computation shows that the value $\tilde{\sigma}_\gamma = 0.5$ is far too small.

Table 2 Summary of Priors (Standard Error Parameters) and Computed ESSs

	$\tilde{\sigma}_\alpha$	$\tilde{\sigma}_\beta$	$\tilde{\sigma}_\gamma$	ESS	ESS_μ	ESS_α	ESS_β	ESS_γ
N1	2.0	2.0	2.0	2.3	<0.001	2.0	3.0	6.0
N2	1.5	1.5	1.5	4.1	<0.001	3.6	5.3	10.6
N3	1.5	1.5	1.0	6.0	<0.001	3.6	5.3	23.9
N4	1.5	1.5	0.5	16.3	<0.001	3.6	5.3	96.1
N5	0.5	0.5	1.5	24.4	<0.001	32.0	48.0	10.6
N6	0.5	0.5	1.0	26.3	<0.001	32.0	48.0	24.0
N7	0.5	0.5	0.5	36.6	<0.001	32.0	48.0	96.1

While each of the hyperparameters $\tilde{\sigma}_\alpha, \tilde{\sigma}_\beta, \tilde{\sigma}_\gamma$ can have an important impact on posterior inferences, they may be difficult to elicit or calibrate. One can use prior ESS graphically to assist in choosing these parameters. The idea is to compute the ESS for different combinations of the hyperparameters $\tilde{\theta}$, similar to Table 2. The ESS values are then plotted as a function of $\tilde{\theta}$. A practical problem is that one must first reduce the dimension of $\tilde{\theta}$ to $d = 2$ to allow one to construct a contour plot of ESS as a function of two hyperparameter values. As a general strategy, we suggest the use of simple restrictions on elements of θ . In this example, we will use $\tilde{\sigma}_\alpha = \tilde{\sigma}_\beta \equiv \tilde{\sigma}$, allowing us to plot ESS contours as a function of the two remaining hyperparameters $(\tilde{\sigma}, \tilde{\sigma}_\gamma)$. Fig. 1 shows the contours for $ESS = 0.5, 1.0, 2.0$, and 5.0 . From Fig. 1 one might, for example, decide to set $\tilde{\sigma}_\alpha = \tilde{\sigma}_\beta = 2.6$ and $\tilde{\sigma}_\gamma = 4.7$ to obtain $ESS = 1.0$.

TWO-AGENT DOSE-RESPONSE MODEL

We consider a design for a Phase I trial to determine an optimal dose combination $X = (X_1, X_2)$ of two cytotoxic agents.^[3] The toxicity probability at X is given by the 6-parameter model

$$\pi(X, \theta) = \frac{\alpha_1 X_1^{\beta_1} + \alpha_2 X_2^{\beta_2} + \alpha_3 (X_1^{\beta_1} X_2^{\beta_2})^{\beta_3}}{1 + \alpha_1 X_1^{\beta_1} + \alpha_2 X_2^{\beta_2} + \alpha_3 (X_1^{\beta_1} X_2^{\beta_2})^{\beta_3}}, \quad (2)$$

where all parameters in $\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, \alpha_3, \beta_3)$ are non-negative. Under this model, if only agent 1 is administered at dose X_1 , with $X_2 = 0$, as in a single-agent Phase I trial, then $\pi(X, \theta) = \pi_1(X_1, \theta_1) = \alpha_1 X_1^{\beta_1} / (1 + \alpha_1 X_1^{\beta_1})$ only depends on X_1 and $\theta_1 = (\alpha_1, \beta_1)$. Similarly, if $X_1 = 0$ then $\pi(X, \theta) = \pi_2(X_2, \theta_2) = \alpha_2 X_2^{\beta_2} / (1 + \alpha_2 X_2^{\beta_2})$ only depends on X_2 and $\theta_2 = (\alpha_2, \beta_2)$. The parameters $\theta_3 = (\alpha_3, \beta_3)$ characterize interactions that may occur when the two agents are used in combination. The model parameter vector thus is partitioned as $\theta = (\theta_1, \theta_2, \theta_3)$. Because

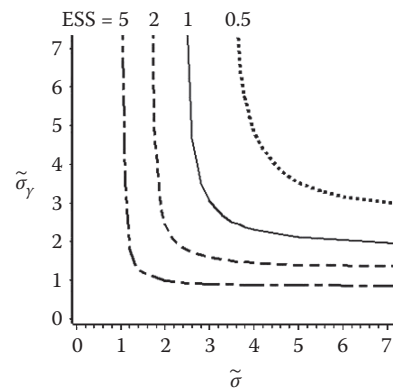


Fig. 1 Contour plots of ESS values for $\tilde{\sigma}$ (x-axis) and $\tilde{\sigma}_\gamma$ (y-axis). Dotted, solid, dashed, dashed-dotted lines show the plots of $ESS = 0.5, 1.0, 2.0$, and 5.0 , respectively

Phase I trials of combinations generally require that each agent previously has been tested alone, it is natural to obtain informative priors on θ_1 and θ_2 , but assume a vague prior on θ_3 . Denoting by $Ga(a,b)$ the gamma distribution with mean a/b and variance a/b^2 , the elicitation process (Thall et al., section 3)^[3] yielded the priors $\alpha_1 \sim Ga(1.74, 4.07)$, $\beta_1 \sim Ga(10.24, 1.34)$ for the effects of agent 1 alone, $\alpha_2 \sim Ga(2.32, 5.42)$, $\beta_2 \sim Ga(15.24, 1.95)$ for the effects of agent 2 alone, and $\alpha_3 \sim Ga(0.33, 0.33)$, $\beta_3 \sim Ga(0.0008, 0.0167)$ for the interaction parameters.

Since doses must be selected sequentially in Phase I trials based on very small amounts of data, an important question is what ESS may be associated with the prior. The ESS methods show that the overall $ESS = 1.5$. However, because informative priors on θ_1 and θ_2 were obtained and a vague prior on θ_3 was desired, it also is important to determine the prior ESS of each subvector. Applying the ESS computation methods yielded $ESS_1 = 547.3$ for θ_1 , $ESS_2 = 756.8$ for θ_2 , and $ESS_3 = 0.01$ for θ_3 . The small value for ESS_3 confirms that the prior on θ_3 reflects little information about the interaction of the two agents. The large numerical discrepancy between $ESS = 1.5$ and $(ESS_1, ESS_2) = (547.3, 756.8)$ is desirable. It reflects the fact that, for each $i = 1, 2$, " θ_i " has a very different meaning in the submodel $\pi_i(X_i, \theta_i)$ parameterized by θ_i alone versus its meaning in the full six-parameter model $\pi(X, \theta)$. See, for, example, Berger and Pericchi.^[4] From a geometric viewpoint, if $\pi(X, \theta)$ is thought of as a response surface varying as a function of the two-dimensional dose (X_1, X_2) , since the edges of the surface correspond to the submodels $\pi_1(X_1, \theta_1)$ where $X_2 = 0$ and $\pi_2(X_2, \theta_2)$ where $X_1 = 0$, the large values of ESS_1 and ESS_2 indicate that the locations of the edges were well known, while the small overall $ESS = 1.5$ says that otherwise very little was known about the surface. In practice, one would report ESS_1 , ESS_2 , ESS_3 , and ESS to the clinician from whom the priors were elicited. The clinician could then judge whether ESS_1 and ESS_2 are reasonable characterizations of his/her prior information about the single agents, and compare ESS to the trial's sample size. In this application, a trial of gemcitabine and cyclophosphamide for advanced cancer, the large values of ESS_1 and ESS_2 were appropriate since there was substantial clinical experience with each single agent, and the small overall ESS also was appropriate because no clinical data on the two agents used together were available and a sample size of 60 patients was planned.

CONCLUSION

An important area of practical application for the proposed prior ESS summaries is early phase clinical trials with small to moderate sample sizes. Concern about inappropriately influential prior information is one of the main impediments to a more widespread use of Bayesian

methods in clinical trials. By computing ESSs, one may avoid the use of an overly informative prior in the sense that inference is dominated by the prior rather than the data. Other uses of ESS values include interpreting or reviewing others' Bayesian analyses or designs, using the ESS values themselves to perform sensitivity analyses in the prior's informativeness, and calibrating the parameters of outcome-adaptive Bayesian designs.

The two examples illustrate four key features of our proposed method, namely that (i) ESS is a readily interpretable index of a prior's informativeness, (ii) it may be useful to compute ESSs for both the entire parameter vector and for particular subvectors, (iii) ESS values may be used as feedback in the elicitation process, and (iv) even when standard distributions are used, it may not be obvious how to define a prior's ESS.

Our ESS method described here is computationally complex. While the actual computational effort is negligible, the choice of a suitable ε -information prior and the evaluation of the prior-to-posterior distance require some problem-specific input from the investigator. That is, it is difficult to completely automate the ESS evaluation. However, the examples given here are intended to provide a basis for interested readers to compute and utilize prior ESS in their problems. A computer program, ESS_RegressionCalculator.R, to calculate the ESS for a normal linear or logistic regression model is available from the website <http://biostatistics.mdanderson.org/SoftwareDownload>.

ACKNOWLEDGMENT

Satoshi Morita's research was supported in part by Grant H20-CLINRES-G-009 from the Ministry of Health, Labour, and Welfare in Japan. Peter Thall's research was partially supported by NCI grant CA 83932.

REFERENCES

1. Morita, S.; Thall, P.F.; Müller P. Determining the effective sample size of a parametric prior. *Biometrics* **2008**, *64*, 595–602.
2. Carlin, J.B. *Assessing the Homogeneity of Three Odds Ratios: A Case Study in Small-Sample Inference*. In: Gatsonis, C.; Robert, E.K.; Carlin, B.; Carriquiry, A.; Gelman, A.; Verdinelli, I.; West, M.; Eds.; *Case Studies in Bayesian Statistics: Volume V*; Springer: New York, 2002, 279–290.
3. Thall, P.F.; Millikan, R.E.; Mueller, P.; Lee, S.J. Dose-finding with two agents in Phase I oncology trials. *Biometrics* **2003**, *59*, 487–496.
4. Berger, J.; Pericchi, L. Objective Bayesian methods for model selection: introduction and comparison (with discussion). In 'Model Selection'; Lahiri, P.; Ed.; Inst. of Math. Stat. Lecture Notes, Monograph Series; vol. 38, 2001.

Bayesian Statistics

Julia A. Varshavsky and Stacy R. Lindborg

Eli Lilly and Company, Indianapolis, Indiana, U.S.A.

Abstract

This entry presents an introduction to the Bayesian school of thought, providing definitions of its basic notions and core concepts and also discusses main ideas on how to use Bayesian analysis for clinical trial monitoring and sample size determination. Key components of the methodology will be demonstrated using examples from the biomedical field.

INTRODUCTION

The Bayesian approach to statistical inference has received increased attention in the scientific community over the past several decades because of the general framework it provides for incorporating new experimental evidence into scientific conclusions. Because of the core principles it rests upon, Bayesian analysis can be particularly useful in the field of biomedical research and development. There is an abundant literature on Bayesian analysis in general, as well as on Bayesian applications to the biopharmaceutical industry. Breslow's review, "Biostatistics and Bayes,"^[1] can serve as an introduction to the subject and links Bayesian analysis to numerous issues arising in clinical practice. Berry and Stangl^[2] provide an extensive collection of papers pertaining to the use of Bayesian methodology in health-related research. Among excellent general references are Berger,^[3] Bernardo and Smith,^[4] Press,^[5] and Gelman et al.^[6] Specific applications with suggested references include design of biomedical experiments,^[7] clinical trial monitoring,^[8–12] evaluation of drug pharmacokinetics and pharmacodynamics,^[13–15] planning trials with multiple endpoints,^[16] metaanalysis,^[17,18] and addressing ethical concerns associated with clinical trials.^[19]

OVERVIEW

In this entry, we will present an introduction to the Bayesian school of thought, providing definitions of its basic notions and core concepts. Key components of the methodology will be demonstrated using examples from the biomedical field. We also present main ideas on how to use Bayesian analysis for clinical trial monitoring and sample size determination. The literature cited contains many references to the other areas of application mentioned.

We will not attempt to discuss the differences between Bayesian and classical approaches in this entry. In the

authors' opinion, the two approaches address completely different sets of scientific questions and should be considered complementary. For a thorough comparison of both philosophies, see Berger,^[3] Carlin and Louis,^[20] or Gelman et al.^[6]

It should be noted that the implementation of Bayesian ideas has become increasingly prevalent in many areas of application because of recent advances in computational tools. Problems that were once just too difficult or too time-consuming to tackle from a Bayesian point of view are now becoming tractable. The last section of the entry contains a brief description of Bayesian computational methods.

BASIC NOTIONS AND BAYES THEOREM

Suppose a researcher conducts an experiment to test certain scientific beliefs, Q . Bayesian analysis provides a tool by which all prior knowledge about Q , $P(Q)$, can be combined with information about Q contained in data from an experiment, resulting in a posterior probability, or distribution of Q conditional on the data, $P(Q|\text{data})$. The core of this informational update is formed by Bayes' theorem. The theorem was first discovered by Rev. Thomas Bayes in the first half of the 17th century and published posthumously in 1763.^[21,22] In its simplest form it states that

$$P(Q|\text{data}) = \frac{P(Q) \times P(\text{data}|Q)}{P(\text{data})} \quad (1)$$

This result is a direct consequence of the laws of probability. In general, Q can be a statistical hypothesis, a distribution of a quantity of interest, or a parameter of this distribution. In the case of a parameter vector θ , Eq. 1 can be rewritten as:

$$\pi(\theta|y) = \frac{\pi(\theta) \times f(y|\theta)}{m(y)} \quad (2)$$

where y represents a vector of data, $f(y|\theta)$ is the likelihood, or conditional density, of the data given θ , $\pi(\theta)$ is the prior distribution of θ , and

$$m(y) = \int f(y|\theta)\pi(\theta)d\theta \quad (3)$$

is called the marginal distribution of Y . The following two examples demonstrate the use of Bayes' theorem for estimation purposes in both a discrete and a continuous case.

Example 1

Suppose a 40-year-old woman arrives for a routine obstetrics/gynecology visit and her physician performs a mammogram to test for the presence of breast cancer. Given a positive result of the test (the data), what is the chance that the patient indeed has breast cancer? Assuming that the woman has no family history of breast cancer and given the patient's age, the chance (prior probability) that she has the breast cancer is small—assume it is 0.1%. Let $\theta = 1$ if the cancer is present and $\theta = 0$ otherwise. Using Bayes' theorem,

$$\begin{aligned} P(\theta = 1 | \text{test is positive}) \\ = \frac{P(\text{test is positive} | \theta = 1) \times P(\theta = 1)}{P(\text{test is positive})} \end{aligned}$$

On the other hand,

$$\begin{aligned} P(\text{test is positive}) &= P(\text{test is positive} | \theta = 1) \\ &\quad \times P(\theta = 1) + P(\text{test is positive} | \\ &\quad \times \theta = 0) \times P(\theta = 0) \end{aligned}$$

Assuming further that the test procedure produces 5% false positives and 1% false negatives,

$$\begin{aligned} P(\text{test is positive}) &= 0.99 \times 0.001 + 0.05 \times (1 - 0.001) \\ &= 0.051 \end{aligned}$$

and, finally

$$\begin{aligned} P(\theta = 1 | \text{test is positive}) &= 0.99 \times 0.001 / 0.051 \\ &\cong 0.0194, \text{ or } 1.94\% \end{aligned}$$

In addition to illustrating the use of Bayes' theorem, this example highlights another important point. While having highly precise medical procedures (with high true-positive and low false-negative rates) is important, it is the “inverse” question—the chance of having the disease given a positive (or negative) result of the test—that is usually of the most interest. The Bayesian paradigm allows us to address this question directly. This seeming discrepancy between the precision of the diagnostic test and the posterior probability becomes especially pronounced for rare diseases (e.g., cancer, AIDS, and tuberculosis).

Example 2

Zyprexa® (olanzapine), a novel drug for the treatment of schizophrenia, was approved by the Food and Drug Administration (FDA) and European Medicinal Evaluation Agency (EMA) in September 1996. After approval, the sponsor, Eli Lilly and Company, was interested in collecting additional safety information. Given that serum prolactin elevations associated with older conventional neuroleptics were well documented in the literature, the sponsor wanted to estimate the impact of prolonged exposure to olanzapine on serum prolactin levels in a schizophrenic population. The mean prolactin level in a schizophrenic population taking olanzapine, θ , was chosen as the safety summary measure. The investigator's prior knowledge of θ was well represented by $N(\mu, \tau^2)$, with $\mu = 18.6$ and $\tau = 20.7$ ng/mL, based on the data previously summarized in the new drug application to the FDA. Let $y = (y_1, \dots, y_n)$ represent collected prolactin levels (ng/mL) from a prospective trial on n patients assessed after 6 months of drug exposure. For simplicity, we shall assume that $Y \sim N(\theta, \sigma^2)$, with $\sigma = 19.1$ ng/mL. Bayes' theorem yields a posterior distribution of θ that is also normal, with the mean $\mu(y)$ and variance v^2 given by:

$$\begin{aligned} \mu(\bar{y}) &= \mu \left(\frac{\sigma^2}{\sigma^2 + n\tau^2} \right) + \bar{y} \left(\frac{n\tau^2}{n\tau^2 + \sigma^2} \right) \text{ and} \\ v^2 &= \frac{\tau^2 \sigma^2}{n\tau^2 + \sigma^2} \end{aligned} \quad (4)$$

where $\bar{y} = n^{-1} \sum_{i=1}^n y_i$. Results from a multicenter double-blind clinical trial^[23] yielded $\bar{y} = 20.97$ ng/mL for $n = 144$ patients with a 6-month postbaseline serum prolactin measurement. The resulting posterior distribution of θ given the prior knowledge and the data is $N(20.96, 1.59^2)$. Hence the investigator could report with 95% probability that, given the currently available knowledge, the mean prolactin level in this patient population is contained between 17.84 and 24.08 ng/mL. Note the straightforward interpretation of this interval, born of conditioning on the data, in contrast to intervals constructed under the repeated sampling paradigm of classical statistics.

The marginal density $m(y)$ that appears in the denominator of Eq. 2 is sometimes called the *predictive distribution* of the data Y , which is also an important notion in Bayesian statistics. This distribution represents information about the value of an observation y given the prior knowledge and the likelihood function. Using the same idea as in Eq. 3, we can arrive at the posterior predictive distribution of new, yet-to-be-observed data, y_{new} , given observed data y . This type of distribution could be valuable, for example, in evaluating safety concerns of drugs under investigation.

Example 3

Clozapine is an atypical antipsychotic that has been associated with an increased risk of agranulocytosis, a

potentially fatal blood disorder defined as an absolute neutrophil count (ANC) $< 500 \mu\text{L}$ and the presence of symptoms of an infectious process. During clinical trials before the approval of olanzapine, there was interest in assessing whether olanzapine was also associated with an increased risk of agranulocytosis. Given the severity of the adverse event and the large amount of data being collected, it is reasonable to assume that physicians would want the ability to predict, for example, the probability of $\text{ANC} < 500$ for future patients conditional on all data collected up to a visit. To answer this question, one could average $f(y_{\text{new}}|\theta)$ with respect to the posterior distribution $\pi(\theta|y)$ that now plays the role of an updated prior, resulting in

$$p(y_{\text{new}}|y) = \int f(y_{\text{new}}|\theta)\pi(\theta|y)d\theta$$

Given this distribution, one could then calculate any probability of interest, for example,

$$p(y_{\text{new}} < 500|y) = \int_0^{500} p(y_{\text{new}}|y)dy_{\text{new}}$$

or report estimates such as 95% probability intervals, to aid in monitoring the safety of patients. One could even imagine setting objective bounds such as, say, $p(y_{\text{new}} < 500|y) > 0.0001$ to assist decisions of a monitoring board. Had this approach actually been used prior to approval, the physician monitoring the predictive distribution would have observed a constant decrease over time in the predictive probability of $\text{ANC} < 500$ as further evidence accumulated.

HYPOTHESIS TESTING

The problems of estimation and prediction are closely related to that of testing statistical hypotheses.

Example 4

Suppose researchers are now interested in testing the performance of olanzapine vs. a conventional neuroleptic competitor, haloperidol, in terms of the ability to reduce negative symptomatology. Such comparisons are often considered in pricing and reimbursement negotiations in European countries as well as for U.S. and Canadian formulary boards and marketing strategies worldwide. Consider a clinical trial where patients were assigned to olanzapine and haloperidol in a double-blinded fashion. The response variable of interest was assessed by the negative score from the Brief Psychiatric Rating Scale (BPRS). Let $Z = (z_1, \dots, z_n)$ and $H = (h_1, \dots, h_n)$ be the observations representing the difference between endpoint and baseline of scores collected from patients on olanzapine

and haloperidol, respectively (here, we assumed for simplicity that an equal number of patients was assigned to each treatment). Defining $d_i = z_i - h_i$, $i = 1, \dots, n$, suppose that $D = (d_1, \dots, d_n) \sim f(d|\theta)$, where θ represents the difference between mean change BPRS negative scores (endpoint minus baseline) of olanzapine and haloperidol, respectively. The test of the hypothesis $H_0: \theta \geq 0$ vs. $H_1: \theta < 0$ is of primary interest.

In the Bayesian paradigm, the choice between H_0 and H_1 is based upon the so-called *posterior odds ratio*:

$$\frac{P(H_1|\text{data})}{P(H_0|\text{data})} = \frac{P(\text{data}|H_1)}{P(\text{data}|H_0)} \times \frac{P(H_1)}{P(H_0)} \quad (5)$$

which represents the evidence in favor of H_1 vs. H_0 after observing the data. The ratio of $P(H_1)$ to $P(H_0)$ is called the *prior odds*. The ratio of posterior to prior odds,

$$B_{10} = \frac{\{P(H_1|\text{data})/P(H_0|\text{data})\}}{\{P(H_1)/P(H_0)\}} \quad (6)$$

is called the *Bayes factor*. It comprises a measure of how much evidence in favor of H_1 has changed after the data have been observed. Note that it represents the *change* in evidence, not an absolute measure of support for a hypothesis; interpreting Bayes factor as the latter can be very misleading.^[24,25]

Consistent with the literature, we will assume that the average of $n = 65$ collected observations, $\bar{D}$, is distributed as $N(\theta, \sigma_D^2/n)$, with $\sigma_D^2 = 11.0$, and that $\theta \sim N(\mu, \tau^2)$, with $\mu = -0.77$ and $\tau^2 = 17.53$. Note that, a priori, $P(H_0) = P(\theta \geq 0) = 0.428$ and $P(H_1) = P(\theta < 0) = 0.572$, and hence the prior odds of the investigator's beliefs about olanzapine's reducing negative symptoms, as compared to haloperidol, are $0.572/0.428 = 1.34$. Suppose that observed data from the prospective clinical trial yields $\bar{d} = -0.18$. The posterior distribution of θ [see Eq. 4] is normal, with the mean equal to -0.186 and the variance equal to 0.168 . Hence the posterior probabilities of the hypotheses are $P(H_0|\text{data}) = P(\theta \geq 0|\text{data}) = 0.326$ and $P(H_1|\text{data}) = P(\theta < 0|\text{data}) = 0.674$, yielding the posterior odds ratio $P(H_1|\text{data})/P(H_0|\text{data}) = 2.07$. We conclude that, based on the prior information and the evidence provided by the data, olanzapine is 2.07 times more likely than haloperidol to reduce negative symptoms as measured by BPRS. Note that the Bayes factor here is 1.55, thus implying that, after conducting the trial, the investigators' beliefs about the odds of olanzapine's reducing negative symptoms, as compared to haloperidol, have increased by 1.55.

The Bayesian approach to model selection and hypothesis testing was pioneered by Jeffreys,^[26] who used it to choose among competing scientific theories. Kass and Raftery^[27] provide an excellent review of the uses of Bayes factors for model selection in a variety of scientific applications.

CHOICE OF PRIOR DISTRIBUTIONS

The process of quantifying knowledge prior to decision making or future experimentation such as in the prolactin experiment is a struggle faced by *all* statisticians. Prior distributions are probabilistic expressions of knowledge possessed *prior* to conducting an experiment. It is perhaps easiest to think about prior distributions as falling into two broad categories, classified according to how the prior information is going to be incorporated into a distributional form or the purpose of the prior. The first category describes the scenario in which it is desirable to measure and incorporate an expert's uncertainty into posterior inference. The second category is characterized by situations in which one does not want to (or cannot) represent an expert's opinion, often referred to as "letting the data speak for themselves."

Scenarios in which one desires to incorporate prior information may arise, for instance, in a situation in which historical data exist and/or an expert opinion is available. In this case, the prior distribution can be formed by direct approximation from the available data or expert opinion. Alternatively, one can first choose a reasonable parametric functional form for the shape of the prior and then estimate or elicit the parameters for this distribution. For example, suppose that a normal distribution was deemed adequate to summarize the prior knowledge. One could then elicit or estimate information such as the center and the upper or lower quartile (quartiles are usually easier to assess than the standard deviation). The elicitation process boils down to a measurement problem in which the unit of measurement is a probability. Because experts are rarely trained in probability theory (because they can be anyone possessing the knowledge prior to an experiment), the elicitation process can be a challenging task, and the elicited information should be checked often for consistency. In the case of the normal distribution, for instance, it may be useful to gather information about *both* quartiles and perhaps some other percentiles of the distribution. For an excellent review and recommendations for eliciting prior distributions for biomedical applications, see Kadane and Wolfson,^[28] O'Hagan,^[29] and Chaloner.^[30] For some standard ways to approximate prior distributions in general, see, for example, Berger.^[3]

A common criticism of these types of priors is their very nature: they are based on subjective expert opinion or subjectively chosen data and thus can change, should another pool of experts or a set of data be chosen. For instance, one certainly wants to avoid the scenario in which an overzealous investigator overestimates the performance of a treatment and thus has an undue effect on the posterior conclusions drawn. To address this problem, Spiegelhalter et al.^[31,32] suggested considering a series of "experts" representing a skeptic, a clinician, the interest of a regulatory agency, and an uninformed or neutral individual. Another way of addressing this criticism is to extend the single

prior to a reasonable class of priors and to investigate the sensitivity of the final conclusion of the experiment to the class of priors chosen. See, for example, Greenhouse and Wasserman,^[11] Lavine et al.,^[33] and Berger^[34,35] for a discussion of issues related to prior robustness. Another argument to keep in mind is that, as the data from the experiment accumulate, the effect of the prior becomes increasingly less pronounced on the posterior. Hence for large sample sizes, usage of a prior becomes merely a way of invoking Bayesian machinery, and its particular choice become inessential.

The elicitation process may also result in a prior that does not prefer any single value of a parameter over any other. This scenario meets the objective of the first category of priors (thus incorporating expert opinion), although it is perhaps more common under the second category (where the impact of the prior information on the posterior is minimized). These types of priors are often called *noninformative* or objective. For example, in a case of a binomial experiment with parameters n and p , a possible noninformative prior on p is a uniform distribution on $[0,1]$. In the case of a normal distribution with a known variance, a reasonable noninformative prior for the mean parameter is a constant over the whole real line. Note that this prior is improper, in the sense that it does not integrate to a finite value, but this fact poses no problem for estimation purposes. Different statistical arguments can give rise to different noninformative priors. If the variance of the normal is unknown, for instance, reasonable choices of noninformative priors are $\pi_1(\theta, \sigma) = 1/\sigma^2$ and $\pi_2(\theta, \sigma) = 1/\sigma$. The former is an example of the so-called Jeffreys prior, based on the Fisher information matrix.^[26] The latter arises from the arguments of Bernardo^[36] and represents a so-called reference prior for the problem. A summary of various types of noninformative priors and corresponding posterior distributions is provided in Box and Tiao.^[37] For a general discussion of noninformative priors, see, for example, Berger,^[3] Bernardo and Smith,^[4] Carlin and Louis,^[20] and Gelman et al.^[6] Kass and Wasserman^[38] provide a summary of the vast literature in which theoretical and practical considerations of noninformative prior distributions are explored and debated.

When employing noninformative priors, it is important to keep in mind that no prior can literally be noninformative or objective; it must be judged as such on a relative basis. The "noninformativeness" it represents is always specific to the information criteria being used and to the scientific and statistical structure of a particular experimental problem. For an insightful discussion of objectivity in the role of science, see Howson and Urbach.^[39] Noninformative priors have long been viewed as attractive and will continue to be important as Bayesian assessments and interpretations become more prevalent in regulatory settings, because of the minimal role they play in the posterior distribution relative to the data.

In trying to accomplish either of the goals associated with the two categories of priors, sometimes priors of a specific functional form are chosen for ease of computation. In “Example 2,” the use of normal data $Y \sim N(\theta, \sigma^2)$, with fixed σ and $N(\mu, \tau^2)$ prior distribution for θ , resulted in a normal posterior distribution for θ , $N(\mu(y), v^2)$. This example illustrates the property of *conjugacy*, in which the posterior distribution follows the same parametric form as the prior distribution. As another example consider binomial data Y resulting from n Bernoulli trials with success probability p . A *conjugate* family of priors for p corresponding to this scenario is a Beta family, $B(a, b)$, where $a, b > 0$ are parameters of the distribution. Specifically, if $Y \sim \text{Binomial}(n, p)$, and a priori $p \sim B(a, b)$, then the posterior distribution for p is $B(a + y, b + n - y)$. More generally, we know that if we sample from the exponential family and our prior distribution is also a member of the exponential family, then the posterior, in turn, will also be from the exponential family. This knowledge removes the necessity to explicitly calculate the marginal density while still allowing the prior to be updated with information contained in the data. Clearly, mathematical convenience should not dictate the form of the prior distribution. However, *mixtures* of conjugates are often mathematically convenient and allow for approximation of prior beliefs in a more flexible manner. For a discussion of conjugate priors, as well as mixtures of conjugate priors and families, see, for example, Bernardo and Smith.^[4]

CLINICAL TRIAL MONITORING

In the standpoint of regulatory, medical, and budget constraints, most controlled clinical trials have a fixed sample size chosen prior to start-up based upon the expected magnitude of drug effectiveness and adverse event profile. Unfortunately, the precision of knowledge available regarding these factors is often minimal prior to conducting a trial, and thus the choice of the sample size may depend on possibly false assumptions. To address the ethical aspects of this dilemma and to assess the possibility of stopping a trial early, partially collected data from a clinical trial are often analyzed during interim analyses. Repeated testing of the null hypothesis in a classical setting leads to an accumulation of the Type I error rate.^[40] Methods arising from the works of Pocock,^[41] O’Brien and Fleming,^[42] Lan and DeMets,^[43] and Jennison and Turnbull^[44] suggest various criteria for adjusting significance levels at interim and at final analyses to address the problem. The main shortcoming of all these approaches is the fact that the final inference depends on whether or not one has entertained the possibility of rejecting the null hypothesis early at an interim analysis, i.e., the stopping rule selected for the trial. In other words, the same set of data may lead to a different conclusion about the significance of the observed results, depending on whether one has looked at part of the data

at an interim analysis. This concept is usually found to be counterintuitive by many practitioners. Bayesian inference, on the other hand, depends only on the observed data and not on the stopping rule used. An excellent discussion of this issue is presented in Berger and Wolpert.^[45] From a Bayesian perspective, one could look at the data as often as necessary without the risk of affecting the final conclusion. The essence of how this monitoring can be carried out may be found in many references (see, for example, Ref. [46] or Ref. [10]) and is demonstrated here through an example.

Recall the setting of “Example 4.” Under the assumptions of the example, the posterior distribution of d , the true treatment difference between olanzapine and haloperidol with respect to negative symptoms, given n observed pairs of observations, is normal $N(\mu(\bar{d}), \rho^{-1})$, with

$$\mu(\bar{d}) = \mu \left[\frac{\sigma^2/n}{(\tau^2 + \sigma^2/n)} \right] + \bar{d} \left[\frac{\tau^2}{(\tau^2 + \sigma^2/n)} \right] \quad \text{and} \\ \rho = \frac{n\tau^2 + \sigma^2}{\tau^2\sigma^2}$$

Note that, in this setting, smaller values of d —say, $\{d < d_1\}$ —would speak to the advantage of olanzapine (i.e., inferiority of haloperidol), while larger values of d —say, $\{d > d_2\}$ —would speak to the advantage of haloperidol (i.e., inferiority of olanzapine). Then the set $\{d_1 < d < d_2\}$ indicates a region of clinical equivalence of the two drugs (in some instances, one may be interested in setting $d_1 = d_2$). When comparing to a proven active control, it is reasonable to suggest that the trial should be stopped early if either treatment is found no worse than the competitor drug with high probability. This requirement means that we would stop after m observed pairs if either posterior probability

$$P(d < d_2 | D_1, \dots, D_m) > 1 - \varepsilon_1 \quad \text{or} \\ P(d > d_1 | D_1, \dots, D_m) > 1 - \varepsilon_2$$

where ε_1 and ε_2 are prespecified, small, positive constants. These requirements translate into the following regions for $\mu(\bar{d})$, $\bar{d} = m^{-1} \sum_{i=1}^m d_i$:

$$\mu(\bar{d}) < d_2 + \rho^{(-1/2)} \phi^{-1}(\varepsilon_1) \quad \text{or} \\ \mu(\bar{d}) < d_1 - \rho^{(-1/2)} \phi^{-1}(\varepsilon_2)$$

That is, one is to stop after m observations if either of the foregoing conditions is satisfied. Here, $\Phi^{-1}(\varepsilon)$ is the ε th percentile of the standard normal distribution.

SAMPLE SIZE DETERMINATION

Because Bayesian analysis allows frequent monitoring of the data being collected, the choice of the sample size prior to conducting an experiment is not necessary. Under the Bayesian paradigm, one can evaluate the decision to

discontinue an experiment with each look at the data. However, considering that all scientists have budgetary constraints and must plan for future expenditures associated with experiments, the choice of the sample size is often required.

A formal approach to sample size determination in Bayesian analysis requires the use of decision-theoretic framework.^[3] The use of this framework requires the introduction of a statistical loss function that specifies a penalty for errors in estimation and a cost of obtaining sampling information. Minimization of the loss function results in the optimal sample size. Here, however, we will present some practical ideas on how to choose the sample size without explicitly specifying loss functions, because the general concept of the latter may seem unnatural to the practitioner. In particular, these ideas rely upon the notion of a so-called highest-posterior-density (HPD) set. Their application to determining sample sizes for binomial proportions, multinomial and normal sampling, can, for example, be found in Adcock,^[47,48] Pham-Gia and Turkkan,^[49] and Joseph et al.^[50] Formally speaking, an HPD set with coverage probability $(1 - \alpha)$ for a posterior density $\pi(\theta|x)$ is defined as a set $A = \{\theta \in \Theta\}$ such that $\int_A \pi(\theta|x) d\theta = 1 - \alpha$, and such that the length of A , $l(A)$, or in higher dimensions, volume, is minimal. For instance, in “Example 2,” the interval [17.48, 24.08] for mean level of serum prolactin (ng/mL) in a schizophrenic population was actually a 95% HPD set.

Typically, an HPD set of length l and coverage probability $(1 - \alpha)$ is sought. Because such a set depends on the data X , which are not known before the experiment, one of the ideas is to average across all possible data. *Average coverage criterion* (ACC) proposes to choose the minimum number of observations, n , satisfying

$$\int_{\mathfrak{N}} \left\{ \int_{A(x,n)} \pi(\theta|x) d\theta \right\} f(x) dx \geq 1 - \alpha \quad \text{and} \quad l(A(x)) = l$$

where $f(x)$ is the data density, $\mathfrak{N}$ is the observational space, and $A(x,n)$ represents the HPD set. Alternatively, one could use an *average length criterion*. In this case, for each data point x in $\mathfrak{N}$, one would first find an HPD set of length $l^*(x,n)$ with coverage probability $(1 - \alpha)$, and then find the minimum n satisfying $\int_{\mathfrak{N}} l^*(x,n) f(x) dx \leq l$. Another alternative, the *worst outcome criterion*, ensures that the coverage probability is at least $(1 - \alpha)$ and $l(A)$ is at most l , regardless of the observed data, by requiring minimum n to be chosen so that

$$\inf_{x \in \mathfrak{N}} \left\{ \int_{A(x,n)} \pi(\theta|x) d\theta \right\} \geq 1 - \alpha \quad \text{and} \quad l(A) = l$$

In the case of “Example 2,” the posterior distribution is normal, and thus $l(x,n)$ depends only on n and not on x .

In addition, $\int_{A(x,n)} \pi(\theta|x) d\theta$ also does not depend on x . Hence all three criteria offer the same answer:

$$n \geq \left(\frac{4}{l^2} \right) \sigma^2 \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) - \frac{\sigma^2}{\tau^2}$$

where $\Phi^{-1}(1 - \alpha/2)$ represents the $(1 - \alpha/2)$ th percentile of the standard normal distribution. A more computationally challenging scenario arises in the case of the binomial data with Beta (conjugate) priors. Joseph et al.^[50] considered an example of a randomized two-arm parallel clinical trial designed to compare two drugs, warfarin and low-molecular-weight heparin, for the treatment of deep vein thrombosis (DVT). Equal numbers of patients, n , were to be randomized to each treatment. Suppose p_1 and p_2 are DVT rates under warfarin and heparin, respectively. Joseph et al.^[50] assumed that the number of patients who would develop DVT in the trial were independent binomial random variables with respective parameter sets (p_1, n_1) and (p_2, n_2) . Based on the previous data, the authors chose Beta(3,11) and Beta(11,52) priors for p_1 and p_2 , respectively. Using $\alpha = 0.05$ and $l = 0.05$ yielded $n_{\text{ACC}} = 1800$, $n_{\text{ALC}} = 1770$, and $n_{\text{WOC}} = 3075$. Computation of these sample sizes were based on the exact likelihood and were performed by drawing Monte Carlo samples from the posterior distribution.^[51–53] The software implementing these computations in S-plus can be found on the World Wide Web (WWW) at the following address:

<http://www.epi.mcgill.ca/~web2/Joseph/software.html>

BAYESIAN COMPUTING

Implementation of Bayesian methodology requires the computation of posterior and marginal distributions that involve integral expressions. The use of conjugate priors is very appealing from that perspective because it results in closed-form expressions for the posterior. In all other cases, one can use methods based on either analytic approximations of Taylor’s theory (see, for example, Schwarz,^[54] Tierney and Kadane^[55]) or numerical quadratures (Naylor and Smith,^[56] Dellaportes and Wright^[57,58]), importance sampling (Hammersley and Handscomb,^[59] Berger^[3]), and Markov chain Monte Carlo techniques (Geman and Geman,^[60] Tanner and Wong,^[61] Gelfand and Smith,^[62] Gamerman,^[63] Cowles and Carlin^[64]). Development of these computational methods has received a lot of attention in recent years and is an active field of current research.^[65]

There is an abundant quantity of available software that implement some of the preceding ideas. Interested readers are referred to Goel^[66] and to Carlin and Louis,^[20] who provide an extensive overview of the currently available Bayesian software. Here, we would like to mention a few popular programs that, in the authors’ view, are likely to be used by scientists in the biomedical field. One such

package, Bayesian inference Using Gibbs Sampling (BUGS), developed by D. Spiegelhalter, A. Thomas, N. Best, and W. Gilks, is built upon the use of Markov chain Monte Carlo techniques. The software and a set of manuals, along with many references on theory and application of the tool, is available through the WWW site of MRC Biostatistics Unit (Cambridge, UK) at:

<http://www.mrc-bsu.cam.ac.uk/bugs/Welcome.html>

for Unix, DOS, and Windows platforms. "Sbayes," developed by L. Tierney at the University of Minnesota, utilizes Laplace approximations for posterior evaluations and consists of a set of functions written in the S programming language. This program is available at:

<http://lib.stat.cmu.edu/S/sbayes>

Both "BUGS" and "Sbayes" are free software. Commercially available packages, SAS and BMDP, also allow some Bayesian computation within their procedures PROC MIXED and BMDP-5V, respectively. These procedures are targeted for the analysis of mixed-effects models,^[67] often used for the analysis of clinical trials, and other scenarios that give rise to repeated-measures data.

ACKNOWLEDGMENTS

The authors wish to thank their colleagues, J.W. Seaman, V. Arora, and J.D. Helderbrand, for many insightful comments and discussions that led to the improvement of both the structure and the focus of the manuscript.

REFERENCES

1. Breslow, N. *Stat. Sci.* **1990**, 5 (3), 269–298.
2. Berry, D.A.; Stangl, D.K. *Bayesian Biostatistics*; Marcel Dekker: New York, 1996.
3. Berger, J.O. *Statistical Decision Theory and Bayesian Analysis*, 2nd Ed.; Springer Verlag: New York, 1985, 74–113.
4. Bernardo, J.M.; Smith, A.F.M. *Bayesian Theory*; Wiley: New York, 1994; 265–285.
5. Press, S.J. *Bayesian Statistics: Principles, Models, and Applications*; Wiley: New York, 1989.
6. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*; Chapman and Hall: London, 1996.
7. Verdinelli, I. *Bayesian Statistics*; Bernardo, J.M.; Berger, J.O., Dawid, A.P., Smith, A.F.M., Eds.; Oxford University Press: London, 1992, 4, 467–481.
8. Berry, D.A. *Stat. Med.* **1985**, 4, 521–526.
9. Berry, D.A. *Drug Inf. J.* **1991**, 25, 345–368.
10. Freedman, L.S.; Spiegelhalter, D.J. *Contr. Clin. Trials* **1989**, 10, 357–367.
11. Greenhouse, J.B.; Wasserman, L. *Stat. Med.* **1995**, 14, 1379–1391.
12. Thall, P.F.; Simon, R. *Contr. Clin. Trials* **15**, 463–481.
13. Gatsonis, C.; Greenhouse, J.B. *Stat. Med.* **1992**, 11, 1377–1389.
14. Racine-Poon, A.; Smith, A.F.M. *Statistical Methodology in Pharmaceutical Sciences*; Berry, D.A., Ed.; Marcel Dekker: New York, 1990; 139–162.
15. Wakefield, J.; Racine-Poon, A. *Stat. Med.* **1995**, 14, 971–986.
16. Berry, D.A. *Bayesian Statistics*; Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M., Eds.; Oxford University Press: Oxford, 1989, 3, 239–270.
17. Berry, D.A. In *A Bayesian Approach to Multicenter Trials and Metaanalysis*. Proceedings of the Biopharmaceutical Section of the ASA, Anaheim, CA, 1990; 1–10.
18. DuMouchel, W. *Statistical Methodology in the Pharmaceutical Sciences*; Berry, D.A., Ed.; Marcel Dekker: New York, 1989, 509–529.
19. *Bayesian Methods and Ethics in a Clinical Trial Design*; Kadane, J.B., Ed.; Wiley: New York, 1996.
20. Carlin, B.P.; Louis, T.A. *Bayes and Empirical Bayes Methods for Data Analysis*; Chapman and Hall: New York, 1996.
21. Bayes, T. *Philos. Trans. R. Soc. Lond.* **1763**, 53, 370–418. Reprinted with an introduction by Barnard, in Ref. 22.
22. Barnard, G. *Biometrika* **1958**, 45, 293–315.
23. David, S.R.; Crawford, A.M.; Breier, A. *Schizophr. Res.* **1998**, 29 (1–2), 1–153.
24. Lindley, D.V.; Smith, A.F.M. *J. R. Stat. Soc., Ser. B* **1972**, 16, 1–42.
25. Lindley, D.V. *Math. Sci.* **1993**, 18, 60–63.
26. Jeffreys, H. *Theory of Probability*, 3rd Ed.; Oxford University Press: London, 1961.
27. Kass, R.E.; Raftery, A.E. *J. Am. Stat. Assoc.* **1995**, 90, 773–796.
28. Kadane, J.B.; Wolfson, L.F. *Statistician* **1998**, 47 (1), 3–21.
29. O'Hagan, A. *Statistician* **1998**, 47 (1), 21–37.
30. Chaloner, K. *Bayesian Biostatistics*; Berry, D.A., Stangl, D.K., Eds.; Marcel Dekker: New York, 1996; 141–156.
31. Spiegelhalter, D.J.; Dawid, A.P.; Lauritzen, S.L.; Cowell, R.G. *Stat. Sci.* **1993**, 8, 219–283.
32. Spiegelhalter, D.J.; Freedman, L.S.; Parmar, M.K.B. *J. R. Stat. Soc., Ser. A* **1994**, 157, 357–416.
33. Lavine, M.; Wasserman, L.; Wolpert, R.L. *J. Am. Stat. Assoc.* **1991**, 86 (416), 964–971.
34. Berger, J.O. *Robustness of Bayesian Analysis*; Kadane, Ed.; Elsevier Science: Amsterdam, 1984, 63–144.
35. Berger, J.O. *J. Stat. Plan. Inference* **1990**, 25, 303–328.
36. Bernardo, J.M. *J. R. Stat. Soc.* **1979**, 41, 113–147.
37. Box, G.E.P.; Tiao, G. *Bayesian Inference in Statistical Analysis*; Wiley Classics: New York, 1973.
38. Kass, R.E.; Wasserman, L.A. *Formal Rules for Selecting Prior Distributions: A Review and Annotated Bibliography*; Technical Report 583; Carnegie-Mellon University: Pittsburgh, PA, 1995.
39. Howson, C.; Urbach, P. *Scientific Reasoning: The Bayesian Approach*; Carnegie-Mellon University; Open Court, IL, 1989.
40. Armitage, P.; McPherson, C.K.; Rowe, B.C. *J. R. Stat. Soc., Ser. A* **1969**, 132, 235–244.
41. Pocock, S.J. *Biometrika* **1977**, 64, 191–199.
42. O'Brien, P.C.; Fleming, T.R. *Biometrics* **1979**, 35, 549–556.

43. Lan, K.K.G.; DeMets, D.L. *Biometrika* **1983**, *70*, 659–663.
44. Jennison, C.; Turnbull, B.W. *J. R. Stat. Soc., Ser. B.* **1989**, *51*, 305–361.
45. Berger, J.O.; Wolpert, R.L. *The Likelihood Principle*, 2nd Ed.; IMS: Hayward, CA, 1988.
46. Lindley, D.V. *Introduction to Probability and Statistics from a Bayesian Viewpoint*; Cambridge Univ. Press: Cambridge, 1965.
47. Adcock, C.J. *Statistician* **1987**, *36*, 155–159.
48. Adcock, C.J. *Statistician* **1988**, *37*, 433–439.
49. Pham-Gia, T.; Turkkan, N. *Statistician* **1992**, *41*, 389–397.
50. Joseph, L.; Wolfson, D.; Du Berger, R. *Statistician* **1995**, *44* (2), 143–154.
51. Albert, J.H. *Am. Stat.* **1993**, *47*, 182–191.
52. Tanner, M. *Tools for Statistical Inference*; Springer-Verlag: New York, 1991.
53. Pham-Gia, T.; Turkkan, N. *Commun. Stat., Theory Methods.* **1993**, *22*, 1755–1771.
54. Schwarz, G. *Ann. Stat.* **1978**, *6*, 461–464.
55. Tierney, L.; Kadane, B. J. *Am. Stat. Assoc.* **1986**, *81*, 82–86.
56. Naylor, J.C.; Smith, A.F.M. *Appl. Stat.* **1982**, *31*, 214–225.
57. Dellaportas, P.; Wright, D.E. *Stat. Comput.* **1991**, *1*, 1–12.
58. Dellaportas, P.; Wright, D.E. *Bayesian Statistics*; Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M., Eds.; Oxford University Press: London, 1992, *4*, 601–606.
59. Hammersley, J.M.; Handscomb, D.C. *Monte Carlo Methods*; Wiley: New York, 1964.
60. Geman, S.; Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, *6*, 721–741.
61. Tanner, M.; Wong, W. J. *Am. Stat. Assoc.* **1987**, *82*, 528–550.
62. Gelfand, A.E.; Smith, A.F.M. *J. Am. Stat. Assoc.* **1990**, *85*, 398–409.
63. Gamerman, D. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*; Chapman and Hall: London, 1997.
64. Cowles, M.K.; Carlin, B.P. *J. Am. Stat. Assoc.* **1996**, *91*, 883–904.
65. Gelfand, A.E.; Smith, A.F.M. *Bayesian Computation*; Wiley: New York, in preparation.
66. Goel, P.K. *Bayesian Statistics*; Bernardo, J.M.; DeGroot, M.H., Lindley, D.V., Smith, A.F.M., Eds.; Oxford University Press: Oxford; 1988, *3*, 173–188.
67. Laird, N.M.; Ware, J.H. *Biometrics* **1982**, *38*, 963–974.

Bayesian Statistics: Medical Device Clinical Trials

Gerry W. Gray

Data-Fi LLC, Silver Spring, Maryland, U.S.A.

Xuefeng Li, Laura Thompson, Yunling Xu, Xu Yan, and Lilly Q. Yue

*Division of Biostatistics, Center for Devices and Radiological Health,
U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*

Abstract

Bayesian statistical methodology has been used for more than 10 years in medical device clinical trials for premarket submissions. In this entry, two major ways of using a Bayesian approach in medical device clinical trials will be discussed in detail with examples. Bayesian methods can be used for adaptive designs and for incorporating prior data.

Keywords: Adaptive design; Hierarchical model; Predictive distribution; Historical data; Operating characteristics.

INTRODUCTION

A medical device is defined in the United States (U.S.) Food, Drug, and Cosmetic Act as “an instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article ... which does not achieve its primary intended purposes through chemical action within or on the body of man or other animals and which is not dependent upon being metabolized for the achievement of its primary intended purposes.”

In the United States, medical devices are regulated by the Food and Drug Administration (FDA) Center for Devices and Radiological Health (CDRH). Medical devices are classified as high, medium, or low risk, and there are two basic pathways for marketing approval in the United States. The first pathway, for higher risk and new technologies, requires demonstration of reasonable assurance of safety and effectiveness of the device. This type of regulatory pathway is called a premarket approval (PMA). The second pathway, mainly for low and medium risk devices, requires demonstration of “substantial equivalence” to a device that is already legally marketed. This type of pathway is called a 510(k) after its section number of the U.S. Food, Drug, and Cosmetic Act.

Since the late 1990s, CDRH has been encouraging the use of Bayesian statistical methods to support approval of medical devices. CDRH issued a guidance document on the use of Bayesian methods in 2010 as well as a guidance in 2016 on adaptive clinical trials and many of which use Bayesian methods. Because medical devices are usually engineered products with many opportunities

for incremental improvement, there are often prior data that can be used. The effects of modifications to a medical device are usually better understood because devices often have a localized known mechanism of action. Therefore, the Bayesian paradigm of learning from evidence as it accumulates is well suited to medical device development.

In the Bayesian framework, prior knowledge about parameters is summarized by the prior density $p(\theta)$, and the information in the observed data is contained in the likelihood $p(y|\theta)$. Using Bayes' theorem, the posterior density for the parameters can be written as

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y).$$

Since $p(y)$ can be regarded as a normalizing constant, we have that

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

That is, the posterior is proportional to the product of the prior and the likelihood. A prior with a high influence on the posterior is said to be “informative.” If there is a large amount of observed data, or if the prior is not informative, then the posterior is primarily determined by the likelihood.

In the following sections, two major ways of using a Bayesian approach in medical device clinical trials will be discussed in detail with examples. One approach uses adaptive design; the other approach incorporates existing prior data.

PLANNING A BAYESIAN CLINICAL TRIAL

Prior Distribution

The prior distribution is a key part of Bayesian inference and represents the information about a parameter that is combined with the probability distribution of new data to yield the posterior distribution, which in turn is used for inferences and decisions. Prior distributions are probabilistic expressions of information possessed prior to conducting a new trial. Prior distributions can be informative or (relatively) non-informative.

Informative prior distributions give preferences to some values of the parameter of interest as being more likely than others. Informative priors may be utilized when historical data related to the device under study are present. Generally, for any previous studies to be considered appropriate as prior information, they should be similar to the proposed study in a number of factors. The process of quantifying prior information for the proposed trial may be challenging. Often, however, the prior distribution can be formed by direct approximation from the available data. Alternatively, one can first choose a reasonable parametric functional form for the shape of the prior and then estimate or elicit the parameters for this distribution.

A non-informative prior distribution, also referred to as a vague or flat prior, may be used when one does not want to use available prior information, or prior information from the previous studies is not available. For example, in the case of a binomial experiment with parameters n and p , a possible non-informative prior on p is a uniform distribution on $[0,1]$. In the case of a normal distribution with a known variance, a reasonable non-informative prior for the mean parameter is a constant over the whole real line or a normal distribution with a large variance.

Determining the Sample Size

The sample size of a Bayesian clinical trial depends on variability of the sample, prior data, assumed statistical model and prior distributions of model parameters, and decision criteria for various actions. Unlike traditional frequentist design, there is usually no closed solution for the sample size. Instead, extensive simulation is usually required to determine the sample size.

To begin the process of determining the sample size for a Bayesian study, often a minimum sample size is first specified based on clinical considerations—for example, sufficient number of patients needed to assess the safety profile of the investigational device. Then, based on available resources, such as the feasibility of enrolling patients and available funding, a maximum sample size is determined. From here on, simulations under various scenarios of the design parameters may be performed to see how potential sample sizes meet certain predetermined operating characteristics, such as type I error rate and

power. The determination of sample size is interrelated with the choice of design parameters.

Assessing the Operating Characteristics of a Bayesian Design

Due to the inherent complexity in the design of a Bayesian clinical trial, specifically when prior data are used and/or a wide variety of adaptations are planned, a thorough evaluation of the operating characteristics is of paramount importance. Evaluation of the operating characteristics usually includes:

- type I error rate
- power
- average sample size and
- if applicable, probability of stopping at each interim look.

In a regulatory environment, it is usually necessary to control type I error rate at an appropriate level. In general, when no prior data are used, the type I error rate is controlled at the customary frequentist level. When prior data are used, the type I error rate is often controlled at a higher level, with consideration given to the credibility of the prior data and the knowledge of potential risk–benefit profile of the investigational device.

An adequate characterization of the operating characteristics of any particular trial design usually needs extensive simulations. Simulations are performed under various more-or-less plausible scenarios of design parameters, evaluating the desirability of the operating characteristics. The process may be iterative, in the sense that sample size, any interim analysis rule, study success criteria, and priors may need to be adjusted many times to achieve the acceptable operating characteristics.

ANALYZING A BAYESIAN CLINICAL TRIAL

Summary of the Posterior Distribution

Conclusions from a Bayesian clinical trial are usually based on the posterior distributions of the parameters of interest. The posterior distribution contains all information from the prior distribution as well as the information from the current trial via the likelihood function. The posterior distribution is usually summarized with posterior mean, standard deviation, and credible interval. Graphical representations of the appropriate distributions are also helpful.

Hypothesis Testing and Interval Estimation

Statistical inference may include hypothesis testing, interval estimation, or both. For medical device clinical trials, the hypotheses are usually one-sided. Therefore, for Bayesian

hypothesis testing, the posterior probability of interest is that the treatment effect is less or greater than a prespecified value. If this posterior probability is greater than a threshold value (for example 97.5%), the null hypothesis is rejected.

Bayesian interval estimates are based on the posterior distribution and are called credible intervals. If the posterior probability that a parameter of interest lies in an interval is $x\%$, then this interval is called an $x\%$ credible interval. Two types of frequently reported credible intervals are highest posterior density intervals and central posterior intervals. A credible interval can also be used for testing a hypothesis in the same way that a confidence interval is used.

Interim Analysis

Interim analyses can be performed to make preplanned adaptations to a trial. Some examples that often occur in medical device trials are interim assessments to decide whether to stop accrual into a trial, whether to stop the trial for success, or whether to stop the trial for futility (additional adaptations are discussed in a later section). For a regulatory submission (as well as good clinical trial practice), the methodology and analyses employed for such interim assessments should be prespecified in the trial's design. For many interim assessments, the (posterior) predictive distribution is useful. The posterior predictive distribution of a new observation, y_{new} , from likelihood $p(y|\theta)$, when y_{old} are data already observed in the trial is

$$p(y_{\text{new}}|y_{\text{old}}) = \int p(y_{\text{new}}|\theta)p(\theta|y_{\text{old}})d\theta.[1]$$

From the predictive distribution, predictive probabilities may be obtained to make interim decisions about the trial.

To stop accruing patients into a trial based on interim data, the predictive probability of study success can be used as a decision criterion. For example, at an interim analysis, the predictive distribution for patients with incomplete follow-up is derived; then the predictive probability of study success is calculated with respect to all accrued patients. If this predictive probability is sufficiently high, the trial may stop accrual and continue follow-up on all patients. The final analysis is performed when all accrued patients have completed the follow-up. Such interim assessments for stopping accrual may be especially useful for trials with long follow-up and slow accrual.

To stop a trial for success based on interim data, both the posterior probability of treatment effect and the predictive probability of study success can be used as criteria. The posterior probability of treatment effect is calculated based on data from enrolled patients with complete follow-up; the predictive probability of study success is calculated with respect to all accrued patients with complete and incomplete follow-up. In general, the criterion to claim

study success is higher using the predictive probability than using the posterior probability because of additional model assumptions required to relate the outcomes for patients with complete follow-up to outcomes to be observed.

For an interim assessment to stop a trial for futility with a planned maximum sample size, one might use the predictive distribution for enrolled patients with incomplete follow-up and for patients to be enrolled within the maximum of the planned sample size. Then, the predictive probability of study success is calculated with respect to the maximum planned sample size. If this predictive probability is too low, the trial may stop for futility.

Sensitivity Analysis

In Bayesian clinical trials, a sensitivity analysis is usually conducted to investigate the effects of deviations from the statistical model and its assumptions on the main results of the trial. Sensitivity analyses might include investigations of:

- deviations from the distributional assumptions;
- alternative functional forms for the relationships in the model;
- alternative prior distributions;
- alternative “hyper prior” parameters in hierarchical models;
- deviations from any “missing at random” assumptions for missing data;
- alternative design parameters, accrual rate, correlations among longitudinal outcomes, etc.

A clinical trial result is considered to be robust if the study conclusions are approximately the same for most of the possible sensitivity analyses.

Model Checking

Model checking is especially important in a Bayesian analysis due to the complexity of the models employed. A tool frequently used in Bayesian model checking is the posterior predictive distribution of hypothetical observations from the trial, under the assumptions of the current model. One can generate new, hypothetical data from the posterior predictive distribution to see how well it “matches” observed data from the trial. The assessment of how well the posterior predictive data match the observed data can be made using the “Bayesian p value,” which is the posterior predictive probability of observing a result at least as extreme as what was observed in the trial.^[1]

In addition, Bayesian deviance measures such as deviance information criterion (DIC)^[2] can be used for model choice by comparing the fit of one model over another. It is calculated as the posterior mean of the model deviance minus the effective number of parameters in the model (p_D), where the effective number of parameters is the decrease in deviance expected from estimating the

parameters of the model. pD can be estimated as the posterior mean of the deviance minus the deviance evaluated at the posterior mean of the parameters.^[1] The minimum DIC estimates the model that will make the best short-term predictions. An useful application of DIC is to compare the fit of a normal random effects distribution to that of a (truncated) Dirichlet process mixture model, if a more flexible random effects distribution is desired.^[3] Alternatively, two models may be compared using Bayes factors.

For additional details on planning and analyzing a Bayesian study, the reader should consult the texts by Carlin and Louis,^[4] Gelman et al.,^[1] and FDA guidances.^[5,6]

BAYESIAN ADAPTIVE DESIGN

The Bayesian approach can offer flexibility in adapting important aspects of a study design, including the sample size, subject allocation ratio, number of treatment arms, etc.^[7] These parameters can be adjusted as information about the performance of the device is accrued, and this can potentially result in shorter trial duration and smaller size. Most importantly, the Bayesian approach allows for the construction of likelihood models that use information obtained at earlier follow-up time points to predict endpoints at later time points. Such use of predictive distributions can be particularly useful for clinical studies in which the follow-up period is long. Adaptive designs may require considerable planning but can pay off enormously. The potential advantages are illustrated in the following paragraphs.

Here, an example of a hypothetical trial to demonstrate superiority of an investigational device to control therapy is presented. Subjects are randomized to the investigational and control devices with a ratio of 2 (investigational) to 1 (control). The primary endpoint of the study is a composite of safety and effectiveness outcomes at 24 months, which classifies patients as an overall success or not. Additionally, interim measurements of the composite primary endpoint are available at 3, 6, and 12 months. To determine the operating characteristics, the success rate for the control therapy is assumed to be 75% (P_C), and for the investigational device, it is assumed to be 85% (P_T). Based on clinical safety considerations, the minimum sample size for the study is 300 subjects. And the maximum sample size is 800 based on the available financial resources. The recruitment rate is projected to be 20 subjects per month.

This Bayesian adaptive trial implements two types of adaptations. The first is sample size adaptation. The first potential sample size adaptation takes place when 300 subjects have been enrolled into the study, with additional potential adaptations every time when 50 subjects are accrued. At each interim analysis for sample size adaptation, one decides whether to stop accrual or to continue. This decision is based on modeling all the follow-up data available, including follow-up data at 3, 6, 12, and

24 months. If the predictive probability of trial success given the enrolled subjects exceeds a prespecified value (chosen as 0.8 for this trial), then the trial stops accrual; otherwise, enrollment will continue to the next stage.

The second adaptation is to potentially stop the trial for an early claim of study success. This adaptation occurs after the trial stops accrual, and it has four planned looks at 0, 6, 12, and 18 months. These looks will be conducted only if at least 150 patients have 24 months of follow-up data. Study success will be claimed if the predictive probability exceeds 0.995 at one of the interim looks. If study success is not claimed at an interim look, then when complete follow-up data are available for all enrolled subjects, study success will be claimed if the posterior probability exceeds 0.985.

The operating characteristics of this Bayesian adaptive trial can be approximated using simulations. To generate patient-level data, patient outcomes at 0, 3, 6, 12, and 24 months are assumed to be correlated. Based on the results from historical trials, the responses for a subject are assumed to follow a Markov chain model—i.e., given the observation at time $(t - 1)$, the observation at time t does not depend on previous observations. The model requires the specification of transition probabilities (see Table 1). Different probabilities of 24-month success can be generated by varying the transition probabilities. The prior distributions were assumed to be non-informative. The data for each subject can then be generated using the beta-binomial distribution. This process was then repeated 10,000 times and the obtained operating characteristics for the design are shown in Table 2.

As can be seen from column 2 of Table 2, study power is over 80% when the 24-month success rate in the treatment group is 85% (P_T). As can be seen from the last row, the type I error rate is 2.1% when $P_T = P_C = 0.75$. Table 2 also shows the average sample size, average number of subjects with complete follow-up at the time trial success is claimed, and the trial duration.

Table 1 Transition probabilities

	Success	Failure
Success at 3 months	0.63	0.37
3-Month success to 6 months	0.9	0.1
3-Month failure to 6 months	0.46	0.54
6-Month success to 12 months	0.9	0.1
6-Month failure to 12 months	0.311	0.689
12-Month success to 24 months	0.9	0.1
12-Month failure to 24 months	0.311	0.689

Table 2 Operating characteristics of proposed Bayesian design

Treatment effect (P_t)	P (trial success)	Average sample size	Average complete follow-up	Trial duration (months)
0.87	0.933	548	363	42.2
0.86	0.884	554	395	43.8
0.85	0.806	562	431	45.6
0.84	0.717	571	466	47.3
0.83	0.606	582	501	49.1
0.75	0.021	694	692	58.6

BAYESIAN DESIGN FOR INCORPORATING HISTORICAL DATA

One of the motivations for using Bayesian methods for medical device clinical trials is the capability to incorporate prior data. Medical device technology develops rapidly, and new versions of a device are developed in an iterative manner. Therefore, if a device is used as a concurrent active control in a study, data from a prior study could be used to augment the active control data. Another situation is where prior data may be available from both the treatment and control devices in the target population. With appropriate downweighting, this prior information might be used to augment a new study.

There are usually several ways to incorporate prior data into the study. A simple way is through a power prior approach that appropriately weights the prior data by raising its historical likelihood to a power.^[8] Several authors have since extended the power prior approach.^[9,10] Another common approach is to use Bayesian hierarchical modeling. A typical hierarchical model has two levels—a patient level and a study level. In this two-level structure, studies have different but related treatment effects or mean outcomes. That is, the studies are exchangeable or exchangeable conditional on observed covariates.

A key step in the specification of the Bayesian hierarchical model is the selection of the prior for the “between-study” variance. It not only determines the extent of borrowing between studies but also influences the operating characteristics of the design. In an ideal situation, data from several historical studies would be available so that the between-study variance could be given a non-informative prior, and the amount of borrowing among studies would be decided primarily by the data. However, when data from only one historical study are available, the prior for the between-study variance can easily dominate its posterior. In other words, the prior for the between-study variance needs to be informative when utilizing historical data from only one study. The determination of such an informative prior is challenging and in some cases can be done through extensive discussions with clinicians and engineers, while relying on any available published information.

The commensurate prior models have been proposed as alternatives to the customary hierarchical model.^[11] The commensurate prior is a structural prior distribution that describes the extent to which a parameter in this study varies about the analogous parameters in the historical study. The approach assumes that this study borrows strength from the historical study in the absence of evidence for heterogeneity. The authors showed that when properly calibrated within a specific context, this approach could yield good operating characteristics.

The following is a hypothetical example of a medical device clinical trial which shows how historical data are incorporated into the control arm of a study using a Bayesian hierarchical model. We consider a prospective randomized active controlled non-inferiority trial. The investigational device is the 2nd generation of a medical device, while the control device is the 1st generation of the device. The primary endpoint is the proportion of subjects who experience an adverse event at one year. The study success criterion is that the treatment arm is non-inferior to the concurrent control arm if the posterior probability $P\{(p_t - p_c) < \delta | \text{data}\}$ is greater than or equal to 0.95, where p_t and p_c represent the primary event rate at 1 year for the treatment and concurrent control arms, respectively, and δ is the non-inferiority margin. A historical study with n_{hc} patients is available for the control device. Since the historical study and this trial are similar in terms of the targeted patient population, definition and measurement of the clinical endpoints, physician training, etc., a hierarchical model to borrow information from the n_{hc} patients treated with the 1st generation device was considered to augment the control arm. The probability distribution for the number of primary endpoint events in the treatment arm, the concurrent control arm, and the historical control arm is assumed to be binomial: $y_t \sim \text{binomial}(p_t, n_t)$, $y_c \sim \text{binomial}(p_c, n_c)$, $y_{hc} \sim \text{binomial}(p_{hc}, n_{hc})$, respectively. Let

$$\mu_t = \log(p_t / (1 - p_t)),$$

$\mu_c + b_c = \log(p_c / (1 - p_c))$, and $\mu_c + b_{hc} = \log(p_{hc} / (1 - p_{hc}))$. Under the hierarchical model, we assume normal random study effects for the concurrent and historical controls: $b_c \sim N(0, \tau)$ and $b_{hc} \sim N(0, \tau)$. We also assume an improper uniform prior for μ_t and μ_c , and a gamma prior $\Gamma(\alpha, \beta)$ for τ , i.e., $\pi_0(\mu_t, \mu_c, \tau) \propto \tau^{\alpha-1} \exp(-\beta\tau)$ where $\alpha > 0$ and $\beta > 0$. So the joint posterior distribution based on this data $\{y_t, y_c, n_t, n_c\}$ and historical data $\{y_{hc}, n_{hc}\}$ is proportional to the product of prior distribution of μ_c , μ_t , and τ , the likelihood function based on this data $\{y_t, y_c, n_t, n_c\}$, and the likelihood function based on the historical control data $\{y_{hc}, n_{hc}\}$.

To calculate the posterior probability $P\{(p_t - p_c) < \delta | \text{data}\}$ based on the above distribution, sampling via the Gibbs sampler can be performed, and a Monte Carlo estimate of the posterior probability can be obtained. Chen et al.^[9] provide theoretical methodology and a computational

algorithm for the incorporation of historical data via a hierarchical modeling approach.

One should note that type I error cannot be stringently controlled at the customary frequentist level or else no historical data will be borrowed for the control arm.^[12] In this example, the type I error rate tends to be greater if the true event rate in the concurrent control arm is less than the rate observed in the historical control arm. In addition, the values of hyper prior parameters have impact on the strength of borrowing from historical information. Fig. 1 illustrates how type I error rate varies under different hyper priors. H_1 , H_2 , and H_3 represent the hyper prior distributions $\Gamma(\alpha, \beta)$ with different values of α and β [e.g., H_1 , H_2 , and H_3 can be $\Gamma(0.0001, 0.0001)$, $\Gamma(0.001, 0.001)$, and $\Gamma(0.01, 0.01)$, respectively]. The assumed true event rate in the concurrent control arm may occur within the interval from p_0 to p_6 , where $p_0 < p_1 < \dots < p_6$, and p_3 is the observed event rate in the historical control arm. The type I error rate under each hyper prior distribution is calculated through simulation when the true event rate in the concurrent control arm is assumed to be p_0 , or p_1 , or ..., or p_6 . From Fig. 1, we can see that type I error rate is controlled around 0.05 when the concurrent control rate is the same as the observed historical control rate (p_3), and type I error rate increases as the concurrent control rate decreases. The type I error rate curve under hyper prior H_1 has the sharpest slope among three different hyper priors because H_1 allows for more strength of borrowing from historical control than the other two hyper priors.

In order to address a potential concern that the type I error rate might be too high, a conditional borrowing strategy can be considered. If the observed event rate in the concurrent control arm is a prespecified amount lower than the rate from the historical control arm, no historical control data will be borrowed. Deciding the cutoff value for not borrowing historical data will entail communication between the sponsor and the regulatory agency. Since the

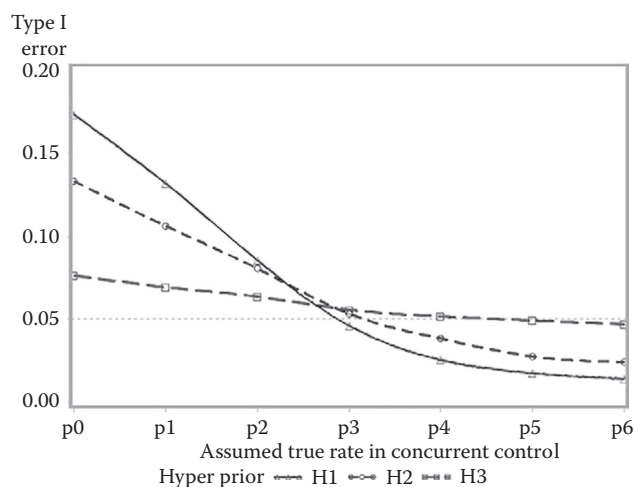


Fig. 1 Simulated type I error rate: Sensitivity analysis of robustness of changing hyper prior parameters

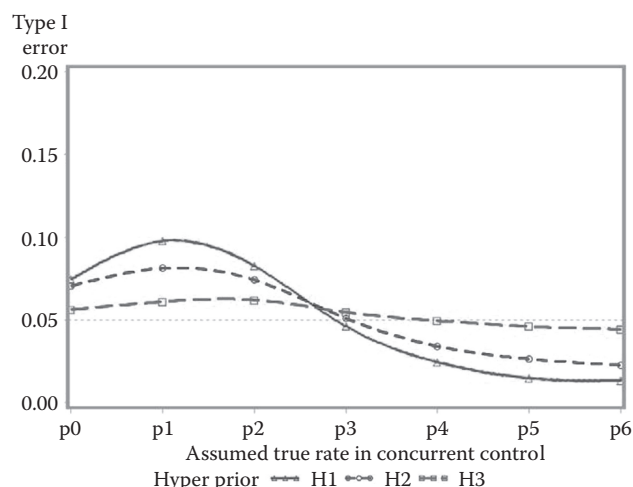


Fig. 2 Simulated type I error rate: Sensitivity analysis of robustness of changing hyper prior parameters using conditional borrowing strategy

choice of the hyper prior may impact the type I error rate as well, simulations are needed to calculate the type I error rate for different cutoff values and different hyper priors. Fig. 2 illustrates how type I error rate varies under different hyper priors using the conditional borrowing strategy that if the observed event rate in the concurrent control arm is less than the cutoff p_2 , then no historical control data will be borrowed. From Fig. 2, we can see that type I error rate stops increasing either at the cutoff point p_2 or at a nearby value p_1 . The maximum value of type I error rate under the hyper prior H_3 is controlled at the smallest level among all three hyper priors. Under this approach, the sample size for the new clinical trial is reduced, and the type I error rate is controlled at the agreed upon level under the hyper prior H_3 .

BAYESIAN DESIGN FOR USING ADULT DATA FOR PEDIATRIC INDICATION

The CDRH aims to promote safe and effective device use in pediatric patients, while ensuring device approvals are based on valid scientific evidence. Unfortunately, there is a paucity of scientific evidence available to substantiate submissions for devices that are indicated for use in the diagnosis or treatment of pediatric patients. Leveraging relevant available clinical data from other age groups (perhaps even adults), when appropriate, may lead to more devices being granted marketing authorization for pediatric indications, which will increase the availability of medical devices with appropriate labeling to support safe and effective device use in pediatric patients. The Guidance “Leveraging Existing Clinical Data for Extrapolation to Pediatric Uses of Medical Devices” issued in 2016 was written to explain how leveraging (or “extrapolation”) of adult data might proceed.^[13]

When the use of extrapolation is determined to be appropriate, there are several options for how to extrapolate from adult data. Available options could depend on whether a prospective study of pediatric patients is needed and feasible, and/or whether sufficiently robust pediatric data can be obtained in other ways, such as from prior studies run by the sponsor, studies in the literature, or pediatric registries.

A hierarchical model that statistically combines adult and pediatric studies might be an option if there are multiple adult studies. However, there are likely to be one or more differences that could prevent the assumption of exchangeability between adult and pediatric studies. If these differences can be identified and measured, it is straightforward to account for them in a hierarchical model. When this is done, we can say that the studies are exchangeable, except for measured differences on certain covariates. Often the differences will be related to the size or ongoing growth of the patient. For example, a device whose effect is related to hormone level may have very different magnitudes of effect for adults than for children because they have different circulating hormone levels. If patients are categorized into low, medium, and high hormone levels, then within each category, the adult studies might be exchangeable with the pediatric study. Presumably, if hormone level is highly associated with the effect of the device, a sponsor is likely to have patient-level data on the level of the circulating hormone in adults. Patient-level information in children would enable the sponsor to construct a model that relates hormone level in the blood to the outcome, and thus condition on hormone level to assume exchangeability.

However, differences between adult and pediatric studies could create performance biases in one study but not another, or they could indicate differences in risk that each age group is willing to take. Consequently, the studies might end up with too many differences to assume exchangeability. However, we still want to borrow from the adult studies. In this case, a *three-level* hierarchical model structure may work^[14] and have the consequence of reduced borrowing across adult and pediatric studies in comparison with a two-level hierarchical model. A three-level hierarchical model is easily extended to include covariates. Below, we show an example when normal distributions are used for the outcome y from the i^{th} patient, in the s^{th} study, and p^{th} population.

In Fig. 3, the outcome y has a mean that depends on covariates, $\mathbf{x}$, which can contain a treatment indicator. The study-level regression coefficients in the p^{th} population come from the same multivariate normal distribution $N(\mu_{\beta_p}, \mathbf{R}_p)$. The between-study precision matrices $\mathbf{R}_p$ describe the extent of borrowing among studies within each population and have a prior Wishart distribution with k degrees of freedom. The degrees of freedom must be set to a minimum value (depending on the dimension of β_{sp}) in order for certain moments of the associated inverse-Wishart distribution

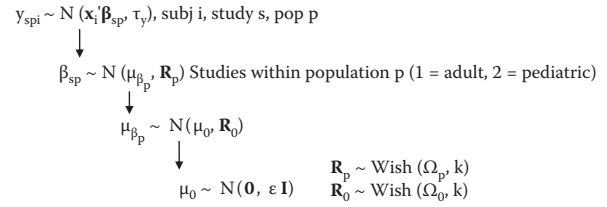


Fig. 3 Three-level hierarchical model for borrowing across populations and studies within populations, adjusting for covariates

to exist (see appendix A in Gelman et al.^[1] for details). The between-population precision matrix $\mathbf{R}_0$ describes the extent of borrowing between the two populations. When only one pediatric study is available, however, there is no information about the variation of the $\beta_{s,2}$ coefficients across pediatric studies (i.e., $\mathbf{R}_2$). In most practical device applications, there will not be more than one pediatric study. A simple solution would be to set $\mathbf{R}_1 = \mathbf{R}_2$. Note that the general model in Fig. 3 can be reduced to the usual two-level hierarchical model by setting the between-population precision matrix $\mathbf{R}_0$ to be “infinite.” That is, there is no extra population variability beyond what appears across the studies.

A common situation occurs when the effect of a treatment on a pediatric population is of a lower (or higher) magnitude than on an adult population, due to the pediatric population having a generally lower or higher mean level of a covariate that influences the treatment effect. Kovalchik, Varadhan, and Weiss^[15] discuss a proportional interactions model (PIM), which can be extended to a three-way interaction among treatment, covariate, and population. The covariate interacts with treatment, but the pediatric population has a different covariate mean level than does the adult population. The PIM can be modified so that the regression coefficients vary by study within populations but that proportionality of the treatment by covariate interaction only varies across populations. This modification could apply to any hierarchical model where there is a population level, a study level, and a patient level, along with measured covariates.

Obviously, there may be numerous ways to borrow from adult data, including methods that are not Bayesian. In this section, we have discussed some general approaches to deal with multiple populations and multiple studies. The Guidance “Leveraging Existing Clinical Data for Extrapolation to Pediatric Uses of Medical Devices”^[13] should be consulted for further information and references.

CONCLUSION

Among the most advantageous features of the Bayesian approach for medical device trials are the design and analysis of adaptive trials using predictive distributions and the capability to incorporate prior information into the analysis.

For adaptive trials, predictive probability can be used to determine how to adapt the trial based not only on the available information but also on the predictive distribution for the outcomes to be observed. Such adaptation has been frequently used for sample size modification using intermediate endpoints. The model to relate the intermediate and final endpoints is of paramount importance to the performance of the adaptive design. It should be carefully chosen at the study design stage, preferably based on prior knowledge. Simulation can be used to demonstrate appropriate operating characteristics for adaptive trials.

Bayesian hierarchical modeling enables synthesizing information from various related data sources with appropriate amounts of uncertainty. Such an approach for borrowing information has been used for formal utilization of historical data and for extrapolation of adult data for inference about pediatric use of the same device. The prior for the variance between studies is very influential to the performance of the study design. It should be carefully considered, preferably supported by simulation showing clinically desirable operating characteristics.

The advantages and flexibility in Bayesian methods are reflected in their increasing use in the medical device arena.^[16] In situations where there is prior information that can be used to augment the current trial, or where a flexible adaptive clinical trial is being considered, Bayesian methods often provide the best analytic approach to the problem.

REFERENCES

1. Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.; Vehtari, A.; Rubin, D.B. *Bayesian Data Analysis*; Chapman & Hall: London, 2014.
2. Spiegelhalter, D.; Best, N.; Carlin, B.; van der Linde, A. Bayesian measures of model complexity and fit (with discussion). *J. Roy. Stat. Soc. B* **2002**, *64* (4), 583–640.
3. Ohlssen, D.; Sharples, L.; Spiegelhalter, D. Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons. *Stat. Med.* **2007**, *26* (9), 2088–2112.
4. Carlin, B.P.; Louis, T.A. *Bayesian Methods for Data Analysis*; Chapman & Hall: London, 2009.
5. Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. Available at <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm> (accessed April 2016).
6. Adaptive Designs for Medical Device Clinical Studies. Available at <http://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm446729.pdf> (accessed October 2016).
7. Berry, S.M.; Carlin, B.P.; Lee, J.J.; Muller, P. *Bayesian Adaptive Methods for Clinical Trials*; CRC Press: Boca Raton, FL, 2011.
8. Ibrahim, J.G.; Chen, M.-H. Power prior distributions for regression models. *Stat. Sci.* **2000**, *15* (1), 46–60.
9. Chen, M.-H.; Ibrahim, J.; Lam, P.; Yu, A.; Zhang, Y. Bayesian design of noninferiority trials for medical devices using historical data. *Biometrics* **2011**, *67* (3), 1163–1170.
10. Hobbs, B.P.; Carlin, B.P.; Mandrekar, S.J.; Sargent, D.J. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* **2011**, *67* (3), 1047–1056.
11. Hobbs, B.P.; Sargent, D.J.; Carlin, B.P. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear model. *Bayesian Anal.* **2012**, *7* (3), 639–674.
12. Pennello, G.; Thompson, L. Experience with reviewing Bayesian medical device trials. *J. Biopharm. Stat.* **2008**, *18* (1), 81–115.
13. Guidance for Leveraging Existing Clinical Data for Extrapolation to Pediatric Uses of Medical Devices. Available at <http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM444591.pdf> (accessed June 2016).
14. Prevost, T.; Abrams, K.; Jones, D. Hierarchical models in generalized synthesis of evidence: An example based on studies of breast cancer screening. *Stat. Med.* **2000**, *19* (24), 3359–3376.
15. Kovalchik, S.; Varadhan, R.; Weiss, C. Assessing heterogeneity of treatment effect in a clinical trial with the proportional interactions model. *Stat. Med.* **2013**, *32* (28), 4906–4923.
16. Campbell, G.; Yue, L.Q. Statistical innovations in the medical device world sparked by FDA. *J. Biopharm. Stat.* **2016**, *26* (1), 3–16.

Binary 2×2 Crossover Trials

Li Zhu, Juan Li, and Eric Chi

Amgen Inc., Thousand Oaks, California, U.S.A.

Abstract

This entry reviews the general features of crossover designs and specific features of 2×2 crossover design and discusses in detail four typical tests of this sort.

INTRODUCTION

Crossover design is a commonly used experimental design for the comparison of treatment effects in clinical trials. By letting subject serve as his or her own control, such design automatically removes subject effect from the comparison. Here we focus on the case where the response to each treatment is binary, namely, each subject either succeeded or failed to response to treatment during certain treatment period. Historically, a variety of statistical tests have been developed to serve different purposes. For example, in some crossover trials with the objective of studying the treatments difference, a special test is desired if the response is affected by the order in which the treatments are applied. A thorough review of these tests can enhance our understanding of the methodologies, and hence aid the selection of optimal test in various practical settings.

This article is organized as follows. The general features of crossover designs are briefly reviewed in section “Crossover Design,” followed by an overview of 2×2 crossover design and some typical tests for binary crossover data in section “Tests for 2×2 Crossover Design”; four typical tests are discussed in detail in the following sections “The McNemar Test,” “The Mainland-Gart Test,” “The Prescott Test,” and “The Hills-Armitage Test”; in the end, some concluding remarks on the comparison of the four tests are included in the discussion section.

CROSSOVER DESIGN

As an appealing alternative to parallel designs, crossover designs are commonly applied in today’s pharmaceutical industry. In contrast to parallel designs where each subject receives one single treatment throughout the study, in crossover trials, subjects are given two or more different treatments in some order. Crossover design represents a special situation where there are no separate comparison groups, and each subject serves as his or her own control. Also, because the same subject receives both treatments, there is no possibility of covariate imbalance. One of

the advantages of crossover design is that it eliminates the between-subject variation. Consequently, it requires smaller sample size than parallel design with the same level of significance in testing the treatment effect. Crossover trials are frequently used to compare treatments for medical conditions that are chronic in nature, subject to temporary relief during treatment, and have a tendency to recur after withdrawal of treatment.^[1] They are also commonly used where the between-subject variation is assumed to be large and/or where subjects are scarce or expensive to use.^[2]

A potential problem in crossover design is the *carryover effect*, or *residual effect*, which is the effect that “carries over” from one study period to the subsequent period in which the treatment is changed. Sometimes a wash-out period is set between the treatment periods to minimize any possible carryover effect. But the appropriateness of setting a wash-out period is still in debate for ethical reasons.^[3] In crossover trials, if a *treatment $\times$ period interaction* is present, the differences between treatments as measured in one period may be different from those measured in a later period. Such treatment $\times$ period interaction may result from the true interaction, namely, the conditions present in the different periods truly affect the size of the differences between treatments; nevertheless the treatment $\times$ period interaction may also be a result of treatments possessing different amount of carryover effects. Generally, designs for more than two treatments can be constructed in such a way that they allow the interaction and carryover effect to be estimated separately, however, it is not easy to differentiate the carryover effect from the interaction under the designs of 2×2 crossover for two treatments.^[3] More details will be discussed in the next section.

TESTS FOR 2×2 CROSSOVER DESIGN

The simplest and most commonly used crossover design is the 2×2 crossover design, in which some subjects will receive the standard therapy first, followed by the new therapy (AB), the others will receive the new therapy first,

followed by the standard therapy (BA), and the durations of treatment period are the same for both groups. Some baseline measurements are taken prior to the start of the trial. Then the trial is usually preceded by a run-in period, which is used for screening purposes and to familiarize subjects with the procedures they will follow during the trial. And between the treatment periods, there is usually a wash-out period to allow the effect of the first period treatment to be washed out of the body before the second period in which the treatment is switched. The procedure is illustrated in Fig. 1.

In many situations it is common to have binary outcomes, typically representing success (denoted as 1) or failure (denoted as 0) of a treatment. The data for a 2 × 2 crossover trial can be represented in a 2 × 4 array of counts as shown in Table 1, where n_{ij} 's are the number of subjects in i th sequence treatment group with j th type of paired response.

In the literature, there are specialized testing procedures proposed to test for specific effects, including treatment, period, treatment × period interaction, and group-by-dependence interaction. Also, the different procedures allow the tests to use different types of outcomes. Some typical binary 2 × 2 crossover tests are summarized in Table 2.

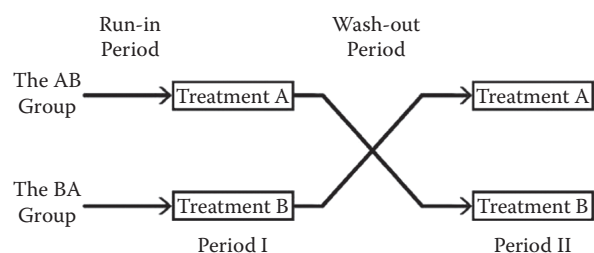


Fig. 1 Diagram of 2 × 2 crossover design

Table 1 2 × 2 Crossover Data

	(0,0)	(0,1)	(1,0)	(1,1)	Total
AB	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
BA	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{..}$

Table 2 Some Binary 2 × 2 Crossover Tests

Test name	Effect to test	Outcomes used	Assumptions		
			Period difference	Treatment × Period interaction	Group-by-dependence interaction
McNemar	Treatment	(0,1), (1,0)	×	×	×
Mainland-Gart	Treatment/Period	(0,1), (1,0)	√	×	×
Prescott	Treatment	(0,0), (1,0) (0,1), (1,1)	√	×	×
Hills-Armitage	Treatment × Period Interaction	(0,0), (1,1)	NA	√	NA

The absence of a treatment × period interaction is a critical assumption in most crossover trials. In the absence of such an interaction, the proportion of (1,1) observations among the concordant pairs should be comparable in the two groups. Hills and Armitage^[4] suggest that a test for this interaction by only using non-preference outcomes.

As mentioned in section “Crossover Design,” the treatment × period interaction and carryover effect cannot be disentangled in 2 × 2 crossover trials. This is because there are only three degrees of freedom under the 2 × 2 crossover design; in other words, we can only include three parameters at most without causing unidentifiability. Out of the three parameters, two are associated with the treatment effect and period effect, so only one is left to be associated with the carryover effect and interaction.^[3] Therefore the treatment × period interaction and carryover effect parameters are intrinsically aliased with each other, and cannot be separated except by *a priori* assumption, such as thinking of the carryover as negligible.^[5]

The dependence structure of the paired outcomes is a question of interest in crossover trials, and is tested as *group-by-dependence interaction*, which characterizes how the response is affected by the allocation of treatment sequences to other patients.^[6] Namely, the group-by-dependence interaction measures the difference of treatment × period interaction between the groups.^[2] So if there is no group-by-dependence interaction, the odds ratio of observing a success in the first period as opposed to that in the second period under the AB group should equal to the odds ratio under the BA group.^[7]

THE McNEMAR TEST

McNemar test is a non-parametric method used on nominal data to determine whether the row and column marginal frequencies are equal. It is applied to 2 × 2 contingency tables with a dichotomous trait with matched pairs of subjects (Table 3). Cells represented in the following manner by the letters a , b , c , and d , the totals across rows and columns marginal totals, and the grand total is represented by n .

Table 3 A 2 × 2 Table Used to in the McNemar Test

		Matched case #2		
		1	0	Total
Matched Case #1	1	<i>a</i>	<i>c</i>	<i>a</i> + <i>c</i>
	0	<i>b</i>	<i>d</i>	<i>b</i> + <i>d</i>
	Total	<i>a</i> + <i>b</i>	<i>c</i> + <i>d</i>	<i>n</i>

Table 4 Discordant Pairs of Outcomes from Table 1

	(0,1)	(1,0)	Total
AB	<i>n</i> ₁₂	<i>n</i> ₁₃	<i>n</i> _{1,2+3}
BA	<i>n</i> ₂₂	<i>n</i> ₂₃	<i>n</i> _{2,2+3}
Total	<i>n</i> _{.2}	<i>n</i> _{.3}	<i>n</i> _{..2+3}

Marginal homogeneity implies that the row totals are equal to the corresponding column totals, which means no treatment effect. The null hypothesis can be simplified as $b = c$.

The McNemar statistic Z is a chi-squared statistic with 1 degree of freedom:

$$Z = \frac{(b - c)^2}{b + c} \sim \chi_1^2.$$

In 2 × 2 crossover studies, if the response to each treatment is independent of the order of application, i.e., in the absence of a period effect, the McNemar test^[8] can be used. Only the discordant pairs of the outcomes as shown in Table 4 are used in the McNemar test.

To conduct the McNemar test, the original data is first converted into the structure as in Table 5. Under the null hypothesis of no treatment effect, given that a subject has shown a preference, the choice is equally likely to have been for either treatment. Out of $n_p = n_{.2} + n_{.3}$ subjects who show a preference, a total of $n_A = n_{13} + n_{22}$ of them prefer treatment A, and a total of $n_B = n_{12} + n_{23}$ of them prefer treatment B. Under the null hypothesis, n_A follows a binomial distribution, i.e.,

$$n_A \sim \text{binomial} \left(n_p, \frac{1}{2} \right), \text{ with mean } E(n_A) = \frac{n_p}{2}, \text{ and}$$

$$\text{Var}(n_A) = \frac{n_p}{4}.$$

Hence, $\frac{Z = (n_A - (1/2)n_p)^2}{\text{Var}(n_A)} = \frac{(n_A - n_B)^2}{n_p}$ has an asymptotic χ_1^2 distribution.

For demonstration purpose, we use Mosteller's data^[9] as an example to perform the McNemar test, as well as all the other tests discussed in this article. Table 6 shows the original Mosteller's data.

Here $n_A = 5 + 4 = 9$, $n_B = 0 + 1 = 1$, and $n_p = 4 + 6 = 10$. Because $n_A \sim \text{binomial}(10, 1/2)$, we have $p = 0.0107$ and

/* McNemar Test */	
data McNemar_data;	
input PrefA \$ PrefB \$ count;	
datalines;	
no no	30
no yes	9
yes no	1
yes yes	60
;	
proc freq data = McNemar_data;	
weight count;	
tables PrefA * PrefB / agree;	
exact mcnemar;	
run;	
McNemar's Test	
Statistic (S)	6.4000
DF	1
Asymptotic Pr > S	0.0114
Exact Pr >= S	0.0215

Fig. 2 SAS code and output for the McNemar test example

doubling this for a two-sided test we get an exact probability of 0.0214. Using the asymptotic method, we compare

$$Z = \frac{(n_A - n_B)^2}{n_p} = \frac{(9 - 1)^2}{10} = 6.4$$

with the χ_1^2 distribution to get a two-sided probability of 0.0114.

The McNemar test can be carried out using SAS frequency procedure. The appropriate SAS code to analyze these data is shown in Fig. 2 with the corresponding output.

Therefore, the results from the McNemar test suggests that the treatment effect is significantly present in this data.

THE MAINLAND-GART TEST

In 2 × 2 crossover trials, each subject receives the two treatments in successive periods, with the order of application being determined randomly. Mainland presents a method of comparing the treatments in the presence of period difference.^[10] Gart^[11] has provided a theoretical interpretation of Mainland's approach in light of a logistic model. This approach was later called the Mainland-Gart test. The data type used in this test is the same as that used in the McNemar test. Namely, the Mainland-Gart test uses only the discordant pairs of the outcomes as illustrated in Table 4.

The theoretical derivation of the test is base on the following logistic models. Let the Bernoulli random variable Y_{ijk} denote the k th subject in the i th treatment group that has the j th type of paired outcome, then we have the logistic models

$$\Pr(Y_{13k_1} = 1) = \frac{e^{\beta_{k_1} + \lambda + \delta}}{1 + e^{\beta_{k_1} + \lambda + \delta}}, \quad \Pr(Y_{22k_2} = 1) = \frac{e^{\beta_{k_2} - \lambda + \delta}}{1 + e^{\beta_{k_2} - \lambda + \delta}},$$

$$\Pr(Y_{12k_1} = 1) = \frac{e^{\beta_{k_1} - \lambda - \delta}}{1 + e^{\beta_{k_1} - \lambda - \delta}}, \quad \Pr(Y_{23k_2} = 1) = \frac{e^{\beta_{k_2} + \lambda - \delta}}{1 + e^{\beta_{k_2} + \lambda - \delta}},$$

Table 5 Data Structure in the McNemar Test

		Prefer B	
		Yes	No
Prefer A	Yes	$n_{.4}$	$n_A = n_{13} + n_{22}$
	No	$n_B = n_{12} + n_{23}$	$n_{.1}$

Table 6 Example: Mosteller's Data

	(0,0)	(0,1)	(1,0)	(1,1)	Total
AB	13	0	5	32	50
BA	17	4	1	28	50
Total	30	4	6	60	100

with $k_1 = 1, \dots, n_{1,2+3}; k_2 = 1, \dots, n_{2,2+3}$. And the parameter λ stands for the period effect, δ is for the treatment effect, and β 's stand for the pair or block effects.

To test the treatment effect, given the total number of successes observed under each treatment, and the total number of subjects in each group, the conditional distribution finally can be derived as hypergeometric (univariate) based on the logistic models aforementioned. Hence the test is equivalent to Fisher's exact test for the usual 2×2 table, i.e., for the test with null hypothesis

$$H_0 : \delta = 0,$$

we have

$$f(n_{13}, n_{22} | n_{13}, n_{23}; \delta = 0, \lambda) = \frac{\binom{n_{1,2+3}}{n_{13}} \binom{n_{2,2+3}}{n_{23}}}{\binom{n_{.,2+3}}{n_{13} + n_{23}}},$$

and the p -value for the one-tailed exact test is

$$P = \frac{\sum_{i=n_{13}}^{n_{1,2+3}} \binom{n_{1,2+3}}{i} \binom{n_{2,2+3}}{n_{13} + n_{23} - i}}{\binom{n_{.,2+3}}{n_{13} + n_{23}}}. \quad (1)$$

On the other hand, we can use the similar idea to test for the period effect. Unlike testing for treatment effect, where the total number of successes observed within each period is fixed, here what is assumed to be fixed is the total number of successes observed under each treatment. Correspondingly, we have

$$H_0 : \lambda = 0,$$

Table 7 Example: Mosteller's Data with Discordant Pairs of Outcomes

	(0,1)	(1,0)	Total
AB	$n_{12} = 0$	$n_{13} = 5$	$n_{1,2+3} = 5$
BA	$n_{22} = 4$	$n_{23} = 1$	$n_{2,2+3} = 5$
Total	$n_{.2} = 4$	$n_{.3} = 6$	$n_{.,2+3} = 10$

$$f(n_{13}, n_{22} | n_{13} + n_{22}; \delta, \lambda = 0) = \frac{\binom{n_{1,2+3}}{n_{13}} \binom{n_{2,2+3}}{n_{22}}}{\binom{n_{.,2+3}}{n_{13} + n_{22}}},$$

$$P = \frac{\sum_{i=n_{13}}^{n_{1,2+3}} \binom{n_{1,2+3}}{i} \binom{n_{2,2+3}}{n_{13} + n_{22} - i}}{\binom{n_{.,2+3}}{n_{13} + n_{22}}}. \quad (2)$$

Furthermore, the treatment effect and period effect can be tested simultaneously:

$$H_0 : \delta = 0, \lambda = 0,$$

by using

$$f(n_{13}, n_{22}; \delta = 0, \lambda = 0) = \binom{n_{1,2+3}}{n_{13}} \binom{n_{2,2+3}}{n_{22}} \left(\frac{1}{2}\right)^{n_{.,2+3}}.$$

If the sample size is large, the joint hypothesis may be tested by

$$\chi^2 = \frac{(2n_{13} - n_{1,2+3})^2}{n_{1,2+3}} + \frac{(2n_{22} - n_{2,2+3})^2}{n_{2,2+3}}.$$

Large values of the test statistic indicates some deviation from the null hypothesis in either or both of the two-sided alternative. For details please see Gart.^[11]

Here we demonstrate an application example by using the Mosteller's data quoted earlier. Table 7 contains only the discordant pairs of outcomes from the full data in Table 6.

From Eq. 1, the one-tailed exact test for treatment effect then yields the p -value:

$$P = \frac{\binom{5}{5} \binom{5}{5+1-5}}{\binom{10}{5+1}} = 0.0238 \quad (3)$$

The result suggests a significant treatment effect. The conclusion is the same as that from McNemar test as shown early. The corresponding SAS code and output are in Fig. 3, and the result is same to that from Eq. 3.

```

/* Mainland-Gart test */
data Gart_data;
input trt_group $ response $ count;
datalines;
AB 01 0
AB 10 5
BA 01 4
BA 10 1
;
proc freq data = Gart_data;
weight count;
tables trt_group * response / chisq;
run;

```

Fisher's Exact Test		
Cell (1,1) Frequency (F)		0
Left-sided Pr <= F		0.0238
Right-sided Pr >= F		1.0000
Table Probability (P)		0.0238
Two-sided Pr <= P		0.0476
Sample Size = 10		

Fig. 3 SAS code and output for the Mainland-Gart test example

Similarly, by Eq. 2, the one-tailed exact test for period effect gives the p -value

$$P = \frac{\binom{5}{5} \binom{5}{5+4-5}}{\binom{10}{5+4}} = 0.5 \quad (4)$$

From Eq. 4, it is indicated that the period effect is not significant based on this data.

THE PRESCOTT TEST

An alternative approach was proposed by Prescott to test for differences between treatments in the presence of period effects.^[12] The approach starts to consider the difference $d = x_1 - x_2$, where x_1 and x_2 are the pair of binary outcomes of a subject from the two treatment period. Then d can take the value 1, 0, -1. The simplest method for testing parametrically the null hypothesis of equality of treatment effects is the t -test.^[13] Using data in Table 1, form T as the difference between the responses in the first and second periods

$$T = n_{13} - n_{12},$$

then use this test statistic to compare the mean difference in the AB sequence with the mean difference in the BA

Table 8 Example: Mosteller's Data with Tied Outcomes

	(0,1)	(0,0) or (1,1)	(1,0)	Total
AB	0	45	5	50
BA	4	45	1	50
Total	4	90	6	100

Table 9 All Possible Configurations in the Prescott Test Example

0	45	5	50
4	45	1	50
4	90	6	100
0	44	6	50
4	46	0	50
4	90	6	100
1	43	6	50
3	47	0	50
4	90	6	100

marginal values. So the Prescott test is essentially a permutation t -test. For the one-tailed test against the alternative that treatment A is more effective, given the row and column totals being fixed, we calculate the probability P of observing any configuration with $T > T_0$, where T_0 denotes the observed value of T . Then

$$P = \frac{\sum_{i=T_0}^{n_3} \sum_{j=0}^{\min(i-T_0, n_2)} \binom{n_3}{i} \binom{n_1+n_4}{n_1-i-j} \binom{n_2}{j}}{\binom{n_{..}}{n_{1.}}}$$

We demonstrate the use of the Prescott test by the following example.^[12] The original data in Table 6 with tied outcomes is in Table 8:

For the Prescott test, the test statistic $T = 5 - 0 = 5$. For fixed marginal values, any of the three configurations in Table 9 satisfy $T \geq 5$:

Thus, the p -value for one-tailed test is

$$Pr(T \geq 5) = \frac{\binom{6}{5} \binom{90}{45} \binom{4}{0} + \binom{6}{6} \binom{90}{44} \binom{4}{0} + \binom{6}{6} \binom{90}{43} \binom{4}{1}}{\binom{100}{50}} = 0.0110$$

sequence. To apply the t -test to our binomial situation, this approach suggests to use the same test statistic T , and determine its distribution by considering the values we would obtain for all possible permutations given the fixed

The conclusion is the same as that from both the McNemar and the Mainland-Gart tests.

Using Cytel Studio™ (CrossOver) similar results as shown in Table 10 can be obtained.

Table 10 Results of the Prescott Test for Treatment Effect from Cytel Studio™

Type	One-Sided <i>p</i> -value	Two-Sided <i>p</i> -value
Asymptotic	0.0060	0.0118
Exact	0.0111	0.0222
Mild <i>P</i> Corrected	0.0061	0.0171

Table 11 Results of the Prescott Test for Period Effect from Cytel Studio™

Type	One-Sided <i>p</i> -value	Two-Sided <i>p</i> -value
Asymptotic	0.2577	0.5154
Exact	0.3732	0.7463
Mild <i>P</i> Corrected	0.2687	0.6419

For testing the period effect, results from Cytel Studio™ (CrossOver) are shown in Table 11. Again, they are consistent with the result from the Mainland-Gart test calculated in Eq. 4.

THE HILLS-ARMITAGE TEST

Most of the tests that are commonly used in the 2 × 2 crossover trials are based on the assumption of absence of treatment × period interaction. Such an assumption can be tested by the Hills-Armitage test^[4,13] that uses only the tied outcomes, i.e., the non-preference observations. The non-preference observations as a part of Table 1 are tabulated in Table 12. The test is developed based on the rationale that the effect of an interaction would lead to the difference in the mean response from the two groups A and B, which would be presented by the relative proportions of subjects falling into the categories (0,0) and (1,1). In other words, the frequencies in Table 12 would be disproportionate if a treatment × period interaction exists. This could be tested in the usual way of testing independence in 2 × 2 contingency tables—chi-square test for large sample and exact test for small sample. Examples are given in the following two paragraphs respectively.

Large sample test of independence in contingency tables such as the Pearson chi-square, the continuity-adjusted

Table 12 2 × 2 Table of Non-Preference Observations for the Hills-Armitage Test

	(0,0)	(1,1)	Total
AB	n_{11}	n_{14}	$n_{1,1} + 4$
BA	n_{21}	n_{24}	$n_{2,1} + 4$
Total	$n_{.1}$	$n_{.4}$	$n_{.,1} + 4$

chi-square for 2 × 2 tables, likelihood-ratio chi-square, and the Mantel-Haenszel chi-square are discussed in details in Agresti.^[14] Here we take the usual Pearson chi-square test for example. The Pearson chi-square statistics based on Table 12 can be written as

$$\chi^2 = \sum_{i=1,2} \sum_{j=1,4} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

Where $\hat{\mu}_{ij} = \left(\frac{n_{i,1+4}}{n_{.,1+4}} \right) \left(\frac{n_{.j}}{n_{.,1+4}} \right)$. Suppose the non-preference data from Table 9 is actually listed below in Table 13. The chi-square statistics calculated as

$$\chi^2 = \frac{(13-15)^2}{15} + \frac{(32-30)^2}{30} + \frac{(17-15)^2}{15} + \frac{(28-30)^2}{30} = 0.8$$

Given the degree of freedom being 1 for the 2 × 2 table, the corresponding *p*-value is 0.3711.

Small sample test for the 2 × 2 table of non-preference observation can be carried out by Fisher's exact test. The cutoff probability based on Table 12 is calculated using the formula

$$P_c = \frac{\binom{n_{1,1+4}}{n_{11}} \binom{n_{2,1+4}}{n_{21}} \binom{n_{.1}}{n_{11}} \binom{n_{.4}}{n_{21}}}{\binom{n_{.,1+4}}{n_{11}} \binom{n_{.,4}}{n_{21}} \binom{n_{.1}}{n_{11}} \binom{n_{.4}}{n_{21}}}$$

This is the hypergeometric distribution. Taking the same data in Table 13 for example, the cutoff probability is

$$P_c = \frac{(45!)(45!)(30!)(60!)}{(90!)(13!)(32!)(17!)(28!)} = 0.1196$$

In order to get the *p*-value for the exact test, we need to calculate the probability in the same way for each possible 2 × 2 data outcome that gives the same margins as that of the observed data in Table 13. And then add up all the extreme probabilities, namely, the probabilities that are less than the cutoff probability. The summation of extreme probabilities will serve as the *p*-value. When sample size is not very small, statistical software is preferred since the analytical calculation can get tedious.

Since the Hills-Armitage test is essentially the usual test for independence of 2 × 2 tables, it can be simply implemented in the SAS frequency procedure. Using the data in Table 13, both the small sample test and the large sample

Table 13 Example Data of Non-Preference Observations

	(0,0)	(1,1)	Total
AB	13	32	45
BA	17	28	45
Total	30	60	90

```

/* Hills-Armitage test */
data tied;
input trt_group $ response $ count;
datalines;
AB 00 13
AB 11 32
BA 00 17
BA 11 28
;
proc freq data = tied;
weight count;
tables trt_group * response;
exact chisq;
run;

```

Chi-Square		0.8000
DF		1
Asymptotic Pr > ChiSq		0.3711
Exact Pr >= ChiSq		0.5027

Fig. 4 SAS code and output for the Hills-Armitage test example

approximation can be conducted. As shown in Fig. 4, if we use the sample SAS code, we can get the results that are the same as previously calculated.

As shown in the output, both the p -value for the exact test and that for the large sample asymptotic test are larger than 0.05, hence the treatment × period interaction is not present in the data in Table 13.

DISCUSSION

McNemar test, Mainland-Gart test, Prescott test, and Hills-Armitage test are the most commonly used tests in 2 × 2 crossover clinical trials with binary outcomes. Hills-Armitage test focuses on testing treatment × period interaction, and the other three tests focus on testing different effects.

The McNemar test is an extremely simple method, but it has very strong assumptions of no period effect and no carry-over effect. In real-life clinical trials, the assumptions do not always hold. To adjust for period effect, the Mainland-Gart test and the Prescott test are more appropriate.

The Mainland-Gart test is derived from the logistic model, the parameters of individual effects are nuisance and eliminated by having $x_1 + x_2 (=1)$ as fixed for each individual, because only the discordant pairs of the data are used. On the other hand, the Prescott test is a non-parametric test, i.e., it is totally distribution-free. The methods reviewed in the previous sections are appropriate for the small sample cases. The versions of large sample test are also provided in the original paper.^[11,12] In terms of study designs, the Mainland-Gart test does not utilize the information that subjects were allocated at random to the two-treatment sequences, so it is applicable to a wider range of studies than the Prescott test. On the other hand,

the Prescott test is directly based on the randomization that will take place in any well-conducted trial. And it has been shown that under the usual randomization setting, the Prescott test is more powerful than the tests that only use the discordant pair of outcomes, including the McNemar test and the Mainland-Gart test. Therefore, the choice of which test to use should depend on the problem of interest and the specific experimental situation.

The Hills-Armitage test is useful to test potential treatment × period interaction. Since the absence of the treatment × period interaction is the assumption of most tests used in binary crossover trials, it is recommended to always use Hills-Armitage test to check whether treatment × period interaction exists.

ACKNOWLEDGMENTS

The authors would like to thank Moraye Bear for her support and helpful guidance.

REFERENCES

1. Backer, M.P.; Balagtas, C.C. Marginal modeling of binary cross-over data. *Biometrics* **1993**, *49*, 997–1009.
2. Kenward, M.G.; Jones, B. A log-linear model for binary cross-over data. *Appl. Stat.* **1987**, *36*, 192–204.
3. Jones, B.; Kenward, M.G. *Design and Analysis of Cross-over Trials*, 2nd Ed.; Chapman and Hall/CRC: London, 2003.
4. Hills, M.; Armitage, P. The two-period cross-over clinical trial. *Br. J. Clin. Pharmacol.* **1979**, *8*, 7–20.
5. Cox, D.R. Interaction. *Int. Stat. Rev.* **1984**, *52*, 1–31.
6. Cox, M.A.A.; Plackett, R.L. Matched pairs in factorial experiments with binary data. *Biom. J.* **1980**, *22*, 697–702.
7. Farewell, V.T. Some remarks on the analysis of crossover trials with binary response. *Appl. Stat.* **1985**, *34*, 121–128.
8. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157.
9. Mosteller, F. Some statistical problems in measuring the subjective response to drugs. *Biometrics* **1952**, *8*, 220–226.
10. Mainland, D. *Elementary Medical Statistics*, 2nd Ed.; Saunders: Philadelphia, PA, 1963.
11. Gart, J.J. An exact test for comparing matched proportions in cross-over designs. *Biometrika* **1969**, *56*, 75–80.
12. Prescott, R.J. The comparison of success rates in crossover trials in the presence of an order effect. *Appl. Stat.* **1981**, *30*, 9–15.
13. Armitage, P. *Statistical Methods in Medical Research*; Blackwell: Oxford, 1971.
14. Agresti, A. *Categorical Data Analysis*; John Wiley & Sons, Inc.: New York, 1990.

Bioassay

Anant M. Kshirsagar

University of Michigan, Ann Arbor, Michigan, U.S.A.

Abstract

The general purpose of a bioassay is to study the potency of a dose from the response it produces. This entry provides statistical methods of performing a bioassay as well as measuring its results.

INTRODUCTION

Finney^[1] defines a bioassay as an experiment for estimating the potency of a drug, material, preparation, or process by means of the reaction that follows its application to living matters. The emphasis in bioassays is on comparing the potencies of treatments rather than estimating the difference between the effects of treatments. For example, Finney points out that an investigation into the effects of different samples of corticotrophin on the ascorbic acid in rat adrenals is not necessarily an assay; it becomes one if the interest lies in using the changes in ascorbic acid for estimating the potencies of the samples in standard units of corticotrophin.

OVERVIEW

The typical bioassay involves a stimulus (for example, a vitamin or a drug) applied to a subject (animal, tissue, or some such experimental unit). The level of the stimulus can be varied and the effect of the stimulus on the subject can be measured in terms of a characteristic we call “response.” The relationship between stimulus and response is a statistical relation subject to a random error. The relationship can be used to study the potency of a dose from the response it produces.

The estimate of potency is always relative to a “standard” preparation of the stimulus. A new test preparation of the stimulus is then assayed to find the mean response to a selected drug. Next we find the dose of the standard preparation that produces the same mean response. The ratio ρ of the dose Z_s of the standard preparation equally effective as this dose Z_M of the test preparation is the relative potency of the test preparation. Thus if $\rho = 4$, 1 unit of the test preparation is as effective as 4 units of the standard preparation.

A statistician must choose a sound design for assaying one or more test preparations. He has to choose the number of levels of doses of each preparation, the number of subjects to be tested at each dose, and the allocation

of the treatments to the units. Sound statistical principles such as randomization, replication, elimination of extraneous sources of variation blocking, and covariates must be employed in designing the experiment for the assay. The data collected are then analyzed and the relative potencies of the test preparation estimated. The statistical techniques that play an important role are analysis of variance and covariance, regression analysis, and linear statistical inference in general. It is necessary to consider carefully the implicit and explicit assumptions about the data and their effect on the conclusions.

There are three types of biological assays: direct assays, indirect assays, and assays based on quantal (“all or nothing”) responses. Direct assays are of fundamental importance. Here the response is measured for a number of doses to establish a statistical dose–response relation, and from this the potency of a test preparation is determined.

PARALLEL-LINE BIOASSAYS

An important type of bioassay is the parallel-line bioassay, in which the dose–response relation is linear for the standard as well as the test preparation and the lines are parallel or that they have a common slope. This linearity can often be achieved by a logarithmic or Box-Cox^[2] type of transformation of the dose and/or the response. The treatments are then applied to the experimental units by using a suitable experimental design, such as randomized blocks, Latin squares, crossover, split-plot, or even an incomplete block design (such as a balanced incomplete block design, partially balanced incomplete block design, or a lattice design). Whether linearity of the dose–response relation is achieved or not, and, if so, whether the slopes of the lines for the standard and test preparations are the same or not, is tested first by what are known as tests of validity, which include the test of parallelism also.

Thus if t_s , t_M denote, respectively, the effects of the standard (S) and the test (M) preparation, the parallel lines representing the dose–response relation can be taken as

$$t_s = \alpha_s + \beta x_s$$

$$t_M = \alpha_M + \beta x_M$$

where x_s, x_M are, respectively, $\log Z_s$ and $\log Z_M$, and Z_s, Z_M are the doses of s and M in original units. If ρ is the relative potency of M with respect to s , then ρZ_M and Z_s produce the same effect, and this leads to

$$\log \rho = \frac{\alpha_M - \alpha_s}{\beta}$$

This can be expressed in terms of the treatment contrasts, using

$$\alpha_M - \alpha_s = \bar{t}_M - \bar{t}_s - \beta(\bar{x}_M - \bar{x}_s)$$

$$\beta = \frac{\sum t_M (x_M - \bar{x}_M) + \sum t_s (x_s - \bar{x}_s)}{\sum (x_M - \bar{x}_M)^2 + \sum (x_s - \bar{x}_s)^2}$$

where $\bar{t}_M, \bar{t}_s$ etc. are the averages and the summation is over the doses of S and M . The contrast $\alpha_M - \alpha_s$ is called the *preparation contrast*; the contrast β is called the *regression contrast*. The contrasts can be estimated from the observations in the experimental design. In complete block designs, these estimates are obtained simply by replacing the treatment effects by their average yields; in incomplete block designs, however, one needs to use adjusted treatment yields for this response, after confirming that the contrasts are estimable. In fact, designs should be chosen so that these two contrasts are unconfounded and are estimated with maximum precision. Mukerjee and Gupta^[3] have studied efficient designs for this purpose and have attempted to unify the results in this area. Finney^[1] and Das and Kulkarni^[4] give several examples of complex and incomplete block designs to be used in parallel-line bioassays. The validity tests are based on orthogonal polynomial contrasts of the treatment effects, such as

$$\sum P_{Sr} t_s \text{ and } \sum P_{Mr} t_s$$

where P_{Sr}, P_{Mr} are orthogonal polynomials of the r th degree in the values of the doses of S and M . In a well-planned experiment, the doses are chosen to be equal in number and also equidistant on a log scale so that tables of orthogonal polynomial can be used to construct these contrasts. If reduction in block size is inevitable, one has to resort to confounding of validity contrasts of higher degrees. Kshirsagar and Yuan^[5] have developed a unified theory of parallel-line bioassays in incomplete block designs. Wang and Kshirsagar^[6] give an illustration of the use of lattice squares—a two-way design with incomplete rows and columns, to estimate the relative potency of several preparations. Recovery of interblock or interrow and intercolumn information can also be

utilized in increasing the precision of the estimates of the preparation and regression contrasts if the blocks (i.e., rows and columns) are chosen at random. A confidence interval for ρ can then be obtained by using Fieller's method^[7] for the ratio of the two contrasts in defining $\log \rho$. The assumption of normal distribution and homogeneity of variances of errors is necessary for carrying out this analysis.

If the linearizing transformation of dose is not logarithmic but of the type $X = Z^\lambda$, the relative potency turns out to be $\rho = (\beta_M/\beta_s)^{1/\lambda}$, provided the dose-response lines of S and M intersect and have slope β_M, β_s for M and S , respectively, and provided the intercepts of α_M, α_s of the two lines are equal. The validity of these assumptions in such slope-ratio assays will have to be tested by testing the departure from linearity of the dose-response regression lines and a test of significance of $\alpha_M - \alpha_s$. It is expected that when $x_s = x_M = 0$, the response should be the same for S and M , and this is tested by introducing "blanks" (neither S nor M) in the experiment.

Völund^[8] has discussed bioassays in the case of multivariate responses. In the case of multiple responses and multiple doses (in which the test or standard preparation is made up of several component drugs or chemicals), composite functions such as discriminant functions or canonical variables could be used, as discussed by Finney^[1] and by Yuan and Kshirsagar.^[9,10]

QUANTAL RESPONSE AND THE TOLERANCE DISTRIBUTION

Sometimes, it is not possible to measure the response and all that can be done is to record whether or not the subject manifests a certain reaction. The response in such cases is quantal, and can be death or any other easily recognizable change in the subjects. Such a change could be irreversible, and a subject can, therefore, be used only once. Such a situation arises in the assaying of insecticides, fungicides, or pesticides or assays of insulin or estrogenic hormones or assays of vitamins. The tolerance of a subject is then defined as the minimum dose that would produce the desired effect. In a direct assay, the tolerance is measured for each subject and the potency is estimated by comparing mean tolerance. In a quantal assay, a selected dose is used, and one can only record whether or not the change occurred and infer from that whether the tolerance was greater or less than the dose used. The potency in quantal assays is estimated by use of the relation between the dose and the proportion of subjects responding. This proportion does not necessarily have a linear regression on dose or its transform. It is customary to assume a suitable distribution for tolerance or log tolerance for approximate linearization. The potency is sometimes characterized by the median lethal dose or, more generally, the median effective dose, symbolized, respectively, by LD50 and ED50. This is a dose that, on an average, produces the desired response in 50% of the subjects. For estimating ED50, the parameters of the tolerance

distribution are estimated first from the data, by using the maximum likelihood estimation method or the minimum χ^2 method or some such suitable method. The assumption of a normal distribution for tolerance leads to what is known as *probit analysis* and the assumption of a logistic distribution leads to *logit analysis*. If $f(z)$ is the probability density function of the tolerance distribution, the probability p that the tolerance of a subject is less than x is

$$p = \int_{-\infty}^x f(z) dz = \int_{-\infty}^y \phi(z) dz$$

where $y = (x - \mu)/\sigma$, μ , σ^2 being, respectively, the mean and variance of the normal distribution of Z and $\phi(z)$ is the p.d.f. of a standard normal variable. If the distribution function of a standard normal distribution is denoted by Φ and its inverse by Φ^{-1} , we then have

$$\Phi^{-1}(p) = \alpha + \beta x$$

where α , β are transforms of μ and σ . One can then estimate ED50 or LD50 by the probit method. The method of scoring^[11] is generally used to solve the maximum-likelihood equations. Several tolerance distributions, such as normal, logistic, binomial, Wilson-Worcester, and angle sigmoid, have been discussed and employed in the literature, and one may refer to Finney,^[1] Govindrajulu,^[12] or Hubert^[13] for further details. The iterative calculations needed to solve the maximum-likelihood equations are laborious, and alternative, simpler but approximate, methods were developed earlier. These include the Spearman-Kärbes, Dragstedt-Behrens, and Reed-Muench.^[1] However, they have lost their merit due to the availability of ready-made computer programs.

Finney^[1] has discussed the principles of planning an assay. He stresses the importance of validity or consistency of an estimate of the relative potency, the economics of the assay design, the necessity of pilot investigations before choosing an optimum design, the simplicity of a symmetric design, and the cost of statistical analysis. He has given specific detailed recommendations for parallel-line and slope ratio assays as well as quantal assays.

For some assays, the measured response is time: usually the time that elapses between the application of the stimulus and the occurrence of some reaction. In such time-response assays the main difficulty is that the assay may end before a response has been measured for every subject. Finney^[1] suggests converting the data to a quantal focus or assigning an arbitrary value as the response for subjects that have not reacted when the assay ends or using a mathematical model for reaction time in such cases.

The relative potency of a test preparation may also change with time due to storage or environmental conditions and the dependence on time of the relative potency can be studied using a growth-curve type of model for the preparation and slope contrast, as in Bandekar and Kshirsagar,^[14] for example. Multivariate bioassays have been discussed by Carter

and Huber,^[15] Williams,^[16] Hui and Rosenberg,^[17] Vølund,^[8] Meisner et al.,^[18] and Srivastava.^[19] For the Bayesian analysis of parallel-line and slope-ratio bioassays, reference may be made to Darley^[20] and Meudoza.^[21] Distribution-free methods for the estimation of relative potency have been discussed by Sen^[22] and Bennett.^[23]

REFERENCES

1. Finney, D.J. *Statistical Methods in Biological Assay*, 3rd Ed.; Charles Griffin: London, 1978.
2. Box, G.E.P.; Cox, D.R. An analysis of transformations. *J. R. Stat. Soc., B* **1964**, *26*, 211–252.
3. Mukerjee, R.; Gupta, S. A efficient design for bioassays. *J. Stat. Plan. Inference* **1995**, *48*, 247–259.
4. Das, M.N.; Kulkarni, G.A. Incomplete block designs for bioassays. *Biometrics* **1966**, *22*, 706–729.
5. Kshirsagar, A.M.; Yuan, W. A united theory for parallel line bioassays in incomplete block designs. *Int. J. Math. Stat. Sci.* **1992**, *1*, 151–171.
6. Wang, W.; Kshirsagar, A.M. Use of lattice square designs in bioassays. *J. Biopharm. Stat.* **1996**, *6*, 185–199.
7. Fieller, E.C. A fundamental formula in the statistics of biological assay and same applications. *Q. J. Pharm. Pharmacol.* **1994**, *17*, 117–123.
8. Vølund, A. Multivariate bioassay. *Biometrics* **1980**, *36*, 225–236.
9. Yuan, W.; Kshirsagar, A.M. Analysis of multivariate parallel-line bioassay with composite responses and composite doses, using canonical correlations. *J. Biopharm. Stat.* **1993**, *3*, 57–71.
10. Kshirsagar, A.M.; Yuan, W. Estimation of relative potency in multivariate parallel line bioassays. *Commun. Stat., Theory Methods* **1993**, *22*, 3355–3361.
11. Rao, C.R. *Linear Statistical Inference and Its Applications*, 2nd Ed.; Wiley: New York, 1973.
12. Govindrajulu, Z. *Statistical Techniques in Bioassay*; Kargesm: New York, 1988.
13. Hubert, J.J. *Bioassay*; Kendall-Hunt: Dubuque, IA, 1980.
14. Bandekar, R.; Kshirsagar, A.M. Time-dependent relative potency. Submitted for publication.
15. Carter, E.M.; Hubert, J.J. Analysis of parallel-line assays with multivariate responses. *Biometrics* **1985**, *41*, 703–710.
16. Williams, D.A. An exact confidence region for relative potency. *Biometrics* **1978**, *34*, 659–661.
17. Hui, S.L.; Rosenberg, S.H. Multivariate slope ratio assay with repeated measurements. *Biometrics* **1985**, *41*, 11–37.
18. Meisner, M.; Kushner, H.B.; Laska, E.M. Combining multivariate bioassays. *Biometrics* **1986**, *42*, 421–427.
19. Srivastava, M.S. Multivariate bioassays, combination of bioassays, and Fieller's theorem. *Biometrics* **1986**, *42*, 131–141.
20. Darley, S.C. A Bayesian approach to parallel line bioassay. *Biometrika* **1980**, *67*, 607–612.
21. Meudoza, M. A Bayesian analysis of slope ratio bioassay. *Biometrics* **1990**, *46*, 1059–1069.
22. Sen, P.K. On the estimation of potency in dilution (direct) assays by distribution free methods. *Biometrics* **1963**, *19*, 532–552.
23. Bennett, B.M. Distribution free methods in bioassay results. *J. Hyg.* **1970**, *76*, 147–162.

Bioavailability and Bioequivalence

Mengdie Yuan and Jingyu Julia Luan

Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

Bioavailability/Bioequivalence (BA/BE) studies are often conducted in support of marketing applications for drugs and biologics. The Hatch-Waxman Act particularly enhanced the generic drug marketing and the evolution of BE studies after 1984. The Code of Federal Regulations lists pharmacokinetic (PK) study, pharmacodynamic (PD) study, clinical endpoint study, and in vitro study as methods for determining BA/BE. In PK studies, three approaches are used to demonstrate BE: average BE, population BE, and individual BE. With the development of generic products for highly variable drugs and narrow therapeutic index drugs, reference-scaled BE approaches are employed. For certain types of drugs, clinical endpoint studies may be the most appropriate way to demonstrate BE, for which the approaches to demonstrating BE are also well developed.

Keywords: Average bioequivalence; Comparative clinical endpoint study; Comparative PK study; Generic drugs; Highly variable drugs; Individual bioequivalence; Narrow therapeutic index drugs; Population bioequivalence.

INTRODUCTION

The US Code of Federal Regulations 21 CFR 320^[1] describes the bioavailability (BA) and bioequivalence (BE) requirements of the US Food and Drug Administration (FDA). It defines BA as “the rate and extent to which the active ingredient or active moiety is absorbed from a drug product and becomes available at the site of action” and BE as “the absence of a significant difference in the rate and extent to which the active ingredient or active moiety in pharmaceutical equivalents or pharmaceutical alternatives becomes available at the site of drug action when administered at the same molar dose under similar conditions in an appropriately designed study.” BA/BE studies are often conducted in support of US marketing applications including Investigational New Drug Applications (INDs), New Drug Applications (NDAs), Biologics Licensing Applications (BLAs), and Abbreviated New Drug Applications (ANDAs). BA studies aim to assess the in vivo pharmacokinetic (PK) performance for drug absorption, distribution, and elimination for a new molecular entity, while BE studies aim to compare the BA of various formulations with regard to the rate and extent of absorption. BE studies occur in both the generic drug development and the new drug development with formulation or strength change. In generic drug development, demonstrating BE became key to allaying concerns that the generic was not as bioavailable as the reference drug.^[2] The Hatch-Waxman Act (1984) established the modern system of generic drugs, authorizing the FDA to approve generic drugs under an

abbreviated process that did not require repetition of the research done by the innovator that established safety and efficacy, enhancing the development of generic drugs and the evolution of BE studies. BE studies are essential in the evaluation of therapeutic equivalence between the test drug product and the reference drug product. Indeed, the two products are considered as therapeutically equivalent, which means that the test product and the reference product are expected to have the same safety and efficacy, if they meet the regulatory criteria of pharmaceutical equivalence and BE. Under 21 CFR 320.24,^[1] the FDA may accept evidence from various approaches to determine BA or BE, including the following:

1. An in vivo test (PK study) in humans in which the concentration of the active ingredient or active moiety, and, when appropriate, its active metabolite(s), in whole blood, plasma, serum, or other appropriate biological fluid are measured as a function of time, or an in vitro test that has been correlated with and is predictive of human in vivo BA data
2. An in vivo test in humans in which the urinary excretion of the active moiety, and, when appropriate, its active metabolite(s), are measured as a function of time
3. An in vivo test (pharmacodynamics [PD] study) in humans in which an appropriate acute pharmacological effect of the active moiety, and, when appropriate, its active metabolite(s), are measured as a function of time if such effect can be measured with sufficient accuracy, sensitivity, and reproducibility

4. Well-controlled clinical trials that establish the safety and effectiveness of the drug product, for the purpose of measuring BA, or appropriately designed comparative clinical trials, for the purpose of demonstrating BE
5. A currently available in vitro test acceptable to FDA that ensures human in vivo BA
6. Any other approach deemed adequate by FDA to measure BA or establish BE

The methods are listed in descending order of accuracy, sensitivity, and reproducibility. Method 1 is particularly applicable to dosage forms intended to deliver the active moiety to the bloodstream. Method 2 is not appropriate where urinary excretion is not the main mechanism of elimination. Method 3 is used when Methods 1 and 2 are not applicable and a method is available for measurement of the acute pharmacological effect. Method 4 is the least accurate, sensitive, and reproducible of the general approaches for measuring BA or demonstrating BE; it is acceptable only when the aforementioned methods cannot be developed and may be considered sufficient when BA/BE is to be determined for dosage forms intended to deliver the active moiety locally, oral dosage forms not intended to be absorbed, or bronchodilators administered by inhalation if the onset and duration of pharmacological activity are defined.^[1]

In general, plasma PK studies provide useful information in assessing BA and demonstrating BE. For locally acting products such as topical dermatological products or those nasal sprays and inhalation products that are intended for local action, traditional plasma PK studies can be useful to evaluate systemic exposure and safety but may not necessarily be relevant for evaluating BE because the drug levels in the systemic circulation may not necessarily be representative of local BA of the drug at or near the site of action. In some situations, comparative clinical endpoint studies have been used to determine BE.^[3] Shah et al. specifically introduced the methods to assess the BA/BE for topical drugs and discussed the challenges.^[4] In this article, we will focus on discussing the study designs and statistical approaches to demonstrating BE in comparative PK studies and comparative clinical endpoint studies.

PK BE STUDY

Study Design

PK BE studies aim to compare the systemic exposure profile of a test drug product (*T*) to that of a reference drug product (*R*). A standard PK BE study design is based on the administration of either single or multiple doses of *T* and *R* to healthy subjects with random assignment to two possible sequences of drug product administration. FDA recommends performing BE studies as single-dose PK

studies for both immediate-release and modified-release drug products because such studies are generally sensitive in assessing the release of the drug substance from the drug product into the systemic circulation.^[5] Useful multiple-dose studies have also been discussed in 21 CFR 320.27.^[1] El-Tahtawy et al. compared single-dose and multiple-dose studies in BE assessment.^[6] Also, clearance, volume of distribution, and absorption in the study are assumed to have less inter-occasion variability compared to the formulation variability so that differences contributed by formulation can be determined.^[5] Generally, non-replicated crossover study designs are recommended for BE studies. However, replicated crossover study designs have the advantage of accounting for the within-subject variability when comparing *T* and *R*, and also the advantage of reducing the number of subjects. In particular, they allow the use of methods based on within-subject variability to assess the BE such as reference-scaled approaches for highly variable (HV) drugs and narrow therapeutic index drugs, as discussed below. Examples of commonly used nonreplicated crossover study design and replicated crossover study designs are shown in Tables 1 and 2. Usually, consecutive periods are separated by an adequate washout period to eliminate the carryover effect. In situations where carryover effects are present, higher order crossover design is recommended.^[7] In some situations such as when the drug has a long half-life, a parallel design may be more appropriate due to the likelihood of dropouts during the required long washout period.^[2] For discussions on more types of BE study design, the interested reader can refer to Chapter 3.4 in Yu and Li.^[2] In addition, the study should enroll enough subjects to sufficiently power the study. Phillips provided a table of sample sizes based on Schuirmann's^[15] two one-sided tests (TOSTs).^[8] Later Liu and Chow provided the approximate formulas for calculation of the sample sizes.^[9]

Table 1 Nonreplicated Design

		Period	
		1	2
Two-period two-sequence crossover			
Sequence	1	<i>T</i>	<i>R</i>
	2	<i>R</i>	<i>T</i>

Table 2 Replicated Design

		Period			
		1	2	3	
Three-way crossover					
Sequence	1	T	R	T	
	2	R	T	R	
		Period			
Four-way crossover		1	2	3	4
Sequence	1	T	R	T	R
	2	R	T	R	T

PK Measures

In BA/BE studies, the blood or plasma concentration is measured after drug administration. Typically, 12–18 samples at various time points will be collected per subject per dose. The concentration–time curve is adopted to profile the drug absorption. As discussed, BE studies aim to compare the BA of various formulations with regard to the rate and extent of absorption. Instead of directly comparing the measures of absorption such as rate constant, rate profile, mean absorption time, mean residence time, peak drug concentration (C_{max}), and time to achieve peak concentration (T_{max}), we normally compare the measures of systemic exposure such as peak drug concentration (C_{max}) area under the curve (AUC). In particular, partial AUC ($pAUC$) truncated at the population median of T_{max} of the reference product can measure the early exposure to the drug; C_{max} can directly measure the peak exposure to the drug; and AUC from time 0 to the last sample point (AUC_{0-t}) or AUC from time 0 to infinity can measure the total exposure ($AUC_{0-\infty} = AUC_{0-t} + C_t/\lambda_z$), where C_t is the concentration at the last sample point and λ_z is a calculated terminal rate constant. The choice of the PK measures for comparison depends on the dosage forms and should be able to capture the relevant differences in the PK profile. The measures of peak exposure and total exposure are commonly used to demonstrate BE. The measures of early exposure are useful where a better control of the drug absorption is essential for drug efficacy and safety.^[10–12] Usually, C_{max} and AUC are informative for immediate-release and modified-release dosage forms, but are not always sufficient for some modified-release dosage forms that exhibit multiphasic PK behavior where the use of $pAUC$ may be appropriate.^[2] Moreover, AUC over a dosing interval at steady state is an appropriate measure of total exposure for steady-state studies.

Statistical Approaches for BE Assessment

BE comparisons are based on an equivalence approach that specifies a function of parameters for testing and a decision criterion to establish BE. In addition, a logarithmic transformation of the PK measures is applied before statistical analysis. Initially, a variety of statistical methods were used to assess BE.^[13] Later, statistical approaches based on hypothesis testing were developed and are employed to this day.^[2] Specifically, the hypotheses of interest can be written as

$$H_0 : \Theta > \theta \text{ versus } H_A : \Theta \leq \theta,$$

where Θ is a function of parameters for comparison between T and R , and θ is the BE margin. BE is assessed by testing the hypotheses at a desired level of significance, say

5%; by rejecting the null hypothesis H_0 , one can conclude the BE between T and R . We conduct the hypothesis testing based on a linear mixed-effects model. Suppose μ_T and μ_R are the logarithmic population mean of the PK measure of interest (e.g., C_{max} , AUC) for T and R , σ_{BT}^2 and σ_{BR}^2 are the between-subject variances of that measure for T and R , and σ_{WT}^2 and σ_{WR}^2 are the within-subject variances for T and R , respectively. The total variances are $\sigma_{TT}^2 = \sigma_{BT}^2 + \sigma_{WT}^2$ and $\sigma_{TR}^2 = \sigma_{BR}^2 + \sigma_{WR}^2$ for T and R , respectively. We allow a constant within-subject correlation between T and R denoted as ρ . The variance component of subject-by-formulation interaction can then be written as

$$\sigma_D^2 = (\sigma_{BT} - \sigma_{BR})^2 + 2(1 - \rho)\sigma_{BT}\sigma_{BR}.$$

In the situation where a small subgroup of the general population have markedly different relative BA of T and R from the general population, or T and R are not bioequivalent for this subgroup but might be bioequivalent for the general population, a subject-by-formulation interaction should be considered.

Currently, there are three approaches to demonstrating BE described in FDA Guidance: average BE (ABE), population BE (PBE), and individual BE (IBE).^[14] Note that ABE is the most frequently used approach. Next, we will introduce these three approaches and their properties.

Average BE

For ABE, we are interested in $\Theta_A = (\mu_T - \mu_R)^2$. Then the hypotheses for testing ABE are

$$H_0 : (\mu_T - \mu_R)^2 > \theta_A^2 \text{ versus } H_A : (\mu_T - \mu_R)^2 \leq \theta_A^2,$$

where θ_A is the ABE margin. FDA currently recommends $\theta_A = \ln(1.25)$.

Schuirman^[15] proposed to conduct TOSTs as follows:

$$H_0 : \mu_T - \mu_R < -\theta_A \quad \text{vs.} \quad H_A : \mu_T - \mu_R \geq -\theta_A$$

and

$$H_0 : \mu_T - \mu_R > \theta_A \quad \text{vs.} \quad H_A : \mu_T - \mu_R \leq \theta_A,$$

where θ_A is the BE margin. Therefore, ABE compares the logarithmic population mean of the PK measure. Note that it is equivalent to conducting hypothesis testing on the geometric mean ratio (GMR) which is the exponential of the difference in the logarithmic mean. Based on TOST, we construct a 90% confidence interval (CI) for estimated $\mu_T - \mu_R$. If the 90% CI is contained within $[-\theta_A, \theta_A]$, then BE is established. The rationale for the choice of the significance level (90%) and the margin (80%–125%) has been intensively discussed in the literature.^[2,16–19]

Population BE

The FDA^[14] recommends the following reference-scaled approach for testing PBE:

When $\sigma_{TR} > \sigma_{T0}$, use

$$\begin{aligned} H_0 : \frac{(\mu_T - \mu_R)^2 + \sigma_{TT}^2 - \sigma_{TR}^2}{\sigma_{TR}^2} > \theta_p \text{ versus} \\ H_A : \frac{(\mu_T - \mu_R)^2 + \sigma_{TT}^2 - \sigma_{TR}^2}{\sigma_{TR}^2} \leq \theta_p, \end{aligned} \quad (1)$$

and when $\sigma_{TR} \leq \sigma_{T0}$, use

$$\begin{aligned} H_0 : \frac{(\mu_T - \mu_R)^2 + \sigma_{TT}^2 - \sigma_{TR}^2}{\sigma_{T0}^2} > \theta_p \text{ versus} \\ H_A : \frac{(\mu_T - \mu_R)^2 + \sigma_{TT}^2 - \sigma_{TR}^2}{\sigma_{T0}^2} \leq \theta_p, \end{aligned}$$

where σ_{T0}^2 is a specified constant for the total variance and θ_p is the PBE margin. The determination of σ_{T0}^2 is based on the maximum allowable population difference ratio (PDR) and $\sigma_{TT}^2 - \sigma_{TR}^2$. Note that the true σ_{TT}^2 is unknown and cannot be observed in studies. Therefore, in practice, we usually compare the observed total standard deviation with a prespecified cutoff point. The choice of the margin θ_p should consider both the ABE and the variance term in PBE. The FDA recommended a PBE margin as

$$\theta_p = \frac{\theta_A^2 + \varepsilon_p}{\sigma_{T0}^2},$$

where ε_p is a variance factor, the value of which is guided by $\sigma_{TT}^2 - \sigma_{TR}^2$. We can construct the 95% upper confidence limit (UCL) for estimated $(\mu_T - \mu_R)^2 + \sigma_{TT}^2 - \sigma_{TR}^2 - \theta_p \sigma_{TR}^2$ or $(\mu_T - \mu_R)^2 + \sigma_{TT}^2 - \sigma_{TR}^2 - \theta_p \sigma_{T0}^2$ depending on the observed total standard deviation for R and compare the 95% UCL with zero. If the 95% UCL is less than or equal to zero, the PBE is established. Note that the interest of PBE lies in the comparison of the PDR, which can be expressed as

$$\begin{aligned} \frac{E(T - R)^2}{E(R' - R)^2} &= \frac{(\mu_T - \mu_R)^2 + \sigma_{TT}^2 + \sigma_{TR}^2}{2\sigma_{TR}^2} \\ &= \frac{1}{2} \times \frac{(\mu_T - \mu_R)^2 + \sigma_{TT}^2 - \sigma_{TR}^2}{\sigma_{TR}^2} + 1 \\ &= \frac{1}{2} \Theta_p + 1, \end{aligned}$$

where $\Theta_p = \frac{(\mu_T - \mu_R)^2 + \sigma_{TT}^2 - \sigma_{TR}^2}{\sigma_{TR}^2}$ is the function of parameters in Hypotheses (1).

PBE is relevant in the context of drug interchangeability when either T or R was prescribed to subjects for the first

time since PBE makes sure that not only the population mean but also the population variability is similar between T and R .^[2] Note that this approach shows discontinuity at the changeover point σ_{T0} . The 95% UCL will jump high at the changeover point as the observed standard deviation for R becomes larger. Also, this approach inflates the type I error rate slightly for a range of scenarios.^[14]

Individual BE

For IBE, we can also use a similar approach as PBE but with the interest lying in the comparison of the individual difference ratios (IDRs) by considering the within-subject variance instead of the total variance. That is,

When $\sigma_{WR} > \sigma_{W0}$, use

$$\begin{aligned} H_0 : \frac{(\mu_T - \mu_R)^2 + \sigma_{WT}^2 - \sigma_{WR}^2 + \sigma_D^2}{\sigma_{WR}^2} > \theta_I \text{ vs.} \\ H_A : \frac{(\mu_T - \mu_R)^2 + \sigma_{WT}^2 - \sigma_{WR}^2 + \sigma_D^2}{\sigma_{WR}^2} \leq \theta_I, \end{aligned} \quad (2)$$

and when $\sigma_{WR} \leq \sigma_{W0}$, use

$$\begin{aligned} H_0 : \frac{(\mu_T - \mu_R)^2 + \sigma_{WT}^2 - \sigma_{WR}^2 + \sigma_D^2}{\sigma_{W0}^2} > \theta_I \text{ vs.} \\ H_A : \frac{(\mu_T - \mu_R)^2 + \sigma_{WT}^2 - \sigma_{WR}^2 + \sigma_D^2}{\sigma_{W0}^2} \leq \theta_I, \end{aligned}$$

where σ_{W0}^2 is a specified constant for the within-subject variance and θ_I is the IBE margin. Similarly, the determination of σ_{W0}^2 is based on the maximum allowable IDR, $\sigma_{WT}^2 - \sigma_{WR}^2$, and σ_D^2 . The value of the θ_I can be chosen as

$$\theta_I = \frac{\theta_A^2 + \varepsilon_I}{\sigma_{W0}^2},$$

where ε_I is a variance factor, the value of which is guided by $\sigma_{WT}^2 - \sigma_{WR}^2 + \sigma_D^2$. The FDA recommends allowing 0.03 for $\sigma_{WT}^2 - \sigma_{WR}^2$ and 0.02 for σ_D^2 , i.e., 0.05 for ε_I .^[14] We can construct the 95% UCL for estimated $(\mu_T - \mu_R)^2 + \sigma_{WT}^2 - \sigma_{WR}^2 + \sigma_D^2 - \theta_I \sigma_{WR}^2$ or $(\mu_T - \mu_R)^2 + \sigma_{WT}^2 - \sigma_{WR}^2 + \sigma_D^2 - \theta_I \sigma_{W0}^2$ depending on the observed within-subject standard deviation for R and compare the UCL with zero. If the 95% UCL is less than or equal to zero, then IBE is established. Similarly, we can show that

IDR is a function of $\Theta_I = \frac{(\mu_T - \mu_R)^2 + \sigma_{WT}^2 - \sigma_{WR}^2 + \sigma_D^2}{\sigma_{WR}^2}$ in

Hypotheses (2).

IBE is useful in the context of drug switchability for a subject who has been taking the test product or the reference product to switch to the other one without compromising the safety or efficacy.^[2]

For feasible models and methods for constructing the 95% UCL for ABE, PBE, and IBE, refer to FDA Guidance.^[14]

Highly Variable Drugs

Drugs with within-subject % coefficient of variation (%CV) greater than 30% for at least one PK measure are considered as highly variable (HV) drugs. A survey of generic products from 2003 to 2005 has suggested that approximately 20% of generic drugs should be considered as HV due to their drug substance dispositional characteristics.^[20] HV drugs tend to have a wider 90% CI compared to normal drugs even with the same mean ratio, and therefore are less likely to pass the BE test. It may turn out that a dramatically large sample size is needed to meet the BE criterion. Given this concern, the FDA's Office of Generic Drugs raised the issue in an Advisory Committee Meeting in 2004, where a reference-scaled approach was recommended. Later, this approach was further investigated and finalized with simulation studies,^[21] and is referred to as reference-scaled average BE (RSABE). Note that the BE study should be a replicated study in order to use the RSABE approach. The RSABE approach for HV drugs is modified from the ABE approach and described next.

When $\sigma_{WR} > \sigma_{W0}$, use

$$H_0: \frac{(\mu_T - \mu_R)^2}{\sigma_{WR}^2} > \frac{\theta_A^2}{\sigma_{WR}^2} \quad \text{vs.} \quad H_A: \frac{(\mu_T - \mu_R)^2}{\sigma_{WR}^2} \leq \frac{\theta_A^2}{\sigma_{WR}^2},$$

and when $\sigma_{WR} \leq \sigma_{W0}$, use

$$H_0: \frac{(\mu_T - \mu_R)^2}{\sigma_{W0}^2} > \frac{\theta_A^2}{\sigma_{W0}^2} \quad \text{versus} \quad H_A: \frac{(\mu_T - \mu_R)^2}{\sigma_{W0}^2} \leq \frac{\theta_A^2}{\sigma_{W0}^2},$$

where σ_{W0}^2 is a specified constant for the within-subject variance and $\theta_A = \ln(1.25)$ is the ABE margin. The FDA proposed the use of cutoff point of $\sigma_{W0} = 0.294$ in the draft guidance on progesterone.^[22] In that guidance, the FDA describes a procedure for BE testing and provides an example of SAS codes to elaborate the RSABE approach. However, the RSABE approach may allow a product to pass the BE test even with the point estimate of the GMR outside [0.8, 1.25]. Therefore, the FDA^[23] recommends imposing a further constraint that the point estimate of the GMR should be contained within [0.8, 1.25]. The constraint is controversial especially when the within-subject %CV exceeds 50%–60% because the BE test will be dominated by the constraint, resulting in the need for an increased sample size to demonstrate BE.

However, the constraint is still adopted by regulatory agencies such as the FDA, European Medicines Agency (EMA), and Australian Therapeutic Goods Administration.^[2]

Narrow Therapeutic Index Drugs

For narrow therapeutic index (NTI) drugs, a small change in the blood concentration can potentially cause therapeutic failure or serious adverse events in the subject. The Code of Federal Regulations 21 CFR 320.33^[1] provides possible definitions of narrow therapeutic index: (1) there is less than a twofold difference in median lethal dose (LD50) and median effective dose (ED50) values, and (2) there is less than a twofold difference in the minimum toxic concentrations and minimum effective concentrations in the blood. There exist numerous reports raising concerns of the BE approaches for certain generic products, for example, levothyroxine, and the switchability of the generic product for the reference product. In particular, 13 US states have considered a list of specific NTI drugs nonsubstitutable.^[2] Therefore, more stringent BE approaches are needed. After careful consideration,^[24,25] the FDA recommends a four-way replicated crossover design to demonstrate BE for NTI drugs and an RSABE approach on mean comparison plus a within-subject variability comparison for BE testing^[2] (see, e.g., the draft guidance on warfarin sodium^[23]).

Mean Comparison

The hypotheses are

$$H_0: \frac{(\mu_T - \mu_R)^2}{\sigma_{WR}^2} > \theta \quad \text{versus} \quad H_A: \frac{(\mu_T - \mu_R)^2}{\sigma_{WR}^2} \leq \theta,$$

where θ is a reference-scaled BE limit. The FDA recommends the use of $\theta = \ln(1/0.9)^2 / \sigma_{W0}^2$ and $\sigma_{W0} = 0.1$ for warfarin sodium.^[23] For different products, θ could be different. Furthermore, the FDA recommends that the NTI drugs also pass the ABE in support of approval.^[2]

Within-Subject Variability Comparison

In addition to the mean comparison, we also need to test the following hypotheses for within-subject variability:

$$H_0: \frac{\sigma_{WT}}{\sigma_{WR}} > \delta \quad \text{vs.} \quad H_A: \frac{\sigma_{WT}}{\sigma_{WR}} \leq \delta,$$

where δ is a limit to control the test variability to be no more than the reference variability. The FDA proposed $\delta = 2.5$ for warfarin sodium.^[23]

In the guidance,^[23] the FDA also describes a procedure for conducting the two above comparisons and provides an example of SAS codes to elaborate the approach. Other agencies such as Health Canada and EMA have considered directly tightening the ABE limits, say [90.00%, 111.11%], for NTI drugs.^[2]

CLINICAL ENDPOINT BE STUDY

Study Design

As discussed earlier, for dosage forms intended to deliver the active moiety locally, oral dosage forms not intended to be absorbed, or bronchodilators administered by inhalation if the onset and duration of pharmacological activity are defined, plasma PK studies may not be relevant to evaluate BE.^[1] Instead, clinical endpoint studies can be used. Generally, clinical endpoint BE studies are randomized, double-blind, placebo-controlled three-arm, parallel groups designed to compare two formulations. The placebo arm is usually but not necessarily included in the clinical endpoint BE studies to evaluate if the study is sufficiently sensitive to detect the clinical effect in the study population. Patients intended to treat are enrolled in the clinical endpoint BE studies, while PK BE studies are conducted on healthy volunteers. Note that the number of subjects enrolled in clinical endpoint BE studies can be much smaller than that in the registration clinical trials in new drug development. Unlike demonstrating efficacy with the registration clinical trials of new drugs, clinical endpoint BE studies are a less sensitive, accurate, and reproducible approach to demonstrating BE compared to other approaches such as PK BE studies. Also, compared to other approaches to demonstrating BE, clinical endpoint BE studies are expensive and time consuming.^[2] However, under some circumstances, clinical endpoint BE studies may be the only choice to demonstrate BE.

Clinical Endpoints

Clinically meaningful endpoints should directly measure how a patient feels, functions, or survives, and represent or characterize the clinical outcome of interest. It is critical to ensure that the proposed clinical endpoint(s) are valid for evaluating treatment of the disease of interest; for a drug with multiple indications, the indication can be accurately diagnosed, and the treatment effect or clinical improvement with regard to this indication is well defined.^[2] In ANDA submissions, there are usually two types of clinical endpoints: continuous endpoints (e.g., number of lesions) and dichotomous endpoints (e.g., clinical success/failure). In this article, we will give a general introduction to the statistical approaches used in assessing BE with regard to these two types of clinical endpoints. See the work of Grosser et al.^[3] for a more comprehensive discussion on BE for generic locally acting drug products.

Statistical Approaches

Let us consider a parallel designed clinical endpoint BE study with three treatments: the test product (T), the reference product (R), and the placebo (P). To demonstrate

BE between T and R , one needs to show that T and R are equivalent, and that both T and R are superior to P .

Equivalence Analysis

The analysis of continuous clinical endpoints is similar to the analysis of PK endpoints, but under parallel design. The compound hypotheses to be tested are as follows:

$$H_0 : \frac{\mu_T}{\mu_R} \leq \theta_1 \text{ or } \frac{\mu_T}{\mu_R} \geq \theta_2 \quad \text{vs.} \quad H_A : \theta_1 < \frac{\mu_T}{\mu_R} < \theta_2,$$

where μ_T is the overall mean of T , μ_R is the overall mean of R , and θ_1 and θ_2 are the BE margins. Note that in generic drug development, θ_1 and θ_2 are recommended by the FDA in product-specified guidance (PSG) for every specific reference product. Generally, $\theta_1 = 0.80$ and $\theta_2 = 1.25$ are used. The null hypothesis H_0 is rejected if the 90% CI for the estimated mean ratio is contained within the interval $[\theta_1, \theta_2]$. Rejection of the null hypothesis H_0 supports the conclusion of equivalence of T and R . Note that for multicenter studies, it is important to investigate the site effect and the site-by-treatment interaction.

It is also worth mentioning that for nasal spray products, we usually adjust the analysis for the baseline value.^[26] Fieller's method^[27] is one of the commonly used methods to calculate the 90% CI for the estimated mean ratio in clinical endpoint BE studies. Exact methods of calculating the 90% CI have also been studied by Locke^[28] and Chow and Shao.^[29] Note that the parametric models generally assume that the data fit the normal distribution. If the assumption is violated, one alternative is to perform a nonparametric test such as rank test. However, it may not be appropriate if the quantity of clinical interest is the mean value instead of the median. Also, nonparametric methods are usually less powerful compared to parametric methods.

For dichotomous endpoints, we consider the compound hypotheses:

$$H_0 : p_T - p_R < -\theta \text{ or } p_T - p_R > \theta \quad \text{vs.} \\ \in H_A : -\theta \leq p_T - p_R \leq \theta,$$

where p_T is the success rate of T , and p_R is the success rate of R . The null hypothesis H_0 is rejected if the 90% CI for the difference in success rates between T and R is contained within the interval $[-\theta, \theta]$. Rejection of the null hypothesis H_0 supports the conclusion that T and R are equivalent. Wald test with Yates' correction^[30] is commonly used for calculating the 90% CI. Other methods such as Z-test have also been used in the literature, which produce similar results as Wald test.

Study Sensitivity

The superiority tests of each active treatment (T or R) over placebo should be conducted at a significance level of 0.05.

For continuous endpoint, we can use an analysis variance model to calculate the p -values for the superiority tests and compare them to 0.05. For dichotomous endpoints, Fisher's exact test or Chi-square test can be used to calculate the p -values for the superiority tests and compare them with 0.05. Failure of superiority tests suggests that the clinical endpoint BE study may not be sensitive enough to detect the difference between T and R .

DISCUSSION

This article introduced the definition of BA/BE, the development of BA/BE, and the methods and types of studies for determining BA/BE. Furthermore, we extensively discussed the study designs and statistical approaches to demonstrating BE in comparative PK studies and comparative clinical endpoint studies. We introduced the ABE, PBE, and IBE approaches in PK BE studies and their properties. Reference-scaled ABE approaches used for HV drugs and NTI drugs were also elaborated upon, referring to the FDA's product specific guidances. Typically used approaches for continuous endpoints and dichotomous endpoints in the clinical endpoint BE studies were presented for both equivalence analysis and superiority analysis. In vitro comparison was not discussed in this article. But it is worth pointing out that physicochemical characteristics can be identified and qualitatively (Q1) and quantitatively (Q2) compared for formulations that are chemically equivalent using in vitro studies. However, this method is only sensitive and reproducible when there is a reasonably well-characterized mechanism of action, and the specific contributions of the various physicochemical characteristics to the therapeutic effect of the active pharmaceutical ingredient at the site of action are well characterized. The interested reader can refer to the FDA's documents,^[31,32] Chow et al.,^[33] and Yu and Li.^[2]

Furthermore, certain products have other additional types of studies recommended to show equivalence between brand and generic. For example, in addition to a PK BE study, the development of transdermal delivery systems (TDSs) also requires two additional studies for efficacy and safety concerns: adhesion study, and skin irritation and sensitization (I/S) study. Adhesion study and skin I/S study are used not only in generic drug development but also in new drug development such as patch size modification. Instead of demonstrating T and R to be equivalent, these types of studies require demonstration that T is not inferior to R , which means T can be statistically better than R . In the adhesion study for generic TDS products, a new approach based on the mean difference has been proposed to replace the previous recommendation of using the mean ratio,^[34] and significantly refined the comparison criterion for well-adhering products. As we can see, the regulatory approaches have been well established for BE assessment. However, with more and more

generic products being developed, the traditional BE methods, e.g., "one-size-fits-all" rule, have been challenged and changed to move toward a more mature direction. The development of statistical approaches for BE assessment of HV and NTI drugs has amply borne this out. Innovative approaches and techniques are also emerging to deal with the new scientific and technical developments in generic products. For example, a new in vitro release test has been studied and recommended for BE assessment of acyclovir in FDA's recent guidance, providing a feasible in vitro comparison for products that are Q1 and Q2 same.^[34] Moreover, BE assessment has also been investigated for other types of drugs such as liposomal drug products, drug products acting locally within gastrointestinal tract, and long-acting injectable products. The interchangeability between a brand-name drug and its generic drugs, and the interchangeability among generic drugs have also been discussed.^[35–37] Many scientific and statistical issues in study design and BE assessment, such as missing data impact^[38] and choice of study population, need future work. For more information, the interested reader can refer to the works of Chow and Liu,^[39] Welage et al.,^[40] and Meredith,^[41] and references therein.

REFERENCES

- 21 CR 320. Bioavailability and Bioequivalence Requirements; <https://www.ecfr.gov/csg-bin/text-idx?SID=57590f8663a81254a39d8ee0f81b8ba0&mc=true&node=pt21.5.320&rgn=div5> (accessed in August 2017).
- Yu, L.X.; Li, B.V. *FDA Bioequivalence Standards*; Springer: New York 2014.
- Grosser, S.; Park, M.; Raney, S.G.; Ranton, E. Determining equivalence for generic locally acting drug products. *Stat. Biopharm. Res.* **2015**, *7* (4), 337–345.
- Shah, V.P.; Maibach, H.I.; Jenner, J. *Topical Drug Bioavailability, Bioequivalence, and Penetration*, 2nd Ed. Springer Science & Business Media: New York, NY, 2014.
- FDA. Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Considerations, 2003; https://www.fda.gov/ohrms/dockets/ac/03/briefing/3995B1_07_GFI-BioAvail-BioEquiv.pdf.
- El-Tahtawy, A.A.; Jackson, A.J.; Ludden, T.M. Comparison of single and multiple dose pharmacokinetics using clinical bioequivalence data and Monte Carlo simulations. *Pharm. Res.* **1994**, *11* (9), 1330–1336.
- Chow, S.C.; Liu, J.P. On assessment of bioequivalence under a higher-order crossover design. *J. Biopharm. Stat.* **1992**, *2* (2), 239–256.
- Phillips, K.F. Power of the two one-sided tests procedure in bioequivalence. *J. Pharmacokinet. Pharmacodyn.* **1990**, *18* (2), 137–144.
- Liu, J.P.; Chow, S.C. Sample size determination for the two one-sided tests procedure in bioequivalence. *J. Pharmacokinet. Biopharm.* **1992**, *20* (1), 101–104.
- Chen, M.L.; Davit, B.; Lionberger, R.A.; Wahba, Z.; Ahn, H.Y.; Yu, L.X. Using partial area for evaluation of

- bioavailability and bioequivalence. *Pharm. Res.* **2011**, 28 (8), 1939–1947.
11. Lionberger, R.A.; Raw, A.S.; Kim, S.H.; Zhang, X.; Yu, L.X. Use of partial AUC to demonstrate bioequivalence of zolpidem tartrate extended release formulations. *Pharm. Res.* **2012**, 29 (4), 1110–1120.
 12. Stier, E.M.; Davit, B.M.; Chandaroy, P.; Chen, M.L.; Fourie-Zirkelbach, J.; Jackson, A.; Kim, S.; Lionberger, R.A.; Mehta, M.; Uppoor, R.S.; Wang, Y.; Yu, L.X.; Conner, D.P. Use of partial area under the curve metrics to assess bioequivalence of methylphenidate multiphasic modified release formulations. *AAPS J.* **2012**, 14 (4), 925–926.
 13. Skelly, J.P. A history of biopharmaceutics in the Food and Drug Administration 1968–1993. *AAPS J.* **2010**, 12 (1), 44–50.
 14. FDA. Statistical Approaches to Establishing Bioequivalence, 2001; <https://www.fda.gov/downloads/drugs/guidances/ucm070244.pdf> (accessed in August 2017).
 15. Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Pharmacodyn.* **1987**, 15 (6), 657–680.
 16. Westlake, W.J. Use of confidence intervals in analysis of comparative bioavailability trials. *J. Pharm. Sci.* **1972**, 61 (8), 1340–1341.
 17. Westlake, W.J. Symmetrical confidence intervals for bioequivalence trials. *Biometrics* **1976**, 32 (4), 741–744.
 18. Westlake, W.J. Statistical aspects of comparative bioavailability trials. *Biometrics* **1979**, 35 (1), 273–280.
 19. Kirkwood, T.B.; Westlake, W.J. Bioequivalence testing—A need to rethink. *Biometrics* **1981**, 37 (3), 589–594.
 20. Davit, B.M.; Conner, D.P.; Fabian-Fritsch, B.; Haidar, S.H.; Jiang, X.; Patel, D.T.; Seo, P.R.; Suh, K.; Thompson, C.L.; Yu, L.X. Highly variable drugs: Observations from bioequivalence data submitted to the FDA for new generic drug applications. *AAPS J.* **2008**, 10 (1), 148–156.
 21. Haidar, S.H.; Davit, B.; Chen, M.L.; Conner, D.; Lee, L.; Li, Q.H.; Lionberger, R.; Makhlof, F.; Patel, D.; Schuirmann, D.J.; Yu, L.X. Bioequivalence approaches for highly variable drugs and drug products. *Pharm. Res.* **2008**, 25 (1), 237–241.
 22. FDA. Draft Guidance on Progesterone, 2011; <https://www.fda.gov/downloads/DrugS/GuidanceComplianceRegulatoryInformation/Guidances/UCM209294.pdf> (accessed in August 2017).
 23. FDA. Draft Guidance on Warfarin Sodium, 2012; <https://www.fda.gov/downloads/DrugS/GuidanceComplianceRegulatoryInformation/Guidances/UCM201283.pdf> (accessed in August 2017).
 24. Jiang, W.; Makhlof, F.; Schuirmann, D.J.; Zhang, X.; Zheng, N.; Conner, D.; Yu, L.X.; Lionberger, R. A bioequivalence approach for generic narrow therapeutic index drugs: evaluation of the reference-scaled approach and variability comparison criterion. *AAPS J.* **2015**, 17 (4), 891–901.
 25. Yu, L.X.; Jiang, W.; Zhang, X.; Lionberger, R.; Makhlof, F.; Schuirmann, D.J.; Muldowney, L.; Chen, M.L.; Davit, B.; Conner, D.; Woodcock, J. Novel bioequivalence approach for narrow therapeutic index drugs. *Clin. Pharmacol. Ther.* **2015**, 97 (3), 286–291.
 26. FDA. Guidance for Industry: Bioavailability and Bioequivalence Studies for Nasal Aerosols and Nasal Sprays for Local Action, 2003; <https://www.fda.gov/downloads/Drugs/.../Guidances/ucm070111.pdf> (accessed in August 2017).
 27. Fieller, E.C. Some problems in interval estimation. *J. R. Stat. Soc. Series B (Methodol.)* **1954**, 16 (2), 175–185.
 28. Locke, C.S. An exact confidence interval from untransformed data for the ratio of two formulation means. *J. Pharmacokinet. Pharmacodyn.* **1984**, 12 (6), 649–655.
 29. Chow, S.C.; Shao, J. An alternative approach for the assessment of bioequivalence between two formulations of a drug. *Biom. J.* **1990**, 32 (8), 969–976.
 30. Yates, F. Contingency tables involving small numbers and the chi-square test. *Suppl. J. R. Stat. Soc.* **1934**, 1 (2), 217–235.
 31. Martinez, M. In Vitro Bioequivalence Pathways; <https://www.fda.gov/downloads/AnimalVeterinary/NewsEvents/WorkshopsConferencesMeetings/UCM444127.pdf> (accessed in August 2017).
 32. Fahmy, R. The Use of In Vitro Bioequivalence as a Regulatory Approach, 2015; <https://www.fda.gov/downloads/AnimalVeterinary/NewsEvents/WorkshopsConferencesMeetings/UCM444128.pdf> (accessed in August 2017).
 33. Chow, S.C.; Shao, J.; Wang, H. In vitro bioequivalence testing. *Stat. Med.* **2003**, 22 (1), 55–68.
 34. FDA. Draft Guidance on Acyclovir, 2016; <https://www.fda.gov/downloads/DrugS/GuidanceComplianceRegulatoryInformation/GuidanceS/UCM428195.pdf> (accessed in August 2017).
 35. Zheng, J.; Chow, S.C.; Yuan, M. On assessing bioequivalence and interchangeability between generics based on indirect comparisons. *Stat. Med.* **2017**, 36 (19), 2978–2993.
 36. Chow, S.C.; Shao, J. Bioequivalence review for drug interchangeability. *J. Biopharm. Stat.* **1999**, 9 (3), 485–497.
 37. Gwaza, L.; Gordon, J.; Welink, J.; Potthast, H.; Hansson, H.; Stahl, M.; Garcia-Arieta, A. Statistical approaches to indirectly compare bioequivalence between generics: A comparison of methodologies employing artemether/lumefantrine 20/120mg tablets as prequalified by WHO. *Eur. J. Clin. Pharmacol.* **2012**, 68 (12), 1611–1618.
 38. Sun, W.; Zhou, L.; Grosser, S.; Kim, C. A meta-analysis of missing data and non-compliance data in clinical endpoint bioequivalence studies. *Stat. Biopharm. Res.* **2016**, 8 (3), 334–344.
 39. Chow, S.C.; Liu, J.P. Current issues in bioequivalence trials. *Drug Inf. J.* **1995**, 29 (3), 795–804.
 40. Welage, L.S.; Kirking, D.M.; Ascione, F.J.; Gaither, C.A. Understanding the scientific issues embedded in the generic drug approval process. *J. Am. Pharm. Assoc.* **1996**, 41 (6), 856–867.
 41. Meredith, P. Bioequivalence and other unresolved issues in generic drug substitution. *Clin. Ther.* **2003**, 25 (11), 2875–2890.

Bioinformatics

Sean D. Mooney

Buck Institute for Age Research, Novato, California, U.S.A.

Jeffrey T. Chang

Institute for Genome Sciences and Policy, Duke University, Durham, North Carolina, U.S.A.

Abstract

This entry covers several important areas in bioinformatics including sequence analysis, genomics, proteomics, systems biology, genotype-to-phenotype mapping, and structural bioinformatics.

Keywords: Bioinformatics; Computational biology; Genomics; Microarray; Natural language processing; Proteomics; Protein structure; Sequence alignment; Sequence analysis; Systems biology.

INTRODUCTION

Broadly, bioinformatics as a field includes research in genomics, proteomics, computer science, pharmacology, drug discovery, and drug development. This review covers several important areas in bioinformatics including sequence analysis, genomics, proteomics, systems biology, genotype-to-phenotype mapping, and structural bioinformatics.

OVERVIEW

Bioinformatics is a broad field that covers many areas of computational research. There is not universal agreement on the definition of bioinformatics. Most narrowly, it refers strictly to the development of algorithms for sequence analysis. It can, however, encompass the development and application of computational tools across the scope of biology, covering molecular biology, cellular biology, and genetics. In this entry, we take the wider view and provide an overview of the major areas of biomedical computation.

Bioinformatics is a relatively new field whose development coincides with advances in personal computing and the development of the Internet. The Internet has given bioinformatics researchers the ability to provide on-line tools and data sets to the research community at large. The National Center for Biotechnology Information (NCBI) is the National Institute of Health's (NIH) public resource for organizing knowledge in biology. It provides access to scientific literature abstracts and citations, curated sequence data sets, and many public data analysis tools. Furthermore, the last few years has seen an incredible rise in genome annotation databases, such as the University of California Santa Cruz Genome Database and the Ensembl Genome Databases.

Research in bioinformatics is disseminated through journals and conferences. The most popular bioinformatics journals include *Bioinformatics*, *PLoS Computational Biology*, the *Journal of Computational Biology*, and *BMC Bioinformatics*, although bioinformatics papers are also published in *Genome Research*, *Genome Biology*, the *Journal of Molecular Biology*, and other common biological journals. In contrast to research publications in some other biology and medical fields, conferences and conference publications are important venues for the publication of bioinformatics. These conferences include the core annual conferences Intelligent Systems for Molecular Biology (ISMB), Research in Computational Molecular Biology (RECOMB), the Pacific Symposium on Biocomputing (PSB), the Computational Systems Biology conference (CSB), and the IEEE International Conference on Bioinformatics and Bioengineering (BIBE). RECOMB and ISMB move from year to year, while PSB is held in Hawaii annually, and CSB is held annually at Stanford University. The most prominent society for bioinformatics is the International Society of Computational Biology (ISCB), which currently has more than 2,500 members.

SEQUENCE ANALYSIS

The rapid development of technology for gene sequencing, motivated by the Human Genome Project, has caused the price for DNA sequencing to plummet, while the quantity of sequence data climbs. Although sequence data is abundant, functional information about genes, such as cellular localization, organ of expression, or time of regulation, is labor-intensive and expensive to acquire. Thus, the major bioinformatic challenge is to develop approaches to store, retrieve, and interpret genetic sequence data. To manage this

data, GenBank provides a repository of all publicly available DNA sequence. It currently contains over 38 billion bases of DNA sequences and is growing exponentially (<http://www.ncbi.nih.gov/Genbank/GenbankOverview.html>). With the recent development of high-throughput sequencing technologies and the production of full genome sequences, these challenges remain as relevant as ever. Previously perfected algorithms are now being adapted and scaled for new uses such as full genome alignment.

Another current bioinformatic challenge is the support of so-called next generation sequencing technologies that can identify millions of short sequences that need to be either mapped to a template genome or assembled into a *de novo* genome. These technologies generate sequence “reads” of 25 bases to more than 1,000 bases, depending on the platform. Each technology has its own data analysis needs, although the underlying principles are the same among them. A fundamental algorithm that underlies sequence analysis compares the similarity between pairs of sequences.

PAIRWISE SEQUENCE COMPARISON

As a result of evolution, many organisms share common ancestors. Homologous sequences, or sequences that share an evolutionary ancestor, between organisms often share function, even if the organisms diverged long ago. For example, nearly half of the genes known to be responsible for human diseases have a counterpart in yeast. Thus the function of an unknown sequence may often be inferred if a homolog has been characterized in another species. The problem then becomes one of finding homologs.

Although the sequences of homologs undergo mutations, enough similarities often remain so that they can still be recognized. Methods to recognize homologs must quantify the similarity between two sequences, while also being sensitive to changes. Particularly for proteins, some changes are more likely than others. For example, evolution is more likely to accept substitutions between similar amino acids, such as changing one hydrophobic amino acid to another. Such substitutions are less likely to have deleterious effects on the structure or function of the protein. Thus, sequence comparison methods should account for the severity of the change in the amino acid.

This problem has been addressed using dynamic programming, an algorithm developed in computer science.^[1] The biological adaptation of this algorithm, called Smith-Waterman after the inventors, finds and scores the optimal alignment between two sequences. A sequence alignment shows the association between each amino acid in one sequence with either an amino acid or a gap in the other. To find the optimal alignment, Smith-Waterman requires (i) a matrix of amino acid substitution scores and (ii) penalty scores for the opening of gaps in a sequence.

The score matrix contains the number of points to award when any two amino acids are aligned, that is, substituted for another. Substitutions between similar amino acids yield higher scores than those between amino acids that are different. One of the oldest matrices still in wide use is the percent accepted mutation (PAM) matrix.^[2] The PAM matrix was created by calculating, from hand-aligned sequences of protein families, the rate at which each amino acid is substituted for another. There are variations of PAM matrices that differ based roughly on evolutionary distance; empirically, the best performing version is PAM250. In this matrix, substituting a LEU for an ILE, two hydrophobic amino acids, is 2. However, changing the LEU to an ASP yields a lower score of −2.

In addition to a score matrix, Smith-Waterman requires a gap penalty. To handle insertion or deletion mutations (for DNA sequences), or inserted domains (for protein sequences), gaps may need to be opened in one sequence. A gap in one sequence can be interpreted either as an insertion in that sequence or a deletion in the other. Gap penalties are calculated with *open* and *extension* parameters. The *open* parameter is the cost of creating a new gap, and the *extension* penalty is the cost of lengthening the gap. Typically, these are found empirically.

In general, Smith-Waterman can comfortably detect homologous protein sequences if they share at least 30% sequence identity.^[3]

A powerful application of Smith-Waterman is to search a database (e.g., all known DNA sequences) for homologs to an unannotated query sequence. Unfortunately, with the size of current databases, such a search is too time-consuming for general use. Therefore, Basic Local Alignment Search Tool (BLAST) searches databases by using a heuristic that approximates Smith-Waterman.^[4] These approaches have also been extended to mapping sequences to the genome sequences, particularly the 10^6 – 10^8 reads with new high throughput sequencing technologies.

FINDING SEQUENCE MOTIFS

Although pairwise sequence alignment can detect homology, another type of sequence analysis is to search for specific subregions of sequences. Often, only a small portion of a protein confers its function. Instead of examining the whole sequence, another approach is to identify first the subsequence responsible for an interesting function, and then create a computational description of those residues, such that the description can be used to recognize other proteins with the same function.

Although there are many ways to represent functional subsequences, the two most popular ones are regular expressions and statistical profiles. PROSITE, a database of protein families and domains, uses manually created regular expressions.^[5] Regular expressions are well

understood, and there are existing algorithms to quickly find instances of these patterns in sequences.

However, regular expression patterns may not encode the subtleties required to recognize some functional sites. Therefore, statistical profiles are also used. Here, hidden Markov models have been found to perform well.^[6] A hidden Markov model encodes information about a protein family in *states* corresponding to residues or groups of residues that share a property. At specific states, certain amino acids may be more acceptable than others. For example, one state may prefer positively charged amino acids and will be more likely to *emit* one than one that is not charged. Thus, a hidden Markov model consists of hidden states that produce observed outputs, which are the sequences that belong to the protein family. Algorithms exist to train a hidden Markov model (calculate the probabilities that states emit amino acids) from a family of proteins, and to score a sequence against the model to determine whether or not it belongs to that family.

COMPARATIVE GENOMICS

Inexpensive gene sequencing technologies have led to the completion of thousands of genomes, including *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*. The availability of genomes enables research in *comparative genomics*, which uses the genomes to discover biology, such as gene structure, gene regulation, and gene function. The main difficulty in comparing genomes is finding corresponding regions in the genomes of two organisms. Therefore algorithms have been developed to align genomes.^[7]

MICROARRAYS

Microarrays can quantitatively measure the relative abundance of short oligonucleotide sequences in a sample. There are two major technologies in use: one type, developed at Stanford, contains cDNA probes spotted on a glass chip, and the other type, developed at Affymetrix, contains oligonucleotide probes synthesized on a chip using a photolithography process. They were originally developed to measure simultaneously the expression levels of all genes in a sample of cells. With microarrays, scientists can assay for the genes that are expressed under experimental conditions, such as starvation, heat shock, or developmental stages. In each experiment, for each gene, microarrays produce a value whose magnitude is proportional to the amount of expressed mRNA present in the sample. However, microarray technology is not specific to gene expression and has also been applied to measure DNA in diverse genomic and epigenomic applications such as the presence of sequence polymorphisms, occupancy of transcription factors, or epigenetic methylation.

Although they have been in use for only decade, they represent an explosion in the amount of biological data available to the researcher.

One application of microarray data is to distinguish between subtypes of tumors.^[8] Identifying the correct tumor subtype can help doctors choose treatments that are more likely to be effective. Because the two main subtypes of diffuse large B-cell lymphoma have different genetic profiles, algorithms can differentiate a tumor based on its profile. Scientists characterized the expression profiles of tumors by using a method called *clustering*.^[9]

Clustering is often called an unsupervised method because it does not require prior knowledge about the data being examined. Given a list of genes, each with an associated expression profile, a clustering algorithm can group genes that exhibit similar profiles. An expression profile for a gene is a list of the expression values from microarrays collected across many different conditions, such as time points in a cell cycle or environmental conditions such as nutrient deficiency. Expression profiles for genes are often represented as a matrix, with genes on one axis and conditions on the other.

To find clusters of genes that exhibit similar patterns of expression, a clustering algorithm employs a *distance function* that quantifies the *distance* between the expression profiles of two genes. The more similar the profiles are, the smaller the distance. Conversely, very different profiles will receive higher scores. Although many distance functions have been applied, the Euclidean distance is used most often.

Applying the distance function to each pair of possible genes results in a matrix that contains the distances between genes. Finally, the clustering algorithm uses the distance matrix groups genes such that each group contains genes that are close to each other. There are two main approaches to this problem: K-means clustering and hierarchical agglomerative clustering. K-means clustering requires a parameter *K*, and produces exactly *K* groups of genes. *K* is often estimated based on prior biological knowledge. Hierarchical agglomerative clustering produces a hierarchical ordering of clusters, such that groups of very close genes are contained within larger clusters.

In addition to clustering, classification methods have been applied to microarray data. In contrast to the previous case, here, the goal is to distinguish between two known classes, such as patients with or without disease. Then, given a patient where it is not known whether they have the disease, the method can predict whether they also have the disease. Although many classification algorithms have been developed, such as Bayesian classifiers or Support Vector Machines, they all draw from the same principles. First is the requirement of a *training set* that contains examples of the two groups to be distinguished. Then, the algorithm is applied on a *test case* that is scored against the training set to identify the most closely related group.

Microarrays often need to be pre-processed before applying classification algorithms. Most importantly, they suffer from the *curse of dimensionality*, whereby the number of genes often exceeds the number of samples available. This causes theoretical problems. Because of the availability of many genes, it may be possible to find genes that can clearly distinguish the two groups based on noisy or arbitrary variations in the training data. Thus, microarrays undergo a pre-processing step that filters out the non-informative genes on the chip.

Because of the ability of microarrays to capture and quantitate the biology underlying disease, they are seen as a potential technology for delivering personalized therapeutics.^[10]

PROTEOMICS

Microarray experiments are exciting because they can simultaneously measure the activity of gene expression and regulation for an entire organism. Another important research area is understanding the activity of the *proteome*, the concentrations, structures, locations, and functions of the protein products of the genome. While experiments are not yet as technologically mature as genomewide microarrays, there are now powerful tools for measuring the diversity of the proteome at a point in time.

Two-dimensional SDS gel electrophoresis methods allow the researcher to separate the protein components of cytosolic extracts. The results of such separations gives a two-dimensional array of spots, each representing (if well separated) a single protein member of cell. Bioinformatics plays an important role in identifying and characterizing the specific spots in a gel. If a spot has not been previously identified or its identity is unclear, mass spectroscopy is a powerful technique for identifying the molecular components of a separated protein. Automated robots are becoming more common for coupling gel separation methods with mass spectroscopy for both separating and characterizing the protein components in a sample.

Proteomics methods can now identify peptides in a sample that is a complex mixture using mass spectrometry. Mass spectra are like “fingerprints” for the peptide that must be interpreted to identify the chemical structure of that peptide. Peptide spectra must be mapped to peptide sequencing with the use of a database driven search engine such as MASCOT, X-TANDEM, or SEQUEST. Search engines utilize large databases of known peptide spectra to make the identification.

The burden of bioinformatics methods is to automatically characterize previously known spots, to choose interesting spots, and to use the results of the mass spectrometry experiments based on genomic data, and to identify any covalent modifications (post-translational) that have occurred.

SYSTEMS BIOLOGY

An exciting goal of post-genomic science is to understand how genes and gene products interact on a cellular level, organ, and systems level. Systems biology is the emergent field aimed at modeling these processes. Cellular processes have been modeled as metabolic networks and gene regulatory (or genetic) networks. Biologically relevant networks are not limited to intracellular processes, but also include the immune system, embryogenesis, and development, as well as the interactions between cells. Bioinformatics' role in systems biology is to discover, characterize, store and simulate biological networks using experimental assay data. Networks quantifying the regulation and interactions of genes and gene products are generally divided into three areas: (i) metabolic networks; (ii) signal transduction networks; and (iii) genetic/transcriptional regulatory networks (GRNs). Because of the complex differences and cultural paths of research, prokaryotic networks are often considered separately than eukaryotic networks, although the underlying principles are generally similar.

The studies of metabolic networks have been undertaken for more than a hundred years. Traditionally, metabolic networks are elucidated by careful study of tissue and cellular extracts. Glycolysis was first investigated as early as 1860 by Louis Pasteur. Some of the early quantitative metabolic studies were undertaken by Hans Krebs, where critical components of the tricarboxylic acid cycle were deduced by determining the reactions that occurred in tissue. His account of this discovery can be read in *Perspectives in Biology and Medicine*.^[11] Today, metabolic networks are characterized through traditional biochemistry assays, and a combination of genomics and proteomics experiments.

To quantitatively analyze metabolic (or any kind of biological) networks, researchers must catalog network data in a manner that is conducive to quantitative studies. Such a resource is MetaCyc.^[12] MetaCyc is used for the study of metabolic network data and is annotated by hand using literature data, and contains nearly 450 pathways from over 150 organisms.

Signal transduction networks describe how molecular signals are received and propagated to the end result. While scientists have a broad understanding of the major backbone of signal transduction pathways, the challenge now is to characterize the interactions that are either transient or context dependent. Here, technological advances in the ability to measure simultaneously the phosphorylation of proteins or the expression levels of genes on a single cell level has provided deeper insight.^[13,14] Coupled with the development of experimental technologies, an active area is the investigation of computational approaches. A diverse array of tools have been proposed, including Bayesian networks, boolean algebra, and differential equations.

GRNs describe how gene transcription, translation, and modification are fundamentally controlled. Bolouri

and Davidson^[15] describe five methods in which genetic regulatory networks are characterized and reconstructed. Those are (i) gene discovery, (ii) regulatory linkage analysis, (iii) identification of transcription factor binding sites, (iv) kinetic data measurement, and (v) simulation of predicted networks.

Inference of phenotypic changes through biological networks is possible through direct simulation. Simulation of biological processes is possible with sufficiently quantitative and completed parameter sets. Several methods are characterized for the simulation and quantification of biological networks. Peleg et al.^[16] have described a method using Petri nets for quantifying biological networks. Random Boolean networks have also been extensively used to model cellular processes.

UNDERSTANDING THE RELATIONSHIPS BETWEEN GENOTYPE TO PHENOTYPE

With the development of powerful genomics assays and the sequencing of large genomic sequences, the study and annotation of genetic variation becomes a significant problem. This includes discovering alleles associated with a phenotype, understanding how these alleles mediate that phenotype, and how that genotype affects response to drug therapies. In humans, the most common form of variation is single nucleotide polymorphisms (SNPs), which occur approximately once every thousand bases in individuals. Other types of variation include small insertions and deletions (indels) and larger changes such as repeats, transposable elements, copy number variants (cnv), and longer insertions and deletions.

Several bioinformatics resources collect mutation data. dbSNP (<http://www.ncbi.nlm.nih.gov/SNP/>) is an interface to public SNP data, and currently contains over nine million submitted SNPs in humans. The HGMD (<http://www.hgmd.org/>) collects disease-associated mutations that are associated with specific loci. NCBI's dbGAP database manage genome wide genetic studies of complex diseases and phenotypes (<http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gap>).

Understanding and predicting which variations are likely to participate in a phenotype is an important problem in bioinformatics. Initial research in this area has focused on SNPs and predicting which SNPs are likely to be deleterious.^[17–19] Several research studies have estimated the percentage of SNPs that are functionally important.^[20–22] These reports suggest that between 20% and 32% of amino acid altering (non-synonymous) mutations are likely to alter or damage the protein in which they occur. Many methods have been developed to predict disease causing mutations.^[23,24]

Pharmacogenomics is the study of genes, diseases, drugs, and the relationships between them. Primary to these efforts is understanding how different genotypes

lead to different patient responses to clinical treatment. To investigate this further, early research has focused on identifying how specific genotypes respond to different therapeutics. PharmGKB (<http://www.pharmgkb.org/>) is a public project that collects genotype and drug interaction data from several NIH-funded centers that are collecting this data.^[25] As soon as sufficient data are collected for different responses, genotypic assays can be developed as clinical diagnostic and treatment aids. Another aspect of this research is understanding the underlying molecular function of genome variation data. Two resources for annotating variation data with molecular function is MutDB (<http://www.mutdb.org/>), LS-SNP (<http://ls-snp.icm.jhu.edu/ls-snp-pdb/>), and PolyPhen (<http://tux.embl-heidelberg.de/ramensky/data/>).

STRUCTURAL BIOINFORMATICS

Proteins (and sometimes RNA) are large macromolecular machines that perform many of the basic functions of the cell. Understanding how the three-dimensional structure of a large macromolecule confers its function is an important problem in biology. Structural analysis bases many of its successes on the fact that the number of naturally evolved scaffolds, or structural folds, is finite. It is estimated that there are fewer than three thousand distinct folds found in nature.^[26] Currently, more than a thousand are known. This enables scientists to apply a guilt-by-association approach to predicting structure and function. An important area of bioinformatics research is collecting, storing, and retrieving structural data. The Protein Data Bank^[27] is a public resource for storing protein structure data, while the Nucleic Acid Database^[28] stores nucleic acid structures. ModBase^[29] is a resource that stores automatically predicted structures using the Modeller software package.^[30]

Several methods are available for annotating the function of a previously uncharacterized protein structure. Many of these methods rely on known functional annotations and detecting similarity in query proteins to those regions that have been previously annotated. A primary method for detecting similarity is through structural alignment—that is, aligning two regions of protein structures to identify similar regions. Structural alignment methods such as Dali^[31] and MinRMS^[32] characterize function based on structurally similar regions. MinRMS provides solutions to two problems. First, it determines which structural superpositions to evaluate, and then it determines which residues should be considered “aligned” or matched. To make the problem computationally tractable, superpositions are sampled by exhaustively evaluating all superpositions based on four atom clusters. All appropriate atoms are then sampled.

Other methods for functional annotation rely on matching similar structural properties. One such example is FEATURE.^[33,34] FEATURE encodes a local environment

of a protein structure as a vector of structural property values. Both supervised and unsupervised approaches can be used to assign function to previously uncharacterized protein structures.

Protein structure prediction methods are important for modeling structural effects in the protein products of genes of unknown structure. A primary consideration when building structural models is ensuring that the scientific question being asked can both be answered by a structural model and that the model you are building will be of high enough resolution to answer the question. There are two primary paths to building a model. First, comparative modeling methods are most often performed because they yield the best predictions. They require a target sequence that has a three-dimensional fold that is already known; in another, similar structure of detectably similar sequence.

Ab initio methods for protein structure prediction traditionally use biophysical first principles to attempt to “fold” a protein or RNA to its native three-dimensional structure. It is theorized that proteins adapt a state representing their minimum free energy. These methods are computationally challenging, because the force field used must be accurate enough to model the minimum free energy correctly, and the computational time required to model the folding is often impossible using even today’s fastest computers.^[35] Common force fields used for the simulation of macromolecules include AMBER,^[36] OPLS,^[37] and CHARMM.^[38] These force fields model the quantum effects of molecules as semi-empirical classical terms, parameterized from the theoretical and experimental properties of small molecules. The AMBER force field contains terms for bonds, angles, dihedrals, van der Waals, and coulombic interactions.

Simulation strategies implementing these force fields often involve focusing on parallelization to extend simulation time. With sufficiently parallel methods and hardware, simulations on the order of hundreds of microseconds are possible.^[39] This is nearing the time required to model adequately the folding of a quickly folding protein.

Homology modeling methods use the known structures of similar sequences to the target sequence to predict structure. These methods are usually performed in four or five steps: (i) fold identification; (ii) threading; (iii) modeling; (iv) model evaluation; and (v) refinement.^[30] Each of these five steps encompasses a large body of current research.

Commonly, choosing a fold or scaffold to use is relatively easy. Easier cases occur when a sequence being modeled has some recognizable degree of similarity with a sequence of known structure. When the sequence has diverged beyond the ability to identify the fold, more sophisticated methods must be employed.

Alignment of the sequence on the structure through the process of threading gives a prediction of how the backbone of the sequence relates to the backbone positions within the structure. Very generally, sequences of greater than 30% identity with a target are aligned by using sequence-based alignment methods, employing pairwise

residue matching scores, such as those outlined in “Overview.” When the difference is below 30% identity, more sophisticated methods may be required. These include statistical nonpairwise potentials,^[40] profile alignments (such as PSI-BLAST),^[41] or hidden Markov models such as those discussed previously.

Building a three-dimensional model based on a scaffold and a threaded alignment is a technically challenging task. The most commonly used package for modeling is Modeller,^[30] which employs a method based on satisfaction of spatial restraints. These spatial restraints are derived from the residue-residue distances in the known structures being modeled upon. After a sufficient number of restraints are determined, simulated annealing is performed to find satisfactory structures.

A notable method for building a model that does not require a similar template is the program Rosetta.^[42] Rosetta uses libraries of short peptide fragments of known structure. These fragment libraries are used to model the structural preferences of a local sequence. These predicted structural fragments are then used to predict the final structure. Modeled structures are often similar enough to deduce some of the topological components of the structure, but functional inference remains a difficult challenge with any modeling method not using a template scaffold as input.

Evaluation of modeled structures is performed by using both first principles, statistics of known structures, and experimental structural assays. Several stand-alone packages are available. The Biotech Validation Suite For Protein Structures (<http://biotech.ebi.ac.uk:8400/>), available at the European Bioinformatics Institute (EBI), provides three programs for structure analysis: Procheck,^[43] Prove, and WhatIf. Each of these applications provides a different aspect of structural analysis. Procheck checks a query structure for discrepancies with geometry, chirality, hydrogen bonds, disulfide bonds, dihedral angles, and so on. Prove performs an analysis of atomic volumes within a query structure. Finally, WhatIf analyzes rotamers, solvent atoms, proline ring puckers, bond angles and lengths, and atomic occupancy. Another package for structure analysis is Prosa II (<http://www.came.sbg.ac.at/Services/prosa.html>). Prosa II works by evaluating a model based on the knowledge of other known structures. Refinements of modeled structures is often performed by using molecular simulation methods or by re-performing one of the model building steps using different parameters.

ANALYZING BIOLOGICAL LITERATURE

Biological knowledge is still primarily disseminated as unstructured free text in textbooks, journal articles, and comments and description fields in databases. The amount of text available is increasing rapidly. MEDLINE currently contains over 12 million citations. Without specialized algorithms,

the knowledge encoded within the literature is inaccessible to computers. Therefore there is an increasing interest in developing bioinformatic algorithms to use the literature.

Having computational access to the knowledge within literature would benefit research and discovery in many ways. In the near term, it helps scientists to analyze their data more quickly by automatically identifying relevant keywords and articles. In addition, literature analysis has been used to improve homology search.^[44] Finally, there are efforts to develop algorithms that will help computers propose new hypotheses and guide experiments. Some have argued that automated analysis of all available literature, the *bibliome*, may uncover new associations that are currently unknown to scientists. This line of research stems from early work by Swanson, who discovered a new treatment for Raynaud's disease by matching its symptoms against drugs known to treat them.^[45]

LINKING BIOLOGICAL DATA TO TEXT

There has been much work in developing text mining algorithms to help make sense of biological data sets. Although there are many types of biological data available, such as structure, sequence, or expression, interpreting the data remains a primary challenge for biologists. Often, this requires literature searches to elucidate interesting features of the data. For aggregated data, such as protein families or genes sharing an expression profile, analyzing all their literature can be a daunting task. For these high-volume analyses, computational approaches can help collect literature relevant to data sets and then summarize interesting aspects of the literature.

Literature associated with protein sequences is often summarized by identifying the keywords that distinguish that set of sequences from a background set. Those keywords are the words occurring statistically more often with those sequences than with a background distribution of sequences. This technique is useful for annotating sequence families. Although there are many possible technical solutions for finding keywords, one of the earliest methods calculated the average frequency of words when they appeared with protein families and compared that against the number of families in which the word occurred.^[46] The interesting keywords were the ones that occurred very often, but with few families. Some of the protein families they annotated this way were *myoglobin*, *ras*, and *phosphoglycerate*.

A related task is to annotate sequences with structured descriptions of function, rather than just keywords. Rather than just identifying the discriminating words, this task aims to select, from a predefined list, the term that best describes the function of a sequence or sequences. One such list of functions is the Gene Ontology (<http://geneontology.org/>), which describes biological functions in three hierarchies, *Molecular Function*, *Biological Process*, and

Cellular Component. Here the task is to identify the *GO code* that best describes the literature associated with a sequence. This has been solved by using statistical *machine learning* algorithms that learn the characteristic features of each GO code, and can match the literature of a sequence based on which features they exhibit.^[47]

EXTRACTING RELATIONSHIPS

Finally, there is much work on identifying relationships present in text. There is much biological research concerned with discovering associations between entities, such as protein-protein binding or protein-drug metabolism. Because of the sheer number of relationships, manually creating lists of all known relationships is difficult. Thus there are bioinformatic algorithms to identify relationships from text and organize them into databases and networks.

The dominant approach for finding related entities is with co-occurrence. The theory behind this method is that related entities will appear in the same abstracts. Thus if two proteins bind to each other, they will occur in the same abstract. Although co-occurrence generates errors (e.g., not all proteins in the same abstract will be related), it works acceptably well for certain applications. One group has applied this method to create automatically regulatory networks.^[48]

TRANSLATIONAL BIOINFORMATICS

The National Institutes of Health (NIH) has recently started an initiative in translational medicine, the tight integration of clinical research and basic research. To enable this, they are funding Clinical and Translational Sciences awards to fund institutes and centers to enable translational research activities. Bioinformatics is a key area for developing translational research.

Important areas for development are in integrating large, de-identified clinical data sets with biochemical, genetic, proteomic, or other complementary data. This is an emergent field and is just being developed. A new conference, the Translational Bioinformatics Summit, is held annually in San Francisco to address research advances in this important area.

CONCLUSION

New topics in bioinformatics will likely continue to follow developments in experimental assays. The recent rise in new bioinformatics methods has not focused on advances in computational power, but on the development of complex assays for surveying the global activity of cells using proteomics and gene expression assays and the rise of the Internet.

Three areas that will continue to be of high importance to bioinformatics research are (i) simulation of biological networks to predict phenotypic changes, (ii) use of basic biology to influence clinical practice, and (iii) automation of complex bioinformatics methods. While these are certainly not the only growth areas, many problems remain unsolved within them.

In conclusion, bioinformatics is a broad and dynamic field that is constantly acquiring new challenges and insights from technology innovation throughout the biosciences.

ACKNOWLEDGMENT

Sean Mooney is funded by NIH R01 LM009722.

REFERENCES

- Smith, T.F.; Waterman, M.S. Identification of common molecular subsequences. *J. Mol. Biol.* **1981**, *147* (1), 195–197.
- Dayhoff, M. *Atlas of Protein Sequence and Structure*; National Biomedical Research Foundation: Washington, DC, 1979, 5.
- Vogt, G.; Etzold, T.; Argos, P. An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *J. Mol. Biol.* **1995**, *249* (4), 816–831.
- Altschul, S.F.; Gish, W.; Miller, W., et al., Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215* (3), 403–410.
- Sigrist, C.J.; Cerutti, L.; Hulo, N., et al., PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* **2002**, *3* (3), 265–274.
- Bateman, A.; Birney, E.; Durbin, R., et al. The Pfam protein families database. *Nucleic Acids Res.* **2002**, *30* (1), 276–280.
- Brudno, M.; Do, C.B.; Cooper, G.M., et al., LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome. Res.* **2003**, *13* (4), 721–731.
- Alizadeh, A.A.; Eisen, M.B.; Davis, R.E., et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. [comment]. *Nature* **2000**, *403* (6769), 503–511.
- Eisen, M.B.; Spellman, P.T.; Brown, P.O., et al., Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U S A* **1998**, *95* (25), 14863–14888.
- Nevins, J.R.; Huang, E.S.; Dressman, H., et al., Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Hum. Mol. Genet.* **2003**, *12* Spec No 2, R153–R157.
- Krebs, H.A. The history of the tricarboxylic acid cycle. *Perspect. Biol. Med.* **1970**, *14* (1), 154–170.
- Karp, P.D.; Riley, M.; Paley, S.M., et al., The MetaCyc database. *Nucleic Acids Res.* **2002**, *30* (1), 59–61.
- Irish, J.M.; Hovland, R.; Krutzik, P.O. Single cell profiling of potentiated phospho-protein networks in cancer cells. *Cell* **2004**, *118* (2), 217–228.
- Geiss, G.K.; Bumgarner, R.E.; Birditt, B., et al., Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotechnol* **2008**, *26* (3), 317–325.
- Bolouri, H.; Davidson, E.H. Modeling transcriptional regulatory networks. *Bioessays* **2002**, *24* (12), 1118–1129.
- Peleg, M.; Yeh, I.; Altman, R.B. Modelling biological processes using workflow and Petri Net models. *Bioinformatics* **2002**, *18* (6), 825–837.
- Mooney, S.D.; Klein, T.E.; Altman, R.B., et al. A functional analysis of disease-associated mutations in the androgen receptor gene. *Nucleic Acids Res.* **2003**, *31* (8), e42.
- Ng, P.C.; Henikoff, S. Predicting deleterious amino acid substitutions. *Genome. Res.* **2001**, *11* (5), 863–874.
- Sunyaev, S.; Ramensky, V.; Koch, I., et al. Prediction of deleterious human alleles. *Hum. Mol. Genet.* **2001**, *10* (6), 591–597.
- Chasman, D.; Adams, R.M. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* **2001**, *307* (2), 683–706.
- Ng, P.C.; Henikoff, S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res* **2002**, *12* (3), 436–446.
- Sunyaev, S.; Ramensky, V.; Bork, P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet.* **2000**, *16* (5), 198–200.
- Li, B.; Vidhya, G.; Krishnan, G., et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* **2009**, *25* (21), 2744–2750.
- Mooney, S., Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Brief Bioinform* **2005**, *6* (1), 44–56.
- Hewett, M.; Oliver, D.E.; Rubin, D.L., et al., PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Res.* **2002**, *30* (1), 163–165.
- Chothia, C. One thousand families for the molecular biologist. *Nature* **1992**, *357* (6379), 543–544.
- Berman, H.M.; Olson, W.K.; Beveridge, D.L., et al., The nucleic acid database. A comprehensive relational database of three dimensional structures of nucleic acids. *Biophysical. J.* **1992**, *63*, 751–159.
- Berman, H.M.; Westbrook, J.; Feng, Z., et al. The nucleic acid database. *Methods Biochem. Anal.* **2003**, *44*, 199–216.
- Sanchez, R.; Sali, A. ModBase: A database of comparative protein structure models. *Bioinformatics*, **1999**, *15* (12), 1060–1061.
- Sali, A.; Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234* (3), 779–815.
- Holm, L.; Sander, C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **1997**, *25* (1), 231–234.
- Jewett, A.I.; Huang, C.C.; Ferrin, T.E. MINRMS: an efficient algorithm for determining protein structure similarity using root-mean-squared-distance. *Bioinformatics* **2003**, *19* (5), 625–634.
- Bagley, S.C.; Altman, R.B. Characterizing the microenvironment surrounding protein sites. *Protein Sci.* **1995**, *4* (4), 622–635.
- Liang, M.P.; Banatao, D.R.; Klein, T.E., et al., Web-FEATURE: An interactive web tool for identifying and

- visualizing functional sites on macromolecular structures. *Nucleic Acids Res.* **2003**, *31* (13), 3324–3327.
35. Duan, Y.; Kollman, P.A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. [comment]. *Science* **1998**, *282* (5389), 740–744.
 36. Cornell, W.D.; Cieplak, P.; Bayly, C.I., et al. A second generation force field for the simulation of proteins and nucleic acids. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
 37. Jorgensen, W.; Tiradorives, J. The OPLS potential functions for proteins - energy minimizations for crystals of cyclic-peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110* (6), 1666–1671.
 38. Brooks, B.R.; Bruccoleri, R.E.; Olafson, B.D., et al. CHARMM – A Program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4* (2), 187–217.
 39. Zagrovic, B.; Snow, C.D.; Shirts, M.R., et al. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.* **2002**, *323* (5), 927–937.
 40. Sippl, M.J.; Weitckus, S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins* **1992**, *13* (3), 258–271.
 41. Altschul, S.F.; Madden, T.L.; Schäffer, A.A., et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search tools. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.
 42. Simons, K.T.; Bonneau, R.; Ruczinski, I., et al., Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **1999**, Suppl 3, 171–176.
 43. Pontius, J.; Richelle, J.; Wodak, S.J. Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J. Mol. Biol.* **1996**, *264* (1), 121–136.
 44. Chang, J.T.; Raychaudhuri, S.; Altman, R.B. Including biological literature improves homology search. *Pac. Symp. Biocomput.* **2001**, *6*, 374–383.
 45. Smalheiser, N.R.; Swanson, D.R. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput. Methods Programs Biomed.* **1998**, *57* (3), 149–153.
 46. Andrade, M.A.; Valencia, A. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* **1998**, *14* (7), 600–607.
 47. Raychaudhuri, S.; Chang, J.T.; Sutphin, P.D., et al. Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res.* **2002**, *12* (1), 203–214.
 48. Jenssen, T.K.; Laegreid, A.; Komorowski, J., et al. A literature network of human genes for high-throughput analysis of gene expression. [comment]. *Nat. Genet.* **2001**, *28* (1), 21–28.

Biologics

P. A. Lachenbruch, A. D. Horne, C. J. Lynch, J. Tiwari, and Susan S. Ellenberg

U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.

Abstract

Biologics comprise products of wide-reaching public health impact, such as vaccines, as well as products, like gene therapies, that are on the cutting edge of science. This entry reviews the basic tenants of biologics and discusses some of the issues and regulatory processes in biologic products.

INTRODUCTION

The Food and Drug Administration (FDA) has six centers that review products for many uses. These centers are as follows:

- Center for Biologics Evaluation and Research (CBER)—regulates biologic products (e.g., blood products, vaccines, monoclonal antibodies, cytokines).
- Center for Devices and Radiological Health (CDRH)—regulates medical devices and radiation products (e.g., X-ray machines, cell separators, bandages).
- Center for Drug Evaluation and Research (CDER)—regulates drugs.
- Center for Food Safety and Applied Nutrition (CFSAN)—regulates foods and food supplements.
- Center for Veterinary Medicine (CVM)—regulates veterinary medical products.
- National Center for Toxicological Research (NCTR)—a research facility devoted to toxicology.

The origin of the regulation of biologics dates back to the Virus, Serum, and Toxin Act of 1902 that established the federal regulation of biologic products.^[1] Many provisions of that Act are still in force today: establishment licensing, product licensing, labeling requirements, facility inspection requirements, revocation or suspension of licenses, and penalties for violations. Initially, and until 1944, biologics were regulated by the Hygienic Laboratory of the Public Health Service. The laboratory promulgated regulations implementing the provisions of the Act. Inspections of manufacturing facilities were to be unannounced and conducted by commissioned medical officers. Samples of products were to be examined for purity and potency.

In 1944, there was a major revision of Federal public health laws. One change made was that “a biological license could [be] issue[d] only upon showing that the product and

establishment met standards to ensure the continued safety, purity, and potency of such products.”^[1] This act also provided for the transfer of biologics regulation to the National Microbiological Institute of the National Institutes of Health (NIH). In the 1950s, biologics regulation was transferred to the Division of Biologics Standards (DBS) of the NIH. In 1972, biologics regulation was transferred to the FDA, where it remains today. In addition to merging the DBS (renamed the Bureau of Biologics) into the FDA, the 1972 merger integrated the regulations of the Public Health Service Act and the Federal Food, Drug, and Cosmetic Act. In 1982, the FDA merged the Bureau of Drugs and the Bureau of Biologics into the Center for Drugs and Biologics; in the late 1980s, these were separated and became the CDER and the CBER.

OVERVIEW

The CBER’s responsibility includes blood and blood products such as immune globulins and clotting factors; therapeutic biologic products such as cytokines, monoclonal antibodies, and cellular and gene therapies; and vaccines and allergenic products. Formally, the Code of Federal Regulations (CFR) defines a biologic as “any virus, therapeutic serum, toxin, antitoxin or analogous product, applicable to the prevention, treatment, and cure of diseases or injuries of man.”^[2] Generally, biologics are products derived from biologic sources, tend to have a large molecular weight, are immunogenic, and exhibit pleiotropic effects (affect many different cell types). They have a prolonged duration of action and are more susceptible to microbial contamination than drug (CDER-regulated) products. Another important difference between biologic and drug products is that the organic materials used to make biologic products are more variable than the substances used in making drug products. The CBER considers the manufacturing process to be a fundamental aspect of the product and requires strict controls on this process. The manufacture of biologic products presents unique regulatory concerns in relation to quality control, consistency, stability, microbial contamination,

The views expressed in this document are those of the authors and do not necessarily represent those of the U.S. Food and Drug Administration.

product administration, and source of the material (serum, cell bank, and tissue). These products also present some special safety concerns related to the modulations of the immune system and the inflammatory response.

Blood and blood products include such items as immune globulin products, clotting factors for the treatment of hemophilia, products that test for human immunodeficiency virus (HIV) antigens, and products that promote cessation of bleeding, among others. The CBER is responsible for regulating the nation's blood banks to ensure a safe blood supply.

The spectrum of biologics for therapeutic applications includes monoclonal antibodies (derived from a single clone, used for diagnostic and therapeutic purposes), cytokines (proteins that mediate the effector phases of natural and specific immunity), hematopoietic growth factors (stimulate the growth and the differentiation of immature leukocytes in the bone marrow), cell and gene therapies, tumor vaccines, xenotherapies (products made from non-human sources), tissue-engineered products, and cell selection devices.

Vaccines are the most widely used class of biologic products that the CBER regulates. These products include polio, measles–mumps–rubella (MMR), diphtheria–tetanus–acellular pertussis (DTaP), *Haemophilus influenzae* type b (Hib), and hepatitis vaccines, as well as other vaccines that are less widely used, such as vaccines to prevent cholera and typhoid. The Office of Vaccines Research and Review within the CBER also regulates allergenic products to prevent or to reduce the effects of hay fever, allergies to grasses, and other allergic conditions.

A recurrent problem in evaluating some of the above types of products or procedures is that of competing risks. For example, in evaluating blood substitutes, a life-threatening safety event may cause the termination of therapy prior to a favorable outcome. When a favorable event (such as shortened hospitalization) and a competing event (such as death) have different implications for efficacy of the product, an assessment of product benefit may not be straightforward. Meinert^[3] noted that “a differential mortality by treatment group may influence the occurrence of nonfatal events. The treatment group with the highest mortality rate may have the lowest nonfatal event rate if death occurs before patients have a chance to develop the nonfatal event of interest.” The competing risk of mortality has little bearing on the evaluation of childhood vaccines, allergenics, and many other biologics.

REGULATORY PROCESS

Product Development and Licensure

The regulation of a biologic begins with the Investigational New Drug (IND) application, filed with CBER after preclinical testing suggests that a product may be safe and has

the desired biologic activity in animals. The clinical investigation of a new biologic, as well as a new drug, generally proceeds in three phases. Phase I studies are the first investigations in humans and focus on the safety, metabolism, and pharmacologic actions of the product, potential dose range, and sometimes preliminary evidence of effectiveness. Phase II studies are generally closely monitored controlled studies to obtain preliminary data on efficacy and side effects, and other risks, and to establish an optimum dose and schedule. Investigators often use data from phase II to determine appropriate endpoints for phase III studies. Phase III studies are larger studies of the efficacy and the safety of the product. They provide information to support risk/benefit considerations and form the basis for product labeling. After the completion of the investigational clinical trials in the IND phase, the sponsor submits a Biological License Application (BLA) to market the product. In order for a license to be issued, the application must show that the product meets standards of safety, purity, potency, and consistency of manufacturing, in addition to demonstrating efficacy. Following licensure, manufacturers are required to continue to monitor the safety of their products and to submit any reported adverse effects to the FDA.

Generic Products

The concept of a generic drug has not been applied to biologic products to date, as a result of the inherent variability in biologic products. Because the manufacturing process is a fundamental characteristic of the product, biologic products that are made differently are considered to be distinct. For example, although there are several DTaP vaccines, none is considered “generic”; they are all considered to be distinctly different products for the same disease indication.

International Harmonization

Since the early 1990s, a consortium of industry and regulatory scientists in the United States, Europe, and Japan has been working to harmonize the requirements for the licensure of drugs and biologics worldwide. This consortium, the International Conference on Harmonization (ICH), has developed a series of documents addressing wide-ranging issues in product development, including aspects of product manufacture, preclinical and clinical testing, postmarketing safety surveillance, and structure of reports to regulatory authorities. These documents have been adopted as regulatory guidances in the three participating regions.

ISSUES IN BLOOD AND BLOOD PRODUCTS

The CBER's Office of Blood Research and Review evaluates proposed studies of new products and the data that result, and conducts research related to the development,

manufacture, testing, and activities of blood products to ensure the safety of the blood supply (e.g., test kits for HIV, hepatitis B and C, syphilis), and those prepared by genetic engineering and synthetic procedures (plasma derivatives, oxygen therapeutics formulated by recombinant technology or purification), in order to establish standards to ensure the continued safety, purity, efficacy, and availability of blood products. The office evaluates study designs and licensing applications related to commercial blood-derived and analogous recombinant products distributed in the United States, such as medical devices used to collect, test, process, or store donated blood; retroviral diagnostic tests including HIV-related tests; and certain human cellular and tissue-derived products.

Many of the statistical approaches for addressing these activities are common to other biomedical evaluations and other centers at the FDA. In many conditions that blood products address, however, the target population is small, so sponsors must contend with limited numbers of patients available for conducting clinical trials. Evaluating the safety of blood and plasma products presents additional challenges. For many plasma products, many donations are pooled from a large number of donors. Thus a single donation of infectious but seronegative plasma or blood can affect many recipients. It is possible to reduce, but not entirely eliminate, the risk of transmitting infection via blood products. Donor questionnaires can identify some donors at elevated risk of having infectious blood; such donations are discarded. The application of nucleic acid testing (NAT) methods can document the presence of infectious agents before antibodies have been produced and therefore reduce the period during which an infectious individual will appear uninfected. The use of such methods permits a greater proportion of infectious donations to be discarded, further minimizing the risk of transmitting infection. Assessing new diagnostic test kits without the benefit of a “gold standard” is problematic, but sometimes unavoidable. In this case, the new test kit is compared with one in current use; but when the results are discordant, it is difficult to ascertain which test kit is superior. This situation leads to distinct statistical problems.^[4]

The primary endpoint or efficacy variable in any clinical trial should provide a convincing basis for comparing treatments. Biologic products sometimes address needs for which there are no currently existing products, so there may be no precedent for selecting an appropriate efficacy endpoint. The difficulty in defining an appropriate endpoint also arises in several types of blood product experiments. For example, fibrin sealants are products that are “painted” on a wound to halt bleeding. Should the endpoint be time-to-bleeding cessation, or the volume of blood loss? Should it be a dichotomous variable taking the value 1 if bleeding stops within some arbitrary time interval and 0 otherwise? What are appropriate comparator groups?

Hemophilia is the condition of a deficiency of clotting Factor VIII, and there are products designed to compensate

for this deficiency. Bleeding is often internal into a joint, however, and therefore not directly measurable. Because the bleeding causes the patient pain, bleeding cessation is measured by the patient’s subjective report that may be biased by the patient’s knowledge of the therapy. Also, it is common to assess multiple episodes of bleeding in such patients, leading to questions of how to account for the correlation among the responses.

In some cases, it has been difficult to conduct controlled trials when the investigational product was thought likely to be much better than existing therapy; thus some products have been assessed and approved on the basis of uncontrolled studies. In such cases, it has been necessary to determine a standard for comparison—either a historical control or an absolute standard.

Statistical Issues in Assessing Biologic Test Kits

The CBER regulates all products, including medical devices, related to assuring the safety of blood and blood products. These include various biologic test kits for detecting HIV, HCV, HAV, and HTLV. Within these larger groups, there are further categories such as HIV-1, HIV-2, and Type O kits. Kits are used for screening, diagnosis, and prognosis; some have been designed for home self-tests and also for quick responses in walk-in clinics. Kits may be designed to detect antigens or antibodies, using serum, plasma, sputum, or urine. The experimental studies for assessing test kits depend upon specific combinations of these reagents. The statistical design and analyses must be selected accordingly.

In evaluating a proposed test kit, the statistical approach consists of either comparing the proposed device to currently licensed devices, or requiring that the estimated sensitivity (the proportion of samples known to be positive that are correctly identified by the proposed kit as positive) and specificity (the proportion of samples known to be negative that are correctly identified by the proposed kit as negative) of the proposed device meet prespecified standards. In some cases, such as assessing a test kit proposed for a quick response, predicted value positive and predictive value negative may be the appropriate primary statistics of interest. Test samples that are indeterminate even after retesting are considered failures for evaluation purposes. Known positive and negative samples are those that had previously been so designated by some gold standard approved by the FDA. Because a kit proposed for screening purposes may be used subsequently to test hundreds of thousands or even millions of samples, it is necessary that the estimates of sensitivity and specificity be precise to as many as four decimal places. This level of precision requires 10,000 or more samples, considerably greater than those used for most other biologic assessments. Because of the required precision, it is preferable to use exact statistical computational methods in contrast to asymptotic approximations in computing estimates.

Predictive value positive (proportion of samples testing positive that are actually positive) and predictive value negative (proportion of samples testing negative that are actually negative) are occasionally the primary statistics of interest, in contrast to sensitivity and specificity. This might be the case, for example, when a result is of immediate importance and the patient may not return for confirmatory tests. A larger sample size is required for estimating predictive value positive and predictive value negative than is required when estimating sensitivity and specificity. Because predictive values are computed by a combination of the sensitivity, specificity, and prevalence, the variance depends on the variance of all three quantities.

Receiver Operating Characteristic (ROC) curves are occasionally proposed to the CBER for assessing biologic test kits. The CBER's practice with regard to ROC assessment has been to use criteria comparable to those provided in guidelines established by the National Committee for Clinical Laboratory Standards.^[5] These are especially useful when the diagnosis of disease depends on a continuous variable. For ROC curves, as with other statistical design issues, it is to the benefit of the study investigators to discuss potential issues, including sample size requirements, with representatives from the CBER before proceeding.

Low Prevalence Conditions

Some blood products are designed to treat rare diseases. Clotting factors for hemophilia, for example, are targeted at a total population of about 40,000 hemophiliacs in the United States. Some subgroups of this population have developed inhibitors to Factor VIII replacement products; treatments for such subgroups are aimed at much smaller target populations.

An even more extreme example is a product proposed for hemophiliacs who have developed inhibitors to the usual Factor VIII products. There are only about 400 patients in North America with this condition. For this product, it is feasible to recruit, at most, 20 patients for a clinical trial. These limited populations make it difficult to conduct clinical trials that can definitively establish efficacy or provide a substantial understanding of the safety profile of the product. Another rare condition is envenomation from snakebites. Only about 8000 people per year have sufficiently severe snakebites to be candidates for treatment. A new product under investigation was expected to provide the same therapeutic effect as the only currently available product, but with the hope that it would have a better adverse event profile. For this rare condition, an appropriate endpoint was needed; because mortality is uncommon, a score based on the severity of the snakebite was used. Because a great deal is known about the currently available product, both with regard to efficacy and safety, it was determined that the new study could be done using historical controls. The study showed that the

new product had a superior safety profile, and its efficacy appeared similar to that of the control, based on comparisons with historical data.

Conducting clinical trials involving few patients may require flexibility regarding some of the preferred characteristics of the ideal clinical trial. For example, it may be necessary to do single-arm studies and to compare results to historical controls. If a historical control group is used, it is important that control subjects meet the same entry criteria as the treated patients. The historical controls should be as contemporary as possible so that treatment characteristics other than the therapy in question are similar. Such rare diseases/problems are not unique to the CBER; indeed, the FDA has an orphan drugs office that helps to regulate therapies for these rare conditions.

Evaluating the Safety of Blood and Blood Products

Some blood and blood products are used to treat patients who donated the blood in the first place. These are called autologous products. Other blood and blood products are from donors other than the recipient (allogeneic products) and may contain viruses or other unwanted particles that could transmit disease to the recipient. Safety is a particular concern with allogeneic products because they are often administered directly into the recipient's bloodstream. Several procedures are used to reduce the risk of viral transmission from blood products:

- Blood collection centers screen donors for current and prior infectious diseases and high-risk lifestyle activities that could pose a risk of viral transmission from their blood or blood products.
- Blood collection establishments test all blood and plasma for HIV, HTLV, hepatitis B, hepatitis C, and other agents. Even with these procedures, occasionally (about one per 34,000 donations) a unit is virally contaminated, primarily because of the "window period" problem. The window period is the interval between the time an individual becomes infected and the time the virus (or antibodies to the virus) is at detectable levels in the blood. A blood or plasma donation made during the window period will not be identified as virally contaminated but may still transmit infection. The probability of a window period donation depends on the length of the window period and the effectiveness of predonation screening.^[6]
- Finally, plasma derivatives undergo viral inactivation treatment prior to distribution. This treatment reduces the risk of transmission of hepatitis B, hepatitis C, and other viruses to extremely low levels. Some nonenvelope viruses such as hepatitis A and parvovirus B19, however, are difficult to inactivate, so not all viral transmissions are affected by such procedures.

ISSUES IN BIOLOGIC THERAPEUTICS

The development of recombinant biologic products to treat human diseases has a relatively short history, although many “traditional” biologic products such as vaccines and blood products have a history of 100 years or more. During the last 10 years, scientific progress in recombinant DNA technology and molecular biology has resulted in new classes of products to treat diseases such as multiple sclerosis, hepatitis C, diabetic ulcers, and non-Hodgkin’s lymphoma and other cancers. The expanded development of therapeutic biologics is a result of new technology that permits molecular biologists to produce large quantities of previously scarce human proteins. Monoclonal antibody products have been developed for a variety of purposes, including the treatment of acute stroke, cancer, rheumatoid arthritis, and Crohn’s Disease; diagnostic imaging; and the prevention and the treatment of rejection of tissue grafts. Growth factor products are used for wound healing and for supporting the immune system in cancer patients receiving chemotherapy. Cytokine products such as the interferons and the interleukins are used to treat cancer, hepatitis C, and multiple sclerosis. Other biologic products have been developed for cystic fibrosis, anemia of chronic renal failure, stroke, myocardial infarction, and genetic disorders.

The fundamental principles of clinical trial and statistical methodologies applied in the area of biologic therapeutics are the same as in other areas of drug development. However, some statistical issues and design questions have been encountered in this area more frequently than others.

In the late 1980s to early 1990s, when alpha interferons, granulocyte colony-stimulating factors (G-CSF), erythropoietin, and other agents were entering Phase III trials, clinical studies were generally designed as superiority trials with concurrent placebo controls because they addressed clinical conditions for which there were no proven treatments. However, as these new agents came into clinical use, the trials of second-generation products were often designed with active control arms. These studies are usually designed now as noninferiority trials, with the objective being to rule out a specified (true) difference δ with specified power. In statistical terms, the null hypothesis is specified as $H_0: \mu_T < \mu_C - \delta$.^[7-9] The first generation of monoclonal antibodies for diagnostic applications as imaging agents has entered Phase III trials, and several have been approved. Clinical trials for diagnostic agents typically obtain images of patients using a conventional diagnostic method and images using the new method. Special problems with these trials include the fact that they are easily unblinded (images produced using radio-labeled monoclonal antibodies appear quite distinct from those produced by traditional imaging modes), and their overall performance is affected by the proportion of diseased patients in the study. Usually, the patient is imaged with both the standard method and the new method, and a comparison is made. These trials require methods that

account for the pairing of images. The traditional measures of efficacy of diagnostic products have been sensitivity and specificity—often considered as ROC curves. The gold standard is usually a histological measurement from a surgical procedure or a biopsy. Predictive values and accuracy are affected by the prevalence of the condition. Indeed, the measures themselves may be inappropriate in some instances. For example, if the objective is to detect the presence of disease in the patient (or an organ), sensitivity and specificity are appropriate. However, if the goal is the detection of lesions (affected areas), it is possible to ascertain the probability that an identified lesion is truly positive (positive predictive value). This may be of importance when the presence of a specific lesion (e.g., liver involvement) is directly related to treatment decisions. For example, in a colon cancer imaging product, the presence of four or more liver lesions is a contraindication to hepatic resection. A difficulty in assessing diagnostic success measures (sensitivity, specificity, negative and positive predictive values) is that areas that are not indicated as affected by the test product are not biopsied.

There is a need to have a carefully planned drug development strategy, and omitting intermediate phases is not generally advisable. The drug development plan may move gradually toward the more traditional paradigm of safety and efficacy evaluations as we identify more biologics that are safe and effective for the treatment of chronic diseases.

ISSUES IN VACCINES

Vaccines may be administered for either preventive or therapeutic indications. Preventive vaccines, which constitute the vast majority of vaccine products, typically are designed to elicit antibodies in healthy individuals to the causative agents for infectious diseases such as polio, diphtheria, tetanus, pertussis, measles, mumps, rubella, influenza, and others. Three major concerns addressed in preventive vaccine evaluation are the extent and the duration of protection from disease or infection, the characteristics of immune response, and the safety profile. Therapeutic vaccines are intended to treat an already existing disease, such as cancer or HIV, in the vaccine recipient. Because therapeutic vaccine trials are designed and statistically analyzed in the same manner as trials for other therapeutic products, they will not be considered further here.

An efficacy trial for a vaccine compares an investigational product to a control in an appropriate population. When the control is a placebo (or a vaccine that has no effect on the disease of interest), the trial can estimate absolute vaccine efficacy. When the control is an already licensed vaccine for the same disease indication, the trial can estimate relative vaccine efficacy.

For regulatory purposes, the main interest is in estimating the *direct* effect of the vaccine on disease incidence. Case definitions (criteria for the diagnosis of disease) must

be resolved in the planning stage. In general terms, vaccine efficacy is customarily estimated from a vaccine efficacy trial as:

$$VE = 1 - IR_v / IR_c$$

where IR_v is the disease risk, incidence rate, or hazard rate in the vaccinated group and IR_c is the corresponding estimate in the control group. VE might refer to absolute or relative efficacy, depending on the control group. There are a variety of methods for constructing confidence intervals around VE.^[10]

In addition to direct effects, vaccines often confer *indirect* effects as well, if the disease is spread by person-to-person contact in the population. In such cases, even persons who did not complete a vaccination series or are unvaccinated still may be protected from the disease because of a reduction in disease transmission. Thus undervaccinated or unvaccinated individuals may be protected from diseases through “herd immunity,” a phenomenon unique to vaccines. The direct estimates of VE obtained from clinical trials may underestimate the overall reduction in disease that might be expected in well-immunized populations postlicensure because such trials do not fully capture the likely indirect effects of vaccination. Much work on the estimation of indirect effects of vaccination has been done.^[11]

Often an investigational vaccine targets a disease for which an already licensed vaccine is available. In this case, the disease incidence may have already been reduced to such a low level that a clinical efficacy trial in that population is not feasible. For example, the recent trials for acellular pertussis vaccines could not be conducted in the United States because the long use of whole cell vaccines reduced the incidence of pertussis to extremely low levels. Thus trials of new acellular pertussis vaccines were conducted in Italy, Germany, and Sweden where vaccination with whole cell pertussis vaccines had been discontinued or was infrequent because of concerns about adverse reactions. Acellular pertussis vaccines were developed to reduce the adverse reaction rate.

Vaccine manufacturers attempt to combine several vaccines into a single product, when possible, to improve the efficiency of administration and to reduce the number of injections individuals must receive. Special issues arise in trials in which currently licensed individual vaccines are physically combined to be given as one injection and the resulting combination product is compared to the vaccines given separately. One concern is that mixing several antigens might result in reduced vaccine efficacy because some antigens may interfere with the immune response to other antigens. Another concern is that the evaluation of efficacy must be based on serological endpoints rather than cases of disease because of low disease incidence as a result of the wide use of the licensed component vaccines. Conducting and interpreting these trials are facilitated if there are prior data showing that certain serum antibody levels are correlated with protection from disease. In such cases,

the serum antibody levels can be compared to assess the noninferiority of the combination product. Unfortunately, such correlates are not known for all vaccines.

A third methodological concern in studies of combination vaccines is the assessment of local reactions. There is a single injection site for the combination product, while there are two (or more) injection sites for the separate products. Thus the separate injection treatment group has more opportunity to experience a local reaction (e.g., redness or swelling at the injection site) compared to the combination group.^[12]

When comparing a possibly improved vaccine (i.e., a vaccine with the same components as an available vaccine, but produced by a different sponsor or by a different manufacturing process and is intended to increase protection and/or to reduce side effects) or a combination vaccine to a standard licensed vaccine, the studies are usually designed as noninferiority trials, using immune responses (serological endpoints) as the primary outcome variable. In a noninferiority trial, in which proportions attaining some prespecified level of immune response are compared, the null hypothesis has the form $H_0: \pi_{v1} < \pi_{v2} - \delta$, where π is a proportion seroconverting or attaining a certain level of response, and δ is the decrease that can be tolerated.^[13] Unfortunately, the protective level of antibody (correlate of protection) is often unknown, thus complicating the establishment of an acceptable margin of equivalence (δ). The use of geometric mean tiers/geometric mean concentrations (GMTs/GMCs) to compare immunogenicity is often proposed because the titers or concentrations have been observed to have log-normal distributions. When GMTs/GMCs are compared, the hypothesis is expressed in terms of a ratio rather than a difference in proportions.^[14,15]

Studies to support consistency of manufacture are required for all products. Such studies are important because, as noted earlier, the manufacturing process is such an integral part of the product's license. These evaluations may be based on the physical characteristics of the vaccines, or may aim to show that successive lots of vaccine will evoke similar immune responses or bioassay results. The latter type, called clinical lot consistency studies, are usually designed and analyzed as equivalence studies (i.e., two-sided) because a difference in either direction is of interest.^[16] It is customary for at least three consecutively produced lots to be compared to establish the consistency of manufacture.

A change in a manufacturing process may occur after the completion of a successful efficacy trial. In this situation, there could be concern that the change may have resulted in a product that is no longer clinically equivalent to the one for which efficacy was demonstrated. This product change could occur if the manufacturing process affects the biologic variability of the product. To allay this concern, a “bridging study” may be conducted in which persons randomized to receive vaccine lots similar or identical to a lot used in the efficacy trial are compared to individuals randomized to receive the newly manufactured lots. Bridging

studies aim to demonstrate the clinical equivalence of the two sets of vaccine lots with respect to immune responses and various safety measures. If the lots are similar regarding these endpoints, then the efficacy and the safety that were observed for the lots used in the efficacy trial are inferred for the newly manufactured lots. A bridging study might also be conducted when there is a change in formulation or dosing schedule, as well as when the “change” is in a population for intended use that is different from the one in which the efficacy trial was conducted.

Bridging studies are usually much smaller than clinical efficacy studies because they rely on (continuous) immunogenicity measures rather than on detection of differences in small proportions acquiring the disease. Differences that may affect the vaccine’s immunogenicity include ethnicity, age, gender, and environmental factors. Generally, the goal of bridging studies is to show that the vaccines will provide similar levels of protection under the different conditions of use (i.e., lots, populations, combinations, etc.).

An evaluation of the safety profile of an investigation vaccine is extremely important. Because preventive vaccines are given to healthy populations, there must be assurance that serious adverse events associated with the vaccine do not occur at an unacceptably high rate. Large studies are usually required to allay this concern (e.g., if 3000 individuals were vaccinated with no serious adverse events being observed, the upper 95% confidence limit for the true rate of serious adverse events would be 0.001). One serious adverse event in a study of 4750 subjects would also yield an upper limit of 0.001. Less serious adverse events can have a somewhat higher bound of acceptability. A common local reaction (e.g., redness) at the vaccination site might be bounded at 0.05 or 0.10. Although local reactions can be attributed to the vaccine in nearly all cases, the causality of systemic events requires the comparison of rates in a controlled trial. The sample size requirement for assessing the common but less serious adverse events in controlled studies (e.g., a doubling of a 10% control event rate) is almost always met, as such differences can be reliably detected in studies of a few hundred people. Less common but more serious events, however, require much larger sample sizes than typically are needed to study efficacy. Ellenberg^[17] has argued that larger controlled trials would increase our knowledge about rare events prior to licensure and would assist in assessing spontaneous reports of events after licensure.

There are several open statistical issues in vaccine evaluation.

- *Evaluation of immune correlates of protection.* These immune response levels are surrogate outcomes for clinical vaccine efficacy that permit the evaluation of new vaccines more efficiently. The usual situation is to use the data from the clinical efficacy trials of the first generation of the vaccine and to attempt to correlate one or more measures of immune response, such as serum antibody levels, with vaccine efficacy. One

issue is determining the quantitative level of immune response that will afford protection from the disease, as well as the qualitative measure of immune response that is predictive of protection. A second issue regards how a surrogate relates to the causal pathway of the disease. Fleming and DeMets^[18] note that surrogates can be useful in phase II trials, but express concern over their use in phase III trials. They provide a number of examples in therapeutic trials in which information on surrogates has been misleading. A major concern is that a factor that affects the true clinical endpoint might not affect the surrogate. In vaccine trials, information on how the surrogate (qualitative measure and quantitative level of immune response) fits in the causal pathway of disease prevention is often quite good, but concerns arise in cases where no clear correlate of protection has been documented.

- *Selection of individuals and/or events to include in the primary efficacy analysis.* It has been a common practice to count cases only in subjects who received a full immunization series (usually with an additional time period to allow effective immune response to the full injection series). Cases occurring in subjects dropping out of the study, or cases of disease occurring before the full series are completed have often been ignored, or counted only in secondary evaluations. There are arguments against this type of analysis. Excluding dropouts affects the randomization and thus raises concerns about the validity of the statistical analysis. If subjects who become ill after the first dose do not complete the series and are then excluded from the efficacy assessment, the efficacy will be overestimated. Similarly, if subjects with adverse reactions fail to complete the series, both safety and efficacy will be overestimated. Thus many would argue for the study to be analyzed by counting all cases in all randomized individuals. This approach preserves the significance level if testing a null hypothesis that $VE \leq V_0$ (where V_0 is fairly large). In this case, the usual procedure is to compute sample sizes under the null hypothesis that $VE = V_0$. It can also be argued that the dropouts may be a nonrandom subset of the group that has been randomized with respect to the probability of disease, and that omitting them may lead to biased estimates of efficacy.

Practically speaking, in most vaccine efficacy trials, which enroll only healthy individuals, compliance with the full immunization series is quite high (often in the 90–95% range) and dropouts as a result of toxicity are not common, so that an intention-to-treat analysis and “full series” analysis often yield similar estimates of efficacy. Even when the two estimates diverge numerically, it has been observed that they tend to lead to similar conclusions about whether the vaccine works well enough (e.g., a VE of 0.85 or a VE of 0.65 indicates substantial efficacy). Thus unlike the situation in therapeutic trials in which the two

analyses may lead to conflicting conclusions about efficacy in part because effect sizes are frequently small and there is a much greater rationale for dropouts being a nonrandom subset of the study population with respect to the study outcome, the two types of analyses are likely to be concordant in vaccine trials. It is currently a common practice to perform both the full series analysis and the intent-to-treat analysis of preventive vaccine efficacy. A detailed discussion of this issue in relation to vaccine efficacy trials can be found elsewhere.^[19]

- *Evaluation of vaccines against a subset of pathogen strains.* In February 2000, the first vaccine against invasive pneumococcal disease for use in infants and children, Wyeth–Lederle’s PREVNAR®, was licensed. PREVNAR® is a seven-valent vaccine, containing polysaccharides from seven disease serotypes thought to be important causative agents, conjugated to a carrier protein. In a large efficacy trial, the vaccine was shown to be 100% effective in preventing disease caused by serotypes in the vaccine. That is, all protocol-defined cases of disease occurred in the placebo arm. No correlation between immune response and protection from disease could be determined.

The dilemma is how new pneumococcal vaccines will be evaluated. With an effective vaccine now available, it may not be possible to conduct a placebo-controlled trial of a new vaccine for the same disease indication. Clinical disease endpoint trials with PREVNAR® as the active control may not be possible, because disease incidence may be too low for required sample sizes to be feasible. It is possible that new pneumococcal vaccines will be evaluated using immune response endpoints, despite the fact that no correlate with protection from disease has been found.

Moreover, some of the new pneumococcal vaccines being developed will contain more than seven serotypes. In this case, the problem is how the new vaccines containing additional serotypes can be shown to be noninferior to PREVNAR®, which does not contain the additional serotypes. To what would the immune responses to the additional serotypes in the new vaccine be compared—a historical control or an external standard? How would the suitability of such comparators be determined? Another problem is that as the number of serotypes in the new vaccines increases (there are many serotypes that are causal agents), so does the statistical problem of multiple testing and its effect on error probabilities. This statistical multiplicity issue would have to be dealt with if the evaluation of a new vaccine were based on serotype-specific efficacy. Some suggestions have incorporated the basic idea of a composite index or score summarizing the overall response to all serotypes received. However, the validity of such measures would need to be assessed.

A related issue pertains to serotype replacement, for there are some pieces of evidence that vaccination against

certain disease serotypes may cause other serotypes to emerge dominant in a population. If a clinical disease endpoint trial were feasible, for example, in another country where disease incidence is still sufficiently high, could evaluation be based on protection from disease because of *any* serotype, as opposed to serotype-specific efficacy? The issue here is the determination of the appropriate outcome for establishing the efficacy of a vaccine. Which outcome has the greatest relevance to public health? Is public health benefit from vaccination best measured by serotype-specific efficacy, or by overall protection from the disease without regard to serotype etiology? A vaccine that is highly effective against diseases caused by serotypes in the vaccine—but does not reduce overall disease incidence in the population because of replacement serotypes—would not seem to be beneficial to public health. These issues are currently being considered.

- *Phase II studies of vaccines to prevent diseases with a behavioral component.* The transmission of some infectious diseases, such as HIV, is highly related to behavior. This leads to special problems when evaluating these vaccines in efficacy trials. For example, behavioral changes could impact the efficacy assessment of these vaccines; individuals who believe that they are protected by a vaccine may engage in more behaviors that expose them to infections. (This situation may be especially prevalent in individuals “self-unblind” by being tested for anti-HIV antibody.) Specifically, if more high-risk behaviors occur (because of a belief that the vaccine is fully protective), the incidence of new HIV infection may even increase, even if the vaccine has a modest protective effect. Similar concerns arise in trials of other diseases, such as genital herpes, which have a large behavioral component in their transmission.

On the other hand, intensive counseling aimed at limiting such behavior is usually considered as a mandatory part of the studies (either as a separate arm, or in both control and vaccinated groups). Such counseling may itself reduce the incidence of infection to an extremely low level, thus requiring much larger studies to demonstrate the efficacy of the vaccine candidate. If such counseling were not provided in clinical practice, the vaccine efficacy might be different (e.g., vaccinated individuals might engage in more high-risk behavior, leading to a lowered efficacy). In this case, a biologically effective vaccine might have a low “use effectiveness” in practice.

PHARMACOGENETICS AND DRUG DEVELOPMENT

Pharmacogenetics is the study of variability in drug response as a result of genetic factors. The field of pharmacogenetics is not new—experimental studies to examine

the genetic control of human drug response started in the early 1950s—but is rapidly expanding. The statistical and mathematical studies of genes in populations and families have a much longer history.^[20–22]

The spectacular advances in human genome sequencing and the development of microarray technology to measure gene expression patterns on a global scale have generated extraordinary levels of scientific excitement and research activity in pharmacogenetics.^[23–25] The major focus of the current research is on the identification of genes and gene products that are associated with various diseases and that may provide targets for new drugs and biologics. Investigators are also searching for genes and allelic variants of genes that affect patients' response to drugs. It may be possible to identify patients who respond to a particular drug (efficacy) without showing any adverse reactions (safety).^[26–29] Methodologies from statistics and bioinformatics are making significant contributions in this field.

Some progress has already been made. HERCEPTIN®, a genetically based monoclonal antibody, was licensed in 1998 for the treatment of patients with metastatic breast cancer whose tumor overexpresses the HER2 protein. This protein is encoded by the *HER2/neu* gene located on human chromosome 17q. The overexpression of HER2 protein is observed in only 15–30% of primary breast cancer patients, however.

Another interesting example is the effect of hepatitis C virus (HCV) genotypes on the efficacy of PEG–Intron/REBETOL in patients with chronic hepatitis C. The treatment response rate (loss of HCV–RNA) for patients with HCV genotype–1 (70–75% of all people with HCV infection) is approximately half of that observed in subjects without viral genotype–1 (41% vs. 75%). In hepatitis C, unlike metastatic breast cancer, all HCV genotypes do respond to PEG–Intron/REBETOL treatment, although some genotypes are more responsive than others. This PEG–Intron/REBETOL combination therapy was licensed for the treatment of hepatitis C by the CBER in 2001.

There are many design, analysis, and regulatory issues that are relevant to future trials incorporating pharmacogenetics. The regulatory approval of genetic tests (for patient screening, entry criteria, evaluation of endpoints, etc.) would involve the design and analyses of studies to estimate diagnostic success measures such as sensitivity, specificity, and predictive values. The number of patients with a specific genetic marker available for a clinical trial will depend on the frequency of that gene in the population; for rare genes, it may take many years to enroll sufficient patients in the trial. Study sample sizes will depend on observed associations between genetic markers and therapeutic response. If the patient population is heterogeneous, consisting of genotypes with different degrees of therapeutic response, the sample size and the trial duration would be increased compared to a study focused on a high-risk population. The differential effects of genotypes on the safety profile of the drug may affect the desirable size of

the safety database. Issues related to multiple comparisons, subgroup analyses, and alpha level may be important considerations in some trials. The examination of statistical issues in these areas is ongoing.^[30–32]

BIOTERRORISM PRODUCTS

In recent years, many countries have sought ways to combat biologic weapons that may be employed against them. Biologic weapons are based on bacteria, viruses, rickettsia, fungi, or toxins produced by these agents and could cause mass morbidity and mortality. The agents thought to be most likely used in this regard are anthrax, smallpox, and plague. Ways of protecting against the impact of bioterrorist attacks include vaccines, antibiotics, and antiviral products. A special challenge arises in the evaluation of these products because the types of exposures that a population might face as a result of a bioterrorist attack generally do not occur in nature. Thus it is not possible to evaluate a vaccine against bioterrorism in a clinical study. Challenge studies, in which vaccinated subjects are deliberately exposed to the agent of concern, have been conducted for some vaccines (e.g., malaria and plague) but only in settings in which clinical disease is 100% curable with immediate treatment. That is not the case for the types of pathogens noted above, and hence challenge studies of vaccines to protect against these diseases would be unethical. For this reason, the development of reliable animal models is critical. For animal studies to be considered to provide adequate evidence of human protection, a sponsor must demonstrate that the selected animal species is a good model of the human response to the vaccine, and must show that a fully vaccinated animal is protected against the specified disease. A draft rule delineating the circumstances in which animal efficacy data will be accepted as a basis for the licensure of new vaccines and related products (e.g., immune globulins) has been proposed.^[33]

WEBSITES FOR FDA INFORMATION AND GUIDANCE DOCUMENTS

There are several guidance documents relating to biologics that are available on the FDA websites. The addresses for the FDA website and the ICH website are:

FDA website: <http://www.fda.gov/cber/guidelines.htm>

ICH website: <http://www.ifpma.org/ich5.html>

CONCLUSION

Biologics comprise products of wide-reaching public health impact, such as vaccines, as well as products, like gene therapies, that are on the cutting edge of science.

Statistical issues can be complex and challenging. With the recent focus on vaccines and immunoglobulins against diseases of bioterrorism, as well as the new developments in pharmacogenetics, the field of biologics promises to be one of the most exciting and fertile areas on the frontiers of scientific investigation.

ACKNOWLEDGMENTS

We thank Toby Silverman, Karen Goldenthal, and Douglas Pratt of CBER for incisive comments on this document. We also thank our other colleagues at CBER for providing us with many insights into the regulation of Biologics.

REFERENCES

- Brady, R.P.; Kracov, D.A. From diphtheria antitoxin to cytokine products: A remarkable scientific journey/an aging regulatory framework. *Regul. Aff.* **1991**, *3*, 105–132.
- Code of Federal Regulations. CFR 21, 600.3(h).
- Meinert, C.L. *Clinical Trials Design, Conduct, And Analysis*; Oxford University Press: New York, 1986.
- Lachenbruch, P.A.; Lynch, C.J. Tests for equivalence in dichotomous variables. *Stat. Med.* **1998**, *17*, 2207–2218.
- Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristics (ROC) Plots; Approved Guideline. In *National Committee for Clinical Laboratory Standards*; December. 1995; Vol. 15, No. 19.
- Schreiber, G.; Busch, M.P.; Kleinman, S.H.; Korelitz, J.J. The risk of transfusion-transmitted viral infections. *N. Engl. J. Med.* **1996**, *334*, 1685–1690.
- Blackwelder, W.C. Proving the null hypothesis in clinical trials. *Control. Clin. Trials.* **1982**, *3*, 345–353.
- Senn, S. Inherent difficulties with active control equivalence studies. *Stat. Med.* **1993**, *12*, 2367–2375.
- Temple, R. Problems in interpreting active control equivalence trials. *Account. Res.* **1996**, *4*, 267–275.
- Ewell, M. Comparing methods for calculating confidence intervals for vaccine efficacy. *Stat. Med.* **1996**, *15*, 2379–2392.
- Haber, M.; Longini, I.M.; Halloran, M.E. Measures of the effects of vaccination in a randomly mixing population. *Int. J. Epidemiol.* **1991**, *20*, 300–310.
- Department of Health and Human Services. *Guidance for Industry: "Guidance for the Evaluation of Combination Vaccines for Preventable Diseases"*; Federal Register, April. 1997. <http://www.fda.gov/cber/guidelines.htm>.
- Blackwelder, W.C. Similarity/equivalence trials for combination vaccines. In *Combined Vaccines and Simultaneous Administration*; Williams, J.C., Goldenthal, K.L., Burns, D.L., Lewis, B.P., Jr., Eds.; Ann. N.Y. Acad. Sci. **1995**, *754*, 321–328.
- Falk, L.A.; Midthun, K.; McBittie, L.D.; Goldenthal, K.L. The Testing and Licensure of Combination Vaccines for the Prevention of Infectious Diseases. In *Combination Vaccines*; Ellis, R., Ed.; Humana Press: Totowa, NJ, 1998.
- Horne, A.D.; Lachenbruch, P.A.; Getson, P.R.; Hsu, H.S. Analysis of studies to evaluate immune response to combination vaccines. *Clin. Infect. Dis.* **2001**, *33* (Suppl. 4), S306–S311.
- Blackwelder, W.C. Proving the null hypothesis in clinical trials. *Control. Clin. Trials.* **1982**, *3*, 345–353.
- Ellenberg, S.S. Safety considerations for new vaccine development. *Pharmacoepidemiol. Drug Saf.* **2001**, *10*, 411–415.
- Fleming, T.R.; DeMets, D. Surrogate end points in clinical trials: Are we being misled? *Ann. Intern. Med.* **1990**, *125*, 605–613. 1006.
- Horne, A.D.; Lachenbruch, P.A.; Goldenthal, K.L. Intent-to-treat analysis and preventive vaccine efficacy. *Vaccine.* **2001**, *19*, 319–326.
- Weber, W.W. *Pharmacogenetics*; Oxford University Press: London, 1997.
- Falconer, D.S. *Introduction to Quantitative Genetics*; The Ronald Press Company: New York, 1967.
- Li, C.C. *Population Genetics*; The University of Chicago Press: Chicago, IL, 1968.
- Emilien, G.; Ponchon, M.; Caldas, C.; Isacson, O.; Maloteaux, J.M. Impact of genomics on drug discovery and clinical medicine. *Q. J. Med.* **2000**, *93*, 391–423.
- Clarke, P.A.; Poole, R.T.; Wooster, R.; Workman, P. Gene expression microarray analysis in cancer biology, pharmacology, and drug development: Progress and potential. *Biochem. Pharmacol.* **2001**, *62*, 1311–1336.
- McLeod, H.L.; Evans, W. Pharmacogenomics: Unlocking the human genome for better drug therapy. *Annu. Rev. Pharmacol. Toxicol.* **2001**, *41*, 101–121.
- Roses, A.D. Pharmacogenetics. *Hum. Mol. Genet.* **2001**, *10*, 2261–2267.
- Roses, A.D. Pharmacogenetics and future drug development and delivery. *Lancet.* **2000**, *355*, 1358–1361.
- Meyer, U. Pharmacogenetics and adverse drug reactions. *Lancet* **2000**, *356*, 1667–1671.
- Wolf, C.R.; Smith, G.; Smith, R.L. Pharmacogenetics. *Br. Med. J.* **2000**, *320*, 987–990.
- Elston, R.C.; Idury, R.M.; Cardon, L.R.; Lichter, J.B. The study of candidate genes in drug trials: Sample size considerations. *Stat. Med.* **1999**, *18*, 741–751.
- Fijal, B.A.; Hall, J.M.; Witte, J.S. Clinical trials in the genomic era: Effects of protective genotypes on sample size and duration of trial. *Control. Clin. Trials* **2000**, *21*, 7–20.
- Cardon, L.R.; Idury, R.M.; Harris, T.J.R.; Witte, J.S.; Elston, R.C. Testing drug response in the presence of genetic information: Sampling issues for clinical trials. *Pharmacogenetics* **2000**, *10*, 503–510.
- Federal Register. Oct. 5. 1999, *164* (192), 21. CFR parts 314 and 601: New drug and biological products; Evidence needed to demonstrate efficacy of new drugs for use against lethal or permanently disabling toxic substances when efficacy studies in humans ethically cannot be conducted; Proposed rule.

Biomarker Development with Applications to Genetics Data: Statistical Tests

Gang Zheng and Jungnam Joo

National Heart, Lung and Blood Institute, Bethesda, Maryland, U.S.A.

Abstract

This entry reviews some recent developments using receiver operating characteristic (ROC) analysis for gene expression microarray data. It also discusses several procedures for obtaining robust test statistics and provides an example of this type of testing.

INTRODUCTION

Biomarker development usually involves several phases. For example, Pepe et al.^[1] provided five phases of biomarker development in cancer research including preclinical exploratory studies, clinical assay development, retrospective repository studies, prospective screening studies, and cancer control studies. These phases can be used for a biomarker development of other diseases (such as cardiovascular disease and asthma) as well. Statistical methods are employed to pursue the specific aims for each of these phases. In early developmental phases of a biomarker, the ability of a biomarker to distinguish diseased from nondiseased subjects needs to be evaluated. Sensitivity and specificity are frequently used measures for a biomarker with binary outcomes in this context. If a biomarker has multiple or continuous outcomes, a receiver operating characteristic (ROC) curve is examined. Standard statistical testing methods are available for these analyses. In retrospective studies, on the other hand, the association between a biomarker and disease is usually tested. A number of statistical test procedures have been developed for testing the association in various situations.

Recent techniques such as gene expression microarrays and proteomics offer novel approaches in the identification of potential markers and promise an improvement in diagnostics and therapeutics development. However, due to the huge number of markers available for analysis, investigators are at great risk of false findings. Therefore, appropriate statistical methods and evaluations are essential. In the first part of this entry, we review some recent developments using ROC analysis for gene expression microarray data. Specifically, we review ROC analysis briefly and how it is applied to gene expression microarray data to evaluate the diagnostic accuracy of biomarkers selected from thousands of markers.

Testing association between a marker and a disease is an important component for detecting genetic markers.

In the second part of this entry, we review robust statistical tests for detecting genetic markers. One feature of the analysis for a genetic biomarker is that optimal statistical tests usually depend on the underlying genetic model, that is, the mode of inheritance corresponding to the recessive, additive, or dominant disease. For many complex diseases, however, the mode of inheritance is unknown. Mis-specification of a genetic model may have substantial impact on genetic marker development. Robust test statistics are useful in this situation. We discuss several robust procedures. To test genetic markers in association studies, several designs are available. We focus on two of them: case-control studies and the case-parents' trio design. The impact of diagnostic errors of cases and controls on an association test using the case-control design is discussed. An example of applying robust tests to linkage analysis using affected sib pairs is also presented. Conclusions are given at the end.

ROC ANALYSIS FOR GENE EXPRESSION DATA

ROC analysis is one of the most popular statistical methods that is used for evaluating the accuracy of diagnostic classification of a biomarker. For a biomarker with binary outcomes, a natural way of evaluating its performance is to examine its sensitivity (true-positive rate) and specificity (true-negative rate). The generalization of the concept of these statistics to continuous (or polytomous) outcomes yields the ROC curve. It is a plot of sensitivity (true-positive rate or TPR) vs. one minus specificity (false-positive rate or FPR), i.e., $\Pr(Y \geq c | D+)$ vs. $1 - \Pr(Y < c | D-) = \Pr(Y \geq c | D-)$ for a series of cut-off values c , where Y denotes the biomarker and $D+$ and $D-$ represent subjects with and without disease, respectively. Here, the assumption that higher values of Y are associated with disease is made. Fig. 1 shows an example of an ROC curve from a simulated continuous marker for 50 case and 50 control subjects.

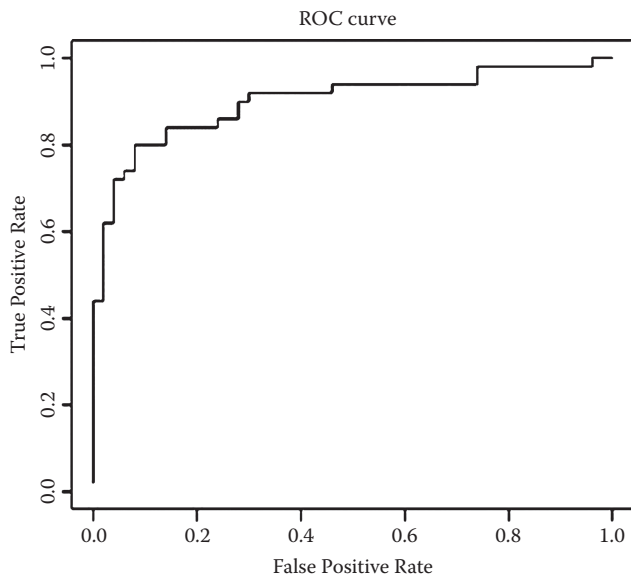


Fig. 1 ROC curve

Mathematically, the ROC curve can be written as $ROC(t) = S_{D+}(S_{D-}^{-1}(t))$ for $t \in (0, 1)$, where S_{D+} and S_{D-} represent the survivor functions for Y in diseased and nondiseased populations, respectively. This curve displays the trade-off between sensitivity and specificity (increase in sensitivity comes with decrease in specificity by lowering cut-off value and vice versa). It is a monotonic increasing function from $[0, 1]$ to $[0, 1]$, and better tests have the ROC curve located closer to the top left corner. Various methods for estimating ROC curves in ordinal rating and continuous data are summarized in Pepe.^[2] Summary measures for evaluating diagnostic accuracy based on ROC curves include the true-positive rate corresponding to a given, acceptable false-positive rate, the area under ROC curve (AUC), and the area under ROC curve in a restricted range of false-positive rates (partial AUC). Evaluating the diagnostic accuracy of a biomarker using prespecified true- and false-positive rates is quite straightforward. For instance, if a biomarker has continuous outcome, one can test whether the true-positive rate is less than a prespecified threshold TPR_0 given an acceptable false-positive rate FPR_0 . From the definition and properties of the ROC curve, this null hypothesis can be written as $H_0 = ROC(FPR_0) = TPR_0$.^[1] If a biomarker is binary, the null hypothesis of $TPR \leq TPR_0$ or $FPR \geq FPR_0$ can be tested.

The AUC can be interpreted as the probability that the class prediction of a randomly chosen diseased subject is more indicative of disease than the class prediction of a randomly chosen nondiseased subject. From this concept, nonparametric estimations of AUC and its standard error are possible with the Mann–Whitney U -statistic.^[3] This index, however, has been criticized since a large part of the area comes from the rightmost part of the curve that includes false-positive rates of no clinical relevance.

This consideration leads to a partial AUC, the area under ROC curve over a range of relevant false-positive rates.^[4,5]

One example of the application of ROC analysis for biomarker development can be found in microarray studies. Gene expression microarray is a technology that monitors the expression levels of thousands of genes in different specimens. Among the wide range of objectives of this novel technology, one common objective is to predict the diagnostic class of a patient based on the gene expression profile of the disease tissue from that patient.^[6] By simultaneous screening of multitude of biomarkers, gene expression microarrays provide a great potential for improving diagnostic classification and therapeutics development. The accuracy of diagnostic classification, however, needs to be closely examined for its clinical applications. In this context, the application of ROC analysis is natural. Here, we review several steps in developing and evaluating diagnostic classification based on gene expression profiles.

The first step is called feature selection or gene selection, in which marker genes that are closely related to disease classes are selected. The number of genes studied in microarray analysis is usually more than several thousand, and a large portion of them is usually not informative. It is, therefore, not efficient to include all of them in a classification model. The most commonly used gene screening procedure is to rank genes based on the value of univariate test statistics. For example, one can consider t or Mann–Whitney statistics for binary classes and F -statistics for polytomous classes. A specified significance level will be used to select significant genes. Various methods for gene selection with consideration of multiple comparisons are proposed and summarized in Dudoit et al.^[7]

In the second step, classification models based on selected genes are determined. Dudoit, Fridlyand, and Speed^[8] provided a comparison of several methods. The methods they compared include nearest neighbors and several variants of linear discriminant analysis, as well as more complicated ones such as classification trees. One of their main conclusions is that the simple classifiers, such as diagonal linear discriminant analysis and nearest neighbor classification, perform well compared to more sophisticated methods.

Evaluating the performance of a classification model may be the most important step for its clinical application. Receiver operating characteristic analysis is most commonly used in this context. Previously mentioned measures (area under ROC curves and partial area under ROC curves) can be utilized to measure the ability of a predictor to distinguish between classes. However, the accuracy of classification cannot be properly evaluated when all samples are used to develop a classification model. If evaluation is performed on the samples that were used to develop a classification model, it is well known that the performance will be overestimated. When independent samples are not available, internal validation is necessary.

Steyerberg et al.^[9] evaluated the performance of several internal validation methods for classification models including split-sample, cross-validation, and bootstrapping approaches. In the split-sample method, data are split into two parts, one to develop the model and the other to evaluate its performance. Cross-validation is an extension of the split-sample method. In this method, the split-sample method is repeated so that all subjects have once served to test the model. The average is taken to evaluate the performance. The most efficient method they found for the logistic regression model is the bootstrapping approach in which bootstrap samples are drawn from the original dataset with replacement. Models are usually developed in bootstrap samples, and they are tested either in the original sample or in those subjects not included in the bootstrap sample. One can use incomplete validation in which the same genes are used throughout the process. However, it is always desirable to use complete validation that repeats the entire model development procedure including gene selection.

As we have reviewed, ROC analysis is useful for evaluating the diagnostic and prognostic classification based on gene expression. But its contribution to biomarker development using gene expression is more global. It has been used to evaluate reasonable statistics for gene selection,^[10] to define expression thresholds that yields reduced misclassification,^[11] and to compare various methods for preprocessing of microarray data.^[12]

ROBUST TESTS FOR BIOMARKER DEVELOPMENT

Two robust tests are useful in detecting genetic markers when the underlying model is unknown. Suppose we have a family of optimal test statistics $\{Z_i : i \in \Lambda\}$, where $\Lambda = \{1, 2, \dots, k\}$ is an index of k underlying models. For example, Z_1 , Z_2 , and Z_3 are optimal test statistics for the recessive, additive, and dominant models ($\Lambda = \{1, 2, 3\}$), respectively. Assume that, under the null hypothesis, each Z_i asymptotically follows a standard normal distribution and that their correlation matrix is given by $\rho_{ij} = \text{Corr}_{H_0}(Z_i, Z_j)$. If all Z_i are significant or none of them are significant at the usual $\alpha = 0.05$ level, we would reject or accept, respectively, the null hypothesis regardless of the underlying model.^[13] In general, however, when some Z_i are significant while others are not, using either Z_i would result in loss of efficiency when the model is mis-specified. Furthermore, the conclusion depends on the choice of test statistics. In practice, robust tests retaining high relative efficiency and being independent of model assumptions would be desirable.

Two robust tests, the maximin efficiency robust test (MERT) and the maximal test (MAX), have been studied in the literature.^[14–16] It has been shown that they are useful in biomarker development.^[17–19] Suppose that the true

model $s \in \Lambda$ with optimal test Z_s is unknown. Define the asymptotic relative efficiency (ARE) of a test statistic Z_i ($i \in \Lambda, i \neq s$) to Z_s as $\text{Eff}(Z_i : Z_s)$. Then, $\inf_{s \in \Lambda} \text{Eff}(Z_i : Z_s)$ is the worst ARE or the test Z_i over all possible true models in Λ . A good robust test should have the best ARE among all worst situations when the true model is unknown. The MERT, denoted by Z_{MERT} , is the test whose ARE reaches $\sup_{i \in \Lambda} \inf_{s \in \Lambda} \text{Eff}(Z_i : Z_s)$, i.e.,

$$\text{Eff}(Z_{\text{MERT}} : Z_s) = \sup_{i \in \Lambda} \inf_{s \in \Lambda} \text{Wff}(Z_i : Z_s)$$

From Gastwirth,^[14,15] when $\rho_0 = \min_{i,j \in \Lambda} \rho_{ij} > \delta > 0$ for some δ , MERT uniquely exists and is the convex combination of all Z_i . In general, obtaining MERT requires a numerical algorithm. In many applications, however, MERT can be written as a linear combination of the extreme pair, i.e., two tests with the minimum correlation. Suppose that the minimum correlation $\rho_0 = \rho_{i_1 i_2}$ is reached at two tests Z_{i_1} and Z_{i_2} , $i_1, i_2 \in \Lambda$. Then, a linear combination of the extreme pair given by

$$Z_{\text{COM}} = \frac{Z_{i_1} + Z_{i_2}}{\{2(1 + \rho_{i_1 i_2})\}^{1/2}} = \frac{Z_{i_1} + Z_{i_2}}{\{2(1 + \rho_0)\}^{1/2}}$$

asymptotically follows a standard normal distribution under the null hypothesis with ARE equal to $\text{Eff}(Z_{\text{COM}} : Z_s) = (1 + \rho_0)/2$. Thus, when $\rho_0 = 0.70$, its ARE is 85%. A large correlation between the extreme pair indicates that the family of models are “close” to each other while a small correlation represents a “wider” family of models. The necessary and sufficient condition for Z_{COM} to be the MERT for the family of statistics $\{Z_i : i \in \Lambda\}$ is that

$$\rho_{i_1 j} + \rho_{j i_2} \geq 1 + \rho_{i_1 i_2} \quad (1)$$

for each $j \in \Lambda$. Condition (1) may not be easy to verify analytically. When the minimum correlation ρ_0 is small, MERT and Z_{COM} may not be sufficient. From Freidlin et al.,^[16] a maximal statistic (MAX) should be used when $\rho_0 < 0.50$ and the MAX and MERT (and Z_{COM}) have similar power robustness when $\rho_0 \geq 0.75$. The MAX defined in Freidlin, Podgor, and Gastwirth^[16] is $Z_{\text{MAX}} = \max_{i \in \Lambda} Z_i$. They also defined several simple versions of MAX such as $Z_{\text{MAX2}} = \max(Z_{i_1}, Z_{i_2})$, $Z_{\text{MAX3}} = \max(Z_{i_1}, Z_{\text{COM}}, Z_{i_2})$, where (Z_{i_1}, Z_{i_2}) is the extreme pair, or $Z_{\text{MAX3}} = \max(Z_{i_1}, Z_{i_2})$, where Z_j corresponds to some model between the extreme pair. For example, $Z_{\text{MAX3}} = \max(Z_{\text{REC}}, Z_{\text{ADD}}, Z_{\text{DOM}})$, where the recessive and dominant model is the extreme pair. The reason for considering MAX3 rather than MAX2 is that using only the extreme pair may not cover all underlying models in a family with a small minimum correlation. Thus, including the test corresponding to a model between the extreme pair would increase the power of the maximal test. In addition, the maximal test statistics over all models is hard to use in applications due to its unavailable asymptotic

distribution. Here, we focus on Z_{MERT} (or Z_{COM}) and Z_{MAX3} . For MAX3, we prefer $Z_{\text{MAX3}} = \max(Z_{i_1}, Z_{\text{COM}}, Z_{i_2})$ for two reasons. First, it takes the middle test that has equal correlation with each of the two extreme ones. Second, in simulation, it depends only on the minimum correlation.

The distribution and p -values of MAX3, $Z_{\text{MAX3}} = \max(Z_{i_1}, Z_{\text{COM}}, Z_{i_2})$, can be found empirically (similarly for MAX2 or other MAX3). Note that, under the null hypothesis, Z_{i_1}, Z_{i_2} asymptotically follow the bivariate normal distribution with zero means, unit variances, and the correlation matrix given by

$$\begin{pmatrix} 1 & \rho_0 \\ \rho_0 & 1 \end{pmatrix}$$

Hence, one generates a random sample of large size (n) from the bivariate normal distribution with the above correlation structure, where ρ_0 is evaluated using the observed data. For each replicate, Z_{COM} and Z_{MAX3} are calculated. So we obtain n Z_{MAX3} , which form an empirical distribution of Z_{MAX3} . The p -values can be obtained using tails of the empirical distribution.

EXAMPLES: APPLICATION OF ROBUST TESTS TO GENETIC ASSOCIATION AND LINKAGE

In genetics, a marker usually refers to a known physical location locus on a chromosome. It has at least two alleles that are different forms of the same gene. We consider a marker with only two alleles, denoted as M and N . This biallelic marker forms three possible genotypes: NN , NM , or MM . Each individual has only one of them at the marker locus. Define penetrances as the probabilities of disease given one of the three genotypes, that is, $f_0 = \Pr(\text{disease} | NN)$, $f_1 = \Pr(\text{disease} | NM)$, and $f_2 = \Pr(\text{disease} | MM)$. If there is no association between the marker and disease, the null hypothesis is $H_0 : f_0 = f_1 = f_2$. Otherwise, the marker is associated with the disease, in which case one of the alleles should have higher risk than the other, say, M is a high-risk allele and N is a normal one. So the alternative hypothesis is $H_a : f_0 \leq f_1 \leq f_2$ where at least one strict inequality holds. Using f_0 as a baseline penetrance, the null and alternative hypotheses can be equivalently stated using genotype relative risks, defined as $r_i = f_i / f_0$ for $i = 1, 2$. In genetics, a genetic model refers to a specific mode of inheritance corresponding to the recessive, additive, or dominant disease. In terms of penetrances, a genetic model is recessive, additive, or dominant when $f_0 = f_1$ (or $r_1 = 1$), $f_1 = (f_0 + f_2)/2$ [or $r_1 = (1 + r_2)/2$], or $f_1 = f_2$ (or $r_1 = r_2$), respectively.

Linkage studies are different from association studies. In linkage analysis, one tests if a marker and an unknown disease trait locus on the same chromosome are more likely to be transmitted together than expected to offspring. Linkage analysis usually requires genotype data

from several generations in order to determine the alleles shared identical by descent (IBD) by family members.

In association studies, many designs have been proposed to detect genetic markers. We review the use of robust statistics in two important designs. One is a family-based design using case–parents' trios^[20,21] and the other is population-based design using unmatched cases and controls.^[25] In linkage analysis, the design using affected sib pairs is illustrated.^[22,23]

Case-Parents' Trio Design

In this design, cases (diseased children) and their parents are ascertained from a population and their genotypes are obtained. For a marker with two alleles M and N , there are six possible parental mating types: (1) $MM \times MM$; (2) $MM \times NM$; (3) $MM \times NN$; (4) $NM \times NM$; (5) $NM \times NN$; and (6) $NN \times NN$, where “ $\times$ ” means mating, and parental genotypes are given on both sides of the mating “ $\times$ ”. The six mating types are given in the first column of Table 1. The second column gives case genotypes for each mating type and the third column is the count of sample size of trios (case and parents) under each mating type. The last three columns contain the probabilities of a case genotype given parental mating type and the disease status of offspring for the recessive (REC), additive (ADD), and dominant (DOM) models, respectively. The derivation of these probabilities can be found in Schaid and Sommer.^[21]

From Table 1, it is seen that mating types 1, 3, and 6 are not informative for the likelihood, as the corresponding conditional probabilities are constant. Thus, inference is based on mating types 2, 4, and 5. (Note that only two mating types are informative for the recessive or dominant model.) Conditional on parental mating and disease status, case genotypes of the six mating types are independent, although the noninformative mating types do not contribute to the likelihood function. Hence, for a given genetic model, the likelihood function is the product of the likelihood functions for the informative mating types with binomial distributions (mating types 2 and 5) and a multinomial distribution (mating type 4). Testing genetic association is equivalent to testing $H_0 : r = 1$. Denote the likelihood function for a given model as $L(r)$. The score test for $H_0 : r = 1$ can be obtained by

$$Z = \frac{\partial \log L(r) / \partial r}{\{-\partial^2 \log L(r) / \partial r^2\}^{1/2}} \Big|_{r=1} \quad (2)$$

For recessive, additive, and dominant models, respectively, (2) can be written as

$$Z_{\text{REC}} = (4n_{22} + 4n_{42} - 2n_2 - n_4) / (4n_2 + 3n_4)^{1/2} \quad (3)$$

$$Z_{\text{ADD}} = (n_{22} + 2n_{42} + n_{51} - n_{21} - 2n_{40} - n_{50}) / (n_2 + 2n_4 + n_5)^{1/2} \quad (4)$$

Table 1 Conditional Probabilities of Case Genotypes Given Parental Mating Types and Offspring Disease Status for Various Genetic Models

Parental mating type	Case genotype	Count	Conditional probability		
			REC ($r_1 = 1, r_2 = r$)	ADD ($r_1 = r, r_2 = 2r - 1$)	DOM ($r_1 = r, r_2 = r$)
1. $MM \times MM$	MM	n_{12}	1	1	1
2. $MM \times NM$	MM	n_{22}	$\frac{r}{r+1}$	$\frac{2r-1}{3r-1}$	$\frac{1}{2}$
	NM	n_{21}	$\frac{1}{r+1}$	$\frac{r}{3r-1}$	$\frac{1}{2}$
3. $MM \times NN$	NM	n_{31}	1	1	1
4. $NM \times NM$	MM	n_{42}	$\frac{r}{r+3}$	$\frac{2r-1}{4r}$	$\frac{r}{3r+1}$
	NM	n_{41}	$\frac{2}{r+3}$	$\frac{1}{2}$	$\frac{2r}{3r+1}$
	NN	n_{40}	$\frac{1}{r+3}$	$\frac{1}{4r}$	$\frac{1}{3r+1}$
5. $NM \times NN$	NM	n_{51}	$\frac{1}{2}$	$\frac{r}{r+1}$	$\frac{r}{r+1}$
	NN	n_{50}	$\frac{1}{2}$	$\frac{1}{r+1}$	$\frac{1}{r+1}$
6. $NN \times NN$	NN	n_{60}	1	1	1

$$Z_{\text{DOM}} = (n_4 + 2n_5 - 4n_{40} - 4n_{50}) / (3n_4 + 4n_5)^{1/2}$$

where $n_2 = n_{21} + n_{22}$, $n_4 = n_{40} + n_{41} + n_{42}$, and $n_5 = n_{50} + n_{51}$.

For recessive, additive, and dominant models, Z_{REC} , Z_{ADD} , and Z_{DOM} correspond to optimal test statistics, respectively. When the underlying genetic model is unknown, Zheng, Freidlin, Gastwirth^[24] showed that REC and DOM is the extreme pair with minimum correlation less than 1/3 and that condition (1) holds. Thus, Z_{COM} is the MERT. In Zheng, Freidlin, Gastwirth,^[24] $Z_{\text{MAX3}} = \max(Z_{\text{REC}}, Z_{\text{ADD}}, Z_{\text{DOM}})$ was studied. Here, we also consider $Z_{\text{MAX3}} = \max(Z_{\text{REC}}, Z_{\text{MERT}}, Z_{\text{DOM}})$. Table 2 presents an empirical power comparison of the three optimal test statistics as well as MERT and two MAX3. The simulation procedure was the same as described in Zheng, Freidlin, Gastwirth.^[24] Hardy–Weinberg equilibrium (HWE) is assumed with allele frequency $p = \Pr(M) = 0.2$, 100 case–parents’ trios and 10,000 replicates. The alternatives are chosen so that the optimal tests have approximately 80% of power. When the underlying model is unknown, the minimum power of each of the five test statistics across the three genetic models is highlighted. MAX3 and MERT have higher minimum powers across five models than optimal tests, which indicates the property of robustness compared to the three optimal statistics. Also, the two MAX3s perform similarly. Hence either one can be used.

Table 2 Empirical Power Estimates for the Optimal and Robust Tests for Trio Designs Under HWE and $p = 0.2$ with 100 Trios and 10,000 Replications

True model	Tests					
	Z_{DOM}	Z_{REC}	Z_{ADD}	Z_{MERT}	Z_{MAX3}^a	Z_{MAX3}^b
NULL	0.048	0.041	0.044	0.051	0.035	0.034
DOM	0.805	0.133	0.759	0.524	0.630	0.648
REC	0.121	0.796	0.459	0.656	0.702	0.698
ADD	0.783	0.297	0.801	0.704	0.671	0.677

DOM: $f_0 = 0.0200, f_1 = 0.0343, f_2 = 0.0343$.

REC: $f_0 = 0.0200, f_1 = 0.0200, f_2 = 0.0510$.

ADD: $f_0 = 0.0200, f_1 = 0.0325, f_2 = 0.0450$.

^a $Z_{\text{MAX3}} = \max(Z_{\text{REC}}, Z_{\text{MERT}}, Z_{\text{DOM}})$.

^b $Z_{\text{MAX3}} = \max(Z_{\text{REC}}, Z_{\text{MERT}}, Z_{\text{DOM}})$.

Unmatched Case–Control Design

The case–control design using population controls is an alternative approach. Data are displayed in Table 3 where cases and controls are sampled independently from their respective populations and their genotypes are obtained.

A linear trend test was proposed to analyze this dataset, which is obtained from a logistic regression model where genotypes are the only covariate.^[25] The genotype is coded using increasing scores as a covariate in the logistic regression model; that is, 0, x , and 1 for NN , NM and MM , respectively, where $0 \leq x \leq 1$. Optimal choices of scores

Table 3 Data of Case–Control Studies for Testing Association Between a Marker and a Disease

	Genotype			Total
	<i>NN</i>	<i>NM</i>	<i>MM</i>	
Cases	r_0	r_i	r_2	R
Controls	s_0	s_i	s_2	S
Total	n_0	n_i	n_2	n

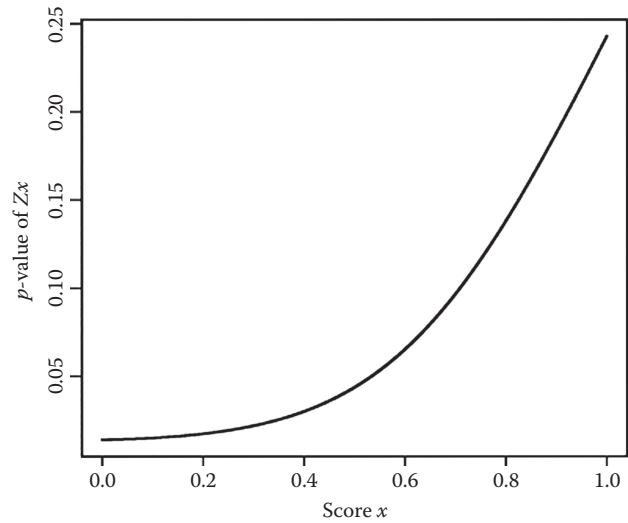
are determined by the underlying genetic model.^[25,26] For recessive, additive, and dominant diseases, optimal scores are $x = 0$, $x = 1/2$, and $x = 1$, respectively. In general, the score x can be any value in the range $[0, 1]$. The trend test^[25] is a function of x given by

$$Z_x = \frac{\sum_{i=0}^2 x_i (Sr_i - Rs_i) / n}{\left\{ (RS/n^3) \left[n \sum_{i=0}^2 x_i^2 n_i - \left(\sum_{i=0}^2 x_i n_i \right)^2 \right] \right\}^{1/2}}$$

where $(x_0, x_1, x_2) = (0, x, 1)$. Under the null hypothesis, Z_x asymptotically follows the standard normal distribution. A procedure using isotonic regression and a pooled adjacent violator algorithm^[27] was used to determine whether or not Z_x is independent of the choice of x ,^[13] that is, whether or not $\min_{x \in [0,1]} Z_x$ is greater than or $\max_{x \in [0,1]} Z_x$ is less than the critical value (1.96 for a two-sided alternative with $\alpha = 0.05$). In either situation, we can reject or accept the null hypothesis regardless of the score x .

In general, when the underlying genetic model is unknown, Freidlin, Zheng, and Gastwirth^[19] studied two robust tests $Z_{\text{COM}} = (Z_0 + Z_1) / \{2(1 + \text{Corr}_{H_0}(Z_0, Z_1))\}^{1/2}$ and $Z_{\text{MAX3}} = \max(Z_0, Z_{1/2}, Z_1)$, and proved that $Z_{\text{COM}} = Z_{\text{MERT}}$. Simulation results in Freidlin, Zing, and Gastwirth^[19] showed similar patterns as in Table 2. To illustrate the use of robust test statistics, we apply them to a real dataset given in Shahbazi et al.^[28] to detect whether the epidermal growth factor gene is a biomarker for malignant melanoma. The data are summarized as $(r_0, r_1, r_2) = (6, 8, 10)$ and $(s_0, s_1, s_2) = (32, 47, 20)$. In Fig. 2, we draw the p -values for Z_x over all possible x . It shows that p -values depend on the choice of x . A conclusion cannot be drawn if there is no scientific knowledge about the underlying model for the choice of x . Hence, robust tests can be used as they are independent of x . From Freidlin, Zheng, and Gastwirth,^[19] p -values for Z_{MERT} and MAX3 were 0.041 and 0.031, respectively.

In case–control studies, the impact of diagnostic errors of cases and controls on trend tests has been recently examined by Zheng and Tian.^[29] Under the null hypothesis of no association between a marker and a disease, the diagnostic errors of cases and controls have no effect on Type I error. Under the alternative hypothesis, however, they have substantial effect on the power of case–control studies. To check this, we use an example given in Zheng and Tian.^[29]

**Fig. 2** p -Values of Z_x over all possible x

Taylor et al.^[30] examined a genetic marker for Alzheimer's disease (AD) in two different studies. The cases in the first study were autopsy-confirmed while in the second study they were clinically diagnosed. The observations for these two studies are given by $(r_0, r_1, r_2) = (67, 95, 45)$, $(s_0, s_1, s_2) = (95, 158, 37)$ and $(r_0, r_1, r_2) = (81, 116, 42)$, $(s_0, s_1, s_2) = (89, 111, 44)$, respectively. Comparing cases and controls with two copies of the risk allele, Taylor et al.^[30] found that the marker is associated with AD in the first study but not in the second. Using the score $x = 0$ in the trend test Z_x , Zheng and Tian^[29] obtained $|Z_0| = 2.659$ ($p = 0.008$) for the first study and $|Z_0| = 0.132$ ($p > 0.05$) for the second one. They further showed that, without diagnostic errors, the power for both studies would be about 95%. If one assumes that the cases in the second study were diagnosed clinically with 95% sensitivity and 95% specificity, then the power of the trend test reduces to 71%. This suggests that, when doing a confirmation study, the diagnosis criterion has to be the same as in the previous study.

Linkage Analysis Using Affected Sib Pairs

Any sib pair shares either 0, 1, or 2 alleles IBD at a marker locus. Denote the probabilities of sharing (0, 1, 2) alleles IBD as $p = (p_0, p_1, p_2)$ with $p_0 + p_1 + p_2 = 1$. For affected sib pairs, the null (no linkage) probabilities are $p_{H_0} = (1/4, 1/2, 1/4)$. Whittemore and Tu^[31] studied the following family of models for $p = (p_0, p_1, p_2)$:

$$\left\{ p : p = (p_0, p_1, p_2) = \lambda(0, x, 1-x) + (1-\lambda)p_{H_0} ; \right. \\ \left. \lambda \in [0,1], x \in [0,1/2] \right\}$$

The null hypothesis is $H_0 : \lambda = 0$ with a nuisance parameter x not defined under the null. Here x is related to the genetic model by $x = 0$ corresponding to rare recessive disease and $x = 1/2$ to additive disease. Among n affected

sib pairs, (n_0, n_1, n_2) are the observed counts for sib pairs sharing (0, 1, 2) alleles IBD. The score test for $H_0^{[31,17]}$ as a function of x can be written as

$$Z_x = \frac{n^{1/2} W_x (\hat{p} - p_{H_0})'}{\left\{ W_x \left(\text{diag}(p_{H_0}) - p_{H_0}' p_{H_0} \right) W_x' \right\}^{1/2}},$$

where $W_x = (0, x/(2(1-x)), 1)$, $p = (n_0/n, n_1/n, n_2/n)$, $i = 0, 1, 2$, and $\text{diag}(p_{H_0})$ is a diagonal matrix with the three elements equal to $1/4, 1/2, 1/4$, respectively. As the minimum correlation is 0.8165 between Z_0 and $Z_{1/2}$, MERT should be used where $Z_{\text{MERT}} = (Z_0 + Z_{1/2})/2(1.8165)^{1/2}$. Other optimal properties of this MERT in terms of power robustness and model selection were discussed by Whittemore and Tu.^[31] For another application in cancer area, see Cerami et al. (2014).^[32]

CONCLUSIONS

In this entry, we reviewed some recent developments for ROC analysis using microarray gene expression data and robust tests for identifying genetic markers. In the first part, several steps for developing and evaluating diagnostic classification based on the gene expression profiles were reviewed. In the second part, several robust tests were reviewed and illustrated, using three examples of detecting a genetic marker in association studies and linkage analysis.

ACKNOWLEDGMENT

The authors would like to thank the Editor of EBS, Dr. Shein-Chung Chow, for inviting us to contribute this entry.

Minor updates have been made by the Editor-in-Chief in order to reflect developments since publication of the *Encyclopedia of Biopharmaceutical Statistics, Third Edition*.

REFERENCES

1. Pepe, M.S.; Etzioni, R.; Feng, Z.; Potter, J.D.; Thompson, M.L.; Thornquist, M.; Winget, M.; Yasui, Y. Phases of biomarker development for early detection of cancer. *J. Natl. Cancer Inst.* **2001**, *93*, 1054–1061.
2. Pepe, M.S. Receiver operating characteristic methodology. *J. Am. Stat. Assoc.* **2000**, *95*, 308–311.
3. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36.
4. Thompson, M.L.; Zucchini, W. On the statistical analysis of ROC curves. *Stat. Med.* **1989**, *8*, 1277–1290.
5. Wieand, S.; Gail, M.H.; James, B.R.; James, K.L. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **1989**, *76*, 585–592.
6. Simon, R. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *Br. J. Cancer* **2003**, *89*, 1599–1604.
7. Dudoit, S.; Yang, Y.W.; Callow, M.J.; Speed, T.P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sin.* **2002**, *12*, 111–139.
8. Dudoit, S.; Fridlyand, J.; Speed, T.P. *Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data*. Technical Report 576; Department of Statistics. University of California—Berkeley: Berkeley, 2000.
9. Steyerberg, E.W.; Harrell, F.E.; Borsboom, G.J.J.M.; Eijkemans, M.J.C.; Vergouwe, Y.; Habbema, J.D.F. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.* **2001**, *54*, 774–781.
10. Lönnstedt, I.; Speed, T.P. Replicated microarray data. *Stat. Sin.* **2002**, *12*, 31–46.
11. Bilban, M.; Buehler, L.K.; Head, S.; Desoye, G.; Quaranta, V. Defining signal thresholds in DNA microarrays: Exemplary application for invasive cancer. *BMC Genomics* **2002**, *3*, 1–19.
12. Cope, L.M.; Irizarry, R.A.; Jaffee, H.A.; Wu, Z.; Speed, T.P. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* **2003**, *1*, 1–10.
13. Zheng, G. Use of max and min scores for trend tests for association when the genetic model is unknown. *Stat. Med.* **2003**, *22*, 2657–2666.
14. Gastwirth, J.L. On robust procedures. *J. Am. Stat. Assoc.* **1966**, *61*, 929–948.
15. Gastwirth, J.L. The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *J. Am. Stat. Assoc.* **1985**, *80*, 380–384.
16. Freidlin, B.; Podgor, M.J.; Gastwirth, J.L. Efficiency robust tests for survival or ordered categorical data. *Biometrics* **1999**, *55*, 883–886.
17. Gastwirth, J.L.; Freidlin, B. On power and efficiency robust linkage tests for affected sibs. *Ann. Hum. Genet.* **2000**, *64*, 443–453.
18. Shih, M.; Whittemore, A.S. Allele-sharing among affected relatives: Nonparametric methods for identifying genes. *Stat. Meth. Med. Res.* **2001**, *10*, 27–55.
19. Freidlin, B.; Zheng, G.; Li, Z.; Gastwirth, J.L. Trend tests for case-control studies of genetic markers: Power, sample sizes and robustness. *Hum. Hered.* **2002**, *53*, 146–152.
20. Spielman, R.S.; McGinnis, R.; Ewens, W.J. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **1993**, *52*, 506–516.
21. Schaid, D.J.; Sommer, S.S. Genotype relative risks: Methods for design and analysis of candidate-gene association studies. *Am. J. Hum. Genet.* **1993**, *53*, 1114–1126.
22. Blackwelder, C.; Elston, R.C. A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet. Epidemiol.* **1985**, *2*, 85–97.

23. Schaid, D.J.; Nick, T.G. Sib-pair linkage tests for disease susceptibility loci: Common test vs. the asymptotically most powerful test. *Genet. Epidemiol.* **1990**, *7*, 359–370.
24. Zheng, G.; Freidlin, B.; Gastwirth, J.L. Robust TDT-type candidate- gene association tests. *Ann. Hum. Genet.* **2002**, *66*, 145–155.
25. Sasieni, P.D. From genotypes to genes: Doubling the sample size. *Biometrics* **1997**, *53*, 1253–1261.
26. Zheng, G.; Freidlin, B.; Li, Z.; Gastwirth, J.L. Choice of scores in trend tests for case–control studies of candidate-gene associations. *Biom. J.* **2003**, *45*, 335–348.
27. Kimeldorf, G.; Sampson, A.R.; Whitaker, L.R. Min and max scorings for two-sample ordinal data. *J. Am. Stat. Assoc.* **1992**, *87*, 241–247.
28. Shahbazi, M.; Pravica, P.; Nasreen, N.; Fakhoury, H.; Fryer, A.A.; Strange, R.C.; Hutchinson, P.E.; Osborne, J.E.; Lear, J.T.; Smith, A.G.; Hutchinson, I.V. Association between functional polymorphism in EGF gene and malignant melanoma. *Lancet* **2002**, *359*, 397–401.
29. Zheng, G.; Tian, X. The impact of diagnostic error on testing genetic association in case–control studies. *Stat. Med.* **2005**, *24*, 869–882.
30. Taylor, A.; Ezquerra, M.; Bagri, G.; Yip, A.; Goumidi, L.; Cotel, D.; Easton, D.; Evans, J.G.; Xuereb, J.; Cairns, N.J.; Amouyel, P.; Chartier-Harlin, M.C.; Brayne, C.; Rubinsztein, D.C. Alzheimer disease is not associated with polymorphisms in the angiotensinogen and renin genes. *Am. J. Med. Genet.* **2001**, *105*, 761–764.
31. Whittemore, A.S.; Tu, I.-P. Simple, robust linkage tests for affected sibs. *Am. J. Hum. Genet.* **1998**, *62*, 1228–1242.
32. Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B.E.; Sumer, S.O.; Aksoy, B.A.; Jacobsen, A.; Byrne, C.J.; Heuer, M.L.; Larsson, E.; Antipin, Y.; Reva, B.; Goldberg, A.P.; Sander, C.; Schult, N. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2012**, *2* (5), 401–404. doi:10.1158/2159-8290.CD-12-009.

Biomarker: Clinical Trials

Mark Chang

Veristat LLC, Southborough, Massachusetts, U.S.A.

Abstract

A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacological responses to a therapeutic intervention. This entry discusses biomarker trial designs with classifier, prognostic, and predictive markers.

INTRODUCTION

What is a biomarker? The National Institutes of Health Workshop^[1] gave the following definitions. A *biomarker* is a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacological responses to a therapeutic intervention. A *clinical endpoint* (or outcome) is a characteristic or variable that reflects how a patient feels or functions, or how long a patient survives. A *surrogate endpoint* is a biomarker intended to substitute for a clinical endpoint. Biomarkers can also be classified as classifier, prognostic, and predictive biomarkers.

A **classifier biomarker** is a marker, e.g., a DNA marker, that usually does not change over the course of study. A classifier biomarker can be used to select the most appropriate target population or even for personalized treatment. For example, a study drug is expected to have effects on a population with a biomarker, which is only 20% of the overall patient population. Because the sponsor suspects that the drug may not work for the overall patient population, it may be efficient and ethical to run a trial only for the subpopulations with the biomarker rather than the general patient population. On the other hand, some biomarkers such as RNA markers are expected to change over the course of the study. This type of markers can be either a prognostic or predictive marker.

A **prognostic biomarker** informs the clinical outcomes, independent of treatment. It provides information about natural course of the disease in individual with or without treatment under study. A prognostic marker does not inform the effect of the treatment. For example, non-small cell lung carcinoma (NSCLC) patients receiving either epidermal growth factor receptor (EGFR) inhibitors or chemotherapy have better outcomes with a mutation than without a mutation. Prognostic markers can be used to separate good and poor prognosis patients at the time of diagnosis. If expression of the marker clearly separates patients with an excellent prognosis from those with a poor prognosis, then the marker can be used to aid the decision

about how aggressive the therapy needs to be. The poor prognosis patients might be considered for clinical trials of novel therapies that will, hopefully, be more effective.^[2] Prognostic markers may also inform the possible mechanisms responsible for the poor prognosis, thus leading to the identification of new targets for treatment and new effective therapeutics.

A **predictive biomarker** informs the treatment effect on the clinical endpoint. A predictive marker can be population-specific: a marker can be predictive for population A but not population B. A predictive biomarker, as compared to true endpoints like survival, can often be measured earlier, easier, and more frequently and is less subject to competing risks. For example, in a trial of a cholesterol-lowering drug, the ideal endpoint may be death or development of coronary artery disease (CAD). However, such a study usually requires thousands of patients and many years to conduct. Therefore, it is desirable to have a biomarker, such as a reduction in post-treatment cholesterol, if it predicts the reductions in the incidence of CAD. Another example would be an oncology study where the ultimate endpoint is death. However, when a patient has disease progression, the physician will switch the patient's initial treatment to an alternative treatment. Such treatment modalities will jeopardize the assessment of treatment effect on survival because the treatment switching is response-adaptive rather than random. If a marker, such as time-to-progression (TTP) or response rate (RR), is used as the primary endpoint, then we will have much cleaner efficacy assessments because the biomarker assessment is performed before the treatment switching occurs.

Biomarkers, as compared to a true endpoint such as survival, can often be measured earlier, easier, and more frequently; are less subject to competing risks; and are less confounded. The utilization of biomarker will lead to a better target population with a larger effect size, a smaller sample size required, and faster decision making. With the advancement of proteomic, genomic, and genetic technologies, personalized medicine with the right drug for the right patient becomes possible.

Conley and Taube^[2] describe the future of biomarker/genomic markers in cancer therapy: “The elucidation of the human genome and fifty years of biological studies have laid the groundwork for a more informed method for treating cancer with the prospect of realizing improved survival. Advanced in knowledge about the molecular abnormalities, signaling pathways, influence the local tissue milieu and the relevance of genetic polymorphism offer hope of designing effective therapies tailored for a given cancer in particular individual, as well as the possibility of avoiding unnecessary toxicity.”

CHALLENGES IN BIOMARKER VALIDATION

Relationships Among Biomarker, Treatment, and True-Endpoint

Validation of biomarker is not an easy task. Validation here refers to the proof of a biomarker to be a predictive marker, i.e., a marker can be used as a surrogate marker. Before we discuss biomarker validations, let's take a close look at the three-way relationships among treatment, biomarker, and the true endpoint. It is important to be aware that the correlations between them are not transitive. In the following example,^[3] we will show that could be is a correlation (R_{TB}) between treatment and the biomarker and a correlation (R_{BE}) between the biomarker and the true endpoint, but there is no correlation (R_{TE}) between treatment and the true endpoint (Fig. 1). Therefore, the proof of a relationship between the treatment and biomarker and a relationship between the biomarker and true endpoint does not lead to the proof of the relationship between the treatment and true endpoint.

Multiplicity and False Positive Rate

Let's further discuss the challenges from a multiplicity point of view. In earlier phases or the discovery phase, we often have a large number of biomarkers to test. Running hypothesis testing on many markers can be done either with a high false positive rate without multiplicity adjustment or a low power with multiplicity adjustment. Also, if model selection

procedures are used without multiplicity adjustment as we commonly see in current practice, the false positive rate could be inflated dramatically. Another source of false positive discovery rate is the so-called publication bias. The last, but not least, source of false positive finding is due to the multiple testing conducted by different companies or research units. Imagine that 200 companies study the same biomarker, even if family-wise type I error rate is strictly controlled at a 2.5% (one-sided) level within each company, there will still be, on average, 5 companies that have positive findings about the same biomarker just by chance.

Validation of Biomarkers

We now realize the importance of biomarker validation and would like to review some commonly used statistical methods for biomarker validation.

Prentice^[4] proposes four operational criteria: (i) treatment has a significant impact on the surrogate endpoint; (ii) treatment has a significant impact on the true endpoint; (iii) the surrogate has a significant impact on the true endpoint; and (iv) the full effect of treatment upon the true endpoint is captured by the surrogate endpoint. Note that this method is for a binary surrogate.^[5] Freedman et al.^[6] argues that the last Prentice criterion is difficult statistically because it requires that the treatment effect is not statistically significant after adjustment of the surrogate marker. They further articulate that the criterion might be useful to reject a poor surrogate marker, but that it is inadequate to validate a good surrogate marker. Therefore they propose a different approach based on the proportion of treatment effect on true endpoint explained by biomarkers and a large proportion required for a good marker. However, as noted by Freedman, this method is practically infeasible due to the low precision of the estimation of the proportion explained by the surrogate. Buyse and Molenberghs^[7] propose the internal validation matrices, which include relative effect (RE) and adjusted association (AA). The former is a measure of association between the surrogate and the true endpoint at an individual level, and the latter expresses the relationship between the treatment effects on the surrogate and the true endpoint at a trial level. The practical use of the Buyse-Molenberghs method raises two concerns: (i) a wide confidence interval of RE requires a large sample size and (ii) treatment effects on the surrogate and the true endpoint are multiplicative and cannot be checked using data from a single trial.

Other methods, such as external validation using meta-analysis and two-stage validation for fast-track programs, also face similar challenges in practice. For further readings on biomarker evaluations, consult Weir and Walley,^[8] who give an excellent review; Case and Qu,^[9] who propose a method for quantifying the indirect treatment effect via surrogate markers, and Alonso et al.,^[10] who propose a unifying approach for surrogate marker validation based on Prentice's criteria.

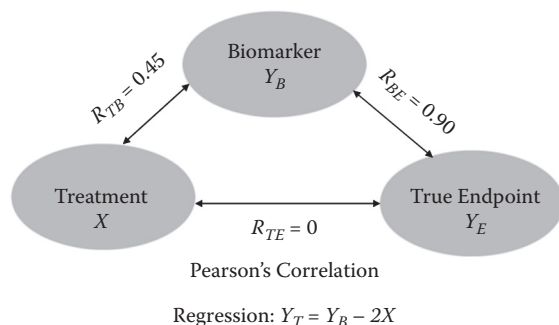


Fig. 1 Treatment-biomarker-endpoint three-way relationship

CLINICAL TRIAL WITH BIOMARKER

In reality, there are many possible scenarios: (i) same effective size for the biomarker and true endpoint, but the biomarker response can be measured earlier; (ii) bigger effective size for the biomarker and smaller for the true endpoint; (iii) no treatment effect on the true endpoint and limited treatment effect on the biomarker; and (iv) treatment effect on the true endpoint only occurs after the biomarker response reaches a threshold. Validation of biomarkers is challenging, and the sample size is often insufficient for the full validation. Given this reality, how we can use biomarker to help our clinical trial designs?

Adaptive Design with Classifier Biomarker

If evidence shows that the test drug might work better in some subpopulations with the biomarker than in those without the biomarker, we can design an adaptive trial in which the treatment effects will be evaluated for each of the subpopulations as well as in the overall population at the interim analysis. Based on the interim results, the second stage trial may be conducted on one of the subpopulations or on the overall populations. The adaptive design allows for combining the data from the first and second stages in the final analysis without undermining the validity and integrity of the trial. Statistically speaking, the family-wise false positive error rate is controlled.

Adaptive Design with Prognostic Biomarker

Insufficiently validated biomarker such as tumor response rate (RR) can be used in oncology trials for interim decision-making about whether to continue to enroll patients or not to reduce the cost. When the response rate in the test group is lower because of the correlation between RR and survival, it is reasonable to believe the test drug will be unlikely to have survival benefit. However, even when the trial stopped earlier due to unfavorable results in response rate, the survival benefit can still be tested. Chang^[3] discusses this for a Non-Hodgkin's Lymphoma (NHL) trial.

Adaptive Design with Predictive Marker

If a biomarker is proved to be predictive, then we can use it to replace the true endpoint from the hypothesis test point of view. In other words, a proof of treatment effect on predictive marker is proof of treatment effect on the true endpoint. However, the correlation between the effect sizes of treatment in the predictive (surrogate) marker and the true endpoints is desirable but unknown. This is one of the reasons that follow-up study on the true endpoint is highly desirable in the NDA accelerated approval program.

SUMMARY AND DISCUSSION

We have discussed the trial designs with classifier, prognostic, and predictive markers. These designs can be used to improve the efficiency by identifying the right population, making decisions earlier to reduce the impact of failure and delivering the efficacious and safer drugs to market earlier. Most recently, biomarker-driven adaptive design has become a very popular trial design for precision medicine.^[11] It, however, should be noted that, full validation of a biomarker is statistically challenging and sufficient validation tools are not available. Fortunately, adaptive designs with biomarkers can be beneficial even when the biomarkers are not fully validated. The technical aspects of adaptive design with biomarkers can be found in Chang's book.^[3]

ACKNOWLEDGEMENTS

Minor updates have been made by the Editor-in-Chief in order to reflect developments since publication of the *Encyclopedia of Biopharmaceutical Statistics, Third Edition*.

REFERENCES

1. De Gruttola, V.G. Considerations in the evaluation of surrogate endpoints in clinical trials: summary of a National Institutes of Health Workshop. *Control. Clin. Trials* **2001**, 22, 485–502.
2. Conley, B.A.; Taube, S.E. Prognostic and predictive marker in cancer. *Dis. Markers* **2004**, 20, 35–43.
3. Chang, M. *Adaptive Design Theory and Implementation Using SAS and R*; Chapman & Hall/CRC, Taylor & Francis: New York, 2007.
4. Prentice, R.L. Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat. Med.* **1989**, 8 (4), 431–440.
5. Molenberghs, G.; Buyse, M.; Burzykowski, T. Burzykowski, T.; Molenberghs, G.; Buyse, M. *The History of Surrogate Endpoint Validation*. In *The Evaluation of Surrogate Endpoint*, Eds.; Springer: New York, 2005.
6. Freedman, L.S.; Graubard, B.I.; Schatzkin, A. Statistical validation of intermediate endpoints for chronic diseases. *Stat. Med.* **1992**, 11, 167–178.
7. Buyse, M.; Molenberghs, G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* **1998**, 54, 1014–1029.
8. Weir, C.J.; Walley, R.J. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Stat. Med.* **2006**, 25, 183–203.
9. Qu, Y.; Case, M. Quantifying the indirect treatment effect via surrogate markers. *Stat. Med.* **2006**, 25, 223–231.
10. Alonso, A.; Molenberghs, G.; Geys, H.; Buyse, M.; Vangeneugden, T. A unifying approach for surrogate marker validation based on Prentice's criteria. *Stat. Med.* **2006**, 25, 205–221.
11. Wang, J.; Chow, S.C.; Chang, M. Biomarker-driven adaptive design for precision medicine. *J. Translational Biomarkers Diagn* **2016**, 2(1), 15–24.

Biopharmaceutics

Iain J. McGilveray

McGilveray Pharmacon, Inc., Nepean, Ontario, Canada

Abstract

This entry focuses on describing the essential features of biopharmaceutical characterization and testing of drug substances and products during their development.

INTRODUCTION

The biopharmaceutics and biopharmaceutical properties of a medicinal/drug product can be considered as the bridging information between physicochemical, pharmaceutical quality, and biological, pharmacological effects that result in therapeutic as well as toxic responses. The terms should not be confused with “biopharmaceuticals,” a term now being increasingly applied to active substances and products derived from biotechnology, such as the “recombinant” proteins.

Although regulatory documents, such as the Food and Drug Administration (FDA) guidelines for new drug applications,^[1] refer to biopharmaceutics and indeed the Code of Federal Regulations (CFR) require^[2] a separate biopharmaceutical review section in the new drug application (NDA), there does not appear to be an official definition of the term. The European regulations and the International Committee on Harmonization (ICH) guidelines have no direct reference to biopharmaceutics.

OVERVIEW

The term *biopharmaceutics* appears to have been first used in 1961 by Wagner^[3] to mean:

encompassing the relation between the nature and intensity of the biological effects observed in animals and man and, 1. the nature of the form of the drug (ester, salt, complex, etc.), 2. the physical state, particle size and surface area, 3. the presence or absence of adjuvants with the drug, 4. the type of dosage form in which the drug is administered, and 5. the pharmaceutical processes used to make the dosage form.

Gibaldi^[4] provided a brief definition of biopharmaceutics: “a major branch of the pharmaceutical sciences concerned with the relationship between the physicochemical properties of a drug in dosage form and the pharmacologic, toxicologic or clinical response after its administration.”

Biopharmaceutics is multidisciplinary, involving specialists in developing the chemistry and manufacturing control dossiers of applications and toxicologists and pharmacologists involved in documenting the pharmacokinetic and sometimes pharmacodynamic responses. Both groups apply appropriate statistical models with the advice of the statistical departments.

PROPERTIES

Biopharmaceutical properties are commonly divided into those intrinsic to the active substance and those of the finished pharmaceutical dosage form. The characteristics are defined during the development stages, which may be for the first entry of the substance or with a line extension using a different dosage form or even when generic products are being prepared. There are tests that are required only during development, and results of other tests that become the lot release specifications for each batch of substance and finished product prepared. In general, many of the physicochemical tests are used for quality control, whereas the pharmacological tests are rarely repeated postapproval.

Most of the literature on biopharmaceutics has been devoted to solid oral dosage forms of drugs that are absorbed from the gastrointestinal tract and are effective systemically. However, other routes, such as inhalation, dermal, and parenteral, also require biopharmaceutical evaluation.

Active (Drug) Substance: Physicochemical Characterization

Solubility

The aqueous solubility of the drug substance is established by obtaining saturated solutions at 37°C over 24 or 48 hr. In addition, pH/solubility profiles, at physiologically relevant values from 1 to 8, are measured. This information indicates if there might be a pH-dependent dissolution and

perhaps effects of food in delaying absorption. Substances with aqueous solubility higher than 1% are considered less likely to show solubility-limited absorption problems.^[5,6]

Intrinsic Dissolution Rate

Because this is surface area-dependent, it is defined as the amount of active substance per unit time that dissolves under standardized conditions of temperature, solvent composition, and liquid/solid interface. Commonly a nondisintegrating disk of drug substance provides the standardized surface area and liquid/solid interface.^[7] The United States Pharmacopeia (USP) has proposed an apparatus and a test to determine intrinsic dissolution.^[8] The cumulative amount dissolved over time per unit area is obtained, and data points were fitted with linear regression to obtain a slope, which is the intrinsic dissolution rate. There is a general agreement that the new drug candidates with intrinsic dissolution rates greater than 1 mg/min/cm² will have fewer absorption problems than those exhibiting rates below this level.^[5]

pK_a , Partition Coefficient, and Permeability

As well as bearing on solubility, the dissociation constant, pK_a , and partition coefficient between aqueous and organic solvents are important influences on the absorption of drugs.^[4] The lipid solubility of drugs is tested in partition experiments between buffer and such solvents as octanol, chloroform, and ethyl acetate. The majority of the organic small molecules are absorbed in the gastrointestinal tract by means of passive diffusion, with the transfer directly proportional to the magnitude of the concentration gradient across the membrane and the lipid/water partition coefficient of the drug. Because most drugs are weak acids or bases, present in the solution in both ionized and unionized forms, the unionized forms are usually lipid-soluble and can diffuse across the cell membrane. This was first proposed in the pH-partition hypothesis^[9] and, although it remains largely true that the pK_a determines the pH for the optimization of the unionized form, which will be more rapidly absorbed, the much larger surface area of the intestinal mucosa leads to a greater rate of absorption, even when the unionized form of the drug prevails at gastric pH.^[10] In addition to partition models, attempts to assess absorbability during development have involved in situ experiments with segments of animal intestine, such as the everted sac.^[11] However, more recently, the biopharmaceutics drug classification system has been promoted as being useful to predict absorbability properties.^[12] This system involves using measured human jejunal permeability^[13] and solubility properties to classify drugs into the following categories:

High solubility–high permeability
 Low solubility–high permeability
 High solubility–low permeability, and
 Low solubility–low permeability

It claims to provide a basis for estimating the absorption of drugs based on these properties of physiologic importance.^[12] This system is being used by the FDA to make decisions on postapproval changes^[14] and is also likely to be expected with the development information in NDAs.^[15] Currently, efforts are being made to search for other tests than human jejunal permeability to describe drug permeability. Such procedures as Caco-2 monolayer systems, rat intestinal models, and human mass balance are under consideration.^[15] Caco-2 increasingly is being used in fast throughput screening of drug candidates.

While many drugs are absorbed via passive diffusion, an increasing number are absorbed by special transporters.^[16] Examples are nutrient, vitamin, and ion transporters. Several cephalosporins and angiotensin-converting enzyme inhibitors are substrates for the intestinal dipeptide transporter. In addition, the role of *P*-glycoprotein countertransport processes in the intestinal mucosa is currently being elucidated and appears to be a significant influence on the bioavailability of some drugs, including verapamil and cyclosporin.^[17]

Particle Size, Surface Area, and Shape

Experiments in the development phase should determine if particle size will be a potential problem and if a batch specification is required. For sparingly soluble drugs, particle size is an important determinant of solubility. Enhancement of bioavailability by micronization to increase surface area was demonstrated for griseofulvin^[18] and spironolactone. However, for hydrophobic powders, micronization does not always lead to an increase in dissolution/absorption because aggregation may occur.^[19] For acid-labile drugs, an increase in particle size can lead to greater degradation in the gastric milieu.

The particle size of the inhalant dosage forms is an extremely important property for delivery to airways. Solution or suspension droplets from the delivery device aim to provide most of a dose with less than 5- μ m particles, sometimes termed as the *respirable volume*.^[20] For drug powder inhalers, powder rheology is complex.

Crystal Properties, Polymorphs, and Solvates

Polymorphism, the different arrangement of molecules in solid state, influencing physicochemical properties, needs to be investigated in development to decide if a batch specification is needed. Often, the less thermodynamically metastable form has more desirable solubility properties than the more stable polymorph. Depending on the rate of conversion of the metastable to stable form, it may be used in dosage forms. Examples include cimetidine^[21] and chlorpropamide.^[22]

Crystalline forms can also exist, with different external crystal shapes, or “habit.” Often, the solvent of crystallization will confer specific crystal properties that can

influence solubility. Although crystalline forms of erythromycin estolate had better dissolution properties than the anhydrous amorphous form, usually the amorphous powder of a drug substance is more soluble and bioavailable than the crystalline form. Examples of better properties with amorphous powders include chloramphenicol stearate, cortisone acetate, and novobiocin.^[23] Also, with ampicillin, the anhydrous form was found to be more soluble and gave greater bioavailability than the trihydrate.^[24] However, for carbamazepine, the dihydrate has better absorption properties than the anhydrous drug.^[25]

Stereochemical Characterization

Stereochemical molecules should be thoroughly investigated during drug development. In general, geometric isomers and diastereoisomers are considered separate chemical entities, unless there are interconversions among them or fixed ratios are important for clinical effect or stability. For enantiomers, the FDA has a 1992 Statement,^[26] and the draft ICH quality guideline 6A^[27] provides some advice. During the development, the chemical and biological stability of the enantiomers should be confirmed and the pharmacokinetic and other pharmacological properties of the individual compounds should be characterized. This will define the specifications required.

Stability

Stability, a key attribute that must be defined early in the development of a new drug candidate, includes degradative chemistry and physical form changes. These can be influenced by the route of synthesis and processing; but from the biopharmaceutical viewpoint, changes in the solubility and rate of dissolution among batches are signals of potential problems. Also, if a drug is unstable in the gastrointestinal milieu, such as degradation by hydrolysis, the alternatives include careful formulation (enteric coating) or a search for an ester resistant to hydrolysis. Although this attribute is defined in the Chemistry and Manufacturing submission file, such as advised in the ICH stability guidance,^[28] there are concerns about the potential effects on biopharmaceutics.

Salts and Prodrugs

While optimizing the stability and the biopharmaceutical properties of new drug candidates toward the formulation of consistent performance, the choice of free acid or base or salt form is important. Usually, the salt form of the weak base or acid drug made with the strong counterion, such as chloride or sodium, respectively, is more soluble and more rapidly dissolved than is the parent compound. Examples of basic drugs for which this prevails are the hydrochloride salts of epinephrine and quinidine; whereas for acids, such as barbitals, the sodium salt is

better dissolved. However, examples of free bases that are more soluble than the hydrochloride salt forms at gastric pH are chlortetracycline and methacycline,^[29] although the related tetracycline is marketed mainly as the hydrochloride salt. An interesting example is naproxen, which is formulated for chronic pain use as the acid, whereas the sodium salt (which is more rapidly dissolved) is prescribed for fast-acting analgesia.

In the preformulation assessment of inhalant dosage forms, the salt form is important.^[30] Many of the drugs formulated in these devices are weak bases prepared as salts, and the pH of solution with strong counterions such as chloride or sulfate could be problematic. Solutions with pH below 2 can cause bronchoconstriction, and thus weaker acid salts may have to be chosen.

Prodrugs are chemical modifications of drugs that are rapidly metabolized to active drug in the body, often at the intestinal villi. They are usually designed to be more lipid-soluble and thus provide better absorption. Ampicillin esters, bacampicillin^[31] and pivampicillin,^[32] were developed to improve the bioavailability of the active parent, ampicillin. Enalapril is the weakly active ethyl ester, converted to the diacid enalaprilat, which is a powerful angiotensin converting enzyme inhibitor.^[33] Valacyclovir is the L-valine ester of the active antiviral, acyclovir, which itself is poorly absorbed, and the de-esterification yields much higher serum concentrations of acyclovir by the oral route.^[34]

Excipients, Formulation Factors, and Manufacture

Overview

Drug product development has been described^[35] as a highly complex craft based on a substantial amount of empirical or semiempirical information in technical areas. The dosage form often assists or enhances bioavailability, with formulation components, but these may also hinder by accident or design (such as in prolonged-release/“sustained-action” products). Formulation development has been described as having three stages: formulation screening, formulation qualification, and production.^a For *screening*, small-scale studies examine the influence of excipients, together with the physicochemical properties of active substance, to target the critical qualitative components and their interactions. In the next stage, *qualification*, effects of quantitative composition, and potential critical variables of process are studied. Finally, in *production*, scale-up and batch-to-batch variables are examined, with validated in-process and final product quality tests, which apply specification standards developed from the knowledge obtained in the previous two stages.

^aDahl, E. Quoted in McGilveray, I.J. *Drug Inf* **1996**, 30, 1029–1037.

Formulation Components

In most cases, these are of pharmacopeial quality, although those not described in official compendia, which may derive from processed food components, should be approved by reference to lists such as the FDA “GRAS” list—generally recognized as safe. The European, Japanese, and U.S. pharmacopeias are working on a harmonization process in parallel with ICH (ICH Q4)^[36] to gain some form of mutual acceptance of excipients as well as for tests and standards (reviewed later). Major excipients are among the monographs being examined in this process. It should be noted that several formulating agents, such as starch, are multifunctional in solid oral dosage forms.

Excipients/Fillers. The amount of filler used relates to the potency of the drug. For low- and submilligram-unit-dose drugs, care is required to avoid adsorbent fillers because the small amount of drug may not be readily released after adsorption on the large filler substance. Lactose and starch are common agents, and variations in their physical properties can be problematic. For example, with salicylic acid, the type of starch used influences dissolution.^[37] Also, with sparingly soluble drugs, the solubility of the excipient influences the dissolution of the active substance. This was shown by the comparison of the dissolution rate of griseofulvin, which was much increased when a soluble carrier, rather than a hydrophobic carrier, was used.^[38]

Disintegrating agents. These substances are added to solid oral dosage forms to cause disintegration of the contents when ingested and exposed to aqueous gastrointestinal fluids. Common agents are processed carboxymethylcelluloses, starch (raw or processed, e.g., sodium glycolate), and Veegum.

Granulating agents and binders. These agents are added to a powdered drug to retain the integrity of the tablet or capsule contents and also to facilitate manufacturing. However, they can also affect dissolution and thereby absorption. Carboxymethylcellulose, gelatins, gums (such as acacia and tragacanth), polyethylene glycols, polyvinyl pyrrolidone (PVP), starch, and Veegum are commonly used agents. Banakar^[23] has reviewed the choice of such substances.

Lubricants. These are processing aids included to improve the flow of bulk granulations and to avoid their adhesion to tablet manufacturing parts. The amount and type of mixing can affect the dissolution properties and absorption of certain drugs because they are predominately hydrophobic compounds such as stearates and talc. Their presence, concentration, and blending with other components have been described^[39] as among the most important factors influencing dissolution/absorption because the substances can form a water-repellent coat

around each granule. More recent work^[40] has confirmed that the amount of magnesium stearate and the time and type of mixing (shear) can be important.

Surfactants (Wetting agents). Surface-active agents such as sodium lauryl (dodecyl) sulfate and polysorbate 80 and hydrophilic polymers such as PVP and Hydroxy Propyl Methyl Cellulose (HPMC) are used to improve the dissolution of sparingly soluble drugs in tablet or capsule formulations. Both their properties and manner of addition can influence the resulting dosage form. The effect on solubilization is concentration-dependent.^[41] Below the critical micelle concentration, drug absorption is enhanced; but above this concentration, a portion of the drug associated with the micelle is unavailable for absorption. These surfactants also influence membrane permeability, and higher concentrations should be used with caution because of potential toxic effects.^[29]

In ophthalmic dosage forms, water-soluble polymers, such as polyvinyl alcohol and hydroxypropylcellulose, have been used at low concentrations to prolong the residence times of the solutions on the eye surface and to provide low-viscosity products that do not interfere with vision.^[42]

Ointments and skin preparations. Preparation of topical skin preparation combines the art of formulation with physicochemical knowledge of the components. Barry^[43] provides an interesting protocol for optimizing a dermatologic formulation. The key question is whether or not percutaneous penetration is desired. The diffusant, vehicle, and state of skin treated and how these factors interact “influence” the penetration. In another monograph,^[44] the tremendous variety of formulation types is considered, including liquid soaks, liniments, lotions, gels, creams, and ointments. These are composed of several bases, including aqueous, powder, and oily components, with thickening agents, emulsifying agents, buffers, antioxidants, and preservatives. Among the ointment bases are hydrocarbon bases (petrolatum); soft and hard paraffins; fats and fixed oils, including mono-, di-, and triglycerides (such as from olive, sesame, and peanut oils); silicones; water-soluble carbonates, and absorption bases such as lanolin, cholesterol, sorbitum monostearate, and wool alcohols. These can be mixed to minimize or optimize skin penetration with particular drugs, and biopharmaceutical tests have been difficult to devise.

For transdermal penetration, both skin patches and ointments have been applied. With many drugs, the limiting absorption-through-the-skin barrier does not provide sufficient dosing, but it has been excellent when low, constant dosing has been required, e.g., estrogen, nicotine, and scopolamine.^[45] Some transdermal systems, as well as ointments, depend on the native rate-limiting skin permeability, and others include rate-limiting membranes. It is easier to devise biopharmaceutical tests for the transdermal systems than for the topical dermatologic agents because

with the transcutaneous absorption, the pharmacokinetics of the drug can be studied in the systemic circulation.

Formulation of inhalants. As with skin preparations, the formulation of inhalation delivery systems is complex.^[30] These include nebulizers, metered-dose, and drug powder inhalers, each providing their own challenge to the developers. The biopharmaceutical evaluation of metered-dose inhalers has been extensively researched in the last decade. Pharmacokinetic valuation is limited because less than 20% of the delivered dose enters through the airways, the residue being swallowed.

Atkins, Barker, and Mathisen^[30] examined the four separate components of metered-dose inhalers: (a) formulation (drug, excipients); (b) container; (c) metering device; and (d) actuator mouthpiece. There are two formulation types, solution and suspension, and both contain propellant. Solutions often contain cosolvents (usually ethanol).

Traditionally, chlorofluorocarbons were used as propellants. However, with the international banning of these ozone-destroying agents, there has to be a reformulation with more ozone-friendly hydroxy- and hydroxy-chlorofluoroalkanes. These have different compatibilities with the surfactant-dispersing agents, and their introduction to formulations has been a challenge after the toxicology screening was complete.

Manufacturing Process

The workshops on scale-up and postapproval changes (SUPAC)^[6,46] indicate that critical manufacturing variables should be researched during the formulation development. Also, the FDA project with the University of Maryland suggested a framework to map out such process variables; the approach is outlined by Shah.^[47] This explains that the critical manufacturing variables include manufacturing, formulation, processing, and equipment that can significantly affect drug release from the product. Examples of formulation variables were noted in “Formulation Components” and include particle size and ranges of magnesium stearate. Equipment differences, such as mixing shear, can be critical, and process variables include compression forces, impeller speed, lubricant blend time, and indeed, order and time of addition of ingredients in the “recipe.” In the complex system of the dosage-form manufacture, multifunctional influences affect dissolution and bioavailability, and knowledge of the key variables and validated tests and standards to control them is vital for consistent quality manufacture. Lukas^[48] has provided a useful review of examples of the manufacturing variables that have been shown to influence dissolution (and sometimes absorption).

Granulation. Granulation usually improves the dissolution performance of sparingly soluble drugs. Wet and dry granulation procedures are both applied. Although

wet granulation has disadvantages, particularly “clumping” for hydrolytically labile drugs, advances in equipment have helped to speed the process and minimize exposure times. Water content can be critical in this process, as an example with naproxen demonstrates.^[49] Less-granulating water exhibited faster dissolution, attributed to smaller granules and less residual moisture. The order of addition of microcrystalline cellulose for granulation was shown to influence disintegration and dissolution.^[40]

Compression. In general, it might be expected that harder tablets from stronger compression would tend to decrease dissolution rates. However, in fact, more complex relationships are described. In compression of powders with different particle sizes, there was an optimum size when bonding and separation of the particles were in equilibrium within a tablet.^[50] The manner of packing of contents also affects capsules, and the modern capsule machines also involve compression. An advice from Abdou^[29] is that a full dissolution/compression study for each granulated formulation is essential to determine the compression parameters that will ensure good dissolution characteristics and adequate bioavailability.

Mixing. Rekhi et al.^[40] and Bergman and Bandelin^[39] have indicated that the amount of lubricant, magnesium stearate, is a key variable in formulation. The type of mixing was shown^[51] to influence this variable in developing cefadroxil capsules. During scale-up, with a constant 1% magnesium stearate composition, a change of type of the capsule-filling machine resulted in a decrease of dissolution. This was found to occur, with the different machine, by a repeated shear in the powder tamping process, leading to the greater coating of the water-repellent lubricant on the drug particles.

Dosage Forms, Compendial Monograph Tests, and Standards

Because “Dosage Forms” is the topic of a separate entry, few details or definitions are provided here. They are defined (with some variations) in the pharmacopeias. Compendial monographs (European, Japanese, and U.S. pharmacopeias) are public standards using validated tests and standards. While applications for new drug products contain proprietary information, the ICH process encourages the use of tests and standards described in these pharmacopeias. This is particularly desired in specifications, as noted in the ICH quality guideline 6 (Q6A and B).^[27]

Dosage Forms

Solutions. These are used primarily in parenteral presentations. In general, the aqueous solutions are considered to have few biopharmaceutical risks. Stability is a key parameter. Unless they have different components, generic

parenteral solutions are waived from in vivo bioequivalence comparisons. Aqueous oral solutions are prescribed for children and the very old. Oily solutions are more subject to biopharmaceutical problems. Cyclosporin provides an example because its formulation in oily solution resulted in erratic absorption in patients.^[52]

Suspensions. Suspensions, which may be for oral or intramuscular administration, are less straightforward in absorption properties than are solutions. Oral suspensions are common as pediatric dosage forms, and these have demonstrated bioavailability problems, as found with phenytoin.^[53] Micronization of the drug resulted in a 50% increase in bioavailability. As well as particle size, viscosity and the use of suspending agents are important quality attributes of suspensions. Although dissolution tests for suspensions have been investigated, they are not yet official.

Capsules. Although less studied than tablets, differences in capsule formulations provided several classic examples of bioequivalence problems. The “Australian phenytoin incident” is one that resulted in disturbing clinical consequences.^[54] This occurred when the filler was changed from calcium sulfate to lactose, yielding higher plasma concentrations of a narrow-therapeutic-range drug that resulted in toxicity to many patients.

There are two main types of capsules: hard gelatin and soft gelatin (soft elastic). More research has examined the properties of hard capsules, which consists of a shell encapsulating powdered or granulated ingredient, which is often precompressed. The release properties are important because the gelatin itself dissolves rapidly (although there have been cross-linking problems^[55]).

Soft gel capsules have been useful as a delivery vehicle for drugs with solubility problems, such as lipophilic compounds. They can be dissolved in some cases in aqueous media with the aid of surfactants, but more usually in polyethylene glycol or oils.^b Formulation is critical because drugs can crystallize, and this has been a very recent problem with ritonavir.^c In drug release, the rupture or burst time of the capsule is the critical point for the dissolution of contents.

Tablets. As the most familiar and convenient of the dosage forms used by the consumer, the often complex formulation of tablets is not evident. However, it is possibly the most studied type of formulation, and there are many examples of bioavailability problems often detected by dissolution. Compression and disintegration are the major determinants of drug release from tablets, and they depend on the eventual release of smaller particles for dissolution.

In the past, sugar coating interfered with disintegration and dissolution; but, presently, coating is rarely problematic, except with some enteric-coated products.

Nonoral dosage. Reference has already been made to the formulation of parenteral, topical, and percutaneous skin and inhalation dosage forms. These range from simple solutions with little biopharmaceutical risk to highly complex drug-device systems for transdermal or metered-dose inhalation. Thorough biopharmaceutical evaluation is therefore an important part of product development.

Compendial Monograph Tests and Standards

As has been noted elsewhere, new drug applications involve setting standards and specifications that are acceptable to regulatory agencies but proprietary to the applicant. However, the ICH process in guideline Q6A^[27] is intended to “assist, to the extent possible, in the establishment of a single set of global specifications for new drug substances and new drug products.” The importance of pharmacopeial methods is promoted as follows:

Pharmacopeial Tests and Acceptance Criteria: References to certain methods are found in pharmacopoeias in each region. Wherever they are appropriate, pharmacopeial methods should be utilized. Whereas differences in pharmacopeial methods and/or acceptance criteria have existed among the regions, as harmonized specification is possible only if the methods and acceptance criteria defined are acceptable to regulatory authorities in all regions, this guideline is dependent on the successful completion of harmonisation of pharmacopeial methods for several attributes commonly considered in the specification for new drug substances or new drug products.

Among the tests that are “essentially harmonized” are dissolution and disintegration apparatuses, and there are efforts to agree on the media and acceptance criteria for those tests. However, it has been noted that the goal of 100% harmonization “would be unrealistic and unattainable.”^[56]

Although, clearly, quality control tests on finished product, as well as in-process controls, have bearing on the biopharmaceutical properties, e.g., content uniformity, the primary in vitro batch tests for biopharmaceutical quality are dissolution and disintegration and, when appropriate, particle size. Dissolution and disintegration have been applied to nonoral as well as oral tablet, capsule, and suspension formulations. Thus the British Pharmacopoeia (BP) has disintegration tests for suppositories and pessaries,^[57] and the USP has dissolution tests for transdermal patches^[58] and indomethacin suppositories.^[58]

Content uniformity. As with most pharmacopeial tests, there is no statistically designed sampling plan for the

^bMorton, F. Quoted by McGilveray, I.J. *Drug Inf J* **1996**, 30, 1029–1037.

^cAbbot Labs. Quoted in *Wall St. Journal*, July 28, **1998**, p. B5.

content uniformity test. Production lot units are to be tested, and the sample size is the same for large or small production runs. Good manufacturing practice expectations and shelf life derived from stability investigations will lead to increased sampling.

The USP has a statement on the lack of sampling plan and statistical advice with their standards. It is noted that “repeats, replicates, statistical rejection of outliers or extrapolation of results to larger populations are neither specified nor proscribed by the compendia.”^[58] However, with its legal status in the United States and Canada, any specimen tested as directed in a monograph, at any time before its date of expiry, shall comply. The tests are somewhat arbitrary in the use of statistics. As an example, the content uniformity tests for some articles, but not others, use relative standard deviations. On the other hand, some dissolution tests have tolerances based on single-unit test results, whereas others have pooled samples.

The USP Content Uniformity tests^[59] are to be applied to products containing 50 mg or less of active ingredient. For amounts above 50 mg, comprising 50% or more by weight of the dosage units, weight variation tests are substituted. For the content uniformity test, 30 units are chosen for testing, from which 10 individual units are first measured by appropriate assay. For tablets, suppositories, and suspensions, the individual unit amounts (% label claim) are obtained and the mean and standard deviation are calculated. If 10 units all lie within the range of 85–115% and the relative standard deviation is 6% or less, the batch is accepted. If 1 unit is outside the 85–115% range but is within 75–125%, then the further 20 units are tested. The second stage requires that no more than 1 unit in the 30 be outside the range 85–115% but not 75–125%. The relative standard deviation for the 30 shall be not more than 7.8%. (The BP test has the same 1 in 30 tolerance but no relative standard deviation.^[57])

For capsules, transdermal systems, and inhalants, the tolerance in the USP^[59] is met at stage 1 if 9 of the 10 units are within 85–115% and no unit is outside 75–125%, then stage 2 requires testing of the remaining 20 units. The stage 2 tolerance is that not more than 3 of the 30 units be outside the range of 85–115% and no unit outside 75–125%. The same relative standard deviation limits prevail as for tablets.

For metered-dose inhalers^[59] in the first stage, 10 units are allowed a wider tolerance of 75–125% and 1 unit up to the range 65–135%. In that case, if 2 or 3 units are outside the 75–125% range, a further 20 units are assayed and not more than 3 of the 30 are outside the 75–125% range; in addition, no unit is outside the 65–135% label claim. There are no relative standard deviation limits provided for this dosage form.

Particle size. This is an active substance test applied to relatively few drugs of limited aqueous solubility when

it has been determined to influence absorption from the final dosage form of the product. Griseofulvin is the classic example (USP). The decision tree II 3 in the ICH Q6A^[27] document is useful for this test with the first question: Is the particle size critical to dissolution solubility or bioavailability? If the answer is yes, criteria have to be developed. Unfortunately, different tests and apparatuses for specific surface area or particle size distribution are not directly comparable. The acceptance criteria are set based on the observed range of variation of three or more lots, taking into account the potential for particle growth. Acceptance criteria will include acceptable particle size distribution in terms of the percent of total particles in given size ranges. The mean, upper, and/or lower particle size limits should be well defined.

Aerosol droplet size and distribution. This is a variation of the drug powder, active substance test that defines the particle in its broadest sense^[30] to include primary drug particles and multiparticulate aggregates. In most of the USP aerosol monographs, particle size is measured, after some priming shots, by actuating the aerosol onto a microscope slide. Particles are examined under standard magnification and field conditions to ensure that most measure less than 5 μm long and to limit those over 10 μm . The USP^[58] also describes two single-stage impactor and one multistage impactor devices for measuring particle size distribution. These tests for respirable volume or dose are being introduced for better quality control of these devices. For salbutamol (albuterol in the United States) pressurized inhalation, the BP^[57] applies a twin-impinger device (one of the single-stage devices in the USP). The number of replicates is not stated, but not less than 35% of the average amount delivered per actuation is deposited in the lower impingement chamber.^[57] The European Pharmacopeia (EP) adopted the BP twin impinger as a general test for aerosols,^[60] and an adaptation has been proposed for nasal sprays.^[61]

Disintegration. The relationship between disintegration, which is a visual test, and dissolution, in which the amount of drug is measured, is complex. Since the 1970s, the USP has attempted to apply dissolution for as many as possible of its solid oral dosage form monographs and has expanded this to many nonoral formulations. In Europe, the issue of dissolution, as discussed by the BP,^[62] is limited to products for which bioavailability problems are more likely. This has led to ICH guideline Q6A, which states that “dissolution testing for immediate-release solid oral dosage products of very water-soluble drug substances may be replaced by disintegration testing.”^[27] It is apparent by comparing the USP and the BP that the former has many fewer monographs, which include a disintegration test, because generally it has been replaced by dissolution.

The disintegration apparatus for tablets and capsules is virtually identical in the United States,^[58] European,^[63]

and Japanese^[64] pharmacopeias. It is called the *basket rack assembly*, with six cells (tubes) held in place with two circular plastic plates and a wire mesh attached to the lower plate to retain particles above a certain size. For some tests, plastic disks are inserted such as pistons in the tubes. Water is the usual immersion fluid, but buffers are used in some monographs.

The set time for passing the visual endpoint test of complete disintegration varies with the monograph; in general, 30 min or less for uncoated tablets, sometimes longer for coated tablets, and with an “acid-resistance” time test for delayed release, enteric-coated tablets.

In stage 1 of the test (USP), six tablets are tested. If these are all disintegrated in the time stated in the monograph, the lot passes (again, there is no given sampling plan). If up to two tablets fail to disintegrate, then (stage 2) a further 12 tablets are tested and the lot passes only if no more than 2 of the 18 total fail to disintegrate completely (an 11% allowance for failure).

The test for buccal tablets allows 4 hr, with the same tolerance as just stated for tablets. For sublingual tablets, each monograph gives a time, but the acceptance criteria are as just stated for tablets.

For delayed-release (enteric-coated) tablets, there is a survival test for 1 hr in simulated gastric fluid. There is zero tolerance for failure among the six tablets, which should show no evidence of disintegration, cracking, or softening. The second part of the test is with simulated intestinal fluid buffer, when, according to the time in the individual monograph, the same acceptance criteria (all 6 in the first stage or 16 of 18 in the second stage) prevail as for tablets.

For capsules, whether hard or soft gelatin, there is a modification of the apparatus, with a removable wire mesh cloth being attached to the top (as well as the bottom) circular plates. The tolerances for all 6 to pass in stage 1 or 16–18 in stage 2 are the same as for tablets, according to the times in the individual monograph.

The BP describes a different apparatus used in Europe for suppositories and pessaries.^[63] Although the usual time is 30 min, for indomethacin, three suppositories are tested and weighed at 90 min when more than 75% must be dissolved.

Dissolution. This may be regarded as the most sensitive of the in vitro biopharmaceutical tests. It is regarded by the USP as a useful surrogate for in vivo bioequivalence.^[58] However, the FDA scale-up and postapproval changes guidelines^[14,65] apply dissolution for only minor changes of formulation, process, or equipment. Its use for bioequivalence comparisons between products is controversial. Nonetheless, it is the primary lot-to-lot quality control test for biopharmaceutical properties. Also, it is applied universally during the formulation development.

The apparatuses are described in the USP^[58] and the EP.^[63] However, because the EP does not include

monographs on specific dosage forms, the BP is an example of a European compendium with specific dissolution standards.

Conventional/Immediate Release. First introduced in the USP during the 1970s, with a modification of the disintegration apparatus just described, two major apparatuses are applied for the dissolution of conventional tablets and capsules, including those with coatings.^[58] These are apparatus 1, called the *rotating basket* method, in which a removable cylindrical wire mesh base is attached to a spindle, and apparatus 2, the *paddle* apparatus, in which a spindle is attached to a paddle. Both rotate in water, buffer, or other medium (according to individual monographs) held at 37 ± 0.5 °C. The volume is usually 900 ml; the stirring rate, usually 50 rpm (paddle) or 100 rpm (basket), is given in the individual monographs. Six cells are usually used in one assembly.

The test procedure involves a sampling plan similar to that for the disintegration tests, with sequential acceptance stages. Six tablets or capsules are tested first; if necessary, up to a total of 24 units may be tested. The pharmacopeias do not impose any other sampling plan but expect the acceptance criteria to be met with any sample from a lot.^[58]

For immediate- (conventional-) release units, the percentage of dose (label claim) to be dissolved (Q) and the time are stated in the individual monographs. The USP, however, for some time has had the acceptance standards shown in Table 1.^[58] There is no request for relative standard deviation or any other statistical manipulation.

The USP has recently introduced the pooled-sample tolerance for some products, avoiding individual assays.^[59] There is again a three-stage testing. The first stage is complete if the average for all 6 units (pooled sample) is not less than $Q + 10\%$. If necessary, the second stage, with another 6 units added, requires the pooled average to be equal or be more than $Q + 5\%$. Finally, if necessary, a further 12 units are tested, and the pooled average of the total 24 units must be equal or be more than Q .

Many manufacturers plot profiles of percent dissolved vs. time and SUPAC.MR^[14] has suggested a formula for comparison of profile f_2 , the similarity factor. Polli, Singh Rekhi, and Shah^[66] have discussed other methods for comparing dissolution profiles. However, these have rarely been validated as discriminating between lots of acceptable and unacceptable bioequivalence, and also they are arbitrary.

For delayed-release (enteric-coated) dosage forms, an acid (gastric fluid) survival test is applied before the conventional-release test for which the acceptance criteria (Table 1) are applied.^[58] The other type of modified-release dosage forms, extended release, is described shortly.

Dissolution of Semisolid Dosage Forms (such as Creams, Gels, Lotions, and Ointments). This in vitro release test is suggested in the SUPAC guideline for non-sterile semisolid dosage forms.^[67] The apparatus includes

Table 1 USP Dissolution Test Acceptance Criteria

Stage	Number tested	Acceptance criteria
S_1	6	Each unit is not less than $Q + 5\%$.
S_2	6	Average of 12 units ($S_1 + S_2$) is equal to or greater than Q , and no unit is less than $Q - 15\%$.
S_3	12	Average of 24 units ($S_1 + S_2 + S_3$) is equal to or greater than Q , not more than 2 units are less than $Q - 15\%$, and no unit is less than $Q - 25\%$.

S_1 —first stage; S_2 —second stage; S_3 —third stage.

an open-chamber diffusion cell (Franz cell) fitted with a synthetic membrane. Diffusion of drug from the topical product to and across the standard area of the membrane is obtained by sequential assay of samples taken from the receptor fluid over the appropriate time intervals, usually to 6 hr. Typically, six cells are tested, and there is a random distribution of test and reference (in this case, the changed formulation vs. the original). Usually, the drugs in the aliquots are quantitated by high-pressure liquid or gas-liquid chromatography.

Because the release of drug from a standard surface is proportional to the square root of time ($\sqrt{t}$), the results are plotted as the amount of drug measured per unit area ($\mu\text{g}/\text{cm}^2$) vs. $\sqrt{t}$, yielding a straight line. The slope of this line reflects the release rate and is formulation-specific. Changes of the slope can be used to monitor between-batch or between-formulation effects. Slopes should involve at least five sample times, and six replicates of each test (and reference) batch are recommended.

The results are obtained in a two-stage manner. In the first stage, six slopes each of test and reference release rates are obtained and a 90% confidence interval (90% CI) for the ratio of the median in vitro release for test and reference is computed. If this 90% CI falls within the limits of 75–133.33%, no further testing is necessary. If this tolerance is not met, 12 additional slopes are obtained for each product and the 90% CI is computed for the total of 18 slopes of each product (original 6 + 12 repeats). Computation from all results should yield a 90% CI of 75–133.33%. It is emphasized in the guideline^[67] that the in vitro release test is useful, with the following caveat:

However, at this time the evidence available for the in vitro: in vivo correlation of (the test) is not as convincing as that for in vitro dissolution as a surrogate for in vivo bioavailability of solid oral dosage forms. Its utility is limited to assess product “sameness” under certain limited scale-up and postapproval changes.

Dissolution of Transdermal Systems. Three apparatuses (5, 6, and 7) are described in the USP^[58] for measuring the release of drug for transdermal patches. They are also described in the EP.^[64] The general release standards suggested include three time points, in terms of the dosing interval D , expressed in hours.

Apparatus 5, “paddle over disk,” is a modification of the USP apparatus 2 (paddle) in which the patch is retained on the bottom of the vessel with a stainless steel disk.

Apparatus 6, “cylinder,” is a modification of USP apparatus 1 in which the rotating basket is replaced with a stainless steel cylinder stirring element, with the patch being retained within the cylinder.

In apparatus 7, “reciprocating disk,” a series of calibrated glass (or inert plastic) vials are set up in rows. The patch is attached to a sample holder that is then suspended from a vertically reciprocating shaker run at about 30 cycles/min. After a set time, the holder moves to the next row of calibrated vessels containing the dissolution medium. Samples are then obtained across time from the rows of calibrated vessels.

All the transdermal release media are kept at $32 \pm 0.5^\circ\text{C}$. The only patch described in the USP to which these release rate tests have been applied is nicotine, and it is product-specific, because all three apparatuses are applied. However, the standards differ for four products. One patch has stated limits of 35–75% in 1 hr, 55–95% in 2 hr, and not less than 73% in 4 hr. The other extreme is a patch set to deliver 31–87% in 2 hr, 62–191% in 12 hr, and 85–261% in 24 hr.

The tolerances are given, as with the immediate-release tablets and capsules, as a possible three-stage testing. Thus if six units are tested and all are within the stated ranges, no further testing is needed. If some units are outside the ranges, stage 2 is run, and of the total of 12 units, no individual value can be outside by more than 10%. If not more than 2 are outside the ranges, then a further 12 units are tested; not more than 2 of the total 24 units can be outside the stated ranges by more than 10%, and none can be outside the average of the stated range by more than 20%.

Release Tests for Extended-Release Articles. The two apparatuses, 1 and 2, just described in detail are applied to test the release from some extended-release products. However, usually the profile of the release of drug is obtained over the labeled dosage interval, usually 12 or 24 hr. A minimum of three time points is chosen to test the usual 6 (first stage) up to 24 (third stage) units. In general, an early time point test for not more than 20% of label to dissolve, a second (mid dosage interval) will check for 50–60% release, and a third sampling, toward the end of the dose interval, will test for 80% release. Tolerances are

established in the individual monographs for accepting or rejecting the profile.

In the USP, beyond apparatuses 1 and 2, two additional apparatuses are described for extended-release articles.^[58] These are the *reciprocating cylinder* and *flow-through* apparatuses. The four apparatuses may be used to develop in vitro/in vivo correlations, also described in the USP. A general tolerance table is provided for extended release, with the ranges being specified in each monograph.

BIOAVAILABILITY AND BIOEQUIVALENCE

Bioavailability and bioequivalence of drug products emerged as important attributes and have been applied to the regulation of medicines for about 25 yr. Bioequivalence was first used in the approval of generic products in Canada in the 1970s^[68] and was defined there in 1973.^[69] However, the publication in 1977^[2] of the U.S. regulations on bioavailability and bioequivalence was a landmark that led to abbreviated new drug applications, in which product efficacy is determined by bioequivalence to a reference listed product.

Definitions

Different authorities offer slight variations on the U.S. Federal Register definitions:^[2]

Bioavailability: The rate and extent to which the active drug ingredient or therapeutic ingredient is absorbed from a drug product and becomes available at the site of action.

Bioequivalent drug products: Pharmaceutically equivalent products that display comparable bioavailability when studied under similar experimental conditions.

Note: The rate and extent of absorption of the test drug do not show a significant difference from the rate and extent of absorption of the reference drug when administered at the same molar dose of the therapeutic ingredient under similar experimental conditions in either a single dose or multiple doses.

Bioavailability is a pharmacokinetic attribute of a drug that depends on absorption, distribution, metabolism, and excretion after administration. *Absolute bioavailability* refers to the comparison of oral with intravenous administration of a drug. Thus presystemic metabolism and first-pass metabolism can be distinguished from poor absorption properties. Drugs with poor absorption will show limited bioavailability. However, drugs that are completely absorbed can also demonstrate poor bioavailability because of the extensive metabolism of the parent. In the latter case, the oral dose shows only a fraction of the intravenous dose in plasma profiles.

Determination of Bioavailability and Bioequivalence

Determination of Bioavailability

Bioavailability is determined in humans after the administration of a drug by measuring the concentration of the active drug and its (major) metabolites in whole blood, plasma, or serum as a function of time after dosing. It can also be estimated by measurement of the amount of drug or its metabolites excreted in urine as a function of time. As just noted, absolute bioavailability is one such key study. Also, a study of the blood and urine profiles related to dose is required to examine dose proportionality (or disproportionality) or change in pharmacokinetic parameters with change in dose.^[1] Other bioavailability studies submitted in NDAs apply the bioequivalence model described next. These include comparison of the pharmaceutical alternatives, such as clinical trial capsule dose form of a drug with the to-be-marketed formulation, and, also, dosage-strength-equivalent studies to show that equivalent doses of different strengths deliver the same amount of drug (e.g., 3×100 mg vs. 1×300 mg tablets).^[1]

Determination of Bioequivalence

The 1992 amendments to the CFR^[70] provide the following in vivo and in vitro approaches, in descending order of accuracy, sensitivity, and reproducibility, for determining the bioequivalence of a drug product.

- 1a. An in vivo test in humans in which the concentration of the active ingredient or active moiety and its active metabolites, in whole blood, plasma, and serum, is measured as a function of time.
- 1b. An in vitro test that has been correlated with and is predictive of human in vivo bioavailability data.
- 1c. An in vivo test in animals that has been correlated with and is predictive of human bioavailability.
2. An in vivo test in humans in which the urinary excretion of the active moiety and its active metabolites are measured as a function of time.
3. An in vivo test in humans in which an appropriate acute pharmacologic effect of the active moiety and its active metabolites are measured as a function of time, if such an effect can be measured with sufficient accuracy, sensitivity, and reproducibility.
4. Appropriately designed comparative clinical trials, for the purposes of demonstrating bioequivalence.

In general, for drugs that are absorbed systematically, approach 1 is the expected study. With some products active against urinary tract infections, approach 2 might be acceptable. For inhalant dosage forms, such as bronchodilators, approach 3 has been applied. Clinical trials (approach 4) have been required for topical drug comparisons.

Design and Implementation of Bioequivalence Studies

Guidances on bioequivalence study design and implementation are available from the FDA on specific drugs. General guidelines are provided by Health Canada;^[71] these have also been described for the U.S. by Dighe and Adams.^[72]

The most common type of study for systematically available drugs is the randomized, single-dose crossover in healthy volunteers. Its major attributes include:

Crossover with a washout period.

A small number of healthy adults, usually 24–36 (12 or more in Canada).^[71]

Single doses of test and reference product.

Appropriate sampling of blood to measure: > 80% of total area under the time—plasma (blood) concentration curve (AUC).

Measures of AUC and peak plasma (blood) concentration ($C_{\max}$) examined by statistical procedures.

For cases *when urine is measured*, again, in healthy volunteers, samples are collected until 95% of the drug is eliminated based on half-life. Usually, urine is collected every 2 hr when a short half-life drug is administered, but this interval can lengthen to 6 hr if sampling extends to 2 days or more. The Canadian guideline^[71] indicates that > 40% of unchanged drug should be eliminated in urine. The amount of drug per sampling period and the cumulative amount of drug excreted are calculated. Also, the rate of excretion in urine is obtained.

For extended-release drugs and some drugs with non-proportional kinetics, *multiple-dose studies* are required:

Crossover.

Healthy volunteers (patients if necessary for ethical reasons).

Dosed to steady state, with a washout period between test and reference dosing.

Predose ($C_{\min}$) concentrations obtained daily.

At steady state, plasma concentrations are obtained to establish the profile of drug (and sometimes metabolites) over time from last dose.

Measures of AUC, $C_{\max}$, $C_{\min}$, and fluctuation are obtained and examined by statistical procedures.

For *pharmacodynamic evaluations of equivalence*, studies may be carried out in patients. An example is for albuterol.^[73] In the case of this bronchodilator, asthmatic patients with moderate disease were admitted to the study. Details of another pharmacodynamic bioequivalence test, the vasoconstrictor assay for topical corticosteroids, are given later.

Statistical Analysis of Bioequivalence Studies

The most comprehensive instructions for statistical analysis of bioequivalence studies are given in the Canadian

guidelines.^[71,74] However, Dighe and Adams^[72] also provide a full description of the U.S. FDA expectations.

Single-Dose Crossover Study

Analysis of variance (ANOVA) must include the randomization scheme for the design, where all subjects randomized into the study are included and identified by code, sequence, and dates of dosing periods for both test and reference.^[71]

The ANOVA should include the appropriate statistical tests of all effects in the model.

All the data for all subjects, including any “dropouts” with measured results, should be presented. Supplementary analysis may be carried out with excluded subjects or data points. Exclusion must be justified.

ANOVA should be carried out on $t_{\max}$ and λ (terminal slope) data and on the logarithmically (ln) transformed AUC_T and AUC_I and $C_{\max}$ data. (AUC_T is to be the last quantifiable point; AUC_I is to infinity).

Reported results should include means and CVs (across subjects) for each product.

The ANOVA containing source, degrees of freedom, sum of squares, mean square, F - and p -values, and the derived intrasubject and intersubject CVs.

Table 2 (from Ref. [71]) provides the ANOVA for the crossover design for ln (AUC_T) for a 16-subject study and derived CVs.

Table 3 provides the AUC ratio estimate, at a 90% confidence interval, showing how these were calculated. ($\bar{X}$ represents the geometric mean.)

Table 2 AUC_T (ng·hr/ml) Analysis—ANOVA for ln (AUC_T)

Source	df	SS	MS	F	PR > F
Seq	1	0.0535	0.0535	0.09	0.770
Subject (Seq)	14	8.4375	0.6027	8.26	< 0.001
Period	1	0.0241	0.0241	0.33	0.574
Form	1	0.1373	0.1373	1.88	0.192
Residual	14	1.0211	0.0729	—	—

Intrasubject CV = $100 \times (\text{MS Residual})^{0.5} / 2 = 100 \times (0.0729)^{0.5} / 2 = 27\%$.

Intersubject CV = $100 \times (\text{MS Subject (Seq)} - \text{MS Residual} / 2)^{0.5} / 2 = 100 \times (0.6027 - 0.0729 / 2)^{0.5} / 2 = 51\%$.

(From Ref. [71].)

Table 3 AUC_T (ng·hr/ml) Analysis—Calculations

Difference = Test $\bar{X}$ – Reference $\bar{X}$ = 5.39 – 5.52 = – 0.13
$SE_{\text{Difference}} = (2MS_{\text{Residual}}/n)^{0.5} = (2 \times 0.0729/16)^{0.5} = 0.0955$
AUC ratio = $100 \times e^{\text{Difference}} = 100 \times e^{(-0.13)} = 88\%$
90% Confidence limits:
Lower, Upper = $100 \times e^{(\text{Difference} \pm t_{0.05,14} \times SE_{\text{Difference}})}$
Lower = $100 \times e^{(-0.13 - 1.761 \times 0.0955)} = 74\%$
Upper = $100 \times e^{(0.13 - 1.761 \times 0.0955)} = 104\%$

The standards for bioequivalence in the U.S. require that the 90% confidence interval for test product vs. reference (innovator) product for AUC and $C_{\max}$ be within the limits of 0.8 and 1.25 (or 80% and 125%). In Canada, only the AUC is required to be within this interval (for uncomplicated drugs) and the $C_{\max}$ ratio only (no confidence interval) within 80–125%.

Multiple-Dose Studies

Again, the Canadian guideline for modified release^[74] provides details from multiple-dose comparisons. Similar to single-dose analysis, information on the ANOVA randomization and sequence is required. The difference is in the number of variables to be listed and tested:

Analysis of $t_{\max}$, λ , and fluctuation = $(C_{\max} - C_{\min})/(AUC_{\tau}) \times 100\%$ should be carried out on the raw scale, while calculations for AUC_{τ} (over dosage interval), AUC_{τ} , and AUC_{τ} , $C_{\max}$ should use the logarithmic (ln) scale. The 90% confidence interval about the mean AUC_{τ} , $C_{\min}$, $C_{\max}$, and fluctuation parameters should be provided.

Table 4 gives the ANOVA for the crossover design model for ln (AUC_{τ}). This analysis gives the appropriate intrasubject variance estimate, MS (Residual), for the calculation of the 90% confidence interval. Any significant effects in the model, other than Subject (Seq), should be investigated. The intrasubject and intersubject CVs should also be calculated.

The AUC_{τ} ratio estimate and its 90% confidence interval are derived in the calculations shown in Table 5 ($\bar{X}$ refers to geometric means). If this study had a balanced design (i.e., an equal number of subjects per sequence), the

Table 4 AUC_{τ} ($\mu\text{g}\cdot\text{hr}/\text{ml}$) Analysis—ANOVA for ln (AUC_{τ})

Source	df	SS	MS	F	PR > F
Seq	1	0.44521	0.44521	1.6849	0.21525
	14	3.69921	0.26423	12.4240	0.00001
	1	0.02422	0.02422	1.1389	0.30394
	1	0.00405	0.00405	0.1905	0.66917
	14	0.29775	0.02127	—	—

Intrasubject CV = $100 \times (\text{MS}_{\text{Residual}})^{0.5} = 100 \times (0.0213)^{0.5} = 15\%$.

Table 5 AUC_{τ} ($\mu\text{g}\cdot\text{hr}/\text{ml}$) Analysis—Calculations

Difference = Test $\bar{X}$ – Reference $\bar{X}$ = 8.0768 – 8.0992 = –0.0225
$SE_{\text{Difference}} = 0.0516$
AUC_{τ} ratio = $100 \times e^{\text{Difference}} = 100 \times e^{-0.0225} = 98\%$
90% Confidence limits:
Lower, Upper = $100 \times e^{(\text{Difference} \pm t_{0.025,14} \times SE_{\text{Difference}})}$
Lower = $100 \times e^{(-0.0225 - 1.761 \times 0.0516)} = 89\%$
Upper = $100 \times e^{(0.0225 + 1.761 \times 0.0516)} = 107\%$

difference would simply be the difference in the arithmetic means of the ln (AUC)s. Because of the fact that the study was not balanced, the least-squares mean estimate for each formulation is used to form this difference, together with the appropriate standard error.

The standards for bioequivalence (U.S.) are that AUC_{τ} , $C_{\max}$, and $C_{\min}$ 90% confidence intervals for test vs. reference (innovator) product be within the limits of 0.8 and 1.25 (80–125%). In Canada, for $C_{\max}$ and $C_{\min}$ only, ratios are required to be within 80–125% and the AUC within the 90% confidence limits.

Food-Effect Bioavailability and Bioequivalence Studies

The presence of food in the gastrointestinal tract about the time of drug administration can affect absorption and other bioavailability measures for many drugs and drug products. Most regulatory authorities require the food challenge information for any new extended-release product because the dose-release mechanism may be either enhanced or impaired by food. This is particularly true of mechanisms influenced by the change of pH in moving down the gut. Karim^[75] presented an excellent paper, providing several decision trees that discussed the importance of food effect studies early in drug development. The FDA-proposed guidance,^[76] in fact, describes the factors Karim discussed.

Both the U.S. and Canadian guidelines^[74] for extended-release dosage forms require single-dose studies of food effects for the approval of new products. Generally, a high-fat meal would be given in one leg of a study and the volunteers would be fasted in the other leg. In the case of a generic extended-release product, the study could be a four-way crossover of test and reference, both with and without food.

In the Canadian guidelines, there is no limit for bioavailability because clinical trials will be provided for a new drug or for a first-entry modified release. However, the food challenge is informational for labeling.

Also, in the case of “dose-dumping,” the product may not be accepted. In the Canadian guidelines for bioequivalence, the product is expected to meet the standards for single-dose studies with and without food, i.e., AUC_{τ} 90% confidence interval of test vs. reference within the range 80–125%, $C_{\max}$ mean ratio from 80% to 125%.

The U.S. draft guidelines of December 1997 examine the food effects in much more detail.^[76] An important section in these guidelines discusses when food effect studies (in bioavailability/bioequivalence) may not be important and covers some “diagnostic factors” to determine whether additional studies of food effect are important. For a drug, such factors include:

High solubility across the intended dose range.

High permeability across the intended dose range (e.g., oral absorption > 90%).

Minimal or no effect of inactive ingredients or absorption of the drug substance in the fasted state (i.e., equivalent profiles in vivo of solution or suspension and test formulation).

For a drug product, the factors include:

Rapidly dissolving drug product.

In vitro dissolution characteristics similar in different pH media and rotation speeds in USP apparatuses 1 and 2.

If all of these factors are positive, absence of a food effect study may be justified. If they are not all positive, then the sponsor must provide justification concerning why a food effect study is not important.

The general design for both BA and BE studies in the draft U.S. guidelines^[76] is the randomized, balanced, single-dose, two-treatment, two-period, and two-sequence crossover. The reference for the bioavailability study of a new drug at the Investigational New Drug stage would be solution or suspension, and this would be tested with and without food to show whether the drug substance is affected. In the preapproval phase, if the drug substance is affected greatly by food, the proposed marketing formulation should also be tested with and without food. For bioequivalence of a substantially changed formulation or a generic, the test formulation and the reference listed drug (or original formulation) should be tested in the presence of food.

The U.S.-proposed guidelines^[74] suggest that the decision documenting whether a food effect is absent or not be based on the ratio of means (population geometric means, based on log-transformed data) of fed and fasted treatments. It is absent if:

AUC, 90% CI fall within 80–125%.

$C_{\max}$, 90% CI fall within 70–143%.

A food effect is concluded when either or both the AUC and $C_{\max}$, 90% CI are found to be outside the given ranges, and the sponsor is asked to indicate the clinical relevance of the magnitude of the observed data.

The Canadian modified-release (MR) guidelines^[74] require only that the AUC in food effect studies be within 80–125% and that, as for fasted studies, the $C_{\max}$ ratio be within 80–125%.

It is of interest that both the BioInternational meetings in 1992^[77] and 1996^[78] had conference statements on the determination of food effects. In the latter, it was noted that the ethnic differences in food are less important in the determination of such effects than caloric content and protein, carbohydrate, and fat ratio (kCal distribution).

Narrow-Therapeutic-Range Drugs in Canada

For drugs designated to be of narrow therapeutic range, bioequivalence requirements are tightened in the Narrow

Therapeutic Range (NTR) guidelines. Thus 95% confidence intervals are imposed and, for both AUC and $C_{\max}$ in fasted and fed conditions, should have test vs. reference means (ln) with 95% CI within 80–125%.^[79]

Individual Bioequivalence

Currently, new guidelines for individual bioequivalence are under review.^[80,81] Studies will involve replicate design, i.e., two treatments of test and two of reference formulation.^[82] This will allow an estimate of intrasubject variation (subject-by-formulation interaction). It is evident that the statistics are complicated^[83] and will involve more assumptions than current regulatory models for bioequivalence. It may lead to improved standards for NTR drugs and a means to deal with highly variable drugs, especially if some scaling factor can be agreed upon.

Pharmacodynamic Measures of Bioequivalence

The U.S. CFR^[2] and EU guidelines on bioavailability and bioequivalence allow the use of pharmacodynamic (PD) comparisons to show the equivalence of products. Other uses of pharmacodynamics involve surrogate endpoints for efficacy, such as blood pressure reduction and CD4 cell counts, and also for investigating the relationship between pharmacokinetic and pharmacodynamic variables. This section will discuss the use of PD in bioequivalence. Examples of PD use in bioequivalence include dermatologic corticosteroids^[84] and albuterol metered-dose inhalers.^[73] The FDA guidelines on corticosteroids^[84] provide great detail on setup and interpretation of studies, and the albuterol studies were discussed at an FDA Advisory Committee meeting in August 1996, where some information was made available.^[73]

The corticosteroid guidelines provide information on the use of the Stoughton–McKenzie vasoconstrictor bioassay^[85] for the assessment of topical corticosteroids. The approach relies on the effect of corticosteroids, which produce vasoconstriction, resulting in blanching of the skin. This is presumed to relate to the amount of drug entering the skin.

The method involves the application of a corticosteroid skin preparation (ointment, cream, or lotion) to skin (forearm) of healthy human subjects, where it is left for 6–18 hr. The vasoconstriction (blanching) of the skin is then monitored by a trained, blinded observer who notes the degree of blanching using a visual rating scale (either 0–3 or 0–4) at the single time point after removal of the formulation (2 hr is common). Chromameter readings of the blanching may offer a more objective assessment than the human observer.^[84]

The key problem of pharmacodynamic studies is that the dose response reaches an asymptote, or effect saturation, and this may be within the range of the application

dose. It is therefore important to identify the linear dose response region. Also, some subjects are more sensitive in showing effects than others; that is, responders have to be differentiated from nonresponders. The FDA recognizes this, and therefore studies to validate the method and choose responders, “pilot studies,” define the pivotal study dose duration on the vasoconstriction assay:

$$E = E_0 + \frac{E_{\max} D}{E_{\max} + D}$$

where E = effect, E_0 = baseline effect, $E_{\max}$ = maximal effect, ED_{50} = half-maximal effect, and dose D is the dose at which the effect is half maximal. The pilot study seeks to establish ED_{50} , D , and D_2 , which are the doses approximately one half and two times ED_{50} , respectively. The values bracket ED_{50} , corresponding to approximately 33% and 67% of maximal response, and they are in the sensitive portion of the dose duration–response curve. Responders are those who provide visual scale readings above 1 on a scale of 0 to 3 or 4.

The data are obtained by chromameter or visual observation and are corrected for baseline value at the site; and, from the individual points on the area under the effect curve, Area Under Effect Curve (AUEC) is measured, such as from 0 to 24 hr. The model is fit by using all the observations of individual subjects to provide ED_{50} and $E_{\max}$ values pooled from the 12 subjects.

The dose-duration information, such as $ED_{50} = 2$ hr, $D_1 = 1$ hr, and $D_2 = 4$ hr, is taken from this pilot study (based on the reference product values) to apply to up to 60 volunteers in the pivotal study to compare test product with reference. As well as using the ED_{50} from the pilot study, within-study day replicates are applied to volunteers, and individual dose-duration responses are obtained. The D_1/D_2 ratios of individual AUEC values obtained are accepted only if they exceed 1.25; values below this designate the volunteer as a “nondetector.” Only “responders” and “detectors” are used in the comparisons of dosage forms.

The data treatment from the study includes only responders, but results from “nondetectors” must be reported.

The AUEC of, say, 0–24 hr is computed for each baseline-adjusted, untreated control, site-corrected dose duration.

“Detectors” are identified from individual subjects whose AUEC values for D_1 and D_2 are both negative (responder) and meet the dose-duration response criterion of 1.25, i.e.,

$$\frac{\text{AUEC at } D_2}{\text{AUEC at } D_1} \geq 1.25$$

where $\text{AUEC at } D_2 = 0.5(\text{AUEC at } D_{2 \text{ left arm}} + \text{AUEC at } D_{2 \text{ right arm}})$ and $\text{AUEC at } D_1 = 0.5(\text{AUEC at } D_{1 \text{ left arm}} + \text{AUEC at } D_{1 \text{ right arm}})$. Only subjects with complete data sets, i.e., duplicated values of D_1 and D_2 and quadruplicate values of test, reference, and untreated, should be included in the data analysis.

All data from all subjects, however, including nondetectors, should be reported (nonresponders would have been eliminated because this is a subject exclusion criterion).

The statistical analysis can be applied only to untransformed results because although AUEC values are usually negative, some are positive. The presence of positive and negative values prevents the use of conventional statistical transformations. Confidence intervals are calculated for Locke’s exact confidence interval method.

The 90% confidence interval is obtained from the ratio of the AUEC response from test to that of reference. Both of the AUEC figures are averages of four replicates. The example from the guidelines given next shows that only 7 of the 12 volunteers were found to be “detectors” (Table 6).

The calculation of the 90% confidence interval for the data set from 7 to 12 subjects from the FDA guidelines (Chapter 1046) is given here. The data used to calculate the confidence interval are the average AUEC values of “detectors” (evaluable subjects) only. The calculation of the confidence interval is facilitated by the calculation of the following intermediate quantities:

$$\bar{X}_T = \frac{1}{n} \sum_{i=1}^n X_{Ti} \quad \bar{X}_R = \frac{1}{n} \sum_{i=1}^n X_{Ri}$$

where n is the number of evaluable subjects, 7 in this example.

Table 6 Average AUEC Values for Subjects Meeting the Dose Duration–Response Criterion $D_2/D_1 \geq 1.25$

Subject	AUEC _(0–24) test product (average)	AUEC _(0–24) reference product (average)
2	– 48.52	– 22.20
3	– 38.99	– 18.65
4	– 7.62	– 22.42
7	0.98	– 10.96
9	– 32.05	– 37.40
11	– 26.18	– 26.73
12	– 11.62	– 12.56

$$\hat{\sigma}_{\text{TT}} = \frac{\sum_{i=1}^n (X_{\text{Ti}} - \bar{X}_{\text{T}})^2}{n-1} \quad \hat{\sigma}_{\text{RR}} = \frac{\sum_{i=1}^n (X_{\text{Ri}} - \bar{X}_{\text{R}})^2}{n-1}$$

$$\hat{\sigma}_{\text{TR}} = \frac{\sum_{i=1}^n (X_{\text{Ti}} - \bar{X}_{\text{T}})(X_{\text{Ri}} - \bar{X}_{\text{R}})}{n-1}$$

These are the sample means, sample variances, and sample covariance for the individual evaluable subject average AUEC data. For the example, these are

$$\bar{X}_{\text{T}} = -23.43 \quad \bar{X}_{\text{R}} = -21.56$$

$$\hat{\sigma}_{\text{TT}} = 323.13 \quad \hat{\sigma}_{\text{RR}} = 80.10 \quad \hat{\sigma}_{\text{TR}} = 78.83$$

We define t as the 95th percentile of the t -distribution for $n - 1$ degrees of freedom. For example, for $n = 7$, t (6 degrees of freedom) is 1.9432. Now we define

$$G = \frac{t^2 \hat{\sigma}_{\text{RR}}}{n \hat{X}_{\text{R}}^2}$$

$G < 1$ is required to have a proper confidence interval. If $G \geq 1$, the study does *not* meet the in vivo bioequivalence requirements. In the example, $G = 0.0930$.

Under the assumption that $G < 1$, we calculate

$$K = \left(\frac{\bar{X}_{\text{T}}}{\bar{X}_{\text{R}}} \right)^2 + \frac{\hat{\sigma}_{\text{TT}}}{\hat{\sigma}_{\text{RR}}} (1 - G) + \frac{\hat{\sigma}_{\text{TR}}}{\hat{\sigma}_{\text{RR}}} \left(G \frac{\hat{\sigma}_{\text{TR}}}{\hat{\sigma}_{\text{RR}}} - 2 \frac{\bar{X}_{\text{T}}}{\bar{X}_{\text{R}}} \right)$$

In the example, $K = 2.791$.

The confidence interval limits may now be calculated:

$$\frac{\left(\frac{\bar{X}_{\text{T}}}{\bar{X}_{\text{R}}} - G \frac{\hat{\sigma}_{\text{TR}}}{\hat{\sigma}_{\text{RR}}} \right) \mp \frac{t}{\bar{X}_{\text{R}}} \sqrt{\frac{\hat{\sigma}_{\text{RR}}}{n} K}}{1 - G}$$

In the example, 90% confidence interval limits are 53.6% and 165.9%, based on the data of 7 evaluable subjects.

The corticosteroid guidelines^[84] state that “the Office of Generic Drugs has not determined at this time (June 1995) the equivalence interval for bioequivalence. The Office requires that an equivalence interval wider than 80–125%, as a public standard, may be necessary.” In fact, this was recognized in the albuterol discussion of the Pharmaceutical Sciences Advisory Committee,^[73] which recommended 70–143% as an acceptable 90% CI for these particular pharmacodynamic response studies (bronchoprovocation).

There is another set of topical skin products guidelines^[86] that advocates dermatopharmacokinetic measurements in place of clinical trials. The statistical calculations are identical to those for AUC, C_{max} , etc. under the preceding immediate-release bioequivalence studies, as is the 90% confidence interval width of 80–125%.

DEVELOPMENT, SCALE-UP, AND POSTAPPROVAL CHANGES, AND IN VITRO (DISSOLUTION)/IN VIVO (PLASMA CONCENTRATION PROFILE) CORRELATIONS

Introduction

The preceding sections have documented, with some examples, the in vitro properties that may influence the biopharmaceutical performance of drugs. It has been noted that solubility, permeability, and rate of dissolution are key properties of the drug substance, and that the dissolution behavior is the key property of the solid dosage form in leading to consistency of in vivo performance as measured by bioavailability or bioequivalence studies. During new drug or new product development (either by changes by innovator or toward introduction of a generic), there will be attempts to minimize expensive in vivo studies with intensive in vitro evaluation. Possibly, some pilot developmental bioequivalence studies will be completed before choosing the pivotal batch for bioequivalence. Such information is helpful in designing specifications (Q6A)^[27] and can also assist in decisions postapproval.

With the SUPAC guidelines, the FDA has attempted to formalize decisions on postapproval changes. The Europeans give much less detail in their guidelines on variations,^[85] but they approach the same problems.

Scale-Up and Postapproval Changes Guidelines

Common Elements

The three SUPAC guidelines issued—immediate-release solid oral dosage forms (SUPAC-IR),^[14] modified-release solid oral dosage forms (SUPAC-MR),^[65] and nonsterile semisolid dosage forms (SUPAC-SS)^[67]—contain some common features as well as advice specific to the dosage-form type. Some common features are:

Currently relates only to scale-up and postapproval changes. (However, principles may be applied to other preapproval situations.)

Level of changes defined—subgroups:

Components and composition.

Site of manufacture.

Batch size (scale-up or scale-down).

Manufacturing equipment.

Manufacturing process.

Each level of change, as well as being defined, has sections listing:

Test documentation comprising:

Chemistry.

Dissolution.

Bioequivalence.

Filing documentation, which includes all stability information.

Changes in components and composition. Level 1. These are changes that are unlikely to have any detectable impact on formulation quality and performance.

Level 2. These are changes that could have a significant impact on formulation quality and performance.

Level 3. These are changes that are likely to have a significant impact on formulation quality and performance.

Changes in site of manufacture. Level 1. These changes consist of site changes within the same facility, but with no changes in:

Manufacturing, components, or composition scale-up.

Equipment.

Standard operating procedures.

Environmental conditions and controls.

Personnel.

Manufacturing batch records.

Level 2. These changes consist of site changes within a contiguous campus or between facilities in adjacent city blocks. There are no other changes—see Level 1.

Level 3. These changes consist of changing the manufacturing site to a different campus. Except for personnel, no other change should be made, as in Level 1.

Changes in batch size (scale-up/scale-down). Level 1 includes changes up to and including a factor of 10 times the size of the pilot/biobatch. However,

Equipment used is of the same design and operating principles.

It is manufactured in full compliance with good manufacturing practices.

Same standard operating procedures and controls are used in the original and changed batches.

Equipment changes. Level 1. These include changes from nonautomated or nonmechanical to automated or mechanical equipment and to alternative equipment of the same design and operating principle.

Level 2. These include changes in equipment to different design and operating principles.

There is an equipment addendum for SUPAC-IR with more details.

Process changes. Level 1. These are changes such as rate of mixing, mixing times, operating speeds, and holding times within the approved application ranges.

Level 2. These are process changes as in Level 1 but outside approved application ranges.

Level 3. These are changes in the type of process used in the manufacture of the product, such as a change from wet granulation to direct compression of powder. Level 3 changes are not expected in the SUPAC-SS category.^[59]

Test documentation. For biopharmaceutical properties in SUPAC, the major decisions in testing concern which situations require bioequivalence studies. Thus the degree of change requires a hierarchy of dissolution studies; when the change is considered likely to have an impact on performance, a study of bioequivalence will be expected between the change batch and the original.

Level 1 changes for IR and MR products will not require additional dissolution studies, except for those in the approved application or in the USP for the particular article. In general, Level 3 changes demand a single-dose bioequivalence comparison of original and changed batches.

Dissolution tests in the IR document^[14] are described in three “cases” of increasing complexity, A, B, and C.

Case A. This uses USP apparatus 1 at 100 rpm, or apparatus 2 at 50 rpm, with 900 ml of 0.1N hydrochloric acid. The units should meet the Q value of 85% in 15 min according to the USP table of tolerances for dissolution.^[58] If changed and original batches meet this standard, it is acceptable.

Case B. This involves a multipoint dissolution profile in the apparatus (1 or 2) and medium in the application or the compendial conditions (if the article is listed), with samples taken at 15, 30, 45, 60, and 120 min or until an asymptote is reached. Profiles of changed and originals are compared using the f_2 equation (see the upcoming section on “Comparison of Profiles”).

Case C. Five separate dissolution profiles are obtained in water, 0.1N hydrochloric acid, and in USP buffer media at pH 4.5, 6.5, and 7.5, with sampling times at 15, 30, 45, 60, and 120 min or until an asymptote is reached. The original and changed batch profiles are compared with the f_2 statistic. If it can be justified, a surfactant can be added to the media to generate the profiles.

SUPAC-IR

The biopharmaceutical drug classification system^[12] is used in SUPAC-IR, and for Level 2 changes in composition for highly soluble, highly permeable drugs, the dissolution case A is applied. If this fails, cases B and C dissolution should be performed. For low-permeability drugs with high solubility, a multipoint dissolution (case B) should be applied. For low-solubility, high-permeability

drugs, case C should be applied. Comparisons should use the f_2 formula. For Level 3 changes in composition, case B dissolution is applied, but an in vivo bioequivalence study is also required unless there is an in vivo/in vitro correlation.

In SUPAC-IR, changes of site Level 3, batch size Level 2, and process Levels 2 and 3, the case B dissolution profile comparison is suggested. If the changes show similarity of dissolution, bioequivalence comparisons are required only for Level 3 changes in composition and process. As well as preapproval, it is likely that these principles of similarity and use of the biopharmaceutical drug classification system will extend to preapproval changes. Bioequivalence studies may be waived when in vitro/in vivo correlations have been established.

SUPAC-MR

The SUPAC-MR application of dissolution requires, for Level 1 change, that dissolution should be compared only by using the method in the approved application or in the USP (if the article is listed). For all Level 2 changes (composition, site, batch size, equipment, and process), extensive dissolution comparison is required. For extended release, for example, in addition to application or compendial release testing, the multipoint dissolution profiles should be obtained in three other media. These may include (but are not restricted to) water, 0.1N hydrochloric acid, and pH 4.5 and 6.8 buffer. Sampling times should be appropriate, such as 1, 2, and 4 hr and every 2 hr until either 80% of the drug is released or an asymptote is reached.

For delayed release (enteric-coated) as well as an acid survival (0.1N hydrochloric acid for 2 hr), multipoint dissolution profiles are to be obtained at three different agitation speeds in buffer at pH 4.5–7.5.

Bioequivalence studies are required in the absence of an in vitro/in vivo correlation for Level 3 changes in composition, site, and manufacturing process; this will be accepted with a single-dose study.

SUPAC-SS

For the SUPAC-SS category of product, dissolution is also considered useful, and has been described in "Dissolution." Levels 2 and 3 changes in composition or manufacturing site and Level 2 changes in equipment, process, or batch size require comparison of drug release slopes.

Comparison of Profiles

The comparison of dissolution is considered important not only for any impact on SUPAC decisions, but also for setting specifications as noted in both the FDA dissolution guidelines^[88] and in the ICH Q6A "specifications" document.^[27] The guidance on dissolution^[79] has a section on comparison of dissolution profiles that describes the similarity factor, f_2 ,

used in some decisions for SUPAC-IR and SUPAC-MR when profiles are being compared. This calculates the percentage difference between the two profiles at each time point and, with an error term, provides a measure of similarity in the percent dissolution. It is a logarithmic reciprocal, square root transformation of the sum of the squared error expressed as:

$$f_2 = 50 \log \left\{ \left[1 + \left(\frac{1}{n} \right) \sum_{t=1}^n (R_t - T_t)^2 \right]^{-0.5} \right\} \times 100$$

where n is the number of time points, R_t is the dissolution value of the reference (prechange batch in SUPAC) at time t , and T_t is the dissolution value of the test (change) batch at time t . Dissolution profiles of 12 units are obtained, such as with a biobatch (R) and a production lot (T), and the f_2 factor is calculated. Values for f_2 above 50 indicate a sameness of the two curves and suggest that the performance of the two batches will be similar.

The dissolution guidelines discuss other comparison procedures, including model-independent multivariate and model-dependent procedures. These have also been discussed by Polli, Singh Rekhi, and Shah.^[66]

In Vitro/In Vivo Correlations

Dissolution, as has already been described in several earlier sections, is a key biopharmaceutical test, and it is widely applied as a quality control test for finished products. However, its discriminating power is significantly enhanced if an in vitro/in vivo correlation can be established such that the dissolution behavior can predict bioavailability and discriminate between acceptable and inequivalent batches. For highly water-soluble drug classes (high-solubility, high-permeability and high-solubility, low-permeability) when the rate-determining step for absorption is not dissolution, correlations are unlikely. However, for low-solubility, high-permeability drugs, correlations are possible. Also, for extended-release drug products, in which the formulation is used to control the release of active drug, correlations are expected. This expectation is formalized in the USP (Chapter 1088)^[58] as well as the EU^[89] and FDA (SUPAC) modified-release guidelines. It also emerges in the ICH specifications document.^[27]

Levels of Correlation

Three correlation levels are defined, in decreasing order of usefulness, for predicting absorption behavior. The relationship in which the in vitro profile reflects the entire plasma drug-concentration profile resulting from the administration of the given drug product is most useful.^[58]

Level A correlation. This represents a point-to-point relationship between dissolution and the in vivo input rate

of the drug from the administered drug batch. The curves may be directly superimposable or depend on a constant offset value. There are different means of calculating the in vivo input rate curve, and, increasingly, mathematical deconvolution is being applied. The correlation is validated by in vitro and in vivo testing extremes of the formulation [slow, intermediate (intended for production), and fast dissolution test batches] with dissolution and from pilot bioavailability studies. A solution or immediate-release product is also given. This allows the differential determination of input curves from solution and sustained-release product to compare with the dissolution curve to provide a point-to-point correlation. The two extremes and the production lots can be used to set specifications and also in justifying postapproval changes. Thus if a change provides a dissolution profile similar to that of the original (within the validated boundaries), the predicted bioavailability will also be expected to be similar.

Level B correlation. This uses the statistical moment analysis. The mean in vitro dissolution time is compared with the mean residence time or the mean in vivo dissolution time calculated from time-plasma concentration profiles. However, because the *shape* of the in vivo curve is not defined from mean residence time, it is less reliable for setting specifications.

Level C correlation. Because this relates only one dissolution time point, such as time for 50% to be dissolved, to one bioavailability variable, such as AUC, $C_{\max}$, or $t_{\max}$, it is of limited value. It is a single-point correlation and does not define the complete shape of the plasma level profile, and it cannot be used in predicting in vivo performance.

CONCLUSIONS

The essential features of biopharmaceutical characterization and testing of drug substances and products during development have been described here. From the earliest point after discovery, pharmaceutical development has focused on providing the optimum dose in the optimum time to the patient in a consistent manner. Thus it is multidisciplinary, including the study of the physicochemical attributes of the substance and the likely dosage form, the pharmacokinetic behavior in animals and humans, and the relationships to pharmacologic and clinical effects. The regulatory authorities are required to assess the development as part of safety and efficacy monitoring and to ensure that valid quality tests and specifications are appropriate to provide consistent product performance. Whereas much research has characterized the biopharmaceutics of solid oral dosage forms, the future challenge is to investigate further other types of dosage forms to assist in the management of preapproval and postapproval changes.

REFERENCES

1. U.S. Food and Drug Administration, Center for Drugs and Biologics. *Guidance for the Format and Content of the Human Pharmacokinetics and Bioavailability Section of an Application*; February 1987.
2. Code of Federal Regulations, Title 21. CFR Part 320. In *Bioavailability and Bioequivalence Requirements*; U.S. Government, 1977.
3. Wagner, J.G.J. *Pharm. Sci.* **1961**, *50*, 359–387.
4. Gibaldi, M. *Biopharmaceutics and Clinical Pharmacokinetics*; Lea & Febiger: Malvern, MA, 1991, 40, preface.
5. Kaplan, S.A. *Current Concepts in the Pharmaceutical Sciences: Dosage Form Design and Bioavailability*; Swarbrick, J., Lea & Febiger: Philadelphia, PA, 1973, 1–31.
6. Skelly, J.P.; Van Buskirk, G.A.; Savello, D.R.; Amidon, G.L.; Arbit, H.M.; Dighe, S.; Fawzi, M.B.; Gonzalez, M.A.; Malic, A.W.; Malinowski, M.; Nedich, R.; Peck, G.E.; Pearce, D.M.; Shah, V.; Shangraw, R.F.; Schwartz, J.B.; Truelove, J. *Pharm. Res.* **1993**, *10*, 313–316.
7. Wood, J.; Syarto, J.; Letterman, H.J. *Pharm. Sci.* **1965**, *54*, 1068.
8. *Pharmacop. Forum* **1998**, *24*, 5617–5619.
9. Shore, P.A.; Brodie, B.B.; Hogben, C.A.M.J. *Pharm. Exp. Ther.* **1957**, *119*, 361–369.
10. Benet, L.A. *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 9th Ed.; Hardman, J.G.; Limbird, L.E., Eds.; McGraw Hill: New York, 1996; 6.
11. Kaplan, S.A.; Cotler, S.J. *Pharm. Sci.* **1972**, *61*, 1361–1364.
12. Amidon, G.L.; Lennernas, M.; Shah, V.P.; Crison, J.R. *Pharm. Res.* **1995**, *12*, 413–420.
13. Lennernas, H.; Ahrenstadt, Ö.; Hällgren, R.; Knutson, L.; Ryde, M.; Paalzow, L.K. *Pharm. Res.* **1992**, *9*, 1243–1251.
14. U.S. FDA. Center for Drug Evaluation and Research. *Guidance for Industry Scale-Up and Post Approval Changes Immediate Release Solid and Dosage Forms*; November 1995.
15. Hussain, A.S. *APPS/CRS/FDA Workshop on Scientific Foundation and Application for the Biopharmaceutical Classification System and In Vitro In Vivo Correlations*; 1997; Crystal City, VA.
16. Hirst, B.H. *Eur. J. Pharm. Sci.* **1997**, *5* (Suppl. 2), 56–57.
17. Lennernas, H. *J. Pharm. Sci.* **1998**, *87*, 403–410.
18. Atkinson, R.M.; Bedford, C.; Child, K.J.; Tomich, E.G. *Nature*, **1962**, *193*, 588, 589.
19. Finholt, P.; Kristiansen, H.; Schmidt, O.; Wold, K. *Medd. Nor. Farm. Selsk.* **1966**, *28*, 17–20.
20. Gonda, I.; Byron, P.R. *Drug Dev. Ind. Pharm.* **1978**, *4*, 243–250.
21. Shibata, J.; Kokubu, H.; Morimoto, L.; Morisaka, K.; Ishida, T.; Inoue, M. *J. Pharm. Sci.* **1983**, *72*, 1436–1438.
22. Ueda, H.; Nambu, N.; Nagai, T. *Chem. Pharm. Bull.* **1984**, *32*, 244–246.
23. Banakar, U. *Pharmaceutical Dissolution Testing*; Marcel Dekker: New York, 1992, 137, 149.
24. Poole, J. *Drug Inf. Bull.* **1968**, *3*, 8–16.
25. Meinsardi, H.; Van den Kleijn, E.; Van den Meijer, J.; Van Rees, H. *Epilepsia* **1975**, *16*, 353–358.
26. Food and Drug Administration, Center for Drug Evaluation and Research. *Policy Statement for the Development of New Stereoisomeric Drugs*; May 1992.

27. International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) In *Quality Guideline 6A. Specifications: Test Procedures and Acceptance Criteria for New Drug Substances and New Drug Products: Chemical Substances*; July 1997.
28. ICH Quality Guideline. *Q1 Stability. A. Stability Testing of New Drugs and Products*; 1993.
29. Abdou, H.M. *Dissolution, Bioavailability and Bioequivalence*; Mack: Easton, PA, 1989, 57, 82, 93.
30. Atkins, P.J.; Barker, N.P.; Mathisen, D. *Pharmaceutical Inhalation Aerosol Technology*; Hickey, A.J., Ed.; Marcel Dekker: New York, 1992, 157–170.
31. Anonymous. *Med. Lett.* **1981**, 23, 49.
32. Foltz, E.L. *Antimicrob. Agents Chemother.* **1970**, 14, 442–446.
33. Ulm, E.H.; Hichens, M.; Gomez, H.J.; Till, A.E.; Hand, E.; Vassil, T.C.; Biollaz, J.; Brunner, H.R.; Schelling, J.L. *Br. J. Clin. Pharmacol.* **1982**, 14, 357–362.
34. Jacobson, M.A.; Gallant, J.; Wang, L.H.; Coakley, D.; Weller, S.; Gary, D.; Squires, L.; Smiley, M.L.; Blum, M.R.; Feinberg, J. *Antimicrob. Agents Chemother.* **1994**, 38, 1534–1540.
35. Robinson, J. Overview of issues. Proceedings of workshop on in vitro dissolution of immediate-release dosage forms: Development of in vivo relevance and quality control issues. *Drug Inf. J.* **1996**, 30, 1030.
36. ICH, Quality Guideline 4 (Q4). In *Pharmacopoeial Harmonization*; 1997.
37. Underwood, T.; Cadwallader, D.E. *J. Pharm. Sci.* **1972**, 61, 239–243.
38. Westerberg, M.; Jonsson, B.; Nyström, C. *Int. J. Pharm.* **1986**, 28, 18–23.
39. Bergman, L.; Bandelin, F. *J. Pharm. Sci.* **1965**, 54, 445–449.
40. Rekhi, G.S.; Eddington, N.D.; Fossier, M.J.; Schwartz, P.; Lesko, L.J.; Augsburger, L.L. *Pharm. Dev. Technol.* **1997**, 2, 11–24.
41. Reigelman, S.; Crowell, W.J. *J. Am. Pharm. Assoc.* **1958**, 47, 115–122.
42. Robinson, J.E. *Rate Control and Drug Therapy*; Prescott, L.F.; Nimmo, W.S., Eds.; Churchill Livingstone: Edinburgh, 1985, 71–82.
43. Barry, B.W. *Percutaneous Absorption: Mechanisms, Methodology, Drug Delivery*; Bronaugh, R.L. Maibach, H.I, Eds.; Marcel Dekker: New York, 1985, 506–507.
44. Barry, B.W. *Dermatologic Formulations: Percutaneous Absorption*; Marcel Dekker: New York, 1983, 295–302.
45. Shaw, J.E.; Theeuwes, F. *Rate Control and Drug Therapy*; Prescott, L.F.; Nimmo, W.S., Eds.; Churchill Livingstone: Edinburgh, 1985, 65–70.
46. Skelly, J.P.; Van Buskirk, G.A.; Arbit, H.M.; Amidon, G.L.; Augsburger, L.L.; Barr, W.H.; Borge, S.; Clevenger, J.; Dighe, S.; Fawzi, M.; Fox, D.; Gonzalez, M.A.; Gray, V.A.; Hoiberg, C.; Leeson, L.J.; Lesko, L.; Malinowski, H.; Nixon, P.R.; Peace, D.M.; Peck, G.; Porter, S.; Robinson, J.; Savello, D.R.; Schwartz, P.; Schwartz, J.B.; Shah, V.P.; Shangraw, R.; Theeuwes, F. *Pharm. Res.* **1993**, 10, 1800–1805.
47. Shah, V.P. *Drug Inf. J.* **1996**, 30, 1085–1089.
48. Lukas, G. *Drug Inf. J.* **1996**, 30, 1091–1104.
49. Gordon, M.S. *Drug Dev. Ind. Pharm.* **1994**, 10, 11–29.
50. Carless, J.; Sheak, A. *J. Pharm. Pharmacol.* **1976**, 28, 17–18.
51. Ullah, I.; Wiley, G.I.; Agharkar, S.N. *Drug Dev. Ind. Pharm.* **1992**, 18, 895–910.
52. Johnston, A.; Marsden, J.T.; Hla, K.K.; Henry, J.A.; Holt, D.W. *Br. J. Clin. Pharmacol.* **1986**, 21, 331–333.
53. Neuvonen, P.J.; Pentikainen, P.J.; Elfving, S.M. *Int. J. Clin. Pharmacol. Biopharm.* **1977**, 15, 84–89.
54. Bochner, F. *J. Neurol. Sci.* **1972**, 16, 48–51.
55. Shiu, G.K. *Drug Inf. J.* **1996**, 30, 1045–1054.
56. Cecil, T.L.; Paul, W.L.; Grady, L.T. *Pharmacop. Forum* **1997**, 23, 3895–3902.
57. *Br. Pharmacop.* **1993**, pp. 159–160, 753, 1091, A159.
58. *U.S. Pharmacop.* **1995**, 23, pp. 9, 803, 1763–1767, 1790–1799, 1924–1929, Ivi.
59. *U.S. Pharmacop.* **1995**, 23 (Suppl. 7), 3894–3895, 3980, (1997).
60. *European Pharmacopoeia*; European Pharmacopoeia Secretariat, Council of Europe: Strasbourg, 1997, 137.
61. Aiache, J.M.; Beyssac, E. *Pharmeuropa* **1997**, 9, 561–565.
62. *British Pharmacopoeia*, 1993; 1995. A416–A419. Addendum.
63. *European Pharmacopoeia*, 3rd Ed.; European Pharmacopoeia Secretariat, Council of Europe: Strasbourg, 1997, 126, 127, 128–133.
64. *Japanese Pharmacopoeia*, 10th Ed.; (English 1991).
65. FDA, Center for Drug Evaluation and Research. *Guidance for Industry SUPAC-MR: Modified Release Solid Oral Dosage Forms*; 1997.
66. Polli, J.E.; Singh Rekhi, G.; Shah, V.P. *Drug Inf. J.* **1996**, 30, 1113–1120.
67. FDA Center for Drug Evaluation and Research. *Guidance for Industry. Non Sterile Semi Solid Dosage Forms. Scale-up and Post-Approval Changes*; 1997.
68. McGilveray, I.J. *Pharmaceutical Bioequivalence*; Welling, P.G.; Tse, F.L.S.; Dighe, S.V., Eds.; Marcel Dekker: New York, 1991, 382–383.
69. Morrison, A.B.; Cook, D.; Casselman, W.G.B. *Can. Med. Assoc. J.* **1973**, 109, 800–802.
70. *Federal Register 21 CFR Part 2*; April 28, 1992, 17981–18001.
71. Health Protection Branch, Drug Directorate Guidelines. *Conduct and Analysis of Bioavailability and Bioequivalence Studies. Part A: Oral Dosage Formulations used for Systemic Effects*; Health Canada: Ottawa, 1992.
72. Dighe, S.V.; Adams, W.P. *Pharmaceutical Bioequivalence*; Welling, P.G.; Tse, F.L.S.; Dighe, S.V., Eds.; Marcel Dekker: New York, 1991, 347–380.
73. *Proceedings of the FDA, Center for Drug Evaluation and Research, Pharmaceutical Advisory Committee*; August 11, 1996.
74. Health Canada, Drugs Programme Guidance. *Conduct and Analysis of Bioavailability and Bioequivalence Studies—Part B: Oral Modified Release Formulations*; Minister of Public Works and Government Services Canada: Ottawa, 1996, 35–130.
75. Karim, A. Bioavailability, Bioequivalence and Pharmacokinetic Studies. International Conference of the FIP: Bio-International '96, Midha, K.K., Nagai, T., Eds.; Business Center for Academic Societies Japan: Tokyo, 1996, 221–230.
76. Food and Drug Administration, Center for Drug Evaluation and Research. *Draft: Guidance for Industry.*

- Food-Effect Bioavailability and Bioequivalence Studies*; October 1997.
77. Conference Statement. In *Bioavailability, Bioequivalence and Pharmacokinetics*, Proceedings of BioInternational 1992; Midha, K.K., Blume, H.H. Medpharm: Bad Homburg; Germany, Eds.; 1993, 21–23, 209–251.
 78. Conference Statement. In *Bioavailability, Bioequivalence and Pharmacokinetic Studies*, Proceedings of FIP BioInternational 1996; Midha, K.K., Nagai, T., Eds.; Business Center for Academic Societies: Tokyo, 1996, 5–6.
 79. Standards for Comparative Bioavailability Studies Involving Drugs with a Narrow Therapeutic Range-oral Dosage Forms. In *Therapeutic Products Programme*; Health Canada: Ottawa, May 7, 1997. Draft.
 80. FDA, Center for Drug Evaluation and Research. *Draft Guidance for Industry. In Vivo Bioequivalence Studies Based on Population and Individual Bioequivalence Approaches*; October 1997.
 81. Williams, R. Background to the Workshop. In *Proceedings of AAPS Workshop on Scientific and Regulatory Issues in Product Quality. Narrow therapeutic Index Drugs/ Individual Bioequivalence*; 1998. Crystal City, VA.
 82. Ekbohm, G.; Melander, H. *Biometrics*; **1989**, 45, 1249–1254.
 83. Schall, R.; Luus, H.E. *Stat. Med.* **1993**, 12, 1109–1124.
 84. FDA, Center for Drug Evaluation and Research, Guidance for Industry. *Topical Dermatologic Steroids: In Vivo Equivalence*; June 2, 1995.
 85. Stoughton, R.B. *Topical Corticosteroids*; Maibach, H.I., Surber, C., Eds.; Karger: Basel, 1992, 42–53.
 86. FDA, Center for Drug Evaluation and Research. Draft Guidance for Industry. In *Topical Dermatologic Drug Product NDAs and ANDAs—In Vivo Bioavailability, Bioequivalence, In Vitro Release, and Associated Studies*; June 1988.
 87. European Economic Community Regulation no. 541/95 Provisions Relating to Variations in Marketing Authorization.
 88. FDA Center for Drug Evaluation and Research. Guidance for Industry. In *Dissolution Testing of Immediate Release Solid Oral Dosage Form*; August 1997.
 89. Commission of the European Communities. *Notes for Guidance. Quality of Prolonged Release and Transdermal Dosage Forms, Section II: Quality of Modified Release and Transdermal Dosage Forms*; 1998.

Biosimilar Studies: Sample Size

Seung-Ho Kang

Department of Applied Statistics, Yonsei University, Seoul, South Korea

Abstract

The two-arm equivalence parallel design has been widely utilized to assess the degree of biosimilarity between a biological product and a reference product. The aim is to show that the difference between two population means of a primary end point is less than a prespecified equivalence margin. A well-known sample size formula for the equivalence trial is expressed as follows:

$$n_1 = kn_2, n_2 = \frac{(z_{\alpha} + z_{\beta/2})^2 \sigma^2}{(\delta - |\mu_T - \mu_R|)^2} (1 + 1/k)$$

where n_1 and μ_T (n_2 and μ_R) are the sample size and the population mean of the primary end point of the biosimilar (reference) group, respectively. σ^2 , δ , and k are the population variance, the equivalence margin, and the sample allocation ratio, respectively. z_{α} is the upper alpha quantile of the standard normal distribution. It is known that this sample size formula is very conservative at times because the formula is derived based on the approximate power rather than the exact power. The exact power based on the sample size computed from the formula to have power of $1 - \beta$ is, in fact, $1 - \beta/2$ under certain conditions. Calculating the sample size precisely and numerically based on the exact power is recommended.

Keywords: Equivalence trial; Margin; Power.

INTRODUCTION

Due to the expiring patents of biological products, the development of generic versions of innovator biologic products has been the subject of increased attention in both biopharmaceutical industry and regulatory agencies.^[1] Sponsors should demonstrate high degrees of similarity in terms of the quality, efficacy, and safety between a biosimilar product and an innovator's biological product in order to obtain approval from the relevant regulatory agencies. In the development of a biosimilar product, a stepwise approach is usually recommended.^[2] A phase III comparative study is usually required in the last step of the stepwise approach, if there are uncertainties about the biosimilarity of the two products based on quality studies and non-clinical studies. Equivalence trials based on the two-arm parallel design are commonly used as phase III comparative studies during the development of biosimilar products.

One of the important issues when designing phase III comparative studies is the calculation of the sample

size. For economic, ethical, and scientific reasons, it is especially important not to overestimate sample sizes.

THE EQUIVALENCE TRIALS BASED ON THE TWO-ARM PARALLEL DESIGN

Although several measures and designs have been studied to evaluate the degree of biosimilarity between a biosimilar product and an innovator's biological product,^[3–7] the most widely used method involves a comparison of two population means based on the two-arm parallel design. Therefore, equivalence trials are frequently used in order to show that the difference between two population means of a primary end point is smaller than a prespecified equivalence margin. For example, the U.S. Food and Drug Administration draft guidance describes the importance of the equivalence trial, as given below:^[2]

A study employing a two-sided test in which the null hypothesis is that either (1) the proposed product is inferior

to the reference product or (2) the proposed product is superior to the reference product based on a pre-specified equivalence margin is the most straightforward study design for accomplishing this objective.

Similarly, the World Health Organization (WHO) guideline also emphasizes the importance of the equivalence trial, as follows:^[8]

Equivalence trials are strongly recommended for medicinal products with a narrow safety margin, such as insulin, to ensure that the SBP (biosimilar product) is not less and not more effective than the RBP (renovator biological product).

In equivalence trials based on the two-arm parallel design, patients are randomized into two groups—the patients in the first group receive a biosimilar product and the patients in the second group receive a reference product. The hypotheses of interest are given by:

$$H_0: |\mu_T - \mu_R| \geq \delta \text{ versus } H_a: |\mu_T - \mu_R| < \delta \quad (1)$$

where $\delta(>0)$ is a prespecified equivalence margin and μ_T and μ_R are the population mean of the primary end point in the biosimilar group and in the reference group, respectively.

THE EXACT AND APPROXIMATE POWER

Let $X_{T,i} (i=1, \dots, n_1)$ and $X_{R,i} (i=1, \dots, n_2)$ represent the continuous primary end points from the biosimilar product in the first group and the reference product in the second group, respectively, which independently follow a normal distribution with means of μ_T and μ_R and a common variance of σ^2 . The common variance σ^2 is assumed to be known for the calculation of the sample size. The hypotheses in Eq. 1 can be decomposed into two one-sided hypotheses, as follows.

$$H_{01}: \mu_T - \mu_R \leq -\delta \text{ versus } H_{a1}: \mu_T - \mu_R > -\delta \quad (2)$$

and

$$H_{02}: \mu_T - \mu_R \geq \delta \text{ versus } H_{a2}: \mu_T - \mu_R < \delta \quad (3)$$

If both H_{01} and H_{02} are rejected at the significance level of α , it can be concluded that H_0 in Eq. 1 is rejected at the significance level of α . Additionally, the two products are claimed to be biosimilar.

Both H_{01} and H_{02} are rejected at the significance level of α , if,

$$Z_L < -z_\alpha \text{ and } Z_U > z_\alpha \quad (4)$$

where

$$Z_L = \frac{\bar{X}_T - \bar{X}_R - \delta}{\sigma \sqrt{1/n_1 + 1/n_2}} \text{ and } Z_U = \frac{\bar{X}_T - \bar{X}_R + \delta}{\sigma \sqrt{1/n_1 + 1/n_2}} \quad (5)$$

where $\bar{X}_T$ and $\bar{X}_R$ are the sample means of the primary end point in the first and the second group, respectively.

Under the alternative hypothesis $H_a: |\mu_T - \mu_R| < \delta$, Kang and Kim^[9] showed that the power of this test is expressed as follows:

$$\begin{aligned} &P(Z_L < -z_\alpha \text{ and } Z_U > z_\alpha | H_a) \\ &= P(Z_L < -z_\alpha | H_a) + P(Z_U > z_\alpha | H_a) \\ &\quad - P(Z_L < -z_\alpha \text{ or } Z_U > z_\alpha | H_a) \end{aligned} \quad (6)$$

$$\geq 2\psi\left(\frac{\delta - |\mu_T - \mu_R|}{\sigma \sqrt{1/n_1 + 1/n_2}} - z_\alpha\right) - 1 \quad (7)$$

Here, ψ is the cumulative distribution function of the standard normal distribution. Note that Eq. 6 provides the exact power, whereas Eq. 7 gives the approximate power. If the sample size is calculated based on the approximate power with Eq. 7, we obtain the following well-known sample size calculation formula for the equivalence trials.^[10]

$$n_1 = kn_2 \text{ and } n_2 = \frac{(z_\alpha + z_{\beta/2})^2 \sigma^2}{(\delta - |\mu_T - \mu_R|)^2} (1 + 1/k) \quad (8)$$

where z_α is the upper alpha quantile of the standard normal distribution (e.g., $z_{0.05} = 1.645$), σ^2 is the population variance of the primary end point, k is the allocation ratio, and n_1 and n_2 are the corresponding sample sizes of the first group and the second group, respectively.

COMPARISON OF THE EXACT AND THE APPROXIMATE POWER

Kang and Kim^[9] compared the exact and the approximate powers based on Eqs. 6 and 7 with the following results.

1. The sample size calculation based on Eq. 8 is very conservative, necessitating larger sample sizes that are actually needed.
2. The exact power based on the sample size calculated by Eq. 8 for power of $1 - \beta$ is actually $1 - \beta/2$ under certain conditions.^[9] In other words, if the sample size is computed from Eq. 8 to have 80% power, the exact power with an identical sample size is, in fact, 90% in many practical cases.

The R code used to calculate the sample sizes based on both the exact and the approximate power level in each case is provided by Kang and Kim.^[9]

CONCLUSION

Although the well-known sample size calculation formula for the equivalence trials is simple, it has been found that the formula is very conservative. For example, in many practical cases, the sample sizes to achieve 80% power based on the approximate power actually produce 90% of the exact power. Therefore, calculating the sample size based on the exact power is recommended. Kang, Jung, and Baik^[11] compared the exact and the approximate powers in the binary end point and reached a similar conclusion.

ACKNOWLEDGMENT

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1D1A1A09916819).

REFERENCES

1. Chow, S.C.; Endrenyi, L.; Lachenbruch, P.A.; Yang, L.Y.; Chi, E. Scientific factors for assessing biosimilarity and drug interchangeability of follow-on biologics. *Biosimilars* **2011**, *1*, 13–26.
2. US FDA. *Guidance for Industry: Scientific Considerations in Demonstrating Biosimilarity to a Reference Product*; Center for Drug Evaluation and Research, U.S. Food and Drug Administration: Silver Spring, MD, 2012.
3. Chow, S.C.; Liu, J.P. Statistical assessment of biosimilar products. *J. Biopharm. Stat.* **2010a**, *20*, 10–30.
4. Chow, S.C.; Hsieh, T.C.; Chi, E.; Yang, J. A comparison of moment-based and probability-based criteria for assessment of follow-on biologics. *J. Biopharm. Stat.* **2010b**, *20*, 31–45.
5. Chow, S.C. *Biosimilars—Design and Analysis of Follow-on Biologics*; CRC Press: Boca Raton, FL, 2013.
6. Kang, S.H.; Chow, S.C. Statistical assessment of biosimilarity based on relative distance between follow-on biologics. *Stat. Med.* **2013**, *32* (3), 382–392.
7. Lei, L.; Olson, K. Evaluating statistical methods to establish clinical similarity of two biologics. *J. Biopharm. Stat.* **2010**, *20*, 62–74.
8. World Health Organization. *Guidelines on Evaluation of Similar Biotherapeutic Products* 2009.
9. Kang, S.H.; Kim, Y. Sample size calculations for the development of biosimilar products. *J. Biopharm. Stat.* **2014**, *24*, 1215–1224.
10. Chow, S.C.; Shao, J.; Wang, H. *Sample Size Calculations in Clinical Research*; CRC Press: Boca Raton, FL, 2003.
11. Kang, S.H.; Jung, H.Y.; Baik, S.H. Sample size calculations for the development of biosimilar products based on binary endpoints. *Commun. Stat. Appl. Methods* **2015**, *22* (4), 389–399.

Biosimilar Studies: Three-Arm Design

Seung-Ho Kang

Department of Applied Statistics, Yonsei University, Seoul, South Korea

Abstract

One of the most commonly used designs for assessing the degree of biosimilarity between a biosimilar product and a reference product is the two-arm parallel design, with which biosimilarity is assessed using the difference between two population means. A problem with this method is that the variability of a reference product among different batches cannot be taken into account accurately. This problem is overcome when using the three-arm parallel design, in which one arm is used for the biosimilar product and the other two arms are for the reference product. A relative distance is employed to evaluate the degree of biosimilarity. The distance between the biosimilar product and the reference product, defined as the absolute mean difference between the two products, becomes the numerator of the relative distance. The distance between the reference products from two different batches becomes the denominator of the relative distance. If the relative distance is less than a pre-specified margin, the two products are claimed to be biosimilar. Statistical tests based on a ratio estimator method were developed to assess the degree of biosimilarity when using the relative distance in the three-arm parallel design.

Keywords: Similarity; Biosimilar; Parallel design; Ratio estimator.

INTRODUCTION

As patents of biological products expire, evaluations of the degree of biosimilarity for these products to meet regulatory approval have received considerable attention. Many studies have been devoted to the search for a feasible biosimilarity criterion and its corresponding measurement technique.^[1–11] In practice, the two-arm-balanced parallel design has been widely utilized to assess the biosimilarity between a biological product and a reference product. In this design, biosimilarity is evaluated with an equivalence test whose hypotheses are given by:

$$\begin{aligned} H_0 : \mu_T - \mu_R \leq -\delta \text{ or } \mu_T - \mu_R \geq \delta \text{ versus} \\ H_A : -\delta < \mu_T - \mu_R < \delta \end{aligned} \quad (1)$$

where $\delta (\delta > 0)$ is a prespecified equivalence margin and μ_T and μ_R are the population means of the biological product and the reference product, respectively. Biosimilarity is assessed using the difference between the two population means. However, it is known that: 1) biosimilar products are generally very sensitive to environmental factors such as light and temperature differences; and 2) a small change at any critical stage of the manufacturing process could cause major changes in clinical outcomes. When the reference products come from two different manufacturing processes (locations) or from different batches from the same manufacturing process, the two-arm parallel design cannot feasibly incorporate the variability

of the reference products. Hence, when the equivalence margin in the two-arm parallel design is larger than the variability of the reference product, this method may arrive at the erroneous conclusion that the two products are biosimilar, although they are, in fact, considerably different. This problem can be resolved if the equivalence margin can be determined by considering the variability of the reference product before any clinical trial begins. However, in practice, this solution may not be feasible, as the variability of a reference product may not be known a priori. In order to solve this problem, the three-arm parallel design was proposed by Kang and Chow^[7] to assess the degree of biosimilarity between a biological product and a reference product.

THREE-ARM PARALLEL DESIGN

We consider three treatments: a biosimilar product (T) and two reference products from different batches (R_1 , R_2). In total, n patients are randomized into the following three groups. First, n_1 patients are assigned to the first group, to receive the biosimilar product T. The patients who are assigned to the second and third groups receive reference products R_1 and R_2 , respectively, with the number of patients in each group denoted by n_2 . The randomization ratio 2:1:1 is used with $n_1 = 2n_2$, and the total sample size is $N = n_1 + 2n_2$. Let Y denote a primary continuous response variable. If the two-arm parallel design is used, it

is possible to divide the reference arm randomly into two groups, R_1 and R_2 .

It is also necessary to define certain distances. Let $d(T, R)$ represents the distance between the biosimilar product T and the reference product R . Similarly, the distance between R_1 and R_2 is denoted by $d(R_1, R_2)$. There exist many possible candidates for specific forms of distance. Examples include:

$$d_1(T, R) = |\mu_T - \mu_R| \quad (2)$$

$$d_2(T, R) = |\mu_T / \mu_R| \quad (3)$$

$$d_3(T, R) = (\mu_T - \mu_R)^2 \quad (4)$$

$$d_4(T, R) = E(Y_T - Y_R)^2 \quad (5)$$

where μ_T is the population mean of Y for the patients with the biosimilar product T and μ_R is defined as $(\mu_{R_1} + \mu_{R_2})/2$, where μ_{R_1} and μ_{R_2} represent the population means of Y in patients who receive reference product R_1 and R_2 , respectively.

Kang and Chow^[7] considered what is known as the relative distance, as defined below, to evaluate the degree of biosimilarity between a biosimilar product and a reference product.

$$rd = \frac{d(T, R)}{d(R_1, R_2)} \quad (6)$$

Note that, because the distance values are non-negative, the relative distances are also non-negative. If the value of the relative distance rd is smaller than a prespecified margin δ ($\delta > 0$) in the three-arm parallel design, the two products are claimed to be a biosimilar. Therefore, the hypotheses for assessing biosimilarity are expressed as:

$$H_0 : rd \geq \delta \text{ versus } H_A : rd < \delta \quad (7)$$

Kang and Chow^[7] employed the absolute mean difference $d_1(T, R) = |\mu_T - \mu_R|$ to evaluate the biosimilarity between two products. We assume that $\mu_{R_1} \neq \mu_{R_2}$, as two different batches of the reference product cannot be completely identical. The relative distance is expressed as:

$$rd = \frac{|\mu_T - \mu_R|}{|\mu_{R_1} - \mu_{R_2}|} = \frac{|\mu_T - (\mu_{R_1} + \mu_{R_2})/2|}{|\mu_{R_1} - \mu_{R_2}|} \quad (8)$$

Hence, the hypotheses in Eq. 7 can be rewritten as:

$$H_0 : \theta \leq -\delta \text{ or } \theta \geq \delta \text{ versus } H_A : -\delta < \theta < \delta \quad (9)$$

where,

$$\theta = \frac{\mu_T - (\mu_{R_1} + \mu_{R_2})/2}{\mu_{R_1} - \mu_{R_2}} \quad (10)$$

The hypotheses in Eq. 9 can be decomposed into two one-sided hypotheses, as follows:

$$H_{01} : \theta \leq -\delta \text{ versus } H_{A1} : -\delta < \theta \quad (11)$$

and

$$H_{02} : \theta \geq \delta \text{ versus } H_{A2} : \theta < \delta \quad (12)$$

STATISTICAL TESTS FOR BIOSIMILARITY

Kang and Chow^[7] proposed two statistical tests in order to test the hypotheses in Eqs. 11 and 12. These were a test based on a ratio estimator and a linearization method. Here, we describe only the statistical test based on the ratio estimator because it has greater power than the linearization method.^[7]

Let $Y_{T,i}$ ($i = 1, 2, \dots, n_1$) and $Y_{R_{k,i}}$ ($k = 1, 2, i = 1, 2, \dots, n_2$) represent the response variables from the biosimilar product in the first group and the reference product in the second and third groups, respectively, with $n_1 = 2n_2$. We assume that $Y_{T,i}$ and $Y_{R_{k,i}}$ ($k = 1, 2$) follow independently normal distributions with means of μ_T and μ_{R_k} ($k = 1, 2$) and common variances of σ^2 . We also assume that the sample size is large enough such that the central limit theorem can be used. Kang and Chow^[7] constructed the following test statistic based on the ratio estimator.

$$\hat{\theta} = \frac{\bar{V}}{\bar{U}} = \frac{\bar{Y}_T - (\bar{Y}_{R_1} + \bar{Y}_{R_2})/2}{\bar{Y}_{R_1} - \bar{Y}_{R_2}} \quad (13)$$

where,

$$\bar{Y}_T = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{T,i} \quad (14)$$

$$\bar{Y}_{R_k} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_{R_{k,i}} \quad \text{for } k = 1, 2 \quad (15)$$

They also showed the asymptotic normality of $\sqrt{n_1}(\hat{\theta} - \theta)$.

$$\sqrt{n_1}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{2\sigma^2}{\mu^2} + 4 \frac{v^2}{\mu^4} \sigma^2\right) \quad (16)$$

where,

$$\mu = \mu_{R_1} - \mu_{R_2} \quad (17)$$

$$v = \mu_T - \frac{(\mu_{R_1} + \mu_{R_2})}{2} \quad (18)$$

The null hypothesis H_{01} in Eq. 11 is rejected if $Z_1 > z_{\alpha}$, where,

$$Z_1 = \frac{\hat{\theta} + \delta}{\frac{s}{\sqrt{n_1}} \sqrt{\frac{2}{(\bar{U})^2} + 4 \frac{(\bar{V})^2}{(\bar{U})^4}}} = \frac{\bar{V} / \bar{U} + \delta}{\frac{s}{\sqrt{n_1}} \sqrt{\frac{2}{(\bar{U})^2} + 4 \frac{(\bar{V})^2}{(\bar{U})^4}}} \quad (19)$$

and z_α is the upper α quantile of the standard normal distribution (e.g., $z_{0.05} = 1.645$) and

$$s^2 = \frac{\left[\sum_{i=1}^{n_1} (Y_{T,i} - \bar{Y}_T)^2 + \sum_{i=1}^{n_2} (Y_{R_1,i} - \bar{Y}_{R_1})^2 + \sum_{i=1}^{n_2} (Y_{R_2,i} - \bar{Y}_{R_2})^2 \right]}{(n_1 + 2n_2 - 3)} \quad (20)$$

Similarly, the null hypothesis H_{02} in Eq. 12 is rejected if $Z_2 < -z_\alpha$ where,

$$Z_2 = \frac{\bar{V} / \bar{U} - \delta}{\frac{s}{\sqrt{n_1}} \sqrt{\frac{2}{(\bar{U})^2} + 4 \frac{(\bar{V})^2}{(\bar{U})^4}}} \quad (21)$$

If each null hypothesis in both Eqs. 11 and 12 is rejected at the significance level of α , the two products are claimed to be biosimilar.

Kang and Chow^[7] derived the asymptotic power function of the statistical test based on the ratio estimator and calculated sample sizes numerically. Simulation studies also showed that the statistical test based on the ratio estimator keeps the empirical type I error rates under the nominal level.

BINARY END POINTS

The statistical test based on the ratio estimator in the three-arm parallel design was extended to three popular metrics for the binary end point—the risk difference, the log relative risk, and the log odds ratio.^[10] Let X_T and X_{R_i} denote the number of events of interest. It is assumed that:

$$X_T \sim B(n_1, p_T) \quad X_{R_i} \sim B(n_2, p_{R_i}), i = 1, 2 \quad (22)$$

and that the three random variables are independent. For the risk difference, the hypotheses used to assess the degree of biosimilarity are expressed as:

$$H_0^d : \theta_d \leq -\delta_d \text{ or } \theta_d \geq \delta_d \text{ versus } H_A^d : -\delta_d < \theta_d < \delta_d \quad (23)$$

where,

$$\theta_d = \frac{p_T - \frac{1}{2}p_{R_1} - \frac{1}{2}p_{R_2}}{p_{R_1} - p_{R_2}} \quad (24)$$

For the log relative risk, the hypotheses for assessing biosimilarity can be written as:

$$H_0^r : \theta_r \leq -\delta_r \text{ or } \theta_r \geq \delta_r \text{ versus } H_A^r : -\delta_r < \theta_r < \delta_r \quad (25)$$

where,

$$\theta_r = \frac{\log(p_T) - \frac{1}{2}\log(p_{R_1}) - \frac{1}{2}\log(p_{R_2})}{\log(p_{R_1}) - \log(p_{R_2})} \quad (26)$$

For the log odds ratio, the hypotheses of assessing biosimilarity can be given by:

$$H_0^o : \theta_o \leq -\delta_o \text{ or } \theta_o \geq \delta_o \text{ versus } H_A^o : -\delta_o < \theta_o < \delta_o \quad (27)$$

where,

$$\theta_o = \frac{\log\left(\frac{p_T}{1-p_T}\right) - \frac{1}{2}\log\left(\frac{p_{R_1}}{1-p_{R_1}}\right) - \frac{1}{2}\log\left(\frac{p_{R_2}}{1-p_{R_2}}\right)}{\log\left(\frac{p_{R_1}}{1-p_{R_1}}\right) - \log\left(\frac{p_{R_2}}{1-p_{R_2}}\right)} \quad (28)$$

Shin and Kang^[10] developed the three statistical tests for evaluating biosimilarity based on the asymptotic normality of $\sqrt{n_1}(\hat{\theta}_d - \theta_d)$, $\sqrt{n_1}(\hat{\theta}_r - \theta_r)$, and $\sqrt{n_1}(\hat{\theta}_o - \theta_o)$.

CONCLUSION

The three-arm parallel design consists of three arms—one arm is for the test product and the other two arms are for reference products from different batches. This design allows us to assess the relative distance in terms of the ratio of the difference between T and R and the difference between R_1 and R_2 . An advantage of the three-arm parallel design is that it uses the relative distance rather than the absolute difference in order to assess the degree of biosimilarity. Kang and Shin^[8] extended the three-arm parallel design to what is known as the $(k+1)$ -arm parallel design for cases in which $k \geq 3$.

ACKNOWLEDGMENT

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2016R1D1A1A09916819).

REFERENCES

1. Chow, S.C.; Liu, J.P. Statistical assessment of biosimilar products. *J. Biopharm. Stat.* **2010**, *20*, 10–30.
2. Chow, S.C.; Hsieh, T.C.; Chi, E.; Yang, J. A comparison of moment-based and probability-based criteria for assessment of follow-on biologics. *J. Biopharm. Stat.* **2010**, *20*, 31–45.

3. Chow, S.C. *Biosimilars - Design and Analysis of Follow-on Biologics*; CRC Press: Boca Raton, FL, 2013.
4. Chow, S.C.; Wang, J.; Endrenyi, L.; Lachenbruch, P.A. Scientific considerations for assessing biosimilar products. *Stat. Med.* **2013**, *32* (3), 370–381.
5. Chow, S.C.; Endrenyi, L.; Lachenbruch, P.A.; Mentre, F. Scientific factors and current issues in biosimilar studies. *J. Biopharm. Stat.* **2014**, *24*, 1138–1153.
6. Hsieh, T.C.; Chow, S.C.; Yang, L.Y.; Chi, E. The evaluation of biosimilarity index based on reproducibility probability for assessing follow-on biologics. *Stat. Med.* **2013**, *32* (3), 406–414.
7. Kang, S.H.; Chow, S.C. Statistical assessment of biosimilarity based on relative distance between follow-on biologics. *Stat. Med.* **2013**, *32* (3), 382–392.
8. Kang, S.H.; Shin, W. Statistical assessment of biosimilarity based on the relative distance between follow-on biologics in the (k+1)-arm parallel design. *Commun. Stat. Appl. Methods.* **2015**, *22* (6), 605–613.
9. Lu, Y.; Zhang, Z.Z.; Chow, S.C. Frequency estimator for assessing of follow-on biologics. *J. Biopharm. Stat.* **2014**, *24*, 1280–1297.
10. Shin, W.; Kang, S.H. Statistical assessment of biosimilarity based on the relative distance between follow-on biologics for binary endpoints. *J. Biopharm. Stat.* **2016**, *26*, 227–239.
11. Yang, L.Y.; Lai, C.H. Estimation and approximation approaches for biosimilar index based on reproducibility probability. *J. Biopharm. Stat.* **2014**, *24*, 1298–1311.

Biosimilarity Assessment

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Fuyu Song

Center for Food and Drug Inspection, China Food and Drug Administration, Beijing, China

Abstract

In recent years, the assessment of biosimilarity between a proposed biosimilar product and an innovative biological product has attracted much attention since the passage of the Biologics Price Competition and Innovation Act in 2009. In this article, we provide an overview of regulatory requirements for review and approval of biosimilar products. Unlike the generic products, the development of biosimilar products is much more complicated because of fundamental differences in functional structures and manufacturing processes. As a result, the criteria and standard methods for the design and analysis of bioequivalence assessment of generic drug products may not be directly applicable to assessing biosimilarity of biosimilar products. In this article, we provide some scientific/statistical considerations for criteria, design, and analysis regarding the assessment of biosimilarity and drug interchangeability of biosimilar products. In addition, we discuss scientific and practical issues raised at the 2010 Food and Drug Administration (FDA) public hearing and the 2011 FDA public meeting on biosimilar products.

Keywords: Biosimilar; Biosimilarity index; Highly similar; Stepwise approach; Totality of the evidence.

INTRODUCTION

Biological drugs make up one of the fastest growing sectors of the pharmaceutical and biotechnology industry (MIDAS, 2010). For example, in 2000, global biologics spending on biosimilars is expected to roar to \$190 billion (see, e.g., Fig. 1). This has given the pharmaceutical and biotechnology industry the opportunity for the development of biosimilars when the innovative biologics are going off patent protection.

Due to the high costs involved in the production and consumption of many drug products, regulatory regimes have been created to balance the intellectual property interests (patent protection) and investment made by innovative (originator) companies, with the need for wider patient access through generic forms of the drugs. In the United States (US), for the traditional chemical (small-molecule) drug market, such a regime was created by the Hatch-Waxman Act. The Hatch-Waxman Act added section 505(j) to the US Federal Food, Drug, and Cosmetic Act. This section and its accompanying regulations created the Abbreviated New Drug Application process, which was designed to provide independent generic firms with a strong incentive to develop and introduce lower cost generic drugs to consumers. By virtually all accounts, the

Hatch-Waxman Act has been extremely successful in bringing cheaper generic products to the market while maintaining incentives for the development and discovery of new drugs. Because of the effectiveness of the Hatch-Waxman Act approach in the context of traditional small-molecule drugs, some have called for applying a similar regime for biologics, given the high costs in this class of drugs. In particular, Rep. Henry Waxman (D-Cal.), one of the original authors of the Hatch-Waxman Act has sponsored the Access to Life-Saving Medicine Act to determine whether such a legal infrastructure was appropriate for regulating the follow-on biologics, and a thorough policy assessment was necessary. This effort has led to the passage of the Biologics Price Competition and Innovation (BPCI) Act.

Following the passage of the BPCI Act, the US Food and Drug Administration (FDA) hosted a public hearing between November 2 and 3, 2010, to obtain public input regarding scientific factors for assessing biosimilarity and drug interchangeability of biosimilar products. After extensive discussions, the FDA developed and circulated three draft guidances on biosimilars on February 9, 2012, and hosted another public hearing to obtain public input and comments on the draft guidances on May 11, 2012.^[1–3] At this Public Hearing, special attention was directed to the discussion of drug interchangeability in terms of the

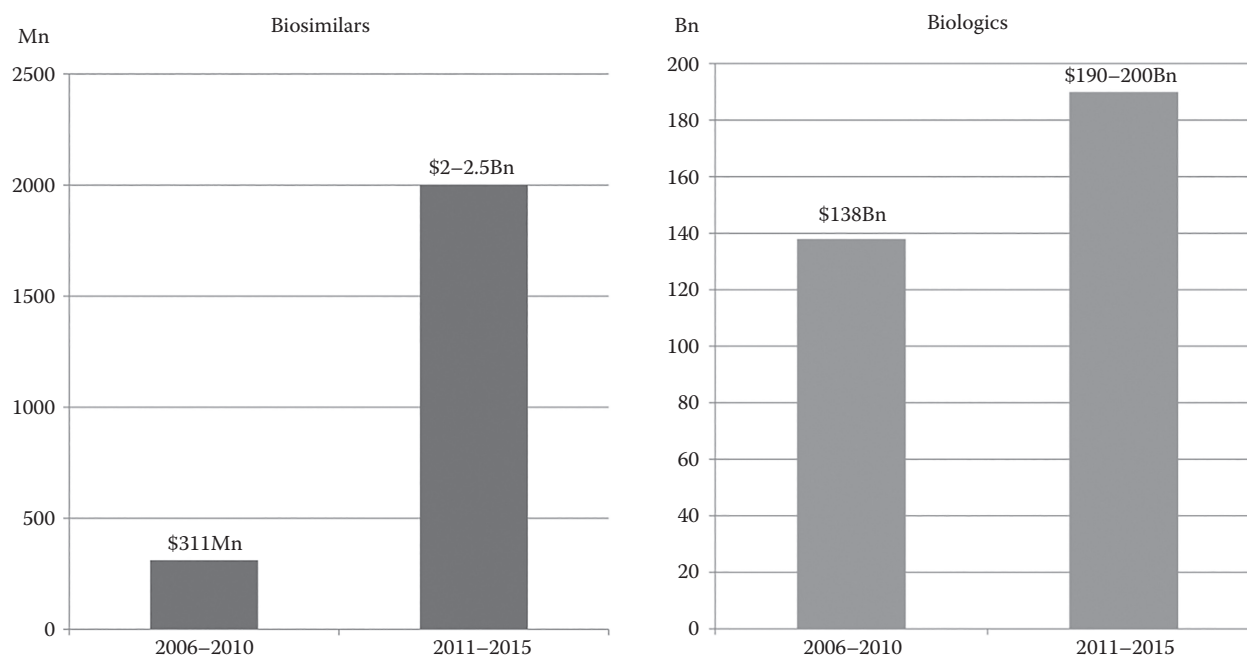


Fig. 1 Global spending on biologics

concepts of switching and alternating as described in the BPCI Act. On March 23, 2010, the BPCI Act (as part of the Affordable Care Act) was written into law, which has given the FDA the authority to approve similar biological drug products. After additional discussions, the draft guidances were finalized in early 2015.

As indicated in the BPCI Act, a biosimilar product is defined as a product that is *highly similar* to the reference product notwithstanding minor differences in clinically inactive components, and there are no clinically meaningful differences in terms of safety, purity, and potency. However, little or no discussion regarding how similar is considered highly similar is given in the BPCI Act.^[4,5] On the other hand, the European Union (EU) has a well-established and well-documented legal and regulatory pathway for the review and approval of biosimilars. Other countries and organizations around the world, including Australia, Canada, China, Japan, and Switzerland, and the World Health Organization (WHO), are also following the same scientific principles for an abbreviated approval pathway for biosimilars. In contrast, the US is at the very beginning of the process, with a legal pathway being discussed by the US Congress for a number of years. The FDA's views on biosimilars can be examined from a publication^[6] and from communications between the FDA and the Congress. These documents seem to indicate that FDA is contemplating similar scientific principles to those established by the EU European Medicines Agency (EMA).

The purpose of this article is to provide a comprehensive review of regulatory requirements, criteria for biosimilarity, statistical methods including power calculations for the determination of sample size that is needed for the assessment of biosimilarity between a proposed biosimilar

product and an innovative biologic product. In addition, a unified approach for demonstrating biosimilarity using the concept of reproducibility is introduced.

REGULATORY REQUIREMENTS

For the assessment of similar biologic products, regulatory requirements from the WHO and different regions such as the EU, the US, and the Asia-Pacific Region such as Japan, Taiwan, South Korea, and China are similar and yet slightly different (Wang and Chow 2012),^[7–11] which are briefly described below. In order to provide a better understanding of regulatory requirements from different regions, a comparison of requirements for the evaluation of biosimilars is given in [Table 1](#).

World Health Organization

As an increasingly wide range of similar biotherapeutic products (SBPs) was under development or already licensed in many countries, WHO formally recognized the need for the guidance for their evaluation and overall regulation in 2007. The *Guidelines on Evaluation of Similar Biotherapeutic Products (SBPs)* were developed and adopted by the 60th meeting of the WHO Expert Committee on Biological Standardization in 2009.^[12] The intention of the guidelines was to provide globally acceptable principles for licensing biotherapeutic products that were claimed to be similar to the reference products that had been licensed based on a full licensing dossier. The scope of the guidelines includes well-established and well-characterized biotherapeutic products that have been

Table 1 Comparison of regulatory requirements

	WHO	Canada	Korea	EU	Japan
Term	SBPs	SEBs	Biosimilars	Biosimilars	Follow-on biologics
Scope		Recombinant protein drugs		Mainly recombinant protein drugs	Recombinant protein drugs
Efficacy	Double-blind or observer-blind; equivalence or non-inferiority design		Equivalence design	Comparability margins should be prespecified and justified	
Reference Product	Authorized in a jurisdiction with a well-established regulatory framework			Authorized in the EU	Authorized in Japan
Stability	1. Accelerated degradation studies 2. Studies under various stress conditions				Not necessary
Purity	Process-related and product-related impurities				
Manufacture	1. Same standards required by the NRA for originator products 2. Full chemistry and manufacture data package				
Physicochemical	1. Primary and higher order structure 2. Posttranslational modifications				
Biological activity	1. Qualitative measure of the function 2. Quantitative measure (e.g., enzyme assays or binding assays)				
Nonclinical studies	1. <i>In vitro</i> (e.g., receptor-binding, cell-based assays) 2. <i>In vivo</i> (PD activity, at least one repeat dose toxicity study, antibody measurements, local tolerance)				
Safety	Pre-licensing safety data and risk management plan				
Principles	1. Generic approach is not appropriate for follow-on biologic 2. Follow-on biologic should be similar to the reference in terms of quality, safety, and efficacy 3. Stepwise comparability approach: Similarity of the SBP to RBP in terms of quality is a prerequisite for reduction of nonclinical and clinical data required for approval 4. Case-by-case approach for different classes of products 5. Pharmacovigilance is stressed				
PK study design and criteria	1. Single-dose, steady-state studies, or repeated determination of PK 2. Crossover or parallel 3. Include absorption and elimination characteristics 4. Traditional 80%–125% equivalence range is used				
PD	PD markers should be selected and comparative PK/PD studies may be appropriate				

marketed for a suitable period of time with proven quality, efficacy, and safety, such as recombinant DNA-derived therapeutic proteins.

Key Principles and Basic Concept

As indicated in the WHO guidelines, one of the most important principles of developing an SBP is the stepwise approach starting with the characterization of quality attributes of the product and followed by nonclinical and clinical evaluations. Manufacturers should submit a full-quality dossier that includes a complete characterization of the product, the demonstration of consistent and robust manufacture of their product, and the comparability exercise between the SBP and the reference biotherapeutic product (RBP) in the quality part, which together serve as the basis for the possible reduction in data requirements in the nonclinical and clinical development. This principle indicates that the data reduction is only possible for the nonclinical and clinical parts of the development program,

and significant differences between the SBP and the chosen RBP detected during the comparability exercise would result in requirement for more extensive nonclinical and clinical data. In addition, the amount of nonclinical and clinical data considered necessary also depends on the class of products, which calls for a case-by-case approach for different classes of products.

Reference Biotherapeutic Product

The choice of the RBP is another important issue covered in the WHO guidelines. Traditionally, National Regulatory Authorities (NRAs) have required the use of a nationally licensed reference product for registering generic medicines, but this may not be feasible in countries lacking nationally licensed RBPs. Thus, additional criteria may be needed to guide the acceptability of using an RBP licensed in other jurisdictions. Considering the choice of the RBP, WHO requires that it should have been marketed for a *suitable* duration and have a volume of marketed use, and be

licensed based on full-quality, safety, and efficacy data. Besides, the same RBP should be used throughout the development of SBP, and the drug substance, dosage form, and route of administration of the SBP should be the same as those of the RBP.

Quality

As mentioned in the preceding section, the comprehensive comparison showing quality similarity between SBP and RBP is a prerequisite for applying the clinical safety and efficacy profile of RBP to SBP; thus, a full-quality dossier for both drug substance and drug product is always required. To evaluate comparability, WHO recommends the manufacturer to conduct a comprehensive physicochemical and biological characterization of the SBP in head-to-head comparisons with RBP. The following aspects of product quality and heterogeneity should be assessed.

Manufacturing Process

The manufacturing process should meet the same standards as required by NRAs for the originator products, and implement Good Manufacturing Practices, modern quality control and assurance procedures, in-process controls, and process validation. The SBP manufacturer should assemble all available knowledge of the RBP with regard to the type of host cell, formulation, and container closure system, and submit a complete description and data package delineating the whole manufacturing process including obtaining target genes and their expression, the optimization and fermentation of gene engineering cells, the clarification and purification of the products, the formulation and testing, aseptic filling, and packaging.

Characterization

Thorough characterization and comparability exercise are required, and details should be provided on the primary and higher order structure, posttranslational modifications, biological activity, process- and product-related impurities, the relevant immunochemical properties, and results from accelerated degradation studies and studies under various stress conditions.

Nonclinical and Clinical Studies

After demonstrating the similarity of SBP and RBP in quality, the proving of safety and efficacy of an SBP usually requires further nonclinical and clinical data. Nonclinical evaluations should be undertaken both *in vitro* (e.g., receptor-binding studies, cell proliferation, cytotoxicity assays) and *in vivo* (e.g., biological/pharmacodynamics (PD) activity, repeat dose toxicity study, toxicokinetic measurements, anti-product antibody titers,

cross-reactivity with homologous endogenous proteins, neutralizing capacity of the product).

In terms of the clinical evaluation, the comparability exercise should begin with pharmacokinetic (PK) and PD studies followed by the pivotal clinical trials. PK studies should be designed to enable detection of potential differences between SBP and RBP. Single-dose, crossover PK studies in a homogeneous population are recommended by WHO. The manufacturer should justify the choice of single-dose studies, steady-state studies, or repeated determination of PK parameters, and the study population. Due to the lack of established acceptance criteria for the demonstration of similar PK between SBP and RBP, the traditional 80%–125% equivalence range is often used. Besides, PD studies and confirmatory PK/PD studies may be appropriate if there are clinically relevant PD markers. In addition, similar efficacy of SBP and RBP has to be demonstrated in randomized and well-controlled clinical trials that should preferably be double-blind or at least observer-blind. In principle, equivalence designs (requiring lower and upper comparability margins) are clearly preferred for the comparison of efficacy and safety of SBP with RBP. Non-inferiority designs (requiring only one margin) may be considered if appropriately justified. WHO also suggest the pre-licensing safety data and the immunogenicity data should be obtained from the comparative efficacy trials.

In addition to the nonclinical and clinical data, applicants also need to present an ongoing risk management and pharmacovigilance plan, since data from preauthorized clinical studies are usually too limited to identify all potential side effects of the SBP. The safety specification should describe the important identified or potential safety issues for the RBP, and any that are specific for the SBP.

In summary, the WHO guidelines on evaluating SBPs represent an important step forward in the global harmonization of biosimilar product evaluation and regulation, and provide a clear guidance for both regulatory bodies and the pharmaceutical industry.

The EU

The EU has pioneered the development of a regulatory system for biosimilar products. The EMA began formal consideration of scientific issues presented by biosimilar products at least as early as January 2001 when an *ad hoc* working group discussed the comparability of medicinal products containing biotechnology-derived proteins as active substances.^[13] In 2003, the European Commission amended the provisions of EU secondary legislation governing requirements for marketing authorization of applications for medicinal products to establish a new category of applications for “similar biological medicinal products”.^[14] In 2005, the EMA issued a general guideline in order to introduce the concept of similar biological medicinal products, to outline the basic principles to

be applied, and to provide applicants with a “user guide,” showing where to find relevant scientific information.^[15] Since then, 14 biosimilar products have been approved by EMA under the pathway. One of the rejected biosimilars is Alpheon (interferon alfa-2a). It was developed by BioPartners GmbH and designed to become a biosimilar of the reference product Roferon-A for the treatment of adult patients with chronic hepatitis C. The EMA refused the marketing authorization for Alpheon due to the difference identified between Alpheon and the reference product, such as impurities, stability, and side effects.

Key Principles and Basic Concept

Unlike WHO’s guideline, which seems to focus more on recombinant DNA-derived therapeutic proteins, EMA’s guidelines clearly indicate that the concept of a “similar biological medicinal product” is applicable to a broad spectrum of products ranging from biotechnology-derived therapeutic proteins to vaccines, blood-derived products, monoclonal antibodies, gene and cell therapy, and so on. However, comparability exercises to demonstrate similarity are more likely to be applied to highly purified products that can be thoroughly characterized such as biotechnology-derived medicinal products. Considering the amount of data submitted, EMA also requires a full-quality dossier, while the comparability exercise at the quality level may allow a reduction of the nonclinical and clinical data requirement compared to a full dossier. In 2011, EMA published a concept paper on the revision of the guideline on similar biological medicinal products.^[16] It emphasizes another main concept that clinical benefit has already been established by the reference medicinal product, and that the aim of a biosimilar development program is to establish similarity to the reference product, not clinical benefit. Besides, a clear definition of “biosimilar,” the feasibility to follow the generic legal basis for some biological products, and the refinement based on the experience gained are recommended.

Reference Biotherapeutic Product

Similar to the regulatory requirement from WHO, EMA requires that the active substance, the pharmaceutical form, the strength, and the route of administration of the biosimilar product be the same as those of the reference product. The same chosen reference medicinal product should be used throughout the comparability program for quality, safety, and efficacy studies during the development of the biosimilar product. One of the major differences between WHO and EMA in terms of the choice of the reference product is that EMA requires the chosen reference medicinal product to be a medicinal product authorized in the community. Data generated from comparability studies with medicinal products authorized outside the community may only provide supportive information.

Quality

In 2006, the Guideline on Similar Biological Medicinal Products Containing Biotechnology-Derived Proteins as Active Substance: Quality Issues was adopted by the Committee for Medicinal Products for Human Use (CHMP),^[17] which addresses the requirements regarding manufacturing processes, the comparability exercises for quality, analytical methods, physicochemical characterization, biological activity, purity, and specifications of the similar biological medicinal product. In 2011, EMA issued a concept paper on the revision of this guideline.^[18] This concept paper proposes that the guideline published in 2006 needs refinements taking account of the evolution of quality profile during the product life cycle, since in the context of a biotherapeutic product claiming or claimed to be similar to another one already marketed, the conclusion of a comparability exercise performed with a reference product at a given time may not hold true from the initial development of the biosimilar product, through marketing authorization, until the product’s discontinuation.

Nonclinical and Clinical Evaluation

The Guideline on Similar Biological Medicinal Products Containing Biotechnology-Derived Proteins as Active Substance: Non-Clinical and Clinical Issues published in 2006 lays down the nonclinical and clinical requirements for a biological medicinal product claiming to be similar to another one already marketed.^[19] The nonclinical section of the guideline addresses the pharmacotoxicological assessment, and the clinical section addresses the requirements for PK, PD, and efficacy studies. Clinical safety studies as well as the risk management plan with a special emphasis on studying the immunogenicity of the biosimilar products are also required. In 2011, EMA published a concept paper on the revision of this guideline,^[20] which indicates several issues that need discussion for a potential revision. First, EMA emphasizes the need to follow the 3R principles (replacement, reduction, and refinement) with regard to the use of animal experiments. Second, a revised version of the guideline will consider a risk-based approach for the design of an appropriate nonclinical study program. Third, the guideline should be clearer considering the need and acceptance of PD markers, and what measures should be taken in case relevant markers are not available. It, however, should be noted that most recently, EMA has issued a biosimilarity guideline in October 2014, making an attempt to address these issues.^[21]

Product Class-Specific Guidelines

The principles of biosimilar drug development discussed in the former sections apply in general to all biological drug products. However, there are no standard data sets that can be applied to the approval of all classes of

biosimilars. Each class of biologic varies in its benefit/risk profile, the nature and frequency of adverse events, the breadth of clinical indications, and whether surrogate markers for efficacy are available and validated. Accordingly, the EMA has developed product class-specific guidelines that define the nature of comparative studies. So far, guidance for the development of biosimilar products has been developed for six different product classes, including erythropoietins, insulins, growth hormones, interferon alfa, granulocyte-colony stimulating factors, and low-molecular-weight heparins (LMWHs),^[22] with three more (interferon beta, follicle stimulation hormone, monoclonal antibodies).^[17,23–28]

European Experiences

As indicated earlier, the EU EMA has issued scientific guidelines on the quality, nonclinical, and clinical standards for the approval of biosimilars. EMA product class-specific guidelines (including erythropoietin, granulocyte-colony stimulating factor, insulin, growth hormone, LMWH, and interferon alfa) are available. According to these product class-specific guidelines, as of December 31, 2010, 14 biosimilar drugs have been approved in Europe. Compared to other regions in the world, Europe holds the highest number of biosimilar approvals. This number is expected to increase as many biologic patents will soon expire in the near future, which will help increase market size and competition among market participants.

Remarks

In summary, the EU has taken a thoughtful and evidence-based approach, and has established a well-documented legal and regulatory pathway for the approval of biosimilar products distinct from the generic pathway. In order to grant a biosimilar product, the EMA requires comprehensive and justified comparability studies between the biosimilar and reference products at the quality, nonclinical, and clinical level, which are explained in detail in the EMA guidelines. The approval pathway of biosimilar products in the EU is based on case-by-case reviews, owing to the complexity and diversity of the biologic products. Therefore, besides the three general guidelines, EMA also developed additional product class-specific guidelines on nonclinical and clinical studies. This approval pathway is now held up as one of the gold standards for authorizing biosimilar products.

North America

US Food and Drug Administration

For the approval of follow-on biologics in the US, current regulations depend on whether the biologic product is approved under the US Food, Drug, and Cosmetic Act (US

FD&C) or is licensed under the US Public Health Service Act (US PHS). For those biologic drugs marketed under the PHS Act, the BPCI Act passed by the US Congress on March 23, 2010, amends the PHS Act to establish an abbreviated approval pathway for biological products that are highly similar or interchangeable with an FDA-authorized biologic drug and gives the FDA the authority to approve follow-on biologics under new section 351(k) of the PHS Act. Some early biologic drugs, such as somatropin and insulin, were approved under the FD&C Act. In this case, biosimilar versions can receive approval for New Drug Applications under section 505(b)(2) of the FD&C Act.

Following the passage of the BPCI Act, in order to obtain the input on specific issues and challenges associated with the implementation of the BPCI Act from a broad group of stakeholders, the US FDA conducted a two-day public hearing on “Approval Pathway for Biosimilar and Interchangeability Biological Product” held on November 2–3, 2010, at the FDA in Silver Spring, Maryland. The scientific issues covered in this public hearing include, but are not limited to, the criteria and design for biosimilarity and interchangeability, comparability between manufacturing processes, patient safety and pharmacovigilance, exclusivity, and user fees.

In practice, there is a strong industrial interest and desire for the regulatory agencies to develop review standards and an approval process for biosimilars rather than an *ad hoc* case-by-case review of individual biosimilar applications. As indicated earlier, for this purpose, the FDA has established three committees to ensure consistency in the FDA’s regulatory approach of follow-on biologics. The three committees are the Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research (CDER/CBER) Biosimilar Implementation Committee (BIC), the CDER Biosimilar Review Committee (BRC), and the CBER BRC. The CDER/CBER BRC focuses on the cross-center policy issues related to the implementation of the BPCI Act. The CDER BRC and CBER BRC are responsible for considering the requests of applicants for advice about proposed development programs for biosimilar products, reviewing biologics license applications (BLAs) that are submitted under section 351(k) of the PHS Act, and managing related issues. Thus, the review process steps of CDER BRC and CBER BRC include the following: (1) an applicant submitting request for advice, (2) internal review team meeting, (3) internal CDER BRC (CBER BRC) meeting, (4) internal post-BRC meeting, and (5) the applicant meeting with CDER (CBER).

Another important issue aroused by the BPCI Act is the interchangeability of biosimilars. Once approved, standard generic drugs can be automatically substituted for the reference product without the intervention of the health-care provider in many states. However, the automatic interchangeability cannot be applied to all biosimilars. In order to meet the higher standard of interchangeability, a sponsor must demonstrate that the biosimilar products

can be expected to produce the same clinical result as the reference product in any given patient.

On February 9, 2012, the FDA announced the publication of three draft guidance documents to assist industry in developing follow-on biologic products, including (1) Scientific Considerations in Demonstrating Biosimilarity to a Reference Product, (2) Quality Considerations in Demonstrating Biosimilarity to a Reference Protein Product, and (3) Biosimilars: Questions and Answers Regarding Implementation of the Biologics Price Competition and Innovation Act of 2009.^[1–3] Subsequently, the FDA hosted another public hearing on the discussion of these draft guidances at the FDA on May 11, 2012. Similar to the requirements of the WHO and EMA, a number of factors are considered important by the FDA when assessing applications for biosimilars, including the robustness of the manufacturing process, the demonstrated structural similarity, the extent to which mechanism of action was understood, the existence of valid, mechanistically related PD assays, comparative PK and immunogenicity, and the amount of clinical data and experience available with the original products. The guidances were later finalized in early 2015. Even though they do not provide clear standards for assessing biosimilar products, they are the first step toward removing the uncertainties surrounding the biosimilar approval pathway in the US.

Recently, following the Advisory Committee's recommendation for approval of a proposed biosimilar (by Novartis) to Amgen's Neupogen® (filgrastim) on January 7, 2015, FDA approved the proposed biosimilar on March 26, 2015. Neupogen® was originally approved by the FDA in 1991 and its patent expired in December 2013. Neupogen® was intended to decrease the incidence of infection, as manifested by febrile neutropenia in patients with nonmyeloid malignancies receiving myelosuppressive anticancer drugs associated with a clinically significant incidence of febrile neutropenia. It should be noted that the approved biosimilar of filgrastim is different from Teva's Neutroval approved by the FDA in August 2012. The approval of Teva's Neutroval was based on a full Biologics License Application (BLA) rather than under the FDA's new biosimilar approval pathway which allows Teva to compete directly with Amgen's filgrastim in the US.

Canada (Health Canada)

Health Canada, the federal regulatory authority that evaluates the safety, efficacy, and quality of drugs available in Canada, also recognizes that with the expiration of patents for biologic drugs, manufacturers may be interested in pursuing subsequent entry versions of these biologic drugs which are called subsequent entry biologics (SEB) in Canada. In 2010, Health Canada issued the "Guidance for Sponsors: Information and Submission Requirements for Subsequent Entry Biologics (SEBs)" whose objective is to provide guidance on how to satisfy the data and

regulatory requirements under the Food and Drugs Act and Regulations for the authorization of SEBs in Canada.^[29]

The concept of a SEB applies to all biologic drug products; however, there are additional criteria to determine whether the product will be eligible to be authorized as SEBs:

1. A suitable reference biologic drug exists that was originally authorized based on a complete data package and has significant safety and efficacy data accumulated.
2. The product can be well characterized by state-of-the-art analytical methods.
3. The SEB can be judged similar to the reference biologic drug by meeting an appropriate set of predetermined criteria.

With regard to the similarity of products, Health Canada requires the manufacturer to evaluate the following factors:

1. Relevant physicochemical and biological characterization data
2. Analysis of the relevant samples from the appropriate stages of the manufacturing process
3. Stability data and impurities data
4. Data obtained from multiple batches of the SEB and reference to understand the ranges in variability
5. Nonclinical and clinical data and safety studies

In addition, Health Canada also has stringent postmarketing requirements including an adverse drug reaction report, periodic safety update reports, and suspension or revocation of notice of compliance. The guidance of Canada shares similar concepts and principles as indicated in the WHO guidelines since it is clearly mentioned in the guidance that Health Canada has the intention to harmonize as much as possible with other competent regulators and international organizations.

Asia-Pacific Region (Japan, Korea, and China)

Ministry of Health, Labour and Welfare

The Japanese Ministry of Health, Labour and Welfare (MHLW) has also been confronted with the new challenge of regulating biosimilar/follow-on biologic products. Based on the similarity concept outlined by the EMA, Japan has published a guideline for the quality, safety and efficacy of biosimilar products in 2009 (MHLW 2009).^[23] The scope of the guideline includes recombinant plasma proteins, recombinant vaccines, PEGylated recombinant proteins, and nonrecombinant proteins that are highly purified and characterized. Unlike the EU, polyglycans such as LMWH have been excluded from the guideline. Another class of products excluded is synthetic peptides,

since the desired synthetic peptides can be easily defined by structural analyses and can be defined as generic drugs. Same as the requirements by the EU, the original biologic should be already approved in Japan. However, there are some differences in the requirements of stability test and toxicology studies for impurities in biosimilars between the EU and Japan. A comparison of the stability of a biosimilar with the reference innovator products as a strategy for the development of biosimilar is not always necessary in Japan. In addition, it is not required to evaluate the safety of impurities in the biosimilar product through nonclinical studies without comparing it with the original product. According to this guideline, follow-on biologics such as Somatropin and Epoetin alfa BS have been approved in Japan.

Korea Food and Drug Administration

In Korea, Pharmaceutical Affairs Act is the high-level regulation to license all medicines including biologic products. The Korean Food and Drug Administration (KFDA) notifications serve as a lower level regulation. Biologic products and biosimilars are subject to the Notification of the regulation on review and authorization of biologic products. The KFDA actively participates to promote a public dialog on the biosimilar issues. In 2008 and 2009, the KFDA held two public meetings and cosponsored a workshop to gather input on scientific and technical issues. The regulatory framework of biosimilar products in Korea is a three-tiered system: (1) Pharmaceutical Affairs Act, (2) Notification of the regulation on review and authorization of biological products, and (3) Guideline on evaluation of biosimilar products.^[30,31] As the Korean guideline for biosimilar products was developed along with that of the WHO,^[12] most of the requirements are similar except for that of the clinical evaluation to demonstrate similarity. The KFDA requires that equivalent rather than noninferior efficacy be shown in order to open the possibility of extrapolation of efficacy data to other indications of the reference product. Equivalence margins need to be pre-defined and justified, and should be established within the range which is judged not to be clinically different from reference products in clinical regards.

China Food and Drug Administration

The Center for Drug Evaluation (CDE) of the China Food and Drug Administration (CFDA) published a draft guidance for comments on October 29, 2014.^[32] The CFDA finalized the draft guidance and published a current trial version on February 28, 2015.^[33] In this version, one significant change is to relax the regulation for the selection of reference product for comparison. In the version for comments, the CFDA requires the reference product to be the originator product authorized by the CFDA. In the current trial version, the CFDA has removed this restriction. In

other words, EU-approved and/or US-licensed reference products can be used as the reference products. In addition, for PK studies, the traditional 80%–125% equivalence range is not mandatory. In the current trial version, the CFDA indicates that an alternative equivalence range (e.g., SABE for highly variable drugs [HVDs]) can be used if scientifically justifiable.

CRITERIA FOR BIOSIMILARITY

As indicated in Chow and Liu,^[34] bioequivalence assessment for generic drug products is possible under the Fundamental Bioequivalence Assumption. It states that if two drug products are shown to be bioequivalent in the drug absorption profile (which is measured in terms of the extent and rate of absorption), it is assumed that they will reach the same therapeutic effect or they are therapeutically equivalent. The Fundamental Bioequivalence Assumption assumes that there is association between PK responses and clinical outcomes. This assumption, however, may not be applicable for the assessment of biosimilarity of biosimilars due to the fundamental differences between generic (small-molecule) drug products and similar biological (large-molecule) drug products (see Table 2). In what follows, several criteria for the primary assessment of bioequivalence or similarity, which have been proposed by the FDA, are discussed for their applicability to the assessment of biosimilarity.

Average Bioequivalence (Biosimilarity)

For the assessment of average bioequivalence (ABE) both *in vivo* and *in vitro*, FDA adopted a one-size-fits-all criterion. That is, for *in vivo* (*in vitro*) bioequivalence

Table 2 Fundamental differences between generic drugs and biosimilars

Generic drugs	Biologic drugs (biosimilars)
Made by chemical synthesis	Made by living cells
Defined structure	Heterogeneous structure
	Mixtures of related molecules
Easy to characterize	Difficult to characterize
Relatively stable	Variable
	Sensitive to environmental conditions such as light and temperature
No issue of immunogenicity	Issue of immunogenicity
Usually taken orally	Usually injected
Often prescribed by a general practitioner	Usually prescribed by specialists

assessment, a test drug product is said to be bioequivalent to a reference drug product if the estimated 90% confidence interval for the geometric mean ratio (GMR) of the primary PK parameters, e.g., area under the blood or plasma concentration–time curve (AUC) and maximum concentration (C_{max}), is totally within the bioequivalence limits of 80.00%–125.00% (90.00%–111.11%). See, e.g., Chow and Liu^[34]; FDA.^[35]

The one-size-fits-all criterion does not take into consideration the therapeutic window (TW) and intra-subject variability of a drug which have been identified to have non-negligible impact on the safety and efficacy of generic drug products compared to the innovative drug products. In the past several decades, this one-size-fits-all criterion has been challenged and criticized by many researchers. It was suggested that flexible criteria in terms of safety (upper bioequivalence limit) and efficacy (lower bioequivalence limit) should be developed based on the characteristics of the drug, its TW, and intra-subject variability (ISV) (see Table 3).

On the other hand, for orally administered drugs with high within-subject variability and wide TW (Class D, HVDs, see Table 3), the regulatory expectation has become, in some cases, more relaxed. For these drugs, the approach of scaled average bioequivalence (SABE) has been proposed. This method is, in fact, a special case of the procedure described earlier for individual BE when the within-subject variation is high ($\sigma_{WR}^2 > \sigma_{W0}^2$). However, the current FDA guidance does not contain special provisions for this class of drugs.

The assessment of ABE focuses on average bioavailability but ignores the variability associated with the PK responses. Thus, two drug products may fail the evaluation of ABE if the variability associated with the PK responses is large even though they have identical means. A drug with large variability is considered highly variable. The FDA defines an HVD as a drug whose within-subject (or intra-subject) variation is larger than or equal to 30%. This definition is based on intra-subject variation, however, rather arbitrary. One of the problematic aspects of this definition is that the estimated within-subject variability depends on the metrics of PK responses such as AUC and C_{max} . Haidar et al.^[36] pointed out that HVDs show variable PK as a result of their inherent properties (e.g., distribution,

systemic metabolism, and elimination) (see also, Haidar et al., Tothfalusi et al., Davit et al.).^[37–39] A drug may have low variability if it is administered intravenously, whereas it can be highly variable after oral administration.

In practice, HVDs often fail to meet the current regulatory acceptance criteria for ABE. In the last decade, the topic for evaluation of bioequivalence for HVDs has received much attention. This topic has been discussed several times at regulatory forums and international conferences, but academics, representatives of pharmaceutical industries, and regulatory agencies failed to reach a consensus until recently, and an approach of SABE was proposed by Haidar et al.^[36,39] The approach of SABE is briefly described below.

Logarithmic means the two drug products, denoted by μ_T and μ_R , respectively, are typically compared. The acceptance of bioequivalence is claimed if the difference between the logarithmic means is between prespecified regulatory limits. The limits (θ_A) are generally symmetrical on the logarithmic scale and usually equal $\pm \ln(1.25)$. Thus, the criterion for ABE can be expressed as follows:

$$-\theta_A \leq \mu_T - \mu_R \leq \theta_A$$

In a bioequivalence study, the individual kinetic responses are evaluated from the measured concentrations. The means of the logarithmic responses of the two formulations are calculated. These sample averages estimate the true population means. A variance is also estimated for each kinetic response. It is a measure of the intra-subject variance but not always identical to it. The FDA suggests that the above ABE could be scaled by a standard deviation as follows:

$$-\theta_s \leq \frac{(\mu_T - \mu_R)}{\sigma_w} \leq \theta_s,$$

where θ_s is the SABE regulatory cutoff. Here the standard deviation (σ_w) is the within-subject standard deviation. In a replicate design, σ_w is generally the within-subject standard deviation of the reference formulation (denoted by σ_{WR}). Thus, the scaling factor of SABE has similar features to the scaling factor of IBE.

Population/Individual Bioequivalence

In the early 1990s, as more generic drug products became available, it was a concern whether the use of generic drug products was safe and whether the approved generic drug products could be used interchangeably. The FDA indicates that an approved generic drug product can be used to substitute the innovative (brand-name) drug product. However, the FDA does not indicate that generic drug products can be used interchangeably. Since generic drug products are approved based on the criterion of the 80/125 rule, there may be a drastic change in blood concentration

Table 3 Classification of drugs

Class	TW	ISV	Example
A	Narrow	High	Cyclosporine
B	Narrow	Low	Theophylline
C	Wide	Low to moderate	Most drugs
D	Wide	High	Chlorpromazine or topical corticosteroids

TW, therapeutic window; ISV, intra-subject variability.

if one shall switch from one generic drug to another. For example, if one switched from a drug that was approved on the lower end of the 80/125 rule (say 80%) to another drug that was approved on the higher end of the 80/125 rule (say 120%), then there would be a sudden 50% increase in blood concentration, which may cause a potential safety concern. In order to address the issue of drug interchangeability in terms of drug prescribability and switchability, between the early 1990s and the early 2000s, the FDA suggested using the concepts of population bioequivalence for addressing drug prescribability and individual bioequivalence for addressing drug switchability.

Let y_T be the PK response from the test product, and y_R and $y_{R'}$ be two identically distributed PK responses from the reference product. Now consider a measure of the relative difference between the mean squared errors of $y_T - y_R$ and $y_R - y_{R'}$. Thus,

$$\theta = \begin{cases} \frac{E(y_T - y_R)^2 - E(y_R - y_{R'})^2}{E(y_R - y_{R'})^2 / 2} & \text{if } E(y_R - y_{R'})^2 / 2 \geq \sigma_0^2 \\ \frac{E(y_T - y_R)^2 - E(y_R - y_{R'})^2}{\sigma_0^2} & \text{if } E(y_R - y_{R'})^2 / 2 < \sigma_0^2 \end{cases}$$

where σ_0^2 is a given constant. If y_T , y_R , and $y_{R'}$ are independent observations from different subjects, then the two drug products show population bioequivalence when $\theta < \theta_p$. On the other hand, if y_T , y_R and $y_{R'}$ are independent observations from the same subject, then the two drug products exhibit individual bioequivalence when $\theta < \theta_i$. Thus, as indicated in section “North America,” for the assessment of individual bioequivalence, the criterion proposed in the FDA guidance^[40] can be expressed as

$$\theta_i = (\delta^2 + \sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2) / \max\{\sigma_{W0}^2, \sigma_{WR}^2\}, \quad (1)$$

where $\delta = \mu_T - \mu_R$, σ_{WT}^2 , σ_{WR}^2 , σ_D^2 are the true difference in means, intra-subject variabilities of the test product and the reference product, and variance component due to subject-by-formulation interaction between drug products, respectively. σ_{W0}^2 is the scale parameter specified by the user or regulator. Similarly, the criterion for the assessment of population bioequivalence suggested in the FDA guidance^[40] is given by

$$\theta_p = (\delta^2 + \sigma_{TT}^2 - \sigma_{TR}^2) / \max\{\sigma_{T0}^2, \sigma_{TR}^2\}, \quad (2)$$

where σ_{TT}^2 , σ_{TR}^2 are the total variances for the test product and the reference product, respectively, and σ_{T0}^2 is the scale parameter specified by the user or regulator.

A typical approach is to construct a one-sided 95% confidence interval for θ_i (θ_p) for the assessment of individual (population) bioequivalence. If the one-sided 95% upper confidence limit is less than the bioequivalence limit of

θ_i (θ_p), we then conclude that the test product is bioequivalent to that of the reference product in terms of individual (population) bioequivalence.

Masking Effect

The goal for evaluation of bioequivalence is to assess the similarity of the distributions of the PK metrics obtained either from the population or from individuals in the population. Under aggregate criteria, however, different combinations of values for the components of the aggregate criterion can yield the same value. In other words, bioequivalence can be reached by two totally different distributions of PK metrics. This is another artifact of the aggregate criteria. At the 1996 advisory committee meeting, it was reported that a 14% increase in the average (ABE only allows 80%–125%) is offset by a 48% difference in the variability, and the test passes IBE but fails ABE. More details regarding individual and population bioequivalence can be found in the works of Chow and Liu^[34] and Chow.^[41]

Profile Analysis for *In Vitro* Bioequivalence Testing

As indicated in the FDA draft guidance for *in vitro* bioequivalence testing, profile analysis using a confidence interval approach should be applied to cascade impactor or multistage liquid impinger for particle size distribution. Equivalence may be assessed based on chi-square differences. The idea is to compare the profile difference between the test product and reference product samples with the profile variation between reference product samples. More specifically, let y_{ijk} denote the observation from the j th subject's i th stage of the k th treatment. Given a sample (j_0) from the test product and two samples (j_0 , j_1) from the reference products and assuming that there are a total of S stages, the profile distance between the test and reference samples is given by

$$d_{TR} = \sum_{i=1}^S \frac{(y_{i0T} - 0.5(y_{i1R} + y_{i2R}))^2}{(y_{i0T} + 0.5(y_{i1R} + y_{i2R}))}.$$

Similarly, the profile variability within reference is defined as

$$d_{RR} = \sum_{i=1}^S \frac{(y_{i1R} - y_{i2R})^2}{0.5(y_{i1R} + y_{i2R})}.$$

For a given triplet sample of (Test, Reference 1, Reference 2), the ratio of d_{TR} and d_{RR} , i.e.,

$$rd = \frac{d_{TR}}{d_{RR}} \quad (3)$$

can then be used as a bioequivalence measure for the triplet samples between the two drug products. For a

selected sample, the 95% upper confidence bound of $E(rd) = E(d_{TR}/d_{RR})$ is then used as a bioequivalence measure for the determination of bioequivalence. In other words, if the 95% upper confidence bound is less than the bioequivalence limit, then we claim that the two products are bioequivalent. The FDA draft guidance recommends a bootstrap procedure to construct the 95% upper bound for $E(rd)$. The procedure is described below.

Assume that the samples are obtained in a two-stage sampling manner. In other words, for each treatment (test or reference), three lots are randomly sampled. Within each lot, ten samples (e.g., bottles or canisters) are sampled. The following is quoted from the 1999 FDA draft guidance regarding the bootstrap procedure to establish profile bioequivalence. For an experiment consisting of three lots each of test and reference products, and with ten canisters per lot, the lots can be matched into six different combinations of triplets with two different reference lots in each triplet. The ten canisters of a test lot can be paired with the ten canisters of each of the two reference lots in $(10 \text{ factorial})^2 = (3,628,800)^2$ combinations in each of the lot triplets. Hence, a random sample of the N -canister pairing of the six Test-Reference 1-Reference 2 lot triplets is needed. rd is estimated by the sample mean of the rd 's calculated for the triplets in ten selected samples of N . Note that the FDA recommends that $N = 500$ be considered.

Similarity Factor for Dissolution Profile Comparison

In vivo bioequivalence studies are surrogate trials for assessing equivalence between test and reference formulations based on the rate and extent of drug absorption in humans to establish similar effectiveness and safety under the Fundamental Bioequivalence Assumption. However, drug absorption depends on the dissolved state of drug product, and dissolution testing provides a rapid *in vitro* assessment of the rate and extent of drug release. Leeson,^[42] therefore, suggested that *in vitro* dissolution testing be used as a surrogate for *in vivo* bioequivalence studies to assess equivalence between the test and reference formulations for postapproval changes. For the comparison of dissolution profiles, the US FDA guidance suggests considering the assessment of (1) the overall profile similarity and (2) similarity at each sampling time point.^[43] Since dissolution profiles are curves over time, Chow and Ki^[44] introduced the concepts of local similarity and global similarity. Two dissolution profiles are said to be locally similar at a given time point if their difference or ratio at the given time point is within some equivalence (similarity) limits, denoted by (δ_L, δ_U) . Two dissolution profiles are considered globally similar if their differences or ratios are within (δ_L, δ_U) across all time points. Note that global similarity is also known as *uniformly similar*. Chow and Ki^[44] suggested

the following similarity limits for comparing dissolution profiles:

$$\delta_L = \frac{Q - \delta}{Q + \delta} \text{ and } \delta_U = \frac{Q + \delta}{Q - \delta},$$

where Q is the desired mean dissolution rate of a drug product as specified in the United States Pharmacopeia and National Formulary individual monograph, and δ is a meaningful difference of scientific importance in mean dissolution profiles of two drug products under consideration. In practice, δ is usually determined by a pharmaceutical scientist.

In order to achieve these two objectives, based on Moore and Flanner,^[45] both the FDA Scale-Up and Post-Approval Change (SUPAC) guidance^[46] and the guidance on dissolution testing^[43] suggest the similarity and difference factors for the assessment of similarity. The similarity factor is then defined as the logarithmic reciprocal square root transformation of 1 plus the mean-squared (the average sum of squares) difference in mean cumulative percentage dissolved between the test and reference formulations over all sampling time points. That is,

$$f_2 = 50 \log \left\{ \left[1 + \frac{Q}{n} \right]^{-0.5} 100 \right\} \quad (4)$$

where

$$Q = \sum_{t=1}^n (\mu_{Rt} - \mu_{Tt})^2,$$

and $\log$ denotes the logarithm base 10.

On the other hand, the difference factor is the sum of the absolute difference in mean cumulative percentage dissolved between the test and reference formulations divided by the sum of the mean cumulative dissolved of the reference formulation.

$$f_1 = \frac{\sum_{t=1}^n |\mu_{Rt} - \mu_{Tt}|}{\sum_{t=1}^n \mu_{Rt}} \quad (5)$$

It should be noted that the definitions of f_1 and f_2 provided by Moore and Flanner,^[45] and in the SUPAC and guidance on dissolution testing are not clear whether they are defined based on the population means or the sample averages. However, following the traditional statistical inference with the ability for evaluation of error probability, we define both f_1 and f_2 based on the population mean dissolution rates. It follows that f_1 and f_2 are the population parameters for the assessment of similarity of dissolution profiles between the test and reference formulations.

The use of the f_2 similarity factor has been discussed and criticized by many researchers.^[47–49] Chow and Shao^[50] pointed out two main problems in using the f_2 similarity factor for assessing similarity between the dissolution profiles of two drug products. The first problem is its lack of statistical justification. Since f_2 is a statistic and, thus, a random variable, $P(f_2 > 50)$ may be quite large when the two dissolution profiles are not similar. However, $P(f_2 > 50)$ can be very small when the two dissolution profiles are similar. Suppose the expected value $E(f_2)$ exists and that we can find a 95% lower confidence bound for $E(f_2)$. Then, a reasonable modification to the f_2 similarity factor approach is to replace f_2 with the 95% lower confidence bound for $E(f_2)$. The second problem with using the f_2 similarity factor is that the f_2 similarity factor assesses neither local similarity nor global similarity, owing to the use of the average of the dissolution data.

STATISTICAL METHODS

For the assessment of biosimilarity of biosimilar products, standard methods for the assessment of bioequivalence for small-molecule drug products are usually considered. These methods include the classical methods such as the confidence interval approach and interval hypothesis testing such as Schuirmann's two one-sided tests procedure, Bayesian methods, and nonparametric methods such as Wilcoxon–Mann–Whitney two one-sided tests procedure. It should be noted that these methods were derived based on a raw data model under a crossover design although they can be easily extended to a parallel group design based on log-transformed data. In practice, it is a concern whether these methods are appropriate for the assessment of biosimilarity of biosimilar products due to fundamental differences between small-molecule drug products and large-molecule biological drug products as described in earlier articles.^[51] As a result, the search for biosimilarity criteria, study endpoints (measures of biosimilarity), and statistical methods has become the center of attention for the assessment of biosimilarity of biosimilar products within regulatory agencies.

Interval Hypotheses

In practice, one of the most widely used designs for assessing biosimilarity between biosimilar products and an innovator biological product is probably either a two-sequence, two-period (2×2) crossover design or a two-arm parallel group design. Under a valid study design, biosimilarity can then be assessed by means of an equivalence test under the following interval hypotheses:

$$H_0: \mu_T - \mu_R \leq \theta_L \text{ or } \mu_T - \mu_R \geq \theta_U \text{ vs. } H_a: \theta_L < \mu_T - \mu_R < \theta_U, \quad (6)$$

where (θ_L, θ_U) are the prespecified equivalence limits (margins), and μ_T and μ_R are the population means of a biological (test) product and an innovator biological (reference) product, respectively. That is, biosimilarity is assessed in terms of the *absolute* difference between the two population means. Alternatively, biosimilarity can be assessed in terms of the *relative* difference (i.e., ratio) between the population means. Note that for the assessment of similarity between small-molecule drug products, average bioequivalence, population bioequivalence, and individual bioequivalence are defined in terms of ratios of appropriate parameters under a crossover design. In practice, since many biological products have long half-lives, a crossover design may not be appropriate in these cases for the assessment of biosimilarity. Instead, a parallel group design is considered more appropriate.

The purpose of this article is to provide a comprehensive summarization of standard statistical methods that are commonly used for the assessment of average bioequivalence for small-molecule drug products under a crossover design. In addition, we will focus on the discussion of statistical methods proposed by Kang and Chow^[52] under a newly proposed three-arm parallel design for investigation of biosimilarity of biosimilar products. The statistical analysis methods proposed by Kang and Chow^[52] consider the relative distance based on the absolute mean differences. Under the three-arm design, patients who are randomly assigned to the first group receive a biosimilar (test) product, while patients who are randomly assigned to the second and third groups receive the innovator biological (reference) product from different batches. The distance between the test product and the reference product is defined by the absolute mean difference between two products. Similarly, the distance between the reference products from two different batches is defined. The relative distance is defined as the ratio of the two distances whose denominator is the distance between the two reference products from different batches. Under the proposed design, Kang and Chow^[52] claim that the two products are biosimilar if the relative distance is less than a prespecified margin.

Classic Methods for Assessing Biosimilarity

Under a crossover design, consider the following statistical model for raw data:

$$Y_{ijk} = \mu + S_{ik} + P_j + F_{j,k} + C_{(j-1,k)} + e_{ijk}, \quad (7)$$

where Y_{ijk} is the response (e.g., AUC) of the i th subject in the k th sequence at the j th period; μ is the overall mean; S_{ik} is the random effect of the i th subject in the k th sequence, where $i = 1, 2, \dots, g$; P_j is the fixed effect of the j th period, where $j = 1, \dots, p$ and $\sum_j P_j = 0$; $F_{(j,k)}$ is the direct fixed effect of the drug product in the k th sequence which is

administered at the j th period, and $\Sigma F_{(j,k)} = 0$; $C_{(j-1,k)}$ is the fixed first-order carryover effect of the drug product in the k th sequence which is administered at the $(j-1)$ th period, where $C_{(0,k)} = 0$ and $\Sigma C_{(j-1,k)} = 0$; and e_{ijk} is the (within-subject) random error in observing Y_{ijk} .

It is assumed that $\{S_{ik}\}$ are independently and identically distributed (i.i.d.) with mean 0, and variance σ_s^2 and $\{e_{ijk}\}$ are independently distributed with mean 0 and variances σ_t^2 where $t = 1, 2, \dots, L$ (the number of formulations to be compared). $\{S_{ik}\}$ and $\{e_{ijk}\}$ are assumed mutually independent. The estimate of σ_s^2 is usually used to explain the intersubject variability, and the estimates of σ_t^2 are used to assess the intra-subject variabilities for the t th drug product.

Confidence Interval Approach

For bioequivalence assessment of small-molecule drug products, the FDA adopts the 80/125 rule based on log-transformed data. The 80/125 rule states that bioequivalence is concluded if the GMR between the test product and the reference product is within the bioequivalence limits of (80.00%, 125.00%), with a certain statistical assurance. Thus, a typical approach is to consider the method of classic (shortest) confidence interval.

Consider a standard two-sequence, two-period cross-over design, under model (7), after the log transformation of the data, let $\bar{Y}_T$ and $\bar{Y}_R$ be the respective least squares means for the test and reference formulations that can be obtained from the sequence-by-period means. The classic (or shortest) $(1 - 2\alpha) \times 100\%$ confidence interval can then be obtained based on the following t statistic:

$$T = \frac{(\bar{Y}_T - \bar{Y}_R) - (\mu_T - \mu_R)}{\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad (8)$$

where n_1 and n_2 are the numbers of subjects in sequences 1 and 2, respectively, and $\hat{\sigma}_d$ is an estimate of the variance of the period differences for each subject within each sequence, which are defined as follows:

$$d_{ik} = \frac{1}{2}(Y_{i2k} - Y_{i1k}), \quad i = 1, 2, \dots, n_k; \quad k = 1, 2.$$

Thus, $V(d_{ik}) = \sigma_d^2 = \sigma_e^2/2$. Under normality assumptions, T follows a central student t distribution with degrees of freedom $n_1 + n_2 - 2$. Thus, the classic $(1 - 2\alpha) \times 100\%$ confidence interval for $\mu_T - \mu_R$ can be obtained as follows:

$$\begin{aligned} L_1 &= (\bar{Y}_T - \bar{Y}_R) - t(\alpha, n_1 + n_2 - 2) \hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \\ U_1 &= (\bar{Y}_T - \bar{Y}_R) + t(\alpha, n_1 + n_2 - 2) \hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \end{aligned} \quad (9)$$

The above $(1 - 2\alpha) \times 100\%$ confidence interval for $\log(\mu_T) - \log(\mu_R) = \log(\mu_T/\mu_R)$ can be converted into a $(1 - 2\alpha) \times 100\%$ confidence interval for μ_T/μ_R by taking an anti-log transformation.

Note that under a parallel group design, a $(1 - 2\alpha) \times 100\%$ confidence interval for μ_T/μ_R can be similarly obtained.

Schirmann's Two One-Sided Tests Procedure

The assessment of average bioequivalence is based on the comparison of bioavailability profiles between the two products. However, in practice, it is recognized that no two drug products will have exactly the same bioavailability profiles. Therefore, if the profiles of the two drug products differ by less than a (clinically) meaningful limit, the profiles of the two drug products may be considered equivalent. Following this concept, Schirmann^[53] first introduced the use of interval hypotheses (6) for assessing average bioequivalence. The concept of interval hypotheses (6) is to show average bioequivalence by rejecting the null hypothesis of average bioinequivalence. In most bioavailability and bioequivalence studies, δ_L and δ_U are often chosen to be $-\theta_L = \theta_U = 20\%$ of the reference mean (μ_R). When the natural logarithmic transformation of the data is considered, the hypotheses corresponding to hypotheses (6) can be stated as

$$H'_0: \mu_T/\mu_R \leq \delta_L \quad \text{or} \quad \mu_T/\mu_R \geq \delta_U \quad \text{vs.} \quad H'_a: \delta_L < \mu_T/\mu_R < \delta_U \quad (10)$$

where $\delta_L = \exp(\theta_L)$ and $\delta_U = \exp(\theta_U)$. Note that the FDA recommends that $(\delta_L, \delta_U) = (80.00\%, 125.00\%)$ for assessing average bioequivalence.

Note that the test for hypotheses in Eq. (10) formulated on the log scale is equivalent to testing for hypotheses (6) on the raw scale. The interval hypotheses (6) can be decomposed into two sets of one-sided hypotheses:

$$\begin{aligned} H_{01}: \mu_T - \mu_R \leq \theta_L \quad \text{vs.} \quad H_{a1}: \mu_T - \mu_R > \theta_L \\ \text{and} \\ H_{02}: \mu_T - \mu_R \geq \theta_U \quad \text{vs.} \quad H_{a2}: \mu_T - \mu_R < \theta_U \end{aligned} \quad (11)$$

The first set of hypotheses is to verify that the average bioavailability of the test formulation is not too low, whereas the second set of hypotheses is to verify that the average bioavailability of the test formulation is not too high. A relatively low (or high) average bioavailability may refer to the concern of efficacy (or safety) of the test formulation. If one concludes that $\theta_L < \mu_T - \mu_R$ (i.e., reject H_{01}) and $\mu_T - \mu_R < \theta_U$ (i.e., reject H_{02}), then it has been concluded that

$$\theta_L < \mu_T - \mu_R < \theta_U.$$

Thus, μ_T and μ_R are equivalent. The rejection of H_{01} and H_{02} , which leads to the conclusion of average bioequivalence, is equivalent to the rejection of H_0 in Eq. (6).

Under hypotheses (6), Schuirmann^[54] introduced the two one-sided tests procedure for assessing the average bioequivalence between drug products. The proposed two one-sided tests procedure suggests the conclusion of equivalence of μ_T and μ_R at the α level of significance if, and only if, H_{01} and H_{02} in Eq. (11) are rejected at a predetermined α -level of significance. Under normality assumptions, the two sets of one-sided hypotheses can be tested with ordinary one-sided t -tests. We conclude that μ_T and μ_R are average equivalent if

$$T_L = \frac{(\bar{Y}_T - \bar{Y}_R) - \theta_L}{\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t(\alpha, n_1 + n_2 - 2)$$

and

$$T_U = \frac{(\bar{Y}_T - \bar{Y}_R) - \theta_U}{\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -t(\alpha, n_1 + n_2 - 2). \quad (12)$$

The two one-sided t -tests procedure is operationally equivalent to the classic (shortest) confidence interval approach; that is, both the classic confidence interval approach and Schuirmann's two one-sided tests procedure will lead to the same conclusion on bioequivalence.

Note that under a parallel group design, Schuirmann's two one-sided tests procedure can be similarly derived with a slightly modification from a paired t -test statistic to a two-sample t -test statistic.

Bayesian Methods

In previous sections, statistical methods for the assessment of biosimilarity were derived based on the sampling distribution of the estimate of the parameter of interest, such as the direct drug effect (i.e., $\theta = \mu_T - \mu_R$), which is assumed to be fixed but unknown. Although statistical inference (e.g., confidence interval and interval hypothesis testing) on the unknown direct drug effect can be drawn from the sampling distribution of the estimate, there is little information on the probability of the unknown direct drug effect being within the equivalence limits (θ_L, θ_U). To have a certain assurance on the probability of the direct drug effect being within (θ_L, θ_U), a Bayesian approach,^[55] which assumes that the unknown direct drug effect is a random variable and follows a prior distribution, is useful.

In practice, before a biosimilar study is conducted, investigators may have some prior knowledge of the drug product under development. As an example, for a PK biosimilar study, according to past experiments, the investigator may have some information on (1) the intersubject and the intra-subject variabilities and (2) the ranges of AUC or C_{max} for the test and reference products. This information can be used to choose an appropriate prior distribution

of the unknown direct drug effect. An appropriate prior distribution can reflect the investigator's belief about the drug products under study. After the study is completed, the observed data can be used to adjust the prior distribution of the direct drug effect, which is called the posterior distribution. Given the posterior distribution, a probability statement on the direct drug effect being within the biosimilarity limits can be made.

A different prior distribution can lead to a different posterior distribution that has an influence on the statistical inference on the direct drug effect. Thus, an important issue in a Bayesian approach is how to choose a prior distribution. Box and Tiao^[55] introduced the use of a locally uniform distribution over a possible range of AUC or C_{max} as a non-informative prior distribution. A non-informative prior distribution assumes that there is an equally likely chance for any two points within the possible range being the true state of the location of the direct drug effect. In this case, the resultant posterior distribution can be used to provide the true state of the location of a direct drug effect. In practice, however, it is also desirable to provide an interval showing a range in which most of the distribution of a direct drug effect will fall. We shall refer to such an interval as a highest posterior density (HPD) interval. The HPD interval is also known as a credible interval^[56] or a Bayesian confidence interval.^[57] An HPD interval possesses the following properties^[55]: (1) the density for every point inside the interval is greater than that for every point outside the interval, and (2) for a given probability distribution, the interval is the shortest. It can be verified that the above two properties imply each other.

In what follows, for illustration purposes, the Bayesian method proposed by Rodda and Davis^[58] is discussed under the following model, which assumes that there are no carryover effects because a washout period of sufficient length can be chosen to completely eliminate the residual effects from one dosing period to the next:

$$Y_{ijk} = \mu + S_{ik} + F_{(j,k)} + P_j + e_{ijk}, \quad (13)$$

where Y_{ijk} , μ , S_{ik} , $F_{(j,k)}$, P_j , and e_{ijk} were defined in Eq. (7). Given the results of a biosimilar study, Rodda and Davis^[58] proposed a Bayesian evaluation to estimate the probability of a clinically important difference (i.e., the probability that the true direct drug effect will fall within the bioequivalent limits is estimated). Under the assumption of normality and equal carryover effects, $\bar{d}_1$, $\bar{d}_2$ and $(n_1 + n_2 - 2)\hat{\sigma}_d^2$ are independently distributed as $N(\theta_1, \sigma_d^2/n_1)$, $N(\theta_2, \sigma_d^2/n_2)$, and $\sigma_d^2 \chi^2(n_1 + n_2 - 2)$, where

$$\bar{d}_{.k} = \frac{1}{n_k} \sum_{i=1}^{n_k} d_{ik}, \quad k = 1, 2$$

$$\theta_1 = \frac{1}{2}[(P_2 - P_1) + (F_T - F_R)],$$

$$\theta_2 = \frac{1}{2}[(P_2 - P_1) + (F_R - F_T)].$$

$$\begin{aligned}\text{Note that } F &= \theta_1 - \theta_2 = (\mu + F_T) - (\mu + F_R) \\ &= \mu_T - \mu_R\end{aligned}$$

Assuming that the non-informative prior distribution for θ_1 , θ_2 , and $\log(\sigma_d)$ is approximately independent and locally uniformly distributed, then the joint posterior distribution of θ_1 , θ_2 , and σ_d^2 , given data $Y = \{Y_{ijk}, i = 1, 2, \dots, n_k; j, k = 1, 2\}$, is

$$p(\theta_1, \theta_2, \sigma_d^2 | \mathbf{Y}) = p(\theta_1 | \sigma_d^2, \bar{d}_1) p(\theta_2 | \sigma_d^2, \bar{d}_2) p(\sigma_d^2 | \hat{\sigma}_d^2), \quad (14)$$

$$\begin{aligned}p(\theta_i | \sigma_d^2, \bar{d}_i) &= N(\bar{d}_i, \hat{\sigma}_d^2/n_i), \quad i = 1, 2, \\ p(\sigma_d^2 | \hat{\sigma}_d^2) &= (n_1 + n_2) \hat{\sigma}_d^2 \chi^2(n_1 + n_2 - 2).\end{aligned}$$

where $\chi^2(n_1 + n_2 - 2)$ is the distribution of the inverse of $\chi^2(n_1 + n_2 - 2)$. Therefore, the joint distribution of $\mu_T - \mu_R (=F)$ and σ_d^2 is given by

$$p(\mu_T - \mu_R, \sigma_d^2 | \mathbf{Y}) = p(\mu_T - \mu_R | \sigma_d^2, \bar{d}_1 - \bar{d}_2) p(\sigma_d^2 | \hat{\sigma}_d^2), \quad (15)$$

where

$$\begin{aligned}p(\mu_T - \mu_R | \sigma_d^2, \bar{d}_1 - \bar{d}_2) &= N\left[\bar{d}_1 - \bar{d}_2, \hat{\sigma}_d^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right] \\ &= N\left[\bar{Y}_T - \bar{Y}_R, \hat{\sigma}_d^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right].\end{aligned}$$

The marginal posterior distribution of F , given data $\mathbf{Y}$, is

$$\begin{aligned}p(\mu_T - \mu_R | \mathbf{Y}) &= \frac{(\hat{\sigma}_d^2 m)^{-1/2}}{B(1/2, v/2) \sqrt{n}} \\ &\quad \left\{ 1 + \frac{[(\mu_T - \mu_R) - (\bar{Y}_T - \bar{Y}_R)]^2}{v \hat{\sigma}_d^2 m} \right\}^{-(v+1)/2}, \quad (16)\end{aligned}$$

where $m = \frac{1}{n_1} + \frac{1}{n_2}$, $v = n_1 + n_2 - 2$, and $-\infty < \mu_T - \mu_R < \infty$.

Thus, we have

$$T_{RD} = \frac{(\mu_T - \mu_R) - (\bar{Y}_T - \bar{Y}_R)}{\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (17)$$

which has a central Student's t -distribution with $n_1 + n_2 - 2$ degrees of freedom. From Eq. (17), the probability of F being within the biosimilarity limits of (θ_L, θ_U) can be estimated by

$$\begin{aligned}P_{RD} &= P\{\theta_L < \mu_T - \mu_R < \theta_U\} \\ &= F_t(t_U) - F_t(t_L),\end{aligned} \quad (18)$$

where F_t is the cumulative distribution function of a central t variable with $n_1 + n_2 - 2$ degrees of freedom, and

$$\begin{aligned}t_U &= \frac{\theta_U - (\bar{Y}_T - \bar{Y}_R)}{\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ \text{and} \\ t_L &= \frac{\theta_L - (\bar{Y}_T - \bar{Y}_R)}{\hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.\end{aligned} \quad (19)$$

The lower and upper limits of the $(1 - 2\alpha) \times 100\%$ HPD interval are, respectively, given by

$$\begin{aligned}L_{RD} &= (\bar{Y}_T - \bar{Y}_R) - t(\alpha, n_1 + n_2 - 2) \hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \\ U_{RD} &= (\bar{Y}_T - \bar{Y}_R) + t(\alpha, n_1 + n_2 - 2) \hat{\sigma}_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.\end{aligned} \quad (20)$$

Hence, it is verified that the $(1 - 2\alpha) \times 100\%$ HPD interval in Eq. (20) is numerically equivalent to the $(1 - 2\alpha) \times 100\%$ classic confidence interval obtained from the sampling theory. However, the interpretation of these two intervals is totally different. For example, a 90% classic confidence interval for F indicates that, in the long run, if the study is repeatedly carried out for numerous times, 90% of the times the interval will contain the unknown direct drug effect $\mu_T - \mu_R$. On the other hand, based on the posterior distribution of $\mu_T - \mu_R$, the chance of $\mu_T - \mu_R$ being within the lower and upper limits of a 90% HPD interval is 90%.

Wilcoxon–Mann–Whitney Two One-Sided Tests Procedure

As described in the previous sections, statistical methods for assessing average biosimilarity between drug products were derived under the assumption that $\{S_{ik}\}$ and $\{e_{ijk}\}$ are mutually independent and normally distributed with mean 0 and variance σ_s^2 and σ_e^2 . Under these normality assumptions, confidence intervals and tests for interval hypotheses were obtained based on either a two-sample t -statistic or an F -statistic. In practice, however, one of the difficulties commonly encountered in comparing drug products is whether the assumption of normality (for raw or untransformed data) or log-normality (for log-transformed data) is valid. If the normality (or log-normality) is seriously violated, the approach based on a two-sample t -statistic or an F -statistic is no longer justified. In this situation, a distribution-free (or nonparametric) method is useful. In this section, a nonparametric version of the two one-sided tests procedure for testing interval hypotheses, namely, Wilcoxon–Mann–Whitney two one-sided tests procedure, is discussed. The Hodges–Lehmann estimator associated with the Wilcoxon rank sum test will be

used to construct a $(1 - 2\alpha) \times 100\%$ confidence interval for $\mu_T - \mu_R$, the difference in average biosimilarity.

Under the standard 2×2 crossover design consisting of a pair of dual sequences (i.e., RT and TR), a distribution-free rank sum test can be applied directly to the two one-sided tests procedure.^[59,60] We shall refer to this approach as Wilcoxon–Mann–Whitney two one-sided tests procedure. Let $\theta = \mu_T - \mu_R$. The two sets of hypotheses in Eq. (6) can then be rewritten as

$$H_{01} : \theta_L^* \leq 0 \quad \text{vs.} \quad H_{a1} : \theta_L^* > 0$$

and

$$H_{02} : \theta_U^* \geq 0 \quad \text{vs.} \quad H_{a2} : \theta_U^* < 0,$$

where $\theta_L^* = \theta - \theta_L$ and $\theta_U^* = \theta - \theta_U$.

Thus, the estimates of θ_L^* and θ_U^* can be obtained as a linear function of period differences d_{ik} , $i = 1, 2, \dots, n_k$, $k = 1, 2$. Let

$$b_{hik} = \begin{cases} d_{ik} - \theta_h & h = L, U, \text{ for subjects in sequence 1} \\ d_{ik} & \text{for subjects in sequence 2.} \end{cases} \quad (21)$$

When there are no carryover effects, the expected value and variance of b_{hik} , where $h = L, U$, $i = 1, 2, \dots, n_k$, and $k = 1, 2$, are given by

$$E(b_{hik}) = \begin{cases} \frac{1}{2}[(P_2 - P_1) + (\theta - 2\theta_h)] & \text{for } k = 1 \\ \frac{1}{2}[(P_2 - P_1) - \theta] & \text{for } k = 2 \end{cases} \quad (22)$$

and

$$V(b_{hik}) = V(d_{ik}) = \sigma_d^2 = \sigma_e^2/2.$$

It can be seen that $E(b_{hi1}) - E(b_{hi2}) = (\theta - \theta_h) = \theta_h^*$.

Thus, for a fixed h , $\{b_{hi1}\}$ and $\{b_{hi2}\}$ have the same distribution except for the difference $(=\theta_h^*)$ in location of the true formulation effect. Here, the Wilcoxon–Mann–Whitney rank sum test^[61,62] for the unpaired two-sample location problem can be directly applied to test each of the two sets of hypotheses given above. Consider the first set of hypotheses that

$$H_{01} : \theta_L^* \leq 0 \quad \text{vs.} \quad H_{a1} : \theta_L^* > 0.$$

The Wilcoxon–Mann–Whitney test statistic can be derived based on $\{b_{Li1}\}$, $i = 1, 2, \dots, n_1$ and $\{b_{Li2}\}$, $i = 1, 2, \dots, n_2$. Let $R(b_{Lik})$ be the rank of b_{Lik} in the combined sample $\{b_{Lik}\}$, $i = 1, 2, \dots, n_k$, $k = 1, 2$. Also, let R_L be the sum of the ranks of the responses for subjects in sequence 1; that is,

$$R_L = \sum_{i=1}^{n_1} R(b_{Li1}).$$

Thus, the Wilcoxon–Mann–Whitney test statistic for H_{01} is given by

$$W_L = R_L - \frac{n_1(n_1 + 1)}{2}.$$

We then reject H_{01} if

$$W_L > w(1 - \alpha). \quad (23)$$

where $w(1 - \alpha)$ is the $(1 - \alpha)$ th quantile of the distribution of W_L . Similarly, for the second set of hypotheses that

$$H_{02} : \theta_U^* \geq 0 \quad \text{vs.} \quad H_{a2} : \theta_U^* < 0,$$

we reject H_{02} if

$$W_U = R_U - \frac{n_1(n_1 + 1)}{2} < w(\alpha), \quad (24)$$

where R_U is the sum of the ranks of $\{b_{Uik}\}$ for subjects in the first sequence. Hence, average bioequivalence is concluded if both H_{01} and H_{02} are rejected; that is,

$$W_L > w(1 - \alpha) \quad \text{and} \quad W_U < w(\alpha). \quad (25)$$

The expected values and variances for W_L and W_U under the null hypotheses H_{01} and H_{02} , when there are no ties, are given by

$$\begin{aligned} E(W_L) &= E(W_U) = \frac{n_1 n_2}{2}, \\ V(W_L) &= V(W_U) = \frac{1}{12} n_1 n_2 (n_1 + n_2 + 1). \end{aligned} \quad (26)$$

When there are ties among observations, average ranks can be assigned to compute W_L and W_U . In this case, however, the expected values and variances of W_L and W_U become

$$\begin{aligned} E(W_L) &= E(W_U) = \frac{n_1 n_2}{2}, \\ V(W_L) &= V(W_U) = \frac{1}{12} n_1 n_2 (n_1 + n_2 + 1 - Q), \end{aligned} \quad (27)$$

where

$$Q = \frac{1}{(n_1 + n_2)(n_1 + n_2 - 1)} \sum_{v=1}^q (r_v^3 - r_v),$$

where q is the number of tied groups and r_v is the size of the tied group v . Note that if there are no tied observations, $q = n_1 + n_2$, $r_v = 1$ for $v = 1, 2, \dots, n$, and $Q = 0$, then Eq. (27) is reduced to Eq. (26). Since W_L and W_U are symmetric about their mean $(n_1 n_2)/2$, we have

$$w(1 - \alpha) = n_1 n_2 - w(\alpha). \quad (28)$$

When $n_1 + n_2$, the total number of subjects, is large (say $n_1 + n_2 > 40$) and the ratio of n_1 and n_2 is close to 0.5, a large sample approximation using the standard normal distribution can be used to approximate for average biosimilarity testing; that is, we may conclude bioequivalence if

$$Z_L > z(\alpha) \text{ and } Z_U < -z(\alpha),$$

where $z(\alpha)$ is the α th quantile of a standard normal distribution and

$$Z_L = \frac{W_L - E(W_L)}{\sqrt{V(W_L)}} = \frac{R_L - \left[\frac{n_1(n_1 + n_2 + 1)}{2} \right]}{\sqrt{\frac{1}{12} n_1 n_2 (n_1 + n_2 + 1)}},$$

$$Z_U = \frac{W_U - E(W_U)}{\sqrt{V(W_U)}} = \frac{R_U - \left[\frac{n_1(n_1 + n_2 + 1)}{2} \right]}{\sqrt{\frac{1}{12} n_1 n_2 (n_1 + n_2 + 1)}}. \quad (29)$$

Note that the variances in Z_L and Z_U should be replaced with that given in Eq. (26) if these are ties.

POWER CALCULATION FOR SAMPLE SIZE

Since power analysis for sample size calculation for bio-similar studies is similar to that for bioequivalence studies, in this section, we will focus on power analysis for sample size calculation for bioequivalence studies. The statistical procedures and/or formulas for bioequivalence assessment for generic drug products can be directly applied to biosimilarity assessment for biosimilar products.

For bioequivalence trials, a traditional approach for sample size determination is to conduct a power analysis based on the 80/20 decision rule. This approach, however, is based on point hypotheses rather than interval hypotheses and therefore may not be statistically valid. Phillips^[63] provided a table of sample sizes based on power calculations of Schuirmann's two one-sided tests procedure using the bivariate noncentral t -distribution. However, no formulas are provided. An approximate formula for sample-size calculations was provided in the work of Liu and Chow.^[64] Under a standard 2×2 crossover design, assuming that $n_1 = n_2 = n$ and $\theta = \mu_T - \mu_R = 0$, the sample size required for achieving $1 - \beta$ power at the α level of significance under interval hypotheses testing (or Schuirmann's two one-sided tests procedure) can be obtained by solving the following equation:

$$n \geq 2 \left[t(\alpha, 2n - 2) + t(\beta/2, 2n - 2) \right]^2 \left(\frac{\hat{\sigma}_d}{\Delta} \right)^2,$$

where $\Delta = 0.2\mu_R$. If the ± 20 rule is used, then the above equation becomes

$$n \geq \left[t(\alpha, 2n - 2) + t(\beta/2, 2n - 2) \right]^2 \left(\frac{CV}{20} \right)^2$$

where $CV = 100 \times \sqrt{2\hat{\sigma}_d^2} / \mu_R$.

In the case where $\theta = \theta_0 \neq 0$, the sample size required for achieving $1 - \beta$ power at the α level of significance under Schuirmann's two one-sided tests procedure can be obtained by solving the following equation:

$$n(\theta_0) \geq 2 \left[t(\alpha, 2n - 2) + t(\beta/2, 2n - 2) \right]^2 \left(\frac{\hat{\sigma}_d}{\Delta - \theta_0} \right)^2,$$

where $\Delta = 0.2\mu_R$. If the ± 20 rule is used, then the above equation becomes

$$n(\theta_0) \geq \left[t(\alpha, 2n - 2) + t(\beta/2, 2n - 2) \right]^2 \left(\frac{CV}{20 - \theta'_0} \right)^2$$

where $\theta'_0 = 100 \times \theta_0 / \mu_R$.

For illustration purpose, Table 4 gives the total sample sizes needed to achieve a desired power for a standard crossover design for various combinations of $\theta = \mu_T - \mu_R$ and CV, where

$$CV = 100 \times \sqrt{2\hat{\sigma}_d^2} / \mu_R.$$

Table 4 Sample size requirements

Power (%)	CV (%)	$\theta = \mu_T - \mu_R$		
		0%	5%	10%
80	10	8	8	16
	14	10	14	26
	18	16	20	42
	22	24	28	62
	26	32	40	86
	30	40	52	114
	34	52	66	146
	38	64	82	180
	40	70	90	200
	40	70	90	200
90	10	10	10	20
	14	14	18	36
	18	20	28	58
	22	28	40	86
	26	40	54	118
	30	52	70	156
	34	66	90	200
	38	80	112	250
	40	90	124	276
	40	90	124	276

In practice, biosimilar studies often utilize higher order crossover designs such as Balaam's design, the two-sequence dual design, and four-period crossover designs with either two sequences or four sequences. A higher order crossover design is defined as a crossover design either whose number of sequences or whose number of periods is larger than the number of treatments to be compared. Sample size requirements for a higher order crossover design can be similarly obtained. Consider the following commonly used high-order crossover designs:

Design 1: Balaam's design—(TT, RR, RT, TR)

Design 2: Two-sequence, three-period dual design—(TRR, RTT)

Design 3: Four-period design with two sequences—(TRRT, RTTR)

Design 4: Four-period design with four sequences—(TTRR, RRTT, TRTR, RTTR)

Let n_i denote the number of subjects in each sequence i having the same value n , and F_v denote the cumulative distribution function of the t distribution with v degrees of freedom. Then the power function, $P_k(\theta)$, of Schuirman's two one-sided tests at the α nominal level for design k is given as follows^[65]:

$$P_k(\theta) = F_{v_k} \left(\left[\frac{\Delta - \theta}{CV \sqrt{\frac{b_i}{n}}} \right] - t(\alpha, v_k) \right) - F_{v_k} \left(t(\alpha, v_k) - \left[\frac{\Delta + \theta}{CV \sqrt{\frac{b_i}{n}}} \right] \right), \text{ for } k = 1, 2, 3, 4,$$

where $v_1 = 4n - 3$, $v_2 = 4n - 4$, $v_3 = 6n - 5$, $v_4 = 12n - 5$, $b_1 = 2$, $b_2 = 3/4$, $b_3 = 11/20$, $b_4 = 1/4$. Similarly, the sample size n required to achieve a $1 - \beta$ power at the α nominal level for each corresponding design k after the logarithmic transformation is determined by the following equations:

$$n \geq b_k \left[t(\alpha, v_k) + t\left(\frac{\beta}{2}, v_k\right) \right]^2 [CV/\ln(1.25)]^2 \text{ if } \delta = 1,$$

$$n \geq b_k \left[t(\alpha, v_k) + t(\beta, v_k) \right]^2 [CV/(\ln(1.25) - \ln \delta)]^2 \text{ if } 1 < \delta < 1.25,$$

and

$$n \geq b_k \left[t(\alpha, v_k) + t(\beta, v_k) \right]^2 [CV/(\Delta - \theta)]^2 \text{ if } 0.8 < \delta < 1.$$

where $\ln$ denotes the natural logarithm and $CV = \sqrt{\exp(\sigma^2) - 1}$, the coefficient of variation in the

multiplicative model, and σ^2 is the residual (within-subject) variance on the log scale. However, because the degrees of freedom are usually unknown, an easy way to find the sample size is to enumerate n .

UNIFIED APPROACH FOR ASSESSING BIOSIMILARITY

Chow^[66,67] proposed the development of a composite index for assessing the biosimilarity of follow-on biologics based on the facts that (1) the concept of biosimilarity for biologic products (made of living cells) is very different from that of bioequivalence for drug products, and (2) biologic products are very sensitive to small changes in the variation during the manufacturing process (i.e., it might have a drastic change in clinical outcome). Although some research on the comparison of moment-based criteria and probability-based criteria for the assessment of (1) average biosimilarity and (2) variability of biosimilarity for some given study endpoints by applying the criteria for bioequivalence is available in the literature (see, e.g., the works of Chow et al. and Hsieh et al.)^[68,69] universally acceptable criteria for biosimilarity are not available in the regulatory guidelines/guidances. Thus, Chow^[66] and Chow et al.^[70] proposed a biosimilarity index based on the concept of the probability of reproducibility as follows:

- Step 1. Assess the average biosimilarity between the test product and the reference product based on a given biosimilarity criterion. For illustration purposes, consider the bioequivalence criterion as a biosimilarity criterion. That is, biosimilarity is claimed if the 90% confidence interval of the ratio of means of a given study endpoint falls within the biosimilarity limits of (80.00%, 125.00%) or (−0.2231, 0.2231) based on raw (original) data or based on log-transformed data.
- Step 2. Once the product passes the test for biosimilarity in Step 1, calculate the reproducibility probability based on the observed ratio (or observed mean difference) and variability.^[71] Thus, the calculated reproducibility probability will take the variability and the sensitivity of heterogeneity in variances into consideration for the assessment of biosimilarity.
- Step 3. We then claim biosimilarity if the calculated 95% confidence lower bound of the reproducibility probability is larger than a prespecified number p_0 , which can be obtained based on an estimate of reproducibility probability for a study comparing a reference product to itself (the “reference product”). We will refer to such a study as an R–R study. We can then claim (local) biosimilarity if the 95% confidence lower bound of the biosimilarity index is larger than P_0 .

In an R–R study, define

$$P_{TR} = P \left(\begin{array}{l} \text{concluding average biosimilarity between the} \\ \text{test and reference products in a future trial,} \\ \text{given that the average biosimilarity based on} \\ \text{ABE criterion has been established in the} \\ \text{first trial} \end{array} \right) \quad (30)$$

A reproducibility probability for evaluating the biosimilarity of the same two reference products, based on the ABE criterion, is defined as follows:

$$P_{RR} = P \left(\begin{array}{l} \text{concluding average biosimilarity of the two} \\ \text{same reference products in a future trial, given} \\ \text{that the average biosimilarity based on ABE} \\ \text{criterion has been established in first trial} \end{array} \right) \quad (31)$$

Since the idea of the biosimilarity index is to show that the reproducibility probability for comparing “a reference product” with “the reference product” is higher in a study than the comparison of a follow-on biologic with the innovative (reference) product, the criterion of an acceptable reproducibility probability (i.e., p_0) for the assessment of biosimilarity can be obtained based on the R–R study. For example, if the R–R study suggests the reproducibility probability of 90%, i.e., $P_{RR} = 90\%$, the criterion of the reproducibility probability for bioequivalence study could be chosen as 80% of the 90% which is $p_0 = 80\% \times P_{RR} = 72\%$.

The biosimilarity index, described previously, has the advantages that (1) it is robust with respect to the selected study endpoint, biosimilarity criteria, and study design, and (2) the probability of reproducibility will reflect the sensitivity of heterogeneity in variance.

Note that the proposed biosimilarity index can be applied to different functional areas (domains) of biological products such as PK, biological activities, biomarkers (e.g., PD), immunogenicity, manufacturing process, and efficacy. An overall biosimilarity index or totality biosimilarity index across domains can be similarly obtained as follows:

Step 1. Obtain $\hat{p}_i$, the probability of reproducibility for the i -th domain, $i = 1, \dots, K$.

Step 2. Define the biosimilarity index $\hat{p} = \sum_{i=1}^K w_i \hat{p}_i$, where w_i is the weight for the i th domain.

Step 3. Claim global biosimilarity if we reject the null hypothesis that $p \leq p_0$, where p_0 is a prespecified acceptable reproducibility probability. Alternatively, we can claim (global) biosimilarity if the 95% confidence lower bound of p is larger than p_0 .

Let T and R be the parameters of interest (e.g., PK response) with means of μ_T and μ_R , for a test product and a reference product, respectively. Thus, the interval

hypotheses for testing the ABE of two products can be expressed as

$$H_0: \theta'_L \geq \frac{\mu'_T}{\mu'_R} \quad \text{or} \quad \theta'_U \leq \frac{\mu'_T}{\mu'_R} \quad \text{vs.} \quad H_a: \theta'_L < \frac{\mu'_T}{\mu'_R} < \theta'_U$$

where (θ'_L, θ'_U) is the ABE limit. For *in vivo* bioequivalence testing, (θ'_L, θ'_U) is chosen to be (80.00%, 125.00%). The above hypotheses can be reexpressed as

$$H_0: \theta_L \geq \mu_T - \mu_R \quad \text{or} \quad \theta_U \leq \mu_T - \mu_R \\ \text{vs.} \quad H_a: \theta_L < \mu_T - \mu_R < \theta_U$$

where μ_T and μ_R are the means of log-transformed data that are equal to the log-transformed values of μ'_T and μ'_R . (θ_L, θ_U) is $(-0.2231, 0.2231)$, which are equal to the log-transformed values of (80.00%, 125.00%). To calculate the reproducibility probability under the above interval hypotheses, the probability of P_{TR} can be expressed as the following when considering a parallel design, since it is a common design for biologic products:

$$P(\delta_L, \delta_U) = P \left(T_L(\bar{Y}_T, \bar{Y}_R, s_T, s_R) > t_{\alpha, dfp} \right. \\ \left. \text{and } T_U(\bar{Y}_T, \bar{Y}_R, s_T, s_R) < -t_{\alpha, dfp} \mid \delta_L, \delta_U \right) \quad (32)$$

where s_T , s_R , n_T , and n_R are the sample standard deviations and sample sizes for the test and reference formulations, respectively. The value of dfp can be calculated by

$$dfp = \frac{\left(\frac{s_T^2}{n_T} + \frac{s_R^2}{n_R} \right)^2}{\left(\frac{s_T^2}{n_T} \right)^2 + \left(\frac{s_R^2}{n_R} \right)^2}$$

$$T_L(\bar{Y}_T, \bar{Y}_R, \hat{\sigma}_T, \hat{\sigma}_R) = \frac{(\bar{Y}_T - \bar{Y}_R) - \theta_L}{\sqrt{\frac{s_T^2}{n_T} + \frac{s_R^2}{n_R}}}$$

$$T_U(\bar{Y}_T, \bar{Y}_R, \hat{\sigma}_T, \hat{\sigma}_R) = \frac{(\bar{Y}_T - \bar{Y}_R) - \theta_U}{\sqrt{\frac{s_T^2}{n_T} + \frac{s_R^2}{n_R}}}$$

$$\delta_L = \frac{\mu_T - \mu_R - \theta_L}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}} \quad \text{and} \quad \delta_U = \frac{\mu_T - \mu_R - \theta_U}{\sigma_d \sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}} \quad (33)$$

σ_T^2 and σ_R^2 are the variances for test and reference formulations, respectively.

The vectors (T_L, T_U) can be shown to follow a bivariate noncentral t -distribution with $n_1 + n_2 - 2$ and dfp degrees of freedom, correlation of 1, and noncentrality parameters

of δ_L and δ_U .^[72,63] Owen^[72] showed that the integral of the above bivariate noncentral t -distribution can be expressed as the difference of the integral between two univariate noncentral t -distributions. Therefore, the power function in Eq. (33) can be obtained by

$$P(\delta_L, \delta_U) = Q_f(t_U, \delta_U; 0, R) - Q_f(t_L, \delta_L; 0, R) \quad (34)$$

where

$$Q_f(t, \delta; 0, R) = \frac{\sqrt{2\pi}}{\Gamma(f/2)2^{(f-2)/2}} \int_0^R G(tx/\sqrt{f} - \delta) x^{f-1} G'(x) dx$$

$$R = (\delta_L - \delta_U) \sqrt{f} / (t_L - t_U), \quad G'(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

$$G(x) = \int_{-\infty}^x G'(t) dt$$

and

$$t_L = t_{\alpha, dfp}, \quad t_U = -t_{\alpha, dfp} \quad \text{and} \quad f = dfp \quad \text{for parallel design.}$$

Note that when $0 < \theta_U = -\theta_L$, $P(\delta_L, \delta_U) = P(-\delta_U, -\delta_L)$.

The reproducibility probabilities increase when the sample size increases and the ratio of means is close to 1, while it decreases when variability increases for the same setting of sample size and means ratio. This shows the impact of variability on reproducibility probabilities. Since true values of δ_L and δ_U are unknown, using the idea of replacing δ_L and δ_U in Eq. (34) by their estimates based on the sample from the first study, the estimated reproducibility probability can be obtained as

$$\hat{P}(\hat{\delta}_L, \hat{\delta}_U) = Q_f(t_U, \hat{\delta}_U; 0, \hat{R}) - Q_f(t_L, \hat{\delta}_L; 0, \hat{R}) \quad (35)$$

where

$$\hat{\delta}_L = \frac{\bar{Y}_T - \bar{Y}_L - \theta'_L}{\sqrt{\frac{s_T^2}{n_T} + \frac{s_R^2}{n_R}}}, \quad \hat{\delta}_U = \frac{\bar{Y}_T - \bar{Y}_L - \theta'_U}{\sqrt{\frac{s_T^2}{n_T} + \frac{s_R^2}{n_R}}},$$

$$\hat{R} = (\hat{\delta}_L - \hat{\delta}_U) \sqrt{f} / (t_L - t_U)$$

CONCLUDING REMARKS

Biologic products or medicines are therapeutic agents made of a living system or organisms. As a number of biologic products will be due to expire in the next few years, the potential opportunity in developing the follow-on products of these originator products may result in the reduction of these products and provide more choices to medical doctors and patients for getting the similar treatment care with lower cost. However, the price reductions versus the originator biologic products remain to be determined, as the advantage of a slightly cheaper price may be outweighed by the hypothetical increased risk of side effects from biosimilar molecules that are not the exact copies of their originators. Unlike traditional small-molecule drug

products, the characteristics and development of biologic products are more complicated and sensitive to many factors. Any small change in the manufacturing process may result in the change of therapeutic effect of the biologic products. The traditional bioequivalence criterion for average bioequivalence of small-molecule drug products may not be suitable for evaluation of the biosimilarity of biologic products. Therefore, in this chapter, we evaluate the biosimilarity index proposed by Chow et al.^[70] for the assessment of the (average) biosimilarity between the innovator and reference products. Results based on both the estimation and Bayesian approaches demonstrate that the proposed method based on the biosimilarity index can reflect the characteristics and impact of variability on the therapeutic effect of biologic products. However, the estimated reproducibility probability based on the Bayesian approach depends on the choice of the prior distributions. If a different prior such as an informative prior is used, a sensitivity analysis may be performed to evaluate the effects of different prior distributions.

The other advantage is that the proposed method can be applied to different functional areas (domains) of biological products such as PK, biological activities, biomarkers (e.g., PD), immunogenicity, manufacturing process, and efficacy, since it is developed based on the probability of reproducibility. Further research will be employed for the development of the statistical testing approach for the evaluation of biosimilarity across domains.

Current methods for the assessment of bioequivalence for drug products with identical active ingredients are not applicable to follow-on biologics due to fundamental differences. The assessment of biosimilarity between follow-on biologics and innovator in terms of surrogate endpoints (e.g., PK parameters and/or PD responses) or biomarkers (e.g., genomic markers) requires the establishment of the Fundamental Biosimilarity Assumption in order to bridge the surrogate endpoints and/or biomarker data to clinical safety and efficacy.

Unlike conventional drug products, follow-on biologics are very sensitive to small changes in variation during the manufacturing process, which have been shown to have an impact on the clinical outcome. Thus, it is a concern whether current criteria and regulatory requirements for the assessment of bioequivalence for drugs with small molecules can be applied also to the assessment of biosimilarity of follow-on biologics. It is suggested that current, existing criteria for the evaluation of bioequivalence, similarity, and biosimilarity be scientifically/statistically evaluated in order to choose the most appropriate approach for assessing biosimilarity of follow-on biologics. It is recommended that the selected biosimilarity criteria should be able to address (1) the sensitivity due to small variations in both location (bias) and scale (variability) parameters, and (2) the degree of similarity that can reflect the assurance for drug interchangeability.

Under the established Fundamental Biosimilarity Assumption and the selected biosimilarity criteria, it is

also recommended that appropriate statistical methods (e.g., comparing distributions and the development of biosimilarity index) be developed under valid study designs (e.g., Design A and Design B described earlier) for achieving the study objectives (e.g., the establishment of biosimilarity at specific domains or drug interchangeability) with a desired statistical inference (e.g., power or confidence interval). To ensure the success of studies conducted for the assessment of biosimilarity of follow-on biologics, regulatory guidelines/guidances need to be developed. Product-specific guidelines/guidances published by EU EMA have been criticized for not having standards. Although product-specific guidelines/guidances do not help to establish standards for the assessment of biosimilarity of follow-on biologics, they do provide the opportunity for accumulating valuable experience/information for establishing standards in the future. Thus, numerical studies are recommended including simulations, meta-analysis, and/or sensitivity analysis, in order to (1) provide a better understanding of these product-specific guidelines/guidances and (2) check the validity of the established Fundamental Biosimilarity Assumption, which is the legal basis for assessing biosimilarity of follow-on biologics.

REFERENCES

1. FDA. *Scientific Considerations in Demonstrating Biosimilarity to a Reference Product*. United States Food and Drug Administration, Silver Spring, MD, 2012.
2. FDA. *Quality Considerations in Demonstrating Biosimilarity to a Reference Protein product*. Food and Drug Administration, Silver Spring, MD, 2012.
3. FDA. *Biosimilars: Questions and Answers Regarding Implementation of the Biologics Price Competition and Innovation Act of 2009*. Food and Drug Administration, Silver Spring, MD, 2012.
4. Chow, S.C. *Biosimilars: Design and Analysis of Follow-On Biologics*; CRC Press, Taylor & Francis Group: New York, 2013.
5. Chow, S.C. Assessing biosimilarity and drug interchangeability of biosimilar products. *Stat. Med.* **2013**, *32*, 361–363.
6. Woodcock, J.; Griffin, J.; Behrman, R.; Cherney, B.; Crescenzi, T.; Fraser, B.; Hixon, D.; Joneckis, C.; Kozlowski, S.; Rosenberg, A.; Schrager, L. The FDA's assessment of follow-on protein products: A historical perspective. *Nat. Rev. Drug Discov.* **2007**, *6*, 437–442.
7. Wang, J. and Chow, S.C. (2012). On regulatory approval pathway of biosimilar products. *Pharmaceuticals*, *5*, 353–368; doi:10.3390/ph5040353.
8. Endrenyi, L.; Declerck, P.; Chow, S.C. *Biosimilar Drug Product Development*; CRC Press, Taylor & Francis Group, New York, **2017**.
9. Chow, S.C.; Wang, J.; Endrenyi, L.; Lachenbruch, P.A. Scientific considerations for assessing biosimilar products. *Stat. Med.* **2013**, *32*, 370–381.
10. Chow, S.C.; Yang, L.Y.; Starr, A.; Chiu, S.T. Statistical methods for assessing interchangeability of biosimilars. *Stat. Med.* **2013**, *32*, 442–448.
11. Chow, S.C.; Song, F.Y.; Endrenyi, L. A note on the Chinese draft guidance on biosimilar products. *Chin. J. Pharm. Anal.* **2015**, *5*, 4–9.
12. WHO. *Guidelines on Evaluation of Similar Biotherapeutic Products (SBPs)*; Geneva, 2009.
13. CPMP. *Guideline on Comparability of Medicinal Products Containing Biotechnology-derived Proteins as Active Substance: Non-clinical and Clinical Issues*; Committee for Proprietary Medicinal Products. EMEA/CPMP/3097/02/Final8, 2001.
14. CD. Commission Directive 2003/63/EC of 25 June 2003 amending Directive 2001/83/EC of the European Parliament and of the Council on the Community Code Relating to medicinal products for human use. *Official Journal of the European Union L* 159/46, **2003**.
15. EMA. *Guideline on Similar biological medicinal products. The European Medicines Agency Evaluation of Medicines for Human Use*; EMEA/CHMP/437/04, London, **2005**.
16. EMA. *Concept paper on the Revision of the Guideline on Similar Biological Medicinal Product*; EMA/CHMP/BMWP/572643, London, **2011**.
17. EMA. *Draft Guideline on Similar Biological Medicinal Products Containing Biotechnology-Derived Proteins as Drug Substance: Quality Issues. The European Medicines Agency Evaluation of Medicines for Human Use*; EMEA/CHMP/49348/05, London, **2006**.
18. EMA. *Concept paper on Revision of the Guideline on Similar Biological Medicinal Products Containing Biotechnology-derived Proteins as Active Substance: Quality Issues*; EMEA/CHMP/BWP/617111, London, **2011**.
19. EMA. *Guideline on Similar Biological Medicinal Products Containing Biotechnology-derived Proteins as Active Substance—Non-clinical and Clinical Issues*; EMEA/CHMP/BMWP/42832, London, **2006**.
20. EMA. *Concept paper on Revision of the Guideline on Similar Biological Medicinal Products Containing Biotechnology-derived Proteins as Active Substance: Non-clinical and Clinical Issues*; EMEA/CHMP/BMWP/572828, London, **2011**.
21. EMA. *Guideline on Similar Biological Medicinal Products*; CHMP/437/04Rev 1, Committee for Medicinal Products for Human Use (CHMP), London, October 23, **2014**.
22. EMA. *Guideline on Similar Biological Medicinal Products. The European Medicines Agency Evaluation of Medicines for Human Use*; EMEA/CHMP/437/04, London, **2006**.
23. MHLW. *Guidelines for the Quality, Safety and Efficacy Assurance of Follow-on Biologics*; Yakushoku shinsahatu 0304007, Tokyo, Japan, **2009**.
24. EMA. *Draft Annex Guideline on Similar Biological Medicinal Products Containing Biotechnology-derived Proteins as Drug Substance – Non Clinical and Clinical Issues—Guidance on Biosimilar Medicinal Products containing Recombinant Erythropoietins*; The European Medicines Agency Evaluation of Medicines for Human Use. EMEA/CHMP/94526/05, London, **2006**.
25. EMA. *Draft Annex Guideline on Similar Biological Medicinal Products Containing Biotechnology-derived Proteins as Drug Substance—Non Clinical and Clinical Issues—Guidance on Biosimilar Medicinal Products*

- containing *Recombinant Granulocyte-Colony Stimulating Factor*; The European Medicines Agency Evaluation of Medicines for Human Use. EMEA/CHMP/31329/05, London, **2006**.
26. EMA. Draft guideline on *Similar Biological Medicinal Products Containing Monoclonal Antibodies*; EMA/CHMP/BMWP/403543/2010, London, **2010**.
 27. EMA. Concept paper on *Similar Biological Medicinal Products Containing Recombinant follicle Stimulation Hormone*; EMA/CHMP/BMWP/94899/2010, London, **2010**.
 28. EMA. Guideline on *Similar Biological Medicinal Products Containing Interferon Beta*; EMA/CHMP/BMWP/652000/2010, London, **2011**.
 29. HC. Guidance for *Sponsors: Information and Submission Requirements for Subsequent Entry Biologics (SEBs)*; Health Canada, Ottawa, ON, **2010**.
 30. Suh, S.K.; Park, Y. Regulatory guideline for biosimilar products in Korea. *Biologicals* **2011**, *39*, 336–338.
 31. KFDA. *Korean Guidelines on the Evaluation of Similar Biotherapeutic Products (SBPs)*. Korean Food and Drug Administration, South Korea, 2011.
 32. CFDA. Draft *Guideline on the Development and Evaluation of Biosimilars* (Chinese version for public comments); Center for Drug Evaluation, China Food and Drug Administration: Beijing, China, October 29, **2014**.
 33. CFDA. Draft *guideline on the development and evaluation of biosimilars* (Chinese version for trial purpose). Center for Drug Evaluation, China Food and Drug Administration: Beijing, China, February 28, **2015**.
 34. Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*, 3rd Ed.; Chapman Hall/CRC Press, Taylor & Francis: New York, **2008**.
 35. FDA. Guidance on: *Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Considerations*; Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, MD, **2003**.
 36. Haidar, S.H.; Davit, B.M.; Chen, M.L.; Conner, D.; Lee, L.; Li, Q.H.; Lionberger, R.; Makhlof, F.; Patel, D.; Schuirmann, D.J.; Yu, L.X. Bioequivalence approaches for highly variable drugs and drug products. *Pharm. Res.* **2008**, *25*, 237–241.
 37. Tothfalusi, L.; Endrenyi, L.; Arieta, A.G. Evaluation of bioequivalence for highly variable drugs with scaled average bioequivalence. *Clin. Pharmacokin.* **2009**, *48*, 725–743.
 38. Davit, B.M.; Conner, D.P.; Fabian-Fritsch, B.; Haidar, S.H.; Jiang, X.; Patel, D.T.; Seo, P.R.H.; Suh, K.; Thompson, C.L.; Yu, L.X. Highly variable drugs: Observations from bioequivalence data submitted to the FDA for new generic drug applications. *AAPS J.* **2008**, *10*, 148–156.
 39. Haidar, S.H.; Makhlof, F.; Schuirmann, D.J.; Hyslop, T.; Davit, B.; Conner, D.; Yu, L.X. Evaluation of a scaling approach for the bioequivalence of highly variable drugs. *AAPS J.* **2008**, *10*, 450–454.
 40. FDA. Guidance on: *Statistical Approaches to Establishing Bioequivalence*; Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, MD, 2001.
 41. Chow, S.C. Individual bioequivalence—A review of FDA draft guidance. *Drug Inf. J.* **1999**, *33*, 435–444.
 42. Leeson, L.J. *In Vitro/in vivo* correlation. *Drug Inf. J.* **1995**, *29*, 903–915.
 43. FDA. *Guidance for Industry: Dissolution Testing of Immediate Release Solid Oral Dosage Forms*; Food and Drug Administration: Rockville, MD, 1997.
 44. Chow, S.C.; Ki, F. Statistical comparison between dissolution profiles of drug products. *J. Biopharm. Stat.* **1997**, *7*, 241–258.
 45. Moore, J.W.; Flanner, H.H. Mathematical comparison of curves with an emphasis on dissolution profiles. *Pharm. Technol.* **1996**, *20*, 64–74.
 46. SUPAC-IR. Food and Drug Administration guideline on: *Immediate Release Solid Oral Dosage Forms: Scale-up and Post Approval Changes: Chemistry, Manufacturing, and Controls, In Vitro Dissolution Testing, and In Vivo Bioequivalence Documentation*; FDA: Rockville, MD, 1995.
 47. Liu, J.P.; Ma, M.C.; Chow, S.C. Statistical evaluation of similarity factor f_2 as a criterion for assessment of similarity between dissolution profiles. *Drug Inf. J.* **1997**, *31*, 1255–1271.
 48. Shah, V.P.; Tsong, Y.; Sathe, P.; Liu, J.P. In vitro dissolution profile comparison – Statistics and analysis of the similarity factor, f_2 . *J. Pharmacol. Res.* **1998**, *15*, 889–896.
 49. Ma, M.; Lin, R.; Liu, J.P. Statistical evaluations of dissolution similarity. *Statistica Sinica* **1999**, *9*, 1011–1027.
 50. Chow, S.C.; Shao, J. A note on statistical methods for assessing therapeutic equivalence. *Controlled Clin. Trials* **2002**, *23*, 515–520.
 51. Chow, S.C. Quantitative evaluation of bioequivalence/biosimilarity. *J. Bioequivalence Bioavailability* **2011**, *S1:002*, 1–8; <http://dx.doi.org/10.4172/jbb.s1-002>.
 52. Kang, S.H.; Chow, S.C. Statistical assessment of biosimilarity based on relative distance between follow-on biologics. *Stat. Med.* **2013**, *32*, 382–392.
 53. Schuirmann, D.J. On hypothesis testing to determine if the mean of a normal distribution is continued in a known interval. *Biometrics* **1981**, *37*, 617.
 54. Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokin. Biopharm.* **1987**, *15*, 657–680.
 55. Box, G.E.P.; Tiao, G.C. *Bayesian Inference in Statistical Analysis*; Addison-Wesley: Reading, MA, 1973.
 56. Edwards, W.; Lindman, H.; Savage, L.J. Bayesian statistical inference for psychological research. *Psychol. Rev.* **1963**, *70*, 193.
 57. Lindley, D.V. *Introduction to Probability and Statistics for a Bayesian Viewpoint, Part II, Inference*; Cambridge University Press: Cambridge, 1965.
 58. Rodda, B.E.; Davis, R.L. Determining the probability of an important difference in bioavailability. *Clin. Pharmacol. Ther.* **1980**, *28*, 247–252.
 59. Hauschke, D.; Steinijans, V.W.; Diletti, E. A distribution-free procedure for the statistical analyses of bioequivalence studies. *Int. J. Clin. Pharmacol. Ther. Toxicol.* **1990**, *28*, 72–78.
 60. Cornell, R.G. The evaluation of bioequivalence using non-parametric procedures. *Commun. Stat. Theory Methods* **1990**, *19*, 4153–4169.

61. Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics* **1945**, *1*, 80–83.
62. Mann, H.B.; Whitney, D.R. On a test of whether one or two random variables is stochastically larger than the other. *Ann. Math. Stat.* **1947**, *18*, 50–60.
63. Phillips, K.F. Power of the two one-sided tests procedure in bioequivalence. *J. Pharmacokin. Biopharm.* **1990**, *18*, 137–144.
64. Liu, J.P.; Chow, S.C. Sample size determination for the two one-sided tests procedure in bioequivalence. *J. Pharmacokin. Biopharm.* **1992**, *20*, 101–104.
65. Chen, K.W.; Chow, S.C.; Li, G. A note on sample size determination for bioequivalence studies with higher-order cross-over designs. *J. Pharmacokin. Biopharm.* **1997**, *25*, 753–765.
66. Chow, S.C. Criteria for assessing biosimilarity for follow-on biologics. Presented at *Current Advanced Statistical Issues in Clinical Trials—Biosimilars*. Taipei, Taiwan, October 2, **2009**.
67. Chow, S.C. On scientific factors of biosimilarity and interchangeability. Presented at the *FDA Public Hearing on Approval Pathway for Biosimilar and Interchangeable Biological Products*. Silver Spring, MD, November 2–3, 2010.
68. Chow, S.C.; Hsieh, T.C.; Chi, E.; Yang, J. A comparison of moment-based and probability-based criteria for assessment of follow-on biologics. *J. Biopharm. Stat.* **2010**, *20*, 31–45.
69. Hsieh, T.C.; Chow, S.C.; Liu, J.P.; Hsiao, C.F.; Chi, E. Statistical test for evaluation of biosimilarity of follow-on biologics. *J. Biopharm. Stat.* **2010**, *20*, 75–89.
70. Chow, S.C.; Endrenyi, L.; Lachenbruch, P.A.; Yang, L.Y.; Chi, E. Scientific factors for assessing biosimilarity and drug interchangeability of follow-on biologics. *Biosimilars* **2011**, *1*, 13–26.
71. Shao, J.; Chow, S.C. Reproducibility probability in clinical trials. *Stat. Med.* **2002**, *21*, 1727–1742.
72. Owen, D.B. A special case of a non-central *t*-distribution. *Biometrika* **1965**, *52*, 437–446.

Biosimilarity of Follow-On Biologics

Shein-Chung Chow

Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, U.S.A.

Jen-pei Liu

Division of Biometry, Department of Agronomy, and Institute of Epidemiology, National Taiwan University, Taipei, Taiwan; and Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan

Abstract

This entry will review the issues surrounding biosimilars, including manufacturing, quality control, clinical efficacy, side effects (safety), and immunogenicity. In addition, it will also attempt to address the question regarding how regulatory agencies and industry regulations are evolving to deal with these issues.

INTRODUCTION

When an innovative (brand-name) drug product goes off patent, pharmaceutical and generic companies may file an abbreviated new drug application (ANDA) for approval of the generic copies of the innovative drug product. In 1984, the U.S. Food and Drug Administration (FDA) was authorized to approve generic drug products under the *Drug Price Competition and Patent Term Restoration Act* (which is also known as the *Hatch and Waxman Act*). For approval of generic drug products, the FDA requires that evidence in average of bioavailability (in terms of rate and extent of drug absorption) be provided through the conduct of bioavailability and bioequivalence studies. As indicated by Chow and Liu,^[1] the assessment of bioequivalence as a surrogate for evaluation of drug safety and efficacy is based on the *Fundamental Bioequivalence Assumption* that if two drug products are shown to be bioequivalent in average bioavailability, they will reach the same therapeutic effect or they are therapeutically equivalent. Under the Fundamental Bioequivalence Assumption, regulatory requirements, study design (e.g., a two-sequence, two-period crossover design or a replicated crossover design), criteria (e.g., 80/125 rule based on log-transformed data), and statistical methods (e.g., Shuirmann's two one-sided tests or confidence interval approach) for assessment of bioequivalence have been well established.^[1–4]

Unlike drug products, the concept for development of generic versions of biologic products, which are usually referred to as follow-on biologics (by the U.S. FDA) or biosimilars (by the European Medicine Agency [EMA] of the European Union) is different. Webber^[5] defines follow-on (protein) biologics as products that are intended to be sufficiently similar to an approved product to permit

the applicant to rely on certain existing scientific knowledge about safety and efficacy of an approved reference product. Under this definition, follow-on products are not only intended to be similar to the reference product, but also intended to be interchangeable with the reference product. As a number of biologic products are due to expire in the next few years, the subsequent production of follow-on products has aroused considerable interest within the pharmaceutical/biotechnological industry as biosimilar manufacturers strive to obtain part of an already large and rapidly growing market. The potential opportunity for price reductions versus the originator biologic products remains to be determined, as the advantage of a slightly cheaper price may be outweighed by the hypothetical increased risk of side effects from biosimilar molecules that are not exact copies of their originators. In this entry, we will focus on the issues surrounding biosimilars, including manufacturing, quality control, clinical efficacy, side effects (safety), and immunogenicity. In addition, we will also attempt to address the question regarding how regulatory agencies and industry regulations are evolving to deal with these issues.

Biosimilars are fundamentally different from generic chemical drugs. Important differences include the size and complexity of the active substance and the nature of the manufacturing process. Unlike classical generics, biosimilars are not identical to their originator products and therefore should not be brought to market using the same procedure applied to generics. This is partly a reflection of the complexities of manufacturing and safety and efficacy controls of biosimilars when compared to their small-molecule generic counterparts.^[6–9] Because biologic products are usually recombinant protein molecules manufactured in living cells,^[10] manufacturing processes for biologic products are

highly complex and require hundreds of specific isolation and purification steps. In practice, it is impossible to produce an identical copy of a biologic product, as changes to the structure of the molecule can occur with changes in the production process. Given that a protein can be modified (e.g., a side chain may be added, the structure may have changed due to protein misfolding, and so on) during the process, different manufacturing processes may invariably lead to structural differences in the final product, which may result in differences in efficacy and may have a negative impact on patient immune responses.

The remainder of this entry is organized as follows. The next section briefly outlines regulatory requirements for approval of biosimilars by the EMEA of the European Union as well as the current position of the FDA. The section “Criteria for Similarity” reviews various criteria for assessment of bioequivalence, similarity, and consistency that appeared in either regulatory guidances and/or literature. Some scientific issues for assessment of biosimilars are discussed in the section “Scientific Issues,” while the section “Assessing Similarity Using Genomic Data” proposes an approach for assessment of similarity using genomic data. The last section provides a brief concluding remark.

REGULATORY REQUIREMENTS

The EMEA^[11–19] has issued a new guideline describing general principles for the approval of similar biological medicinal products or biosimilars in the European Union. The guideline is accompanied by four concept papers that outline areas in which the agency intends to provide more targeted guidance. Specifically, the concept papers discuss approval requirements for four classes of human recombinant products containing erythropoietin, human growth hormone, granulocyte-colony stimulating factor, and insulin. The guidelines consist of a checklist of documents published to date relevant to data requirements for biological pharmaceuticals. It is not clear regarding what specific scientific requirements the guidelines will be applied to biosimilar applications, nor how the agency will treat innovator data contained in the reference product dossiers. Although the guidelines provide a useful summary of the biosimilar legislation and previous E.U. publications, it provides few answers to the issues.

On the other hand, approval of follow-on biologics in the United States depends on whether the biologic product is approved under the U.S. Food, Drug, and Cosmetic Act (U.S. FD&C) or whether it is licensed under the U.S. Public Health Service Act (U.S. PHS). As indicated, some proteins are licensed under the PHS Act, whereas others are approved under the FD&C Act. For products approved under a new drug application (NDA) (U.S. FD&C Act), generic version of the products can be approved under an ANDA, e.g., under Section 505(b)(2) of the FD&C Act.

For products that are licensed under a biologics license application (BLA) (U.S. PHS Act), there exists no abbreviated BLA. As pointed out by Woodcock et al.,^[20] for assessment of similarity of follow-on biologics, the FDA would consider the following factors: (i) the robustness of the manufacturing process, (ii) the degree to which structural similarity could be assessed, (iii) the extent to which mechanism of action was understood, (iv) the existence of valid, mechanistically related pharmacodynamic assays, (v) comparative pharmacokinetics, (vi) comparative immunogenicity, (vii) the amount of clinical data available, and (ix) the extent of experience with the original product.^[21–23] A typical example would be a recent regulatory approval of Omnitrope® (Somatropin), which was approved in 2006 under section 505(b)(2) of FD&C Act. Omnitrope® was approved based on the following evaluations: (i) physicochemical testing that established highly similar structure to Genotropin; (ii) new non-clinical pharmacology and toxicology data specific to Omnitrope®; (iii) pharmacokinetic (PK), pharmacodynamic (PD), and comparative bioavailability data; (iv) clinical efficacy and safety data from comparative controlled trials and from long-term trials with Omnitrope®; and (v) vast clinical experience and a wealth of published literature concerning the clinical effects (safety and effectiveness) of human growth hormone. The approval of Omnitrope® is based on an ad hoc case-by-case review of individual biosimilar application. In practice, there is a strong industrial interest and desire for the regulatory agencies to develop review standards and an approval process for biosimilars than an ad hoc case-by-case review of individual biosimilar applications. Along these lines, the FDA has indicated that the following guidances are currently under development: include (i) guidance for industry on scientific considerations in demonstrating the safety and effectiveness of follow-on protein products, and (ii) guidance for industry on chemistry, manufacturing, and control (CMC) issues for follow-on protein products.

CRITERIA FOR SIMILARITY

For comparison between drug products, some criteria for assessment of bioequivalence, similarity (e.g., dissolution profiles comparison), and consistency (e.g., comparison between manufacturing processes) are available in either regulatory guidelines/guidances and/or literature. These criteria, however, can be classified into (i) absolute change versus relative change, (ii) aggregated versus disaggregated, or (iii) moment-based versus probability-based. This section briefly reviews different categories of criteria.

Absolute Change Versus Relative Change

In clinical research and development, for a given study endpoint, post-treatment absolute change from baseline or post-treatment relative change from baseline is usually

considered for comparison between treatment groups. A typical example would be the study of weight reduction in obese patient population. In practice, it is not clear whether a clinically meaningful difference in terms of absolute change from baseline can be translated to a clinically meaningful difference in terms of relative change from baseline. Sample size calculation based on power analysis in terms of absolute change from baseline or relative change from baseline could lead to a very different result.

Current regulation for assessment of bioequivalence between drug products in terms of average bioavailability is based on relative change. In other words, we conclude (average) bioequivalence between a test product and a reference product if the 90% confidence interval for the ratio of means of the primary pharmacokinetic response such as area under the blood or plasma-concentration time curve (AUC) between the two drug products is totally within 80% and 125%. Note that regulatory agencies suggest that a log-transformation be performed before data analysis for assessment of bioequivalence.

Aggregated Versus Disaggregated

As indicated by Chow and Liu,^[1] bioequivalence can be assessed by evaluating differences in averages, intra-subject variabilities, and variance due to subject-by-formulation interaction between drug products *separately*. Individual criteria for assessment of differences in averages, intra-subject variabilities, and variance due to subject-by-formulation interaction between drug products are referred as to the disaggregated criteria. If the criterion is a single summary measure composed of these individual criteria, it is called the aggregated criterion.

For assessment of bioequivalence in average bioavailability (ABE), most regulatory agencies including the FDA recommend the use of disaggregate criterion based on average bioavailability. In other words, bioequivalence is concluded if the average bioavailability of the test formulation is within (80%, 125%) that of the reference formulation, with a certain assurance. Note that the EMEA^[24] and World Health Organization (WHO)^[25] use the same equivalence criterion of 80–125% for the log-transformed pharmacokinetic responses such as AUC. However, for Cmax, in certain cases, the EMEA and WHO allow a wider interval of 75–133% for the ratio of average bioavailability to address any safety and efficacy concerns for patients switched between formulations. If a wider interval is used, it must be pre-specified in the protocol. More details can be found in Chow and Liu.^[1]

On the other hand, for assessment of population bioequivalence (PBE) and individual bioequivalence (IBE), the following aggregated criteria are often considered. For assessment of individual bioequivalence, criterion proposed in the FDA guidance^[3] can be expressed as

$$\theta_I = (\delta^2 + \sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2) / \max\{\sigma_{W0}^2, \sigma_{WR}^2\}, \quad (1)$$

where $\delta = \mu_T - \mu_R$, $\sigma_{WT}^2, \sigma_{WR}^2, \sigma_D^2$ are the true difference in means, intra-subject variabilities of the test product and the reference product, and variance due to subject-by-formulation interaction between drug products, respectively. σ_{W0}^2 is the scale parameter specified by the user. Similarly, the criterion for assessment of population bioequivalence suggested in the FDA guidance^[3] is given by

$$\theta_p = (\delta^2 + \sigma_{TT}^2 - \sigma_{TR}^2) / \max\{\sigma_{T0}^2, \sigma_{TR}^2\}, \quad (2)$$

where $\sigma_{TT}^2, \sigma_{TR}^2$ are the total variances for the test product and the reference product, respectively and σ_{T0}^2 is the scale parameter specified by the user.

A typical approach is to construct a one-sided 95% confidence interval for $\theta_I(\theta_p)$ for assessment of individual (population) bioequivalence. If the one-sided 95% upper confidence limit is less than the bioequivalence limit of $\theta_I(\theta_p)$, we then conclude that the test product is bioequivalent to that of the reference product in terms of individual (population) bioequivalence. More details regarding individual and population bioequivalence can be found in Chow and Liu.^[1]

Moment Based Criteria Versus Probability Based Criteria

Schall and Luus^[26] propose the moment-based and probability-based measures for the expected discrepancy in pharmacokinetic responses between drug products. The moment-based measure suggested by Schall and Luus^[26] is based on the following expected mean-squared differences:

$$d(Y_j; Y_{j'}) = \begin{cases} E(Y_T - Y_R)^2 & \text{if } j = T \text{ and } j' = R \\ E(Y_R - Y_R')^2 & \text{if } j = R \text{ and } j' = R. \end{cases} \quad (3)$$

For some pre-specified positive number r , one of probability-based measures for the expected discrepancy is given as^[26]

$$d(Y_j; Y_{j'}) = \begin{cases} P\{|Y_T - Y_R| < r\} & \text{if } j = T \text{ and } j' = R \\ P\{|Y_R - Y_R'| < r\} & \text{if } j = R \text{ and } j' = R. \end{cases} \quad (4)$$

$d(Y_T; Y_R)$ measures the expected discrepancy for some pharmacokinetic metric between test and reference formulations, and $d(Y_R; Y_R')$ provides the expected discrepancy between the repeated administrations of the reference formulation. The role of $d(Y_R; Y_R')$ in formulation of bioequivalence criteria is to serve as a control. The rationale is that the reference formulation should be bioequivalent to itself. Therefore, for the moment-based measures, if the test formulation is indeed bioequivalent to the reference formulation, then $d(Y_T; Y_R)$ should be very close to $d(Y_R; Y_R')$. It follows that the criteria are functions of the difference (or ratio) between $d(Y_T; Y_R)$ and $d(Y_R; Y_R')$. Bioequivalence

is concluded if they are smaller than some pre-specified limit. On the other hand, for probability-based measures, if the test formulation is indeed bioequivalent to the reference formulation, as compared with $d(\mathcal{R}_R; \mathcal{Y}_R')$, $d(\mathcal{R}_T; \mathcal{Y}_R)$ should be relatively large. As a result, bioequivalence is assumed if the criteria based on the probability-based measure is greater than some pre-specified limit.

Similarity Factor for Dissolution Profile Comparison

In vivo bioequivalence studies are surrogate trials for assessing equivalence between test and reference formulations based on the rate and extent of drug absorption in humans to establish similar effectiveness and safety under the fundamental bioequivalence assumption. However, drug absorption depends on the dissolved state of drug product and dissolution testing provides a rapid *in vitro* assessment of the rate and extent of drug release. Leeson,^[27] therefore, has suggested that *in vitro* dissolution testing be used as a surrogate for *in vivo* bioequivalence studies to assess equivalence between the test and reference formulations for post-approval changes. For comparison of dissolution profiles, FDA guidance suggests considering the assessment of (i) the overall profile similarity and (ii) similarity at each sampling time point.^[28] In order to achieve these two objectives, based on Moore and Flanner,^[29] both the FDA scale-up and post-approval changes (SUPAC) guidance^[30] and guidance on dissolution testing^[28] suggest the similarity and difference factor for assessment of similarity. The similarity factor is then defined as the logarithm of the reciprocal square root transformation of 1 plus the mean-squared (the average sum of squares) difference in mean cumulative percentage dissolved between the test and reference formulations over all sampling time points. That is,

$$f_2 = 50 \log \left\{ \left[1 + \frac{Q}{n} \right]^{-0.5} 100 \right\}, \quad (5)$$

where

$$Q = \sum_{i=1}^n (\mu_{Ri} - \mu_{Ti})^2,$$

where log denotes the logarithm based on 10.

On the other hand, the difference factor is the sum of the absolute difference in mean cumulative percentage dissolved between the test and reference formulations divided by the sum of the mean cumulative dissolved of the reference formulation.

$$f_1 = \frac{\sum_{i=1}^n |\mu_{Ri} - \mu_{Ti}|}{\sum_{i=1}^n \mu_{Ri}} \quad (6)$$

It should be noted that the definitions of f_1 and f_2 provided by Moore and Flanner,^[29] and in the SUPAC and guidance on dissolution testing are not clear whether they are defined based on the population means or the sample averages. However, following the traditional statistical inference with ability for evaluation of error probability, we define both f_1 and f_2 based on the population mean dissolution rates. It follows that f_1 and f_2 are population parameters for assessment of similarity of dissolution profiles between the test and reference formulations.

Consistency in Manufacturing Process/Quality Control

Tse et al.^[31] have proposed a statistical quality control (QC) method to assess a proposed an index to test consistency between raw materials (which are from different resources) and/or between final products manufactured by different manufacturing processes. The consistency index is defined as the probability that the ratio of the characteristics (e.g., potency) of the drug products produced by two different manufacturing processes is within a pre-specified limit of consistency. The consistency index closes to 1 indicates that the characteristics of the drug products from the two manufacturing processes are almost identical. The idea for testing consistency is to construct a 95% confidence interval for the proposed consistency index under a sampling plan. If the constructed 95% confidence lower limit is greater than a pre-specified QC lower limit, then we claim that the final products produced by the two manufacturing processes are consistent.

Let U and W be the characteristics of the drug products from two different manufacturing processes, where $X = \log U$ and $Y = \log W$ follows normal distributions with means μ_X , μ_Y and variances V_X , V_Y , respectively. Similar to the idea of using $P(X < Y)$ to assess reliability in statistical quality control,^[32–33] Tse et al.^[31] have proposed the following probability as an index to assess the consistency between the two different manufacturing processes:

$$p = P\left(1 - \delta < \frac{U}{W} < \frac{1}{1 - \delta}\right),$$

where $0 < \delta < 1$ and is defined as a limit that allows for consistency. Tse et al.^[31] refer p as the consistency index. Thus p tends to 1 as δ tends to 1. For a given δ , if p is close to 1, materials U and W are considered to be identical. It should be noted that a small δ implies the requirement of high degree of consistency between material U and material W . In practice, it may be difficult to meet this narrow specification for consistency. Tse et al.^[31] have proposed the following quality control (QC) criterion: If the probability that the lower limit $LL(\hat{p})$ of the constructed $(1 - \alpha) \times 100\%$ confidence interval of p is greater than or equal to a pre-specified quality control lower limit, say, QC_L , exceeds a pre-specified number β (say $\beta = 80\%$), then we claim that

U and W are consistent or similar. In other words, U and W are consistent or similar if $P(QC_L \leq LL(\hat{p})) \geq \beta$, where β is a pre-specified constant.

SCIENTIFIC ISSUES

Biosimilarity in Biological Activity

Pharmacological or biological activity is an expression describing the beneficial or adverse effects of a drug on living matter. When the drug is a complex chemical mixture, this activity is exerted by the substance's active ingredient or pharmacophore but can be modified by the other constituents. The main kind of biological activity is a substance's toxicity. Activity is generally dosage-dependent, and it is not uncommon to have effects ranging from beneficial to adverse for one substance when going from low to high doses. Activity depends critically on fulfillment of the absorption, distribution, metabolism, and excretion (ADME) criteria.

Note that the new European Community (EU) *Pharmaceutical Review* legislation published on April 30, 2004, amended the EU community code on medicinal products to provide for the approval of biosimilars based on less preclinical and clinical data than had been required for the original reference product. The complexity of the protein and knowledge of its structure-function relationships determine the types of information needed to establish similarity.

Similarity in Size and Structure

In practice, various *in vitro* tests such as the assessments of the primary amino acid sequence, charge, and hydrophobic properties are performed to compare the structural aspects of biosimilars with their originator molecules. However, there are concerns about whether *in vitro* tests can be predictive of biological activity *in vivo*, due to the fact that there are significant differences in biological activity despite similarities in size and structure. Besides, it is difficult to assess biological activity adequately as few animal models are able to provide data that can be extrapolated for an accurate and reliable prediction of biological activity in human. Thus, controlled clinical trials remain the most reliable means of demonstrating similarity between a biosimilar molecule and the originator product in the clinic.

The Problem of Immunogenicity

Given that all biologic products are biologically active molecules derived from living cells and have the potential to evoke an immune response, immunogenicity is probably the most critical safety concern for assessment of biosimilarity of follow-on biologics. Common possible causes of immunogenicity include, but are not limited to,

(i) sequence differences between therapeutic protein and endogenous protein; (ii) non-human sequences or epitopes; (iii) structural alterations; (iv) storage conditions; (v) purification during the manufacturing process; (vi) formulation (e.g., surfactants, etc.); (vii) route, dose, and frequency of administration; and (viii) patient status, such as genetic background. Thus, the following questions are necessarily asked when assessing biosimilarity between biological products: (i) What is the immunogenic potential of the therapeutic protein? (ii) What is the impact of the generating antibodies to the self protein? (iii) What is the impact of immunogenicity in preclinical toxicity (e.g., pharmacokinetic levels and dose limiting toxicity)? (iv) What is the impact of immunogenicity of the therapeutic protein on safety? (v) What are the risk evaluation and mitigation strategy process required by regulatory agencies such as the FDA?

The immune responses to biologic products can lead to (i) anaphylaxis, (ii) injection site reactions, (iii) flu-like symptoms, and (iv) allergic responses. Note that one of the most serious adverse events occurs when neutralizing antibodies to product react with endogenous proteins that have a unique physiological role. The risk of immunogenicity can be reduced through stringent testing of the products during its development. However, it should be noted not only that immunogenicity in animals does not predict immunogenicity in the clinical trials, but also that analytical techniques may not detect changes that may impact immunogenicity. Therefore, the immunogenicity of a biological product depends heavily upon the product quality attributes such as the physical, structural, and functional properties of the active pharmaceutical ingredients as well as excipients, container closure, and delivery system. It turns out that similarity of the acceptable ranges of these quality attributes is crucial to the evaluation of equivalence between biosimilar and innovator's products.

Manufacturing Process

Unlike small-molecule drug products, biological products are made of living cell. Thus, manufacturing of biologic products is a very complicated process, which involves the steps of (i) cell expansion, (ii) cell production (in bioreactors), (iii) recovery (through filtration or centrifugation), (iv) purification (through chromatography), and (v) formulation. A small discrepancy at each step (e.g., purification) could lead to a significant difference in the final product, which might cause drastic change in clinical outcomes. Thus, process control and validation play an important role for the success of the manufacturing of biological products. In addition, because at each step (e.g., purification), different methods may be used at different biological manufacturing processes (within the same company or at different biotech companies), tests for consistency are necessarily performed. Note that at the step of purification, the following chromatography media or resin are commonly

considered: (i) gel filtration, (ii) ion exchange, (iii) hydrophobic interaction, (iv) reversed phase normal phase, and (v) affinity. Thus, at each step of manufacturing process, primary performance characteristics should be identified, control, and tested for consistency for process control and validation.

Statistical Considerations

Fundamental Biosimilarity Assumption

Because of complexity of the biosimilar products, unlike its chemical generics, although the EMEA approved Omnitrope® and Valtropin® in 2006 and the FDA approved Omnitrope® in 2006, both regulatory agencies required clinical equivalence trials. For example, the application documents of Valtropin® submitted to the EMEA consists of the following clinical evaluations:

- i. One pivotal bioequivalence study in healthy volunteers. ($n = 24$)
- ii. One phase-III pivotal clinical equivalence trial for efficacy and safety in children with growth hormone deficiency (GHD). ($n = 147$)
- iii. One phase-III trial for efficacy and safety in children with Turn syndrome (TS). ($n = 30$)
- iv. One phase-III trial for efficacy and safety in children with TS conducted in Korea. ($n = 60$)
- v. One phase-III trial for efficacy and safety in adults with GHD. ($n = 92$)

Therefore, substantial clinical information was required by both regulatory agencies for the approval of Valtropin® as compared to one pivotal bioequivalence study and possible one food interaction study required by the chemical generics. In fact, the amount of clinical information requested by the regulatory agencies is almost equivalent to a NDA or BLA of an innovator's biological products. Consequently, reduction of the price of the biosimilar products is severely limited and the affordability and accessibility of the biological products, including biosimilar products, for patients in need is seriously hampered.

However, similar to the chemical generic drug products, approval for the biosimilar drug products can be treated as the evaluation of post-approval changes (International Conference on Harmonization Q5E, 2005),^[23] and the post-approval change is the change of drug manufacturers. Therefore, if the biosimilar drug products and the corresponding innovator's biological products appear highly similar. In addition, based on the accumulated experience relevant information and data, minute differences observed in the product characteristics are expected to have no adverse effect of safety and efficacy profiles. In these circumstances, biosimilar drug products and an innovator's products can be considered similar. Therefore, except for the traditional pivotal bioequivalence study, no further data

from pivotal phase-III trials should be requested. However, the above statement is based on a crucial assumption that the at least one of the product characteristics are reliable predictors of the safety and efficacy profiles of the biological products. As a result, the Fundamental Biosimilarity Assumption (FBA) is

When a biosimilar product is claimed to be biosimilar to an innovator's product based on some well-defined product characteristics and is therapeutically equivalent provided that the well-defined product characteristics are validated and reliable predictors of safety and efficacy of the products.

For the chemical generic products, the well-defined product characteristics are the exposure measures for early, peak, and total portions of the concentration-time curve. The FBA assumes that the equivalence in the exposure measures implies therapeutically equivalent. However, due to the complexity of the biosimilar drug products, one has to verify that some validated product characteristics are indeed reliable predictors of the safety and efficacy. It follows that the design and analysis for evaluation of equivalence between the biosimilar drug product and an innovator products are substantially different from those of the chemical generic products.

Study Design

Because some of the biological products such as therapeutic antibody or pegylated proteins have a long half-life, equivalence in terms of absorption/bioavailability may not be sufficient. Demonstration of equivalence on clearance and half-life may be required for assess the risk of difference in elimination rate. As a result, the traditional crossover designs may not be optimal for evaluation of equivalence between follow-on or biosimilar and innovator biological products. On the other hand, if the well-defined and validated product characteristics are pharmacokinetic/pharmacodynamic (PK/PD) responses, it is then very important to investigate the extrapolation ability of equivalence in PK responses to equivalence in PD, and to equivalence in efficacy responses. In order to ensure the internal validity of treatment comparisons, PK, PD, and efficacy response should be evaluated simultaneously in the same trials. We consider Design (a) in Fig. 1 proposed below:

Design (a) is a two-group parallel design in patients for the PKPD/efficacy bridging study with the disease which the innovator biological product indicates for. After meeting the inclusion and exclusion criteria, patients are randomly divided into two groups. PD/efficacy/safety will be evaluated for the first group of the patients (validation set). Additional PK responses will be assessed for the second group of the patients (training set). A randomization in a 1:1 ratio will be performed separately for each group. The sample size of the second group will be large enough to

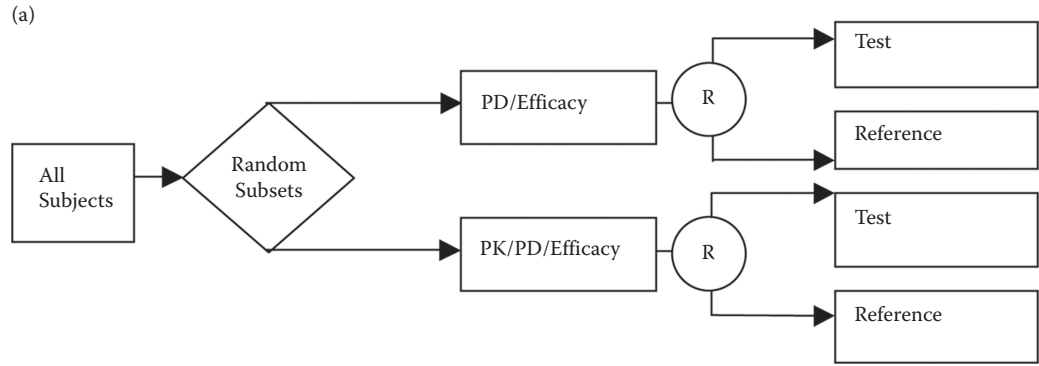


Fig. 1 Diagram of PK/PD/efficacy bridging study

provide sufficient power for evaluation of bioequivalence based on PK responses. Calibration models will be built based on the PK/PD/efficacy response obtained from the patients in the training set. The PD/efficacy data from the validation set will be used to provide independent assessment of extrapolation ability of equivalence in PK responses to equivalence in PD/efficacy responses.

Design (a) may require quite large sample size because of simultaneously evaluation of extrapolation ability of equivalence in PK to equivalence in PD and to equivalence in efficacy. One way to resolve this issue is to adopt the design for dose-response trials for evaluation of extrapolation ability of equivalence in some well-defined product characteristics to equivalence in efficacy. This design is referred to as Design (b) shown in Fig. 2. Design (b) in fact consists of two dose-response trials: one for the biosimilar product and one for the innovator’s biological product, each with at least three dose levels with a placebo group. Eligible patients are first randomized into biosimilar or innovator groups. Within each group, patients are randomized again to receive

one of the doses for the respective products. Well-defined product characteristics and primary efficacy endpoints are evaluated for all patients at their respective doses. Suppose that a statistically significant relationship as represented by a simple linear regression equation can be established between the well-defined product characteristics and the primary efficacy endpoint through dose levels for the innovator’s product, perhaps after a suitable transformation. If a similar linear relationship can be also obtained for the biosimilar product, and its corresponding linear regression equation is very close to the one for the innovator’s product, then equivalence in efficacy based on the primary efficacy endpoint may be claimed. Because the innovator’s product has been approved by the regulatory agencies due to its confirmed efficacy, therefore the objective of Design (b) is not to establish the efficacy of either biological products but to establish the similar patterns of the relationship between the well-defined product characteristics and primary efficacy endpoint for the two products. As a result, the sample size of Design (b) can be reduced significantly.

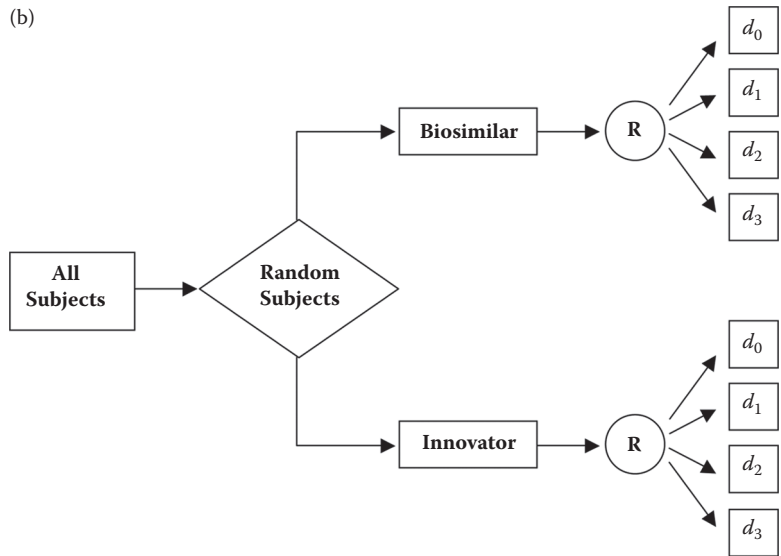


Fig. 2 Design (b) for evaluation of extrapolation ability

Alternative Criteria for Biosimilarity

Due to the complexity, heterogeneity, and complication mechanisms of biological drug products, difference in variability between biosimilar and innovator biological products in PK, PD, and clinical responses will be much larger than the difference observed between the conventional generic and innovator chemical drug product. Therefore, average bioequivalence alone is not sufficient to establish equivalence between the follow-on and innovator biological products. On the other hand, because of the masking effect, the aggregate metrics for population and individual bioequivalence fail to address the closeness of the distributions of the responses between the follow-on and innovator biological products.^[34,35] Disaggregate metrics can address the masking effect suffered by the aggregate metrics and find the sources of inequivalence. However, determination of individual equivalence margins with different interpretation is not an easy task. In addition, because of multi-parameters involved, any procedures based on disaggregate metric for evaluation of equivalence between follow-on and innovator biological products will tend to be conservative, especially in the small samples. Furthermore, all current methods derived from the probability-based, moment-based, aggregate, or disaggregate criteria are based on the normality assumption that is either extremely difficult to verify or simply not true. To resolve the above-mentioned dilemmas for evaluation of equivalence, the following concept of stochastic equivalence or stochastic non-inferiority is proposed.

Let $F(x)$ and $G(y)$ be the cumulative distribution functions of the responses for biosimilar and innovator biological products, respectively. Assuming that a large response value indicates a better efficacy, the follow-on and innovator biological products are said to be stochastically equivalent (two-sided) if the absolute difference between $F(x)$ and $G(x)$ is within some pre-specific margins for all x . In other words, metric $\theta = \sup|F(x) - G(x)|$ and hypothesis for equivalence becomes

$$\begin{aligned} H_0 : \sup|F(x) - G(x)| \geq \eta, \text{ for some } x \\ \text{vs.} \\ H_1 : \sup|F(x) - G(x)| < \eta, \text{ for all } x \end{aligned} \quad (7)$$

Similarly, the biosimilar product is said to be stochastically non-inferior to the innovator counterpart if the difference between $F(x)$ and $G(x)$ is greater than $-\eta$. The corresponding hypothesis is given as

$$\begin{aligned} H_0 : \sup[F(x) - G(x)] \leq -\eta, \text{ for some } x \\ \text{vs.} \\ H_1 : \sup[F(x) - G(x)] > -\eta, \text{ for all } x \end{aligned} \quad (8)$$

However, the hypotheses in Eq. (7) and Eq. (8) can only be used for evaluation of equivalence with respect to one study endpoint such as AUC or some primary efficacy endpoint. They cannot be utilized to assess whether the equivalence in product characteristics such as AUC can be extrapolated to equivalence in primary efficacy endpoints. Both well-defined product characteristics and primary efficacy endpoints are measured for each patient. Therefore, group means of well-defined product characteristic can be computed for each dose level for biosimilar and innovator's product respectively. Using the group means of the well-defined characteristic as the independent variable, a simple linear regression equation can be fit to the primary efficacy endpoint (dependent variable) for biosimilar and innovator's biological products respectively. It follows that the concept of the relative potency in the parallel-line bioassay can be then employed to investigate extrapolation ability of equivalence in product characteristic to equivalence in efficacy.^[36] In other words, if the relative potency between the biosimilar and innovator's biological products is within some predefined margins, then it can be concluded that equivalence in product characteristic be extrapolated to equivalence in efficacy. Fig. 3 depicts the application of the parallel line assay to evaluation of the extrapolation ability of equivalence in product characteristic to equivalence in efficacy. Let ρ be the relative potency of the biosimilar product to the innovator's biological product. The hypothesis of extrapolation ability is given below:

$$H_0 : \rho \leq \rho_L \text{ or } \rho \geq \rho_U \quad \text{vs} \quad H_A : \rho_L < \rho < \rho_U, \quad (9)$$

where $0 < \rho_L < 1 < \rho_U$.

Statistical Methods

Although the methods based on Kolmogorov-Smirnov type of statistics have extensively investigated,^[37] relatively little literature exists on the statistical tests for

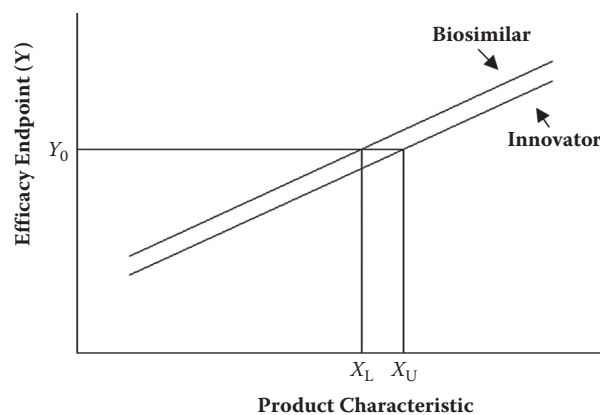


Fig. 3 Parallel line assays to evaluation of extrapolation ability

stochastic equivalence or non-inferiority. One method is to employ the naïve asymptotic confidence band for $\theta = \sup|F(x) - G(x)|$ or $\theta = \sup[F(x) - G(x)]$ as the test statistics for hypotheses Eq. (7) and Eq. (8). If the $(1-2\alpha)100\%$ confidence band is totally contained with the band formed by the equivalence margins $(-\eta, \eta)$, then equivalence between the biosimilar and biological products can be concluded at the α significance level. Similarly, if the $(1-\alpha)100\%$ lower confidence band is above lower band formed by the lower margin $-\eta$, then non-inferiority of biosimilar product to the innovator's biological product can be established at the α significance level. Derivation of the test statistics for hypotheses in Eq. (7) and Eq. (8) at the boundary margins of the null hypothesis and the corresponding distribution and confidence interval require further research. However, permutation and bootstrap technique can also be used to find the distribution of the test statistics and the corresponding confidence intervals empirically.

Because of the nature of Design (a) proposed above, it is in fact an active control trial without a placebo control arm where the follow-on biological product is the test treatment, and its innovator counterpart is the active control treatment. Assuming that the innovator biological product was approved due to its superior efficacy over placebo, the equivalence in PD/efficacy is in fact the equivalence in relative efficacy as compared to the (putative) placebo of the follow-on biological product with the innovator counterpart. On the other hand, the prior information of the comparison of the innovator biological product to the placebo can be incorporated into determination of equivalence margins and the evaluation of equivalence between the follow-on and innovator biological products. The Bayesian design proposed by Simon^[38] may be applied to derive the procedures for assessment of equivalence based on PD/efficacy endpoints.

Calibration models have been used to correlate the surrogate responses with the true endpoints.^[39] Because PK, PD, and efficacy responses are random variables, mixed models was suggested to assessing surrogates as trial endpoints.^[40] The measurement error models can be used to establish calibration models.^[41] With the established calibration model obtained from the data of the training set of Design (a),

$$P_{\text{efpk}} = \Pr(\text{equivalence in efficacy responses} | \text{equivalence in PK responses})$$

and

$$P_{\text{pdpk}} = \Pr(\text{equivalence in PD responses} | \text{equivalence in PK responses})$$

will be estimated for evaluation extrapolation ability of the PK responses using the data from the validation set of Design (a). If P_{efpk} is sufficiently high, then the equivalence in PK responses can be extrapolated to the equivalence in

efficacy, then no further phase III clinical evaluation of biosimilar product based on efficacy responses may not be required. On the other hand, if P_{efpk} is low, then equivalence in PK responses cannot predict the equivalence in efficacy responses and phase III clinical trials for evaluation of follow-on biological product is required. Under Design (a), sample size required for the bioequivalence evaluation based on PK responses of the training set may be different for that the validation set for assessment of extrapolation. It is of interest to compare the sample size required by Design (a) to the total sample size required for the full clinical evaluation of biosimilar product.

For Design (b), the standard statistical method for analysis of parallel line assays can be used to construct the $(1-2\alpha)100\%$ confidence interval for the relative potency in the following steps:

- Step 1: Fit a linear regression equation to the primary efficacy endpoint with the group mean of the product characteristic at each dose level as the independent variable separately for the biosimilar and innovator's biological products. This may be done after a suitable transformation.
- Step 2: If the estimate of the slope of any one product is not significant at the predefined level, then one must conclude that no simple relationship can be established between the product characteristic and the primary endpoint and hence, a full clinical evaluation of the biosimilar product is required. Otherwise, go to Step 3.
- Step 3: Test whether the two estimated simple linear regressions are parallel at the predefined significance level. If the two estimated linear regressions are not parallel, then further clinical evaluation of the biosimilar product is warranted. Otherwise, proceed to Step 4.
- Step 4: Compute estimated relative potency and its corresponding $(1-2\alpha)100\%$ confidence interval. If the $(1-2\alpha)100\%$ confidence interval for the relative potency is within the predefined margins (ρ_L, ρ_U) , then equivalence in the product characteristic can be extrapolated to equivalence in the primary efficacy endpoint at the α significance level. Otherwise, further clinical investigation of the biosimilar product is needed.

The objective of application of Design (b) is not to establish the efficacy of biosimilar product but rather to test whether the relative potency is within some predefined limits. Therefore the required sample size for Design (b) is determined upon the test for positive slope and equivalence margins. Therefore, the sample size required by Design (b) may be fewer than that of Design (a) or of the full clinical evaluation of biosimilar product. However, further theoretical work or simulation studies are needed to address this issue.

ASSESSING SIMILARITY USING GENOMIC DATA

Although factors such as age, gender, education, socioeconomic status, smoking habit, weight, sexual orientation, and underlying disease characteristics at the baseline may contribute to the variation among patients, one of the most important reasons is the genetic or genomic variations among trial participants. As a result, due to genetic variations and genetic-by-environmental interaction, patients respond differently to the same treatment or therapeutic regimen. After completion of the Human Genome Project, the disease targets at the molecular level can be identified and hence biochip products based on heritable DNA markers, mutations, and expression patterns for detection of diseases using the microarray technology is possible. Genomic technologies such as DNA sequencing, mRNA transcript profiling, and comparative genomic hybridization have increased the possibility of identifying those patients who are most likely to benefit from a molecularly targeted drug, and this also indicates an increasing importance of diagnostic tests for identification of molecular targets and an increasing demand for targeted clinical trials conducted for the individualized treatment of patients. A new generation of molecularly targeted agents has been developed and approved by regulatory agencies such as the FDA and EMEA. Many of these drugs benefit only a subset of treated patients and may be overlooked by the traditional, broad-eligibility approach to randomized clinical trials. A major portion of targeted drugs are biological products. It follows that evaluation of biosimilar products should take into account the genomic information. Chow et al.^[42] proposed methods for evaluation bioequivalence using genomic data. Although their methods were derived for the chemical generics, the same principles can be applied to evaluation of biosimilar products. Similarly, the genomic information can be incorporated into the linear regression equations as possible covariates to reduce the variability for estimation and inference about the relative potency.

Because biological products are peptides or protein products with primary, secondary, tertiary, and quaternary structures, immunogenicity is an extremely important safety issue. On the other hand, unlike chemical generic products, no two biological products are exactly the same, even for the different batches of any innovator's product. It follows that evaluation of immunogenicity of biosimilar product is more important and difficult than chemical generic products and usually requires large clinical trials. However, identification of potential immunodominant positions and prediction of antigenic variants may provide a way to evaluation of immunogenicity of biosimilarity product without extensive clinical immunogenicity trials.^[43,44] Because the linear sequence of the primary structure usually determines the tertiary structure, one can conduct pairwise comparisons between the amino acid sequences

of the biosimilar product with those of the innovator's biological product and use the antigenic distance as a measure for evaluation of the similarity of the amino acid sequences between the biosimilar and innovator's products. If the potential antigenic sites of the amino acid sequences are known predictors of antigenic variants, this information along with other data can be used for equivalence evaluation of the amino acid sequences between the biosimilar and innovator's biological products. The cut-off margin based on the antigenic distance is 4, as recommended by Lee and Chen.^[45] However, more stringent margin could be applied if the potential antigenic sites are known to induce serious immunogenicity reactions.

CONCLUDING REMARKS

Because of the size and complexity of the active ingredients and the nature of manufacturing process, biological products are different from the traditional chemical drugs with small molecular weights. Many have pointed out that no two biological products are the same.^[6,9] It is also important to know that traditional chemical generics are not exact the same as their corresponding innovator's product either. They are different in the excipients and their compositions and manufacturing methods and processes. In addition, no two batches of the same innovator's biological products are the same. This is why International Conference on Harmonization (ICH) issued Q5E Guideline on Comparability of Biotechnological/Biological Products Subject to Changes in their Manufacturing Process^[46] issued in 2005 to address these issues the post-approval changes. The same principles for evaluating the post-approval changes for the innovator's biological products can and should be also applied to assess the similarity between the biosimilar and innovator's biological products because the change is the manufacturers of the product.

Under the authorization of the Drug Price Competition and Patent Restoration Act (Hatch-Waxman Act), in 1984 the FDA started to approve traditional chemical generic drug products through the ANDA process. Although at the beginning of the implementation of Hatch-Waxman Act there were serious concerns about whether the chemical generic products could deliver equivalent efficacy and safety, experience accumulated over the past 25 years indicates that the fundamental bioequivalence assumption is a sound basis for approval chemical generic products without conducting costly and time-consuming clinical trials. However, no similar legislation has been passed in the United States or the European Union to address those biological products that are much more expensive than the traditional chemical drug products. Therefore, approval of biosimilar products in Europe is based on some lower-ranking guidance that fails to provide specific scientific requirements. Approvals of Valtropin® and Omnitrope® by the FDA and the EMEA require the long-term clinical and safety data

obtained from multi-county, multi-center randomized clinical trials. These two applications are in fact two full-blown NDAs. Therefore, reduction in price for Valtropin® and Omnitrope® is very limited. Furthermore, due to the requirement of costly and long-term clinical trials, accessibility of biosimilar products to the needed patients is unnecessarily and severely delayed. It hopes that both the U.S. Congress and the E.U. Parliament can pass the necessary legislation required for approval of biosimilar products.

One key scientific issue is the search for product characteristics that are strongly correlated with efficacy, immunogenicity, and safety so that the fundamental biosimilar assumption can be verified. The crucial statistical methodology that should be rapidly developed is to evaluate the extrapolation ability from equivalence in product characteristics to equivalence in efficacy or immunogenicity. Until the above issues are satisfactorily and adequately addressed, biosimilar products will be still approved on an individual basis in accordance with the requirement of clinical data without any possibility of reducing price and increasing patients' accessibility.

DISCLAIMER

The views expressed in this entry are personal opinions of the authors and may not necessarily represent the position of Duke University School of Medicine and/or National Taiwan University.

REFERENCES

1. Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*, 3rd Ed.; Chapman Hall/CRC Press, Taylor & Francis: New York, 2008.
2. Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* **1987**, *15*, 657–680.
3. FDA. *Guidance on Statistical Approaches to Establishing Bioequivalence*; Center for Drug Evaluation and Research, the US Food and Drug Administration: Rockville, MD, 2001.
4. FDA. *Guidance on Bioavailability and Bioequivalence Studies for Orally Administered Drug Products – General Considerations*; Center for Drug Evaluation and Research, the US Food and Drug Administration: Rockville, MD, 2003.
5. Webber, Keith O. *Biosimilars: Are We There Yet?* Presented at Biosimilars 2007, George Washington University: Washington, DC, 2007.
6. Schellekens, H. How similar do 'biosimilar' need to be? *Nat. Biotechnol.* **2004**, *22*, 1357–1359.
7. Chirino, A.J.; Mire-Sluis, A. Characterizing biological products and assessing comparability following manufacturing changes. *Nat Biotechnol* **2004**, *22*, 1383–1391.
8. Crommelin, D.; Bermejo, T.; Bissig, M.; Damianns, J.; Kramer, I.; Rambourg, P.; Scroccaro, G.; Strukelj, B.; Tredree, R.; Ronco, C. Biosimilars, generic versions of the first generation of therapeutic proteins: do they exist? *Contrib. Nephrol.* **2005**, *149*, 287–294.
9. Roger, S.D.; Mikhail, A. Biosimilars: opportunity or cause for concern? *J. Pharm. Sci.* **2007**, *10*, 405–410.
10. Kuhlmann, M.; Covic, A. The protein science of biosimilars. *Nephrol. Dial. Transplant.* **2006**, *21* (suppl 5), v4–v8.
11. EMEA. *Note for Guidance on Comparability of Medicinal Products Containing Biotechnology-derived Proteins as Drug Substance – Non Clinical and Clinical Issues*; The European Medicines Agency Evaluation of Medicines for Human Use, EMEA/CHMP/3097/02: London, 2003.
12. EMEA. *Rev. 1 Guideline on Comparability of Medicinal Products Containing Biotechnology-derived Proteins as Drug Substance – Quality Issues*; The European Medicines Agency Evaluation of Medicines for Human Use, EMEA/CHMP/ BWP/3207/00/Rev 1: London, 2003.
13. EMEA. *Guideline on Similar Biological Medicinal Products*; The European Medicines Agency Evaluation of Medicines for Human Use, EMEA/CHMP/437/04: London, 2005.
14. EMEA. *Draft Guideline on Similar Biological Medicinal Products Containing Biotechnology-derived Proteins as Drug Substance: Quality Issues*; The European Medicines Agency Evaluation of Medicines for Human Use, EMEA/CHMP/49348/05: London, 2005.
15. EMEA. *Draft Annex Guideline on Similar Biological Medicinal Products Containing Biotechnology-derived Proteins as Drug Substance – Non Clinical and Clinical Issues – Guidance on Biosimilar Medicinal Products containing Recombinant Erythropoietins*; The European Medicines Agency Evaluation of Medicines for Human Use, EMEA/CHMP/94526/05: London, 2005.
16. EMEA. *Draft Annex Guideline on Similar Biological Medicinal Products Containing Biotechnology-derived Proteins as Drug Substance – Non Clinical and Clinical Issues – Guidance on Biosimilar Medicinal Products containing Recombinant Granulocyte-Colony Stimulating Factor*; The European Medicines Agency Evaluation of Medicines for Human Use, EMEA/CHMP /31329 /05: London, 2005.
17. EMEA. *Draft Annex Guideline on Similar Biological Medicinal Products Containing Biotechnology-derived Proteins as Drug Substance – Non-Clinical and Clinical Issues – Guidance on Biosimilar Medicinal Products containing Somatropin*; The European Medicines Agency Evaluation of Medicines for Human Use, EMEA/CHMP/94528/05: London, 2005.
18. EMEA. *Draft Annex Guideline on Similar Biological Medicinal Products Containing Biotechnology-derived Proteins as Drug Substance – Non Clinical and Clinical Issues – Guidance on Biosimilar Medicinal Products containing Recombinant Human Insulin*; The European Medicines Agency Evaluation of Medicines for Human Use, EMEA/CHMP/32775/05: London, 2005.

19. EMEA. *Guideline on the Clinical Investigating of the Pharmacokinetics of Therapeutic Proteins*; The European Medicines Agency Evaluation of Medicines for Human Use, EMEA/CHMP/89249/04: London, 2005.
20. Woodcock, J. The FDA's assessment of follow-up protein products: a historical perspective. *Nature Review Drug Discovery* **2007**, *6*, 437–442.
21. ICH. *Q5C Guideline on Quality of Biotechnological Products: Stability Testing of Biotechnological/Biological Products*; Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, the US Food and Drug Administration: Rockville, MD, 1996.
22. ICH. *Q6B Guideline on Test Procedures and Acceptance Criteria for Biotechnological/Biological Products*; Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, the US Food and Drug Administration: Rockville, MD, 1999.
23. ICH. *Q5E Guideline on Comparability of Biotechnological/Biological Products Subject to Changes in Their Manufacturing Process*; Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, the US Food and Drug Administration: Rockville, MD, 2005.
24. EMEA. *Note for Guidance on the Investigation of Bioavailability and Bioequivalence*; The European Medicines Agency Evaluation of Medicines for Human Use, EMEA/EWP/QWP/1401/98: London, 2001.
25. WHO. *World Health Organization Draft Revision on Multisource (Generic) Pharmaceutical Products: Guidelines on Registration Requirements to Establish Interchangeability*; Geneva, Switzerland, 2005.
26. Schall, R.; Luus, H.G. On population and individual bioequivalence. *Statistics in Medicine* **1993**, *12*, 1109–1124.
27. Leeson, L.J. *In Vitro/in vivo* correlation. *Drug Inf. J.* **1995**, *29*, 903–915.
28. FDA. *Guidance for Industry: Dissolution Testing of Immediate Release Solid Oral Dosage Forms*; The United States Food and Drug Administration: Rockville, MD, 1997.
29. Moore, J.W.; Flanner, H.H. Mathematical comparison of curves with an emphasis on dissolution profiles. *Pharm. Technol.* **1996**, *20*, 64–74.
30. SUPAC-IR. The United States Food and Drug Administration Guideline Immediate Release Solid Oral Dosage Forms: Scale-Up and Postapproval Changes: Chemistry, Manufacturing, and Controls, In Vitro Dissolution Testing, and In Vivo Bioequivalence Documentation, Rockville, MD, 1995.
31. Tse, S.K.; Chang, J.Y.; Su, W.L.; Chow, S.C.; Hsiung, C.; Lu, Q. Statistical quality control process for traditional Chinese medicine. *J. Biopharm. Stat.* **2006**, *16*, 861–874.
32. Enis, P.; Geisser, S. Estimation of the probability that $Y < X$. *J. Am. Stat. Assoc.* **1971**, *66*, 162–168.
33. Church, J.D.; Harris, B. The estimation of reliability from stress-strength relationships. *Technometrics* **1970**, *12*, 49–54.
34. Liu, J.P. Statistical evaluation of individual bioequivalence. *Commun. Stat. Theory Methods* **1998**, *27*, 1433–1451.
35. Carrasco, J.L.; Jover, L. Assessing individual bioequivalence using structural equation model. *Stat. Med.* **2003**, *22*, 901–912.
36. Finney, D.J. *Statistical Method in Biological Assay*, 3rd Ed.; Oxford University Press: New York, 1979.
37. Serfling, R.J. *Approximation Theorems of Mathematical Statistics*; John Wiley: New York, 1980.
38. Simon, R. Bayesian design and analysis of active control clinical trials. *Biometrics* **1999**, *55*, 484–487.
39. Sargent, D.F.; Wieand, S.; Haller, D.G., et al. Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies. *J. Clin. Oncol.* **2005**, *23*, 8664–8670.
40. Korn, E.L.; Albert, P.S.; McShane, L.M. Assessing surrogated as trial endpoints using mixed models. *Stat. Med.* **2005**, *24*, 163–182.
41. Cheng, C.L.; Van Ness, J.W. *Statistical Regression with Measurement Errors*; Arnold: London, 1999.
42. Chow, S.C.; Shao, J.; Li, L. Assessing bioequivalence using genomic data. *J. Biopharm. Stat.* **2004**, *14*, 869–880.
43. Lee, M.S.; Chen, M.C.; Liao, Y.C.; Hsiung, A.G. Identifying potential immunodominant positions and predicting Antigenic variants of influenza A/H3N2 viruses. *Vaccine* **2007**, *25*, 8133–8139.
44. Liao, Y.C.; Lee, M.S.; Ko, C.Y.; Hsiung, C.A. Bioinformatics models for predicting variants of influenza A/H3N2 viruses. *Bioinformatics* **2008**, *24*, 505–512.
45. Lee, M.S.; Chen, M.C. Predicting antigenic variants of influenza A/H3N2 viruses. *Emerg. Infect. Dis.* **2004**, *10*, 1385–1390.
46. ICH. E5 Guideline on Ethnic Factors in Acceptability of Foreign Data. *Fed. Regist.* **1998**, *83*, 31790–31796.

Bivariate Zero-Inflated Poisson Populations: Statistical Test for Homogeneity

Siu-Keung Tse and Hak-Keung Yuen

Department of Management Sciences, City University of Hong Kong, Hong Kong, China

Abstract

The theme of this entry is on the discussion of the testing of treatment difference in the occurrence of a study end point. In particular, a randomized parallel-group comparative clinical trial with repeated responses under the assumption that the responses follow a bivariate zero-inflated Poisson (ZIP) distribution is considered. Result based on likelihood ratio test for homogeneity of two bivariate ZIP populations is presented. Approximate formula for sample size calculation is also obtained, which achieves a desired power for detecting a clinically meaningful difference under an alternative hypothesis. An example concerning the comparison of treatment effect in an addictive clinical trial in terms of the number of days of illicit drug use during a month is given for illustration purpose.

Keywords: Bivariate ZIP model; Erosion counts; Likelihood ratio test; Repeated responses; Sample size calculation.

INTRODUCTION

In clinical trials, the occurrence of certain events is commonly used as a surrogate end point for assessment of clinical outcomes (safety and/or efficacy) of a test treatment under the assumption that the surrogate end point is predictive of clinical outcomes. Commonly considered occurrences of certain events in clinical trials include, but are not limited to, the number of tumors observed during the course of the study in cancer trials, the number of erosions seen in gastrointestinal (GI) studies, the number of days of antibiotic use in infectious diseases (ID) clinical trials (which is an indication of efficacy), and the number of days of illicit drug use in addictive drug abuse trials (which is predictive of the treatment effect on addiction of drug abuse). These counts are usually considered as markers for predictive of clinical outcomes (i.e., safety and/or efficacy), if they are not directly related to clinical outcomes. For example, the number of tumors in cancer trials could be an indication of metastasis of the cancer. A large number of erosions most likely will lead to gastroduodenal ulcers and consequently serious GI toxicity (which is considered as adverse events for safety concern). The reduction of the number of days of antibiotic use is an indication of efficacy of the test treatment for treating certain infections. A small percent of illicit drug use is an indication of recovery from the addiction of drug abuse.

These counts are usually assumed to follow a Poisson distribution and can be based to assess whether there is any difference in mean counts between treatment groups. However, these counts could include a large proportion of zeros, and the proportion of zeros could be as high as 60%

in practice. Thus, a Poisson model may not be able to provide a good fit to the data. To accommodate this unusually large proportion of zeros, a zero-inflated Poisson (ZIP) model can be used for an accurate and reliable assessment of the treatment effect.

ZIP model has been used commonly to model the occurrence of unusually large number of zeros. Neyman^[1] and Feller^[2] first introduced the concept of zero-inflation to address the problem of excess zeros. Since then, there have been extensive studies related to the development of ZIP modeling. Van den Broek^[3] discussed the test of zero-inflation and proposed to use the Rao's score test for hypotheses testing. A simulation study was conducted to show that the score test was appropriate. Properties and inference of the ZIP distribution, including the maximum likelihood estimates (MLEs) of the parameters, can be found in the study by Gupta, Gupta and Tripathi.^[4] A review of the related literature was given in the study by Böhning^[5] and examples from different disciplines were also discussed. Bayesian analysis of ZIP can be found in the study by Angers and Biswas.^[6] Furthermore, Lambert;^[7] Böhning et al.;^[8] Dobbie and Welsh;^[9] and Lee, Wang, and Yau^[10] discussed the examples of ZIP regression model under various settings. More specifically, Jansakul and Hinde^[11] modified the score test to the regression case where the zero-mixing probability is dependent on covariates. A review of published works on the standard ZIP model can be found in Lee et al.,^[12] which also discussed the use of a multilevel ZIP regression model incorporating random effects to account for the data dependence arising from repeated measures or longitudinal studies.

For the comparison of two ZIP models, Lachenbruch^[13,14] discussed the hypotheses testing of two two-part models, where the framework incorporated the comparison of two ZIP models. In particular, these studies compared the size, the power, and the required sample size to achieve a desired power for several methods including the t-test and the Wilcoxin–Mann–Whitney rank sum test. Given two ZIP models, there are three possible scenarios: 1) there is no difference between treatments in the group of zeros; 2) there is no difference between treatments in the group of non-zeros; and 3) there are no differences between treatments in both groups of zeros and non-zeros. Tse et al.^[15] assumed that these counts follow a ZIP distribution and test the null hypothesis that there is no treatment difference in mean counts between treatment groups at the end of the study. Under a ZIP distribution, Tse et al.^[15] derived valid statistical tests under the above three possible scenarios.

However, their tests focus on the assessment at the end of the study and ignore the treatment effect during the process. In practice, multiple assessments are often considered to monitor the performance of the test treatment under investigation to determine when the test treatment will achieve the optimal therapeutic effect prior to the end of the treatment duration. Majumdar and Gries^[16] discussed a Bayesian approach to model bivariate count data with excess zeros. They used simulations to compare the performance of their Bayesian and classical ZIP approaches based on regression models. These models do not give a direct way to derive the required sample size to achieve a desired power for detecting a clinically meaningful difference under an alternative hypothesis. However, in many clinical applications, outcomes with unusually large number of zeros are often taken repeatedly on the same individual. These repeated outcomes certainly possess a certain correlated structure and analysis using univariate ZIP models would produce misleading conclusion, since the correlated structure is not properly addressed in the analysis process. Thus, there is a need to motivate the search of proper multivariate ZIP (MZIP) model to model these data. Early results on the discussion of MZIP model can be found in the study by Skellman,^[17] which are based on the idea of extending the univariate Poisson binomial distribution to a compound distribution of the binomial and the Poisson. Li et al.^[18] discussed several other ways to construct MZIP distribution and presented real-life applications of the MZIP in equipment-fault detection. However, as inherent in all multivariate distributions, the number of model parameters would increase rapidly in order to describe the covariance structure of the multivariate observations as the dimension increases. Large number of model parameters not only requires a large number of observations but also imposes difficulty in the estimation of the model parameters. Therefore, it is desirable to choose a MZIP model such that the corresponding number of model parameters is in a manageable size.

Yuen, Chow, and Tse^[19] proposed a bivariate model with five parameters and derived valid statistical test

for assessment of treatment effect, which takes multiple assessments into consideration. In particular, they proposed a bivariate ZIP model to compare the effects of two treatments in a two-armed clinical trial and derived approximate formula to determine the required sample size to achieve a desired power for two given bivariate ZIP models.

This entry aims to summarize the results developed by Yuen, Chow, and Tse.^[19] In the next section, a bivariate ZIP model with five parameters is presented and the corresponding notations are given in details. Results of the likelihood ratio test under a randomized parallel-group design comparing a test treatment with a control assuming that the responses follow a bivariate ZIP distribution are given in section “Likelihood Ratio Tests for the Comparison of Treatments.” The section “Determination of Sample Size” presents an approximate formula for sample size calculation which achieves a desired power for detecting a clinically meaningful difference at a prespecified level of significance. To illustrate the application of the results, section “An Example” presents an example concerning the comparison of randomization effects in terms of the number of days using illicit drug during a month is given for illustration purpose. Some concluding remarks are given in the section “Conclusion.”

A BIVARIATE ZIP MODEL

A bivariate ZIP (Y_1, Y_2) can be constructed from a mixture of a point mass at $(0, 0)$, two univariate Poisson distributions with parameters λ_1 and λ_2 , and a bivariate Poisson distribution with parameters $(\lambda_{10}, \lambda_{20}, \lambda_{00})$ as follows:

$$\begin{aligned} (Y_1, Y_2) &\sim (0, 0), && \text{with probability } p_0 \\ &\sim (\text{Poisson}(\lambda_1), 0), && \text{with probability } p_1 \\ &\sim (0, \text{Poisson}(\lambda_2)), && \text{with probability } p_2 \\ &\sim \text{bivariate Poisson}(\lambda_{10}, \lambda_{20}, \lambda_{00}), && \\ &&& \text{with probability } p_3 = 1 - p_0 - p_1 - p_2 \end{aligned} \quad (1)$$

Let $\lambda_1 = \lambda_{10} + \lambda_{00}$ and $\lambda_2 = \lambda_{20} + \lambda_{00}$. The probability function of the bivariate ZIP (Y_1, Y_2) is given as:

$$\begin{aligned} P(Y_1 = 0, Y_2 = 0) &= p_0 + p_1 \exp(-\lambda_1) + p_2 \exp(-\lambda_2) \\ &\quad + p_3 \exp(-\lambda_{10} - \lambda_{20} - \lambda_{00}) \\ P(Y_1 = y_1, Y_2 = 0) &= [p_1 \lambda_1^{y_1} \exp(-\lambda_1) + p_3 \lambda_{10}^{y_1} \exp(-\lambda_{10} - \lambda_{20} - \lambda_{00})] / y_1! \\ P(Y_1 = 0, Y_2 = y_2) &= [p_2 \lambda_2^{y_2} \exp(-\lambda_2) + p_3 \lambda_{20}^{y_2} \exp(-\lambda_{10} - \lambda_{20} - \lambda_{00})] / y_2! \\ P(Y_1 = y_1, Y_2 = y_2) &= p_3 \sum_{j=0}^{\min(y_1, y_2)} \lambda_{10}^{y_1-j} \lambda_{20}^{y_2-j} \lambda_{00}^j \frac{\exp(-\lambda_{10} - \lambda_{20} - \lambda_{00})}{(y_1 - j)! (y_2 - j)! j!} \end{aligned} \quad (2)$$

It should be noted that the marginal distribution of Y_1 is a univariate ZIP which is a mixture of point mass at 0 with probability $p_0 + p_2$ and a Poisson distribution with parameter λ_1 with probability $1 - (p_0 + p_2)$; similarly, Y_2 is a mixture of point mass at 0 with probability $p_0 + p_1$ and a Poisson distribution with parameter λ_2 with probability $1 - (p_0 + p_1)$. Furthermore, in many applications, Y_1 and Y_2 can be assumed to be identically distributed, i.e., $\lambda_1 = \lambda_2$. To simplify notations, let $\lambda_1 = \lambda_2 = \lambda$, then $\lambda_{10} = \lambda_{20} = \lambda_0$ and $\lambda_{00} = \lambda - \lambda_0$. Thus, λ , λ_0 , p_0 , p_1 , and p_2 are the parameters required in defining the model.

Suppose that a clinical study is conducted to compare the effect between a test treatment (T) and a control treatment (C). Assume that n_T and n_C subjects are randomly assigned to the test treatment and control treatment, respectively. Let (Y_{iC1}, Y_{iC2}) be the observed responses corresponding to the i^{th} subject in the control treatment group, $i = 1, 2, \dots, n_C$. Similarly, (Y_{iT1}, Y_{iT2}) are the observed responses corresponding to the i^{th} subject in the test treatment group, $i = 1, 2, \dots, n_T$. Furthermore, assume that (Y_{iC1}, Y_{iC2}) follows a bivariate ZIP model with parameters $\lambda_c, \lambda_{0c}, p_{0c}, p_{1c}$, and p_{2c} for the control treatment, and (Y_{iT1}, Y_{iT2}) follows a bivariate ZIP model with parameters $\lambda_T, \lambda_{0T}, p_{0T}, p_{1T}$, and p_{2T} for the test treatment.

To facilitate the comparison of the two treatments, consider the following transformation of parameters.

$$\begin{aligned}\theta_1 &= \lambda_{0T} - \lambda_{0C} \text{ and } \theta_6 = \lambda_{0C} \Rightarrow \lambda_{0T} = \theta_1 + \theta_6 \\ \theta_2 &= \lambda_T - \lambda_C \text{ and } \theta_7 = \lambda_C \Rightarrow \lambda_T = \theta_2 + \theta_7 \\ \theta_3 &= p_{0T} - p_{0C} \text{ and } \theta_8 = p_{0C} \Rightarrow p_{0T} = \theta_3 + \theta_8 \\ \theta_4 &= p_{1T} - p_{1C} \text{ and } \theta_9 = p_{1C} \Rightarrow p_{1T} = \theta_4 + \theta_9 \\ \theta_5 &= p_{2T} - p_{2C} \text{ and } \theta_{10} = p_{2C} \Rightarrow p_{2T} = \theta_5 + \theta_{10}\end{aligned}\quad (3)$$

Given observations (y_{ij1}, y_{ij2}) , $j = T, C$, $i = 1, 2, \dots, n_j$, define the index sets:

$$\begin{aligned}I_{0j} &= \{i : y_{ij1} = 0 \text{ and } y_{ij2} = 0; i = 1, 2, \dots, n_j\} \\ I_{1j} &= \{i : y_{ij1} \neq 0 \text{ and } y_{ij2} = 0; i = 1, 2, \dots, n_j\} \\ I_{2j} &= \{i : y_{ij1} = 0 \text{ and } y_{ij2} \neq 0; i = 1, 2, \dots, n_j\} \\ \text{and } I_{3j} &= \{i : y_{ij1} \neq 0 \text{ and } y_{ij2} \neq 0; i = 1, 2, \dots, n_j\}\end{aligned}\quad (4)$$

For $i \in I_{kj}$, $j = T, C$ and $k = 0, 1, 2, 3$, let P_{ijk} denotes the probability that the observed response is (y_{ij1}, y_{ij2}) . Then, P_{ijk} is given by:

$$\begin{aligned}P_{iC0} &= \theta_8 + (\theta_9 + \theta_{10}) \exp(-\theta_7) + (1 - \theta_8 - \theta_9 - \theta_{10}) \\ &\quad \exp(-(\theta_6 + \theta_7)) \\ P_{iC1} &= \left[\theta_9 \theta_7^{y_{iC1}} \exp(-\theta_7) + (1 - \theta_8 - \theta_9 - \theta_{10}) \theta_6^{y_{iC1}} \right. \\ &\quad \left. \exp(-(\theta_6 + \theta_7)) \right] / y_{iC1}! \\ P_{iC2} &= \left[\theta_{10} \theta_7^{y_{iC2}} \exp(-\theta_7) + (1 - \theta_8 - \theta_9 - \theta_{10}) \theta_6^{y_{iC2}} \right. \\ &\quad \left. \exp(-(\theta_6 + \theta_7)) \right] / y_{iC2}!\end{aligned}$$

$$\begin{aligned}P_{iC3} &= (1 - \theta_8 - \theta_9 - \theta_{10}) \sum_{j=0}^{\min(y_{iC1}, y_{iC2})} \theta_1^{y_{iC1} + y_{iC2} - 2j} \\ &\quad (\theta_7 - \theta_6)^j \frac{\exp(-(\theta_6 + \theta_7))}{(y_{iC1} - j)!(y_{iC2} - j)!j!}\end{aligned}\quad (5)$$

and

$$\begin{aligned}P_{iT0} &= (\theta_3 + \theta_8) + (\theta_4 + \theta_5 + \theta_9 + \theta_{10}) \\ &\quad \exp(-(\theta_2 + \theta_7)) + (1 - u) \exp(-v) \\ P_{iT1} &= \left[(\theta_4 + \theta_9) (\theta_2 + \theta_7)^{y_{iT1}} \exp(-(\theta_2 + \theta_7)) \right. \\ &\quad \left. + (1 - u) (\theta_1 + \theta_6)^{y_{iT1}} \exp(-v) \right] / y_{iT1}! \\ P_{iT2} &= \left[(\theta_5 + \theta_{10}) (\theta_2 + \theta_7)^{y_{iT2}} \exp(-(\theta_2 + \theta_7)) \right. \\ &\quad \left. + (1 - u) (\theta_1 + \theta_6)^{y_{iT2}} \exp(-v) \right] / y_{iT2}! \\ P_{iT3} &= (1 - u) \sum_{j=0}^{\min(y_{iT1}, y_{iT2})} (\theta_1 + \theta_6)^{y_{iT1} + y_{iT2} - 2j} \\ &\quad (\theta_2 + \theta_7 - \theta_1 - \theta_6)^j \frac{\exp(-v)}{(y_{iT1} - j)!(y_{iT2} - j)!j!}\end{aligned}\quad (6)$$

with $u = \theta_3 + \theta_4 + \theta_5 + \theta_8 + \theta_9 + \theta_{10}$ and $v = \theta_1 + \theta_2 + \theta_6 + \theta_7$.

Based on these observed responses, the log-likelihood function $\ln L$ is given by:

$$L = \prod_{k=0}^3 \prod_{j=C, T} \prod_{i \in I_{kj}} P_{ijk} \Rightarrow \ln L = \sum_{k=0}^3 \sum_{j=C, T} \sum_{i \in I_{kj}} \ln P_{ijk}\quad (7)$$

Denote $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10})$. Then, the MLEs $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4, \hat{\theta}_5, \hat{\theta}_6, \hat{\theta}_7, \hat{\theta}_8, \hat{\theta}_9, \hat{\theta}_{10})$ of θ can be found by solving the likelihood equations $\frac{\partial \ln L}{\partial \theta_i} = 0$;

$i = 1, \dots, 10$, which are derived by taking the first order derivatives of the log-likelihood function given in Eq. 6 with respect to the unknown parameters. Then, the MLEs of the model parameters $\lambda_c, \lambda_{0c}, p_{0c}, p_{1c}$, and p_{2c} for the control and $\lambda_T, \lambda_{0T}, p_{0T}, p_{1T}$, and p_{2T} for the test treatment can be found easily using the invariance principle. Without loss of generality, assume that both the control and test treatments have the same sample size, i.e., $n_C = n_T = n$. Denote the Fisher information matrix by:

$$I(\theta) = E \left[- \frac{\partial^2 \ln L}{\partial \theta \partial \theta^T} \right]\quad (8)$$

Let,

$$I_1(\theta) = \frac{1}{n} I(\theta)\quad (9)$$

then $I_1(\theta)$ is the Fisher information per unit. The inverse of $I(\theta)$ given in Eq. 8 is the asymptotic covariance matrix of the MLE $\hat{\theta}$ of the model parameters θ . Thus, the

corresponding diagonal elements of $I^{-1}(\theta)$ would be the asymptotic variances of the MLE $\hat{\theta}$. Statistical inference of the individual parameters can be drawn based on this information.

LIKELIHOOD RATIO TESTS FOR THE COMPARISON OF TREATMENTS

In many clinical applications, it is common to test whether there is any difference between a control treatment and a test treatment. If the responses are bivariate ZIP distributed, the problem is formulated as the testing of the following hypotheses:

$$H_0 : \lambda_C = \lambda_T, \lambda_{0C} = \lambda_{0T}, p_{0C} = p_{0T}, p_{1C} = p_{1T}, p_{2C} = p_{2T} \text{ versus } H_A : \text{not all are equal} \quad (10)$$

However, using the transformation given in Eq. 3, this is equivalent to test:

$$H_0 : \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0 \text{ versus } H_A : \text{Some of these } \theta_i \text{ are not equal to 0} \quad (11)$$

Likelihood ratio test method is used to develop test procedure for these hypotheses. Denote $\theta_0 = (0, 0, 0, 0, 0, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10})$ and $\hat{\theta}_0 = (0, 0, 0, 0, 0, \hat{\theta}_{06}, \hat{\theta}_{07}, \hat{\theta}_{08}, \hat{\theta}_{09}, \hat{\theta}_{010})$, where $\hat{\theta}_{0i}, i = 6, \dots, 10$ are the MLE of $\theta_i, i = 6, \dots, 10$ under H_0 .

Partition the matrix $i(\theta)$ given in Eq. 9 into:

$$i(\theta) = \begin{pmatrix} i_{aa}(\theta) & i_{ab}(\theta) \\ i_{ba}(\theta) & i_{bb}(\theta) \end{pmatrix}_{10 \times 10} \quad (12)$$

$$\text{where } i_{aa}(\theta) = -\frac{1}{n} E \left(\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right)_{i=1, \dots, 5; j=1, \dots, 5} \quad (13)$$

$$i_{bb}(\theta) = -\frac{1}{n} E \left(\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right)_{i=6, \dots, 10; j=6, \dots, 10} \quad (14)$$

$$\text{and } i_{ab}(\theta) = -\frac{1}{n} E \left(\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right)_{i=1, \dots, 5; j=6, \dots, 10} \quad (15)$$

Based on the results of Cox and Hinkley,^[20] the likelihood ratio test statistic for the hypotheses given in Eq. 11 is:

$$W = 2 \left[\ln L(\hat{\theta}) - \ln L(\hat{\theta}_0) \right] \quad (16)$$

It is easy to verify that the regularity conditions stated in the study by Lehman and Casella^[21] are satisfied. Based on Wilks Theorem, the likelihood ratio statistic W asymptotically follows a central χ^2_5 under the null hypothesis H_0 and a

non-central chi-squared distribution with a non-centrality parameter δ under the alternative hypothesis H_A , where:

$$\delta^2 = n(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) \begin{bmatrix} i_{aa}(\theta) - i_{ab}(\theta) i_{bb}^{-1}(\theta) i_{ba}(\theta) \end{bmatrix} \begin{pmatrix} \theta_1, \theta_2, \theta_3, \theta_4, \theta_5 \end{pmatrix}^T \quad (17)$$

Thus, the null hypothesis H_0 is rejected at a significance level α if $W > \chi^2_5(\alpha)$, where $\chi^2_5(\alpha)$ is the upper α -quantile of a χ^2_5 distribution. Based on this result, for a given level α , the critical region is given by:

$$\left\{ (y_{ij1}, y_{ij2}), i = 1, 2, \dots, n_j; j = T, C, \mid W > \chi^2_5(\alpha) \right\} \quad (18)$$

and the power function of the test is given by:

$$1 - \beta = P \left\{ W > \chi^2_5(\alpha) \mid H_A \right\} \quad (19)$$

DETERMINATION OF SAMPLE SIZE

In clinical study, it is of interest to determine the required sample size such that the corresponding sample size can achieve a desired power $(1 - \beta)$ for detecting a clinically meaningful difference of the parameters at a prespecified level of significance α . Thus, based on the power function given in Eq. 19, the required sample size n can be determined in such a way that the resulting sample would achieve a given power $(1 - \beta)$. Using normal approximation to the probability given in Eq. 19 (details can be found in Johnson, Kotz and Kemp^[22]):

$$(2 + \delta^2)^{\frac{1}{2}} \left[\left(\frac{W - 1/3}{2 + \delta^2} \right)^{\frac{1}{2}} - \left(1 - \frac{1}{3(2 + \delta^2)} \right)^{\frac{1}{2}} \right] \quad (20)$$

or equivalently,

$$\left(W - \frac{1}{3} \right)^{\frac{1}{2}} - \left(\frac{5 + 3\delta^2}{3} \right)^{\frac{1}{2}} \quad (21)$$

approximately follows a standard normal distribution. Therefore, based on Eq. 19,

$$\left(\chi^2_5(\alpha) - \frac{1}{3} \right)^{\frac{1}{2}} - \left(\frac{5 + 3\delta^2}{3} \right)^{\frac{1}{2}} = -z(\beta) \quad (22)$$

where $z(\beta)$ is the upper β -quantile of a standard normal distribution. Simplification of Eq. 22 gives a formula of the sample size for achieving a $(1 - \beta)$ power as follows:

$$n \geq C^{-1} \left\{ \left[\left(\chi^2_5(\alpha) - \frac{1}{3} \right)^{\frac{1}{2}} + z(\beta) \right]^2 - \frac{5}{3} \right\} \quad (23)$$

where $C = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) [i_{aa}(\theta) - i_{ab}(\theta)i_{bb}^{-1}(\theta)i_{ba}(\theta)] (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)^T$.

For illustration purpose, Yuen, Chow, and Tse^[19] listed the required sample size n for selected combinations of parameter values using $\alpha = 0.05$ and $1 - \beta = 0.8$. In general, their results suggest that the difference of λ_C and λ_T plays an important role in determining the required sample size to discriminate the two bivariate ZIP models. When λ_C and λ_T are close, it requires a very large sample size to achieve the nominal power level.

Since the results presented in Eq. 23 are based on approximation, Yuen, Chow, and Tse^[19] conducted a simulation study to assess the accuracy of these results. Their study showed that the simulated powers are higher than the nominal power level for cases with sample sizes larger than 20. Thus, the required sample sizes derived by the approximation results would give the desired power level, albeit in a somewhat conservative manner. For cases with sample sizes less than 20, the simulated power is lower than the nominal value. There are two possible causes leading to this pattern. One plausible explanation is that the derived sample sizes may be too small to support the estimation of the relatively large number of parameters in the model since there are 10 unknown parameters in the model. On the other hand, it may also due to the fact that the asymptotic approximation results only work for cases with relatively large sample size, say, at least 20.

AN EXAMPLE

A clinical trial for evaluation of a test treatment for patients with addictive drug or substance abuse to illustrate the application of the proposed statistical test described in the previous section. Addiction is a chronic, often relapsing brain disease that causes compulsive drug seeking and use, despite harmful consequences to the addicted individuals and those around him or her. Drug or substance

abuse and addiction have negative consequences for individuals and for society such as family disintegration, loss of employment, failure in school, domestic violence, and child abuse.

In this clinical study, 156 qualified subjects were randomly assigned to receive a test treatment and another 139 subjects were assigned to receive an active control treatment in conjunction with behavioral therapy. The treatment duration was for six months. Clinical assessment such as measurements for social behavior and safety implications of drug abuse and addiction was done at 3 months and 6 months (end point visit). In addition, individual patient diary was used to capture the number of days of illicit drug use at the third month and the sixth month. At the end of the study, of the 156 subjects under the test treatment, 51 of them are with zero values (0, 0); on the other hand, 45 of the 139 control subjects are with (0, 0) values. The corresponding data are listed in Table 1.

In particular, the bivariate ZIP model defined in the section "A Bivariate ZIP Model" is used to model these bivariate data and the likelihood ratio test developed in the section "Likelihood Ratio Tests for the Comparison of Treatments" is used to assess whether there is any difference in the therapeutic effect between the two treatments, namely, control and test treatments. Thus, the assessment is done by comparing the two bivariate ZIP models. In particular, following the model formulation outlined in section III, the hypotheses of interests are:

$$H_0: \lambda_C = \lambda_T, \lambda_{0C} = \lambda_{0T}, P_{0C} = P_{0T}, P_{1C} = P_{1T}, P_{2C} = P_{2T} \text{ versus } H_A: \text{not all are equal} \quad (24)$$

which is equivalent to:

$$H_0: \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0 \text{ versus } H_A: \text{Some of these } \theta_i \text{ are not equal to 0} \quad (25)$$

Based on the data, the unrestricted MLEs of the parameters are $\hat{\theta} = (-0.666, -0.777, 0.001, -0.105, -0.001, 3.034,$

Table 1 Data used in the example

I Observed data under control treatment ($n_c = 139$)

45 subjects are with values (0, 0) and the others are listed in the following.

(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)
(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 2)	(0, 4)	(0, 11)	(0, 13)
(1, 0)	(1, 0)	(1, 0)	(1, 0)	(1, 0)	(1, 0)	(1, 0)	(1, 0)	(1, 0)	(1, 0)
(1, 0)	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 4)	(2, 0)	(2, 0)	(2, 0)	(2, 0)
(2, 1)	(2, 1)	(2, 2)	(2, 2)	(2, 3)	(2, 5)	(2, 6)	(2, 10)	(2, 13)	(3, 0)
(3, 0)	(3, 0)	(3, 1)	(3, 3)	(3, 5)	(3, 9)	(3, 11)	(4, 0)	(4, 0)	(4, 0)
(4, 0)	(4, 3)	(4, 4)	(4, 10)	(5, 0)	(5, 0)	(5, 4)	(5, 4)	(5, 6)	(5, 11)
(6, 0)	(6, 0)	(6, 2)	(6, 8)	(8, 0)	(8, 1)	(8, 4)	(8, 10)	(9, 1)	(9, 4)
(9, 8)	(9, 10)	(9, 14)	(11, 6)	(12, 0)	(12, 6)	(12, 12)	(13, 0)	(13, 9)	(13, 11)
(13, 13)	(14, 0)	(14, 9)	(14, 13)						

(Continued)

Table 1 Data used in the example (*Continued*)II Observed data under test treatment ($n_T = 156$)

51 subjects are with values (0, 0) and the others are listed in the following.

(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 1)	(0, 2)	(0, 2)	(0, 2)
(0, 2)	(0, 2)	(0, 2)	(0, 2)	(0, 2)	(0, 3)	(0, 3)	(0, 3)	(0, 3)	(0, 4)
(0, 4)	(0, 5)	(0, 10)	(1, 0)	(1, 0)	(1, 0)	(1, 1)	(1, 1)	(1, 1)	(1, 1)
(1, 2)	(1, 2)	(1, 2)	(1, 2)	(1, 2)	(1, 2)	(1, 3)	(1, 4)	(1, 4)	(1, 8)
(1, 10)	(2, 0)	(2, 0)	(2, 0)	(2, 0)	(2, 0)	(2, 1)	(2, 1)	(2, 1)	(2, 2)
(2, 3)	(2, 3)	(2, 4)	(2, 4)	(2, 4)	(2, 6)	(2, 6)	(3, 0)	(3, 0)	(3, 0)
(3, 0)	(3, 1)	(3, 2)	(3, 3)	(3, 3)	(3, 3)	(3, 4)	(3, 6)	(4, 0)	(4, 0)
(4, 2)	(4, 3)	(4, 3)	(4, 4)	(4, 6)	(4, 7)	(4, 8)	(4, 9)	(5, 1)	(5, 1)
(5, 2)	(5, 3)	(5, 4)	(5, 5)	(6, 0)	(6, 4)	(6, 6)	(6, 14)	(8, 0)	(8, 0)
(8, 2)	(8, 5)	(8, 8)	(8, 11)	(9, 4)	(9, 8)	(10, 0)	(10, 5)	(10, 8)	(11, 1)
(11, 8)	(11, 9)	(11, 13)	(13, 10)	(13, 13)					

4.6737, 0.320, 0.212, 0.137). Consequently, the MLEs and the corresponding standard errors (listed inside the brackets) of the model parameters are summarized in the following:

	λ_0	λ	p_0	p_1	p_2
Control	3.034 (0.143)	4.673 (0.111)	0.320 (0.002)	0.212 (0.002)	0.137 (0.002)
Treatment	2.368 (0.067)	3.896 (0.036)	0.321 (0.001)	0.107 (0.001)	0.136 (0.001)

Under H_0 , the MLEs are $\hat{\theta}_0 = (0, 0, 0, 0, 0, 2.623, 4.246, 0.321, 0.157, 0.136)$.

Furthermore, the test statistic $W = 2 \times [-1232.51 - (-1239.77)] = 14.53$, which is larger than $\chi^2_5(0.05) = 11.07$. Therefore, we reject H_0 at a 5% level of significance. The results indicate that there is a significant difference in therapy effect between the test treatment and the control treatment.

CONCLUSION

This entry discusses the problem of testing treatment difference in the occurrence of a study end point in a randomized parallel-group comparative clinical trial with repeated responses. These occurrence counts usually could include a large proportion of zeros and are usually modeled by a ZIP model to give an accurate assessment of the treatment effect. However, there is a need to extend a univariate ZIP model to the multivariate case, since repeated responses are observed in the study. The works in this entry demonstrate that a bivariate ZIP model can be used to compare the effect of two treatments in a two-armed clinical trial. In particular, likelihood ratio test is derived to test the homogeneity of the two bivariate ZIP populations. An approximate formula to determine the

required sample sizes to achieve a desired power is given. Although the discussion of the example and the results given are based on only two repeated responses, it can be shown that the model can be readily extended to more than two repeated responses. However, the number of parameters would inevitably increase, which would require a relatively large data set and impose considerable amount of difficulty on the estimation of the model parameters. In any event, the proposed model provides a viable tool for practitioners to assess whether a new treatment is more effective than another treatment when the responses are repeated counts with usually large proportion of zeros.

REFERENCES

1. Neyman, J. On a new class of contagious distributions applicable in entomology and bacteriology. *Ann. Math. Stat.* **1939**, *10*, 35–57.
2. Feller, W. On a general class of contagious distributions. *Ann. Math. Stat.* **1945**, *12*, 389–400.
3. Van den Broek, J. A score test for zero-inflation in a Poisson distribution. *Biometrics*. **1995**, *51*, 738–743.
4. Gupta, P.L.; Gupta, R.C.; Tripathi, R.C. Analysis of zero-adjusted count data. *Comput. Stat. Data An.* **1996**, *23*, 207–218.
5. Böhning, D. Zero-inflated poisson models and C.A. Man: A tutorial collection of evidence. *Biometrical J.* **1998**, *40*, 833–843.
6. Angers, J.F.; Biswas, A. A Bayesian analysis of zero-inflated generalized Poisson model. *Comput. Stat. Data An.* **2003**, *42*, 37–46.
7. Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **1992**, *34*, 1–14.
8. Böhning, D.; Dietz, E.; Schlattmann, P.; Mendonca, L.; Kirchner, U. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *J. Roy. Stat. Soc. Ser. A* **1999**, *162*, 195–209.

9. Dobbie, M.J.; Welsh, A.H. Modelling correlated zero-inflated count data. *Aust. NZ J. Stat.* **2001**, *43*, 431–444.
10. Lee, A.H.; Wang, K.; Yau, K.K.W. Analysis of zero-inflated Poisson data incorporating extent of exposure. *Biometrical J.* **2001**, *43*, 963–975.
11. Jansakul, N.; Hinde, J.P. Score tests for zero-inflated poisson models. *Comput. Stat. Data An.* **2002**, *40*, 75–96.
12. Lee, A.H.; Wang, K.; Scott, J.A.; Yau, K.K.W.; McLachlan, G.J. Multi-level zero-inflated Poisson regression modeling of correlated count data with excess zeros. *Stat. Methods Med. Res.* **2006**, *15*, 47–61.
13. Lachenbruch, P.A. Comparisons of two-part models with competitors. *Stat. Med.* **2001**, *20*, 1215–1234.
14. Lachenbruch, P.A. Power and sample size requirements for two-part models. *Stat. Med.* **2001**, *20*, 1235–1238.
15. Tse, S.K.; Chow, C.S.; Lu, Q.S.; Cosmatos, D. Testing homogeneity of two zero-inflated poisson populations. *Biometrical J.* **2009**, *51*, 159–170.
16. Majumdar, A.; Gries, C. Bivariate zero-inflated regression for count data: a bayesian approach with application to plant counts. *Int. J. Biostat.* **2010**, *6*, 27.
17. Skellman, J.C. Studies in statistical ecology: 1. Spatial pattern. *Biometrika* **1952**, *39*, 346–362.
18. Li, C.S.; Lu, J.C.; Park, J.; Kim, K.; Brinkley, P.A.; Peterson, J.P. Multivariate zero-inflated poisson models and their applications. *Technometrics* **1999**, *41*, 29–38.
19. Yuen, H.K.; Chow, S.C.; Tse, S.K. On statistical test for homogeneity of two bivariate zero-inflated poisson populations. *J. Biopharm. Stat.* **2015**, *25*, 44–53.
20. Cox, D.R.; Hinkley, D.V. *Theoretical Statistics*; Chapman and Hall: London, 1970.
21. Lehmann, E.L.; Casella, G. *Theory of Point Estimation*; Springer: New York, 1998.
22. Johnson, N.; Kotz, S.; Kemp, A. *Univariate Discrete Distributions, 2nd Ed.*; John Wiley: New York, 1992.

Blinded Sample Size Re-Estimation

Chien-Hua Wu

Department of Applied Mathematics, Chung-Yuan Christian University, Chung-Li, Taiwan

Abstract

The sample size calculation plays an important role in clinical trials. Estimation of sample size in clinical trials requires knowledge of parameters. When there is uncertainty about some of the parameters, it can be challenging to determine the number of subjects required for the study. This article provides a method via the transition probability matrix to re-estimate the sample size based on interim data while the blind procedure is preserved. The examples of categorical response, continuous response and survival data are provided to illustrate the use of proposed methods.

Keywords: Sample Size; Transition Probability Matrix; Continuous Response; Categorical Response; Survival Analysis.

INTRODUCTION

All human subject researches, particularly in large confirmatory Phase III trials, must provide the evidence on the efficacy and safety of medical interventions for filing new drug application submission. The clinical trial protocol blueprints the protection of safeties, rights and welfare of the human subjects, and promises the quality of data collection. Interim analyses play an integral role in clinical trials for the possibility of stopping the trial early for safety concerns, and/or futility, and/or early evidence of efficacy. The 1998 International Conference on Harmonization guidance for industry “E9 Statistical Principles for Clinical Trials”^[1] states that all staff involved in the conduct of the trial should remain blind to the results of analyses, because of the possibility that their attitudes to the trial will be modified and cause changes in the characteristics of patients to be recruited or biases in treatment comparisons. To ensure the confidentiality of interim analysis, the techniques of sample size re-estimation without breaking the blind are proposed by Gould and Shih^[2] using EM algorithm for continuous responses and Shih and Zhao^[3] stratifying the randomization procedure for binary outcomes. Bristol and Shurzinske^[4] propose a method to estimate the variance. Kieser and Friede^[5] provide the procedures for blinded sample size adjustment that do not affect the type I error rate. Xing and Ganju^[6] uses the randomization block to estimate the variance without breaking the blind. Ganju and Xing^[7] extend previous work to estimate the within-group variance of a continuous endpoint. Masking the treatment effect, Shih^[8] perturbs unblinding to estimate the variance. Friede and Kieser^[9] compare the blinded sample size re-estimation procedures with respect to bias and variance of the variance estimators. Using the

dummy strata, Wu^[10] extends the work of Shih and Zhao^[3] to continuous endpoints. Additionally, Wu^[11] suggests the sponsor may stop the trial early because efficacy is unequivocally established in a crossover design. Without loss integrity of clinical trials, the randomization blocks with the primary outcomes are used to estimate the sample size in the middle of clinical trial masking the randomization codes in the released data.

CATEGORICAL RESPONSE

Suppose we have a discrete variable y taking p categories and another variable having b treatment groups in the original data, and a discrete variable x taking p categories and a variable with a randomization blocks in the released data. The responses are relocated from the treatment groups of the original data to the randomization blocks of the released data via the transition probability matrix. Let $\mathbf{y}_j = (y_{j1}, \dots, y_{jp})'$ be the random vector at the j th group satisfied $y_{ik} = 1$ or 0 and $\sum_{k=1}^p y_{jk} = 1$ in the original data, where $j = 1, 2, \dots, b$. Let $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$ be the random vector at the i th block satisfied $x_{ik} = 1$ or 0 and $\sum_{k=1}^p x_{ik} = 1$ in the released data, where $i = 1, 2, \dots, a$. For k th category, a transition probability matrix Θ is defined as a matrix composed of elements θ_{ij} such that the probability

$$\begin{aligned}\pi_{x_{ik}} &= E(x_{ik}) = P(x_{ik} = 1) \\ &= \sum_{j=1}^b P(x_{ik} = 1 | y_{jk} = 1) P(y_{jk} = 1) \\ &= \sum_{j=1}^b \theta_{ij} \pi_{y_{jk}}\end{aligned}$$

where $\theta_{ij} = P(x_{ik} = 1 | y_{jk} = 1)$ and $\pi_{yjk} = P(y_{jk} = 1)$. For simplicity,

$$\begin{pmatrix} \pi_{x_{1k}} \\ \pi_{x_{2k}} \\ \vdots \\ \pi_{x_{ak}} \end{pmatrix} = \begin{pmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1b} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2b} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{a1} & \theta_{a2} & \cdots & \theta_{ab} \end{pmatrix} \begin{pmatrix} \pi_{y_{1k}} \\ \pi_{y_{2k}} \\ \vdots \\ \pi_{y_{bk}} \end{pmatrix} \quad (1)$$

where $k = 1, 2, \dots, p$. Note that $\sum_{k=1}^p \pi_{x_{ik}} = 1$, $\sum_{k=1}^p \pi_{y_{jk}} = 1$, and $\sum_{j=1}^b \theta_{ij} = 1$.

Letting $\pi_x = (\pi_{x_{i1}}, \pi_{x_{i2}}, \dots, \pi_{x_{ip}})'$ and $\pi_y = (\pi_{y_{j1}}, \pi_{y_{j2}}, \dots, \pi_{y_{jp}})'$, Eq. (1) becomes

$$\begin{pmatrix} \pi_{x_1} \\ \pi_{x_2} \\ \vdots \\ \pi_{x_a} \end{pmatrix} = \begin{pmatrix} \theta_{11} I_p & \theta_{12} I_p & \cdots & \theta_{1b} I_p \\ \theta_{21} I_p & \theta_{22} I_p & \cdots & \theta_{2b} I_p \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{a1} I_p & \theta_{a2} I_p & \cdots & \theta_{ab} I_p \end{pmatrix} \begin{pmatrix} \pi_{y_1} \\ \pi_{y_2} \\ \vdots \\ \pi_{y_b} \end{pmatrix}$$

$$= \begin{pmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1b} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2b} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{a1} & \theta_{a2} & \cdots & \theta_{ab} \end{pmatrix} \otimes I_p \begin{pmatrix} \pi_{y_1} \\ \pi_{y_2} \\ \vdots \\ \pi_{y_b} \end{pmatrix}$$

$$= \Theta \otimes I_p \begin{pmatrix} \pi_{y_1} \\ \pi_{y_2} \\ \vdots \\ \pi_{y_b} \end{pmatrix} \quad (2)$$

or $\pi_x = (\Theta \otimes I_p) \pi_y$, where I_p denotes the identity matrix of size p . If $a = b$ and $\Theta \otimes I_p$ is a nonsingular matrix, we have

$$\begin{aligned} \pi_y &= (\Theta \otimes I_p)^{-1} \pi_x \\ &= (\Theta^{-1} \otimes I_p^{-1}) \pi_x \\ &= (\Theta^{-1} \otimes I_p) \pi_x. \end{aligned} \quad (3)$$

If $a = b = p = 2$, the estimator for Eq. (3) is the same as the estimator provided by Shih and Zhao^[3]. The unbiased estimators of π_{y11} and π_{y21} are given by

$$\begin{pmatrix} \hat{\pi}_{y11} \\ \hat{\pi}_{y21} \end{pmatrix} = \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\pi}_{x11} \\ \hat{\pi}_{x21} \end{pmatrix}.$$

Letting $\theta_{11} = \theta_{22} = \theta$, this becomes

$$\begin{pmatrix} \hat{\pi}_{y11} \\ \hat{\pi}_{y21} \end{pmatrix} = \begin{pmatrix} \theta & 1-\theta \\ 1-\theta & \theta \end{pmatrix}^{-1} \begin{pmatrix} \hat{\pi}_{x11} \\ \hat{\pi}_{x21} \end{pmatrix}$$

$$= \frac{1}{2\theta-1} \begin{pmatrix} \theta & -(1-\theta) \\ -(1-\theta) & \theta \end{pmatrix} \begin{pmatrix} \hat{\pi}_{x11} \\ \hat{\pi}_{x21} \end{pmatrix}.$$

If we have n_i observations in i th block, by the Central Limit Theorem, this implies

$$\hat{\pi}_{x_{i1}} - \pi_{x_{i1}} \xrightarrow{D} N\left(0, \pi_{x_{i1}}(1-\pi_{x_{i1}})/n_i\right).$$

for $i = 1, 2$. By the delta method, we have

$$\hat{\pi}_{y11} - \pi_{y11} \xrightarrow{D} N\left(0, V(\hat{\pi}_{y11})\right)$$

and

$$\hat{\pi}_{y21} - \pi_{y21} \xrightarrow{D} N\left(0, V(\hat{\pi}_{y21})\right)$$

where $V(\hat{\pi}_{y11}) = \frac{1}{(2\theta-1)^2} \left(\theta^2 \frac{\pi_{x11}(1-\pi_{x11})}{n_1} + (1-\theta)^2 \frac{\pi_{x21}(1-\pi_{x21})}{n_2} \right)$

and

$$V(\hat{\pi}_{y21}) = \frac{1}{(2\theta-1)^2} \left((1-\theta)^2 \frac{\pi_{x11}(1-\pi_{x11})}{n_1} + \theta^2 \frac{\pi_{x21}(1-\pi_{x21})}{n_2} \right).$$

Thus,

$$(\hat{\pi}_{y11} - \hat{\pi}_{y21}) - (\pi_{y11} - \pi_{y21}) \xrightarrow{D} N\left(0, V(\hat{\pi}_{y11}) + V(\hat{\pi}_{y21})\right).$$

To test whether there is a difference between the mean response rates of the test drug and the reference drug, the following hypotheses are usually considered:

$$H_0 : \pi_{y11} = \pi_{y21} \text{ versus } H_a : \pi_{y11} \neq \pi_{y21}.$$

We reject the null hypothesis at the α level of significance if

$$\left| \frac{\hat{\pi}_{y11} - \hat{\pi}_{y21}}{\sqrt{V(\hat{\pi}_{y11}) + V(\hat{\pi}_{y21})}} \right| > z_{\alpha/2}$$

Under the alternative hypothesis $H_a : \pi_{y11} \neq \pi_{y21}$, the power of the above test is approximately

$$\Phi \left(\frac{\pi_{y11} - \pi_{y21}}{\sqrt{V(\hat{\pi}_{y11}) + V(\hat{\pi}_{y21})}} - z_{\alpha/2} \right).$$

As a result, the sample size needed for achieving a power of $1 - \beta$ can be obtained by the following equation:

$$\frac{\pi_{y11} - \pi_{y21}}{\sqrt{V(\hat{\pi}_{y11}) + V(\hat{\pi}_{y21})}} - z_{\alpha/2} = z_{\beta}.$$

After some algebra, this leads to

$$\begin{aligned} & \frac{(2\theta - 1)^2 \left(\frac{\pi_{y11} - \pi_{y21}}{z_{\beta} + z_{\alpha/2}} \right)^2}{\theta^2 + (1 - \theta)^2} \\ &= \frac{\pi_{x11}(1 - \pi_{x11})}{n_1} + \frac{\pi_{x21}(1 - \pi_{x21})}{n_2}. \end{aligned}$$

If $n = n_1 = n_2$, then the blind sample size is given by

$$\begin{aligned} n &= \left(\pi_{x11}(1 - \pi_{x11}) + \pi_{x21}(1 - \pi_{x21}) \right) \\ &\times \frac{\theta^2 + (1 - \theta)^2}{(2\theta - 1)^2} \left(\frac{z_{\beta} + z_{\alpha/2}}{\pi_{y11} - \pi_{y21}} \right)^2. \end{aligned} \quad (4)$$

Comparing to the unblind sample size^[12]

$$m = \left(\pi_{y11}(1 - \pi_{y11}) + \pi_{y21}(1 - \pi_{y21}) \right) \left(\frac{Z_{\beta} + Z_{\alpha/2}}{\pi_{y11} - \pi_{y21}} \right)^2,$$

we need extra information in order to estimate the blind sample size. That is, the information of θ , π_{x11} , and π_{x21} . If $\theta = 1$, $\pi_{y11} = \pi_{x11}$ and $\pi_{y21} = \pi_{x21}$, then $n = m$. In fact, the blind sample size is greater than or equal to the unblind sample size in the balance design, which is shown in the following theorem.

Theorem 1 $n \geq m$.

Proof 1 Since $\pi_{x11} = \theta\pi_{y11} + (1 - \theta)\pi_{y21}$ and $\pi_{x21} = (1 - \theta)\pi_{y11} + \theta\pi_{y21}$, the blind sample of Eq. (4) can be expressed as

$$\begin{aligned} n &= \left(\pi_{x11}(1 - \pi_{x11}) + \pi_{x21}(1 - \pi_{x21}) \right) \\ &\times \frac{\theta^2 + (1 - \theta)^2}{(2\theta - 1)^2} \left(\frac{Z_{\beta} + Z_{\alpha/2}}{\pi_{y11} - \pi_{y21}} \right)^2 \\ &= \left(\pi_{y11}(1 - \pi_{y11}) + \pi_{y21}(1 - \pi_{y21}) \right) \\ &\times \frac{\theta^2 + (1 - \theta)^2}{(2\theta - 1)^2} \left(\frac{Z_{\beta} + Z_{\alpha/2}}{\pi_{y11} - \pi_{y21}} \right)^2 \\ &= \left(\pi_{y11}(1 - \pi_{y11}) + \pi_{y21}(1 - \pi_{y21}) \right) \\ &\times \frac{1}{2} \left(1 + \frac{1}{(2\theta - 1)^2} \right) \left(\frac{Z_{\beta} + Z_{\alpha/2}}{\pi_{y11} - \pi_{y21}} \right)^2 \\ &\geq \left(\pi_{y11}(1 - \pi_{y11}) + \pi_{y21}(1 - \pi_{y21}) \right) \left(\frac{Z_{\beta} + Z_{\alpha/2}}{\pi_{y11} - \pi_{y21}} \right)^2 \\ &= m \end{aligned} \quad (5)$$

due to the fact that $\frac{1}{(2\theta - 1)^2} \geq 1$.

Corollary 1 The blind sample size is getting close to the unblind sample size as θ approaches to either 0 or 1.

Proof 2 The result is concluded in Eq. (5).

Non-inferiority or Superiority Trials

Suppose an investigator plans to demonstrate the study drug is superior (or non-inferior) to the competitor. The null hypothesis states that the new treatment is not superior (or inferior) to the active control at certain level of margin, say δ . The alternative hypothesis states that the study drug is superior (non-inferior) to the competitor. It can be written in terms of mathematical formulae such as

$$H_0 : \pi_{y11} - \pi_{y21} \leq \delta \quad \text{versus} \quad H_a : \pi_{y11} - \pi_{y21} > \delta$$

where δ is the superiority or non-inferiority margin. The margin $\delta > 0$ can be used to show that the new treatment is superior by a fixed amount δ . Setting the margin $\delta < 0$, we claim that the active treatment is less effective than the control treatment by no more than δ if the null hypothesis is rejected. We reject the null hypothesis at the α level of significance if

$$\left| \frac{\hat{\pi}_{y11} - \hat{\pi}_{y21} - \delta}{\sqrt{V(\hat{\pi}_{y11}) + V(\hat{\pi}_{y21})}} \right| > Z_{\alpha}$$

Under the alternative hypothesis $H_a : \pi_{y11} - \pi_{y21} > \delta$, the power of the above test is approximately

$$\Phi \left(\frac{\pi_{y11} - \pi_{y21} - \delta}{\sqrt{V(\hat{\pi}_{y11}) + V(\hat{\pi}_{y21})}} - Z_{\alpha} \right).$$

As a result, the sample size needed for achieving a power of $1 - \beta$ can be obtained by the following equation:

$$\frac{\pi_{y11} - \pi_{y21} - \delta}{\sqrt{V(\hat{\pi}_{y11}) + V(\hat{\pi}_{y21})}} - Z_{\alpha} = Z_{\beta}.$$

After some algebra, this leads to

$$\begin{aligned} & \frac{(2\theta - 1)^2 \left(\frac{\pi_{y11} - \pi_{y21} - \delta}{Z_{\beta} + Z_{\alpha}} \right)^2}{\theta^2 + (1 - \theta)^2} \\ &= \frac{\pi_{x11}(1 - \pi_{x11})}{n_1} + \frac{\pi_{x21}(1 - \pi_{x21})}{n_2} \end{aligned}$$

If $n = n_1 = n_2$, then the blind sample size is given by

$$\begin{aligned} n &= \left(\pi_{x11}(1 - \pi_{x11}) + \pi_{x21}(1 - \pi_{x21}) \right) \\ &\times \frac{\theta^2 + (1 - \theta)^2}{(2\theta - 1)^2} \left(\frac{Z_{\beta} + Z_{\alpha}}{\pi_{y11} - \pi_{y21} - \delta} \right)^2. \end{aligned} \quad (6)$$

Equivalence Trials

The objective of this section is to demonstrate that two treatments have no clinically meaningful difference. The

null hypothesis and alternative hypothesis can be written in the following form

$$H_0: |\pi_{y11} - \pi_{y21}| \geq \delta \quad \text{versus} \quad H_a: |\pi_{y11} - \pi_{y21}| < \delta$$

where δ is the equivalent margin. The equivalence test are performed using two one-sided tests. We reject the null hypothesis at the α level of significance if

$$\frac{\hat{\pi}_{y11} - \hat{\pi}_{y21} - \delta}{\sqrt{V(\hat{\pi}_{y11}) + V(\hat{\pi}_{y21})}} < -Z_\alpha$$

and

$$\frac{\hat{\pi}_{y11} - \hat{\pi}_{y21} - \delta}{\sqrt{V(\hat{\pi}_{y11}) + V(\hat{\pi}_{y21})}} < -Z_\alpha$$

Under the alternative hypothesis $H_a: |\pi_{y11} - \pi_{y21}| < \delta$, the power of the above test is approximately

$$2\Phi\left(\frac{\delta - |\pi_{y11} - \pi_{y21}|}{\sqrt{V(\hat{\pi}_{y11}) + V(\hat{\pi}_{y21})}} - Z_\alpha\right) - 1.$$

As a result, the sample size needed for achieving a power of $1 - \beta$ can be obtained by the following equation:

$$\frac{\delta - |\pi_{y11} - \pi_{y21}|}{\sqrt{V(\hat{\pi}_{y11}) + V(\hat{\pi}_{y21})}} - Z_\alpha = Z_{\beta/2}.$$

After some algebra, this leads to

$$\begin{aligned} & \frac{(2\theta - 1)^2}{\theta^2 + (1 - \theta)^2} \left(\frac{\delta - |\pi_{y11} - \pi_{y21}|}{Z_{\beta/2} + Z_\alpha} \right)^2 \\ &= \frac{\pi_{x11}(1 - \pi_{x11})}{n_1} + \frac{\pi_{x21}(1 - \pi_{x21})}{n_2} \end{aligned}$$

If $n = n_1 = n_2$, then the blind sample size is given by

$$\begin{aligned} n &= \left(\pi_{x11}(1 - \pi_{x11}) + \pi_{x21}(1 - \pi_{x21}) \right) \\ &\times \frac{\theta^2 + (1 - \theta)^2}{(2\theta - 1)^2} \left(\frac{Z_{\beta/2} + Z_\alpha}{\delta - |\pi_{y11} - \pi_{y21}|} \right)^2. \end{aligned}$$

Example

McIntyre et al. (2005)^[16] conducted a multicenter, randomized controlled trial to compare buccal midazolam with rectal diazepam for emergency-room treatment of children aged 6 months and older presenting to hospital with active seizures and without intravenous access. The results have shown that therapeutic success was 56% for buccal midazolam and 27% for rectal diazepam. Suppose now we wish to design a new study to

investigate a new treatment with buccal midazolam as the control. We expect this new treatment to increase the chances of success to 66%. The sponsor has confident about the incidence rate of 56% in the control group, but they are not sure about the 66% incidence rate in the test group. Thus, an interim analysis masking the randomization codes is planned to allow for sample size re-estimation based on observed block difference. Here, the sample proportions for each randomization block are $\hat{\pi}_{x11} = 0.59$ and $\hat{\pi}_{x11} = 0.64$. We wish to have 80% power and a two-sided significance level of 5%. Assuming the superiority margin is 0.05. Considering the transition probability $\theta = 0.8$, the estimated blind sample size of Eq. (6) is evaluated by

$$\begin{aligned} \hat{n} &= \left(\pi_{x11}(1 - \pi_{x11}) + \pi_{x21}(1 - \pi_{x21}) \right) \\ &\frac{\theta^2 + (1 - \theta)^2}{(2\theta - 1)^2} \left(\frac{Z_\beta + Z_\alpha}{\pi_{y11} - \pi_{y21} - \delta} \right)^2 \\ &= (0.59(1 - 0.59) + 0.64(1 - 0.64)) \\ &\frac{0.8 + (1 - 0.8)^2}{(2 \times 0.8 - 1)^2} \left(\frac{0.84 + 1.64}{0.66 - 0.56 - 0.05} \right)^2 \\ &= 2207 \text{ subjects for each block.} \end{aligned}$$

CONTINUOUS RESPONSE

Suppose we have a randomization blocks and b treatment groups. Let y_j denote the original response in the j th group for $j = 1, 2, \dots, b$. Let $x_i = \sum_{j=1}^b I_{ij} y_j$ be the relocated response from one of the treatment groups to the i th block, where

$$I_{ij} = \begin{cases} 1, & \text{if the subject relocates} \\ & \text{the } j\text{th group to the } i\text{th blocks,} \\ 0, & \text{otherwise.} \end{cases}$$

Note that $\sum_{j=1}^b I_{ij} = 1$ for $i = 1, 2, \dots, a$. Collecting all of x 's, y 's and I 's, this can expressed in a linear model. That is,

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_a \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} & \cdots & I_{1b} \\ I_{21} & I_{22} & \cdots & I_{2b} \\ \vdots & \vdots & \vdots & \vdots \\ I_{a1} & I_{a2} & \cdots & I_{ab} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_b \end{pmatrix} \quad (7)$$

or $\mathbf{x} = \mathbf{I}\mathbf{y}$. Taking the expectation of both sides of Eq. (7), we have

$$\boldsymbol{\mu}_x = E(\mathbf{x}) = E(\mathbf{I}\mathbf{y}) = E(\mathbf{I})E(\mathbf{y}) = \Theta\boldsymbol{\mu}_y \quad (8)$$

if $\mathbf{I}$ and $\mathbf{y}$ are independent. The transition probability matrix is given by

$$\Theta = \begin{pmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1b} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2b} \\ \vdots & \vdots & \vdots & \vdots \\ \theta_{a1} & \theta_{a2} & \cdots & \theta_{ab} \end{pmatrix} \quad (9)$$

where the random vector $\mathbf{I}_i = (I_{i1}, I_{i12}, \dots, I_{ib})'$ has the probability vector $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i12}, \dots, \theta_{ib})'$ and $\sum_{j=1}^b \theta_{ij} = 1$ for all i . If $a = b$ and the transition probability matrix Θ is nonsingular, the mean vector of treatment responses in Eq. (8) becomes

$$\boldsymbol{\mu}_y = \Theta^{-1} \boldsymbol{\mu}_x. \quad (10)$$

The vector estimator $\hat{\boldsymbol{\mu}}_y$ of (10) is unbiased if we have unbiased estimator $\hat{\boldsymbol{\mu}}_x$. Considering $a = b = 2$ and setting $\theta_{11} = \theta_{22} = \theta$, Eq. (10) becomes

$$\begin{pmatrix} \mu_{y1} \\ \mu_{y2} \end{pmatrix} = \begin{pmatrix} \theta & 1-\theta \\ 1-\theta & \theta \end{pmatrix}^{-1} \begin{pmatrix} \mu_{x1} \\ \mu_{x2} \end{pmatrix} \\ = \frac{1}{2\theta-1} \begin{pmatrix} \theta & -(1-\theta) \\ -(1-\theta) & \theta \end{pmatrix} \begin{pmatrix} \mu_{x1} \\ \mu_{x2} \end{pmatrix}.$$

Let x_{ik} be the response observed from the k th subject in the i th block, where $k = 1, \dots, n$ and $i = 1, 2$. The unbiased estimators of μ_{x1} and μ_{x2} are $\hat{\mu}_{x1} = \bar{x}_1$ and $\hat{\mu}_{x2} = \bar{x}_2$, respectively. If x_{ik} 's are normally distributed and independent with mean μ_{xi} and common variance σ_x^2 , then the estimators

$$\hat{\mu}_{y1} = \frac{1}{2\theta-1} (\theta \hat{\mu}_{x1} + (1-\theta) \hat{\mu}_{x2})$$

and

$$\hat{\mu}_{y2} = \frac{1}{2\theta-1} ((1-\theta) \hat{\mu}_{x1} + \theta \hat{\mu}_{x2})$$

are also normally distributed with mean μ_{y1} and μ_{y2} , respectively, and common variance $\frac{1}{(2\theta-1)^2} (\theta^2 + (1-\theta)^2) \sigma_x^2 / n$. This follows that

$$\hat{\mu}_{y1} - \hat{\mu}_{y2} \sim N \left(\mu_{y1} - \mu_{y2}, \frac{2}{(2\theta-1)^2} (\theta^2 + (1-\theta)^2) \sigma_x^2 / n \right).$$

Note that in the balanced design the population treatment variance

$$\sigma_y^2 = \frac{1}{(2\theta-1)^2} (\theta^2 + (1-\theta)^2) \sigma_x^2 \quad (11)$$

if $\sigma_{y1} = \sigma_{y2} = \sigma_y$. The objective is to test whether there is a difference between the mean responses of the test drug and the placebo control or active control. Hence, the following hypotheses are considered:

$$H_0 : \mu_{y1} = \mu_{y2} \quad \text{versus} \quad H_a : \mu_{y1} \neq \mu_{y2}$$

When σ_x^2 is known, the null hypothesis H_0 is rejected at the significance level α if

$$\left| \frac{\hat{\mu}_{y1} - \hat{\mu}_{y2}}{\sqrt{\frac{2}{(2\theta-1)^2} (\theta^2 + (1-\theta)^2) \sigma_x^2 / n}} \right| > Z_{\alpha/2}$$

Under the alternative hypothesis that $H_a : \mu_{y1} \neq \mu_{y2}$, the power of the above test is

$$\begin{aligned} 1 - \beta &= P \left(\left| \frac{\hat{\mu}_{y1} - \hat{\mu}_{y2}}{\sqrt{\frac{2}{(2\theta-1)^2} (\theta^2 + (1-\theta)^2) \sigma_x^2 / n}} \right| > Z_{\alpha/2} \right) \\ &= P \left(Z < -Z_{\alpha/2} - \frac{\mu_{y1} - \mu_{y2}}{\sqrt{\frac{2}{(2\theta-1)^2} (\theta^2 + (1-\theta)^2) \sigma_x^2 / n}} \right) \\ &\quad + P \left(Z < -Z_{\alpha/2} + \frac{\mu_{y1} - \mu_{y2}}{\sqrt{\frac{2}{(2\theta-1)^2} (\theta^2 + (1-\theta)^2) \sigma_x^2 / n}} \right) \\ &= \Phi \left(-Z_{\alpha/2} - \frac{\mu_{y1} - \mu_{y2}}{\sqrt{\frac{2}{(2\theta-1)^2} (\theta^2 + (1-\theta)^2) \sigma_x^2 / n}} \right) \\ &\quad + \Phi \left(-Z_{\alpha/2} + \frac{\mu_{y1} - \mu_{y2}}{\sqrt{\frac{2}{(2\theta-1)^2} (\theta^2 + (1-\theta)^2) \sigma_x^2 / n}} \right) \end{aligned}$$

if $\mu_{y1} > \mu_{y2}$. Ignoring the term of the right hand side of the equation, i.e.

$$\Phi \left(-Z_{\alpha/2} - \frac{(\mu_{y1} - \mu_{y2})}{\sqrt{\frac{2}{(2\theta-1)^2} (\theta^2 + (1-\theta)^2) \sigma_x^2 / n}} \right) \leq \alpha/2, \text{ the power}$$

can be expressed as

$$1 - \beta \approx \Phi \left(-Z_{\alpha/2} + \frac{\mu_{y1} - \mu_{y2}}{\sqrt{\frac{2}{(2\theta-1)^2} (\theta^2 + (1-\theta)^2) \sigma_x^2 / n}} \right).$$

Similarly, the power can be approximated by

$$1 - \beta \approx \Phi \left(-Z_{\alpha/2} - \frac{\mu_{y1} - \mu_{y2}}{\sqrt{\frac{2}{(2\theta-1)^2} (\theta^2 + (1-\theta)^2) \sigma_x^2 / n}} \right).$$

if $\mu_{y1} < \mu_{y2}$. In summary, we have the following result

$$1 - \beta \approx \Phi \left(-Z_{\alpha/2} + \frac{|\mu_{y1} - \mu_{y2}|}{\sqrt{\frac{2}{(2\theta-1)^2} (\theta^2 + (1-\theta)^2) \sigma_x^2 / n}} \right).$$

The sample size needed to achieve power $1 - \beta$ can be obtained by solving the following equation

$$\frac{|\mu_{y1} - \mu_{y2}|}{\sqrt{\frac{2}{(2\theta - 1)^2}(\theta^2 + (1 - \theta)^2)\sigma_x^2/n}} - Z_{\alpha/2} = Z_{\beta}.$$

The blind sample size can be expressed as

$$n = \frac{(Z_{\alpha/2} + Z_{\beta})^2}{(\mu_{y1} - \mu_{y2})^2} \frac{2}{(2\theta - 1)^2} (\theta^2 + (1 - \theta)^2) \sigma_x^2. \quad (12)$$

The population variance of the randomization blocks can be estimated by the pool sample variance in the middle of clinical trial, unaware of which group subjects belong to. That is,

$$\hat{\sigma}_x^2 = \frac{\sum_{i=1}^2 \sum_{k=1}^n (x_{ik} - \hat{x}_i)^2}{2n - 2}.$$

The blind sample size of Eq. (12) is then estimated by

$$\hat{n} = \frac{(z_{\alpha/2} + z_{\beta})^2}{(\mu_{y1} - \mu_{y2})^2} \frac{2}{(2\theta - 1)^2} (\theta^2 + (1 - \theta)^2) \hat{\sigma}_x^2. \quad (13)$$

The probability θ provided in the protocol is the proportion of the first treatment assigned in the first block. Taking the expectations on both side, we have

$$\begin{aligned} E(\hat{n}) &= \frac{(z_{\alpha/2} + z_{\beta})^2}{(\mu_{y1} - \mu_{y2})^2} \frac{2}{(2\theta - 1)^2} (\theta^2 + (1 - \theta)^2) E(\hat{\sigma}_x^2) \\ &= \frac{(z_{\alpha/2} + z_{\beta})^2}{(\mu_{y1} - \mu_{y2})^2} \frac{2}{(2\theta - 1)^2} (\theta^2 + (1 - \theta)^2) \sigma_x^2 \\ &= \frac{2(z_{\alpha/2} + z_{\beta})^2 \sigma_y^2}{(\mu_{y1} - \mu_{y2})^2} \\ &= E(\tilde{n}) \end{aligned}$$

by Eq. (11). Here, the the unblind sample size^[12] is

$$\hat{n} = \frac{2(Z_{\alpha/2} + Z_{\beta})^2 \hat{\sigma}_y^2}{(\mu_{y1} - \mu_{y2})^2}$$

where $\hat{\sigma}_y^2 = \sum_{i=1}^2 \sum_{k'=1}^2 (y_{i'k'} - \bar{y}_{i'})^2 / (2n - 2)$ is an unbiased estimator of σ_y^2 . Its variance is given by

$$V(\tilde{n}) = \left[\frac{2(Z_{\alpha/2} + Z_{\beta})^2}{(\mu_{y1} - \mu_{y2})^2} \right]^2 \frac{2\sigma_y^4}{2n - 2}$$

due to the fact that $V\left(\frac{(2n - 2)\hat{\sigma}_y^2}{\sigma_y^2}\right) = V(\chi^2(2n - 2))$ and $V(\chi^2(2n - 2)) = 2(2n - 2)$.

The variance of $\hat{n}$ Eq. (13) is evaluated by

$$\begin{aligned} V(\tilde{n}) &= \left[\frac{(z_{\alpha/2} + z_{\beta})^2}{(\mu_{y1} - \mu_{y2})^2} \frac{2}{(2\theta - 1)^2} (\theta^2 + (1 - \theta)^2) \right]^2 \\ &\quad \times \frac{2\sigma_x^4}{2n - 2} \\ &= \left[\frac{2(z_{\alpha/2} + z_{\beta})^2}{(\mu_{y1} - \mu_{y2})^2} \right]^2 \frac{1}{(2\theta - 1)^4} (\theta^2 + (1 - \theta)^2)^2 \\ &\quad \times \frac{2\sigma_x^4}{2n - 2} \\ &= \left[\frac{2(z_{\alpha/2} + z_{\beta})^2}{(\mu_{y1} - \mu_{y2})^2} \right]^2 \left(\frac{\theta^2 + (1 - \theta)^2}{(2\theta - 1)^2} \sigma_x^2 \right)^2 \\ &\quad \times \frac{2}{2n - 2} \\ &= \left[\frac{2(z_{\alpha/2} + z_{\beta})^2}{(\mu_{y1} - \mu_{y2})^2} \right]^2 \frac{2\sigma_y^4}{2n - 2} \\ &= V(\hat{n}). \end{aligned}$$

In conclusion, we have an unbiased estimator of blind sample size and its variance is the same as the variance of the unblind sample size.

ANOVA Model

In clinical trial, the number of the randomization blocks is most likely more than the number of treatments (i.e. $a > b$). Thus, the inverse of the transition probability matrix Θ in Eq. (10) is not held. The sample size can be re-evaluated alternatively by using the analysis of variance model if Θ is not a square matrix. Suppose the response x_{ij} for j th subject in the i th block follows a unknown distribution with mean μ_{xi} and variance η^2 for $j = 1, 2, \dots, n_i$. This can be written in a linear form as

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_a \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n1} & \mathbf{0}_{n1} & \cdots & \mathbf{0}_{n1} \\ \mathbf{0}_{n2} & \mathbf{1}_{n2} & \cdots & \mathbf{0}_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{na} & \mathbf{0}_{na} & \cdots & \mathbf{1}_{na} \end{pmatrix} \begin{pmatrix} \mu_{x1} \\ \mu_{x2} \\ \vdots \\ \mu_{xa} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_a \end{pmatrix} \quad (14)$$

or simplified as

$$\mathbf{x} = \mathbf{A}\boldsymbol{\mu}_x + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, \dots, \epsilon_{ini})'$ for $i = 1, 2, \dots, a$. Here, the error term $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \dots, \boldsymbol{\epsilon}_a)'$ is distributed with mean $\mathbf{0}$

and covariance matrix $\eta^2 I$. A vector of 1's with size n_i is denoted by $\mathbf{1}_{ni}$ and a vector of 0's with size n_i by $\mathbf{0}_{ni}$. Substituting Eq. (8) into Eq. (14), this model becomes

$$\mathbf{x} = A\Theta\boldsymbol{\mu}_y + \epsilon.$$

If the transition probability matrix Θ_{ab} is a known full rank matrix and $a \geq b$, we have a full-rank model for which the means of true responses $\boldsymbol{\mu}_y$ can be estimated by the method of least squares. That is, the least square estimator (LSE) is given by

$$\tilde{\boldsymbol{\mu}}_y = \left((A\Theta)' A\Theta \right)^{-1} (A\Theta)' \mathbf{x}. \quad (15)$$

It is interesting to note that LSE is the same as the estimator of Eq. (10). That is,

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_y &= \left((A\Theta)' A\Theta \right)^{-1} (A\Theta)' \mathbf{x} \\ &= \left(\Theta' A' A\Theta \right)^{-1} \Theta' A' \mathbf{x} \\ &= \left(\Theta' \text{diag}(n_1, n_2, \dots, n_a) \Theta \right)^{-1} \Theta' A' \mathbf{x} \\ &= \Theta^{-1} \text{diag}(1/n_1, 1/n_2, \dots, 1/n_a) \Theta^{-1} \Theta' A' \mathbf{x} \\ &= \Theta^{-1} \text{diag}(1/n_1, 1/n_2, \dots, 1/n_a) A' \mathbf{x} \\ &= \Theta^{-1} (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_a)' \\ &= \hat{\boldsymbol{\mu}}_y \end{aligned}$$

if the square matrix Θ is a nonsingular matrix. Suppose $\mathbf{x}$ is distributed $N(A\Theta\boldsymbol{\mu}_y, \eta^2 I)$ and C is $q \times b$ of rank $q \leq b \leq a$. Rencher and Schaafje^[13] shows that the statistic

$$F = \frac{SSH/q}{SSE / \left(\sum_{i=1}^a n_i - b \right)} \quad (16)$$

is distributed as $F\left(q, \sum_{i=1}^a n_i - b\right)$ under $H_0: C\boldsymbol{\mu}_y = 0$, where

$$SSH = (C\tilde{\boldsymbol{\mu}}_y)' \left(C \left((A\Theta)' A\Theta \right)^{-1} C' \right)^{-1} (C\tilde{\boldsymbol{\mu}}_y)$$

and

$$SSE = (\mathbf{x} - A\Theta\tilde{\boldsymbol{\mu}}_y)' (\mathbf{x} - A\Theta\tilde{\boldsymbol{\mu}}_y).$$

The null hypothesis H_0 is rejected at the α level of significance if

$$F = \frac{SSH/q}{SSE / \left(\sum_{i=1}^a n_i - b \right)} > F_{\alpha, q, \sum_{i=1}^a n_i - b},$$

where $F_{\alpha, q, \sum_{i=1}^a n_i - b}$ is the α upper quantile of the F-distribution with q and $\sum_{i=1}^a n_i - b$ degrees of freedom. Under the alternative hypothesis, the test statistic F is distributed as $F_{\alpha, q, \sum_{i=1}^a n_i - b, \lambda}$, where

$$\lambda = \frac{1}{2\eta^2} \boldsymbol{\mu}_y' C' \left(C \left((A\Theta)' A\Theta \right)^{-1} C' \right)^{-1} C \boldsymbol{\mu}_y.$$

If $n_1 = n_2 = \dots = n_a = n$, then $A'A = In$. The non-central parameter becomes

$$\lambda = \frac{1}{2\eta^2} \boldsymbol{\mu}_y' C' \left(C(\Theta'\Theta)^{-1} C' \right)^{-1} C \boldsymbol{\mu}_y. \quad (17)$$

If we have $C, \boldsymbol{\mu}_y, \Theta$ and η^2 , the sample size n can be obtained by solving Eq. (17). Given a, b, α, q and β , the parameter λ can be obtained by solving the following power equation

$$P(F > F_{\alpha, q, na-b} | H_a, \lambda) = 1 - \beta \quad (18)$$

where the random variable F follows the non-central F distribution with degrees of freedom q and $na - b$, and non-centrality parameter λ . However, this power equation cannot be solved for λ because of unknown sample size n . Two unknown parameters n and λ can be obtained by solving two nonlinear Eqs. (17) and (18). Sample size algorithms^[14] is employed to obtain the blind sample size. Given $a, b, \alpha, \beta, C, \boldsymbol{\mu}_y, \Theta$ and η^2 , the following actions should occur to obtain the sample size.

Step 1: Provide an initial sample size.

Step 2: Obtain the critical value $F_{\alpha, q, na-b}$ such that $P(F > F_{\alpha, q, na-b} | H_0) = \alpha$, where the test statistic F follows a central F distribution with degrees of freedom q and $na - b$.

Step 3: Calculate the power $P(F > F_{\alpha, q, na-b} | H_a, \lambda)$, where the random variable F follows the non-central F distribution with degrees of freedom q and $na - b$, and non-centrality parameter $\lambda = \frac{n}{2\eta^2} \boldsymbol{\mu}_y' C' \left(C(\Theta'\Theta)^{-1} C' \right)^{-1} C \boldsymbol{\mu}_y$.

Step 4: If the power in Step 3 is smaller than the desired power $1 - \beta$, increase the sample size by one unit. Otherwise, reduce the sample size by one unit.

Step 5: Do the iterations from Step 2 to Step 4 until the sample size converges.

Non-inferiority or Superiority Trials

The objective of this section is to show that the new treatment is statistically not inferior (or superior) to the standard approved product. The hypothesis of interesting is

$$H_0: \mu_{y1} - \mu_{y2} \leq \delta \quad \text{versus} \quad H_a: \mu_{y1} - \mu_{y2} > \delta$$

where δ is the non-inferiority/superiority margin. When σ_x^2 is known, the null hypothesis H_0 is rejected at the significance level α if

$$\frac{\hat{\mu}_{y1} - \hat{\mu}_{y2} - \delta}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}} > Z_\alpha.$$

Rejecting the null hypothesis implies that the treatment mean is greater than the reference mean by at least the margin δ . Under the alternative hypothesis that $H_a: \mu_{y1} - \mu_{y2} > \delta$, the power of the test can be expressed as

$$\begin{aligned} 1 - \beta &= P\left(\frac{\hat{\mu}_{y1} - \hat{\mu}_{y2} - \delta}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}} > Z_\alpha\right) \\ &= P\left(Z > Z_\alpha - \frac{\mu_{y1} - \mu_{y2} - \delta}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}}\right) \\ &= P\left(Z < -Z_\alpha + \frac{\mu_{y1} - \mu_{y2} - \delta}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}}\right) \\ &= \Phi\left(-Z_\alpha + \frac{\mu_{y1} - \mu_{y2} - \delta}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}}\right). \end{aligned}$$

The sample size needed to achieve power $1 - \beta$ can be obtained by solving the following equation

$$-Z_\alpha + \frac{\mu_{y1} - \mu_{y2} - \delta}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}} = Z_\beta.$$

After some algebra, the blind sample size is given by

$$n = \frac{(z_\alpha + z_\beta)^2}{(\mu_{y1} - \mu_{y2} - \delta)^2} \frac{2}{(2\theta-1)^2} (\theta^2 + (1-\theta)^2) \sigma_x^2. \quad (19)$$

Equivalence Trials

The objective of the equivalence study is to test the following hypotheses

$$H_0: |\mu_{y1} - \mu_{y2}| \geq \delta \quad \text{versus} \quad H_a: |\mu_{y1} - \mu_{y2}| < \delta.$$

where δ is the equivalence margin. This is the same as to conduct two one-sided tests simultaneously. That is

$$H_{01}: \mu_{y1} - \mu_{y2} \geq \delta \quad \text{versus} \quad H_{a1}: \mu_{y1} - \mu_{y2} < \delta$$

and

$$H_{02}: \mu_{y1} - \mu_{y2} \leq -\delta \quad \text{versus} \quad H_{a2}: \mu_{y1} - \mu_{y2} > \delta.$$

Suppose σ_x^2 is known. The null hypotheses H_{01} and H_{02} are rejected at the significance level α if

$$\frac{\hat{\mu}_{y1} - \hat{\mu}_{y2} - \delta}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}} = -Z_\alpha$$

and

$$\frac{\hat{\mu}_{y1} - \hat{\mu}_{y2} + \delta}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}} > Z_\alpha,$$

respectively. Note that there is no need to correct the overall significant level α due to the fact that both null hypotheses are rejected for the equivalent study. Under $H_{a1}: \mu_{y1} - \mu_{y2} < \delta$, the power can be written by

$$\begin{aligned} 1 - \beta &= P\left(\frac{\hat{\mu}_{y1} - \hat{\mu}_{y2} - \delta}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}} < -Z_\alpha\right) \\ &= P\left(Z < -Z_\alpha + \frac{\delta - (\mu_{y1} - \mu_{y2})}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}}\right) \\ &= \Phi\left(-Z_\alpha + \frac{\delta - (\mu_{y1} - \mu_{y2})}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}}\right) \end{aligned}$$

Under $H_{a2}: \mu_{y1} - \mu_{y2} > \delta$, the power for the second hypothesis is given by

$$\begin{aligned} \text{power} &= P\left(\frac{\hat{\mu}_{y1} - \hat{\mu}_{y2} + \delta}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}} > -Z_\alpha\right) \\ &= P\left(Z > Z_\alpha + \frac{-\delta - (\mu_{y1} - \mu_{y2})}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}}\right) \\ &= P\left(Z < -Z_\alpha + \frac{\delta + \mu_{y1} - \mu_{y2}}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}}\right) \end{aligned}$$

$$= \Phi \left(-Z_{\alpha} + \frac{\delta + \mu_{y1} - \mu_{y2}}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}} \right).$$

Under the alternative hypothesis $H_a: |\mu_{y1} - \mu_{y2}| < \delta$, the overall power is given by

$$\begin{aligned} 1 - \beta &= \Phi \left(-z_{\alpha} + \frac{\delta - (\mu_{y1} - \mu_{y2})}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}} \right) \\ &\quad + \Phi \left(-z_{\alpha} + \frac{\delta + \mu_{y1} - \mu_{y2}}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}} \right) - 1 \\ &\geq 2\Phi \left(-Z_{\alpha} + \frac{\delta - |\mu_{y1} - \mu_{y2}|}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}} \right) - 1 \\ &= 1 - \beta \end{aligned}$$

To achieve the desired power $1 - \beta$, the blind sample size can be evaluated by solving the following equation

$$-Z_{\alpha} + \frac{\delta - |\mu_{y1} - \mu_{y2}|}{\sqrt{\frac{2}{(2\theta-1)^2}(\theta^2 + (1-\theta)^2)\sigma_x^2/n}} = Z_{\beta/2}.$$

This leads to

$$n = \frac{(Z_{\alpha} + Z_{\beta/2})^2}{(\delta - |\mu_{y1} - \mu_{y2}|)^2} \frac{2}{(2\theta-1)^2} (\theta^2 + (1-\theta)^2) \sigma_x^2.$$

Example

Suppose we plan to duplicate the clinical trial of chronic dizziness in Taiwan. This study is originally conducted by Yardley et al. (2004)^[17]. It is a single blind randomized, controlled trial with 20 study center in southern England. The objective of this study is to evaluate the effectiveness of nurse-delivered vestibular rehabilitation in primary care for patients with chronic dizziness. One hundred seventy patients with chronic dizziness who were randomly assigned to vestibular rehabilitation or usual medical care. Each patient received 30- to 40-minute appointment with a primary care nurse. The nurse taught the patient exercises to be carried out daily at home, with the support of a treatment booklet. The primary

endpoint is the self-reported dizziness handicap inventory at 6 month. Assuming the treatment difference between the vestibular rehabilitation group and the usual medical care groups is about 5 points difference. There may be considerable the difference between ethnicity which may influence the variability in the forthcoming trial. If the variance is unknown prior to the start the trial, the sample size should be calculated at some point during the study. We now conduct a trial with a sample size re-estimation based on the blinded variance estimation at middle of the study. The blind sample variance based upon 100 evaluable patients is $\hat{\sigma}_x^2 = 200$. Considering the transition probability $\theta = 0.8$, the estimated blind sample size of Eq. (13) with 5% type I error and 80% power is determined by

$$\begin{aligned} \hat{n} &= \frac{(z_{\alpha/2} + z_{\beta})^2}{(\mu_{y1} - \mu_{y2})^2} \frac{2}{(2\theta-1)^2} (\theta^2 + (1-\theta)^2) \hat{\sigma}_x^2 \\ &= \frac{(1.96 + 0.84)^2}{5^2} \frac{2}{(1.6-1)^2} (0.8^2 + (1-0.8)^2) 200 \\ &= 238 \text{ subjects per block.} \end{aligned}$$

SURVIVAL ANALYSIS FOR EXPONENTIAL DISTRIBUTION

Considering the lifetime T and a fixed censoring time $L - a$, where a stands for the entry time. The exact lifetime T will be observed only if $T \leq L - a$. The random variable δ indicates whether the lifetime T is observed or not. That is,

$$\delta = \begin{cases} 1, & \text{if } T \leq L - a \\ 0, & \text{if } T > L - a. \end{cases}$$

The probability density function is $f_T(t) = \lambda e^{-t\lambda}$ if the lifetime T is assumed to be exponentially distributed with the parameter λ . Let $X = \min(T; L - a)$. If $\delta = 1$, then $T \leq L - a$. This implies that $X = \min(T, L - a) = T$. Thus, $f_X(x) = f_T(x) = \lambda e^{-x\lambda}$. If $\delta = 0$, then $T > L - a$. and $X = \min(T, L - a) = L - a$. The survival function of random variable X is given by

$$\begin{aligned} S_X(x) &= P(X > x) \\ &= P(\min(T, L - a) > x) \\ &= P(T > L - a > x) \\ &= \int_x^{\infty} \lambda e^{-t\lambda} dt \\ &= e^{-x\lambda}. \end{aligned}$$

The joint probability density function of the random variables x and δ can be written by

$$g(x, \delta) = \begin{cases} \lambda e^{-x\lambda}, & \text{if } \delta = 1 \\ e^{-x\lambda}, & \text{if } \delta = 0. \end{cases} \quad (20)$$

This is equivalent to state that $g(t, \delta = 1) = \lambda e^{-t\lambda}$ if $0 < t < L$, and $g(t, \delta = 0) = e^{-t\lambda}$ if $0 < L < t$. The function of (20) can be simplified as

$$g(x, \delta) = (\lambda e^{-x\lambda})^\delta (e^{-x\lambda})^{1-\delta} = \lambda^\delta e^{-x\lambda}.$$

Assuming that the entry time a follows the probability density function

$$h(a) = \frac{\gamma e^{-\gamma a}}{1 - e^{-\gamma L_0}}, \quad 0 \leq a \leq L_0.$$

The likelihood function is given

$$L(\lambda) = \frac{\gamma e^{-\gamma a}}{1 - e^{-\gamma L_0}} \lambda^\delta e^{-x\lambda}. \quad (21)$$

Suppose there is n_i subjects of i th block with lifetimes $T_{i1}, T_{i2}, \dots, T_{ini}$ and entry times $a_{i1}, a_{i2}, \dots, a_{ini}$. Each subject has a fixed censoring time $L_{ik} = L - a_{ik} > 0$, where $k = 1, 2, \dots, n_i$ and $i = 1, 2, \dots, g$. If T_{ik} 's are assumed to be independently and identically distributed exponentially with the same parameter λ_{xi} , the likelihood function of λ_{xi} is given by

$$\begin{aligned} L(\lambda_{xi}) &= \prod_{k=1}^{n_i} \frac{\gamma e^{-\gamma a_{ik}}}{1 - e^{-\gamma L_0}} (\lambda_{xi})^{\delta_{ik}} e^{-\lambda_{xi} L_{ik}} \\ &= \frac{\gamma^n e^{-\gamma \sum_{i=1}^g \sum_{k=1}^{n_i} a_{ik}}}{(1 - e^{-\gamma L_0})^n} (\lambda_{xi})^{\sum_{k=1}^{n_i} \delta_{ik}} e^{-\sum_{k=1}^{n_i} \lambda_{xi} L_{ik}}. \end{aligned}$$

Taking the logarithm, we have

$$\begin{aligned} \log L(\lambda_{xi}) &= \log \frac{\gamma^n e^{-\gamma \sum_{i=1}^g \sum_{k=1}^{n_i} a_{ik}}}{(1 - e^{-\gamma L_0})^n} \\ &\quad + \sum_{k=1}^{n_i} \delta_{ik} \log \lambda_{xi} - \sum_{k=1}^{n_i} \lambda_{xi} L_{ik}. \end{aligned} \quad (22)$$

Taking the derivative of Eq. (22) with respect to λ_{xi} and then setting to zero, we have

$$\frac{\partial \log L}{\partial \lambda_{xi}} = \sum_{k=1}^{n_i} \delta_{ik} \frac{1}{\lambda_{xi}} - \sum_{k=1}^{n_i} x_{ik} = 0.$$

Solving for λ_{xi} , the maximum likelihood estimator (MLE) is given by

$$\hat{\lambda}_{xi} = \frac{\sum_{k=1}^{n_i} \delta_{ik}}{\sum_{k=1}^{n_i} x_{ik}}.$$

The Fisher's information can be obtained by taking the second derivative of log function of Eq. (21) firstly

$$\frac{\partial^2 \log L}{\partial \lambda_{xi}^2} = -\frac{\delta_{ik}}{\lambda_{xi}^2}$$

and then taking the expectation. That is,

$$\begin{aligned} I(\lambda_{xi}) &= -E \left(\frac{\partial^2 \log L}{\partial \lambda_{xi}^2} \right) \\ &= E \left(\frac{\delta_{ik}}{\lambda_{xi}^2} \right). \end{aligned}$$

Here

$$\begin{aligned} E(\delta_{ik}) &= \sum_{\delta_{ik}=0}^1 \delta_{ik} P(\delta_{ik}) \\ &= P(\delta_{ik} = 1) \\ &= 1 - P(\delta_{ik} = 0) \\ &= 1 - P(0 < a_i < L_0, T_i > L - a_i) \\ &= 1 - \int_0^{L_0} \int_{L-a_i}^{\infty} \frac{\gamma^n e^{-\gamma a_i}}{1 - e^{-\gamma L_0}} \lambda_{xi} e^{-\lambda_{xi} t_i} dt_i da_i \\ &= 1 - \int_0^{L_0} \frac{\gamma e^{-\gamma a_i}}{1 - e^{-\gamma L_0}} e^{-\lambda_{xi} (L-a_i)} da_i \\ &= 1 + \frac{\gamma e^{-\lambda_{xi} L} \left(1 - e^{(\lambda_{xi} - \gamma) L_0} \right)}{(\lambda_{xi} - \gamma)(1 - e^{-\gamma L_0})}. \end{aligned}$$

The variance for $\hat{\lambda}_{xi}$ is

$$V(\hat{\lambda}_{xi}) = \frac{1}{I(\lambda_{xi})} = \frac{\hat{\lambda}_{xi}^2}{\hat{E}(\delta_{ik})} = \frac{\lambda_{xi}^2}{\gamma e^{-\lambda_{xi} L} \left(1 - e^{(\lambda_{xi} - \gamma) L_0} \right) + (\lambda_{xi} - \gamma)(1 - e^{-\gamma L_0})}.$$

When $\gamma = 0$, we define $h(a) = 1/L_0$. Then

$$E(\delta_{ik}) = \frac{e^{-\lambda_{xi} L} - e^{\lambda_{xi} (L-L_0)}}{\lambda_{xi} L_0}. \text{ Consequently,}$$

$$V(\lambda_{xi}) = \frac{1}{I(\lambda_{xi})} = \frac{\lambda_{xi}^2}{\hat{E}(\delta_{ik})} = \frac{\lambda_{xi}^2}{1 + \frac{e^{-\lambda_{xi} L} - e^{-\lambda_{xi} (L-L_0)}}{\lambda_{xi} L_0}}.$$

By Slutsky's theorem, the MLE satisfies

$$\sqrt{n_i} (\hat{\lambda}_{xi} - \lambda_{xi}) \xrightarrow{D} N(0, V(\hat{\lambda}_{xi})).$$

Letting $\lambda_x = (\lambda_{x1}, \lambda_{x2}, \dots, \lambda_{xg})'$ and $n_1 = n_2 = \dots = n_g = n$, we have

$$\sqrt{n} (\hat{\lambda}_x - \lambda_x) \xrightarrow{D} N(0, V(\hat{\lambda}_x))$$

where $V(\hat{\lambda}_x) = \text{diag}(V(\hat{\lambda}_{x1}), \dots, V(\hat{\lambda}_{xg}))$. Considering the transition probability matrix model

$$\lambda_x = \Theta \lambda_y,$$

Where $\lambda_y = (\lambda_{y1}, \lambda_{y2}, \dots, \lambda_{yg})'$ and the transition probability matrix

$$\Theta = \begin{pmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1g} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2g} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{g1} & \theta_{g2} & \cdots & \theta_{gg} \end{pmatrix}$$

If the transition probability matrix Θ is a nonsingular matrix, then

$$\lambda_y = \Theta^{-1} \lambda_x.$$

Hence,

$$\sqrt{n} \left(C \hat{\lambda}_y - C \lambda_y \right) \xrightarrow{D} N \left(0, C \Theta^{-1} V \left(\hat{\lambda}_x \right) \Theta^{-1'} C' \right)$$

where C is a known $q \times g$ coefficient matrix of rank $q \leq g$. In particular, if $g = 2$, $C = (1 \ -1)$ and $\theta_{11} = \theta_{12} = \theta$, then

$$\sqrt{n} \left(\hat{\lambda}_{y1} - \hat{\lambda}_{y2} - (\lambda_{y1} - \lambda_{y2}) \right) \xrightarrow{D} N \left(0, V \left(\hat{\lambda}_{y1} - \hat{\lambda}_{y2} \right) \right).$$

Hence,

$$\begin{aligned} V \left(\hat{\lambda}_{y1} - \hat{\lambda}_{y2} \right) &= C \Theta^{-1} V \left(\hat{\lambda}_x \right) \Theta^{-1'} C' \\ &= \frac{1}{(2\theta - 1)^2} \left(V \left(\hat{\lambda}_{x1} \right) + V \left(\hat{\lambda}_{x2} \right) \right) \end{aligned}$$

We reject the null hypothesis $H_0 : \lambda_{y1} = \lambda_{y2}$ at the α level of significance if

$$\sqrt{n} \left| \frac{\hat{\lambda}_{y1} - \hat{\lambda}_{y2}}{\sqrt{V \left(\hat{\lambda}_{y1} - \hat{\lambda}_{y2} \right)}} \right| > Z_{\alpha/2}$$

Under the alternative hypothesis $H_a : \lambda_{y1} \neq \lambda_{y2}$, the power of the above test is approximately

$$\Phi \left(\frac{\sqrt{n} (\lambda_{y1} - \lambda_{y2})}{\sqrt{V \left(\hat{\lambda}_{y1} - \hat{\lambda}_{y2} \right)}} - Z_{\alpha/2} \right).$$

As a result, the sample size needed for achieving a power of $1 - \beta$ can be obtained by the following equation:

$$\frac{\sqrt{n} (\lambda_{y1} - \lambda_{y2})}{\sqrt{V \left(\hat{\lambda}_{y1} - \hat{\lambda}_{y2} \right)}} - Z_{\alpha/2} = Z_{\beta}.$$

This leads to

$$\begin{aligned} n &= V \left(\hat{\lambda}_{y1} - \hat{\lambda}_{y2} \right) \left(\frac{Z_{\alpha/2} + Z_{\beta}}{\lambda_{y1} - \lambda_{y2}} \right)^2 \\ &= \frac{1}{(2\theta - 1)^2} \left(V \left(\hat{\lambda}_{x1} \right) + V \left(\hat{\lambda}_{x2} \right) \right) \left(\frac{Z_{\alpha/2} + Z_{\beta}}{\lambda_{y1} - \lambda_{y2}} \right)^2 \end{aligned}$$

where

$$V \left(\hat{\lambda}_{x_i} \right) = \frac{\lambda_{x_i}^2}{\gamma e^{-\lambda_{x_i} L} \left(1 - e^{(\lambda_{x_i} - \gamma) L_0} \right) + \frac{1}{(\lambda_{x_i} - \gamma)(1 - e^{-\gamma L_0})}}$$

for $i = 1, 2$.

Non-inferiority or Superiority Trials

Let λ_{y1} and λ_{y2} are the hazard rates of the control and test drug, respectively. A smaller hazard rate is considered in favor of the test drug. The hypothesis for the superiority/non-inferiority is given by

$$H_0 : \lambda_{y1} - \lambda_{y2} \leq \delta$$

versus the alternative hypothesis

$$H_a : \lambda_{y1} - \lambda_{y2} > \delta$$

where δ stands for the superiority or non-inferiority margin. The non-inferiority test is to show that the new treatment is at least as good as the approved product. The superiority test is to show that the new treatment is better than the standard product.

We reject the null hypothesis $H_0 : \lambda_{y1} - \lambda_{y2} \leq \delta$ at the α level of significance if

$$\frac{\sqrt{n} \left(\hat{\lambda}_{y1} - \hat{\lambda}_{y2} - \delta \right)}{\sqrt{V \left(\hat{\lambda}_{y1} - \hat{\lambda}_{y2} \right)}} > Z_{\alpha}.$$

Rejection of the null hypothesis indicates that the test drug is superior to the standard treatment if the margin $\delta > 0$ or the test drug is non-inferior to the standard treatment if the margin $\delta < 0$. Under the alternative hypothesis $H_a : \lambda_{y1} - \lambda_{y2} > \delta$, the power of the above test is approximately

$$\Phi \left(\frac{\sqrt{n} (\lambda_{y1} - \lambda_{y2} - \delta)}{\sqrt{V \left(\hat{\lambda}_{y1} - \hat{\lambda}_{y2} \right)}} - Z_{\alpha} \right).$$

As a result, the sample size needed for achieving a power of $1 - \beta$ can be obtained by the following equation:

$$\frac{\sqrt{n} (\lambda_{y1} - \lambda_{y2} - \delta)}{\sqrt{V \left(\hat{\lambda}_{y1} - \hat{\lambda}_{y2} \right)}} - Z_{\alpha} = Z_{\beta}.$$

After some algebra, this leads to

$$n = V(\hat{\lambda}_{y1} - \hat{\lambda}_{y2}) \left(\frac{Z_\alpha + Z_\beta}{\lambda_{y1} - \lambda_{y2} - \delta} \right)^2$$

$$= \frac{1}{(2\theta - 1)^2} \left(V(\hat{\lambda}_{x1}) + V(\hat{\lambda}_{x2}) \right) \left(\frac{Z_\alpha + Z_\beta}{\lambda_{y1} - \lambda_{y2} - \delta} \right)^2.$$

Equivalence Trials

The new treatment is equivalent to the standard product.

The hypothesis we are interested in the equivalence trial is given by

$$H_0: |\lambda_{y1} - \lambda_{y2}| \geq \delta \quad \text{versus} \quad H_a: |\lambda_{y1} - \lambda_{y2}| < \delta$$

where δ stands for the equivalence margin. Breaking up the absolute values in the original hypothesis into two one-sided tests, we reject the null hypothesis at the α level of significance if

$$\frac{\sqrt{n}(\hat{\lambda}_{y1} - \hat{\lambda}_{y2} - \delta)}{\sqrt{V(\hat{\lambda}_{y1} - \hat{\lambda}_{y2})}} < -Z_\alpha$$

and

$$\frac{\sqrt{n}(\hat{\lambda}_{y1} - \hat{\lambda}_{y2} - \delta)}{\sqrt{V(\hat{\lambda}_{y1} - \hat{\lambda}_{y2})}} > Z_\alpha.$$

Under the alternative hypothesis $H_a: |\lambda_{y1} - \lambda_{y2}| < \delta$, the power of the above test is approximately

$$2\Phi \left(\frac{\sqrt{n}(\delta - |\lambda_{y1} - \lambda_{y2}|)}{\sqrt{V(\hat{\lambda}_{y1} - \hat{\lambda}_{y2})}} - Z_\alpha \right) - 1.$$

As a result, the sample size needed for achieving a power of $1 - \beta$ can be obtained by the following equation:

$$\frac{\sqrt{n}(\delta - |\lambda_{y1} - \lambda_{y2}|)}{\sqrt{V(\hat{\lambda}_{y1} - \hat{\lambda}_{y2})}} - Z_\alpha = Z_{\beta/2}.$$

After some algebra, this leads to

$$n = V(\hat{\lambda}_{y1} - \hat{\lambda}_{y2}) \left(\frac{Z_\alpha + Z_{\beta/2}}{\delta - |\lambda_{y1} - \lambda_{y2}|} \right)^2$$

$$= \frac{1}{(2\theta - 1)^2} \left(V(\hat{\lambda}_{x1}) + V(\hat{\lambda}_{x2}) \right)$$

$$\times \left(\frac{Z_\alpha + Z_{\beta/2}}{\delta - |\lambda_{y1} - \lambda_{y2}|} \right)^2.$$

Example

Suppose that a sponsor is planning to conduct a two-stage adaptive design trial to investigate a new treatment against the control. That is to select patient population at the first stage and then at the second stage to confirm the efficacy based on the hazard ratio between treatment and control. An interim analysis is planned to evaluate the sample size when this study is completed at the first stage. This trial at the second stage is designed to last for 3 years ($L = 3$) years with 1 year accrual ($L_0 = 1$). Assuming that uniform patient entry for both groups. That is, $\gamma = 0$. The sample variance of $\hat{\lambda}_{x_i}$ is given by

$$\hat{V}(\hat{\lambda}_{x_i}) = \frac{\hat{\lambda}_{x_i}^2}{1 + \frac{e^{-\hat{\lambda}_{x_i}L} - e^{-\hat{\lambda}_{x_i}(L-L_0)}}{\hat{\lambda}_{x_i}L_0}}.$$

Assuming that the hazard rates are $\lambda_{y1} = 1$ and $\lambda_{y2} = 2$ for the treatment group and control group, respectively. At the end of first stage, we have the sample hazard rates $\hat{\lambda}_{x1} = 1.1$ and $\hat{\lambda}_{x2} = 1.9$. The corresponding variances are $\hat{V}(\hat{\lambda}_{x1}) = 1.30$ and $\hat{V}(\hat{\lambda}_{x2}) = 3.65$. If the transition probability $\theta = 0.8$, the sample size for the second stage is calculated by

$$\hat{n} = \frac{1}{(2\theta - 1)^2} \left(\hat{V}(\hat{\lambda}_{x1}) + \hat{V}(\hat{\lambda}_{x2}) \right) \left(\frac{Z_{\alpha/2} + Z_\beta}{\lambda_{y1} - \lambda_{y2}} \right)^2$$

$$= \frac{1}{(2 \times 0.8 - 1)^2} (1.30 + 3.65) \left(\frac{1.96 + 1.28}{2 - 1} \right)^2$$

$$= 145 \text{ per block.}$$

CONCLUSION

In clinical trial, the interim analyses should be specified in the protocol if there is ethical and/or economic concern. Tharmanathan et al.^[15] point out a significant rise in the number of trials reporting use of interim analysis over last decade. An independent Data Monitoring Committee is usually established to review the interim analysis results and make recommendations to the sponsor about any action needed based on unblind results. It becomes a challenge problem if the sponsor itself plans to re-evaluate the sample size without breaking the blind in the middle of trial. It is achievable using the released data that contains the primary response and the randomization block. The transition probability matrix plays a central role in this setting. The transition probability matrix is essentially a transformation matrix between the treatment groups and the randomization blocks. A randomization scheme is desired to obtain the treatment allocation ratios of this matrix. The sample size is getting smaller as this matrix approaches to the identity matrix. The sample size also depends on the hypothesis. The calculated sample size for a non-inferiority trial is usually the smallest of the three hypothesis. The superiority comparison will have a sample

size that is much larger. In general, a equivalence trial will have a sample size that is larger than a non-inferiority trial. It must beware of $\mu_1 - \mu_2 > \delta$ for the sample size calculation of the non-inferiority or superiority trial.

In the balanced design with two treatments and two blocks, the blind sample size is comparable to the unblind sample size in terms of mean and variance. Stratified randomization for setting a probability transition matrix is recommended where the randomization scheme is performed separately within each dummy stratum, which is proposed by Shih and Zhao^[3]. For example, suppose equal allocation is planned in a two-armed trial (treatment A and B). If the patients are assigned to treatment A with specified probabilities, say 2/3 and 1/3, for first dummy stratum and second dummy stratum, respectively, the transition probability matrix here is $\Theta = \begin{pmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix}$.

The randomization procedure is designed as follows. A block size of 12 with an allocation ratio of 2:1 would lead to random assignment of 8 subjects to treatment A and 4 to the treatment B for first stratum, vice versa for second stratum. The subjects are randomly assigned to either treatment A or B with equal probability by swapping over stratum 1 and 2. In this situation, the randomization plan may look like:

Stratum 1: AABABAAAABAB... for odd number of patients.

Stratum 2: ABBBABABABBB... for even number of patients.

The treatments for first 24 patients are AAABBBABBBAA BAAABAABBABBB...

ACKNOWLEDGEMENTS

This research was partially supported by a research grant from the Ministry of Science and Technology in Taiwan (Grant No. MOST 105-2118-M-033-002).

REFERENCES

1. Lewis, J.A. Statistical principles for clinical trials (ICHE9): An introductory note on an international guideline. *Statistics in Medicine* **1999**, *18* (15), 1903–1942.
2. Gould, A.L.; Shih, W.J. Sample size re-estimation without unblinding for normally distributed outcomes with

- unknown variance. *Communications in Statistics - Theory and Methods* **1992**, *21* (10), 2833–2853.
3. Shih, W.J.; Zhao, P.L. Design for sample size re-estimation with interim data for double-blind clinical trials with binary outcomes. *Statistics in Medicine* **1997**, *16* (17), 1913–1923.
4. Bristol, D.R.; Shurzinske, L. Blinded sample size adjustment. *Drug Information Journal* **2001**, *35* (4), 1123–1130.
5. Kieser, M.; Friede, T. Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Statistics in Medicine* **2003**, *22* (23), 3571–3581.
6. Xing, B.; Ganju, J. A method to estimate the variance of an endpoint from an on-going blinded trial. *Statistics in Medicine* **2005**, *24* (12), 1807–1814.
7. Ganju, J.; Xing, B. Re-estimating the sample size of an on-going blinded trial based on the method of randomization block sums. *Statistics in Medicine* **2009**, *28* (1), 24–38.
8. Shih, W.J. Two-stage sample size reassessment using perturbed unblinding. *Statistics in Biopharmaceutical Research* **2009**, *1* (1), 74–80.
9. Friede, T.; Kieser, M. Blinded sample size re-estimation in superiority and non-inferiority trials: Bias versus variance in variance estimation. *Pharmaceutical Statistics* **2013**, *12* (3), 141–146.
10. Wu, C.H. Sample size re-estimation without breaking the blind in clinical trial. *Journal of Biopharmaceutical Statistics* **2015**, *25* (5), 1114–1129.
11. Wu, C.H. Randomized response model versus mixture model in the crossover design for clinical trials. *Communications in Statistics - Theory and Methods* **2016**, *45* (15), 4611–4627.
12. Chow, S.C.; Shao, J.; Wang, H. *Sample Size Calculations in Clinical Research*. Marcel Dekker: New York, 2003.
13. Rencher, A.C.; Schaafje, G.B. *Linear Models in Statistics*. 2nd Ed.; Wiley: New York, 2008.
14. Wu, C.H.; Wan, S.M.; Yang, Y.C.; Huang, C.Y. Sample size algorithms in clinical trials. *Drug Information Journal* **2008**, *42* (5), 429–439.
15. Tharmanathan, P.; Calvert, M.; Hampton, J.; Freemantle, N. The use of interim data and Data Monitoring Committee recommendations in randomized controlled trial reports: Frequency, implications and potential sources of bias. *BMC Medical Research Methodology* **2008**, *8*, 1–12.
16. McIntyre, J.; Robertson, S.; Norris, E.; McIntyre, J.; Robertson, S.; Norris, E.; Appleton, R.; Whitehouse, W.P.; Phillips, B.; Martland, T.; Berry, K.; Collier, J.; Smith, S.; Choonara, I. Safety and efficacy of buccal midazolam versus rectal diazepam for emergency treatment of seizures in children: A randomised controlled trial. *Lancet* **2005**, *366* (9481), 205210.
17. Yardley, L.; Donovan-Hall, M.; Smith, H.E.; Walsh, B.M.; Mullee, M.; Bronstein, A.M. Effectiveness of primary care-based vestibular rehabilitation for chronic dizziness. *Annals of Internal Medicine* **2004**, *141* (8), 598–605.

Blinding

Shein-Chung Chow

Duke University, School of Medicine, Durham, North Carolina, U.S.A.

Jun Shao

University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.

Abstract

Blinding is an approach used in clinical trials meant to eliminate bias by blocking the identity of treatments. This entry focuses on different types of clinical trial blinding and discusses the importance of this practice in obtaining honest results.

INTRODUCTION

Although randomization is used to prevent bias from a statistical assessment of the study drug, it does not guarantee that there will be no bias caused by subjective judgment in reporting, evaluation, data processing, and statistical analysis. This subjectivity is a result of the knowledge of the identity of the treatments. Because this subjective and judgmental bias is directly or indirectly related to treatment, it can seriously distort statistical inference on the treatment effect. In practice, it is extremely difficult to quantitatively assess such bias and its impact on the assessment of the treatment effect. In clinical trials, it is therefore important to consider an approach to eliminate such bias by blocking the identity of treatments. Such an approach is referred to as blinding.

TYPES OF BLINDING

Blinding in clinical trials can be classified into four types: open label, single blind, double blind, and triple blind.^[1]

Open Label

An open-label study is a clinical trial where no blinding is employed; that is, both the investigators and the patients know which treatment the patients receive. Because patients may psychologically react in favor of the treatments if they are aware of which treatment they receive, serious bias may eventually occur. Therefore open-label trials are generally not recommended for comparative clinical trials. In current practice, open-label trials are not accepted as adequate, well-controlled clinical trials for providing substantial evidence for approval by most regulatory agencies including the Food and Drug Administration (FDA). However, under certain circumstances, open-label trials are necessary. Spilker^[2] provided a list of situations

and circumstances in which open-label trials may be conducted. Ethical consideration is always an important factor, or perhaps the only factor that is used to determine whether a trial should be conducted in an open-label fashion. For example, Phase I studies on dose escalation for the determination of the maximum tolerable dose of drugs in treating terminally ill cancer patients are usually open-labeled. Clinical trials for evaluation of the effectiveness and safety of a new surgical procedure are usually conducted in an open-label fashion. This is because it is clearly unethical to conduct a double-blind trial with a concurrent control group in which patients are incised under a general anesthesia to stimulate the surgical procedure. Note that premarketing and postmarketing surveillance studies are usually open-labeled. The purpose of premarketing surveillance studies is to collect the data of efficacy and safety with respect to the duration of exposure of a broader patient population to the test drug, while the objective of postmarketing surveillance studies is to monitor the safety and tolerability of the drug product.

Single Blind

In practice, a single-blind trial is referred to as a trial in which the patient's treatment assignment is not known to the patient only. As compared with open-label trials, single-blind studies offer a certain degree of control and the assurance of the validity of clinical trials. However, by knowing which treatment the patients receive, the investigator's clinical evaluation may be biased.

Double Blind

A double-blind trial is a trial in which neither the patients nor the investigators (study centers) are aware of patients' treatment assignment. Note that the "investigator" could mean all of the health-care personnel, which include the study center, the contract laboratories, and the other

consulting experts for evaluation of effectiveness and safety of patients in a broader sense. In addition to the patients and the investigators, if all members of clinical project team associated with the study are also blinded, then the clinical trial is said to be triple blinded. These members include the project clinicians, clinical research associates (CRA), statisticians, programmers, and data coordinators. In addition to patients’ treatment assignment, the blindness also applies to concealment of the overall results of the trial. In practice, although the project clinicians, CRA, statisticians, programmers, and data coordinators usually have access to the individual patient’s data, they are generally not aware of the treatment assignment for each patient. In addition, the overall treatment results, if any (such as interim analyses), will not be made available to the patients, the investigators, the project clinicians, CRA, statisticians, programmers, and data coordinators until a decision is made at an appropriate time.

Triple Blind

A triple-blind study with respect to blindness can provide the highest degree for the validity of a controlled clinical trial. Hence it provides the most conclusive unbiased evidence for the evaluation of the effectiveness and safety of the therapeutic intervention under investigation.

Note that triple blinding is generally reserved for large cooperative, multicenter studies monitored by a committee; however, it has been applied to company monitors to ensure they remain unaware of treatment allocation. For the current practice, double blinding has become a gold standard for clinical research because it provides the greatest probability for reducing bias.

THE INTEGRITY OF BLINDING

In practice, even with the best intentions for preserving blindness throughout a clinical trial, blindness can sometimes be breached for various reasons. One method to determine whether the blindness is seriously violated is to ask patients to guess their treatment codes during the study or at the conclusion of the trial prior to unblinding. In some cases, investigators are also asked to guess the patients’ treatment codes. Once the guesses are recorded on the case report forms and entered into the database, the integrity of blinding can be tested. When the integrity of blinding is doubtful, adjustments to statistical analysis should be made.

Data on guessing treatment codes are typically of the form given in Tables 1 and 2 in the following two examples.

The first example is a 1-year double-blind placebo-controlled study conducted by the National Institutes of Health to evaluate the distinguish between the prophylactic and therapeutic effects of ascorbic acid for the common

cold.^[3] A two-group parallel design was used. At the completion of the study, a questionnaire was distributed to everyone enrolled in the study so that they could guess which treatment they had been taking. Results from the 190 subjects who completed the study are given in Table 1.

The second example is a double-blind placebo-controlled trial with a two-group parallel design for evaluation of the effectiveness of an appetite suppressant in weight loss in obese women.^[4] Table 2 lists the data on patients’ guesses of the treatment codes.

To test the integrity of blinding, we need to define a null hypothesis H_0 . Consider a single-site parallel design comparing $a \geq 2$ treatments. Let A_i be the event that a patient guesses that he/she is in the i th group, $i = 1, \dots, a$. Let $A_a + 1$ be the event that a patient does not guess (or answers “do not know”). Finally, let B_j be the event that a patient is assigned to the j th group, $j = 1, \dots, a$. If the hypothesis

$H_0: A_i \text{ and } B_j \text{ are independent for any } i \text{ and } j$

holds, then the blindness is considered to be preserved. The hypothesis H_0 can be tested using the well-known Pearson’s chi-square test under the contingency tables constructed based on observed counts such as those given by Tables 1 and 2. Analysis on investigators’ guesses of patients’ treatment codes can be similarly performed.

A straightforward calculation using data in Tables 1 and 2 results in observed Pearson’s chi-square statistics 31.3 and 22.5, respectively. Thus the null hypothesis of independence of A_i and B_j is rejected at a very high significance level (the p values are smaller than 0.001). That is, in both examples, the blindness is not preserved.

Table 1 Results of Patients’ Guess on Treatment for the Prophylactic Use

Patient’s guess	Actual assignment	
	Ascorbic acid	Placebo
Ascorbic acid	40	11
Placebo	12	39
Do not know	49	39
Total	101	89

From Ref. [3].

Table 2 Results of Patients’ Guess on Treatment for Weight Loss

Patient’s guess	Actual assignment	
	Active drug	Placebo
Active drug	19	3
Placebo	3	16
Do not know	2	6
Total	24	25

From Ref. [4].

ANALYSIS UNDER BREACHED BLINDNESS

When the test of the integrity of blinding resulted in a significant result, analyzing data by ignoring this result may lead to a biased result; that is, the integrity of blinding is doubtful. In what follows, we introduce a method of testing treatment effects by incorporating the data of patients' guesses of their treatment codes.^[5] The idea is to include patient's guess as a factor in the analysis of variance (ANOVA) for the treatment effects.

Suppose that the study design is a single-site parallel design comparing $a \geq 2$ treatments. If the blindness is preserved, then treatment effects can be tested using the one-way ANOVA. Let γ be the factor of guessing treatment codes, which has b levels. Including factor γ in the analysis with patients' guessing data and assuming that patients' guessing is independent of their response value, we can test treatment effects by using a two-way ANOVA. If the study is a multicenter trial, then including factor γ leads to a three-way ANOVA.

There are different ways to construct the variable γ . One way is to use guessing treatment $i, i = 1, \dots, a$, as the first a levels of γ and not guessing (do not know) as the last level. Hence $b = a + 1$. Another way is to use guessing correctly, guessing incorrectly, and not guessing as three levels for γ and thus $b = 3$.

Even if the original design is balanced, i.e., each treatment (and center) has the same number of patients, the two-way ANOVA or three-way ANOVA after including factor γ is not balanced. Hence methods for unbalanced ANOVA are necessarily considered. In the following, we illustrate the idea using a single study site, i.e., two-way ANOVA.

Let x_{ijk} be the response from the k th patient under the i th treatment with the j th guessing status, where $i = 1, \dots, a, j = 1, \dots, b, k = 1, \dots, n_{ij}$ and n_{ij} is the number of patients in the (i, j) th cell. Let $\bar{x}_{ij}$ be the sample mean of the patients in the (i, j) th cell, $\bar{x}_i$ be the sample mean of the patients under treatment i , $\bar{x}_j$ be the sample mean of patients with guessing status j , $\bar{x}$ be the sample mean of all patients n_j be the number of patients under treatment i , n_j be the number of patients with guessing status j , and n be the total number of patients. Define

$$R(\mu) = n\bar{x}^2$$

$$R(\mu, \tau) = \sum_{i=1}^a n_i \bar{x}_i^2$$

where τ denotes treatment effect and μ denotes the overall mean,

$$R(\mu, \gamma) = \sum_{j=1}^b n_j \bar{x}_j^2$$

$$R(\mu, \tau, \gamma, \tau \times \gamma) = \sum_{i=1}^a \sum_{j=1}^b n_{ij} \bar{x}_{ij}^2$$

where $\tau \times \gamma$ denotes the interaction between τ and γ , and

$$R(\mu, \tau, \gamma) = \sum_{i=1}^a n_i \bar{x}_{i-}^2 + Z' C^{-1} Z$$

where Z is a $(b-1)$ vector whose j th component is

$$n_{.j} \bar{x}_{.j} - \sum_{i=1}^a n_{ij} \bar{x}_{i-}, \quad j = 1, \dots, b-1$$

and C is a $(b-1) \times (b-1)$ matrix whose j th diagonal element is

$$n_{.j} - \sum_{i=1}^a n_{ij}^2 / n_i$$

and (j, l) th off-diagonal element is

$$- \sum_{i=1}^a n_{ij} n_{il} / n_i$$

Let

$$R(\tau \times \gamma \mid \mu, \tau, \gamma) = R(\mu, \tau, \gamma, \tau \times \gamma) - R(\mu, \tau, \gamma)$$

$$R(\tau \mid \mu) = R(\mu, \tau) - R(\mu)$$

$$R(\gamma \mid \mu) = R(\mu, \gamma) - R(\mu)$$

$$R(\tau \mid \mu, \gamma) = R(\mu, \tau, \gamma) - R(\mu, \gamma)$$

$$R(\gamma \mid \mu, \tau) = R(\mu, \tau, \gamma) - R(\mu, \tau)$$

$$SSE = \sum_{i=1}^a \sum_{j=1}^3 \sum_{k=1}^{n_{ij}} x_{ijk}^2 - R(\mu, \tau, \gamma, \tau \times \gamma)$$

and s be the number of nonzero n_{ij} s. An ANOVA table according to Searle^[6] can be constructed (Table 3).

An F ratio (in the last column of Table 3) is said to be significant at the α level if it is larger than the $(1 - \alpha)$ th quantile of the F distribution with denominator degrees of freedom $n - s$ and the numerator degrees of freedom given by the number in the third column of the same row. Note that $F(\tau \mid \mu)$ is the F ratio for testing τ effect (treatment effect) adjusted for μ and ignoring γ , whereas $F(\tau \mid \mu, \gamma)$ is the F ratio for testing τ effect adjusted for both μ and γ . These two F ratios are the same in a balanced model but are different in an unbalanced model. Similar discussion can be made for $F(\gamma \mid \mu)$ and $F(\gamma \mid \mu, \tau)$.

Because of its imbalance, the interpretation of the results given by F ratios in the ANOVA table is not straightforward. Consider first the case where the interaction $F(\tau \times \gamma \mid \mu, \tau, \gamma)$ is not significant. Table 4 lists a total of 14 possible cases according to the significance of F ratios $F(\tau \mid \mu)$, $F(\tau \mid \mu, \gamma)$, $F(\gamma \mid \mu)$, and $F(\tau \mid \mu)$. The suggestion from Searle^[6] regarding which effects should

Table 3 ANOVA for Treatment Effects Under Breached Blindness

Source	Sum of squares	Degrees of freedom	F ratio
τ after μ	$R(\tau \mu)$	$a - 1$	$F(\tau \mu) = \frac{R(\tau \mu)/(a-1)}{SSE/(n-s)}$
γ after μ and τ	$R(\gamma \mu, \tau)$	$b - 1$	$F(\gamma \mu, \tau) = \frac{R(\gamma \mu, \tau)/(b-1)}{SSE/(n-s)}$
γ after μ	$R(\gamma \mu)$	$b - 1$	$F(\gamma \mu) = \frac{R(\gamma \mu)/(b-1)}{SSE/(n-s)}$
τ after μ and γ	$R(\tau \mu, \gamma)$	$a - 1$	$F(\tau \mu, \gamma) = \frac{R(\tau \mu, \gamma)/(a-1)}{SSE/(n-s)}$
Interaction	$R(\tau \times \gamma \mu, \tau, \gamma)$	$s - a - b + 1$	$F(\tau \times \gamma \mu, \tau, \gamma) = \frac{R(\tau \times \gamma \mu, \tau, \gamma)/(s-a-b+1)}{SSE/(n-s)}$
Error	SSE	$n - s$	
Total	SS(TO)	$n - 1$	

be included in the model is given in the second-last column of Table 4. However, our purpose is slightly different; that is, we are interested in whether the treatment effect τ is significant regardless of the presence of the effect γ . Our recommendations in these 14 cases are given in the last column of Table 4, which is interpreted as follows. When both $F(\tau|\mu)$ and $F(\tau|\mu, \gamma)$ are significant (the first 4 rows of Table 4), regardless of whether γ effect is significant or not, the conclusion is easy to make; that is, the treatment effect is significant. In the next three cases (rows 5–7 of Table 4), $F(\tau|\mu)$ is not significant but $F(\tau|\mu, \gamma)$

is significant, indicating that the treatment effect cannot be clearly detected by ignoring γ but once γ is included in the model as a blocking variable, the treatment effect is significant. In these three cases we conclude that the treatment effect is significant. When $F(\tau|\mu, \gamma)$ is not significant but $F(\gamma|\mu)$ is significant, it indicates that once γ is fitted into the model, the treatment effect is not significant; that is, the treatment effect is distorted by the γ effect. In such cases (rows 8–11 in Table 4), we cannot conclude that the treatment effect is significant. In the last three cases (rows 12–14 of Table 4), both $F(\gamma|\mu)$ and $F(\tau|\mu, \gamma)$ are

Table 4 Conclusions on the Significance of the Treatment Effect When $F(\tau \times \gamma|\mu, \tau, \gamma)$ is Insignificant

Significance of F ratio				Effects to be included significance in the model ^[6]	Conclusion: significance of the treatment effect
Fitting τ and then γ after τ		Fitting γ and then τ after γ			
$F(\tau \mu)$	$F(\gamma \mu,\tau)$	$F(\gamma \mu)$	$F(\tau \mu,\gamma)$		
Yes	Yes	Yes	Yes	τ, γ	Yes
Yes	Yes	No	Yes	τ, γ	Yes
Yes	No	Yes	Yes	τ	Yes
Yes	No	No	Yes	τ	Yes
No	Yes	Yes	Yes	τ, γ	Yes
No	Yes	No	Yes	τ, γ	Yes
No	No	No	Yes	τ, γ	Yes
No	Yes	Yes	No	γ	No
No	No	Yes	No	γ	No
Yes	Yes	Yes	No	γ	No
Yes	No	Yes	No	τ, γ	No
Yes	No	No	No	τ	Yes
No	Yes	No	No	τ, γ	No
No	No	No	No	None	No

Table 5 Mean Weight Loss (kg)

Patient's guess	Actual assignment	
	Active drug	Placebo
Active drug	9.6	2.6
Placebo	3.9	6.1
Do not know	12.2	5.8
Total	9.1	5.6

From Ref. [4].

not significant. If $F(\tau|\mu)$ is significant but $F(\gamma|\mu, \tau)$ is not (row 12 of Table 4), it indicates that γ has no effect and the treatment effect is significant. On the other hand, if $F(\gamma|\mu, \tau)$ is significant but $F(\tau|\mu)$ is not (row 13 of Table 4; a case “should happen somewhat infrequently” according to Ref. [6]), we cannot conclude that the treatment effect is significant. Finally, when neither $F(\tau|\mu)$ nor $F(\gamma|\mu, \tau)$ is significant, then neither treatment effect nor γ effect is significant (row 14 of Table 4).

The analysis is difficult when the interaction $F(\tau \times \gamma|\mu, \tau, \gamma)$ is significant. In general, we cannot conclude that the treatment effect is significant when $F(\tau \times \gamma|\mu, \tau, \gamma)$ is significant. Analysis conditional on the value of γ may be carried out to draw some partial conclusions. An illustration is given in the following example.

AN EXAMPLE

Observed mean weight loss (in kilograms) in the problem of the effectiveness of an appetite suppressant in weight loss in obese women^[4] is shown in Table 5. Recall that the analysis in “The Integrity of Blinding” shows that in this example, the blindness is not preserved with a high significance.

Source	R	df	R/df	F ratio	p value
τ after μ	$R(\tau \mu)$	1	160.2	6.43	0.015
y after μ and τ	$R(y \mu, \tau)$	2	47.2	1.90	0.0162
y after μ	$R(y \mu)$	2	41.9	1.68	0.198
τ after μ and y	$R(\tau \mu, y)$	1	17.9	0.72	0.401
Interaction	$R(\tau \times y \mu, \tau, y)$	2	61.3	2.46	0.907
Error	SSE	43	24.9		

First, consider the analysis with y = guessing correctly, guessing incorrectly, and not guessing. Then,

$\bar{x}_{11} = 9.6$	$n_{11} = 19$	$\bar{x}_{21} = 6.1$	$n_{21} = 16$
$\bar{x}_{12} = 3.9$	$n_{12} = 3$	$\bar{x}_{22} = 2.6$	$n_{22} = 3$
$\bar{x}_{13} = 12.2$	$n_{13} = 2$	$\bar{x}_{23} = 5.8$	$n_{23} = 6$
$\bar{x}_{1..} = 9.1$	$n_{1..} = 24$	$\bar{x}_{2..} = 8.0$	$n_{2..} = 35$
$\bar{x}_{2..} = 5.6$	$n_{2..} = 25$	$\bar{x}_{3..} = 3.3$	$n_{3..} = 6$
$\bar{x} = 7.3$	$n = 49$	$\bar{x}_3 = 7.4$	$n_3 = 8$

The resulting ANOVA table (see Table 3) is given by

Source	R	df	R/df	F ratio	p value
τ after μ	$R(\tau \mu)$	1	160.2	6.43	0.015
y after μ and τ	$R(y \mu, \tau)$	2	72.1	2.90	0.066
y after μ	$R(y \mu)$	2	32.6	1.31	0.280
τ after μ and y	$R(\tau \mu, y)$	1	115.2	4.63	0.037
Interaction	$R(\tau \times y \mu, \tau, y)$	2	12.6	0.51	0.604
Error	SSE	43	24.3		

It seems that the interaction $F(\tau \times \gamma|\mu, \tau, \gamma)$ is not significant and both treatment-effect F ratios $F(\tau|\mu)$ and $F(\tau|\mu, \gamma)$ are significant. Thus according to the previous discussion (see Table 4), we can conclude that the treatment effect is significant, regardless of whether the effect of γ is significant or not.

However, the conclusion may be different if we consider the levels of γ to be guessing active drug, guessing placebo, and not guessing. With this choice of levels of γ , the sample means are given by

$\bar{x}_{11} = 9.6$	$n_{11} = 19$	$\bar{x}_{21} = 2.6$	$n_{21} = 3$
$\bar{x}_{12} = 3.9$	$n_{12} = 3$	$\bar{x}_{22} = 6.2$	$n_{22} = 16$
$\bar{x}_{13} = 12.2$	$n_{13} = 2$	$\bar{x}_{23} = 5.8$	$n_{23} = 6$
$\bar{x}_{1..} = 9.1$	$n_{1..} = 24$	$\bar{x}_{2..} = 8.7$	$n_{2..} = 22$
$\bar{x}_{2..} = 5.6$	$n_{2..} = 25$	$\bar{x}_{3..} = 5.7$	$n_{3..} = 19$
$\bar{x} = 7.3$	$n = 49$	$\bar{x}_3 = 7.4$	$n_3 = 8$

Note that $\bar{x}_{1i}$ are unchanged but $\bar{x}_{2j}$ are changed with this new choice of levels of γ . The corresponding ANOVA table is given at the bottom of the page.

Although $F(\tau|\mu)$ remains the same, the value of $F(\tau \times \gamma|\mu, \tau, \gamma)$ is much larger than that in the previous case. The p value corresponding to $F(\tau \times \gamma|\mu, \tau, \gamma)$ is 0.097, which indicates that the interaction between the treatment and γ is marginally significant. If this interaction is ignored, then we may conclude that the treatment effect is significant because the only significant F ratio is $F(\tau|\mu)$. But no conclusion can be made if the interaction effect cannot be ignored. In the presence of interaction, a subgroup analysis (according to the levels of γ) may be useful.

Subgroup sample mean comparisons can be made as indicated in Fig. 1. Figure 1 displays six subgroup sample means $\bar{x}_{ij}, i = 1, 2, j = 1, 2, 3$. Fig. 1(A) considers the situation where γ = guessing correctly, guessing incorrectly, and not guessing, while Fig. 1(B) considers the situation where γ = guessing active drug, guessing placebo, and not guessing. The two sample means (dots) corresponding to the same γ level are connected by a straight line segment. In part (A) of the figure, although the three line segments have different slopes, the slopes have the same sign; furthermore, every pair of two line segments either does not cross or crosses slightly. This indicates that in the situation considered by part (A) of the figure, there is no significant

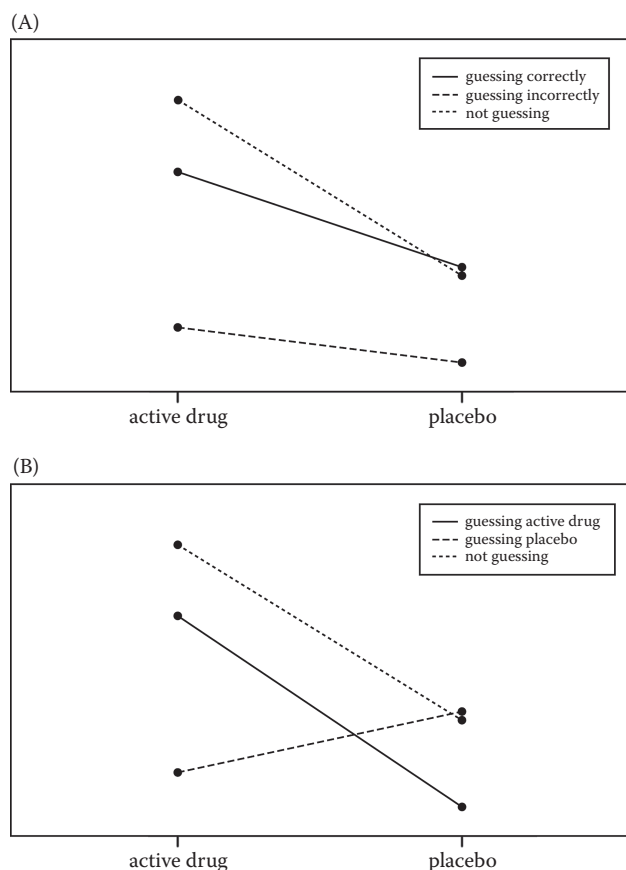


Fig. 1 Subgroup sample mean comparison

interaction and the treatment effect is evident. On the other hand, the slopes of the line segments in part (B) of Fig. 1 have different signs and two line segments considerably cross, which indicates that the interaction is significant and that we cannot draw an overall conclusion on the treatment effect in the situation considered by part (B) of the figure.

REFERENCES

1. Chow, S.C.; Liu, T.P. *Design and Analysis of Clinical Trials*, 2nd Ed.; John Wiley and Sons: New York, 2003.
2. Spilker, B. *Guide to Clinical Trials*; Raven Press: New York, 1991.
3. Karlowski, T.R.; Chalmers, T.C.; Frenkel, L.D.; Kapikian, A.Z.; Lewis, T.L.; Lynch, J.M. Ascorbic acid for the common cold: A prophylactic and therapeutic trial. *J. Am. Med. Assoc.* **1975**, *231*, 1038–1042.
4. Brownell, K.D.; Stunkard, A.J. The double-blind in danger: Untoward consequences of informed consent. *Am. J. Psychiatr.* **1982**, *139*, 1487–1489.
5. Chow, S.C.; Shao, J. *Statistics in Drug Research-Methodology and Recent Development*, Marcel Dekker, Inc.: New York, 2002.
6. Searle, S.R. *Linear Models*, John Wiley: New York, 1971.

Bootstrap, The

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Jun Shao

University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.

Abstract

This entry details the bootstrap approach to statistical analysis by providing several examples of this method to test its validity and applications.

INTRODUCTION

The process of a statistical analysis based on a set of observed data can be briefly described as follows. We believe that the observed data are from a postulated (parametric or nonparametric) model. Unknown quantities in the postulated model are estimated based on the observed data. Estimators are assessed using some accuracy measures (e.g., variance or mean squared error), and further inference may be made based on these estimators and their accuracy measures and/or distributional properties.

OVERVIEW

In the traditional approach, accuracy measures and distributions of estimators are estimated by first deriving their theoretical (approximate) formulas and then by estimating unknown quantities in these formulas. The derivations involved may be difficult, complicated, and/or tedious. For example, when the chosen estimator is $g(\bar{X})$, where $\bar{X}$ is the sample mean of some independent and identically distributed (iid) random vectors and g is a differentiable function, applying Taylor's expansion and the Central Limit Theorem leads to approximate formulas of the variance and the distribution of $g(\bar{X})$, i.e., $g(\bar{X})$ is approximately normally distributed with a mean of $g(\mu)$ and a variance of $[\nabla g(\mu)]'V(\bar{X})\nabla g(\mu)$, where ∇g is the vector of partial derivatives of g , $\mu = E(\bar{X})$, and $V(\bar{X})$ is the variance–covariance matrix of $\bar{X}$. The derivation of ∇g may be quite involved if the problem is high-dimensional and the function g is defined implicitly or in a complicated manner. As an example, consider the estimation of the poverty line (low income cutoff) of the population in the study of income shares or wealth distributions. For the i th sampled family, let z_i be the expenditure on necessities, let y_i be the total income, and let x_{it} ($t = 1, \dots, T$) be the variables such as urbanization category and family size. Then:

$$\log z_i = \gamma_1 + \gamma_2 \log y_i + \sum_{t=1}^T \beta_t x_{it} + \text{error} \quad (1)$$

where γ and β are unknown parameters. Let γ_0 be the overall proportion of income spent on necessities. Then the poverty line θ can be defined as the solution of:

$$\log[(\gamma_0 + 0.2)\theta] = \gamma_1 + \gamma_2 \log \theta + \sum_{t=1}^T \beta_t x_{0t} \quad (2)$$

for a particular set of $x_{01}, \dots, x_{0T}$.^[1] Suppose that the regression parameters in Eq. 1 are estimated by the least squares estimators $\hat{\gamma}_j$ and $\hat{\beta}_t$ and γ_0 is estimated by $\bar{z}/\bar{y}$, where $\bar{z}$ and $\bar{y}$ are the sample means of z_i and y_i , respectively. Then a natural estimator $\hat{\theta}$ of θ is the solution of Eq. 2 with γ_j and β_t replaced by $\hat{\gamma}_j$ and $\hat{\beta}_t$, respectively. Let $X_i = (z_i, y_i, \log z_i, \log y_i, x_{i1}, \dots, x_{iT})$. Then $\hat{\theta}$ can be written as $g(\bar{X})$ for an implicit function g . The derivation of ∇g is complicated. Even if g is not implicit, the derivation of the derivatives is tedious when T is large. More examples can be found in Shao and Steel^[2] and Shao and Wang.^[3]

Suppose that we can generate B (a large number) data sets from the postulated model. Then, we may replace the derivations in the traditional approach by empirical estimates based on the B data sets. For example, instead of using the theoretical formula for the variance of an estimator, we can calculate the same estimator repeatedly B times and use the sample variance based on the resulting B estimates. We may call this the simulation approach.

In practice, however, generating multiple data sets is impossible because the postulated model usually involves unknown quantities. What we can do is to substitute unknown quantities in the postulated model by some estimates (based on the original data) and then apply the simulation approach. This combination of substitution and simulation is exactly the *bootstrap* method proposed by Efron,^[4] which replaces complicated theoretical derivations

by repeatedly generating data and computing estimates. (In “Generating Bootstrap Data” and “Using Bootstrap Data,” we shall illustrate how to use the bootstrap in the previously discussed poverty line problem.) Because of the availability of high-speed computers, the bootstrap has become very popular and attractive to researchers in recent years.

There are three key issues in applying the bootstrap method. First, we need to know how the bootstrap data sets are generated. Second, we need to choose a procedure of using the bootstrap data in inference. Finally, we need to establish the validity of the bootstrap, although the work required in establishing the validity of the bootstrap is different from the derivations required by the traditional approach.

GENERATING BOOTSTRAP DATA

In general, bootstrap data should be generated by mimicking the process of generating the original data.

Parametric Bootstrap

Suppose that the original data are from a parametric population $P\theta$, where θ is a vector of unknown parameters. If $\hat{\theta}$ is the chosen estimator of θ based on the original data, then bootstrap data should be generated from $P_{\hat{\theta}}$. For example, if $X_1, \dots, X_n$ are iid with a density of

$$\frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right)$$

where μ and $\sigma > 0$ are unknown parameters and f is a known density, then the bootstrap data may be generated from the density

$$\frac{1}{\hat{\sigma}} f\left(\frac{x - \hat{\mu}}{\hat{\sigma}}\right)$$

where $\hat{\mu}$ and $\hat{\sigma}$ are estimators of μ and σ , respectively, based on $X_1, \dots, X_n$ (e.g., the maximum likelihood estimators). This is referred to as the parametric bootstrap.

Nonparametric Bootstrap

When the original data $X_1, \dots, X_n$ are iid from a distribution $\hat{F}$ from a nonparametric family, a simple nonparametric bootstrap generates bootstrap data from the empirical distribution $\hat{F}$ (constructed using $X_1, \dots, X_n$), which amounts to generating a simple random sample with replacement from the observed values $X_1, \dots, X_n$.

In the comparison of K treatments, the original data are independently generated from K populations (i.e., $X_{k1}, \dots, X_{kn_k}$ are iid from F_k , $k = 1, \dots, K$). In such a case, a

stratified bootstrap should be applied (i.e., for each k , we generate a simple random sample from $X_{k1}, \dots, X_{kn_k}$, and the bootstrap data are generated independently across k).

Longitudinal Bootstrap

In medical studies, frequently data are longitudinal in the sense that repeated (dependent) measurements are obtained from each sampled subject. The bootstrap data should also have the same or very similar dependence structure. The simplest way is to use the subject as the primary unit in generating bootstrap data. For example, suppose that a 2×2 crossover design is used for comparing two treatments, A and B ; i.e., n subjects are randomly assigned to two groups (of sizes n_1 and n_2) and subjects in group 1 (or group 2) receive two different treatments on the order of A and B (or B and A). Let Y_{kij} be the j th observation from the i th subject in the k th group. Then, each subject has a vector of two observations $X_{ki} = (Y_{ki1}, Y_{ki2})$. Bootstrap data can be generated by taking a simple random sample with replacement from the vectors $X_{k1}, \dots, X_{kn_k}$, independently across k .

Bootstrapping Linear Models

Suppose that the original data follow a linear model:

$$E(y_i | Z_i) = \beta' Z_i, \quad i = 1, \dots, n \quad (3)$$

where Z_i is the i th value of a covariate, β is a vector of unknown parameters, $E(y|z)$ denotes the conditional expectation of y given z , and y_i is independent. Note that model (1) is a special case of model (3).

If Z_i is a random observation, then a nonparametric bootstrap, which is also referred to as bootstrapping pairs, generates bootstrap data by taking a simple random sample with replacement from the pairs $(y_1, Z_1), \dots, (y_n, Z_n)$. This bootstrap is actually the same as that in “Nonparametric Bootstrap.”

Under model (3), it is often more efficient to apply the following semiparametric bootstrap. Suppose that the regression parameter β in Eq. 3 is estimated by $\hat{\beta}$ (e.g., the least squares estimator). Assume that given Z_i s, $e_i = y_i - \beta' Z_i$, $i = 1, \dots, n$, are iid with a mean of 0 and that:

$$\sum_{i=1}^n \hat{e}_i = 0$$

where $\hat{e}_i = y_i - \hat{\beta}' Z_i$ is the i th residual. Then, a bootstrap sample of e s can be generated by taking a simple random sample with replacement from the residuals $\hat{e}_1, \dots, \hat{e}_n$ and the i th bootstrap y value is obtained by adding $\hat{\beta}' Z_i$ to the i th bootstrap value of e . This bootstrap is commonly referred to as bootstrapping residuals.

Without-Replacement Bootstrap

In sample surveys, the original dataset is often sampled without replacement from a finite population. Consider for simplicity the case where $X_1, \dots, X_n$ is a simple random sample without replacement from a finite population with N units. Ideally, one should generate a bootstrap sample by taking a simple random sample without replacement from $X_1, \dots, X_n$, where the bootstrap sampling fraction is the sample as the original sampling fraction n/N . Such a bootstrap procedure, however, is often impractical. For example, if $n = 100$ and $N = 10,000$, then $n/N = 1\%$ and the bootstrap sample size has to be 1.

The following without-replacement bootstrap method was proposed in Gross^[5] and Chao and Lo.^[6] Assume for simplicity that $N = kn$ with an integer k . We first create a pseudo-population of size N by replicating $X_1, \dots, X_n$ exactly k times and then generate a bootstrap sample by taking a simple random sample of size n without replacement from the pseudo-population. Note that the bootstrap sampling fraction is the same as the original sampling fraction.

Other modified without-replacement bootstrap procedures can be found in Sitter.^[7]

Bootstrapping Imputed Data

Nonresponse often occurs in medical studies and sample surveys. One popular method of handling nonresponse is imputation. Suppose that we have data (X_i, a_i) , $i = 1, \dots, n$, where $a_i = 1$ indicates that X_i is a respondent and $a_i = 0$ indicates that X_i is an imputed value (i.e., the original value of X_i is missing). Usually, the imputation procedure is constructed so that estimators calculated by treating imputed values as observed data and using formulas for the case of no response are still valid (i.e., almost unbiased and consistent). However, applying the bootstrap by treating imputed values as observed data usually produces wrong results.

A bootstrap procedure that takes imputation into account can be described as follows.^[8,9] Suppose that the original data are iid. First, we generate $X_1^*, \dots, X_n^*$ by taking a simple random sample with replacement from $X_1, \dots, X_n$. Let a_i^* be the a value corresponding to X_i^* . If $a_i^* = 1$, we set $\tilde{X}_i^* = X_i^*$. If $a_i^* = 0$, we obtain an imputed $\tilde{X}_i^*$ by using the same imputation procedure for the original dataset. The resulting bootstrap dataset is then $\tilde{X}_1^*, \dots, \tilde{X}_n^*$.

USING BOOTSTRAP DATA

Bootstrap data are usually used in variance estimation and/or construction of confidence sets.

Bootstrap Variance Estimators

Consider the variance estimation problem for a given estimator $\hat{\theta}$. Let X^* be a bootstrap dataset generated

according some procedure (e.g., one of the methods in “Generating Bootstrap Data”). Let X^{*b} , $b = 1, \dots, B$, be iid copies of X^* , and let $\hat{\theta}^{*b}$ be the same as $\hat{\theta}$ but based on the bootstrap dataset X^{*b} . The bootstrap variance estimator for $\hat{\theta}$ is:

$$v_B = \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}^{*b} - \bar{\theta}^*)^2$$

where $\bar{\theta}^*$ is the average of $\hat{\theta}^{*b}$, $b = 1, \dots, B$.

Consider for example the estimation of poverty line described in the “Introduction.” A bootstrap variance estimator for $\hat{\theta}$ can be obtained by generating bootstrap dataset X^{*b} according to one of the procedures in “Bootstrapping Linear Models” and by computing $\hat{\theta}^{*b}$ by solving Eq. 2 based on X^{*b} , $b = 1, \dots, B$. Although the amount of computation is large, the computation is routine once a program of solving Eq. 2 is available.

Bootstrap Confidence Sets: Percentile Type

Let θ be a real valued parameter of interest. Consider the problem of constructing a confidence bound or interval for θ . Efron^[10] provided the following bootstrap percentile method. Let $\hat{\theta}^{*b}$ be defined as in “Bootstrap Variance Estimators” and let $\hat{\theta}^*(\alpha)$ be the “sample” 100 α th percentile of $\hat{\theta}^{*b}$, $b = 1, \dots, B$. Then, a $1 - \alpha$ upper (or power) confidence bound for θ is $\hat{\theta}^*(1 - \alpha)$ (or $\hat{\theta}^*(\alpha)$) and a $1 - 2\alpha$ confidence interval for θ is $[\hat{\theta}^*(\alpha), \hat{\theta}^*(1 - \alpha)]$.

To improve the bootstrap percentile confidence bounds and intervals, Efron^[10] proposed the bootstrap bias-corrected percentile confidence bounds and intervals. The bootstrap bias-corrected percentile lower confidence bound is $\hat{\theta}^*(\beta)$, where $\beta = \Phi(\Phi^{-1}(\alpha) + 2z_0)$, Φ is the standard normal distribution function, $z_0 = \Phi^{-1}(c/B)$, and c is the number of $\hat{\theta}^{*b}$ values that are less or equal to $\hat{\theta}$. Note that $z_0 = 0$ (i.e., the bias-corrected percentile method is the same as the percentile method) if $\hat{\theta}$ is the median of the bootstrap estimators $\hat{\theta}^{*b}$, $b = 1, \dots, B$.

Efron^[11] proposed a bootstrap accelerated bias-corrected percentile method that further improves the bootstrap bias-corrected percentile method. However, applications of the bootstrap accelerated bias-corrected percentile method involve some derivations that may be complicated.

Bootstrap Confidence Sets: t-Type

Confidence sets for θ are often constructed using the distribution of $\sqrt{n}(\hat{\theta} - \theta)$ or $\sqrt{n}(\hat{\theta} - \theta)/\hat{\sigma}$ where $\hat{\sigma}^2$ is an estimator of the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta)$. If the 100(1 - α)th percentile of the distribution of $\sqrt{n}(\hat{\theta} - \theta)$ is known and denoted by $z_{1-\alpha}$, then a $1 - \alpha$ lower confidence bound for θ is:

$$\hat{\theta} - n^{-1/2} z_{1-\alpha}$$

If the $100(1 - \alpha)$ th percentile of the distribution of $\sqrt{n}(\hat{\theta} - \theta)/\hat{\sigma}$ is known and denoted by $t_{1-\alpha}$, then a $1 - \alpha$ lower confidence bound for θ is:

$$\hat{\theta} - n^{-1/2} \hat{\sigma} t_{1-\alpha}$$

However, $z_{1-\alpha}$ and $t_{1-\alpha}$ are typically unknown.

If we approximate the distribution of $\sqrt{n}(\hat{\theta} - \theta)$ by the empirical distribution based on the bootstrap estimators $\sqrt{n}(\hat{\theta}^{*b} - \hat{\theta})/\hat{\sigma}^{*b}$, $b = 1, \dots, B$, then $z_{1-\alpha}$ can be estimated by $z_{1-\alpha}^*$, the 100α th sample percentile of $\sqrt{n}(\hat{\theta}^{*b} - \hat{\theta})/\hat{\sigma}^{*b}$, $b = 1, \dots, B$ and the resulting $1 - \alpha$ bootstrap lower confidence bound for θ is:

$$\hat{\theta} - n^{-1/2} \hat{\sigma} z_{1-\alpha}^* \quad (4)$$

If we approximate the distribution of $\sqrt{n}(\hat{\theta} - \theta)/\hat{\sigma}$ by the empirical distribution based on the bootstrap estimators $\sqrt{n}(\hat{\theta}^{*b} - \hat{\theta})/\hat{\sigma}^{*b}$, $b = 1, \dots, B$, then $t_{1-\alpha}$ can be estimated by $t_{1-\alpha}^*$, the 100α th sample percentile of $\sqrt{n}(\hat{\theta}^{*b} - \hat{\theta})/\hat{\sigma}^{*b}$, $b = 1, \dots, B$, and the resulting $1 - \alpha$ bootstrap lower confidence bound for θ is:

$$\hat{\theta} - n^{-1/2} \hat{\sigma} t_{1-\alpha}^* \quad (5)$$

Bootstrap upper confidence bounds and intervals can be similarly obtained.

The bound in Eq. 4 is referred to as the hybrid bootstrap confidence bound and the bound in Eq. 5 is referred to as the bootstrap- t confidence bound. Although the bootstrap- t method is generally better than the hybrid bootstrap method, the bootstrap- t method requires a consistent estimator $\hat{\sigma}$ that can be easily computed.

VALIDITY OF THE BOOTSTRAP

The following are some cases where the validity of the bootstrap has been established in the literature:

1. The parametric bootstrap. In a parametric problem, the parametric bootstrap ("Parametric Bootstrap") is usually valid. Consider the example in ("Parametric Bootstrap" with $\mu = E(X_1)$, $\sigma^2 = \text{Var}(X_1)$, $\hat{\mu} = \bar{X}$ (the sample mean), and $\hat{\sigma}^2 =$ the sample variance. Then, the bootstrap- t confidence bound

$$\bar{X} - n^{-1/2} \hat{\sigma} t_{1-\alpha}^*$$

is an exact $1 - \alpha$ confidence bound for μ (i.e., the probability that μ is larger than this bound is exactly $1 - \alpha$, regardless of the values of the unknown parameter).

2. Nonparametric bootstrap (stratified or longitudinal) for functions of sample means. When the

nonparametric bootstrap ("Nonparametric Bootstrap" and "Longitudinal Bootstrap") is considered, the bootstrap is asymptotically valid when the estimator $\hat{\theta}$ is a smooth function of (several) sample means. Thus in the estimation of poverty line ("Introduction"), the bootstrap is asymptotically valid if bootstrapping pairs ("Bootstrapping Linear Models") are used.

3. Bootstrapping residuals in linear models. For bootstrapping residuals under the linear model (3) ("Bootstrapping Linear Models"), the bootstrap is asymptotically valid when the estimator is a smooth function of the least squares estimator of β .
4. Bootstrapping imputed data. For bootstrapping imputed data ("Bootstrapping Imputed Data"), the bootstrap is asymptotically valid when the estimator is a smooth function of sample means.

Other theoretical results can be found in Hall^[12] and Shao and Tu.^[13]

A theoretical study is needed to show the validity of the bootstrap in situations where no theoretical confirmation has been made. The theoretical work required in establishing the validity of the bootstrap is typically different from the theoretical derivations required by the traditional approach. For example, in the estimation of poverty line ("Introduction"), for the validity of the bootstrap we only need to show that $\hat{\theta}$ is a differentiable function of sample means based on $X_i = (z_i, y_i, \log z_i, \log y_i, x_{i1}, \dots, x_{iT})$, $i = 1, \dots, n$. The exact forms of the partial derivatives of this perhaps very complicated function are not needed in applying the bootstrap, whereas they have to be explicitly derived if the traditional approach (Taylor's expansion method) is applied.

However, establishing the theoretical validity of the bootstrap is sometimes a difficult research problem (e.g., Refs. [12,13]).

There are many results showing that the bootstrap is better than some traditional or nonbootstrap methods such as the normal approximation.^[12] However, whether the bootstrap is better depends on which method the bootstrap is compared to. In setting confidence bounds, for example, the bootstrap- t confidence bounds are shown to be asymptotically more accurate than the confidence bounds obtained by using normal approximation, but are asymptotically equivalent to the confidence bounds obtained by using the one-term Edgeworth or the Cornish-Fisher expansion.^[12] Thus the reason why the bootstrap is used is not because "the bootstrap is better," but because the bootstrap can replace the complicated derivation of the Edgeworth or the Cornish-Fisher expansion by repeated computations.

APPLICATIONS IN BIOEQUIVALENCE TESTING

In vivo bioequivalence testing based on pharmacokinetic responses (such as area under the blood or plasma

concentration–time curve) is considered as a surrogate for a clinical evaluation of the therapeutic equivalence between two drug products, a test drug (T) and a reference drug (R) (e.g., R is a brand name drug and T is a generic copy). In its 1997 draft guidance,^[14] the U.S. Food and Drug Administration (FDA) started to consider the population bioequivalence (PBE) and the individual bioequivalence (IBE). The PBE addresses drug prescribability, which is referred to as the choice for prescribing an appropriate drug for a physician's new patients among the drug products available, while the IBE refers to drug switchability, which is related to the switch from a drug product to an alternative drug product within the same patient. In assessing the PBE and the IBE, the key statistical issue is to set a confidence bound for a function of population means and variance components.

We focus on the IBE. The study design is the following 2×4 crossover design. n subjects are randomly assigned to two sequences (groups) of sizes n_1 and n_2 . In the first sequence, subjects receive treatments T and R at four periods on the order of TRTR, while in the second sequence, subjects receive treatments at four periods on the order of RTRT. Let y_{ijk} be the observed response (or log response) of the i th subject in the k th sequence at j th period, where $i = 1, \dots, n_k$, $j = 1, \dots, 4$, $k = 1, 2$. The following statistical model is assumed:

$$y_{ijk} = \mu + F_l + W_{ijk} + S_{ikl} + e_{ijk} \quad (6)$$

where μ is the overall mean; F_l is the fixed effect of the l th formulation ($l = T, R$ and $F_T + F_R = 0$); W_{ijk} s are fixed period, sequence, and interaction effects ($\sum_k W_{ijk} = 0$, where $\bar{W}_{lk}$ is the average of W_{ijk} s with fixed (l, k) , $l = T, R$; S_{ikl} is the random effect of the i th subject in the k th sequence under formulation l ; and (S_{ikT}, S_{ikR}) , $i = 1, \dots, n_k$, $k = 1, 2$, are iid bivariate normal random vectors with a mean of 0 and an unknown covariance matrix:

$$\begin{pmatrix} \sigma_{BT}^2 & \rho\sigma_{BT}\sigma_{BR} \\ \rho\sigma_{BT}\sigma_{BR} & \sigma_{BR}^2 \end{pmatrix}$$

e_{ijk} s are independent random errors distributed as $N(0, \sigma_{wl}^2)$; and S_{ikl} s and e_{ijk} s are mutually independent.

Define

$$\theta = \frac{\delta^2 + \sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2}{\max\{\sigma_0^2, \sigma_{WR}^2\}}$$

where σ_0^2 is a given constant specified by the FDA, $\delta = F_T - F_R$ and $\sigma_D^2 = \sigma_{BT}^2 + \sigma_{BR}^2 - 2\rho\sigma_{BT}\sigma_{BR}$ is the variance of $S_{ikT} - S_{ikR}$, which is referred to as the variance because of the subject-by-formulation interaction. According to the FDA,^[14,15] two drug formulations can be claimed to be bioequivalent if the hypothesis:

$$H_0 : \theta \geq \theta_0$$

is rejected at the 5% significance level, where θ_0 is a given constant specified by the FDA.

In the FDA,^[14] the bootstrap method is proposed for setting a 95% upper confidence bound $\hat{\theta}_U$ for θ so that we reject H_0 if $\hat{\theta}_U < \theta_0$.

Let $\hat{\delta}$, $\hat{\sigma}_{Bl}^2$ and $\hat{\sigma}_{wl}^2$, $l = T, R$, be estimators of δ , σ_{Bl}^2 , and σ_{wl}^2 , $l = T, R$, respectively. If these estimators are obtained using the restricted maximum likelihood method, then the parametric bootstrap in “Parametric Bootstrap” can be applied to generate bootstrap data. If these estimators are obtained using the method of moments, then the longitudinal nonparametric bootstrap method in “Longitudinal Bootstrap” can be applied to generate bootstrap data.

In the FDA,^[14] the bootstrap percentile method (“Bootstrap Confidence Sets: Percentile Type”) is recommended without any indication of why it is recommended. Since 1997, theoretical and empirical research on the validity of the bootstrap confidence bounds has been carried out.^[16,17] In particular, Shao et al.^[17] showed that the bootstrap method produces asymptotically valid results. For example, when $\hat{\delta}$, $\hat{\sigma}_{Bl}^2$, and $\hat{\sigma}_{wl}^2$ are moment estimators, they are smooth functions of some sample means, in which case the asymptotic validity of the bootstrap has been established (“Validity of the Bootstrap”).

On the other hand, some nonbootstrap methods in assessing IBE have been found.^[18,19] Because the required theoretical derivation is done in Hyslop et al.,^[19] replacing derivations by computations is no longer necessary and thus it is natural for the FDA^[15] to adopt the method of Hyslop et al. (the nonbootstrap approach) for assessing the IBE in its 2001 guidance. This does not mean that the bootstrap is worse than the method of Hyslop et al. When properly used, the bootstrap should be as accurate as, or even better, than the Hyslop et al. method.

Unfortunately, the FDA incorrectly applied the Hyslop et al. method to the PBE problem. One of the key assumption in using the Hyslop et al. method is the independence of the estimated variance components, which is true in the IBE problem but not true in the PBE problem (details can be found in Ref. [20]). On the other hand, the application of the bootstrap does not require this assumption. This shows another advantage of using the bootstrap over a nonbootstrap method that requires theoretical formulas. Note that theoretical formulas are frequently derived under some assumptions that vary from problem to problem. If one does not carefully check these assumptions, one may incorrectly use some formulas. The bootstrap, however, is frequently insensitive to the violation of assumptions because the bootstrap does not directly use these theoretical formulas. Of course, the bootstrap may also be misused. But the example of IBE and PBE shows that using the bootstrap may avoid the kind of mistake made by the FDA.

REFERENCES

1. Mantel, H.J.; Singh, A.C. Standard errors of estimates of low income proportions: A proposed methodology. technical report, Statistics Canada **1991**.
2. Shao, J.; Steel, P. Variance estimation for imputed survey data with non-negligible sampling fractions. *J. Am. Stat. Assoc.* **1999**, *94*, 254–265.
3. Shao, J.; Wang, H. Sample correlation coefficients based on survey data under regression imputation. *J. Am. Stat. Assoc.* **2002**, *97* (458), 544–552.
4. Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **1979**, *7*, 1–26.
5. Gross, S. Median Estimation in Sample Surveys. In *Proceedings of the Section on Survey Research Methods*; American Statistical Association: Alexandria, VA, 1980, 81–184.
6. Chao, M.T.; Lo, S.H. A bootstrap method for finite populations. *Sankya*, **1985**, *47*, 399–405.
7. Sitter, R.R. A resampling procedure for complex survey data. *J. Am. Stat. Assoc.* **1992**, *87*, 755–765.
8. Efron, B. Missing data, imputation, and the bootstrap. *J. Am. Stat. Assoc.* **1994**, *89*, 463–479.
9. Shao, J.; Sitter, R.R. Bootstrap for imputed survey data. *J. Am. Stat. Assoc.* **1996**, *91*, 1278–1288.
10. Efron, B. Nonparametric standard errors and confidence intervals (with discussions). *J. Can. Stat.* **1981**, *9*, 139–172.
11. Efron, B. Better bootstrap confidence intervals (with discussions). *J. Am. Stat. Assoc.* **1987**, *82*, 171–200.
12. Hall, P. *The Bootstrap and Edgeworth Expansion*; Springer-Verlag: New York, 1992.
13. Shao, J.; Tu, D. *The Jackknife and Bootstrap*; Springer-Verlag: New York, 1995.
14. FDA. *In Vivo Bioequivalence Studies Based on Population and Individual Bioequivalence Approaches*; Division of Bioequivalence, Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, MD, 1997.
15. FDA. *Guidance for Industry on Statistical Approaches to Establishing Bioequivalence*; Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, MD, 2001.
16. Shao, J.; Chow, S.; Wang, B. The bootstrap procedure for individual bioequivalence. *Stat. Med.* **2000**, *19*, 2741–2754.
17. Shao, J.; Kübler, J.; Pigeot, I. Consistency of the bootstrap procedure in individual bioequivalence. *Biometrika*. **2000**, *87*, 573–585.
18. Wang, W. On testing of individual bioequivalence. *J. Am. Stat. Assoc.* **1999**, *94*, 880–887.
19. Hyslop, T.; Hsuan, F.; Holder, D.J. A small sample confidence interval approach to assess individual bioequivalence. *Stat. Med.* **2000**, *19*, 2885–2897.
20. Chow, S.; Shao, J.; Wang, H. Statistical tests for population bioequivalence. *Stat. Sin.* **2002**, 539–554.

Botanical Drug Product Development: Scientific Issues

Annpey Pong

Merck Research Laboratories, Rahway, New Jersey, U.S.A.

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

A botanical drug product is often recognized as a traditional Chinese (herbal) medicine (TCM). The use of TCM in humans for treating various diseases has a history of a few thousand years, although not much convincing scientific evidence (documentations) regarding clinical safety and efficacy are available. Thus, how to effectively and scientifically develop a promising TCM the Western way has become an important issue in public health. The purpose of this entry is to provide an overview of scientific and/or regulatory issues that are commonly encountered in botanical drug product or TCM development. These issues include, but are not limited to, intellectual property (IP), variation (or consistency) in raw materials, component-to-component interactions, animal studies, matching placebo and calibration of study endpoints in clinical trials, packaging insert, and transition from experience-base to evidence-base clinical practice.

Keywords: Traditional Chinese herbal medicine; Dynamic balance or harmony; Utilization ratio of the extracts (URE); Experience-base clinical practice; Matching placebo.

INTRODUCTION

In recent years, as more and more innovative drug products are going off patents, the search for new or alternative medicines such as botanical drug products that can treat critical and/or life-threatening diseases has become the center of attention of many pharmaceutical companies and research organizations such as the United States National Institutes of Health (NIH). Botanical drug products are often referred to as traditional Chinese herbal medicine (TCM). This leads to the development of TCM, especially for those intended for treating critical and/or life-threatening diseases such as cancer. The use of TCM in humans for treating various diseases has a history of a few thousand years, although not much convincing scientific evidence (documentations) regarding clinical safety and efficacy are available. Thus, how to effectively and scientifically develop a promising TCM the Western way has become an important issue in public health.

A Western medicine (WM) (e.g., a small molecular chemical drug product) often contains a single active ingredient, while a TCM (e.g., a botanical drug product) usually consists of *multiple* active and inactive components. These multiple active and inactive components may not be characterized and their relationships are usually unknown. Thus, in practice, it is of great concern that whether a TCM can be scientifically evaluated the Western way due to some fundamental differences between a WM and a TCM.

The purposes of this entry are: (i) provide a comparison between a WM and a traditional TCM in terms of the fundamental differences; (ii) provide an overview of regulatory requirements for botanical drug product development based on a guidance published in 2004,^[1] (iii) discuss critical scientific and/or regulatory issues that are commonly encountered during the development of a botanical drug product or TCM.

FUNDAMENTAL DIFFERENCES

The process for pharmaceutical/clinical research and development of a chemical drug product or WM is well established, and yet it is a lengthy and costly process.^[2] This lengthy and costly process is necessary to ensure the efficacy, safety, quality, stability and reproducibility of the drug product under investigation. For the development of a botanical drug product (or TCM), one may consider directly applying this well-established process. However, this process may not be feasible due to some fundamental differences.^[3] These fundamental differences are summarized in [Table 1](#), and briefly described below.

Medical Theory and Mechanism

TCM has a long history of holistic medical system encircling the entire scope of human experience. It combines the

Table 1 Fundamental Differences between a WM and a TCM

Description	Western Medicine	Traditional Chinese Herbal Medicine
Active ingredient	Single	Multiple
Dose	Fixed	Flexible
Diagnostic procedure	Objective; validated	Subjective; not validated
Therapeutic index	Well-established	Not well-established
Medical mechanism	Specific organs	Global dynamic balance/harmony among organs
Medical perception	Evidence-base	Experience-base
Statistics	Population	Individual

use of Chinese herbal medicines, acupuncture, massage, and therapeutic exercises such as *Qi-gong* (the practice of internal *air*) and *Tai-gie* for both treatment and prevention of disease.

The elements of TCM can also help to describe the etiology of disease including six exogenous factors (i.e., wind, cold, summer, dampness, dryness, and fire), seven emotional factors (i.e., anger, joy, worry, grief, anxiety, fear, and fright), and other pathogenic factors. Once all of the information are collected and processed into a logical and workable diagnosis, the traditional Chinese medical doctor can determine the treatment approach.

Under the medical theory and mechanism described above, Chinese doctors believe that all of the organs within a healthy subject should reach the so-called *global dynamic balance or harmony* among organs. An experienced Chinese doctor usually assesses the causes of global imbalance before a TCM with flexible doses is prescribed to fix the problem. This approach is sometimes referred to as a personalized (or individualized) medicine approach.

Medical Practice

Different medical perceptions regarding signs and symptoms of certain diseases could lead to a different diagnosis and treatment for the diseases under study. For example, the signs and symptoms of type 2 diabetic subjects could be classified as the disease of *thirsty reduction* by Chinese doctors. The disease of type 2 diabetes is not recognized by Chinese medical literature although they have the same signs and symptoms as the well-known disease of thirsty reduction. This difference in medical perception and practice has an impact on the diagnosis and treatment of the disease.

In addition, therapeutic effect of WMs is usually sooner than TCMs. TCMs are often considered for patients who have chronic diseases or non-life-threatening diseases. For critical and/or life-threatening diseases such as cancer or stroke, TCMs are often used as the second line or third line treatment with no other alternative treatments. In many cases such as patients with later phase of cancer, TCMs are often used in conjunction with WMs without the knowledge of the primary care physicians.

Techniques of Diagnosis

The Chinese diagnostic procedure for patients with certain diseases consists of four major techniques, namely, inspection, auscultation and olfaction, interrogation, and pulse taking and palpation. All these diagnostic techniques aim mainly at providing an objective basis for differentiation of syndromes by collecting symptoms and signs from the patient.

The Chinese diagnostic procedure is subjective, with large between rater variability. This subjectivity and variability will have an impact not only on the patient’s evaluability but also the prescribability of TCM. For evaluation of a WM, objective criteria based on some well-established clinical study endpoints are usually considered.

In clinical trials, evaluation is usually based on some validated tools such as laboratory tests. Test results are then evaluated against some normal ranges for abnormality. Thus, it is suggested that the Chinese diagnostic procedure must be validated in terms of validity and reliability before it can be used for evaluation of clinical efficacy and safety of the TCM under investigation.

Treatment

TCM prescriptions typically consist of a combination of several components. The combination is usually determined based on the medical theory of global dynamic balance (or harmony) among organs, and the observations from the Chinese diagnostic procedure. The dose and treatment duration are usually flexible. This concept leads to the concept of so-called personalized (or individualized) medicine, which minimizes intra-subject variability. Most WMs contain a single active ingredient. After drug discovery, an appropriate formulation is necessarily developed so that the drug can be delivered to the site action in an efficient way. At the same time, an assay is necessarily developed to quantitate the potency of the drug. The drug is then tested on animals for toxicity, and humans for pharmacological activities. Unlike the WMs, TCMs usually consist of multiple components with certain relative proportions among the components. As a result, the typical

approach for evaluation of single active ingredient for WM is not applicable.

In practice, one may suggest evaluating the TCM component by component. However, this is not feasible due to the following difficulties. First, in practice, analytical methods for quantitation of individual components are often not tractable. Thus, the pharmacological activities of these components are not known. In practice, it is not known which relative proportions among these components can lead to the optimal therapeutic effect of the TCM. In addition, the relative component-to-component and/or component by food interactions are usually unknown, which may have an impact on the evaluation of clinical efficacy and safety of the TCM.

Fixed Dose versus Flexible Dose

Most WMs are usually administered in a fixed dose. On the other hand, a TCM is often prescribed as an individualized flexible dose.

The approach of WM with a fixed dose is a population approach to minimize the between subject (or inter-subject) variability, while the approach to TCM with an individualized flexible dose is to minimize the variability within each individual. In practice, it is a concern whether an individual flexible dose is compatible with a Western evaluation of the TCM. An individualized flexible dose depends heavily upon the subjective judgment, which may vary from one doctor to another. As a result, although an individualized flexible dose does minimize intra-subject variability, the variability from one doctor to another could be huge, and hence non-negligible.

REGULATORY REQUIREMENTS

In 2004, the United States (US) Food and Drug Administration (FDA) published a guidance on botanical drug products, which explains when a botanical drug may be marketed under an over-the-counter drug monograph, when an FDA regulations require approval for marketing of New Drug Application (NDA), and when Investigational New Drug Applications (INDs) for botanical products currently lawfully marketed as foods in US.^[1] In addition, the CDER of FDA also published Manual of Policies and Procedures (MAPP): Review of Botanical Drug Products, which describes the review process in CDER for INDs and NDAs for botanical drug products.^[4] The 2004 MAPP recommends that the guidance entitled *Guidance for Industry – INDs for Phase 2 and Phase 3 Studies, Chemistry, Manufacturing, and Controls Information* be consulted for preparing chemistry, manufacturing, and control (CMC) information that would be submitted for phase 2 and phase 3 studies required for botanical drug product development conducted under INDs in US.^[5]

In FDA guidance 2003, the INDs for botanical drug products should include protocols, CMC, pharmacological and toxicology information, and previous human experience with the product (see Table 2). Requirements for CMC and non-clinical safety assessment are given in Table 3 and Table 4, respectively. Note that CMC requires animal safety test which referred to an acute animal toxicity test applied only to injectable drug product. For matching placebo, the FDA requires that the components of any placebo used must be described. For phase 3 clinical studies, the FDA requires that the following information that (i) description of product and documentation of human experience, (ii) chemistry, manufacturing, and controls, (iii) non-clinical safety assessment, (iv) bioavailability and clinical pharmacology, and (v) clinical considerations must be provided.

Current Review Process for Botanical Products

Regulatory review of botanical submission include CMC information review, clinical pharmacology/biopharmaceutics information review, non-clinical pharmacology and toxicology

Table 2 Basic Format for IND for Botanical Drug Products

Cover sheet
Table of contents
Introductory statement and general investigational plan
Investigator's brochure
Protocols
Chemistry, manufacturing, and controls
Pharmacological and toxicology information
Previous human experience with the product

Table 3 CMC Requirements for INDs of Botanical Drug Products

Botanical raw material
Botanical drug substance and product
Animal safety test
Placebo
Labeling
Environmental assessment

Table 4 Requirements for Non-clinical Safety Assessment

Repeat-dose general toxicity studies
Nonclinical pharmacokinetic/toxicokinetic studies
Reproductive toxicology
Genotoxicity studies
Carcinogenicity studies
Special pharmacology/toxicology studies
Regulatory consideration

information review, and medical/statistical information review. For each botanical submission, FDA establishes a botanical review team (BRT) to assist the review divisions for review of the botanical submission. The BTR review covers (i) biology of the medicinal plants-identification, potential misuse of related species, (ii) pharmacology of the botanical product-activity/toxicology in old documents and new tests, (iii) prior human experiences with the botanical product-past clinical use and relevance to current setting. The botanical team will perform pharmacognosy review throughout the IND and NDA process and botanical-specific medical review. Although regulatory requirements for botanical products are similar to those for chemical drugs in principle, there are many differences in policy issues, which are summarized below:

Purification and Identification - Botanical drugs are derived from vegetable matter and are usually prepared as complex mixtures. Their chemical constituents are not always well defined. In many cases, the active constituent in a botanical drug is not identified, nor is its biological activity well characterized. A new botanical drug (containing multiple chemical constituents) may qualify as a new chemical entity. Both purification and identification of the active ingredients in botanicals are optional and not required. In the initial stage of clinical studies of a botanical drug, it is generally not necessary to identify the active constituents or other biological markers or to have a chemical identification and assay for a particular constituent or marker. Identification by spectroscopic and/or chromatographic fingerprinting and strength by dry weight (weight minus water or solvents) can be acceptable alternatives.

Test and Control - Because of the complex nature of a typical botanical drug and the lack of knowledge of its active constituent(s), FDA may rely on a combination of tests and controls to ensure the identity, purity, quality, strength, potency, and consistency of botanical drugs. These tests and controls include (i) multiple tests for drug substance and drug product (e.g., spectroscopic and/or chromatographic fingerprints, chemical assay of characteristic markers, and biological assay), (ii) raw material and process controls (e.g., strict quality controls for the botanical raw materials and adequate in-process controls), and (iii) process validation (especially for the drug substance).

Individualized Treatments - Botanical therapies are highly individualized with variations in relative contents of multiple plant ingredients tailored for each patient. A sponsor may not submit a separate IND for every change in composition, if similar patients are being treated for the same indication. Studies can be designed to take into account individualized treatments. Multiple formulations can be included in one IND if they are being studied under a single clinical trial. It is important that the IND provide the rationale for using multiple formulations and the criteria used to assign patients to different treatment regimens.

Toxicity - Many medicinal plants with therapeutical potential are quite toxic. Well-known examples of safety issues concerning botanicals include the nephrotoxicity associated with herbal preparations containing aristolochic acid and the hepatotoxicity associated with comfrey products containing pyrrolizidine alkaloid. Other examples include the cardiovascular and central nervous system effects associated with yohimbe and the hepatotoxicity associated with germander and chapparal. When the potential benefit of an investigational drug outweighs its risk in the intended patient population, clinical trials may be allowed to proceed under an IND.

Prior Human Experience - The Guidance also stipulates that because many botanicals have been used as medicine in alternative medical systems for long time, the prior human experiences may substitute for animal toxicology studies in the preliminary safety evaluation of IND studies. The sponsor is encouraged to provide as much data as possible, and the review team for the botanical drug IND generally will accept all available information for regulatory consideration.

Priority - FDA assigns the same level of priority to botanical drug products as to other drugs with respect to meeting with IND sponsors and NDA applicants. For clinical data to support marketing approval, there should be no difference between botanical and non-botanical drugs. The Guidance also provides two flow charts for: 1) the regulatory approaches for marketing botanical drug products in [Fig. 1](#) (attachment B, FDA 2004), and 2) the information to be provided in an IND for a botanical drug in [Fig. 2](#) (attachment A, FDA, 2004).

SCIENTIFIC ISSUES IN BOTANICAL DRUG (OR TCM) DEVELOPMENT

The critical issues that are often encountered during the development of a botanical drug product (or TCM) are briefly summarized below.

Intellectual Property (IP)

Song (2011) indicated that innovation of TCM can be divided into two categories: one is the *self-innovation* of TCM and the other is innovation based on knowledge and techniques of TCM.^[6] Both innovations have been presented with different questions of IP derived from past events. As Song pointed out, under market economy conditions where private rights are legitimate and interests must be maximized, a proper intellectual property legal system should be established for the protection of the innovation of TCM. However, majority of self-innovation of TCM generally do not seek for protection of IP due to the nature of conservativeness of Chinese people and most importantly lack of confidence in current patent system for IP protection.

Attachment B: Information to be provided in an IND for a botanical drug

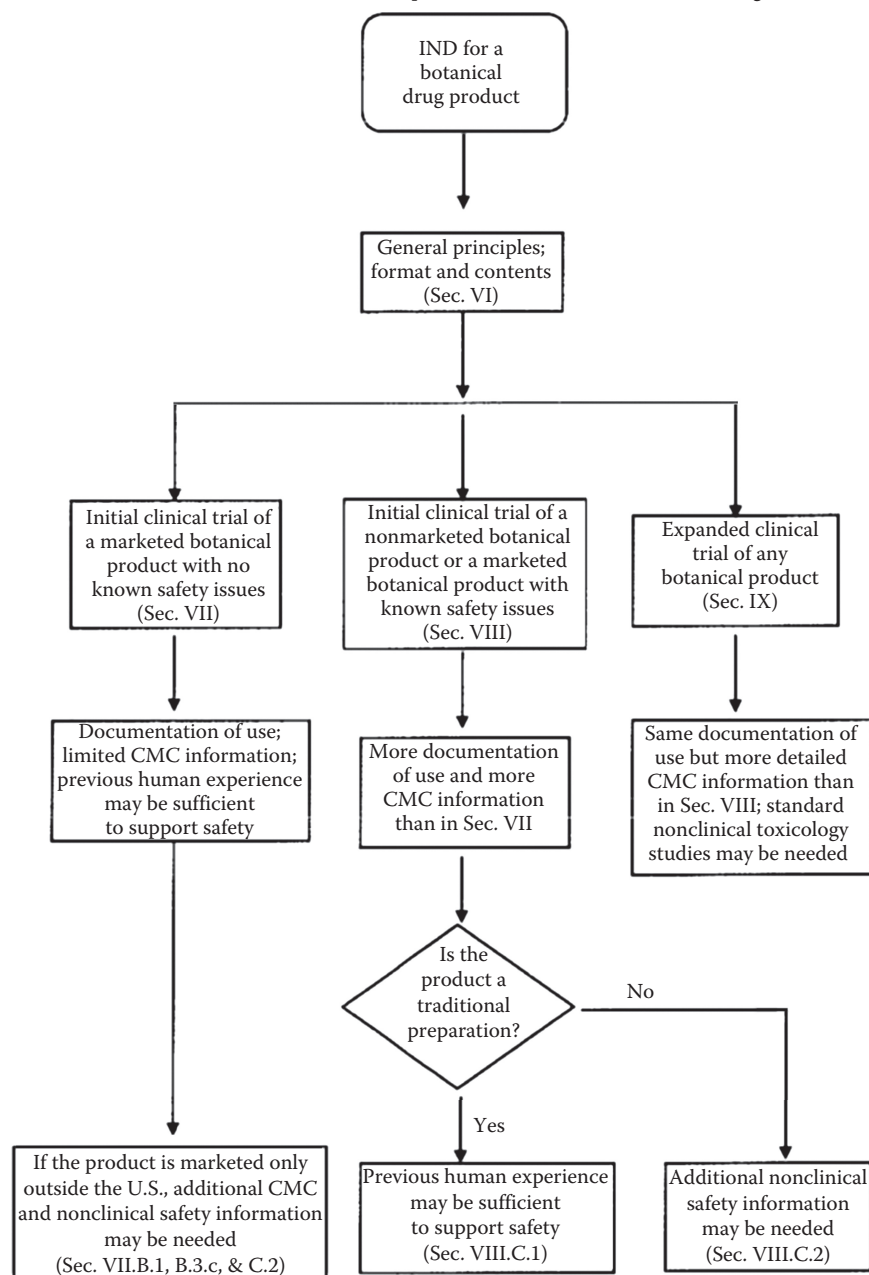


Fig. 1 Flow chart for approval pathway of an IND for a botanical drug product

Hsiao (2007) pointed out that in the current patent system, an invention has to satisfy the examiner in many aspects including its novelty, inventive step, industrial applicability and enablement.^[7] An invention has to pass the novelty test before proceeding to other steps. Failure to satisfy the novelty requirement will render the invention non-patentable. In the past several decades, novelty has barred many Chinese herbal medicines from patentability because they are either based on traditional formulas or have the same medicinal use and are already available to the public.

As indicated by the Taiwan Intellectual Property Office (TIPO) Guidelines for Patent Examination, an invention is

referred to as any creation of technical concepts by utilizing the rules of nature. TIPO adopts the standard of absolute novelty in the sense that an invention has to be new and nothing similar to those that have appeared before the date of filing. TIPO has established a database of classics and traditional formulas that are in the public domain. Along this line, there are several possible types of end products of traditional Chinese medicine deriving from traditional knowledge, methods of treatments, products and processes are patentable.^[8] Despite the controversies of patenting traditional medicines, the patenting of products and processes of Chinese herbal inventions is not a novel idea.

Attachment A: Regulatory approaches for marketing botanical drug products

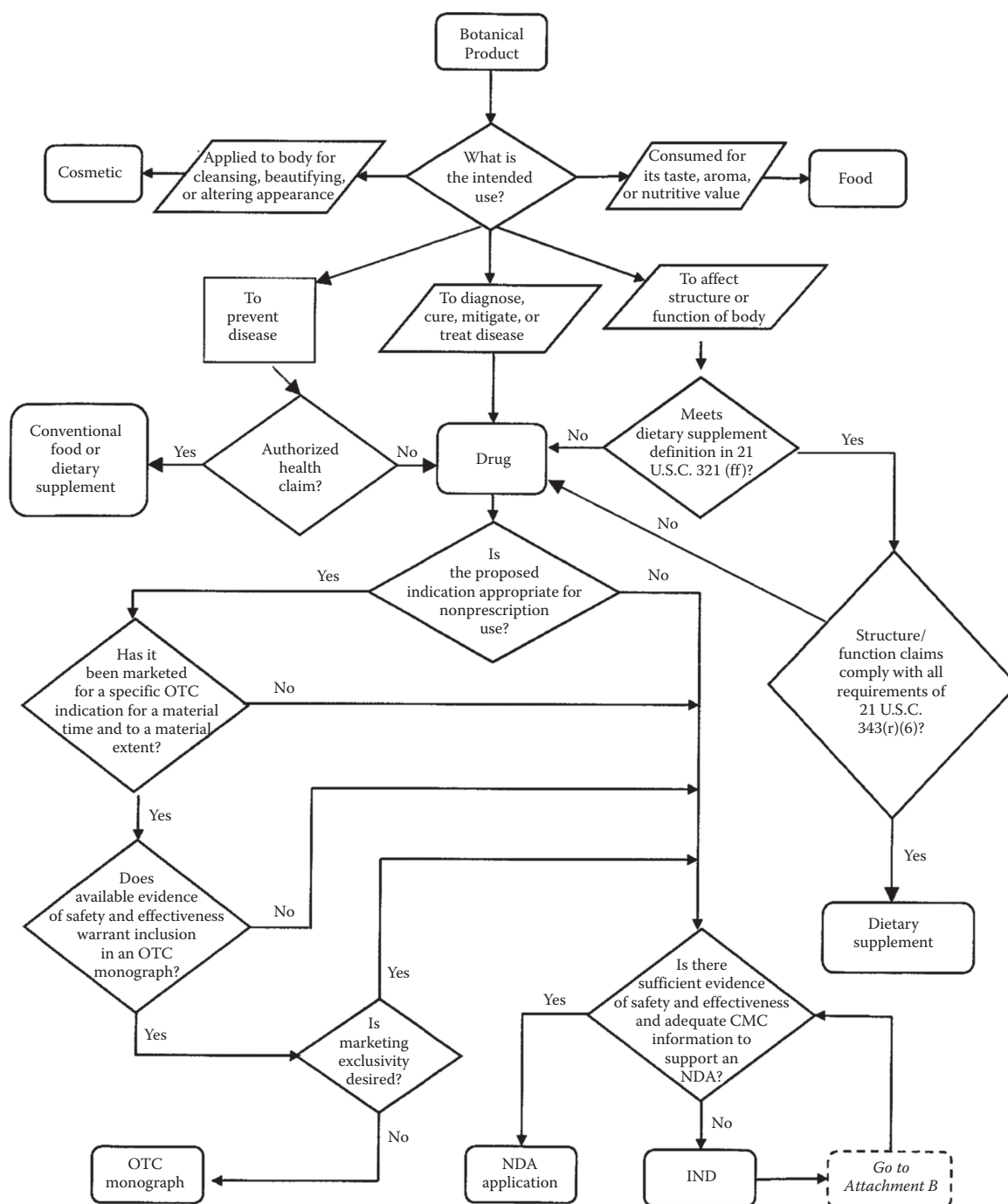


Fig. 2 Flow chart for regulatory approaches for marketing botanical drug products

Variation (Consistency) in Raw Materials

One of the critical issues in TCM manufacturing is variation in raw materials. The variation in raw materials, which may be due to the fact that they come from different regions, climates, and time of harvest, could have a negative impact on the quality of the TCM and consequently

the safety and efficacy of the TCM. Yan and Qu (2013) indicated that the efficacy of a TCM depends upon the combined effects of its components.^[9] Variation in chemical composition between batches of TCM has been always the deterring factor of achieving consistency in efficacy. These components, however, may or may not be correlated and may have component-to-component interaction which

will have an impact on achieving optimal therapeutic effect of the TCM. In practice, unfortunately, the correlations among the components and their relative ratios for achieving optimal therapeutic effect are often unknown.

Yan and Qu (2013) indicated that batch mixing process can significantly reduce the batch-to-batch variation in TCM extracts by mixing them in a *well-designed* proportion, which is referred to as utilization ratio of the extracts (URE).

They proposed a batch mixing optimization model to reduce variation in raw materials and improve quality of the TCM under development. Their method is briefly described below. Suppose that there are t (target) components with controlled contents in the mixture. Let s be the number of storage tanks, i.e., the maximum number of batches of extracts that can be stored. Also let $x = (x_1, x_2, \dots, x_s)$ be the amounts of each stored batch of extract used for mixing and $u = (u_1, u_2, \dots, u_s)$, the amounts of each batch of extract stored. Define $a_k = (a_{1k}, a_{2k}, \dots, a_{tk})'$, the contents of the constituents in the k th batch of extract stored. For unidentified constituents, test values such as peak areas on the chromatographic fingerprint obtained by certain analytical methods can be used. Thus, we have $A = (a_1, a_2, \dots, a_s)$, a $t \times s$ matrix that consists of the contents of the components in the batches of extracts stored. Now, let b be the amount of mixture needed in each batch and (L_i, U_i) be the minimum and maximum content limits of the i th component. Yan and Qu (2013) proposed to maximize

$$\max \sum_{i=1}^s x_i^2, \quad (1)$$

which maximizes the sum of squares of the used amounts of each batch. Therefore, the optimization model in this study is to maximize the value of (1) under the following constraints

$$\sum_{i=1}^s x_i = b, \quad (2)$$

where (2) determines the amount of mixture needed for the follow-up process. The contents of the constituents in the mixture are assumed to be the weighted average of the contents in the extracts, where the weights are the amounts of each batch used for mixing;

$$\min L' \leq \frac{A \cdot x'}{b} \leq \max U', \quad (3)$$

where $\max U = (\max U_1, \max U_2, \dots, \max U_t)$, the maximum content limits for the components in the mixtures and $\min L = (\min L_1, \min L_2, \dots, \min L_t)$, the minimum content limits for the constituents in the mixtures. (3) ensures that the contents are in the range of their limits, and,

$$0 \leq x' \leq u', \quad (4)$$

where (4) is the maximum amount of each batch that can be used for mixing. Note that the above optimization model can be solved by quadratic programming (QP) algorithm. In TCM manufacturing, batch mixing can be conducted with the extracts obtained from previous processes like decoction, concentration, and purification. The extracts created are first stored in the storage tanks. Batch mixing ratios are calculated according to the optimization model and mixtures are created for the follow-up process.

Component-to-Component Interactions

Unlike most Western medicines, TCM often consists of a number of components whose pharmacological activities may or may not be quantitatively characterized. Besides, the relative proportions (ratios) of the components for achieving optimal therapeutic effect are also unknown in addition to possible component-to-component (drug-to-drug) interaction among these components. Thus, the study of determining the relative ratios of components for achieving optimal therapeutic effect has become the key to the success of the TCM development. To study the main effects and interactions of these components, some commonly employed designs are briefly described below.

Full factorial design – A full factorial design is a design that consists of all possible different combinations of one level from each factor. If there are l_k levels for the k th factor X_k , the corresponding full factorial design is called a general $l_1, l_2, \dots, l_K$ factorial design. In practice, a factorial design is expressed in terms of a number of arrays (or runs) that indicate the levels of each factor. For example, for a typical 2^4 factorial design, the arrangement of the arrays is given in the following standard order (see Table 5). The first column of the design matrix consists of successive minus (–) and plus (+) signs, the second column of successive pairs of (–) and (+) signs, the third column of four (–) signs followed by four (+) signs, and so on. In this 2^4 factorial design, there are four factors at two levels, with a total of $N = 2^4 = 16$ runs. The two levels of each factor are conventionally denoted by (+) and (–). A full factorial design provides estimates not only for main effects but also for interactions with maximum precision. The main effects and interaction effects can easily be obtained using a table of contrast coefficients and/or Yate's algorithm.^[10,11]

Fractional factorial design – A fractional factorial design is a design that consists of a fraction of a full factorial experiment. For example, a $\left(\frac{1}{2}\right)^p$ fraction of a 2^K factorial design is called a 2^{K-p} fractional factorial design. For a 2^{4-1} fractional factorial design, the eight observations cannot provide independent estimates for the 16 effects alone (from the full 2^4) but for some confounding effects, such as the sum of a main effect and a three-factor interaction that are confounded with each other. In practice, however, the three-factor or higher-factor interactions are usually negligible (see Table 6). In this case, a fractional factorial

Table 5 A Full 2^4 Factorial Design

Run	Design matrix				Y
	X_1	X_2	X_3	X_4	
1	–	–	–	–	Y_1
2	+	–	–	–	Y_2
3	–	+	–	–	Y_3
4	+	+	–	–	Y_4
5	–	–	+	–	Y_5
6	+	–	+	–	Y_6
7	–	+	+	–	Y_7
8	+	+	+	–	Y_8
9	–	–	–	+	Y_9
10	+	–	–	+	Y_{10}
11	–	+	–	+	Y_{11}
12	+	+	–	+	Y_{12}
13	–	–	+	+	Y_{13}
14	+	–	+	+	Y_{14}
15	–	+	+	+	Y_{15}
16	+	+	+	+	Y_{16}

Table 6 A 2^{4-1} Fractional Factorial Design

Run	Design matrix				Y
	X_1	X_2	X_3	$X_4 = X_1X_2X_3$	
1	–	–	–	–	Y_1
2	+	–	–	+	Y_2
3	–	+	–	+	Y_3
4	+	+	–	–	Y_4
5	–	–	+	+	Y_5
6	+	–	+	–	Y_6
7	–	+	+	–	Y_7
8	+	+	+	+	Y_8

design is useful in estimating the main effects. In practice, a fractional factorial design is useful when there are many factors to be studied because it is almost impossible to perform a full factorial design even at two levels.

Central composite design – A central composite design is a full factorial design or a fractional factorial design augmented by a $\pm\alpha$ level at each of the K factors and n central points. The central composite design consists of one center point, eight points on the cube (a 2^3 factorial arrangement), and six star points. It should be noted that a central composite design with $K = 2$, $\alpha = 1$, and $n = 1$ reduces to a 3^2 factorial design (see Table 7). For a full 2^K factorial design, although the design provides independent estimates for the $2^K - 1$ effects, it does not give an estimate of the experimental error unless some runs are repeated. Unlike the full 2^K factorial design, the central composite design provides an estimate

Table 7 Central Composite Design for $K = 3$ and $n = 1$

Run	X_1	X_2	X_3
1	–1	–1	–1
2	1	–1	–1
3	–1	1	–1
4	1	1	–1
5	–1	–1	1
6	1	–1	1
7	–1	1	1
8	1	1	1
9	0	0	0
10	α	0	0
11	$-\alpha$	0	0
12	0	α	0
13	0	$-\alpha$	0
14	0	0	α
15	0	0	$-\alpha$

of the experimental error. The experimental error is usually estimated based on n observations at the central point.

Animal Studies

Since TCM often consists of multiple components, it is unquestionable that many components are successful in suppressing different types of diseases in humans. However there does not appear to be any scientific documentations, which meets stringent Western criteria for safety and efficacy, to support their use in humans. Thus, in the pharmaceutical development, animal studies such as mice, rats, rabbits, dogs, or monkeys for testing toxicity and evaluation of efficacy are required for regulatory review and approval before they can be used in humans.

However, there are tremendous debates regarding whether animal studies are necessary for the development of TCM, especially for those have been used in for thousands of years. Most Chinese doctors suggest that it would be better to focus on clinical development, instead of going back to animal studies to obtain evidence for a precise pharmacological profile. It would be better to focus on safety and efficacy by starting with clinical trials. It is also suggested that studies on mechanisms and active ingredients should be conducted after safety and efficacy have been confirmed.

For development of Western medicines, animal studies (models) are often used to screen experimental drugs before they enter human trials. This is animal models can provide pharmaceutical scientists the insight how do the drugs work in the living system and evaluate the toxicity at various doses assuming that animal model is predictive of human model. Although there have been controversial debates over which diseased animal models should

be used to screen new drugs before they can proceed to the clinical trials, there are certain animal models in different therapeutic areas remain good standards for drug screening used in the pharmaceutical industry. One of the concerns is that apparent physiological and genetic differences between human and animals, which clearly indicate that the diseased models are different from the human equivalent.

Matching Placebo in Clinical Trials

In recent years, randomized controlled trials (RCT) have been recognized as the gold standard in clinical trials for evaluation of safety and efficacy of a test compound under investigation. One of the important components in RCTs is blinding. The ultimate goal of clinical trials is to achieve a double-blind design to avoid any possible operational biases due to the knowledge of the treatment assignment. Qi et al. (2008) conducted a comprehensive review on the validity of matching placebo used in blinded clinical trials for Chinese herbal medicine in recent years and related patents^[12]. The review was conducted based on a database called the *Wanfang Database*, which contains a total of 827 Chinese journals of medicine and/or pharmacy, from 1999 to 2005 and 598 full-length articles related to clinical trials utilizing matching placebo. Qi et al. (2008) concluded the review that researchers in Chinese medicine commonly ignored the quality of the matching placebo in comparison to the test drug. This may have led to substantial bias in clinical trials. As a result, Qi et al. urged that quality specifications and evaluation of the matching placebo must be developed and carefully evaluated in order to reduce possible bias in randomized controlled TCM trials.

Fai et al. also pointed out that in many clinical trials of Chinese herbal medicines, it is very difficult to make a quality matching placebo to achieve the purpose of blinding.^[13] Ideally, the characteristics of the test drug and matching placebo should be identical in color, appearance, smell and taste. The quality matching placebo should be identical to the test drug in physical form, sensory perception, packaging, and labeling, and it should have no pharmaceutical activity. For this purpose, Fai et al. developed a placebo capsule to match a herbal medicine in terms of its physical form, chemical nature, appearance, packaging and labeling. Based on the assessment results, the developed placebo capsule assessment results suggested that the placebo was found satisfactory in these aspects. Thus, it concluded that a matching placebo could be created for a RCT involving herbal medicine.

It should be noted that the preparation of matching placebo is extremely important to maintain the integrity of blinding for avoid any possible operational biases that may be introduced due to the knowledge of the treatment assignment. The oral dosage form of capsule is often considered for preparation of matching placebo for clinical

trials involving Chinese herbal medicines as it may remove the strong smell and taste of the herbal medicines. However, one of the major challenges is that patients or clinicians will reveal the treatment assignments if they break the capsules. Thus, standard operating procedures (SOP) for preventing patients and clinicians from breaking the capsules are necessary developed.

Calibration of Study Endpoints

The primary study endpoints for assessment of safety and effectiveness of a TCM are usually assessed by a quantitative instrument or the four diagnostic procedures by experienced Chinese doctors. The assessment by a quantitative instrument has been criticized in many ways. First, it may not capture the true health of status of the patients with diseases under study (e.g., by asking wrong questions). Second, it may not detect the effect of the test treatment under investigation. The assessment based on a quantitative instrument by experienced Chinese doctors is not only subjective, but also lack of validity. Consequently, the reliability of the assessment is a concern, especially when there is evidence of large rater-to-rater variability.

Thus, although the quantitative instrument is developed by the community of Chinese doctors and is considered a gold standard for assessment of safety and effectiveness of the TCM under investigation, it may not be accepted by the Western clinicians not only due to the lack of validity and reliability, but also the interpretation of the assessment. Thus, for modernization or Westernization of TCMs, whether the subjective quantitative instrument can accurately and reliably assess the safety and effectiveness of the TCM is a concern for development of TCM. In practice, it is then suggested that a clinical trial be conducted to calibrate the subjective quantitative assessment against either life events or well-established clinical endpoints that are commonly used in assessment of Western medicines. The clinical trials should consist of two arms: one arm will include subjects with diseases under study diagnosed by the subjective quantitative instrument and the other arm will include subjects diagnosed by Western diagnostic or testing procedures. Each subject post-treatment will be assessed by both Chinese doctors using the quantitative instrument and Western clinicians based on the well-established and widely accepted study endpoints.^[14]

Package Insert

For the research and development of a TCM, before a TCM clinical trial is conducted, the following questions are commonly asked.

1. Will the TCM clinical trial be conducted by Chinese doctors alone, Western clinicians alone, Western clinicians who have some background of Chinese

herbal medicine alone, or both Chinese doctors and Western clinicians?

2. Will traditional Chinese diagnostic and/or trial procedures be used throughout the TCM clinical trial?
3. Upon approval, is the TCM intended for use by Chinese doctors or Western clinicians?

With respect to the first two questions, if the TCM clinical trial is to be conducted by Chinese doctors alone, the following questions arise. First, should the Chinese diagnostic procedure be validated in order to provide an accurate and reliable assessment of the TCM? In addition, it is of interest to determine how an observed difference obtained from the Chinese diagnostic procedure can be translated to the clinical endpoint commonly used in similar WM clinical trials with the same indication. These two questions can be addressed statistically by the calibration and validation of the Chinese diagnostic procedure with respect to some well-established clinical endpoints for evaluation of Western medicines. If the TCM clinical trial is to be conducted by Western clinicians or Western clinicians who have some background of Chinese herbal medicine, the standards and consistency of clinical results as compared to those WM clinical trials are ensured. However, the good characteristics of TCM may be lost during the process of the conduct of the TCM clinical trials. On the other hand, if the TCM clinical trial is to be conducted by both Chinese doctors and Western clinicians, difference in medical practice and/or possible disagreement regarding the diagnosis, treatment, and evaluation are major concerns.

For the third question, if the TCM is intended for use of Chinese doctors but it is conducted by Western clinicians, difference in perception regarding how to prescribe the TCM is of great concern. The preparation of a package insert based on the clinical data could be a major issue, not only to the sponsor but also to regulatory authorities. Similar comments apply to the situation where the TCM is intended for use of Western clinicians, but the trial is conducted by Chinese doctors.

As a result, it is suggested that the intention of use (i.e., labeling for the indication) be clearly evaluated when planning a TCM clinical trial. In other words, the sponsor needs to determine whether the TCM is intended for use of Western clinicians only, Chinese doctors only, or both Western clinicians and Chinese doctors at the planning stage of a TCM clinical trial, for an adequate package insert of the target diseases under study.

Transition from Experience-Base to Evidence-Base Clinical Practice

TCMs have been in practice for thousands of years. Many of the commonly used TCMs were found safe and efficacious. However, evidence of safety and efficacy of these TCM were not documented for scientific evaluation. Unlike evidence-based clinical data, the experience-based

clinical information has been criticized for lacking of scientific validity and reliability for assessment of safety and efficacy of the TCMs. As a result, experience-based clinical information is considered not only subjective, but also biased and misleading.

In practice, how to collect relevant and important clinical data from experience-based clinical practice is then of particular interest to clinical scientists for development of TCMs. Most Chinese doctors resist to (i) collect clinical data that they are not familiar with, and (ii) cooperate with Western doctors to collect further information from their patients due to fundamental differences in culture and clinical theory, perception, and practice. Thus, the transition from experience-based clinical practice to evidence-based clinical practice requires careful communication and planning. This transition is essential for achieving the ultimate goal of modernization and/or Westernization of TCMs.

CONCLUSION

One of the key issues in botanical drug product or TCM development is to clarify the difference between Westernization of TCM and modernization of TCM. For Westernization of TCM, we follow regulatory requirements at critical stages of the process for pharmaceutical development including drug discovery, formulation, laboratory development, animal studies, clinical development, manufacturing process validation and quality control, regulatory submission, review, and process despite the fundamental differences between WM and TCM. For modernization of TCM, it is suggested that regulatory requirements should be modified in order to account for the fundamental differences between WM and TCM. In other words, we still ought to be able to see if TCM is really working with modified regulatory requirements using Western clinical trials as a standard for comparison.

In practice, it is recognized that WMs tend to achieve the therapeutic effect sooner than that of TCMs for critical and/or life-threatening diseases. TCMs are found to be useful for patients with chronic diseases or non-life-threatening diseases. In many cases, TCMs have shown to be effective in reducing toxicities or improving safety profile for patients with critical and/or life-threatening diseases. As a strategy for TCM research and development, it is suggested that (i) TCM be used in conjunction with a well-established WM as a supplement to improve its safety profile and/or enhance therapeutic effect whenever possible, and (ii) TCM should be considered as the second line or third line treatment for patients who fail to respond to the available treatments. However, some sponsors are interested in focusing on the development of TCM as a dietary supplement due to (i) the lack or ambiguity of regulatory requirements, (ii) the lack of understanding of the medical theory/mechanism of TCM, (iii) the confidentiality

of non-disclosure of the multiple components, and (iv) the lack of understanding of pharmacological activities of the multiple components of TCM.

Since TCM consists of multiple components which may be manufactured from different sites or locations, the post-approval consistency in quality of the final product is both a challenge to the sponsor and a concern to the regulatory authority. As a result, some post-approval tests, such as tests for content uniformity, weight variation, and/or dissolution and (manufacturing) process validation, must be performed for quality assurance before the approved TCM can be released for use.

A comprehensive summarization of the issues in botanical drug product development can be found in Chow (2015).^[15]

REFERENCES

1. FDA. *Guidance for Industry – Botanical Drug Products*. The United States Food and Drug Administration: Rockville, MD, 2004.
2. Chow, S.C. and Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*. 3rd Ed., Chapman and Hall/CRC Press, Taylor & Francis: New York, 2008.
3. Chow, S.C., Pong, A., and Chang, Y.W. On traditional Chinese medicine clinical trials. *Drug Information Journal* **2006**, *40*, 395–406.
4. MAPP. *CDER/FDA Manual of Policies and Procedures (MAPP): Review of Botanical Drug Products*. Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, MD, June, 2004.
5. FDA. *Guidance for Industry – INDs for Phase 2 and Phase 3 Studies, Chemistry, Manufacturing, and Controls Information*. Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, MD, May 2003.
6. Song, X. New problems of intellectual property during innovation of traditional Chinese medicine. *World Science and Technology* **2011**, *13* (3), 466–469.
7. Hsiao, J.I.H. Patent protection for Chinese herbal medicine product invention in Taiwan. *The Journal of World Intellectual Property* **2007**, *10*, 1–21.
8. Koon, W.S., Lo, M.W., Marsden, J.E. and Ross, S.D. The genesis trajectory and heteroclinic connections, AAS/AIAA Astrodynamics Specialist Conference, Girdwood, AK, 1999, AAS99–451, 1999.
9. Yan, B.J. and Qu, H.B. An approach to maximize the batch mixing process for improving the quality consistency of the products made from traditional Chinese medicines. *Journal of Zhejiang University Science B* **2013**, *14*, 1041–1048.
10. Myers, R.H. *Response Surface Methodology*. Allyn and Bacon: Boston, MA, 1976.
11. Hicks, C. *Fundamental Concepts in the Design of Experiments*. 3rd Edition, CBS College Publishing: New York, 1982.
12. Qi, G.D., We, D.A., Chung, L.P., and Fai, C.K. Placebos used in clinical trials for Chinese herbal medicine. *Recent Pat Inflamm Allergy Drug Discov* **2008**, *2*, 123–127.
13. Fai, C.K., Qi, G.D., Wei, D.A., and Chung, L.P. Placebo preparation for the proper clinical trial of herbal medicine—requirements, verification and quality control. *Recent Pat Inflamm Allergy Drug Discov* **2011**, *5*, 169–174.
14. Hsiao, C.F., Tsou, H.H., Pong, A., Liu, J.P., Lin, C.H., Chang, Y.J., and Chow, S.C. Statistical validation of traditional Chinese diagnostic procedure. *Drug Information Journal* **2009**, *43*, 83–95.
15. Chow SC. *Quantitative Methods for Traditional Chinese Medicine Development*. Chapman and Hall/CRC Press: New York, 2015.

Bracketing Design

John R. Murphy

Eli Lilly and Company, Indianapolis, Indiana, U.S.A.

Abstract

Bracketing is the design of a stability schedule so that at any time point only the samples on the extremes, are tested. This entry provides a brief overview of this concept.

INTRODUCTION

Bracketing is a special term used in stability design. Since the amount of data required to demonstrate the stability characteristics of a product can often be large and expensive to generate, methods to reduce the size of stability studies have been proposed. As defined in the Glossary of the International Conference on Harmonization (ICH) Stability Guideline Q1A,^[1] *bracketing* is: “The design of a stability schedule so that at any time point only the samples on the extremes, for example, of container size and/or dosage strengths, are tested.” The Glossary further explains that the design assumes that the intermediate condition samples are represented by the extremes and that bracketing may be particularly applicable where strengths are related closely in composition, such as same formulation but different tablet sizes or different amounts filled into capsules. For bracketing across containers, the guideline indicates that bracketing designs may be applicable if the composition of the container and the type of closure are the same.

OVERVIEW

The U.S. Food and Drug Administration (FDA) draft stability guidance^[2] provides more detail about when bracketing may be applied without justification and when bracketing must be justified with theory and/or data. The FDA draft guidance provides an example, reproduced

in [Table 1](#), of a bracketing design for a case where there are three strengths made from a common granulation to be packaged in three different sizes of HPDE bottles with different counts: 30 counts, C1; 100 counts, C2; and 200 counts, C3.

The FDA draft guidelines also cautions that if the statistical analysis determines that there is dissimilarity between the extremes, the intermediate sizes/strengths should be considered to be no more stable than the least stable extreme. This could be a significant drawback to the use of bracketing depending of course, on the particular situation.

It seem clear that bracketing offers the potential for reducing the number of samples in registration stability studies. The question naturally arises whether bracketing has any application in validation studies and marketed product stability studies. For example, would it be possible to bracket across packages in the validation? Even if the all validation lots must be produced, would it be possible to place only some of the package/strength combinations on stability using bracketing? Since bracketing is a relatively new idea for the design of stability studies, these types of questions will continue to surface and will be discussed over the next few years.

CONCLUSION

Bracketing seems to be an idea that many stability studies can utilize to good advantage. However, it is

Table 1 Example of a Bracketing Design

Batch	1									2									3								
	100 mg			200 mg			300 mg			100 mg			200 mg			300 mg			100 mg			200 mg			300 mg		
Strength	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
Container/ Closure																											
Sample on Stability	X		X				X	X	X	X						X	X	X	X						X		X

also clear that the regulatory agencies worldwide will move forward with some caution, and proposed protocols will need to be cleared almost on a case-by-case basis at first. The ICH Stability Guideline and the FDA Draft Stability Guidance both encourage consultation with the appropriate regulatory body to get agreement about applying bracketing in the stability study before the protocol is executed.

REFERENCES

1. International Conference on Harmonization. *Harmonized Tripartate Guidelines: Stability Testing of New Drug Substances and Products*; Q1A (Step5 document effective 1 January 1998).
2. Food and Drug Administration (CDER and CBER). *Guidance for Industry: Stability Testing of Drug Substances and Drug Products*; Draft Guidance, June 1998.

Bridging Studies

Jen-pei Liu

Division of Biometry, Department of Agronomy, and Institute of Epidemiology, National Taiwan University, Taipei, Taiwan; and Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan

Abstract

As globalization becomes the key to survival of pharmaceutical companies, issues concerning the design and the analysis of bridging studies with respect to geotherapeutics pose new challenges to statisticians. This entry addresses some of these challenges and some methods of overcoming them.

INTRODUCTION

For the marketing approval of a pharmaceutical product, sponsors are required to provide substantial evidence of effectiveness and safety from adequate and well-controlled clinical trials. On the other hand, the U.S. Food and Drug Administration (FDA) recommends that usually at least two so-called “pivotal trials” in the same targeted patient population be performed to confirm the reproducibility of evidence on efficacy, safety, and dose response. However, after the pharmaceutical product is approved by the original region such as the United States or European Union, sponsors might seek the registration of the product in the new region. Due to the possible differences on ethnicity and clinical practice of the new region, the concerns that these differences may have an impact on the safety, efficacy, dose, and dosing regimen have limited the willingness for the regulatory authority in the new region, as a whole, to accept the foreign clinical data generated in the original region. Consequently, the regulatory authority in the new region has often requested the replication of all, or much, of the foreign clinical data in the new region. This extensive duplication of clinical evaluation in the new region not only requires valuable development resources but also delays the availability of the new pharmaceutical product to patients in need in the new region. To resolve this dilemma, the International Conference on Harmonization (ICH) has published a tripartite E5 guidance titled “Ethnic Factors in the Acceptability of Foreign Clinical Data” to address the above issues.^[1]

According to the ICH E5 guidance, ethnic factors are factors relating to races or larger populations grouped according to common traits and customs. Ethnic factors may be classified as intrinsic or extrinsic. Intrinsic ethnic factors are factors that help to define and to identify a subpopulation and may influence the ability to extrapolate clinical data between regions. Examples of intrinsic factors include genetic polymorphism, age, gender, height,

weight, lean body mass, body composition, and organ dysfunction. On the other hand, extrinsic ethnic factors are factors associated with the environment and culture in which a person lives. Extrinsic factors tend to be less genetically and more culturally and behaviorally determined. Examples of extrinsic factors include the social and cultural aspects of a region, such as a medical practice, diet, use of tobacco, use of alcohol, exposure to pollution and sunshine, socioeconomic status, and compliance with prescribed medications; particularly important are the reliance on studies from a different region, practices in clinical trial design, and conduct.

The objective of the guidance is to provide a framework for the evaluation of the impact of ethnic factors on the efficacy and the safety of a pharmaceutical product at a particular dosage or dose regimen. In addition, it describes regulatory strategies of minimizing duplication for clinical data and the requirement of bridging evidence for extrapolation of foreign clinical data to a new region. According to the ICH E5 guidance, a bridge data package consists of (i) selected information from the Complete Clinical Data Package (CCDP), which is relevant to the population of the new region; and, (ii) if needed, a bridging study to extrapolate the foreign efficacy and/or safety data to the new region. In other words, bridging evidence is actually provided either in the CCDP generated during the clinical drug development program for submission to the original region, or in a bridging study conducted in the new region after the pharmaceutical product is approved in the original region. If the bridging evidence provided in the CCDP cannot allow the extrapolation of foreign clinical data to a new region, then a bridging study should be conducted in the new region to generate a limited amount of clinical data to bridge the clinical data between the two regions. As a result, the ICH defined a bridging study as a supplemental study performed in the new region to provide pharmacodynamic (PD) or clinical data on efficacy, safety, dosage, and dose regimen in the new region that will

allow the extrapolation of the foreign clinical data to the new region. Bridging studies can also include additional pharmacokinetic (PK) information.

NATURE AND TYPES OF BRIDGING STUDIES

In general, the types of bridging studies required depend upon the ethnic sensitivity of the pharmaceutical agent, experience of the drug class, extrinsic ethnic factors, and ethnic differences between the new and original regions. From the ICH E5 guidance, the following table is a summary of types of bridging studies with respect to the above-mentioned factors:

<i>Medicine</i>	<i>Region</i>	<i>Medical practice</i>	<i>Drug class</i>	<i>Clinical experience</i>	<i>Bridging studies</i>
Insensitive	—	Similar	—	—	No
Sensitive	Similar	—	—	Sufficient	No
Sensitive	Dissimilar	Similar	Familiar	—	PD
Dosage	—	Different	Unfamiliar	Insufficient	CCT ^a

^aControlled clinical trials.

If the populations between the two regions are similar, medical practical practices are also similar, and sufficient experience with the pharmaceutical is available, too, the extrapolation of foreign clinical data to the new region might be allowed without a bridging study even when the medicine is ethnically sensitive. On the other hand, from the above table, if the pharmaceutical is ethnically sensitive and the populations of the two regions are not similar but the medical practice and the conduct of the clinical trials are similar and the drug class is also a familiar one in the new region, then a controlled pharmacodynamic (PD) study with a well-established pharmacological end point may generate the necessary data to bridge the foreign clinical data. However, the uncertainty of the dose used in the new region, insufficient clinical experience, dissimilar medical practice, or unfamiliar drug class may warrant that controlled clinical trials be conducted in the new region to obtain the necessary bridging data. Occasionally, the bridging safety studies might be needed to accurately determine the rates of the relatively common adverse events in the new region and to detect some serious adverse events. However, in general, the relationship among a PK, PK/PD, and clinical bridging studies is not always clear-cut. The translation of similarity observed from PK studies to similarity or reproducibility in clinical trials requires further research. This type of research can start by examining systematically the relationship among PK, PD, and clinical data of the approved drugs by drug class.

As pointed out by the ICH E5 guidance, bridging studies might not be necessary in some situations for the extrapolation of the foreign clinical data to the new region. Chow et al.^[2] proposed a method to evaluate the necessity

of bridging studies that is based on the concept of reproducibility and generalizability of the foreign clinical data. Reproducibility is the probability of reaching the positive treatment effect in the second clinical trial with the same patient population given the result of the positive treatment effect for the first trial. On the other hand, generalizability is the probability of reaching the positive treatment effect in the second clinical trial with the different patient population given the result of the positive treatment effect for the first trial. For the calculation of generalizability, the differences in population mean and variance between the new and original regions are characterized by a sensitivity index. Chow et al.^[2] proposed the following sequential strategy for the necessity of bridging studies:

- Step 1: From the CCDP, summarize the results and calculate the reproducibility probability. If the reproducibility probability is greater than some prespecified level set by the regulatory authority, then stop and suggest that bridging studies, from a statistical viewpoint, are not required for the evaluation of the ability of extrapolation of the foreign clinical data to the new region. Otherwise go to Step 2.
- Step 2: Identify the sensitivity indices.
- Step 3: Compute the generalizability probability. Compare the generalizability with the regulatory requirement to determine the necessity of bridging studies and the types of bridging studies to be conducted in the new region.

The ICH E5 guidance stresses that the regulatory authority in the new region request only additional bridging data necessary to assess the ability to extrapolate the foreign clinical data from the CCDP to the new region. As a result, the ability for extrapolation is the key to the evaluation of bridging data generated in the new region. The ICH E5 describes two situations where the data generated from a bridging study can provide scientific evidence for allowing the extrapolation of the foreign clinical data to the new region:

- Situation 1: If the bridging study shows that dose response, safety, and efficacy in the new region are similar, then the study is readily interpreted as capable of bridging the foreign clinical data.
- Situation 2: If a bridging study indicates that a different dose in the new region results in a safety and efficacy profile that is not too substantially different from that derived in the original region, it will often be possible to extrapolate the foreign clinical data to the new region, with appropriate dose adjustment.

EXTRAPOLATION AND SIMILARITY

Although the ICH E5 guidance clearly states that the assessment of the ability of extrapolation of the foreign data relies on the similarity of dose response, efficacy, and safety between the new and original regions, with or without dose adjustment, it does not provide a precise definition of similarity. Different interpretations of similarity have been proposed.^[3-5] They include positive treatment effect, equivalence, non-inferiority, and consistency.

For simplicity, we only consider the problem for assessment of similarity on efficacy for comparing a test medicine with a placebo control. Similar ideas can be directly applied to the evaluation of similarity with respect to either dose response or safety. Let $\mathcal{Y}_{ijk}$ be the some efficacy response variable for patient k receiving treatment j in region i , $k = 1, \dots, K$; $j = T$ (test), P (placebo); and $i = O$ (original), N (new). For the sake of illustration, we assume that approximately $\mathcal{Y}_{ijk}$ is independently distributed as a normal distribution with mean μ_{ij} and known variance σ^2 . The treatment effect for region i is defined as:

$$\Delta_i = \mu_{iT} - \mu_{iP}, \quad i = O, N$$

The concept of the positive treatment effect is defined as follows. Because the test medicine has already been approved in the original region due to its proven efficacy against placebo control, if the data collected from the bridging studies in the new region demonstrate superior efficacy of the test medicine over placebo control, then the efficacy of the new region is claimed to be similar to that of the original region. In other words, under the concept of the positive treatment, the effect for similarity can be evaluated through the following hypothesis:

$$H_0 : \Delta_N \leq 0 \quad \text{vs.} \quad H_a : \Delta_N > 0 \quad (1)$$

However, the concept of the positive treatment effect does not really assess the similarity of the magnitude of efficacy between the two regions. Therefore Liu et al.^[3] applied the concept of average bioequivalence to the evaluation of similarity for the extrapolation of foreign clinical data. The idea is that the similarity with respect to efficacy should be interpreted as the difference of treatment effects between regions within some clinically acceptable limit, say δ . The relationship between δ and overall treatment effect Δ can be expressed as $\delta = f\Delta$. δ is clinically acceptable and meaningful only if $0 < f < 1$ because similarity dictates that the difference of treatment effects between regions should be smaller than the overall treatment effect. Let $\theta = (\mu_{NT} - \mu_{NP}) - (\mu_{OT} - \mu_{OP})$, then the corresponding hypothesis can be formulated as the following two-sided equivalence hypothesis:

$$H_0 : \theta \leq -\delta \quad \text{or} \quad \theta \geq \delta \quad \text{vs.} \quad H_a : -\delta < \theta < \delta \quad (2)$$

On the other hand, the non-inferiority test may also be appropriate for evaluating the similarity between the two regions because the efficacy of the new region can be claimed to be similar to, if it is not inferior to that of, the original region. The one-sided non-inferiority hypothesis is given as:

$$H_{0L} : \theta \leq -\delta \quad \text{vs.} \quad H_a : \theta \geq -\delta \quad (3)$$

Shih^[5] proposed the idea of the consistency for evaluation of the similarity of the efficacy between the new and original regions. Let $\mathbf{w} = \{w_1, \dots, w_H\}$ be the results of the H reference studies that were conducted in the original region and let v be the result of the bridging study conducted in the new region. The result of the bridging study is said to be consistent with the previous results $\mathbf{w}$ if:

$$p(v|\mathbf{w}) \geq \min \left\{ p(w_h|\mathbf{w}), h = 1, \dots, H \right\} \quad (4)$$

where $p(u|\mathbf{w})$ is the predictive probability function of the study u given the previous result $\mathbf{w}$.

The concept of consistency in Eq. 4 can be reformulated as

$$|\Delta_N - \Delta| < v \quad (5)$$

where Δ is the mean of the treatment effects over the H previous studies. Under the normal assumption, $v^2 = \max \{ (\Delta_h - \Delta)^2, h = 1, \dots, H \}$.

DESIGN AND ANALYSIS

The goal of analysis of the clinical data generated by the local bridging study is not only to provide the information on the dose response, efficacy, and safety with the data in the patient population of the new region alone, but also to evaluate the similarity of dose response, efficacy, and safety with the foreign clinical data by the clinical trials conducted in the original region. As a result, the bridging evidence provided by this strategy consists of the clinical data by the local bridging study as well as those by the clinical trial in the original region, which are not concurrently generated and hence are not internally valid. To minimize the bias on similarity inference, the clinical data by the new and the original regions must be at least externally valid. It follows that the bridging study in the new region should replicate the randomized parallel dose response design as close to those of the original region as possible with respect to study design, inclusion and exclusion criteria, dosage, dosage regimen, and control. The difference between the bridging study and trials conducted in the original region should be limited only to the region and sample size, or possibly end points and duration.

Liu et al.^[3] proposed a hierarchical model to evaluate the similarity in terms of equivalence or non-inferiority on the results between the bridging study and the original region. Under the proposed hierarchical model, the maximum likelihood estimate (MLE) of θ and its asymptotic variance can be obtained to construct a $(1 - 2\alpha)100\%$ confidence interval for y . The similarity of the efficacy provided by the bridging study in the new region and foreign clinical data can be claimed if the $(1 - 2\alpha)100\%$ confidence interval for y is completed within $(-\delta, \delta)$. Similarly, the efficacy of the bridging study in the new region is not inferior to that provided by the foreign clinical data if the lower limit of the $(1 - 2\alpha)100\%$ confidence interval for y is greater than $-\delta$. If the variance is the same for all trials in the original region and the new region, and if treatment allocation is 1:1, for type I and type II error rates controlled at α and β levels, respectively, the total sample size for the bridging study is given as

$$n_N \geq 1 / \left\{ (f/CV)^2 [z(\alpha) + z(\beta)]^2 - (1/N) \right\} \quad (6)$$

where $f = \delta/(\mu_{OT} - \mu_{OP})$, $CV = 2\sigma/(\mu_{OT} - \mu_{OP})$, N is the total number of the patients in CCDP provided by the original region, and σ^2 is the common variance. Under the hierarchical model, the sample size required for bridging studies sometimes is forbiddingly large to practically conduct bridging studies in the new region. In addition, one of the objectives of a bridging study is to minimize an unnecessary duplication of clinical data in the new region. Therefore it is imperative to borrow the strength and to incorporate the information on dose response, efficacy, and safety from the CCDP generated in the original region into the analysis of the data obtained from the bridging study in the new region. Liu et al.^[4] suggest applying the empirical Bayesian method to the assessment of similarity in terms of a positive treatment effect. They proposed the following model:

$$Y = \mu_p(1 - x) + \{\mu_{NT}Z + \mu_{OT}(1 - Z)\}X + \varepsilon \quad (7)$$

where μ_p is the common placebo mean for both new and original regions; $\mu_{NT}(\mu_{OT})$ is the treatment mean for the new (original) region; X , being 1(0), indicates the treatment (placebo) group; and Z , being 1(0), indicates the new (original) region. Given the data collected from the bridging study and prior distribution on $(\mu_p, \mu_{NT}, \mu_{OT})$, similarity on efficacy in terms of a positive treatment effect for the new region can be concluded if the posterior probability:

$$Pr\{\mu_{NT} - \mu_p > 0 | \text{bridging data and prior}\} > 1 - \alpha$$

for some prespecified $\alpha > 0$.

Under further assumption of $\mu_{NT} = \mu_{OT}$, the sample size for the bridging study can be expressed in terms

of the sample size in the CCDP and the strength of the evidence of the positive treatment effect provided by the CCDP in the original region. In addition, this approach assumes the same placebo mean for both new and original regions. It follows that the sample size for the bridging study decreases when the number of patients assigned to the treatment group increases. For establishing similarity using the concept of consistency using normal responses, let w_h be the standardized between-group difference in means ($h = 1, \dots, H$). If the total sample size for each of H reference studies is reasonably large, then w_h approximately follows a normal distribution with common mean Δ and variance 1 ($h = 1, \dots, H$). Under the assumption of vague prior for Δ , the predictive probability function for the standardized between-group difference v for the bridging study follows a normal distribution with mean w and variance $(H + 1)/H$, where w is the mean of $w_1, \dots, w_H$.

Shih^[5] showed that $p(v|w) \geq \min\{p(w_h|w), h = 1, \dots, H\}$ if and only if $(v - w)^2 \leq \lambda = \max\{(w_h - w)^2, h = 1, \dots, H\}$. In addition, let R denote the expanse of all consistent trials $R = \{v: (v - w)^2 \leq \lambda\}$. The cover of this expanse is given by the predictive probability $Pr(R) = 1 - 2\Phi\{-[\lambda H/(H + 1)]^{1/2}\}$, where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal distribution. $Pr(R)$ is a probability that quantifies the chance that the results of the bridging study in the new region will be consistent with the experience of the reference studies in the CCDP provided by the original region. If $Pr(R)$ is greater than some prespecified limit set by the regulatory authority, say 90%, Shih^[5] suggests that there is no need to test what has already been shown in the CCDP. Therefore the conventional sample size and power calculation are no longer relevant in this situation. On the other hand, if the reference studies in the CCDP fail to provide sufficient support that $Pr(R)$ meets the regulatory requirement, the bridging study will be the confirmation trial in the new region and hence conventional sample size and power calculation will be required.

Liu et al.^[6,7] proposed a Bayesian approach to evaluation of bridging studies. On the other hand, to avoid the bias due to between-study comparison, the methods in a sequential nature have been proposed to assess bridging studies.^[8,9] Liu et al.^[10] suggest the concept of ICH E5 can be also applied to evaluation of acceptability of foreign bioequivalence data for generic drugs. Most importantly, in September 2006, ICH issued a guidance on the Ethnic Factors in the Acceptability of Foreign Data—Questions and Answers.^[11]

CONCLUSION

The ICH E5 guidance provides the regulatory requirements that the acceptability of foreign clinical data should be evaluated based on the similarity on dose response, efficacy, and safety between the new and original regions. However, it does not provide a precise definition of

similarity. Although various interpretations of similarity have been suggested, no consensus has been reached. Until a scientifically sound and statistically valid definition of similarity is defined, all currently proposed statistical methodology on the bridging studies cannot be installed as the regulatory implementation for the evaluation of extrapolation of the foreign clinical data to the new region. As globalization becomes the key to survival of pharmaceutical companies, issues concerning the design and the analysis of bridging studies with respect to geotherapeutics pose new challenges to statisticians. These issues include the designs of global clinical trials versus multinational bridging study, frequentist approach versus Bayesian approach, use of genomic information, selection of end points for the bridging study, determination of equivalence limits, and sample size requirements. Details regarding the resolutions and/or discussions of these issues can be found in Liu, Chow and Hsiao (2012).^[12]

ACKNOWLEDGEMENTS

Minor updates have been made by the Editor-in-Chief in order to reflect developments since publication of the *Encyclopedia of Biopharmaceutical Statistics, Third Edition*.

REFERENCES

1. ICH. *International Conference on Harmonisation Tripartite Guidance E 5, 1997; Ethnic Factors in the Acceptability of Foreign Data*; The U.S. Federal Register: Washington, DC, 1998, 31790–31796.
2. Chow, S.C.; Shao, J.; Hu, O.Y.P. Assessing sensitivity and similarity in bridging studies. *J. Biopharm. Stat.* **2002**, *12*, 385–400.
3. Liu, J.P.; Hsueh, H.-M.; Chen, J.J. Sample size requirement for evaluation of bridging evidence. *Biometr. J.* **2002**, *44*, 965–981.
4. Liu, J.P.; Hsueh, H.-M.; Hsiao, C.F. Bayesian approach to evaluation of the bridging studies. *J. Biopharm. Stat.* **2002**, *12*, 401–408.
5. Shih, W.J. Clinical trials for drug registrations in Asian-Pacific countries: Proposal for a new paradigm from a statistical perspective. *Control. Clin. Trials* **2001**, *22*, 357–366.
6. Liu, J.P.; Hsueh, H.M.; Hsiao, C.F. A Bayesian non-inferiority approach to evaluation of bridging studies. *J. Biopharm. Stat.* **2004**, *14*, 291–300.
7. Hsiao, C.F.; Hsu, Y.Y.; Liu, J.P. Use of prior distributions for Bayesian evaluation of bridging studies. *J. Biopharm. Stat.* **2007**, *17*, 109–121.
8. Hsiao, C.F.; Xu, J.Z.; Liu, J.P. A group sequential approach to evaluation of bridging studies. *J. Biopharm. Stat.* **2003**, *13*, 793–810.
9. Hsiao, C.F.; Xu, J.Z.; Liu, J.P. A two-stage design for bridging studies. *J. Biopharm. Stat.* **2005**, *15*, 75–84.
10. Liu, J.P. Bridging bioequivalence studies. *J. Biopharm. Stat.* **2004**, *14*, 857–868.
11. ICH. *International Conference on Harmonisation Tripartite Guidance E 5, 2006; Ethnic Factors in the Acceptability of Foreign Data – Questions and Answers*; US Food and Drug Administration: Rockville.
12. Liu, J.P.; Chow, S.C.; Hsiao, C.F. (Ed) *Design and Analysis of Bridging Studies*. Taylor & Francis: New York, 2012.

Bridging Studies: Statistical Methods

Yuh-Jenn Wu

Department of Applied Mathematics, Chung Yuan Christian University, Taoyuan, Taiwan

Chieh Chiang, Chi-Tian Chen, Hsiao-Hui Tsou, and Chin-Fu Hsiao

Institute of Population Health Sciences, National Health Research Institutes, Miaoli, Taiwan

Abstract

After the test product has been approved for commercial marketing in the original region based on its proven efficacy and safety, and if the sponsor wants to apply the new drug registration to a new region, a bridging study may be conducted in the new region to provide pharmacodynamic or clinical data on efficacy, safety, dosage and dose regimen to allow extrapolation of the foreign clinical data to the population of the new region. In this entry, we reviewed some statistical methods for bridging studies. In the first part, we reviewed the strategy for determining whether a bridging study is needed for providing substantial evidence in the new region based on the clinical results observed in the original region. In the last three part, we reviewed three statistical methods for assessing the similarity between a bridging study conducted in a new region and studies conducted in the original region.

Keywords: Bridging study; Complete clinical data package (CCDP); Similarity.

INTRODUCTION

The idea of bridging study came from the ICH E5 (1) entitled “Ethnic Factors in the Acceptability of Foreign Clinical Data.” The purpose of the ICH E5 is to facilitate the registration of medicines among ICH regions including European Union (EU), the United States of America (USA), and Japan by recommending a framework for evaluating the impact of ethnic factors on a medicine’s effect such as its efficacy and safety at a particular dosage and dose regimen. By ICH E5, bridging study is defined as a supplementary study conducted in the new region (e.g., Asian Pacific Region) to provide pharmacodynamic or clinical data on efficacy, safety, dosage and dose regimen to allow extrapolation of the foreign clinical data (e.g., data from EU or USA) to the population of the new region. Therefore, a bridging study is usually conducted in the new region only after the test product has been approved for commercial marketing in the original region (e.g., the United States of America or European Union) based on its proven efficacy and safety. Although use of bridging studies for bringing an approved drug product from the original region to a new region has attracted sponsors as well as regulatory authorities in recent years, the key issue however lies on how the ethnic factors influence clinical outcomes for evaluation of efficacy and safety.

The ICH E5 guideline provides regulatory strategies of minimizing duplication of clinical data and requirement of bridging evidence for extrapolation of foreign clinical data to a new region. To do so, the bridging data package should

include (1) information including PK data and any preliminary PD and dose-response data from the complete clinical data package (CCDP) that is relevant to the population of the new region and if needed, (2) bridging study to extrapolate the foreign efficacy data and/or safety data to the new region. If the study medicines are insensitive to ethnic factors, the type of bridging studies (if needed) will depend upon experience with the drug class and upon the likelihood that extrinsic ethnic factors could affect the medicine’s safety, efficacy, and dose-response. On the other hand, for medicines that are ethnically sensitive, bridging study is usually needed since the populations in two regions are different.

The ICH E5 guideline points out that evaluation of the ability of extrapolation of the foreign clinical data relies upon the similarity of dose response, efficacy, and safety between the new and original regions either with or without dose adjustment. Several statistical procedures have been proposed to assess the similarity based on the additional information from the bridging study and the foreign clinical data in the CCDP. In this entry, we will review some statistical methods for bridging studies.

REPRODUCIBILITY AND GENERALIZABILITY

Shao and Chow (2) proposed the concepts of reproducibility and generalizability probabilities for assessment of bridging studies. For convention, we focus on the trials for comparing a test product and a placebo control. Let n_t and

n_p represent the numbers of patients recruited for the test product and the placebo respectively in new region. Let X_i and Y_j be the efficacy responses for patients i and j receiving the test product and the placebo control respectively, $i = 1, \dots, n_T, j = 1, \dots, n_p$. We assume that both X_i 's and Y_j 's are normally distributed with variance σ^2 . Let μ_{NT} and μ_{NP} be the population means of the test and placebo, respectively, and let $\Delta_N = \mu_{NT} - \mu_{NP}$. Subsequently, the hypothesis of testing for the treatment effect is given as

$$H_0 : \Delta_N = 0 \text{ vs. } H_A : \Delta_N \neq 0. \quad (1)$$

The test statistic for (1) under H_0 is taken as

$$T_N = \frac{\bar{x}_T - \bar{y}_p}{\sqrt{\frac{(n_T - 1)s_T^2 + (n_p - 1)s_p^2}{(n - 2)} \left(\frac{1}{n_T} + \frac{1}{n_p} \right)}} \sim t_{n-2},$$

where $\bar{x}_T = \sum_{i=1}^{n_T} x_i / n_T$, $\bar{y}_p = \sum_{j=1}^{n_p} y_j / n_p$, s_T^2 and s_p^2 are the sample variances based on the data from the test and placebo groups respectively, and t_{n-2} represents the t distribution with degrees of freedom $n - 2$ with $n = n_T + n_p$. More specifically, we reject H_0 at the α level of significance if and only if $|T_N| > t_{n-2, 1-\alpha/2}$, where $t_{n-2, 1-\alpha/2}$ is 100(1 - $\alpha/2$)% percentile of the t -distribution with $n - 2$ degrees of freedom. Consequently, the power of T_N is given by

$$p(\theta) = P\left(|T_N| > t_{n-2, 1-\alpha/2} \mid \theta\right) = 1 - \mathfrak{I}_{n-2}\left(t_{n-2, 1-\alpha/2} \mid \theta\right) + \mathfrak{I}_{n-2}\left(-t_{n-2, 1-\alpha/2} \mid \theta\right), \quad (2)$$

where

$$\theta = \frac{\Delta_N}{\sigma \sqrt{\frac{1}{n_T} + \frac{1}{n_p}}}$$

and $\mathfrak{I}_{n-2}(\cdot \mid \theta)$ denotes the cumulative distribution function of the non-central t distribution with $n - 2$ degrees of freedom and the non-centrality parameter θ .

Let T_O represent the test statistics obtained from the original region and t_O the value of T_O based on the data observed from the original region. When the ethnic difference is negligible, we can replace θ in Eq. (2) by its estimate t_O and derive the estimated power

$$\hat{p} = 1 - \mathfrak{I}_{n-2}\left(t_{n-2, 1-\alpha/2} \mid t_O\right) + \mathfrak{I}_{n-2}\left(-t_{n-2, 1-\alpha/2} \mid t_O\right)$$

which consequently can be defined as a reproducibility probability for a future clinical trial with the same patient population. When the ethnic difference is notable, Shao and Chow (2) recommend to calculate the estimated reproducibility probability based on a set of values of θ . If all the estimated reproducibility probabilities obtained based on this set of values of θ meets regulatory requirement,

then the sponsor may stop and conclude that clinical bridging studies are not needed and clinical data in the original region is capable of bridging to the new region. Conversely, if the estimated reproducibility probabilities fail to meet regulatory requirement, then the sponsor may need to prepare for a bridging study.

TEST FOR CONSISTENCY

Shih (3) proposed a method for assessment of consistency to determine whether the study is capable of bridging the foreign data to the new region. Again, we only focus on the trials for comparing a test product and a placebo control. Suppose there are K reference studies in the CCDP. Based on the K historical reference studies, the test product has been already approved in the original region due to its proven efficacy against placebo control.

Let x_{ij} be some efficacy responses for the j^{th} patient receiving the test product in the i^{th} historical trial, $i = 1, \dots, K$, and $j = 1, \dots, m_i$ and y_{ij} the efficacy responses for j^{th} patient receiving the placebo control in the i^{th} historical trial, $i = 1, \dots, K$, and $j = 1, \dots, n_i$. Then the treatment group difference in means for the i^{th} study is

$$\omega_i = \bar{x}_i - \bar{y}_i,$$

where $\bar{x}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij}$ and $\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$ are the sample means of the m_i and n_i observations in the drug and the placebo group, respectively. Consequently, the standardized treatment difference in means for the i^{th} study can be expressed by

$$T_i = \frac{\omega_i}{s_i \sqrt{\frac{1}{m_i} + \frac{1}{n_i}}}, \quad (3)$$

where s_i is the pooled sample standard deviation of the i^{th} trial. Note that T_i approximately follows a normal distribution with mean μ and variance 1 with sufficient sample sizes. Let $T = (T_1, \dots, T_K)$ be the results of the K reference studies. Also let $\mathbf{w} = (\omega_1, \dots, \omega_K)$.

Let x_{Br} and y_{Bs} be the efficacy responses for patients r and s receiving the test product and the placebo control respectively from the bridging study conducted in the new region, $r = 1, \dots, m_B$, and $s = 1, \dots, n_B$. Let v be the treatment group difference in means for the bridging study. That is,

$$v = \bar{x}_B - \bar{y}_B,$$

where $\bar{x}_B = \frac{1}{m_B} \sum_{r=1}^{m_B} x_{Br}$ and $\bar{y}_B = \frac{1}{n_B} \sum_{s=1}^{n_B} y_{Bs}$. Let T_B be the standardized treatment difference in means of the bridging study conducted in the new region. That is

$$T_B = \frac{v}{s_B \sqrt{\frac{1}{m_B} + \frac{1}{n_B}}},$$

where s_B is the pooled sample standard deviation of the bridging study.

Shih (3) constructed the predictive probability function, $p(T_B | T)$, which provides a measure of plausibility of T_N given the previous results T and also constructed the predictive probability functions, $p(T_i | T)$, for $i = 1, \dots, K$. The result T_N from the bridging study is consistent with the reference result T if and only if

$$p(T_N | T) \geq \min \{p(T_i | T), i = 1, \dots, K\}.$$

Similar to Shih (3), Tsou et al. (4) construct the predictive probability function, $p(v | \mathbf{w})$, which provides a measure of plausibility of v given the previous results $\mathbf{w}$. In addition, $p(\omega_i | \mathbf{w})$ for $i = 1, \dots, K$ can also be constructed. The result v is consistent with the reference result $\mathbf{w}$ if and only if

$$p(v | \mathbf{w}) \geq \rho_B \min \{p(\omega_i | \mathbf{w}), i = 1, \dots, K\},$$

for some pre-specified $\rho_B > 0$. Here ρ_B represents the magnitude of consistency trend.

WEIGHTED Z-STATISTICS

Lan, Soo et al. (5) introduced the weighted Z-tests in which the weights may depend on the prior observed data for the design of bridging studies. As seen in Section 2, T_O and T_N are approximately distributed as a standard normal distribution, $N(0, 1)$, for reasonable sample sizes. The weighted Z-test with constant $0 \leq w \leq 1$, is expressed as

$$Z_w = \sqrt{w}T_O + \sqrt{1-w}T_N.$$

The condition $|Z_w| > z_{1-\alpha/2}$ implies the bridging study conducted in the new region is consistent with the study conducted in the original region. Note that the weighted Z-test combines two Z-values from two different studies conducted in different ethnic populations.

USE OF PRIOR INFORMATION FOR BAYESIAN EVALUATION

To minimize unnecessary duplication of generating clinical data in the new region, the sponsor may borrow “strength” from the information on dose response, efficacy, and safety from the CCDP in the original region and incorporate them into the analysis of the additional data obtained from the bridging study. Hsiao, Hsu et al. (6) therefore proposed

a Bayesian approach by considering a mixture model for the prior information of Δ_N in Section 2 to synthesize the data generated by the bridging study and foreign clinical data generated in the original region for assessment of similarity. More specifically, the proposed mixture model of the prior information for Δ_N is a weighted average of two priors as given below

$$\pi(\Delta_N) = \gamma \pi_1(\Delta_N) + (1-\gamma) \pi_2(\Delta_N), \quad (4)$$

where $0 \leq \gamma \leq 1$. In (4), $\pi_2(\cdot)$ is a normal prior with mean θ_0 and variance σ_0^2 summarizing the foreign clinical data about the treatment difference obtained from the CCDP, whereas $\pi_1(\cdot) \equiv c$. Here $\pi_1(\cdot)$ is a noninformative prior which represent no information will be borrowed from the CCDP.

Given the data from the bridging study and prior information, similarity on efficacy in terms of a positive treatment effect for the new region can be concluded if the posterior probability of similarity

$$P_{SP} = P(\mu_{NT} - \mu_{NP} > 0 | \text{bridging data and prior}) > \tau,$$

for some pre-specified τ . Here the value of τ may be determined by the negotiation between the sponsor and the regulatory agency of the new region, and should generally ensure that posterior probability of similarity is at least 80%. On the other hand, selection of weight, γ should reflect the difference in both intrinsic and extrinsic ethnic factors defined in ICH E5 between the new and original regions.

CONCLUSION

In this entry, we reviewed some statistical methods for bridging studies. In the first part, we reviewed the strategy proposed by Shao and Chow (2) for determining whether a bridging study is needed for providing substantial evidence in the new region based on the clinical results observed in the original region. In the last three part, we reviewed statistical methods developed by Shih (3), Lan, Soo et al. (5), and Hsiao, Hsu et al. (6) for assessing the similarity between a bridging study conducted in a new region and studies conducted in the original region. Determination of sample size required for the bridging study can also be found in their work.

ACKNOWLEDGMENTS

The authors would like to thank the Editor of EBS, Professor Shein-Chung Chow, for inviting us to contribute this entry.

REFERENCES

1. ICH. *International Conference on Harmonisation: Tripartite Guidance E5 Ethnic Factors in the Acceptability of Foreign Data*. The U.S. Federal Register 1998, 83, 31790–31796.
2. Shao, J.; Chow, S.C. Reproducibility probability in clinical trials. *Statistics in Medicine* **2002**, *21* (12), 1727–1742.
3. Shih, W.J. Clinical trials for drug registration in Asian-Pacific countries: proposal for a new paradigm from a statistical perspective. *Controlled Clinical Trials* **2001**, *22* (4), 357–366.
4. Tsou, H.H.; Chien, T.Y.; Liu, J.P.; Hsiao, C.F. A consistency approach to evaluation of bridging studies and multiregional trials. *Statistics in Medicine* **2011**, *17*, 2171–2186.
5. Lan, G.K.K.; Soo, Y.W.; Siu, C.T.; Wang, M. The use of weighted Z-tests in medical research. *Journal of Biopharmaceutical Statistics* **2005**, *15* (4), 625–39.
6. Hsiao, C.F.; Hsu, Y.Y.; Tsou, H.H.; Liu, J.P. Use of prior information for Bayesian evaluation of bridging studies. *Journal of Biopharmaceutical Statistics* **2007**, *17* (1), 109–121.

Calibration

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

This entry discusses calibration for obtaining analytical results and its statistical validity and also summarizes statistical methods for assessment of performance characteristics for assay validation.

INTRODUCTION

When a new pharmaceutical compound is discovered, most regulatory agencies such as the United States Food and Drug Administration (FDA) require that an analytical method or test procedure for quantitation of the active ingredients of the compound be developed and validated before it can be applied to animal or human subjects.^[1] The FDA current Good Manufacturing Practice (cGMP) requires that test methods, which are used for assessing compliance of pharmaceutical products with established specifications, must meet proper standards of accuracy and reliability. The General Chapter of the United States Pharmacopeia and National Formulary (USP/NF) defines the validation of analytical methods as the process by which it is established, in laboratory studies, that performance characteristics of the methods meet the requirements for the intended analytical application.^[2] The performance characteristics include accuracy, precision, limit of detection (LOD), limit of quantitation (LOQ), selectivity (or specificity), linearity, range, and ruggedness. These performance characteristics are often used for assessment of accuracy and reliability of the test results obtained from the analytical methods or the test procedures. The analytical application may be referred to as a drug potency, which is usually based on gas chromatography (GC) or high-performance liquid chromatography (HPLC) for potency and stability studies, immunoassays such as radioimmunoassay (RIA) for the *in vitro* activity of an antibody or antigen, or a biological assay for the *in vivo* activity such as median effective dose (ED_{50}).

To assure the accuracy and reliability of the test results obtained from the analytical method or the testing procedure based on some instruments such as HPLC, calibration of the instrument employed in the analytical method or the testing procedure is critical. In the next section, statistical methods for establishment of the standard curve in calibration for assay validation are introduced. “Calibration and Analytical Results” discusses calibration for obtaining analytical results and its statistical validity (bias and mean-squared error). “An Example” provides a summary

of statistical methods for assessment of performance characteristics for assay validation as described in the USP/NF.

STANDARD CURVE

As indicated earlier, when a new pharmaceutical compound is discovered, analytical methods or test procedures are necessarily developed for determining the active ingredients of the compound in compliance with USP/NF standards for the identity, strength, quality, purity stability, and reproducibility of the compound. An analytical method or test procedure is usually developed based on instruments such as HPLC. The cGMP indicates that an instrument must be suitable for its intended purposes and be capable of producing valid results with certain degrees of accuracy and reliability. Therefore the instrument must be calibrated, inspected, and routinely checked according to written procedures. As a result, instrument calibration is essential in order to meet the established written procedures or specifications.

A typical approach to instrument calibration is to have a number of known standard concentration preparations $\{x_i, i = 1, \dots, n\}$ put through the instrument to obtain the corresponding responses $\{y_i, i = 1, \dots, n\}$. Based on $\{x_i, y_i, i = 1, \dots, n\}$, an estimated calibration curve can be obtained by fitting an appropriate statistical model between y_i and x_i . The estimated calibration curve is usually referred to as the “standard curve.” For a given unknown sample x_0 with a response of y_0 , the concentration of x_0 can be estimated based on the established standard curve by the x value that corresponds to y_0 . Alternatively, an estimate of x_0 can also be obtained by applying the inverse regression with x_i and y_i switched. Chow and Shao^[3] studied the difference between the estimates of x_0 obtained by using the standard curve and the inverse regression.

Model Selection

For an analytical method or a test procedure, the precision and reliability of the analytical result or the test result of an

unknown sample rely on the validity and efficiency of the established standard curve that is constructed based on the selected statistical model. Therefore it is critical to select an appropriate statistical model to calibrate the instrument. Let x_i be the i th standard concentration preparation and y_i be the corresponding response from the instrument, $i = 1, \dots, n$. In practice, the following models are commonly used for establishment of standard curve:

Model	Description
1	$y_i = x_i\alpha + \beta x_i + e_i$
2	$y_i = \beta x_i + e_i$
3	$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + e_i$
4	$y_i = \alpha x_i^\beta e_i$
5	$y_i = \alpha e^{\beta x_i} e_i$

where α , β , β_1 , and β_2 are unknown parameters and e_i 's are independent random errors with $E(e_i) = 0$ and finite $\text{Var}(e_i)$ in models 1–3, and $E(\log e_i) = 0$ and finite $\text{Var}(\log e_i)$ in models 4 and 5.

Model 1 is a simple linear regression model that is probably the most commonly used statistical model for establishment of standard curves. When the standard curve passes through the origin, model 1 reduces to model 2. Model 3 indicates that the relationship between y_i and x_i is quadratic. When there is a nonlinear relationship between y_i and x_i , models 4 and 5 are useful. As a matter of fact, both model 4 and model 5 are equivalent to a simple linear regression model after a logarithmic transformation. It can be seen that under each of the above models, although the standard curve can be obtained by fitting an ordinary linear regression, the resultant estimates of the concentration of the unknown sample are different. Thus how to select an appropriate model among the five models has been a challenge for scientists in this area. For example, Box and Hill^[4] used a Bayesian approach to determine the true model among several candidate models. Some statistical tests of hypotheses for discriminating among models were developed by Cox^[5,6] and Atkinson.^[7] Akaike^[8] suggested an information measure to discriminate among several models. Borowiak^[9] gave a comprehensive discussion on the general issues of model discrimination for nonlinear regression models.

Because models 1–4 are polynomials and model 5 can be approximated by a polynomial, Ju and Chow^[10] recommended the following ad hoc procedure for model selection using a sequential hypotheses testing approach:

Step 1: If the number of levels of standard concentration is less than 4, go to the next step; otherwise, start with the following polynomial regression model

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + e_i \quad (1)$$

Let p_{34} be the p value for testing

$$H_{034} : \beta_3 = \beta_4 = 0$$

based on the ordinary least-squares fitting of model 1. If p_{34} is greater than a predetermined level of significance, go to the next step; otherwise, go to Step 4.

Step 2: Because β_3 and β_4 are not significantly different from zero, model 1 reduces to

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + e_i \quad (2)$$

We then select a model among models 1–3. Let p_2 be the p value for testing

$$H_{02} : \beta_2 = 0$$

If p_2 is greater than a predetermined level of significance, model 3 is chosen; otherwise, go to the next step.

Step 3: If β_2 is not significantly different from zero, model 2 reduces to

$$y_i = \alpha + \beta_1 x_i + e_i$$

In this case, we select a model between models 1 and 2 by testing

$$H_{01} : \beta_1 = 0$$

If the p value for testing H_{01} is smaller than the predetermined level of significance, model 1 is chosen; otherwise, model 2 is selected.

Step 4: We select model 4 or model 5. Because models 4 and 5 have the same number of parameters, we would select the model with a smaller residual sum of squares.

Alternatively, Tse and Chow^[11] proposed a selection procedure based on the R^2 statistic and the mean-squared error of the resultant calibration estimator of the unknown sample. Their idea can be described as follows. First, fit the five candidate models using data $\{x_i, y_i, i = 1, \dots, n\}$. Then, eliminate those candidate models that fail the lack-of-fit test or significance test. For the remaining models, calculate the corresponding R^2 value, where

$$R^2 = \frac{\text{SSR}}{\text{SSR} + \text{SSE}}$$

and SSR and SSE are the sum of squares due to the regression and the sum of squares of residuals, respectively. The R^2 represents the proportion of the total variability of the responses explained by the model. It is often used as an indicator for the goodness of fit of the selected statistical model in analytical research in drug research and development.^[12] Let $R_{\max}^2$ be the maximum of the R^2 values among the remaining models. For each model, define $r = R^2/R_{\max}^2$. Eliminate those models with r value less than r_0 , where r_0 is a predetermined critical value. For the remaining models, identify the best model by comparing the mean-squared errors of the calibration estimator of the unknown concentration at x_0 based on the candidate models that would produce a given target response y_0 .

Weight Selection

For the establishment of standard curve in calibration of an instrument, larger variability is often observed in the response of a higher standard concentration preparation. If this is the case, the ordinary least-squares estimators may not be efficient. Alternatively, a weighted least-squares method is often considered to incorporate the heterogeneity of the variability. Let $Y = (y_1, \dots, y_n)'$ and $\varepsilon = (e_1, \dots, e_n)'$ under models 1–3 and $Y = (\log y_1, \dots, \log y_n)'$ and $\varepsilon = (\log e_1, \dots, \log e_n)'$ under models 4 and 5. Then, the five models described above can be written as

$$Y = X\theta + \varepsilon$$

where θ is a p vector of parameters, X is the $n \times p$ matrix whose i th row is x_i and θ and X_i for each model are given below.

Model	θ	X_i
1	$(\alpha, \beta)'$	$(1, x_i)'$
2	β	x_i
3	$(\alpha, \beta_1, \beta_2)'$	$(1, x_i, x_i^2)'$
4	$(\log \alpha, \beta)'$	$(1, \log x_i)'$
5	$(\log \alpha, \beta)'$	$(1, x_i)'$

Let W be the $n \times n$ diagonal matrix whose i th diagonal element is the weight w_i . Then, a weighted least-squares estimator of θ is

$$\hat{\theta} = (X'WX)^{-1}X'WY \quad (3)$$

Appropriate weights are usually selected so that the variance of the response at each standard concentration preparation is stabilized. Selection of an appropriate weight depends on the pattern of the standard deviation of the response at each standard concentration preparation. Let SD_x be the standard deviation of the response corresponding to the standard concentration preparation x . In practice, the following three types of weights are commonly employed for establishment of standard curve:

SD_x	Weight
σ_e	1
$\sigma_e \sqrt{x}$	$1/x$
$\sigma_e x$	$1/x^2$

where $\sigma_e > 0$ is an unknown parameter. When SD_x is a constant across all standard concentration preparations under study, no weight is necessary and W becomes the identity matrix, in which case $\hat{\theta}_w$ in formula 3 is the same as the ordinary least-squares estimator. If SD_x is proportional to $\sqrt{x}$ (or x), an appropriate choice for w_i is $1/x_i$ (or $1/x_i^2$).

For each of these three weights, the weighted least-squares estimator $\hat{\theta}_w$ is an unbiased estimator of θ with

$$\text{Var}(\hat{\theta}_w) = (X'WX)^{-1}X'W\text{Var}(\varepsilon)WX(X'WX)^{-1} \times \sqrt{b^2 - 4ac} \quad (4)$$

which reduces to $\sigma_e^2 (X'WX)^{-1}$ if the correct weight is used (e.g., $SD_x = \sigma_e \sqrt{x}$ and $w_i = 1/x_i$ is chosen).

Ju and Chow^[10] suggested to use the following model for weight selection:

$$|r_i| = \gamma_0 + \gamma_1 \sqrt{x_i} + \gamma_2 x_i + u_i, i = 1, \dots, n$$

where r_i is the i th residual from the ordinary least-squares estimation of θ and γ_j 's are unknown parameters. If we fail to reject the null hypothesis of

$$H_0 : \gamma_1 = \gamma_2 = 0$$

then no weight is necessary. When the above null hypothesis is rejected, we select the weight of $1/x^2$ if

$$SS(b_2 | b_0, b_1) > SS(b_1 | b_0, b_2)$$

where $SS(b_2 | b_0, b_1)$ is the extra sum of squares due to the inclusion of the term $\gamma_2 x$ provided that γ_0 and $\gamma_1 \sqrt{x}$ are already in the model. $SS(b_1 | b_0, b_2)$ is similarly defined. Otherwise, we select the weight of $1/x$.

Alternatively, we may select a weight function by minimizing an estimated $\text{Var}(\hat{\theta}_w)$ given by model 4 because the true variance of $\hat{\theta}_w$ is minimized when the correct weight is selected. Let $W_j, j = 0, 1, 2$, be the weight matrix W corresponding to the selection of weight functions 1, $1/x$, $1/x^2$, respectively. An approximately unbiased estimator of $\text{Var}(\hat{\theta}_{W_j})$ is

$$D_j = (X'W_jX)^{-1}X'W_j\hat{V}W_jX(X'W_jX)^{-1}$$

where $\hat{V}$ is the $n \times n$ diagonal matrix whose i th diagonal element is r_i^2 and r_i is the i th residual from the ordinary least-squares estimation. Then, the weight function corresponding to W_j is selected if $D_j = \min\{D_0, D_1, D_2\}$. It will be shown in "Calibration and Analytical Results" that choosing W by minimizing $\text{Var}(\hat{\theta}_w)$ leads to the most efficient calibration.

Note that weight selection affects the efficiency, not the bias of the estimation of θ , because any weighted least-squares estimator is unbiased as long as a correct model is chosen. On the other hand, if a wrong model is used, then all weighted least-squares estimators are biased, and they are not good estimators even if the best weights are used. Thus weight selection can be performed after a suitable model is chosen. At the model selection stage, the ordinary least-squares method can be used for simplicity.

CALIBRATION AND ANALYTICAL RESULTS

For a given unknown sample concentration x_0 with a response of y_0 , the analytical result or assay result of the unknown sample, denoted by $\hat{x}_0$, is the estimated value of x_0 obtained by finding the x value corresponding to y_0 on the established standard curve. For example, if model 1 is used, then

$$\hat{x}_0 = (y_0 - \hat{\alpha}) / \hat{\beta}$$

where $\hat{x}$ and $\hat{\beta}$ are the ordinary or weighted least-squares estimators of α and β . In general,

$$\hat{x}_0 = h(y_0, \hat{\theta}_w)$$

where $\hat{\theta}_w$ is the weighted least-squares estimator of θ given by formula 3 under one of the five models in the subsection “Weight Selection” and h is a function depending on the model. The functions h for the five models described above are given as follows:

Model	θ	$h(y_0, \theta)$
1	$(\alpha, \beta)'$	$(y_0 - \alpha) / \beta$
2	β	y_0 / β
3	$(\alpha, \beta_1, \beta_2)'$	$[-\beta_1 \pm \sqrt{\beta_1^2 - 4\beta_2(\alpha - y_0)}] / (2\beta_2)$
4	$(\log \alpha, \beta)'$	$(y_0 / \alpha)^{1/\beta}$
5	$(\log \alpha, \beta)'$	$(\log y_0 - \log \alpha) / \beta$

Note that under model 3, there may be two x values corresponding to a single y_0 . One of these x values, however, is usually outside of the study range of x and thus the other x value is $\hat{x}_0$.

In general, $\hat{x}_0$ is not exactly equal to x_0 because of two sources of variability: the variation due to the response y_0 and the statistical variation in the construction of the standard curve. The variability of an estimator is usually assessed by its bias and mean-squared error. For $\hat{x}_0$, however, its exact first and second moments may not exist because h is a nonlinear function of θ . Thus we may use asymptotic (large n) bias and mean-squared error of $\hat{x}_0$ to assess its variability. The asymptotic bias of $\hat{x}_0$ is

$$\text{bias}_{x_0} = E_{y_0}[h(y_0, \theta)] - x_0$$

Where E_{y_0} is the expectation with respect to y_0 . Under models 1 and 2, h is linear in y_0 and, hence

$$\text{bias}_{x_0} = h(E(y_0), \theta) - x_0 = 0$$

that is, $\hat{x}_0$ is asymptotically unbiased. Similarly, $\hat{x}_0$ is asymptotically unbiased under model 5. Under model 3 or 4, $\hat{x}_0$ is slightly biased because h is nonlinear in y_0 . Similarly, the asymptotic mean-squared error of $\hat{x}_0$ is

$$\text{mse}_{x_0} = E_{y_0}[h(y_0, \theta) - x_0]^2 + E_{y_0}\left[H(y_0, \theta)' \text{Var}(\hat{\theta}_w) H(y_0, \theta)\right] \quad (5)$$

where $H(y_0, \theta) = \partial h(y_0, \theta) / \partial \theta$. Under model 1, for example,

$$\text{mse}_{x_0} = \frac{1}{\beta^2} \left\{ \text{Var}(y_0) + E_{y_0} \left[\left(1, \frac{y_0 - \alpha}{\beta} \right) \text{Var}(\hat{\theta}_w) \left(1, \frac{y_0 - \alpha}{\beta} \right)' \right] \right\}$$

And in the case of constant weight function, it reduces to

$$\text{mse}_{x_0} = \frac{\sigma_e^2}{\beta^2} \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2 + \sigma_e^2 / \beta^2}{S_{xx}} \right] \quad (6)$$

where $\bar{x}$ is the average of x_i 's and $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$.

A result similar to formula 6 is given by Buonaccorsi.^[13]

Formula 5 also indicates that we should select the weight matrix W by minimizing $\text{Var}(\hat{\theta}_w)$ which yields the most efficient estimator $\hat{x}_0$. Note that formula 5 is still valid even if an incorrect weight function is used. However, this formula is not valid if a wrong model is selected.

Formula 5 or 6 can be used to determine the sample size n in establishing the standard curve. For example, Buonaccorsi and Iyer^[14] suggested three criteria for choosing n : V optimality, which concerns mse_{x_0} ; AV optimality, which concerns an average of mse_{x_0} over x_0 with respect to a distribution of x_0 ; and M optimality, which concerns the maximum of mse_{x_0} over a range of x_0 .

When there are some independent replicates at x_0 , which yield several estimates of x_0 , the asymptotic mean-squared error of $\hat{x}_0$ can be estimated. Without replicates, we can still estimate mse_{x_0} but some derivations are necessary. We illustrate this in the following example.

AN EXAMPLE

Table 1 contains 12 standard concentration preparations and their corresponding absorbances (the y responses) in a potency test of a drug product. The response y_0 corresponding to the unknown sample concentration x_0 is also included in Table 1. The results in Table 1 are in the actual order of the runs.

Ju and Chow's model and weight selection procedure described above was applied to select a model and weight function, which results in model 1 with constant variance. The standard curve is

$$y = -2.358 + 5.429x$$

with $R^2 = 0.966$ and $\hat{\sigma}_e = 3.987$. With $y_0 = 90.044$, we obtain that

$$\hat{x}_0 = 17.020$$

Table 1 Calibration data

Concentration ($\mu\text{g ml}^{-1}$)	Absorbance (area)
17.65(L)	97.485
17.65(L)	95.406
22.06(M)	121.200
22.06(M)	121.968
26.47(H)	142.346
26.47(H)	145.464
x_0	90.044
26.47(H)	141.835
26.47(H)	135.625
22.06(M)	113.814
22.06(M)	112.890
17.65(L)	89.872
17.65(L)	90.964

Note that under model 1 with constant variance, $\hat{x}_0$ is asymptotically unbiased with mse_{x_0} given by formula 6. mse_{x_0} can be estimated by replacing σ_e^2/β^2 in formula 6 by $\hat{\sigma}_e^2/\hat{\beta}^2 = 3.987^2/5.429^2 = 0.539$. Formula 6 also involves x_0 . If x_0 in formula 6 is replaced by $\hat{x}_0$ then we obtain the following estimate of mse_{x_0} :

$$\widehat{\text{mse}}_{x_0} = \frac{\hat{\sigma}_e^2}{\hat{\beta}^2} \left[1 + \frac{1}{n} + \frac{(\hat{x}_0 - \bar{x})^2 + \hat{\sigma}_e^2/\hat{\beta}^2}{S_{xx}} \right] = 0.674$$

A slightly different estimator according to Buonaccorsi^[13] is

$$\widetilde{\text{mse}}_{x_0} = \frac{\hat{\sigma}_e^2}{\hat{\beta}^2} \left[1 + \frac{1}{n} + \frac{(\hat{x}_0 - \bar{x})^2}{S_{xx}} \right] = 0.672$$

which is motivated by the fact that

$$\begin{aligned} E_{y_0} (\hat{x} - \bar{x})^2 &= E_{y_0} \left(\frac{y_0 - \hat{\alpha}}{\hat{\beta}} - \bar{x} \right)^2 \\ &= \frac{E_{y_0} (y_0 - \hat{\alpha} - \hat{\beta}\bar{x})^2}{\hat{\beta}^2} \\ &\approx \frac{E_{y_0} (y_0 - \alpha - \beta\bar{x})^2}{\beta^2} \\ &= (x_0 - \bar{x})^2 + \sigma_e^2/\beta^2 \end{aligned}$$

Note that $\widetilde{\text{mse}}_{x_0}$ is slightly more conservative than $\widehat{\text{mse}}_{x_0}$.

APPLICATION IN ASSAY VALIDATION

As indicated by the USP/NF, the validation of an analytical method or a test procedure can be carried out by assessing a set of performance characteristics or analytical

validation parameters. These validation parameters include accuracy, precision, limit of detection, limit of quantitation, selectivity, range, linearity, and ruggedness. Shah et al.^[15] recommended that an analytical method be considered validated in terms of accuracy if the mean value is within $\pm 15\%$ of the actual value. Similarly, it is also suggested that the analytical method be considered validated in terms of precision if the precision around the mean value does not exceed a 15% coefficient of variation (CV). One may use the same 15% criterion for other validation parameters, except for the limit of quantitation, where it should not deviate from the actual value by more than 20%.

In practice, the validation of an analytical method is usually carried out by the following steps: First, it is important to develop a prospective protocol that clearly states the validation design, sampling procedure, acceptance criteria for the performance characteristics to be evaluated, and how the validation is to be carried out. Second, collect the data and document the experiment, including any violations from the protocol that may occur. The data should be audited for quality assurance. The collected data are then analyzed based on appropriate statistical methods. Appropriate statistical methods are referred to as those methods that can reflect the validation design and meet the study objective. Finally, draw a conclusion regarding whether the analytical method is validated based on the statistical inference drawn about the accuracy, precision, and ruggedness of the assay results. In this section, we define each validation parameter and provide a description of standard statistical method for assessment of the validation parameter.

Accuracy

The accuracy of an analytical method or a test procedure is defined as the closeness of the result obtained by the analytical method or the test procedure to the true value. In practice, the bias is often considered as the primary measure for closeness in the assessment of assay accuracy.

In the validation of an analytical method, the accuracy of an analytical method is usually assessed through the conduct of a recovery study. In a recovery study, the analytical method is applied to samples or mixtures of excipients, which usually consist of known amounts of analyte that are added both above and below the normal levels expected in the samples. The analytical method employed recovers the active ingredients from the excipients. The recovered added amount of active ingredients are usually expressed as a percent of a label claim. A recovery study is typically conducted on three separate days. Three assay results are obtained on each day. Let y_{ij} be the assay result obtained by the analytical method when the known amount of analyte is x_{ij} , where i is the index for day. The percent recovery of x_{ij} is defined to be $z_{ij} = y_{ij}/x_{ij}$, which is assumed to follow a one-way random effects model

$$z_{ij} = \mu_z + D_i + e_{ij}, \quad i = 1, 2, 3, \quad j = 1, 2, 3 \quad (7)$$

where μ_z is an unknown parameter, D_i 's are independently distributed as $N(0, \sigma_D^2)$, e_{ij} 's are independently distributed as $N(0, \sigma_e^2)$ and D_i 's and e_{ij} 's are mutually independent. Note that σ_D^2 and σ_e^2 are known as the between-day (or day-to-day) variability and the within-day (or intraday) variability, respectively.

Using the criteria recommended by Shah et al.,^[15] the analytical method is validated in terms of accuracy if $85\% < \mu_z < 115\%$. Thus if a 95% confidence interval based on the recovery data is within the interval (85%, 115%), then the analytical method is considered validated. In practice, the largest possible bias (i.e., either the lower or upper limit of the 95% confidence interval for the mean) is often reported in assay validation for accuracy.

Because model 7 is a balanced one-way random effects model, the best unbiased estimator of μ_z is $\bar{z} = \sum_{ij} z_{ij} / 9$. A 95% confidence interval for μ_z is

$$\left(\bar{z} - t_{0.975;2} s_D / \sqrt{3}, \bar{z} + t_{0.975;2} s_D / \sqrt{3} \right)$$

where $s_D^2 = \sum_{i=1}^3 (\bar{z}_i - \bar{z})^2$, $\bar{z}_i = \sum_{j=1}^3 z_{ij} / 3$, and $t_{0.975;r}$ is the 97.5th percentile of the t distribution with r degrees of freedom.

The accuracy of an analytical method can also be assessed by a statistical analysis of the standard curve corresponding to the analytical method. For example, when the standard curve is given by a simple linear regression model, i.e.,

$$y_i = \alpha + \beta x_i + e_i$$

Bohidar^[16] and Bohidar and Peace^[17] suggested that the accuracy of an analytical method be assessed by testing the following hypotheses:

$$\begin{array}{ll} H_{01} : \beta = 0 & \text{vs. } H_{11} : \beta \neq 0. \\ H_{02} : \beta = 1 & \text{vs. } H_{12} : \beta \neq 1. \\ H_{03} : \alpha = 0 & \text{vs. } H_{13} : \alpha \neq 0. \end{array}$$

The first set of hypotheses is used to quantify a significant linear relationship between recovered (or found) and added amounts (input). This is because an analytical method cannot be validated if no relationship between found and input exists. The second set of hypotheses is used to verify whether the change in 1 unit of the added amount will result in the change in 1 unit of the recovered amount. The third set of hypotheses is to make sure that the analytical method does not recover any active ingredient if it is not added.

Precision

The precision of an analytical method or a test procedure is referred to as the degree of closeness of the result obtained by the analytical method or the test procedure to the true value. In practice, the total variability associated with the

assay result is often considered as the primary measure for the assessment of assay precision.

The assessment of assay precision can be carried out by using the data from a recovery study. Under model 7, the total variability, the between-day (or day-to-day) variability, and the within-day variability of the analytical method can be estimated. Using the method of analysis of variance (ANOVA), we obtain the following estimator of within-day variance σ_e^2 :

$$\hat{\sigma}_e^2 = \frac{1}{6} \sum_{i=1}^3 \sum_{j=1}^3 (z_{ij} - \bar{z}_i)^2$$

and the following estimator of between-day variance σ_D^2 :

$$\hat{\sigma}_D^2 = s_D^2 - \hat{\sigma}_e^2 / 3$$

Because an ANOVA estimate may lead to a negative estimate for the between-day variance σ_D^2 , the method of restricted maximum likelihood (REML), which truncates the ANOVA estimate at zero, is usually considered as an alternative to the ANOVA estimate. To avoid a negative estimate, it is recommended that the method proposed by Chow and Shao^[18] be used.

The total variability of an analytical method, $\sigma_T^2 = \sigma_D^2 + \sigma_e^2$, can be estimated by

$$\hat{\sigma}_T^2 = \frac{1}{8} \sum_{ij} (z_{ij} - \bar{z})^2$$

If an analytical method is considered validated in terms of precision when the total variability is less than a predetermined σ_0^2 , then we may claim that the analytical method is validated if $8\hat{\sigma}_T^2 / \chi_{0.95;8}^2 < \sigma_0^2$, where $\chi_{0.95;r}^2$ is the 95th percentile of the chi-square distribution with r degrees of freedom and $8\hat{\sigma}_T^2 / \chi_{0.95;8}^2$ is a 95% confidence bound for σ_T^2 .

Other Performance Characteristics

Other performance characteristics of an analytical method or a test procedure that are commonly considered in assay validation include ruggedness, specificity, linearity, range, limit of detection, and limit of quantification. These performance characteristics are briefly described below. More details can be found in Refs. [2, 16–18].

The ruggedness of an analytical method usually refers to the degree of reproducibility of the analytical results obtained by analysis of the same sample under a variety of normal test conditions such as different laboratories, different instruments, different analysis, different lots of reagents, different elapse times, different assay temperatures, and different days. Ruggedness is often used to assess the influence of uncontrollable factors or the degree of reproducibility on assay performance. Typical approaches for assessing assay ruggedness include the one-way nested random-effects model and the two-way

crossed-classification mixed model. For the assessment of assay ruggedness, it should be noted, however, that the classical analysis of variance method may produce negative estimates for the variance components and that the sum of best estimates of variance components may not be the best estimate of the total variability. In these situations, methods proposed by Chow and Shao^[19] and Chow and Tse^[20] are useful.

The specificity, which is also known as selectivity, is defined as the ability of an analytical method to accurately measure the analyte, specifically in the presence of components that may be expected to be present in the sample matrix. Specificity is often expressed as the difference in assay results obtained by analysts of samples containing added impurities, degradation products, related chemical compounds, or placebo ingredients to those from samples without added substances. Hence specificity is a measure of the degree of interference in analysis of complex sample matrix. An acceptable method should be free of any significant interference by substances known to be present in the drug product.

The linearity of an analytical method is defined as the ability to elicit assay results that are directly proportional to the concentration of analyte in samples within a given range. The range of an analytical method is the interval between the upper and lower levels (inclusive) of analyte that have been determined with accuracy, precision, and linearity using the method as written.

The limit of detection is defined as the lowest concentration of analyte in a sample that can be detected, but not necessarily quantitated, under the experimental conditions specified. In other words, LOD denotes the amount of analyte that can be reliably distinguished from background with certain assurance. Let x be the amount of analyte and y be the response. Suppose that $x = 0$ corresponds to the background and the standard curve passes through the origin. Then, the following $100(1 - a)\%$ upper confidence limit of the mean response at background,

$$LC = z_{1-a}\sigma_0$$

can be viewed as a critical limit (LC) for detecting an analyte from background, where z_{1-a} is the $(1 - a)$ th quantile of the standard normal distribution and σ_0^2 is the variance of a response at background. An amount of analyte x is considered to be detectable from the background if and only if the probability that a response at x falls below the critical limit is less than a . Thus the limit of detection, LOD, for a given analyte is the x value corresponding to $LC = z_{1-a}\sigma_D$, i.e.,

$$LOD = \frac{LC + z_{1-a}\sigma_D}{\beta} \quad (8)$$

where β is the slope of the calibration curve and σ_D is the standard deviation of the response at the detectable analyte in the sample. Alternatively, the ICH guideline

for validation of analytical procedures suggests that the concept of the signal-to-noise ratio be applied for estimating the detection limit.^[21] The concept of the signal-to-noise ratio is to compare signals obtained from samples with known low concentrations of analyte with those of blank (background) samples and establishing the minimum concentration at which the analyte can be reliably detected. The ICH guideline indicates that a signal-to-noise ratio of 3.3:1 is generally considered acceptable for establishing the detection limit. This leads to

$$LOD = \frac{3.3\sigma_D}{\beta} \quad (9)$$

where β is the slope of the calibration curve. It should be noted that formula 8 is equivalent to formula 9 if $a = 5\%$ and $\sigma_0 = \sigma_D$.

The limit of quantitation is the lowest concentration of an analyte in a sample that can be determined with acceptable precision and accuracy under the experimental condition specified. In practice, the lowest concentration of an analyte in a sample is usually defined as that for which the relative standard deviation is 10%. This leads to a signal-to-noise ratio of 0.1. As a result, the limit of quantitation can be obtained by

$$LOQ = \frac{10\sigma_D}{\beta}$$

It should be noted, however, that although a relative standard deviation of 10% is considered an acceptable standard in chemical assays in practice, 10% RSD is often unachievable in biological assays. About 20–25% variability is generally sought in binding assays, while much higher levels are more realistic in bioassay.

CONCLUSION

The validation of an analytical method or a test procedure is usually carried out by assessing a set of performance characteristics or analytical validation parameters. These validation parameters include accuracy, precision, limit of detection, limit of quantitation, selectivity, range, linearity, and ruggedness. The calibration of the instrument employed in the analytical method of test procedure plays an extremely important role to assure the accuracy and reliability of the results obtained from the analytical method or test procedure.

REFERENCES

1. FDA. *Guideline for Submitting Samples and Analytical Data for Methods Validation*; Center for Drug Evaluation and Research, The United States Food and Drug Administration: Rockville, MD, 1987.

2. USP/NF. *The United States Pharmacopeia 24 and the National Formulary 19*; The United States Pharmacopoeial Convention: Rockville, MD, 2000.
3. Chow, S.C.; Shao, J. On the difference between the classical and inverse methods of calibration. *J. R. Stat. Soc. Ser. C* **1990**, *39*, 219–228.
4. Box, G.E.P.; Hill, W.J. Discrimination among mechanistic models. *Technometrics* **1967**, *9*, 57–71.
5. Cox, D.R. Tests of Separate Families of Hypotheses. In *Proceedings of 4th Berkeley Symposium*; 1961; Vol. 1, 105–123.
6. Cox, D.R. Further results on tests of separate families of hypotheses. *J. R. Stat. Soc. Ser. B* **1962**, *24*, 406–424.
7. Atkinson, C.A. A test for discrimination between models. *Biometrika* **1969**, *56*, 337–347.
8. Akaike, H. A new look at statistical model discrimination. *IEEE Trans. Automat. Contr.* **1973**, *19*, 716–723.
9. Borowiak, D.S. *Model Discrimination for Nonlinear Regression Models*; Marcel Dekker, Inc.: New York, 1989.
10. Ju, H.L.; Chow, S.C. A procedure for weight and model selection in assay development. *J. Food Drug Anal.* **1996**, *4*, 1–12.
11. Tse, S.K.; Chow, S.C. On model selection for standard curve in assay development. *J. Biopharm. Stat.* **1995**, *5*, 285–296.
12. Anderson-Sprecher, R. Model comparison and R. *Am. Stat.* **1994**, *48*, 113–116.
13. Buonaccorsi, J.P. Design considerations for calibration. *Technometrics* **1986**, *28*, 149–156.
14. Buonaccorsi, J.P.; Iyer, H.K. Optimal designs for ratios of linear combinations in general linear model. *J. Stat. Plan. Inference* **1986**, *73*, 345–356.
15. Shah, V.P.; Midha, K.K.; Dighe, S.; McGilveray, I.J.; Skelly, J.P.; Yacobi, A.; Layoff, T.; Viswanathan, C.T.; Cook, C.E.; McDowall, R.D.; Pittman, K.A.; Spector, S. Analytical methods validation: Bioavailability, bioequivalence and pharmaceutical studies. *Pharm. Res.* **1992**, *9*, 588–592.
16. Bohidar, N.R. Statistical aspects of chemical assay validation. *Proc. Biopharm. Sect. Am. Stat. Assoc.* **1983**, 57–62.
17. Bohidar, N.R.; Peace, K. Pharmaceutical Formulation Development. In *Biopharmaceutical Statistics for Drug Development*; Peace, K., Ed.; Marcel Dekker, Inc.: New York, **1988**, 149–229. Chapter 4.
18. Chow, S.C.; Liu, J.P. *Statistical Design and Analysis in Pharmaceutical Science*; Marcel Dekker, Inc.: New York, 1995.
19. Chow, S.C.; Shao, L. A new procedure for estimation of variance components. *Stat. Probab. Lett.* **1988**, *6*, 349–355.
20. Chow, S.C.; Tse, S.K. On variance estimation in assay validation. *Stat. Med.* **1991**, *10*, 1543–1553.
21. ICH. *Validation of Analytical Procedures: Methodology*; Tripartite International Conference on Harmonization Guideline, 1996.

Canadian Health Products and Food Branch (HPFB) and Therapeutic Products Directorate (TPD)

Dongsheng Tu

NCIC Clinical Trials Group, Queen's University, Kingston, Ontario, Canada

Abstract

The objective of this entry is to provide an overview of the activities within HPFB that are related to biopharmaceutical products intended for human use with a focus on TPD.

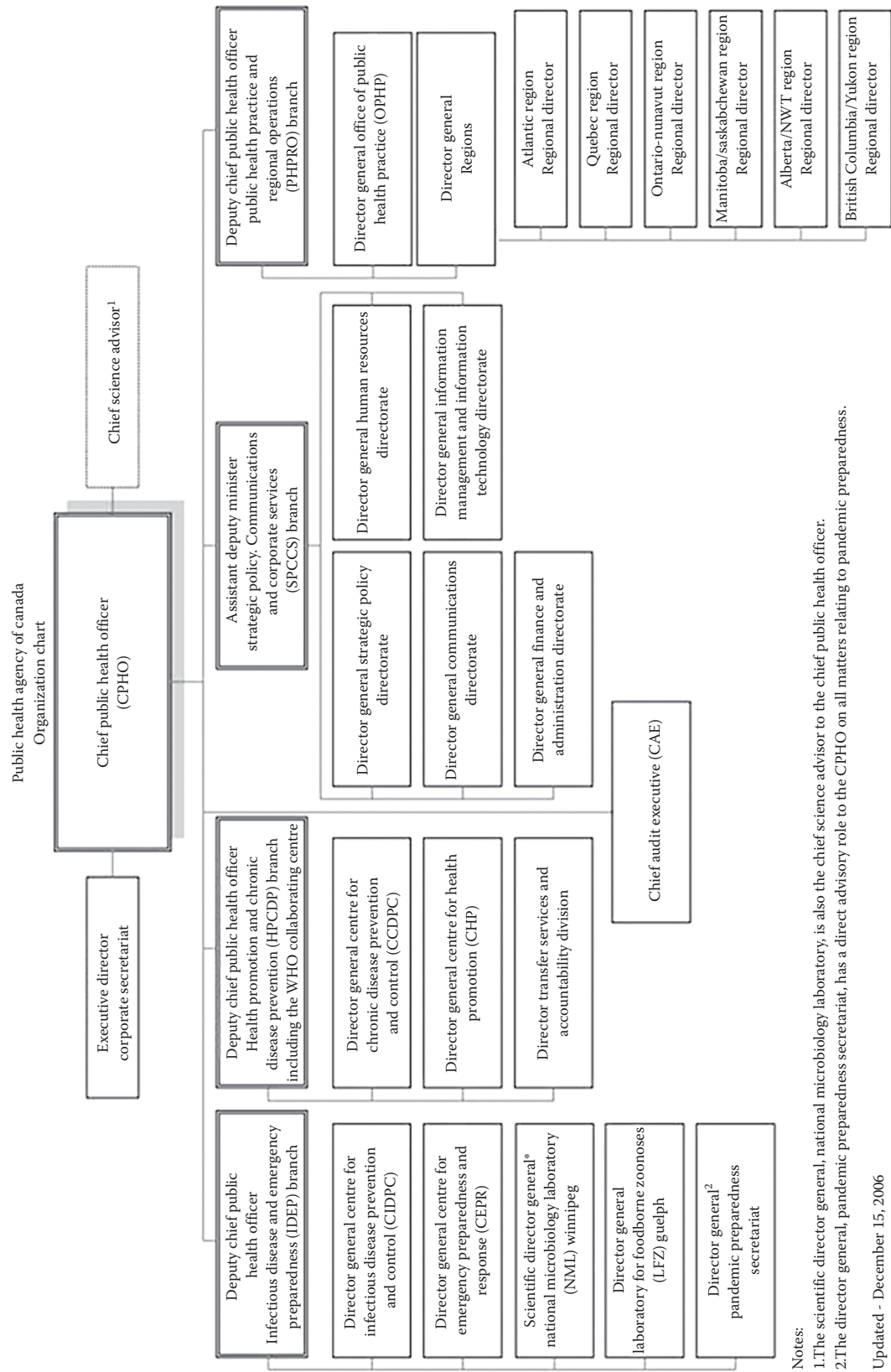
INTRODUCTION

Health Canada, the Canadian federal department with a mission to help the people of Canada maintain and improve their health, had a major organizational realignment in July of 2000. The former Health Protection Branch (HPB) was divided into several new branches. The Health Products and Food Branch (HPFB) (*) is one of these new branches; its mandate is to take an integrated approach to managing the health-related risks and benefits of health products and food by minimizing health risk factors to Canadians while maximizing the safety provided by the regulatory system for health products and food and promoting conditions that enable Canadians to make healthy choices and providing information so that they can make informed decisions about their health. To fulfill this mandate, HPFB is divided into several offices and directorates. Principal directorates in HPFB include: (i) the Biologics and Genetic Therapies Directorate (BGTD), which is responsible for the regulation of biological and radiopharmaceutical drugs, including blood and blood products, viral and bacterial vaccines, genetic therapeutic products, tissues, organs, and xenografts; (ii) the Food Directorate, which is responsible for the protection and the improvement of the health of the people of Canada through science-based policies and programs related to safe and nutritious food; (iii) the HPFB Inspectorate, which is responsible for the management of inspection, investigation, monitoring activities, and enforcement strategies related to the fabrication, packaging/labeling, testing, importation, distribution, and wholesale of regulated health products for human and veterinary use; (iv) the Natural Health Products Directorate (NHPD), which is responsible for ensuring that Canadians have ready access to health products that are safe, effective, and of high quality by maintaining proper labeling and by implementing a regulatory framework that supports freedom of choice and cultural diversity; (v) the Therapeutic Products Directorate (TPD), which is responsible for the regulation of pharmaceutical drugs, medical devices, and

other therapeutic products available to Canadians; (vi) the Veterinary Drugs Directorate (VDD), which is responsible for ensuring the safety of foods such as milk, meat, eggs, fish, and honey from animals treated with veterinary drugs and for ensuring that veterinary drugs sold in Canada are safe and effective for animals; and (vii) the Marketed Health Products Directorate (MHPD), which is responsible for post-markets surveillance of all marketed health products regulated by HPFB. The objective of this entry is to provide an overview of the activities within HPFB that are related to biopharmaceutical products intended for human use with a focus on TPD.

DRUG REGULATIONS AND GUIDANCES IN CANADA

The TPD derives its legislative authorities from a range of federal statutes. The most important one is the Food and Drugs Act, which is the regulatory base for the activities in TPD. The current Food and Drug Act provides the authority to prohibit the sale of drugs and devices that are manufactured under unsanitary conditions or that are adulterated; to prohibit the false or misleading advertising, labeling, packaging, processing, and sale of drugs and devices; and to prohibit the sale of drugs and devices that do not comply with prescribed or professed standards. According to the Food and Drug Act, a “drug” includes any substance or mixture of substances manufactured, sold, or represented for use in the diagnosis, treatment, mitigation or prevention of disease, disorder or abnormal physical state, or the symptoms thereof, in human or animal; the restoration, correction, or modification of organic functions in human or animal; or the disinfections in premises in which food is manufactured, prepared, or kept. A “device” means any article, instrument, apparatus, or contrivance, including any component, part, or accessory thereof, manufactured, sold, or represented for use in the diagnosis, treatment, mitigation, or prevention of a disease, disorder



or abnormal physical state, or the symptoms thereof, in man or animal; the restoration, correction, or modification of a body function or the body structure of man or animal; the diagnosis of pregnancy in humans or animals; or the care of humans or animals during pregnancy and at and after birth of the offspring, including care of the offspring (a contraceptive device but not a contraceptive drug is included). Therefore the range of therapeutic products encompassed by the Food and Drug Act is very broad, which includes: pharmaceuticals, including prescription and over-the-counter drugs; homeopathics, herbal remedies, and nutraceuticals; biological products, such as vaccines, blood and blood products, certain hormones and enzymes, tissues and organs, and xenografts; medical devices, such as medical and dental implants, medical equipment and instruments, test kits for diagnosis, and contraceptive devices; disinfectants; low-risk products, such as sunscreen, antiperspirants, and toothpaste; and narcotic, controlled, and restricted drugs. The Food and Drugs Act was invoked in 1920, after the Inland Revenue Act of 1876, which dealt mostly with the adulteration of alcohol but without the definition of the drugs, and Adulteration Act of 1884, which for the first time defined what was a drug, and the adulteration and conditions under which adulteration might take place.^[1,2] It was amended in 1953 so that it had control of the manufacture, distribution, and sale of drugs except narcotics. In 1977, it took under its control also the drugs listed under the Proprietary or Patent Medicine Act, which was passed in 1909 because of concerns over the efficacy and the safety of secret ingredient drugs.

The regulatory control of drugs in Canada is specified in the Regulations to the Food and Drugs Act, which have the force of law and are issued under the authority of the Food and Drugs Act. It is also the vehicle for introducing technical changes to the Act and may include a revocation of an existing regulation, the amendment of an existing regulation, or a new regulation. According to the Food and Drug Regulations, new pharmaceutical drugs and biological products (referred only as “drugs” from now on) cannot be marketed in Canada without a Notice of Compliance (NOC) from Health Canada verifying that the manufacture and the sale of the new drug are in compliance with Food and Drugs Act and Regulations. Medical devices, cosmetics, and “old” drugs must also meet departmental safety standards.

There were many amendments to the Food and Drug Act and Regulations. An important one was made in June 2001, which introduced: (i) a 30-day default review period for an application to sell a drug for the conduct of human drug clinical trials in Canada in Phases I–III of development; (ii) clear and transparent requirements to sponsors for application, information, amendments, notification, labeling, record keeping, and adverse drug reaction reporting; (iii) an inspection system against internationally accepted good clinical practice principle and good manufacturing

practices; and (iv) a clear authority to refuse an application, to suspend, or to cancel the sale of drugs for use in clinical trials in Canada where they do not meet the updated regulatory requirements.

There are also some other federal legislations and regulations that are important to TPD: for example, Patented Medicines (Notice of Compliance) Regulations require that a company wishing to sell a generic version of another company’s drug must address that company’s patents before obtaining marketing approval to sell the generic version, and the Controlled Drugs and Substances Act (CDSA), which came into effect in 1997 and replaced the Narcotic Control Act of 1961 and parts of the Food and Drugs Act, prohibits the possession, double doctoring, trafficking, possession for the purpose of trafficking, importation, exportation, and possession for the purpose of export and production of listed psychotropic substances.

Except for the Food and Drugs Act and Regulations, there are also many guidance and policy documents issued by the TPD. These documents do not have the force of the law, but reflect the thinking of TPD on some matters interesting to sponsors and provide directions to some specific issues. A detailed list of TPD guidelines and policies can be found on the website: <http://www.hc-sc.gc.ca/dhp-mps/prodpharma/legislation/index-eng.php>.

RELATIONSHIP BETWEEN THERAPEUTIC PRODUCTS DIRECTORATE AND INTERNATIONAL CONFERENCE ON HARMONIZATION

Canada, represented by the TPD, is the sole observer country in the International Conference on Harmonization of Technical Requirements for the Registration of Pharmaceuticals for Human Use (ICH), which allows the TPD to participate in discussions of both the steering committee and the expert working groups of ICH. The TPD is committed to the consultation on, and the implementation of, ICH products (guidelines and other documents) in a timely, transparent, and expeditious manner. The general guiding principle is that ICH products will be implemented “as is” (i.e., “without modification”). In the circumstances when the scope and the subject matter of TPD guidelines are not entirely consistent with those of the ICH guidances, the ICH guidances adopted by the TPD will take precedence and corresponding TPD guidelines may be amended or withdrawn, depending on the extent of inconsistencies. For example, the TPD 1989 guideline “Conduct of Clinical Investigations” was superseded by the ICH guidance E6, “Good Clinical Practice: Consolidated Guideline,” adopted by the TPD in 1997.

But in exceptional circumstances, an ICH product maybe amended by TPD, based on considerations of special (e.g., indigenous) populations or the potential impact

of difference in the practice of medicine or in risk assessment/management models, or may not be adopted if the adoption would hinder its ability to ensure the safety and health of Canadians. Sometimes, the TPD may find it necessary to clarify the principles expressed in the ICH guidance in order to provide sufficient operational detail to the industry and the stakeholders regarding technical requirements. In these circumstances, directorate-specific statements will be captured in the TPD foreword and/or addendum to the ICH guidances. For example, in the TPD foreword to ICH guidance E6, there is a TPD-specific statement that deals with the principles outlined in the section of ICH guidance E6 on the responsibilities and functions of the Institutional Review Board/Independent Ethics Committee. The statements emphasize that while ICH principles will generally be respected, “it is recognized that (Canadian) provincial/regional policies and laws may supersede them.” Also, because ICH guidance M3 recognizes the regional differences in the timing of reproduction toxicity studies to support the inclusion of women of childbearing potential in clinical trials and because there is no general principle outlined, a statement in the TPD foreword clarifies that, in Canada, the principles outlined in the TPD guideline “Inclusion of Women in Clinical Trials” should be followed. The expansion of the scope of the TPD-adopted ICH guidances beyond that expressed in these guidances is usually handled through the publication of separate, parallel TPD guidance documents. For example, TPD guidances “Identification, Qualification, and Control of Related Impurities in Existing Drugs” and “Identification, Qualification, and Control of Related Impurities in New Drugs” were developed by the TPD as companion documents to ICH guidances Q3A and Q3B, respectively, to provide further guidance on TPD’s requirements regarding the identification, qualification, and control of related impurities in “existing” and new drugs, respectively.

All the ICH guidances adopted by the TPD with its foreword can be found on the website: <http://www.hc-sc.gc.ca/dhp-mps/prodpharma/applic-demande/guide-ld/ich/index-eng.php>.

REVIEWING PROCESS AND STRUCTURE OF THE THERAPEUTIC PRODUCTS DIRECTORATE

As mentioned before, the TPD is the national authority that evaluates and monitors the safety, effectiveness, and quality of drugs, medical devices, and other therapeutic products available to Canadians. In Canada, prior to the commencement of clinical trials involving drugs intended for humans and conducted outside the approved indication, the sponsor is required to submit a Clinical Trial Application (CTA), which requests permission to distribute the drug to responsible clinical investigators who are

named in the application, to the TPD for review. Some of the information contained in a CTA includes the results from preclinical tests, production methods, dosage form, and information regarding the investigators who will be conducting the study. As mentioned before, the regulation requires the TPD to review the CTA within 30 days. If the TPD does not respond with a letter of objection within the 30-day period, the sponsor may proceed with the trial. If clinical trial results showed that the drug has potential therapeutic value that outweighs the risks associated with its use (e.g., adverse effects, toxicity), the sponsor may choose to file a New Drug Submission (NDS) with the TPD to seek the approval of marketing the drug in Canada. The NDS contains information and data about the drug’s safety, effectiveness, and quality. It includes the results of the preclinical and clinical studies, the details regarding the manufacture of the drug, packaging, and labeling details, and the information regarding therapeutic claims and side effects. After receiving the application, respective reviewing bureaus within the TPD first screen all the materials submitted for acceptability. If deficiencies are identified during the screening, a Screening Deficiency Notice identifying the deficiencies will be sent to the sponsor. Otherwise, the TPD will start performing a thorough review of submitted information, sometimes using external consultant and advisory committees, which evaluate the safety, efficacy, and quality data to assess the potential benefits and risks of the drug. The TPD also reviews the information that the sponsor proposes to provide to health care practitioners and consumers about the drug (e.g., the label, product brochure). If, upon the completion of the review, the conclusion is that the benefits outweigh the risks and that the risks can be mitigated, the drug is issued an NOC, as well as a Drug Identification Number (DIN), which permits the sponsors to market the drug in Canada and indicates the drug’s official approval in Canada. In addition, Health Canada laboratories may test certain biological products before and after the authorization to sell in Canada has been issued. This is done through its Lot Release Process, in order to monitor the safety, efficacy, and quality. If there is insufficient evidence to support the safety, efficacy, or quality claims, the TPD will not grant a marketing authorization for the drug. If this happens, the sponsor has the opportunity to supply additional information, to resubmit its submission at a later date with additional supporting data, or to appeal the TPD’s decision. The length of time for review depends on the product being submitted and the size and quality of the submission, and is influenced by the TPD’s workload and human resource.^[3] Currently, the process for the review of a drug takes an average of 18 months from the time that a submission is accepted until the TPD makes a marketing decision. For most of the NDS for the new active substance, the TPD has set 300 calendar days as the target performance standard for the review.^[4] For some

promising drug products for life-threatening or severely debilitating conditions, such as cancer, AIDS, or Parkinson's disease, for which there are few effective therapies available in the market, the TPD has a Priority Review Process in place, which allows for a faster review (with a target of 180 days). There is also a Special Access Program (SAP) administered by the TPD, which allows physicians to gain access to drugs that are not currently available in Canada. Following approval by the SPA, a physician may prescribe such a drug to specified patients, if it is the physician's belief that conventional therapies have failed or are inappropriate. The drug is only released after the TPD has determined that the need is legitimate and that a qualified physician is involved. Once a new drug is on the market, regulatory controls continue (maybe by different directorates in the HPFB such as the HPFB Inspectorate and the MHPD). The distributor of the drug must report any new information received concerning serious side effects, including the failure of the drug to produce the desired effect. The distributor must also notify the HPFB about any studies that have provided new safety information. The HPFB monitors adverse events, investigates complaints and problem reports, maintains postapproval surveillance, and manages recalls, should the necessity arise.^[5] In addition, the HPFB licenses most drug production sites and conducts regular inspections as a condition for licensing. However, certain products such as natural health and homeopathic remedies, some veterinary drugs, and vitamin and mineral supplements are not subjected to these requirements.

For a medical device that is classified in Classes II–IV based on the risk associated with their use (Class I devices presents the lowest potential risk, e.g., thermometer, and Class VI devices presents the greatest potential risk, e.g., pacemakers), a Medical Device License (MDA) is required before they can be sold. Class I devices are monitored through establishment licenses, which require the establishment to provide assurance to the TPD that regulatory requirements related to postproduction activities have been met. For a device in Class II, III, and IV, an MDA is issued after the TPD reviews the Medical Device License Application (MDLA) submitted by a sponsor and finds the information provided in MDLA as meeting the requirements in the Food and Drug Act and Medical Devices Regulations. If an MDL is not issued by the TPD, the sponsor can resubmit the application and supply additional information or appeal the TPD's decision. The length of the review varies depending on the class of the device: Class III and IV license applications have a target review time of 75 calendar days, and Class II license application have a 15-calendar day target. There is also a SAP for medical devices, which allows doctors to gain access to medical devices that have not been licensed in Canada in emergency situations or when conventional therapies have failed, are

unavailable, or are unsuitable to treatment of a patient. After medical devices are licensed, they are monitored by the HPFB to ensure their continued safety and effectiveness. The license can be suspended, or the sponsor may be requested to recall or refit a medical device if it is found to be no longer safe and effective.

To make the above process more efficient, based on several external reviews,^[2,3] the TPD has also undergone major organization changes in recent years. The current structure of TPD includes three reviewing bureaus responsible for the clinical, pre-clinical, and labeling review of drug submissions indicated for use in specific therapeutic areas (the Bureau of Metabolism, Oncology, and Reproductive Sciences; the Bureau of Gastroenterology, Infection, and Viral Diseases, which is also responsible for the evaluation of pre-market applications and the management of all issues related to non-prescription drugs; and the Bureau of Cardiology, Allergy, and Neurological Sciences) and a Medical Devices Bureau that manages and conducts clinical trial reviews, pre-market reviews, licensing, and post-market assessments of medical devices; conducts research related to medical devices; conducts pre-market and post-market laboratory investigations of medical devices; and manages the SAP for medical devices. There is also a Bureau of Pharmaceutical Science that is responsible for the chemistry and manufacturing review as well as the evaluation of clinical comparative bioavailability data for all pharmaceutical products and a Bureau of Policy, Science, and International Programs that manages and coordinates the Directorate's processes for policy development, regulatory affairs, submission performance reporting, international policy development and liaison, information dissemination, expert advice and evaluation of Directorate policies, and regulations and processes. Biostatisticians are housed in the Office of Science within the Bureau of Policy, Science, and International Programs and provide statistical support for all the activities within the directorate.

CONCLUSION

In this entry, we have provided a brief description of the regulatory foundation, reviewing process of, and structure of the HPFB and TPD. As with any other organization, the structure of the HPFB and TPD is changing frequently to better meet external and internal demands and serve Canadians. Several new initiatives are being developed to amend Food and Drug Act and Regulations and to improve the drug review process. Readers are advised to check the website of both HPFB (<http://www.hc-sc.gc.ca/ahc-asc/branch-dirgen/hpfb-dgpsa/index-eng.php>) and TPD (<http://www.hc-sc.gc.ca/dhp-mps/prodpharma/index-eng.php>) for updated information about these two organizations and these initiatives.

ACKNOWLEDGEMENTS

Minor updates have been made by the Editor-in-Chief in order to reflect developments since publication of the *Encyclopedia of Biopharmaceutical Statistics, Third Edition*.

REFERENCES

1. Hpb. *Health Protection and Drug Laws*; Health and Welfare Canada: Ottawa, 1988.
2. Gagnon, D. *Working in Partnerships... Drug Review for the Future: Review of Canadian Drug Approval System*; Health and Welfare Canada: Ottawa, 1992.
3. Rawson, N.S.B. Drug reviews in Canada: a comparison with Australia, Sweden, the United Kingdom, and the United States. *Drug Inf. J.* **1998**, 32, 1133–1141.
4. Tpd. *Policy on Management of Drug Submissions*; Therapeutic Products Directorate: Ottawa, 2001.
5. Appel, W.C. Postmarketing surveillance in Canada. *Drug Inf. J.* **1996**, 30, 655–659.

Cancer Chemotherapy: Maximum Tolerable Dose

Jen-pei Liu

Division of Biometry, Department of Agronomy, and Institute of Epidemiology, National Taiwan University, Taipei, Taiwan; and Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan

Abstract

Statistically, the maximum tolerable dose is defined as some percentile of a tolerance distribution or dose-response curve in terms of the presence or absence of dose-limiting toxicity. This entry details how MTD can be determined in accordance with patient safety.

Keywords: Continual reassessment method; Dose limiting toxicity; Single-stage design; Two-stage design.

INTRODUCTION

The primary scientific objective of the evaluation of new chemotherapeutic agents in cancer patients during the Phase I clinical development is to employ an efficient, reliable, yet practical dose-finding design to search the highest dose with an acceptable and manageable safety profile for use in subsequent Phase II trials. This dose with an acceptable and manageable safety profile is usually referred to as the maximum tolerable dose (MTD). The unacceptable or unmanageable safety profile is generally called the dose-limiting toxicity (DLT), which is predefined by some criteria such as grade 3 or greater hematological toxicity according to the U.S. National Cancer Institute's (NCI) Common Toxicity Criteria (CTC). In short, the MTD is the highest possible but still tolerable dose with respect to some pre-specified dose limiting toxicity.^[1,2]

DEFINITION

Statistically, the MTD is defined as some percentile of a tolerance distribution or dose-response curve in terms of the presence or absence of DLT. In other words, the MTD is defined as the dose where a specified proportion of patients, say p_0 , experience DLT. Storer^[3] indicated that the value of p_0 is usually in the range from 0.1 to 0.4. Let Y be the binary response such that $Y = 1$ denote the occurrence of a predefined DLT and $\{d_i, i = 1, 2, \dots, I\}$ be a set of fixed dose levels. The relationship between the occurrence of the DLT and dose level d can be described by the following model:

$$h[P(x, \theta)] = \alpha + \beta x$$

where $h(\cdot)$ is some function of the probability of the occurrence of the DLT, x is the dose level taking one of the values d_i , and $\theta = (\alpha, \beta)'$.

The MTD is then defined as

$$x_m = (k_p - \alpha) / \beta,$$

where $k_p = h(p_0)$.

One of the commonly used functions is the logistic function. It follows that $h(p_0) = \text{logit}(p_0) = \ln[p_0/(1 - p_0)]$, where $\ln$ denotes the natural logarithm. On the other hand, the dose levels employed in phase I cancer trials for determination of the MTD are in general derived from information on animal studies. Storer^[3] pointed out that a commonly employed starting dose level is from one-tenth to one-third of the mouse LD10. In addition, the dose levels are usually selected to be approximately equally spaced on the logarithmic scale. Schneiderman,^[4] for example, suggested the use of the modified Fibonacci sequence of the diminishing multipliers of $\{2, 1.67, 1.5, 1.4, 1.33, \dots\}$.

DESIGN AND ANALYSIS

General Considerations

The primary objective of Phase I cancer trials is to establish the MTD with an adequate precision. The following considerations are important for the selection of an appropriate design for estimation of the MTD:

1. The patients are critically ill. Some of them are even in the terminal stage of the disease and the test chemotherapeutic agent may be the last hope for the patients.

2. The number of patients available for phase I cancer trials is relatively small.
3. The patient population is usually rather heterogeneous because phase I cancer trials might enroll terminal patients with different cancers.
4. Phase I cancer trials can be viewed as a screening process where cytotoxic agents with a tolerable safety profile are selected and their MTDs are determined with a minimal number of patients in a minimal amount of time.
5. Most chemotherapeutic agents generally can induce serious, irreversible, and life-threatening toxicity. Phase I cancer trials usually establish the MTD from below. In fact, regulatory agencies sometimes dictate the dose for the first patient.

Designs for phase I cancer trials generally can be classified into three categories: single-stage, two-stage, and Bayesian design. The single-stage and two-stage designs are up-and-down designs where the doses are adjusted either upward (escalation) or downward (de-escalation) within the pre-specified set of the fixed dose levels. The Bayesian approach chooses the dose level for the next patients by minimizing some measure based on the difference between the current estimate of probability for the occurrence of DLT and p_0 .

Single-Stage Design

There are three commonly employed single-stage designs for phase I cancer trials.^[1,5,6]

Design A—Standard Dose-Escalation Design

- Step 1: A group of three patients are treated at the initial dose d_1 .
- Step 2: If no pre-specified DLT is observed in all three patients, the dose for the next group of three patients is escalated to the next higher dose level. Otherwise, the next group of three patients is treated at the same dose level.
- Step 3: The dose of the next group of three patients is escalated to the next higher dose level if the pre-specified DLT is observed in at most one patient of the six patients from both Steps 1 and 2. Otherwise the trial stops.
- Step 4: Steps 2 and 3 are repeated with two consecutive groups of three patients until the trial stops.

A flowchart for Design A is given in Fig. 1.

Traditionally, if the study stops at dose level d_i , then the MTD is estimated as the next lower dose level d_{i-1} . If the trial stops at the dose level where only three patients were treated, then additional three patients should be enrolled at the next lower dose level for a total of six patients. Therefore, the MTD is the highest dose where at most one of six patients developed DLT.

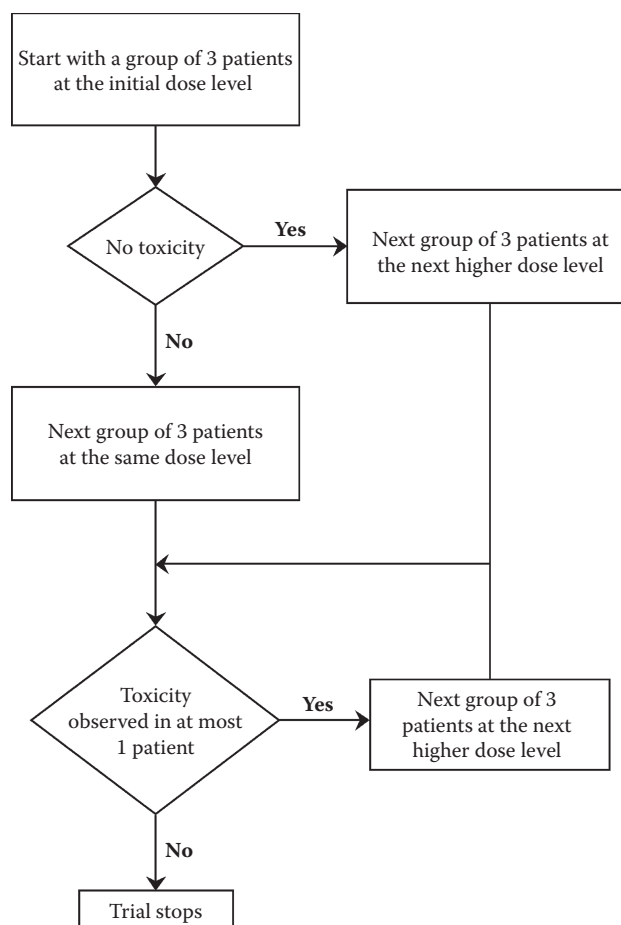


Fig. 1 Flowchart for design A

It can be seen that the standard dose-escalation design only allows dose level to be escalated upward. Consequently, the starting dose level is the lowest dose level d_1 , and hence many patients are treated at doses well below the therapeutically meaningful level. Therefore, the MTD obtained from this design may be too conservative. Another drawback of this design is that because too many patients are treated at the lower dose levels, it might take a long time for the MTD to be reached. To overcome these two shortcomings of the traditional design, Storer^[5] proposed the following three designs to allow both escalation and de-escalation and to eliminate the need to use the lowest dose level as the starting dose.

Design B

- Step 1: A single patient starts at a preselected dose level.
- Step 2: If no pre-specified DLT is observed, then the next patient is treated at the same dose level. Otherwise, the next patient is treated at the next lower dose level.
- Step 3: If no pre-specified DLT occurs in both consecutive patients, then the next patient is treated at the next higher dose level. If the DLT is

observed in both consecutive patients, then the dose of the next patient is decreased to the next lower dose level. Otherwise, the trial stops.

Step 4: Repeat Steps 2 and 3 until the trial stops.

A flowchart for Design B is given in Fig. 2.

Design C

Design C is similar to Design B except that the dose level is decreased to the next dose level if the pre-specified DLT is observed in one patient, and escalation to the next higher dose level occurs only when no DLT is observed in two consecutive patients.

Design D

- Step 1: A group of three patients are treated at the initial dose $d_i, i = 1, 2, \dots, I$.
- Step 2: The dose of the next group of three patients is escalated to the next higher dose level if no DLT is observed in all three patients, or stays at the same dose level if the pre-specified DLT is observed in one patient, or decreases to the next

lower level if the DLT occurs in more than one patient.

Step 3: The trial stops when the pre-specified number of patients has completed the study.

A flowchart for Design D is given in Fig. 3.

In a simulation study, Storer^[5] reported that none of the single-stage designs performs well in an arbitrary dose-response setting with fixed sample sizes. In addition, the standard dose-escalation design even frequently failed to provide a convergent estimate for MTD. As a result, Storer^[1,5] and Simon et al.^[7] have considered a variety of combinations of traditional single-stage designs to achieve the following objectives:

- 1. To address the flaws of the standard dose-escalation design by minimizing the number of patients treated at subtherapeutic dose levels.
- 2. To concentrate sampling around the MTD.
- 3. To reduce the duration of the study.
- 4. To obtain information about interpatient variability and cumulative toxicity.

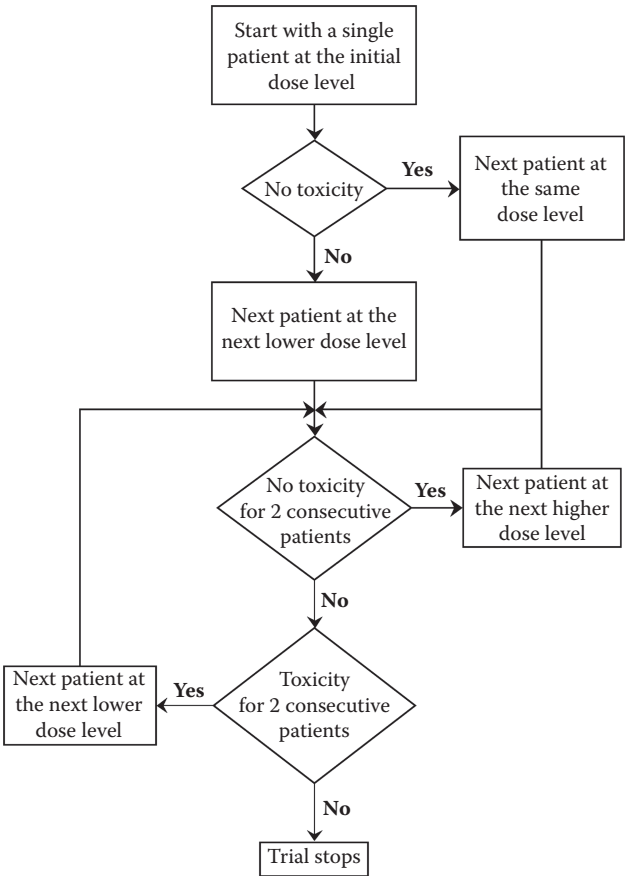


Fig. 2 Flowchart for design B

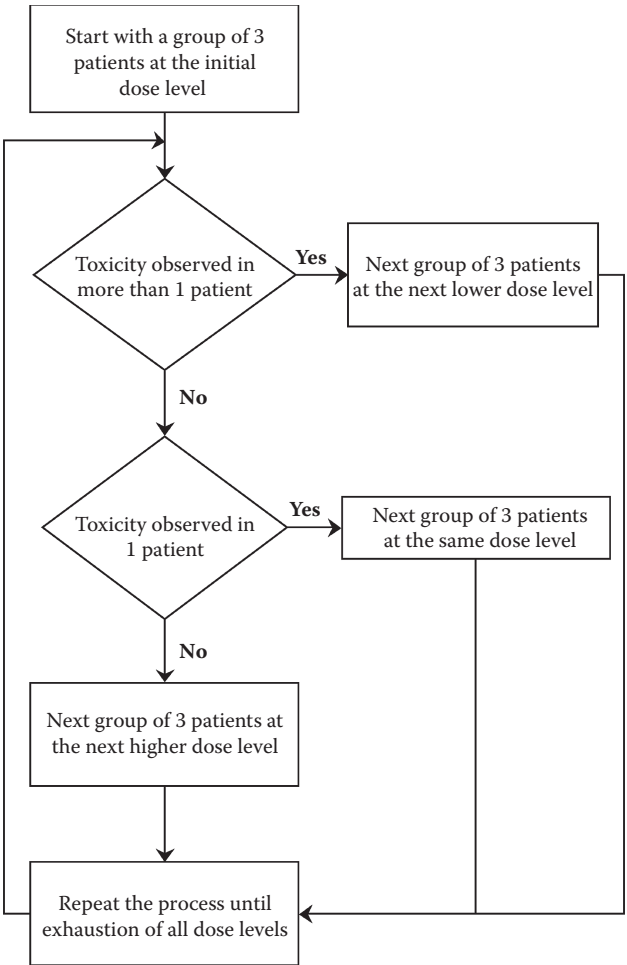


Fig. 3 Flowchart for design D

Two-Stage Design

Design BD

- Stage 1: Employ Design B until the first patient experiences DLT. Then start de-escalation with the next lower dose level until the first patient does not experience DLT.
- Stage 2: Continue the trial with Design D where the initial dose is the dose level at which the first patient does not experience DLT at Stage 1.^[5,6]

Accelerated Titration Designs

For the accelerated titration designs (ATDs), Simon et al.^[7] further grouped the toxicity grades based on the NCI's CTC into two classes: moderate toxicity for grade 0–2 and DLT for grade 3 and above. In addition, the patients will receive at least three courses of treatment in the accelerated titration designs. This type of designs consists of an initial accelerated stage and a standard dose-escalation stage.

Accelerated titration design 1 (ATD1). Initial Accelerated Stage: Cohorts of one new patient per dose level, starting at a preselected initial dose level d_1 . Dose escalation stops and reverts to the standard dose-escalation design when the first instance of the DLT is observed at the first course in one patient or two instances of moderate toxicities at the first course are observed. Dose escalation uses 40% dose-step increments. Standard Dose-Escalation Stage: Expand the cohort for current dose level to three patients and continue the trial using the traditional escalation design in a cohort of three patients. The intrapatient dose escalation is allowed if grade 0–1 toxicity is observed in the previous trial. The dose is de-escalated if grade 3 toxicity or above occurred in the previous course. Otherwise, stay at the same dose level.

Accelerated titration design 2 (ATD2). ATD2 is the same as ATD1 except that 100% dose steps are used for the initial accelerated stage.

Accelerated titration design 3 (ATD3). This design is the same as ATD2 except for the switch to the standard dose-escalation design when the first instance of the DLT is observed at any course in any patient.

Continual Reassessment Method

As mentioned above, the patients in cancer phase I trials are those with terminal cancer and at high risk of death. The characteristics of the chemotherapeutic agents evaluated in cancer phase I trials are the following: (i) they produce fatal toxicity at higher dose level; (ii) they yield no effectiveness at lower dose levels; and (iii) except for

some scarce animal data, no information about the dosing range is available. The standard dose-escalation design and one-stage or two-stage up-and-down designs update the dose for the next patient based on information on the occurrence of DLT from the current and previous patients. These designs, however, fail quantitatively using a model to combine prior information about the MTD and the information from the patients collected in the Phase I trial. To overcome these drawbacks, O'Quigley et al.^[8] have proposed the continual reassessment method (CRM) to use the Bayesian approach to estimate the MTD.

Recall that the MTD is the dose associated with the $100p_0$ percentile of the dose-response relationship with respect to the occurrence of a predefined DLT. Let $p_1, p_2, \dots, p_l$ be the initial guesses of probabilities corresponding to the set of fixed dose levels $d_1, d_2, \dots, d_l$, which are assumed to be monotonically increasing with the dose level. The dose-response relationship can then be characterized by a single-parameter, logistic regression model

$$\text{logit}[P(x_i, \beta)] = \alpha + \beta x_i$$

where the intercept α is assumed fixed, say 3,^[9–11] the slope β is to be estimated, and x_i is updated after the information of the occurrence of the DLT for the current patient becomes available using $x_i = \text{logit}^{-1}[P(x_p, \beta)]$.

- Step 1: Determine the dose-response model to be employed in the study and select a set of fixed dose levels with their corresponding occurrence probabilities. Finally, choose a fixed sample size n , say 18–24.
- Step 2: Select a prior distribution $g(\beta)$ that reflects the current brief of the investigator about the dose-response relationship. Determine the dose for the first patient as the dose level that produces the prior probability of DLT closest to the targeted probability p_0 .
- Step 3: After the result of the occurrence of DLT for the current patient at dose level x_{i-1} becomes available, obtain the Bayesian estimate of β with respect to quadratic error loss function as

$$\beta_i = E(\beta | \Lambda_{i-1}) = \int \beta f(\beta | \Lambda_{i-1}) d\beta,$$

where $f(\beta | \Lambda_{i-1})$ is the posterior density,

$$f(\beta | \Lambda_{i-1}) = q_{\Lambda_{i-1}}(\beta) g(\beta) / \int q_{\Lambda_{i-1}}(z) g(z) dz,$$

$q_{\Lambda_{i-1}}(\beta)$ is the likelihood function such that

$$q_{\Lambda_{i-1}}(\beta) = \prod [P(x_j, \beta)]^{y_j} [1 - P(x_j, \beta)]^{1-y_j}$$

and $\Lambda_{i-1} = \{(x_1, y_1), \dots, (x_{i-1}, y_{i-1})\}$, the cumulative data up to the patient receiving dose level x_{i-1} .

Step 4: Determine the dose level for the next patient by the dose level that minimizes the absolute loss function $|P(x_i, \beta_i) - p_0|$ with $x_i = \text{logit}^{-1}[P(x_i, \beta_i)]$, $i = 1, 2, \dots, I$.

Step 5: Repeat step 4 until the result of the last patient is known.

Step 6: The MTD x_m is estimated as the dose level for the hypothetical $n + 1$ patient. The corresponding probability for the occurrence of DLT can then be estimated as $P(x_m, \beta_L)$ with the corresponding $100(1 - \alpha)\%$ credibility interval given as $[P(x_m, \beta_L), P(x_m, \beta_U)]$, where (β_L, β_U) is the $100(1 - \alpha)\%$ credibility interval based on the posterior probability $f(\beta|\Lambda_m)$.

The CRM presents a revolutionary approach to estimating the MTD as compared to the traditional up-and-down method. It estimates the MTD from a continuous spectrum of doses whereas the other methods choose their MTD from a discrete set of pre-specified dose levels. In addition, the CRM utilizes a statistical model to synthesize the prior belief about the MTD and cumulative information from the trial for selection of the dose level for the next patient. However, despite these advantages, as pointed out by Korn et al.,^[2] Goodman et al.,^[9] and Ahn,^[10] the original CRM is not free of problems in its implementation for estimation of the MTD in cancer Phase I trials. First, the dose level for the next patient can be determined only after the result of the DLT for the current patient becomes available. On the other hand, the CRM only uses a cohort size of one patient for the dose adjustment. Consequently, it takes much longer time to complete the trial than the traditional dose-escalation design and most of other up-and-down designs. Secondly, the CRM might start the trial with an initial dose above the lowest dose that is often one tenth the LD10 in mice. This possibility of the initial dose above the lowest dose level for the first patient makes many clinicians uncomfortable and reluctant to implement the CRM. Another special feature of the CRM that also reduces its acceptance is the possibility that dose can be escalated for more than one dose level. This can result in that more patients may be treated at high dose levels. Møller^[12] demonstrated that the second dose can be allocated at dose level 10 even when the first patient has no DLT at the lowest dose level 1 as the initial dose.

Goodman et al.,^[9] Ahn,^[10] Ishizuka and Ohashi,^[11] O'Quigley and Shen,^[13] Heyd and Carlin,^[14] Thall et al.,^[15] and Zohar and Chevret^[16] have suggested various ways to overcome the above-mentioned shortcomings of the CRM. These improvements can be summarized as below:

1. The trial always starts with the lowest dose level as the initial dose for the first patient.
2. The size of cohorts increases to three patients.
3. The magnitude of dose escalation is limited to one dose level only between two adjacent cohorts.

4. Different accrual strategies with delayed patient outcomes are used.
5. The likelihood approach is used.
6. Different stopping rules are employed.

In addition, the CRM was modified to incorporate other information and to achieve other objectives. Ishizuka and Ohashi^[11] and Piantadosi^[17] illustrated methods to incorporate the additional pharmacokinetic data into the CRM. Legedza and Ibrahim^[18] proposed a longitudinal design for Phase I cancer trials based on the CRM to investigate the cumulative toxicity of chemotherapeutic agents. Kramar et al.^[19] applied the CRM to a combination of two drugs. Gasparini and Eisele^[20] suggest the use of the product-of-beta prior for the curve-free application of the CRM. Recently, Thall et al.,^[21] through the framework of the CRM, have developed a method for dose finding on feasibility and toxicity in T-cell infusion trials. Piantadosi et al.^[22] and Dougherty et al.^[23] have provided excellent illustrations for practical implementation of the CRM with real examples. Since the publication of the original CRM by O'Quigley et al.,^[8] numerous simulations have been performed to empirically compare the traditional dose-escalation design, the up-and-down designs, and the original CRM and its various modifications. The result can be found in Korn et al.,^[2] Storer,^[6] Goodman et al.,^[9] Ahn,^[10] O'Quigley,^[24] and Korn et al.^[25] However, the comparison between the accelerated titration design and the CRM deserves further investigation.

CONCLUSION

Although the standard dose-escalation design and most of up-and-down designs are implicitly based on a dose-response model, they all select the MTD operationally without actually modeling the relationship between the occurrence of DLT and dose based on the data collected in the trial. The reason for failure of the traditional likelihood approaches such as the delta method, Fieller's theorem, and likelihood ratio method to provide an adequate estimate of the MTD and its corresponding confidence interval is that they are unable to take into account the adaptive and sequential nature of a discrete Markov chain for these designs.^[1] Storer^[1,5,6] has proposed a two-parameter, logistic regression mode to fit the data collected from Design BD. On the other hand, as mentioned above, the CRM employs a single-parameter model to fit the data after the completion of the study. It should be noted that for the CRM, the model for updating the dose levels will be different from the model for the dose-response model to describe the relationship between the DLT and dose. In addition, Shen and O'Quigley^[26] have shown that the CRM will provide consistent estimates of the MTD even under model misspecification. Most recently, Hather and Mackey^[27] have provided some notable properties

of the standard dose-escalation design. A special issue of the *Journal of Biopharmaceutical Statistics* summarizes the recent development and challenges for estimation of MTD in early phase cancer clinical trials.^[28]

REFERENCES

1. Storer, B.E. Small-sample confidence sets for the MTD in a Phase I clinical trials. *Biometrics* **1993**, *49*, 1117–1125.
2. Korn, E.L.; Midthune, D.; Chen, T.T.; Rubinstein, L.V.; Christian, M.C.; Simon, R. A comparison of two phase I designs. *Stat. Med.* **1994**, *13*, 1799–1806.
3. Storer, B.E. Phase I Trials. In *Encyclopedia of Biostatistics*; Armitage, P.; Colton, T.; Eds.; John Wiley and Sons: New York, 1997, 3365–3370.
4. Schneiderman, M.A. Mouse to Man: Statistical Problems in Bring a Drug to Clinical Trial. In *Proc. of Fifth Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, CA, 1967; Vol. 4.
5. Storer, B.E. Design and analysis of phase I trials. *Biometrics* **1989**, *45*, 925–937.
6. Storer, B.E. An evaluation of phase I clinical trial designs in the continuous dose-response setting. *Stat. Med.* **2001**, *20*, 2399–2408.
7. Simon, R.; Freidlin, B.; Rubinstein, L.; Arbuck, S.G.; Collins, J.; Christian, M.C. Accelerated titration designs for phase I trials in Oncology. *J. Natl. Cancer Inst.* **1997**, *89*, 1138–1147.
8. O'Quigley, J.; Pepe, M.; Fisher, L. Continual reassessment method: A practical design for Phase I clinical trials in cancer. *Biometrics* **1990**, *46*, 33–48.
9. Goodman, S.N.; Zahurak, M.L.; Piantadosi, S. Some practical improvements in the continual reassessment method for phase I studies. *Stat. Med.* **1995**, *14*, 1149–1161.
10. Ahn, C. An evaluation of phase I cancer clinical trial design. *Stat. Med.* **1998**, *17*, 1537–1549.
11. Ishizuka, N.; Ohashi, Y. The continual reassessment method and its applications: A Bayesian methodology for phase I cancer clinical trials. *Stat. Med.* **2001**, *20*, 2661–2681.
12. Møller, S. An extension of the continual reassessment methods using a preliminary up-and-down design in a dosing finding study in cancer patients in order to investigate a greater range of doses. *Stat. Med.* **1995**, *14*, 911–922.
13. O'Quigley, J.; Shen, L.Z. Continual reassessment method: A likelihood approach. *Biometrics* **1996**, *52*, 673–684.
14. Heyd, J.M.; Carlin, B.P. Adaptive design improvement in the continual reassessment method for phase I studies. *Stat. Med.* **1999**, *18*, 1307–1321.
15. Thall, P.E.; Lee, J.J.; Tseng, C.H.; Estey, E.H. Accrual strategies for phase I trials with delayed patient outcome. *Stat. Med.* **1999**, *18*, 1155–1169.
16. Zohar, S.; Chevret, S. The continual reassessment method: Comparison of Bayesian stopping rules for dose-ranging studies. *Stat. Med.* **2001**, *20*, 2827–2843.
17. Piantadosi, S.; Liu, G.H. Improved designs for dose escalation studies using pharmacokinetic measurements. *Stat. Med.* **1996**, *15*, 1605–1618.
18. Legedza, A.T.R.; Ibrahim, J.G. Longitudinal design for phase I clinical trials using the continual reassessment method. *Control Clin. Trials* **2000**, *21*, 574–588.
19. Kramar, A.; Lebecq, A.; Candalh, E. Continual reassessment methods in phase I trials of the combination of two drugs in oncology. *Stat. Med.* **1999**, *18*, 1849–1864.
20. Gasparini, M.; Eisele, J. A curve-free method for phase I clinical trials. *Biometrics* **2000**, *56*, 609–6015.
21. Thall, P.E.; Sung, H.G.; Choudhury, A. Dose-finding based on feasibility and toxicity in T-cell infusion trials. *Biometrics* **2001**, *57*, 914–921.
22. Piantadosi, S.; Fisher, J.D.; Grossman, S. Practical implementation of a modified continual reassessment method for dose-finding trials. *Cancer Chemother. Pharmacol.* **1998**, *41*, 429–436.
23. Dougherty, T.B.; Prosche, V.; Thall, P.F. Maximal tolerable dose of nalmefene in patients receiving epidural fentanyl and dilute bupivacaine for postoperative analgesia. *Anesthesiology* **1999**, *92*, 1010–1016.
24. O'Quigley, J. Another look at two phase I clinical trial designs. *Stat. Med.* **1999**, *18*, 2683–2690.
25. Korn, E.L.; Midthune, D.; Chen, T.T.; Rubinstein, L.V.; Christian, M.C.; Simon, R. Commentary. *Stat. Med.* **1999**, *18*, 2691–2692.
26. Shen, L.Z.; O'Quigley, J. Consistency of continual reassessment method under model misspecification. *Biometrika* **1996**, *83*, 395–405.
27. Hather, G.J.; Mackey, H. Some notable properties of the standard oncology phase I designs. *J. Biopharm. Stat.* **2009**, *19*, 543–555.
28. Special Issue on Statistical Issues in Early Phase Cancer Clinical Trials, Special editors: Hyslop, T.; Jung, S.H. *J. Biopharm. Stat.* **2009**, *19* (3).

Cancer Trials

Clet Niyikiza and Douglas E. Faries

Eli Lilly and Company, Indianapolis, Indiana, U.S.A.

Abstract

This review presents the major statistical issues encountered in cancer clinical trials, summarizes the standard approaches, and presents recent proposals regarding the design and analysis of cancer trials.

INTRODUCTION

Owing to the severity of many types of cancer, the critical need for improvement in therapies, and the large amount of research in this area, efficient designs and analyses for cancer clinical trials are of great importance. This review presents the major statistical issues encountered in cancer clinical trials, summarizes the standard approaches, and presents recent proposals regarding the design and analysis of cancer trials. The review is organized by clinical trial phases, followed by a discussion of statistical analysis methods and efficacy endpoints in cancer trials.

PHASE I CLINICAL TRIALS

Introduction

The primary objectives of phase I cancer trials are to estimate the maximum tolerated dose (MTD) of the compound using a specific schedule and then to select a dose level for subsequent, phase II trials. In addition, the trials provide information regarding the types and severities of toxicity, the pharmacology of the compound, and initial data on efficacy. Because of the toxicities of cancer compounds, participants of phase I cancer trials are typically patients with advanced stages of the disease, patients for whom alternative treatment options have been exhausted, rather than normal volunteers. Also, the dose level for efficacy is often limited by toxicity. The need to avoid toxicity and yet treat patients at a potentially therapeutic dose presents a challenge in the design of phase I trials. Multiple phase I trials are typically completed for a given compound in order to investigate effects from various dose and schedule combinations or to investigate combinations of treatments.

The Traditional Design

Selection of Dose Levels

The initial dose level for phase I trials is typically 1/10 of the LD₁₀ in mice, expressed in mg/m² units, although it may be smaller, based on toxicities observed in other species. This dose is expected to be a conservative dose, which will not cause any significant toxicity. Additional dose levels, based upon the modified Fibonacci series, are in increments of 2.0, 1.67, 1.5, 1.4, 1.33 (and 1.33 thereafter) times the preceding dose.^[1–3]

Dose Escalation Rule

Prior to the trial, specific criteria for determining “significant toxicity” are selected. Typically, this is based on toxicity grades, such as toxicity grade 3 or worse. The traditional design^[2,4,5] begins by assigning three patients to an initial dose level. If no significant toxicity is observed in any of the three patients treated at a particular dose level, then the next three patients are treated at the next higher dose level. If a toxicity is observed in one of three patients at a particular dose level, then three additional patients are assigned to the same dose level. If none of the additional three patients experience significant toxicity, then the next three patients are assigned to the next higher dose level. Once two patients experience significant toxicities at a particular dose level, the trial ends. That dose level, or the dose level below, is selected as the MTD. Because this is a sequential procedure, the sample size varies. Typically, at most, 25 patients are included in a phase I trial, although if the starting dose is too conservative, this number could be higher.

Issues with the Traditional Design

While the traditional approach has been the standard for many years, it has been increasingly criticized for multiple

reasons.^[6–8] If toxicity in many patients is proportional to that observed in animal studies, then seven dose escalations are required to reach the MTD using the modified Fibonacci series approach. A review of past trials^[7] indicated high variability in the number of dose escalations needed, and several trials requiring at least 12 escalations were noted. The result of a large number of escalations is that a large number of patients will receive a subtherapeutic dose level. In addition, the length of the study and the extended use of limited research resources results in a delay in bringing new compounds to the market and a reduction in the number of new compounds that can be tested in man. The final estimate of the MTD from a traditional design also has poor statistical properties. One has little information regarding the proportion of patients who would have a toxicity at the traditional design estimated MTD. It does not estimate any particular percentile of the dose–toxicity function and is dependent upon the starting dose, the true dose–toxicity function, and the number of dose levels examined. In addition, no accounting for sampling error is taken into account when estimating the MTD. Poor estimates of the MTD can lead to poor selection of doses for patients in subsequent trials, resulting in delays or potential discontinuation of promising compounds. For these reasons, several novel designs for phase I cancer trials have been proposed over the last 12 years.

New Designs

Pharmacologically Guided Designs

In pharmacologically guided designs,^[6,7,9] dose escalations are based upon pharmacokinetic analyses. The goal of these designs is to quickly escalate patients to achieve a concentration similar to the target concentration based on animal studies. The underlying premises are that toxicity is correlated with concentrations of drug in plasma and that the relationship between toxicity and drug exposure holds across species. A review of previous studies indicated that the relationship between drug concentrations and toxicity from mouse to man is more consistent across compounds than the relationship between doses and toxicity. Retrospective analyses demonstrated a significant savings in numbers of patients and time to complete phase I trials as compared to the traditional approach. To utilize such a procedure, one must have a sensitive drug assay, and interspecies pharmacodynamic differences must be small. Results of trials using pharmacologically guided designs are included in Refs.^[10,11]

Modified Up-and-Down Designs

Several procedures, based upon the original up-and-down procedure,^[12] have been proposed.^[13] These procedures address the issues of treating too many patients at subtherapeutic doses and the lengthy time to complete traditional

trials by allowing for more flexibility in the number of patients required prior to dose escalation and in the increments between dose levels and by allowing decreases in dose levels.

Proposals have included two-stage designs. Two-stage designs incorporate an accelerated escalation scheme until the first significant toxicity is observed, and then return to a more conventional escalation scheme. Proposed accelerated escalation schemes include keeping 100% increments between dose levels and/or allowing an increase in dose after a single nontoxic response or after two nontoxic responses. The MTD is defined as a specified percentile of the dose–toxicity function; once the trial is complete, the MTD may be estimated by fitting a logistic dose–toxicity model. Simulation studies have demonstrated that these designs can reduce the number of patients treated at subtherapeutic levels, reduce the time needed to complete trials, and provide improved estimates of the MTD as compared to traditional trials. The main concern with these approaches is the increased risk of treating patients at highly toxic dose levels, although simulation studies suggest this increase is limited.

Recently, a two-stage up-and-down design that incorporates data from multiple courses of treatment as well as inpatient dose escalation has been proposed.^[13] In this approach, an accelerated escalation in stage I is discontinued when any significant toxicity is observed (any course for any patient) or when the second instance of moderate toxicity is observed. Simulations indicate that this approach does not increase the chance of administering a highly toxic dose, but has lesser gains in reduction of patients at subtherapeutic levels and time to study completion than other up-and-down designs.

Continual Reassessment Method and Modifications

The continual reassessment method (CRM)^[8] utilizes a Bayesian approach to estimate the MTD after each successive patient. The MTD is first defined as a percentile of the dose–toxicity function. In addition, a prior distribution for the MTD and a dose–toxicity model is selected based upon prior data and/or expert opinion (or a noninformative prior may be selected). After each successive patient, an updated estimate of the MTD is obtained by computing the posterior distribution of the MTD. Each patient is then assigned to the dose level closest to the current estimate of the MTD. With this approach, each patient is treated close to the estimate, based on all currently available information, of the MTD. Thus a reduced number of patients are treated at subtherapeutic levels. In addition, simulations have demonstrated a reduced number of patients needed to complete the trial as well as a more efficient estimate of the MTD.^[8,14] As with the accelerated up-and-down designs, concerns have been raised that the CRM will result in treating more patients with toxic dose levels. This is especially

true with the CRM, because the proposed starting dose is the prior estimate of the MTD, not necessarily a conservative dose level. Another criticism with the CRM is the need to obtain a prior distribution, and the added complexity in computations that make implementation difficult without statistical support.

Several modifications of the CRM have recently been proposed.^[15–18] These proposals attempt to retain many of the advantages of the CRM while limiting the potential for administering toxic doses. Modifications to the CRM include starting at a conservative dose level as in the traditional approach, limiting dose escalations to those allowed in the traditional approach, always using the closest dose level lower than the current MTD estimate, and using an initial up-and-down procedure prior to the CRM. The references provide one example of a recent completed trial using a modified version of the CRM.^[19] To address the criticism with the prior distribution and the complexity of the Bayesian computations, recent literature proposes the use of maximum-likelihood estimation.^[20,21] To obtain the updated estimates of the MTD after each successive patient, maximum-likelihood estimation using a specified dose–toxicity function is performed. However, the use of a seed dataset or some other initial dose escalation is required as unique maximum-likelihood estimates will not exist until certain data conditions are met.^[22] The advantage of this approach is the ease in computations as compared to the CRM. Drawbacks include the need for a seed dataset, just as the CRM requires the use of a prior distribution.

Other Recent Work

Other recent proposals include approaches incorporating all levels of observed toxicity grades (rather than dichotomizing into the presence or absence of significant toxicity)^[23] and using estimates of parameters of pharmacodynamic models to guide the dose escalation.^[24–26]

Discussion

In summary, phase I cancer trials provide ethical and statistical challenges different than phase I trials in normal volunteers. The traditional method uses a conservative starting dose and escalates using cohorts of three patients at each dose. Recent literature has demonstrated the poor statistical properties of the traditional approach and is full of novel approaches offering varying improvements. The selection of a procedure for a particular trial depends upon many factors, including the compound being studied, the amount and type of available information regarding the toxicities of the compound, the type of toxicities expected, and the patient population for the trial. Because of the small sample sizes, simulation studies are typically required in order to compare competing procedures and to determine the operating characteristics of a procedure.

In most of the newer approaches, the MTD is defined as a percentile of the dose–toxicity relationship. Thus statistically the problem reduces to estimating the percentile of a distribution function. The choice of the specific percentile to be estimated again depends on many factors, although values from 0.2 to 0.33 have been cited in the literature.

PHASE II CLINICAL TRIALS

Introduction

The primary objective of a phase II cancer trial is to determine whether a new experimental treatment is sufficiently active, relative to standard treatments, to warrant the conduct of a large, randomized phase III comparative trial. Phase II cancer studies are important for two reasons. First, they are numerous and, as a result, many cancer patients receive their primary treatment in the context of such studies. Second, they often determine if and what phase III studies are ultimately conducted. Phase II trials are of two basic types. The first type is composed of trials to determine whether a single agent has antitumor activity against a particular type of cancer and provides a rough estimate of the level of that activity. The second type includes phase II trials whose objective is to determine whether the level of activity of a particular combination of active agents is sufficiently promising to warrant a randomized phase III evaluation of the combination against the standard treatment. The first type of phase II trials has received considerable statistical attention in the literature. However, the statistical methods developed for the first type of phase II trials have not been entirely adequate for the second type of phase II trials. Phase II combination trials require explicit consideration of the results of the standard treatment. This comparative nature of phase II trials of combination regimen has often been ignored. For more details, see Ref. [27].

Phase II Cancer Clinical Trial Designs

Initiation of phase II clinical trials generally assumes that a “safe” dose regimen has emerged from the phase I clinical trial experience. The selected dose regimen’s efficacy must then be tested in phase II patients with advanced disease while further characterizing its safety profile. Therapeutic efficacy is often primarily evaluated on the basis of “regression probability,” i.e., the probability that an eligible patient receiving the treatment regimen will experience a tumor regression as defined in the protocol. Historically, phase II study design strategy has been “modality-oriented” as well as “disease-oriented.” With the modality-oriented strategy, patients with different primary tumor types were treated with the same dose regimen. Some of the oldest anticancer agents, such as cyclophosphamide, 5-FU, methotrexate, and adriamycin, were initially developed with this strategy. The most significant issue with the modality-oriented

strategy was that factors associated with a given disease, such as cancers of lung, pancreas, head and neck, or breast, introduced bias in the trial results. Increasingly, refined understanding of prognostic factors associated with specific types of cancers has rendered the modality-oriented strategy impractical and strengthened the case for disease-oriented strategy. With the disease-oriented strategy, the investigator determines several key factors that can affect the outcome of the trial. They include the type of disease to study, where in the course of disease treatment to intervene, and what end result is deemed appropriate for the trial. Prognostic factors that may affect response to treatment are then accounted for in the trial design. Almost all of the statistical methodology put forward for phase II cancer trials has been developed under the disease-oriented strategy.

The Single-Stage Design: The Classical Approach

The classical approach for designing and conducting phase II cancer trials has been the single-stage design.^[28] Typically, a total of N evaluable patients are accrued, treated, and observed. Then the number, S , of patients who experience a tumor regression is determined. From the data, one can obtain the point estimate of the incidence of tumor regression probability or tumor response rate, given by the ratio S/N . In addition to the derivation of tumor response rate, it is commonly of interest to formulate a testing procedure to decide whether or not the treatment regimen deserves further investigation. This exercise amounts to specifying in addition to the total number of evaluable patients N , the single-stage design parameters C . The choice of C is made so that if more than C of the N evaluable patients are tumor responders, the new treatment regimen is declared effective, and ineffective otherwise.

The Single-Stage Design: The Bayesian Approach

Phase II clinical trials—typically single-arm studies with the objective to decide whether a new treatment E is sufficiently promising relative to a standard treatment, S —are inherently comparative although a standard therapy arm is usually not included. Rarely is the uncertainty about the response rate θ_s of S made explicit, either in planning the trial or interpreting the results. Thall and Simon^[29] have put forward some practical Bayesian guidelines for deciding whether the new therapy E is promising relative to the standard therapy S in the setting where patient response is binary and the data are monitored continuously. The rationale for such an approach stems from practical observation. In order to implement a phase II trial, the clinician must specify a single value of the patient's response rate θ_s to S . In many cases, there is uncertainty regarding θ_s . For example, it is not uncommon for a clinician to give a range of values when asked to provide the response rate of the standard therapy S . Under such circumstances, it would

be in fact inappropriate to insist on a single value of θ_s for planning purposes, because the resulting design and analysis would treat a variable quantity as if it was a constant. The authors point out that a more realistic approach should explicitly account for the clinician's uncertainty regarding θ_s in the planning of the phase II trial and the interpretation of the results. Briefly, the efficacy of an experimental treatment E is evaluated relative to that of a standard treatment S based on data from an uncontrolled trial of E , and informative prior for θ_s , a targeted improvement for E , and a noninformative prior for θ_E . A noninformative prior for θ_E is appropriate here, because in practice little is known about the efficacy of E prior to the trial, and one wishes to avoid designs that allow strong prior opinion to be used in lieu of empirical evidence. No explicit specification of a loss function is required. The trial continues until E is shown with a high posterior probability to be either promising or not promising, or until a predetermined maximum sample size is reached. The design provides decision boundaries, a probability distribution for the sample size at termination, and operating characteristics under fixed response rate for E . Two extensions of this decision structure were also proposed.^[29] The first gives criteria from early termination of trials unlikely to yield conclusive results, based on the marginal predictive distribution of the observed response rate. The second allows early termination only if the experimental treatment E is not found to be promising compared to the standard treatment S . For further details on Bayesian approaches in phase II cancer trials, we encourage the reader to consult the references on the topic.^[29]

Sequential and Multistage Designs

The rationale for these types of designs stems from ethical concerns. Investigators have become, over the years, increasingly aware of the need for early termination of phase II trials when early data tend to support the hypothesis of treatment ineffectiveness, and for early reporting of results when early data support the hypothesis of effectiveness of a new treatment. Early termination, with or without a data-driven statistical hypothesis test, has the effect of inflating the false-positive (type I) and/or false-negative (type II) error probabilities above prespecified tolerable limits.

Strictly Sequential Designs. Wald's strictly sequential design^[30] is perhaps the most well-known early tool used in designs that permits early termination of a trial while maintaining overall control of type I and type II errors. Here no fixed sample size is previously specified. The decision to stop or continue patient accrual is made after each new patient is entered. The method allows for the control of type I and type II error probabilities. Drawbacks of strictly sequential design have been discussed by a number of authors.^[31] They include administrative burdens of maintaining a constant vigil over the data, delays between patient entry and response determination, and

delay between response determination and the reporting of response to the trial sponsor.

Multistage Designs. Multistage designs are a compromise between strictly sequential designs and the single-stage design. A “full” sample size is declared up front, but patients are entered into the trial in batches. After each batch is accrued, a decision is made either to stop the trial or to continue to the next stage on the basis of the number of responding patients observed up to that point. Multistage designs also allow for the control of type I and type II error probabilities. The major drawback is mostly the administrative problems associated with monitoring the trial so as to order the incoming observations properly and to minimize the nonrandomness in the sequence of incoming-patient data.

Several articles that present a one-sample multiple-testing procedure, developed explicitly to handle phase II cancer trials, have appeared in the statistical literature. The most extensively used multistage designs were put forward by Gehan in the early 1960s,^[32] Fleming in the early 1980s,^[33] and Simon in the late 1980s.^[34]

Gehan^[32] designed a two-stage procedure in which early stopping occurs if no response is observed in the first n_1 patients, where n_1 is chosen to be sufficiently large that such a lack of response would be inconsistent, at a given significance level, with the hypothesis of activity of the new treatment. If any responses are observed in the first n_1 patients, an additional n_2 patients are entered, where the total number of patients $N = n_1 + n_2$ is chosen so that the true antitumor activity of the new treatment can be estimated with approximately a prespecified precision. In practice, however, Gehan’s rule has been (and probably still is) greatly misunderstood by clinicians. Most of them believe that the rule is a hypothesis testing procedure rather than an estimation procedure, and that for any specified response rate θ_0 the first n_1 patients will be sufficient for a phase II trial. The second sample of n_2 patients is rarely mentioned.

Fleming’s one-sample-group sequential design^[33] is perhaps the most widely used for phase II trials by cancer investigators. A K -stage multiple testing procedure, the design employs a standard single-stage procedure at the last test, allows for early termination, and essentially preserves the size and power. Simplicity of the single-stage procedure is captured and preserved through the specification of easily calculated acceptance and rejection points for each of the K tests. Simon^[34] introduced an optimized version of this type of group sequential design by minimizing expected sample size subject to upper bounds on the error probabilities.

PHASE III CLINICAL TRIALS

Phase III trials are the last step in the planned evaluation of a new treatment. They have one of the following

objectives:^[28,35] (1) to determine the effectiveness of a new treatment relative to placebo (disease natural history), (2) to determine the effectiveness of a new treatment with regard to the best standard therapy (active control), (3) to determine if a new treatment is as effective as the standard treatment but is associated with less severe toxicity. Typical primary endpoints in a phase III cancer trial are response rate, time-to-an-event measure, such as time-to-disease progression, time-to-treatment failure, or time-to-death. Recently, novel endpoints such as clinical benefit,^[36] as well as quality of life have been explored in phase III cancer trials for advanced, generally symptomatic disease where palliation is deemed a clinically meaningful treatment outcome. However, hard endpoints, such as survival and time-to-disease progression, remain the most preferred endpoints in phase III cancer trial. As a result, cancer clinical research has been one of the most important breeding grounds for some of the most important and creative statistical developments in the area of survival analysis.

Unstratified Designs

With unstratified design, patients are randomly assigned to treatment arms as they become available. As the trial sample size increases, the number of patients per treatment arm becomes approximately the same, but at any point, an imbalance between the number of patients per treatment could occur by chance. This is most likely to be the case in small sample trials. Because, for the sake of statistical power, one would like to ensure approximately equal number of patients per treatment arm, the randomization is “blocked” or “restricted.” A block is a sequence of n patients in which each study treatment is allocated an equal number of times. Because of potentially severe toxicities that can be associated with most cancer therapies, treatments are generally not blinded to clinical investigators in order to ensure the best patient care during the course of a phase III cancer trial. This constitutes the major drawback to blocked randomization, because an investigator may figure out the block size and thus know exactly what treatment will be assigned to the last patient(s) of the block and, therefore, can select the patients for a given treatment. This is especially true for single-center studies, where only one or two investigators are enrolling the patients. Simon^[37] and Pocock^[38] give an elegant discussion of a number of methods to limit this kind of bias, the simplest of which is to vary the dimension of the blocks and to keep investigators blinded to their sizes.

Stratified Designs

When factors other than treatment that can influence the response to treatment are known, they must be accounted for when evaluating the meaning of an observed difference between treatment outcomes. The ideal scenario is to account for the possible influence of the prognostic factors

by distributing them equally between the treatment groups. The random assignment of patients to treatment arms is therefore performed in each stratum of patients. The most significant drawback is that stratification by more than two factors, or more than two levels, quickly leads to impractical design. Indeed, the number of strata become too large and, as a result, an imbalance between treatment arms with respect to the prognostic factors can occur simply by chance. Excessive stratification in itself is not useful, because, in the long run, it amounts to no stratification and unblocked randomization.^[37,39,40] Also, the complexity of randomization within a large number of strata is conducive to administrative errors.

Adaptive Stratified Designs

In the presence of a large number of strata, the Pocock and Simon^[39] “minimization” approach is perhaps the most effective and commonly used way to avoid severe imbalance in the distribution of patients on treatment arms with respect to known prognostic factors. The reader is encouraged to consult this reference for more details on the minimization techniques.

The minimization procedure is not without inconvenience. Indeed, if an investigator is aware of patient treatment assignment with respect to prognostic factors, the subject can predict the next patient treatment assignment. However, this is likely to be difficult to achieve in the case of multicenter trials with a centralized coordination, where the investigator has no access to information outside subject’s own institution, at least before the end of the trial.

The types of designs just described are by no means the only ones encountered in phase III cancer clinical trials, but they are probably the most frequently used. Factorial and crossover designs are rare but have been used in cancer phase III trials.

PHASE II/III DESIGNS WITH MULTIPLE EXPERIMENTAL ARMS

Recently, a number of authors have put forward new designs for the selection of treatments to be tested in randomized clinical trials.^[41–46] Motivations behind these new designs that interface phase II and phase III cancer trials are ethical, practical, and efficiency-driven in nature. In a program of clinical trials where several experimental treatments are of interest, it may be strategically worthwhile to first identify the single best version of the experimental treatments and then to compare it to a standard control therapy. This is particularly desirable if one would be satisfied with one real advance. The goal is to minimize the number of patients expected to be accrued to new regimens that do not offer a tumor-shrinkage or survival benefit over the standard therapy. Examples of clinical settings where such designs are useful include evaluation of combination

regimens, because conventional nonrandomized phase II trials of combined modality regimens are often unreliable or difficult to interpret, and evaluation of different dosing regimens of the same drug. Interfacing phase II and phase III trials has an added benefit of reducing the time for protocol development, review, and implementation, among others, and takes advantage of the efficiency of early rejection of treatment regimen(s) without stopping the entire study.

These new designs rely heavily on statistical ranking and selection theory^[47,48] for choosing the best among the new regimens. This attractive statistical methodology has not yet been fully leveraged by investigators either in early drug development or in phase II/phase III cancer trial designs. The total sample size needed to select the best among several candidate treatments can be substantially less than that needed to test the classical null hypothesis of equality.

STATISTICAL METHODS AND CANCER TRIAL ENDPOINTS

Tumor response rate remains the most common primary endpoint in phase II cancer trials and is handled with the binomial model.^[28,33] In phase III trials, however, response rate is increasingly reduced to secondary endpoint status and replaced by time-to-event endpoints, such as time-to-progressive disease and survival.

The increasing role of time-to-event endpoints as a more reliable tool for assessing efficacy and safety merits of new therapies, especially in frontline settings, has prompted in recent years a prolific surge in the development of new survival analysis methods. These methods nicely complement or refine the popular Kaplan–Meier nonparametric product-limit method^[49] of estimating survival probability at a given time point and Cox regression modeling.^[50,51] Because of the increasing importance of time-to-event endpoints in phase II and phase III cancer trials, it is worthwhile to highlight some of the most recent survival analysis tools in the statistician’s toolbox.

“Time-to-event” data are routinely encountered in clinical trials. They reflect a time measured from randomization or start of therapy to the date of a well-defined, clinically important event. Examples of such events include death, progression of the disease, off study, response (positive or negative), and relapse. The most common time-to-event data are right-censored, because a patient’s information may be lost to follow-up before the event of interest occurs, arising from a number of reasons, such as dropout, completion of the study, or death from other causes. It is typically assumed that the events must occur in a finite time and censoring occurs at random.

Basic methods used in the analysis of time-to-event data have traditionally consisted of estimation and comparison of survival curves. Although parametric models

are available, the nonparametric Kaplan–Meier method for estimating survival distribution function and the Logrank and Wilcoxon tests for comparing two survival distribution functions remain popular. This is the case also for the semiparametric Cox proportional hazard (PH) regression model with one or more fixed cofactors.

The advantages of the Cox PH model are twofold: First, the model does not require an exact parametric specification of the survival distribution. Second, parameters have a rather straightforward and useful interpretation: the risk ratio for an x -unit increase in a given cofactor, all other factors fixed, is given by $\exp(\beta x)$, where β is the regression coefficient for the cofactor of interest in the model. The model has also its disadvantages. It is limited to fixed cofactors. It is limited to univariate time-to-event variables. Finally, the underlying PH assumption may not be realistic in many cases.

The counting process approach to survival analysis, first formulated by Aalen,^[52] provided the break needed for extensions of the Cox regression model. Aalen outlined a unified theoretical framework for application of point processes, especially counting processes, for time-to-event analysis. His work precipitated the development and application of important and useful multivariate time-to-event models throughout the 1980s.^[53–56] It also led to improved computational methods that facilitated the use of time-dependent cofactors as well as time-to-event vectors.

The Anderson–Gill (AG), Prentice–William–Peterson (PWP), and Wei–Lin–Westfield (WLW) models share five nice features: (1) all three models are semiparametric, multiplicative-hazard models (with exponential relative risk); (2) for each model, a common vector parameter β of regression coefficients can be estimated; (3) the risk ratio for an x -unit increase in a given cofactor, all other factors fixed, is given by $\exp(\beta x)$, where β is the regression coefficient for the cofactor of interest in the model, (4) they allow for more than one time-to-event (dependent) variable, and (5) all three models are now fairly easy to implement using readily available software, such as SAS®.^[57]

There are also subtle but important differences among AG, PWP, and WLW models. First, the AG and PWP models differ in two respects: (1) for the PWP model, patients that have not had the “first” event are not considered at risk for the second and subsequent events, and (2) the PWP model has distinct underlying hazards for each event. Second, the AG and PWP models assume that the distinct events will be interpreted as sequential with independent increments, conditional on cofactors. Finally, the WLW model requires neither a sequential interpretation of the events nor independent increments between the events.

In summary, the AG, PWP, and WLW models are relatively new but immensely useful survival analysis tools in the investigator’s toolbox. The assumption of sequential events with independent increments in the AG and PWP models are strong and should be seriously assessed if these models are to be appropriately implemented. The WLW

model can be applied to any data for which the assumption of exponential relative risk is reasonable. As straightforward generalizations of the Cox model, all three models are relatively easy to interpret in light of the complexity of the problem. To date, multivariate time-to-event models have been rarely implemented in clinical trials. Yet the three models discussed can be applied relatively easily using the SAS® PHREG procedure.^[57] They offer a unique exploratory tool for evaluating not only the relationship between a vector of time to events and fixed or time-dependent cofactors, but also the interrelationships between time to events.

CONCLUSIONS

We have shared here our own experience in designing, implementing, analyzing, and reporting cancer clinical trials. We apologize for the omission and superficial treatment of many important areas. We hope that the discussion will point the reader in the right direction for a more comprehensive treatment of the different topics introduced. Our intention has been to focus more on statistical aspects unique to cancer clinical investigation. For readers interested in a much broader perspective, Simon^[27] offers a superb account of the state of progress in statistical methodology for clinical trials.

REFERENCES

1. Schneiderman, M.A. Proc. Fifth Berkeley Symp. Math. Stat. Probab. **1965**, 4, 855–866.
2. Geller, N.L. Drug Inf. J. **1990**, 24, 341–349.
3. Simon, R.; Freidlin, B.; Rubinstein, L.; Arbus, S.G.; Collins, J.; Christian, M.C. J. Natl. Cancer Inst. **1997**, 89 (15), 1138–1147.
4. Von Hoff, J.K.; Clark, G. *Cancer Clinical Trials: Methods and Practice*; Buyse, M.E., Staquet, M.J., Sylvester, R.J., Eds.; Oxford University Press: New York, 1984.
5. Storer, B.E.; DeMets, D. J. Clin. Res. Drug Dev. **1987**, 1, 121–130.
6. Collins, J.M.; Zaharko, D.S.; Dedrick, R.L.; Chabner, B.A. Cancer Treat. Rep. **1986**, 70, 73–80.
7. Collins, J.; Grieshaber, C.; Chabner, B. J. Natl. Cancer Inst. **1990**, 82, 1321–1326.
8. O’Quigley, J.; Pepe, M.; Fisher, L. Biometrics **1990**, 46, 33–48.
9. EORTC Pharmacokinetics and Metabolism Group. Eur. J. Cancer Clin. Oncol **1987**, 23, 1083–1087.
10. Foster, B.J.; Newell, D.R.; Graham, M.A.; Gumbrell, L.A.; Jenns, K.E.; Kaye, S.B.; Calvert, A.H. Eur. J. Cancer **1992**, 28 (2–3), 463–469.
11. Gianni, L.; Vigano, L.; Surbone, A.; Ballinari, D.; Casali, P.; Tarella, C.; Collins, J.; Bonadonna, G. J. Natl. Cancer Inst. **1990**, 82 (6), 469–477.
12. Dixon, W.J.; Mood, A.M. Ann. Math. Stat. **1948**, 36, 800–807.

13. Storer, B.E. *Biometrics* **1989**, *45*, 925–937.
14. Chevret, S. *Stat. Med.* **1993**, *12*, 1093–1108.
15. Faries, D. J. *Biopharm. Stat.* **1994**, *4* (2), 147–164.
16. Goodman, S.N.; Zahurak, M.L.; Piantadosi, S. *Stat. Med.* **1995**, *14*, 1149–1161.
17. Korn, E.L.; Midthune, D.; Chen, T.T.; Rubinstein, L.V.; Christian, M.C.; Simon, R. *Stat. Med.* **1994**, *13* (18), 1799–1806.
18. Moller, S. *Stat. Med.* **1995**, *14*, 911–922.
19. Rinaldi, D.A.; Burris, H.A.; Dorr, F.A.; Woodworth, J.R.; et al. *J. Oncol.* **1995**, *13* (11), 2842–2850.
20. O'Quigley, J.; Shen, L.Z. *Biometrics* **1996**, *52* (2), 673–684.
21. Murphy, J.R.; Hall, D.L. *J. Biopharm. Stat.* **1997**, *7* (4), 635–647.
22. Silvapulle, M.J. *J. R. Stat. Soc., Ser. B* **1981**, *43* (3), 310–313.
23. Gorden, N.H.; Willson, J.K.V. *Stat. Med.* **1992**, *11*, 2063–2075.
24. Mick, R.; Ratain, M.J. *J. Natl. Cancer Inst.* **1993**, *85*, 217–223.
25. Ratain, M.J.; Mick, R.; Schilsky, R.L.; et al. *J. Natl. Cancer Inst.* **1993**, *85*, 1637–1643.
26. Piantadosi, S.; Liu, G. *Stat. Med.* **1996**, *15*, 1605–1618.
27. Simon, R. *Stat. Med.* **1991**, *10*, 1789–1817.
28. Byse, M.E.; Staquet, M.J.; Sylvester, R.J. *Cancer Clinical Trials: Methods and Practice*; Oxford University Press: Oxford, 1984.
29. Thall, P.F.; Simon, R. *Control. Clin. Trials* **1994**, *15*, 463–481.
30. Wald, A. *Sequential Analysis*; Wiley: New York, 1947.
31. Byar, D.P.; Simon, R.M.; Friedewald, W.T.; Schlesselman, J.J.; DeMets, D.L.; Ellenberg, J.H.; Gail, M.H.; Ware, J.H. *New Engl. J. Med.* **1976**, *295*, 74–80.
32. Gehan, E.A. *J. Chronic. Dis.* **1961**, *13*, 346–353.
33. Fleming, T.R. *Biometrics* **1982**, *38*, 143–151.
34. Simon, R. *Control. Clin. Trials* **1989**, *10*, 1–10.
35. Simon, R. *Cancer, Principles and Practice of Oncology*; DeVita, V.T., Hellman, S., Rosenberg, S.A., Eds.; JB Lippincott: Hagerstown, PA, 1982, 198–225.
36. Andersen, J.S., et al. ASCO program proceedings. Abstract 1600 **1994**.
37. Simon, R. *Biometrics* **1979**, *35*, 503–512.
38. Pocock, S.J. *Biometrics* **1979**, *35*, 183–197.
39. Pocock, S.J.; Simon, R. *Biometrics* **1975**, *31*, 103–115.
40. Zelen, M. *Cancer Therapy: Prognostic Factors and Criteria of Response*; Staquet, M., Ed.; Raven Press: New York, 1975, 1–35.
41. Thall, P.F.; Simon, R.; Ellenberg, S.S. *Biometrika* **1988**, *75*, 303–310.
42. Thall, P.F.; Simon, R.; Ellenberg, S.S. *Biometrics* **1989**, *45*, 537–548.
43. Schaid, D.J.; Wieand, S.; Therneau, T.M. *Biometrika* **1990**, *77*, 507–513.
44. Simon, R.; Thall, P.F.; Ellenberg, S.S. *Stat. Med.* **1994**, *13*, 417–429.
45. Liu, P.Y.; Dahlberg, S.; Crowley, J. *Biometrics* **1993**, *49*, 391–398.
46. Chen, T.T.; Simon, R. *Biometrics* **1993**, *49*, 753–761.
47. Gibbons, J.D.; Olkin, I.; Sobel, M. *Selecting and Ordering Populations—A New Statistical Methodology*; Wiley: New York, 1977.
48. Mukhopadhyay, N.; Solanky, K.S.T. *Selection and Ranking Procedures—Second-Order Asymptotics*; Marcel Dekker: New York, 1994, Vol. 142.
49. Kaplan, E.L.; Meier, P. *J. Am. Stat. Assoc.* **1958**, *9*, 457–481.
50. Cox, D.R. *J. R. Stat. Soc., B* **1972**, *34*, 187–220.
51. Cox, D.R. *Biometrika* **1975**, *62*, 269–276.
52. Aalen, O.O. Ph.D. Dissertation, University of California at Berkeley, 1975.
53. Andersen, P.K.; Gill, R.D. *Ann. Stat.* **1982**, *10*, 1100–1120.
54. Prentice, R.L.; Williams, B.J.; Peterson, A.V. *Biometrika* **1981**, *68*, 372–379.
55. Wei, L.J.; Lin, D.Y.; Weissfeld, L. *J. Am. Stat. Assoc.* **1989**, *84*, 1065–1073.
56. Lin, D.Y. *Stat. Med.* **1994**, *13*, 2233–2247.
57. SAS. SAS Institute Inc.: Cary, NC, 1990.

Carcinogenicity Studies of Pharmaceuticals

Karl K. Lin

U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

The purpose of this entry is to describe some commonly used methods of statistical analysis of tumor data and the interpretation of results of rodent carcinogenicity studies of pharmaceuticals. This entry concentrates only on the analysis of tumor data, with adjustment for possible differences in survival among treatment groups.

INTRODUCTION

The assessment of the risk of exposure to a new drug in humans usually begins with an assessment of the risk of the drug in animals. It is required by law that the sponsor of a new drug conducts nonclinical studies in animals to assess the pharmacological actions, the toxicological effects, and the pharmacokinetic properties of the drug in relation to its proposed therapeutic indications or clinical uses. Studies in animals designed for the assessment of the toxicological effects of the drug include acute, subacute, subchronic and chronic toxicity studies, carcinogenicity studies, reproduction studies, and pharmacokinetic studies. The assessment of the risk of drug exposure in humans includes an important assessment of the risk for carcinogenicity in lifetime tests in mice and rats.

It is recognized that there are distinctive differences in chemical and pharmacological nature between environmental chemicals and human pharmaceuticals. Because of those differences in nature, the design of experiments, the methods of analysis, and the interpretation of the results used in rodent carcinogenicity studies of pharmaceuticals are not totally the same as those used in carcinogenicity studies of environmental chemicals. For example, it has been recommended in the literature that the highest dose used for carcinogenicity studies of environmental chemicals should be close to the toxicity-based endpoint maximum tolerated dose (MTD). In addition to the traditional approach using toxicity-based endpoints, other criteria have been proposed by the International Conference on

Harmonization (ICH) for high-dose selection for carcinogenicity studies of human pharmaceuticals. These include approaches using pharmacodynamic endpoints, pharmacokinetic endpoints, saturation of absorption, and maximum feasible dose.^[1] It is also recognized that knowledge of the pharmacology, the pharmacokinetics, and the metabolic disposition of the pharmaceuticals in humans may be available and can be used in dose selection, in the determination of target organs, and in assessing the relevance of animal findings to humans.

Traditional long-term animal carcinogenicity studies usually are conducted on both sexes of mice and rats for the majority of those animals' normal life spans. Treatment on the animals usually lasts for about 2 years after weaning. A standard, traditional carcinogenicity study without planned interim sacrifices usually uses 50 animals per sex in each of four treatment groups: control, low, medium, and high. To save time and money, the ICH^[2] has approved experimental approaches to the evaluation of carcinogenic potential that may obviate the necessity for the routine conduct of two long-term rodent carcinogenicity studies for those pharmaceuticals that need such evaluation. In the document, the ICH made provision for drug sponsors either to continue to conduct two long-term rodent carcinogenicity studies or to use the alternative approach by conducting only one long-term rodent carcinogenicity study (preferably in rats) and a short- or medium-term rodent test system to evaluate the carcinogenic potential of their drug products. The short or medium rodent test systems include the use of models, such as models of initiation-promotion in rodents, transgenic rodents, or newborn rodents, that provide insight into carcinogenic endpoints *in vivo*.

The primary purpose of a long-term animal carcinogenicity experiment is to evaluate the oncogenic potential of a new drug when it is administered to animals for the majority of their normal life span. Because of the complicated nature of the rodent carcinogenicity experiment, the statistical analysis in this area involves tests

The views expressed here are those of the author and not necessarily those of the U.S. Food and Drug Administration. The techniques or methods of analysis and the interpretation of results described are based on the author's assessment of current literature, his consultations with outside experts, his personal research, and his best scientific judgment, but these are issues where consensus among experts in the evaluation of animal carcinogenicity studies does not always exist.

and estimations of other data, such as survival, food consumption, body weight, organ weight, clinical chemistry, and hematology, in addition to pathology data. The analyses of the other data provide the information about the design validity of a study, the need for adjustment for mortality differences among treatment groups in the final analysis of tumor data, and the possible use of scientific hypotheses about the mechanisms of carcinogenesis. Because of the limitation of space, this entry concentrates only on the analysis of tumor data, with adjustment for possible differences in survival among treatment groups.

In a rodent carcinogenicity study of a new drug using a series of increasing dose levels, statistical tests for positive trends in tumor rates are usually of greatest relevance and interest in the evaluation of the oncogenic potential of the drug. However, as is explained in a subsequent section, in some situations, pairwise comparisons are considered more appropriate than trend tests for carcinogenicity evaluation based on pharmacological justifications.

The purpose of this entry is to describe some commonly used methods of statistical analysis of tumor data and the interpretation of results of rodent carcinogenicity studies of pharmaceuticals. “Tests for Trend and Difference in Tumor Incidence in Chronic Studies Using Peto Tests” and “Tests for Trend and Difference in Chronic Studies Using Poly- k Tests” discuss methods of statistical analysis in long-term carcinogenicity studies. “Analysis of Data from Transgenic Mouse Carcinogenicity Studies” discusses methods of data analysis of the more recently developed alternative test systems for the evaluation of carcinogenic potential of pharmaceuticals. “Analysis of Data from Studies using Dual Control Groups” discusses the arguments for and against the use of this type of study design and methods of analysis of such data. “Interpretation of Results of Statistical Tests” discusses how the results should be interpreted. The techniques or methods of analysis, the interpretation of results described in the following sections are based on the author’s assessment of current literature, his consultations with outside experts, his internal research, and his best scientific judgment, but these are issues where consensus among experts in the evaluation of animal carcinogenicity studies does not always exist.

TESTS FOR TREND AND DIFFERENCES IN TUMOR INCIDENCE IN CHRONIC STUDIES USING PETO TESTS

As just mentioned, in a rodent carcinogenicity study of a new drug using a series of increasing dose levels, statistical tests for positive trends or differences in tumor rates are usually of greatest relevance and interest in the evaluation of the oncogenic potential of the drug. There are different methods proposed for analyzing tumor data from

a rodent carcinogenicity study. General discussions on various methods of tumor data analysis can be found in the statistical literature.^[3–16] The methods of analysis discussed in sections “Adjustment of Tumor Rates for Intercurrent Mortality,” “Contexts of Observation of Tumor Types,” “Tests of Data of Incidental Tumors,” “Tests of Data of Fatal Tumors,” “Tests of Data of Tumors Observed in Both Incidental and Fatal Contexts,” “Tests of Data of Mortality-Independent Tumors,” and “Exact Tests” are from a previously published article by Lin and Ali^[11] with or without modifications.

Adjustment of Tumor Rates for Intercurrent Mortality

Intercurrent mortality refers to all deaths unrelated to a tumor or class of tumors to be analyzed for evidence of carcinogenicity. Like human beings, older rodents have a manyfold higher probability of developing or dying of tumors than those of younger age. Therefore, in the analysis of tumor data, it is essential to identify and adjust for possible differences in intercurrent mortality (or longevity) among treatment groups to eliminate or reduce biases caused by these differences. It has been pointed out that “the effects of differences in longevity on numbers of tumor-bearing animals can be very substantial, and so, whether or not they [the effects] appear to be [very substantial], they should routinely be corrected when presenting experimental results.”^[12] The following examples clearly demonstrate this important point.

Example 1

Consider an experiment consisting of one control group and one treated group of 100 animals each of a strain of mice.^[12] A very toxic but not carcinogenic new drug was administered to the animals in the diet for 2 years. Assume that the spontaneous incidental tumor rates for both groups are 30% at 15 months and 80% at 18 months of age and that the mortality rates at 15 months for the control and the treated groups are 20% and 60%, respectively, due to the toxicity of the drug. The results of the experiment are summarized in Table 1.

If one looks only at the overall tumor incidence rates of the control and the treated groups (70% and 50%,

Table 1 Effects of Mortality Differences on Tumor Incidence Rates (Example 1)

	Control			Treated		
	<i>T</i>	<i>D</i>	%	<i>T</i>	<i>D</i>	%
15 months	6	20	30	18	60	30
18 months	64	80	80	32	40	80
Total	70	100	70	50	100	50

T = incidental tumors found at necropsy; *D* = deaths.

respectively) without considering the significantly higher early deaths in the treated group caused by the toxicity of this new drug, one might misinterpret the apparently significant ($p = 0.002$, one-tailed) decrease in the treated group in this tumor type. However, the one-tailed p value is 0.5 when the survival-adjusted prevalence method^[12] is used.

Example 2

Assume that the design used in this experiment is the same as the one used in the experiment of Example 1. However, assume that the tested new drug in this example induces an incidental tumor that does not directly or indirectly cause animals' deaths, in addition to having severe toxicity, as in the previous example. Assume further that the incidental tumor prevalence rates for the control and treated groups are 5% and 20%, respectively, before 15 months of age, and 30% and 70%, respectively, after 15 months of age; and that the mortality rates at 15 months are 20% and 90% for the control and the treated groups, respectively.^[6] The results of this experiment are summarized in Table 2.

The age-specific tumor incidence rates are significantly higher in the treated group than those in the control group. The survival-adjusted prevalence method yielded a one-tailed p value of 0.003, revealing a clear tumorigenic effect of the new drug. The overall tumor incidence rates, however, are 25% for the two groups. Without adjusting for the significantly higher early mortality in the treated group, the positive finding would be missed.

Before analyzing the tumor data, the intercurrent mortality data should be checked to see if the survival distributions of the treatment groups are significantly different or if there exists a significant trend for survival. The Cox test,^[6,13,17] the generalized Wilcoxon or Kruskal-Wallis test,^[13,18,19] and the Tarone trend tests^[12,20,21] are routinely used to test for heterogeneity in survival distributions and significant dose-response relationships (trends) in survival.

It is recommended by Peto et al.^[12] that whether or not survival among treatment groups is significantly different, tumor rates should routinely be adjusted for survival when presenting experimental results.

Table 2 Effects of Mortality Differences on Tumor Incidence Rates (Example 2)

	Control			Treated		
	<i>T</i>	<i>D</i>	%	<i>T</i>	<i>D</i>	%
Before 15 months	1	20	5	18	90	20
After 15 months	24	80	30	7	10	70
Total	25	100	25	25	100	25

T = incidental tumors found at necropsy; *D* = deaths.

Contexts of Observation of Tumor Types

The choice of a survival-adjusted method to analyze tumor data depends on the role that a tumor plays in causing the animal's death. Tumors can be classified as "incidental," "fatal," and "mortality-independent (or observable)," according to the contexts of observation described in Peto et al.^[12] Tumors that are not directly or indirectly responsible for the animal's death but are merely observed at the autopsy of the animal after it has died of an unrelated cause are said to have been observed in an *incidental* context. Tumors that kill the animal, either directly or indirectly, are said to have been observed in a *fatal* context. Tumors, such as skin tumors, whose detection occurs at times other than when the animal dies are said to have been observed in a *mortality-independent (or observable)* context. To apply a survival-adjusted method correctly, it is essential that the context of observation of a tumor be determined as accurately as possible. "The distinction between fatal and incidental tumors is important because it is essential to distinguish between a chemical that reduces survival by shortening the time to tumor onset or the time to death following tumor onset (a real carcinogenic effect), and one that also reduces survival, but for which tumors are observed earlier simply because animals are dying of competing causes (a noncarcinogenic effect)."^[4]

Different statistical techniques have been proposed for analyzing data on tumors observed in different contexts of observation. For example, the prevalence method, the death-rate method, and the onset-rate method are recommended for analyzing data on tumors observed in incidental, fatal, and mortality-independent contexts of observation, respectively, in Peto et al.^[12] In this paper, Peto et al. demonstrate the possible biases resulting from misclassification of incidental tumors as fatal tumors or of fatal tumors as incidental tumors.

Tests of Data of Incidental Tumors

The *prevalence* method described in the paper by Peto et al.^[12] is routinely used by statisticians in testing for positive trends in prevalence rates of incidental tumors. The method can be described briefly as follows.

The prevalence method focuses on the age-specific tumor prevalence rates to correct for intercurrent mortality differences among treatment groups in the test for positive trends or differences in incidental tumors. The experimental period is partitioned into a set of intervals plus interim (if any) and terminal sacrifices. The incidental tumors are then stratified by those intervals of survival times. The selection of the partitions of the experimental period does not matter very much as long as the intervals "are not so short that the prevalence of incidental tumors in the autopsies they contain is not stable, nor yet so large that the real prevalence in the first

half of one interval could differ markedly from the real prevalence in the second half.^{12]}

In each time interval, for each group, the observed and the expected numbers of animals with a particular tumor type found in necropsies are compared. The expected number is calculated under the null hypothesis that there is no dose-related trend. Finally, the differences between the observed and the expected numbers of animals found with the tumor type after their deaths are combined across all time intervals to yield an overall test statistic using the method described in the paper of Mantel and Haenszel.^[22]

The following derivation of the Peto prevalence test statistic uses the notations in Table 3. Let the experimental period be partitioned into the following m intervals, $I_1, I_2, \dots, I_m$. As mentioned before, interim (if any) and terminal sacrifices should be treated as separate intervals. The number of autopsied animals expected to have the particular incidental tumor in group i and interval k , under the null hypothesis that there is no treatment effect, is:

$$E_{ik} = O_{.k} P_{ik}$$

The variance-covariance of $(O_{ik} - E_{ik})$ and $(O_{jk} - E_{jk})$ is:

$$V_{ijk} = P_{ik} (\delta_{ij} - P_{jk})$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } I = j \\ 0 & \text{otherwise.} \end{cases}$$

Define

$$O_i = \sum_k O_{ik}$$

$$E_i = \sum_k E_{ik}$$

$$V_{ij} = \sum_k V_{ijk}$$

The test statistic T for the positive trend in the incidental tumor is defined as:

$$T = \sum_i D_i (O_i - E_i)$$

with estimated variance $V(T) = \sum_i \sum_j D_i D_j V_{ij}$ where D_i is the dose level of the i th group. Under the null hypothesis of equal prevalence rates among the treatment groups, the statistic $Z = \frac{T}{[V(T)]^{1/2}}$ is approximately distributed as a standard normal.

As mentioned earlier, to use the prevalence method, the experimental period should be partitioned into a set of intervals plus interim (if any) and terminal sacrifices. The following partitions (in weeks) are used most often in 2-year studies: (i) 0–50, 51–80, 80–104, interim sacrifice (if any), and terminal sacrifice; (ii) 0–52, 53–78, 79–92, 93–104, interim sacrifice (if any), and terminal sacrifice (proposed by National Toxicology Program [NTP]); and (iii) partition determined by the “ad hoc runs” procedure described in Peto et al.^[12]

The data of liver hepatocellular adenoma of male mice from a carcinogenicity experiment of a new drug are used as examples to explain the prevalence method for testing the positive trend in tumor rates of an incidental tumor.

Table 3 Notations Used in the Derivation of Peto Prevalence Test Statistics

	Group	0	1	...	i	...	r	
Interval	Dose	D_0	D_1	...	D_i	...	D_r	Sum
I_1	R_1	O_{01}	O_{11}	...	O_{i1}	...	O_{r1}	$O_{.1}$
		P_{01}	P_{11}	...	P_{i1}	...	P_{r1}	$P_{.1}$
I_2	R_2	O_{02}	O_{12}	...	O_{i2}	...	O_{r2}	$O_{.2}$
		P_{02}	P_{12}	...	P_{i2}	...	P_{r2}	$P_{.2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_k	R_k	O_{0k}	O_{1k}	...	O_{ik}	...	O_{rk}	$O_{.k}$
		P_{0k}	P_{1k}	...	P_{ik}	...	P_{rk}	$P_{.k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_m	R_m	O_{0m}	O_{1m}	...	O_{im}	...	O_{rm}	$O_{.m}$
		P_{0m}	P_{1m}	...	P_{im}	...	P_{rm}	$P_{.m}$

R_k = number of animals that have not died of the tumor type of interest but come to autopsy in timeinterval k .

P_{ik} = proportion of R_k in group i .

O_{ik} = observed number of autopsied animals in group i and interval k found to have the incidental tumor type.

$O_{.k} = \sum_i O_{ik}$.

Table 4 Data on Liver Hepatocellular Adenoma of Male Mice

Time interval (weeks)	Group											
	Control			Low			Medium			High		
	<i>T</i>	<i>N</i>	%	<i>T</i>	<i>N</i>	%	<i>T</i>	<i>N</i>	%	<i>T</i>	<i>N</i>	%
0–50	0	6	0	0	2	0	0	2	0	0	4	0
51–80	1	26	4	1	18	6	3	17	18	1	13	8
81–106	4	37	11	2	14	14	2	14	14	7	19	37
Terminal sacrifice	2	31	6	5	16	31	3	17	18	4	14	29
Total	7	100	7	8	50	16	8	50	16	12	50	24

T = number of necropsies with the particular tumor.

N = number of necropsies during a time interval.

% = percentage of necropsies with the particular tumor.

There were four treatment groups in the mouse experiment. The control group had 100 animals; the three treated groups had 50 animals each. The dose levels used were 0, 10, 20, and 40 mg/kg/day for the control, low, medium, and high groups, respectively. The study lasted for 106 weeks. In this example, the study period was partitioned into four intervals: 0–50, 51–80, 81–106, and terminal sacrifice. The numbers of animals died and necropsied and the numbers of necropsied animals with liver hepatocellular adenoma by treatment group in each interval are included in Table 4.

The observed rates and the expected rates of the tumor type calculated under the null hypothesis that there is no trend (or drug-induced increase) are shown in Table 5. The expected tumor rates in each interval were calculated in the following way. First, the tumor rate for the interval using data of all treatment groups in the interval was estimated. For example, the estimated tumor rate for the interval 51–80 weeks is $6/74 = 0.0811$. Second, the expected rates for individual groups in the interval were calculated by multiplying the numbers of necropsies by the estimated

tumor rate. For the interval 50–81 weeks, the expected tumor rates for the control, low, medium, and high groups are $26 \times 6/74 = 2.11$, $18 \times 6/74 = 1.46$, $17 \times 6/74 = 1.38$, and $13 \times 6/74 = 1.05$, respectively.

The values of the test statistic *T* and its variance *V(T)* for the data of the five intervals calculated by the formulas listed earlier are included in Table 6. It is noted that the first interval, 0–50 weeks, did not contribute to the overall test result because none of the 14 animals that died during the time interval developed liver hepatocellular adenoma. The overall result shows a statistically significant positive trend in tumor rates for this tumor (with one-sided *p*-value = 0.002). Also, as mentioned earlier, this method uses normal approximation in the tests for positive trend or difference in tumor prevalence rates. The accuracy of the normal approximation depends on the number of tumor occurrences in each group in each interval, the number of intervals used in the partitioning, and the mortality patterns. The approximation may not be stable and reliable when the numbers of tumor occurrences across treatment groups are small. In this situation, an exact permutation trend test based on an extension of the hypergeometric distribution (to be discussed in the section “Exact Tests”) is used to test for the positive trend in tumor prevalence rates.

Table 5 Observed and Expected Tumor Rates for Liver Hepatocellular Adenoma of Male Mice

Time interval (weeks)	Observed and expected rates	Group			
		Control	Low	Medium	High
0–50	Observed	0	0	0	0
	Expected	0	0	0	0
51–80	Observed	1	1	3	1
	Expected	2.11	1.46	1.38	1.05
81–106	Observed	4	2	2	7
	Expected	6.61	2.50	2.50	3.39
Terminal sacrifice	Observed	2	5	3	4
	Expected	5.56	2.87	3.05	2.51
Total	Observed	7	8	8	12
	Expected	14.28	6.83	6.93	6.95

Expected tumor rates were calculated under the null hypothesis that there is no trend.

Table 6 Test Statistics, Their Variances, *z* Values, and *p*-value of Peto Prevalence Analysis of Incidental Tumors for Liver Hepatocellular Adenoma of Male Mice

Time interval (weeks)	<i>T</i> statistic, <i>T</i>	Variance of <i>T</i> stat, <i>V(T)</i>	<i>z</i> = <i>T</i> / [<i>V(T)</i>] ^{0.5}	<i>p</i> value
0–50	–	–	–	–
51–80	25.6756	1116.583	0.7683	0.2211
81–106	129.2857	3091.314	2.3253	0.0100
Terminal sacrifice	79.7435	2445.855	1.6124	0.0534
Overall total	234.7048	6653.752	2.8773	0.0020

The *z* and *p*-value columns do not add up to the totals. The *z* and *p*-values for the overall total row were calculated based on the *T* and *V(T)* for the row.

Tests of Data of Fatal Tumors

The death-rate method described in Peto et al.^[12] should be utilized to test for the positive trend or difference in tumors observed in a fatal context.

The notations of “Tests of Data of Incidental Tumors,” with some modifications, will be used in this section to derive the test statistic of the death-rate method. Now let $t_1 < t_2 < \dots < t_m$ be the time points when one or more animals died of the fatal tumor of interest. These time points replace the intervals used in the prevalence method. The notations in Table 3 are redefined as follows:

- R_k = number of animals of all groups just before t_k .
- P_{ik} = proportion of R_k in group i (same as in prevalence method).
- O_{ik} = observed number of animals in group i dying of the fatal tumor of interest at time t_k .
- $O_{.k} = \sum_i O_{ik}$.

As in the prevalence method, the test statistic T for the positive trend in the fatal tumor is defined as: $T = \sum_i D_i(O_i - E_i)$ with estimated variance $V(T) = \sum_i \sum_j D_i D_j V_{ij}$ where D_i , O_i , E_i , and V_{ij} are defined as in “Tests of Data of Incidental Tumors.”

Under the null hypothesis of equal tumor rates across the treatment groups, the statistic $Z = \frac{T}{[V(T)]^{1/2}}$ is distributed approximately as standard normal.

Tests of Data of Tumors Observed in Both Incidental and Fatal Contexts

When a tumor is observed in a fatal context for some animals and is also observed in an incidental context for other animals in the experiment, data of the incidental tumor part should be analyzed by the prevalence method, and data of the fatal tumor part should be analyzed by the death-rate method using all animals as at risk of death with the tumor. Results from the different methods can then be combined to yield an overall result. The combined overall result can be obtained simply by adding together either the separate observed frequencies, the expected frequencies, and the variances, or the separate T statistics and their variances.^[12]

Tests of Data of Mortality-Independent Tumors

Tumors observed in a mortality-independent context, such as skin tumors and mammary gland tumors, that are visible and/or can be detected by palpation in living animals should be analyzed using the onset-rate method. The onset-rate method for mortality-independent tumors and the death-rate method for fatal tumors are essentially the same in principle except that the endpoint in the onset-rate method is the occurrence of such a tumor (e.g., skin tumor

reaching some prespecified size) rather than the time or cause of the animal's death.

In the onset-rate method, all those animals that, although still alive, have developed the particular mortality-independent tumor and, hence, are no longer at risk for such a tumor are excluded from the calculation of the numbers of animals at risk. The R_k , P_{ik} , and O_{ik} described in “Tests of Data of Fatal Tumors” are now redefined as follows for the onset-rate method:

- R_k = number of animals alive and free of the mortality-independent tumor of interest in all groups just before t_k .
- P_{ik} = proportion of R_k in group i (same as in death-rate method).
- O_{ik} = observed number of animals in group i found to have developed the mortality-independent tumor of interest at time t_k .

The test statistic T and its estimated variance $V(T)$ are the same as those defined in the death-rate method.

Exact Tests

As mentioned in previous sections, the prevalence method, the death-rate method, and the onset-rate method use normal approximation in the test for the positive trend in tumor incidence rates. Mortality patterns, the number of intervals used in the partitioning of the study period, and the numbers and patterns of tumor occurrence in each individual interval have effects on the accuracy of the normal approximation. It is also well known that the approximation results may not be stable and reliable and that they tend to underestimate the exact p values when the total numbers of tumor occurrence across treatment groups are small.^[23] In this situation, the exact permutation trend test should be used to test for the positive trend.^[6] The exact permutation trend test is a generalization of the Fisher's exact test to a sequence of $2 \times (r + 1)$ tables. The exact permutation trend test procedure described next is for tumors observed in an incidental context. However, the positive trends in incidence rates of tumors observed in a fatal or in a mortality-independent context can be tested in a similar way. In those cases, the number of $2 \times (r + 1)$ tables will be equal to the number of time points when one or more animals died of a particular fatal tumor or when one or more animals developed a particular mortality-independent tumor. Fairweather et al.^[24] contains a discussion on the limitations of applying exact methods to fatal tumors.

The exact method is derived by conditioning on the row and column marginal totals of each of the $2 \times (r + 1)$ tables formed from the partitioned data set of Table 3. Consider the k th interval, I_k , (in Table 3) and rewrite it as in Table 7. Let the column totals C_{0k} , C_{1k} , ..., C_{rk} and the row totals $O_{.k}$ and $A_{.k}$ be fixed. Define $P_{ik} = C_{ik}/R_k$. Then the quantities $E_{ik} = O_{.k}P_{ik}$, $V_{ijk} = P_{ik}(\delta_{ij} - P_{jk})$, E_i , and $V(T)$ (defined in

Table 7 Data in the k th Time Interval, I_k , are Written as a $2(r+1)$ Table

	Group	0	1	...	i	...	r	
	Dose	D_0	D_1	...	D_i	...	D_r	Total
Number with tumor		O_{0k}	O_{1k}	...	O_{ik}	...	O_{rk}	$O_{.k}$
Number without tumor		A_{0k}	A_{1k}	...	A_{ik}	...	A_{rk}	$A_{.k}$
Total		C_{0k}	C_{1k}	...	C_{ik}	...	C_{rk}	R_k

“Evaluation of Validity of Designs of Negative Studies”) are all known constants.

Now let y be the observed value of $Y = \sum D_i O_i$, where $O_i = \sum_k O_{ik}$, the total number of tumor-bearing animals with the tumor of interest in treatment group i . Then (under conditioning on the column and row marginal totals in each table) the observed significance level

$$\begin{aligned}
 p \text{ value} &= P\left[\sum D_i O_i \geq y\right] = P\left(\sum_i D_i \sum_k O_{ik} \geq y\right) \\
 &= P\left(\sum_k \sum_i D_i O_{ik} \geq y\right) \\
 &= P\left(\sum_k Y_k \geq \sum_i y_k\right) = P(Y \geq y)
 \end{aligned}$$

where $Y = \sum Y_k = \sum_k \sum_i D_i O_{ik}$, and $y = \sum y_k$, the observed value of Y . The p value $P(Y \geq y)$ is computed from the exact permutational distribution of Y . Given the observed row and column marginal totals in a $2 \times (r+1)$ table, all possible tables having the same marginal totals can be generated. Let S_k ($k = 1, 2, \dots, K$) be the set of all such tables generated from the k th observed table. From a set of K tables taking one from each S_k and assuming independence between the K tables, the foregoing expression for the p value can now be written as $p \text{ value} = \sum [P(Y_1 = y_1) \dots P(Y_k = y_k)]$ where $y_k = \sum_i D_i O_{ik}$ ($k = 1, 2, \dots, K$), the sum is over all sets of K tables such that $y_1 + y_2 + \dots + y_K \geq y$, the observed value of Y , and $P(Y_k = y_k)$ is the conditional probability given the marginal totals in the k th table, i.e.,

$$P(Y_k = y_k) = \frac{\left(\begin{matrix} C_{0k} \\ O_{0k} \end{matrix} \right) \left(\begin{matrix} C_{1k} \\ O_{1k} \end{matrix} \right) \dots \left(\begin{matrix} C_{rk} \\ O_{rk} \end{matrix} \right)}{\left(\begin{matrix} R_k \\ O_{.k} \end{matrix} \right)}$$

The following example illustrates how the probabilities are computed in the exact permutation trend test described earlier. Consider an experiment with three treatment groups (control, low, and high) with dose levels $D_0 = 0$, $D_1 = 1$, and $D_2 = 2$, respectively. Suppose the study period is partitioned into the intervals 0–50, 51–80, and 81–104 weeks and the terminal sacrifice week. Consider a tumor type (classified as incidental) with data as in Table 8. As all the observed tumor counts (i.e., O s) in the first two time intervals are zero, the data for these intervals will not contribute anything to the test statistic and these intervals may be ignored. The observed subtables formed from the last two intervals are given in Table 9.

Now generate all possible tables from observed subtable 1. As the marginal totals are fixed, these tables may be generated by distributing the total tumor frequency $O_{.1} (= 2)$ among the three treatment groups. Thus, each table will correspond to a configuration of this distribution of $O_{.1}$. The configurations, the values of Y_1 , and the $P(Y_1 = y_1)$ are shown in Table 10.

Table 8 Hypothetical Tumor Data for Exact Permutation Trend Test

Time interval		Dose level			Total
		0	1	2	
0–50	O	0	0	0	0
	C	1	3	3	7
51–80	O	0	0	0	0
	C	4	5	7	16
81–104	O	0	0	2	2
	C	10	12	15	37
Terminal sacrifice	O	0	1	0	1
	C	35	30	25	90

O = observed tumor count; C = number of animals necropsied.

Table 9 Observed Subtables from the Hypothetical Tumor Data of Table 8

Observed subtable 1					Observed subtable 2				
Dose	0	1	2	Total	Dose	0	1	2	Total
O	0	0	2	$2 = O_{.1}$	O	0	1	0	$1 = O_{.2}$
A	10	12	12	$35 = A_{.1}$	A	35	29	25	$89 = A_{.2}$
C	10	12	15	$37 = R_1$	C	35	30	25	$90 = R_2$

Table 10 All Possible Configurations of O_1 and the Corresponding Hypergeometric Probabilities

Configuration	y_1	$P(Y_1 = y_1)$
0,0,2	4	0.15766
0,2,0	2	0.09910
2,0,0	0	0.06757
0,1,1	3	0.27027
1,0,1	2	0.22523
1,1,0	1	0.18018

Table 11 All Possible Configurations of O_2 and the Corresponding Hypergeometric Probabilities

Configurations	y_2	$P(Y_2 = y_2)$
0,0,1	2	0.27778
0,1,0	1	0.33333
1,0,0	0	0.38889

To illustrate the computation of y_1 and $P(Y_1 = y_1)$, consider the last row. Here

$$y_1 = D_0 \times 1 + D_1 \times 1 + D_2 \times 0 \\ = 0 \times 1 + 1 \times 1 + 2 \times 0 = 1$$

and

$$P(Y_1 = 1) = \frac{\binom{10}{1} \binom{12}{1} \binom{15}{0}}{\binom{37}{2}} = \frac{10 \times 12 \times 2}{37 \times 36} \\ = 0.18018$$

The configurations and probabilities obtained from observed subtable 2 are given in Table 11. Note that the first configuration (0,0,2) in Table 10 corresponds to the observed subtable 1 with a value of $y_1 = (0 \times 0) + (1 \times 0) + (2 \times 2) = 4$ and a probability of 0.15766, and the second configuration (0,1,0) in Table 11 corresponds to the observed subtable 2 with a value of $y_2 = (0 \times 0) + (1 \times 1) + (0 \times 0) = 1$ and a probability of 0.33333. Thus, the observed value of $y = y_1 + y_2 = 4 + 1 = 5$. Now the exact p value (right-tailed) is calculated as follows:

$$P(Y = Y_1 + Y_2 \geq 5) = P(Y_1 = 4, Y_2 = 1) \\ + P(Y_1 = 4, Y_2 = 2) \\ + P(Y_1 = 3, Y_2 = 2) \\ = 0.15766 \times 0.33333 + 0.15766 \\ \times 0.27778 + 0.27027 \\ \times 0.27778 \\ = 0.17142$$

For the purpose of comparison, it may be noted that the normal approximated p -value for the data set in our example is 0.0927.

TESTS FOR TREND AND DIFFERENCE IN TUMOR INCIDENCE IN CHRONIC STUDIES USING POLY-K TESTS

As mentioned previously, in the analysis of tumor data, it is essential to identify and adjust for possible differences in intercurrent mortality among treatment groups in order to eliminate or reduce biases caused by these differences. However, most tumors, except those that can be detected by palpation and visual inspection, such as skin tumors, are discovered only at the time of the animal's death. The exact tumor onset times are censored and, therefore, are unidentifiable and unobservable. "Without direct observations of the tumor onset times, the desired survival adjustment usually is accomplished by making assumptions concerning tumor lethality, cause of death, multiple sacrifices, or parametric models."^[3]

There is considerable discussion in the literature on statistical tests for positive trend in tumor incidence rate based on those assumptions and models. Some of those proposed statistical tests are not practical because they are based on either uncommon designs with serial sacrifices or mathematical assumptions that are not realistic and/or cannot be verified. The widely used prevalence method, the death-rate method, and the onset-rate methods for analyzing incidental, fatal, and mortality-independent tumors, respectively, described in the previous sections, rely on the information on tumor lethality and cause of death. That is, the choice of a survival-adjusted method to analyze tumor data depends on the role that a tumor plays in causing the animal's death.^[8,12]

There are situations in which sponsors do not include the tumor lethality and cause-of-death information in their statistical analyses and in their submitted electronic data sets. Under those situations, statistical reviewers in regulatory agencies either treat all tumors as incidental or, if provided, follow the cause-of-death information determined by reviewing pharmacologists and toxicologists in the agencies. There are consequences in misclassifying tumors and misusing survival-adjusted statistical tests. The use of the prevalence method will reject the null hypothesis of no positive trend less frequently than it should as the lethality of a tumor increases.^[3,12] This will increase the probability of failing to detect true carcinogens.

The poly-3 and poly-6 (in general, poly- k) tests^[3,25-28] have been proposed for testing linear trends in tumor rates. These tests are basically modifications of the survival-unadjusted Cochran-Armitage test^[29-31] for linear trend in tumor rate. If the entire study period is considered as one interval, the data of a particular tumor will be in the form

Table 12 Data Using the Entire Study Period as an Interval

Group	0	1	...	i	...	r	
Dose	D_0	D_1	...	D_i	...	D_r	Total
Number with tumor	O_0	O_1	...	O_i	...	O_r	O
Number without tumor	A_0	A_1	...	A_i	...	A_r	A
Total	C_0	C_1	...	C_i	...	C_r	R

of Table 12. The notations in Table 12 to be used to explain these tests are the same as those in Table 7 except that the k th interval is now the entire study period. The second subscript, k (for the k th interval), was dropped from the notations.

The Cochran-Armitage test statistic for linear trend^[30] in tumor rate is defined as:

$$\chi_{CA}^2 = \frac{R\{R\sum O_i D_i - O\sum C_i D_i\}^2}{O(R-O)\{R\sum C_i D_i^2 - (\sum C_i D_i)^2\}} \text{ or } \frac{\{\sum D_i(O_i - E_i)\}^2}{\sum E_i D_i^2 - (\sum E_i D_i)^2 / O}$$

where $O = \sum O_i$, $A = \sum A_i$, $R = \sum C_i$, and $E_i = OC_i/R$. The test statistic χ_{CA}^2 is distributed approximately as χ^2 on one degree of freedom. The Cochran-Armitage linear trend test is based on a binomial assumption that all animals in the same treatment group have the same risk of developing the tumor over the duration of the study. However, as mentioned previously, the animal's risk of developing the tumor increases as study time increases. The assumption is no longer valid if some animals die earlier than others. It has been shown that as long as the mortality patterns are similar across treatment groups, the Cochran-Armitage test is still valid although it might be slightly less efficient than a survival-adjusted test.^[3] However, if the mortality patterns are different across treatment groups, the Cochran-Armitage test can give very misleading results.

The poly-3 test adjusts the differences in mortality among treatment groups by modifying the number of animals at risk in the denominators in the calculations of overall tumor rates in the Cochran-Armitage test to reflect "less-than-whole-animal contributions for decreased survival."^[25] The modification is made by defining a new number of animals at risk for each treatment group. The number of animals at risk for the i th treatment group C_i^* is defined as $C_i^* = \sum w_{ij}$ where w_{ij} is the weight for the j th animal in the i th treatment group, and the sum is over all animals in the group.

Bailer and Portier^[25] proposed applying the weight w_{ij} as follows:

$$w_{ij} = \begin{cases} 1 & \text{to animals dying with the tumor} \\ \left(\frac{t_{ij}}{t_{\text{sacr}}}\right)^3 & \text{to animals dying without the tumor.} \end{cases}$$

where t_{ij} is the time of death of the j th animal in the i th treatment group, and t_{sacr} is the time of terminal sacrifice.

The power of 3 used in the weighting is from the observation that tumor incidence can be modeled as a polynomial of order 3. Similarly, the poly-6 test (or the general poly- k test) assigns the weight, $w_{ij} = (t_{ij}/t_{\text{sacr}})^6$ [or $w_{ij} = (t_{ij}/t_{\text{sacr}})^k$] to animals dying without the tumor when the tumor incidence is close to a polynomial of order 6 (or order k).

The class of poly- k tests is carried out by replacing the C_i s by the new numbers of animals at risk C_i^* s in the calculation of the foregoing Cochran-Armitage test statistic.

Bieler and Williams^[26] showed that the poly-3 trend test is anticonservative when tumor incidence rates are low and treatment toxicity is high. They proposed a test called the *ratio trend test* that is another modification to the Cochran-Armitage linear trend test. The ratio trend test employs the adjusted quantal response rates calculated in Bailer and Portier^[25] and the delta method^[32] in the estimation of the variance of the adjusted quantal response rates, $p_i^* = O_i/C_i^*$. The ratio trend test treats both the numerator and the denominator of the adjusted quantal response rate as being subject to random variation.

The computational formula for Bieler-Williams ratio trend (modified C-A) test statistic is given as follows:

$$\chi_{BW}^2 = \frac{\sum m_i p_i^* D_i - (\sum m_i D_i)(\sum m_i p_i^*) / \sum m_i}{\left\{c \left[\sum m_i D_i^2 - (\sum m_i D_i)^2 / \sum m_i \right]\right\}^{1/2}}$$

where

- $c = \sum \sum (r_{ij} - r_i)2/[R - (r + 1)]$.
- $m_i = C_i^*/C_i$.
- $r_{ij} = y_{ij} - p^* w_{ij}$.
- y_{ij} = tumor response indicator (0 = absent at death, 1 = present at death) for the j th animal in the i th group.

It is noted that the notation c in the above formulas is not the same as C_i , the number of animals in group i

(see Table 12), and the notation r for the number of treatment $r + 1$ is not the same as r_{ij} and r_i defined above.

Bieler and Williams^[26] further showed that for tumors with low background rates, the ratio trend test yielded actual type I errors close to the nominal levels used and was observed to be less sensitive than the poly-3 trend test to misspecification of the shape of tumor incidence function and the magnitude of treatment toxicity.

The classes of poly- k tests and the ratio trend test adjust for differences in survival, do not need the information on cause of death, and require only a (the terminal) sacrifice. They are also relative robust to (not affected greatly by) tumor lethality. Results of simulation studies^[3,25] show that these tests performed very well under many conditions simulated. Those tests should be used to replace the asymptotic tests that depend on the information on tumor lethality and on cause of death when the information is not available. As in the case of the Peto tests, it is well known that the approximation results of poly- k tests may also not be stable and reliable and that they tend to underestimate the exact p -values when the total numbers of tumor occurrence across treatment groups are small. Exact versions of the above the poly- k tests should be used for testing data from tumor types with small numbers of tumor-bearing animals. However, because these tests use risk sets based on all animals in each treatment group, the computations involved in the exact tests will be extensive.

Statistical reviewers of carcinogenicity studies in the Center for Drug Evaluation and Research (CDER) of Food and Drug Administration (FDA) have recently decided to move toward the replacement of the Peto tests by the poly- k tests. The following are the reasons for FDA statistical reviewers to change the statistical tests in their reviews.

As mentioned previously, the Peto tests require the information of cause-of-death of individual tumors on individual animals. The information may not be available, or even is available but may not be accurate and reliable since it is difficult for pathologists to determine objectively if a tumor causes the death of an animal. Misclassifications of the cause-of-death information of tumor types can produce misleading results as also mentioned previously.

Most of the time, the Peto tests use the partitions of study duration into sets of fixed time intervals for all tumor types for the practical purpose in data analysis. However, this practice in the Peto tests becomes problematic when some treatment groups were terminated early. In those situations, the numbers of treatment groups in individual time intervals will be different, and there is no clear way to combine the results of individual time intervals to obtain the overall results of the tests.

For incidental tumor data analysis, the data in a time interval containing 0/0s, that is, no death, even just for a single group, were excluded in the analysis. This can result in excluding a big amount of data. Although this problem can be overcome by using different partitions of the study duration, such as using the ad hoc runs procedure

described in the Peto et al. paper, for different incidental tumor types to avoid time intervals containing 0/0s, it may be impractical in terms of time and efforts needed to complete the statistical review of carcinogenicity studies of a new drug submission.

In some widely used commercially developed computer programs, there is a problem in the exact permutation test for trend in incidence in a tumor that is fatal for some animals and is incidental for the other animals. When a tumor types is observed under fatal and incidental contexts of observation, the permutation of the fatal tumor rates will affect the denominators of the incidental tumor rates because a death is changed from fatal to incidental in one treatment group has to be compensated by the change of a death from incidental to fatal in the treatment group in which the original fatal tumor bearing animal is moved to. Table 13 contains an example illustrates the above problem.

One permutation of the fatal incidences moving the fatal tumor-bearing animal from the medium group to the control group will result in the changes in the incidental incidences shown in Table 14.

However, those computer systems just use the fatal and incidental incidences in Table 15 to do the computation without adjusting the changes in the denominators of the incidental incidences. Using frequency tables without proper changes of denominators of the incidental incidence rates will produce invalid final overall results of exact tests.

Since the poly- k tests do not require the cause-of-death information, the above problem of the exact test for a tumor

Table 13 Example Data of Original Fatal and Incidental Incidences in a Given Time Interval of a Given Tumor Type

	Control	Low	Medium	High
Fatal	0/45	0/42	1/42	0/42
Incidental	3/10	0/12	2/12	1/15

Table 14 One Permutation of the Fatal Incidences Resulting in Changes in the Numerators and Denominators of the Incidental Incidences

	Control	Low	Medium	High
Fatal	1/45	0/42	0/42	0/42
Incidental	3/9	0/12	2/13	1/15

Table 15 One Permutation of the Fatal Incidences without Changing the Denominators of the Incidental Incidences

	Control	Low	Medium	High
Fatal	1/45	0/42	0/42	0/42
Incidental	3/10	0/12	2/12	1/15

that is observed under both fatal and incidental contexts of observation in the Peto tests won't be a problem in the exact version of the poly-*k* tests.

ANALYSIS OF DATA FROM TRANSGENIC MOUSE CARCINOGENICITY STUDIES

The high cost (between \$2 and \$3 million) and long time (a minimum of three years) needed to conduct a standard long-term in-vivo carcinogenicity study and the increased insight into the mechanisms of carcinogenicity due to the advances made in molecular biology have led to alternative in-vivo approaches to the assessment of carcinogenicity. People also argue that that genetically altered mice are better animal surrogates for human cancer because they carry some specifically activated oncogenes that are known to function in both human and animal cancers. The International Conference on Harmonization (ICH) has developed a document, accepted by the United States and other countries, titled—"Guidance on Testing for Carcinogenicity of Pharmaceuticals" Federal Register, vol. 63, 8983–8986, 1998) (ICH, 1998).^[2] The guidance outlines experimental approaches to the evaluation of carcinogenic potential that may obviate the necessity for the routine use of two long-term rodent carcinogenicity studies, allowing sponsors either to continue to conduct two long-term rodent carcinogenicity studies or to use the alternative approach of conducting one long-term rodent carcinogenicity study together with a short- or medium-term rodent test. The short- or medium-term rodent test systems include studies such as initiation-promotion in rodents, transgenic rodents, or newborn rodents that provide rapid observation of carcinogenic endpoints in vivo.

Studies using transgenic mice have become the most important alternative to carcinogenicity testing among the short- or medium-term rodent test systems recommended in the ICH guideline. Many new studies of known carcinogens and non-carcinogens from previous two-year bioassays using transgenic rodents have been carried out by the National Toxicological Program (NTP) and by the International Life Science Institute (ILSI) to evaluate the specificity and sensitivity of the alternative test system.^[33] Different strains of transgenic mice (models) have been proposed and used in the alternative system of testing carcinogenicity of pharmaceutical and environmental chemical compounds. We also have seen and reviewed carcinogenicity studies of pharmaceuticals using transgenic mice submitted by drug companies. The following are the main strains (models) having been proposed and used in the studies mentioned above: (i) p53+/- transgenic mice (with knockout of one of the two alleles of the tumor suppression gene p53), (ii) Tg.AC transgenic mice (with genetically initiated skin to induce epidermal papillomas in response to dermal or oral exposure to chemical agents and act as a reporter phenotype of the activities of the tested

chemicals), (iii) Tg.rasH2 transgenic mice (with five or six copies of the stable human c-Ha-ras gene—they were first developed and patented in Japan), and (iv) XPA-/- repair deficient mice (developed in Europe).

The standard study protocol described below has been used in the above studies.

1. Study duration: 26 weeks
2. A positive control group in addition to the regular three or four treatment groups (negative control, low, medium, and high)
3. 15–30 animals per sex/treatment group.
4. Tissues/organs of p53+/-, Tg.rasH2, and XPA-/- repair deficient mice died or terminally sacrificed are microscopically examined for neoplastic and non-neoplastic lesions.
5. In studies using Tg.AC transgenic mice, only data (both incidence rates and numbers of papillomas observed over time) of tumor type of skin papillomas are collected and tested for drug effects.

In general, the exact and asymptotic tests for trend and difference in tumor incidence with or without requirement of cause-of-death information for the traditional two-year study can be applied to carcinogenicity studies using p53+/-, Tg.rasH2, and XPA-/- transgenic mice since the endpoints in these two types of study are the same.

Because of the major differences in designs used—that is, use of 15–30 animals per sex/group and only one or two treated groups and a small number of tumor types developed in animals in the new type of study—the decision rules used in the two-year studies may have to be modified in order to maintain a desirable level of overall false positive rate. Also, because only 15–30 animals per sex/group are used, the power of the trend and pairwise comparison tests should be evaluated.

Methods for analyzing the numbers of skin papilloma in studies using Tg.AC transgenic mice are somewhat different from those for studies using the above models. Incidence rates of skin papilloma (proportions of animals with the tumor) can be analyzed by the same exact or asymptotic tests for trend and difference in incidence with or without requirement of cause-of-death information in regular two-year studies. Individual weekly counts of skin papilloma data can be analyzed by Mann-Whitney and Jonckheere tests for pairwise difference and positive trend, respectively.

However, the more recently developed method by Dunson et al.^[34] that uses data of all time points in one analysis has been using by the statistical reviewers in the FDA's Center for Drug Evaluation and Research (CDER).

The Dunson model for analyzing the number of papillomas is expressed as follows:

$$\begin{aligned} \mu_{ij} &= E(Y_{ij} | M_{ij}, b_i, t_i, d_i) = \exp\{\beta_1 + (b_i + \gamma_1)t_i d_i\} \quad \text{if } M_{ij} = 0 \\ \text{or} \quad \mu_{ij} &= E(Y_{ij} | M_{ij}, b_i, t_i, d_i) = \exp(\beta_2 + \gamma_2 d_i) \quad \text{if } M_{ij} \neq 0 \end{aligned}$$

where

μ_{ij} is the Poisson mean of the number of papillomas on the back of mouse i observed at week j .

$Y_{ij} = M_{i,j+1} - M_{ij}$ is the increase in the maximum papilloma burden for mouse i between week $j-1$ and j . It is assumed that Y_{ij} has a Poisson distribution with mean μ_{ij} .

M_{ij} = maximum ($Z_{i1}, \dots, Z_{i,j-1}$) is the maximum papilloma burden observed for mouse i prior to week j .

Z_{ij} is the number of detectable papillomas on the back of mouse i at week j .

b_i is a mouse specific susceptibility variable. It is assumed to have a normal distribution with mean zero and variance σ^2 .

$t_j = j/T$, and T is the duration of the study.

d_i is the dose level on the log scale for mouse i .

β_1 and β_2 are the intercept parameters related to the rate of appearance of spontaneous papillomas.

γ_1 and γ_2 are the slope parameters associated with exposure.

The first expression of the model relates to tumor onset. If γ_1 is zero, then the dosing of the drug does not affect the probability of an animal developing its initial papilloma during the study. Alternatively, a large estimated value of γ_1 implies that the exposed animals develop papillomas more rapidly, or equivalently, a higher proportion of the exposed animals developed papillomas during the study. The second expression of the model relates to tumor multiplicity. If $\gamma_2 = 0$, then the dosing of the drug does not affect the probability of developing additional papillomas during the study, once an initial papilloma has occurred. The mouse specific variable accounts for possible heterogeneity in response among the animals.

ANALYSIS OF DATA FROM STUDIES USING DUAL CONTROL GROUPS

There are two categories of studies with dual control groups. The first category (Category A) includes studies using an untreated control group and a vehicle control group. The second category (Category B) includes studies that use two identical control groups.^[35,36]

The main reasons for using an untreated control and a vehicle control in a study in Category A are to check if the vehicle substance has effects, such as spontaneous tumor incidence and pattern, body weight, and food consumption (in dietary studies) on the test animals, and to make sure that the control animals are subjected to the same method and the route of exposure (e.g., by gavage or by injection) experienced by the treated animals so that all animals will be under equal physiological responses and stress (i.e., to isolate the treatment effect from other possible effects).^[6,37]

There are arguments for and against using two identical controls in a study in Category B. The arguments for using

this type of design are to employ the results from the two identical controls as a quality control mechanism for identifying unsuspected biases in the study^[6] and to evaluate the biological significance of increases in tumor incidence in the treated groups (i.e., true increases vs. noises). However, as described later, difficulties are encountered in statistical analysis of data from a study in this category. Also, there is a concern about the possible existence of extrabinomial within-study variability between the two identical groups that will result in an inflation of false-positive rates. These difficulties and concerns form the basis for the arguments against the use of designs with two identical controls.

Statisticians and pharmacologists/toxicologists should work together to decide which of the two control groups is appropriate for the analysis of data from a study in Category A. There are situations in which only analyses of data of the vehicle control and the treated groups are meaningful. There are other situations in which three sets of analysis are performed: control 1 versus treated groups, control 2 versus treated groups, and control 1 plus control 2 versus treated groups. As the concerns about the possible effects of the vehicle substance on the test animals are the reasons for using the vehicle control in addition to the untreated control, it is also of interest to compare the mortality, the tumor rates, the body weight, and the food consumption (in dietary studies) between the two control groups.

Depending on mortality and tumor rates, data from dual control groups may or may not be combined for statistical analysis of data from studies in Category B. If comparisons of the dual controls show no major differences in mortality and tumor rate, then the data of the two controls are combined to form a single control group in subsequent analyses.^[26] If the data show evidences of major differences in mortality or tumor incidence between the identical controls, then three tests for each tumor/organ combination are carried out: control 1 versus treated groups, control 2 versus treated groups, and control 1 plus control 2 versus treated groups.

There is the question of how to interpret the results of a study in Category B in the second case. Two approaches to the question exist. The first approach is that a trend or a difference in tumor rate is considered as significant only if it is significant at the levels of significance described in "Adjustment for the Effect of Multiple Tests (Control over False-Positive Error)" in all the three tests. The second approach is that the trend or the difference is considered significant as long as any one of the three tests shows a significant result. The first approach may be very conservative in the sense that the null hypothesis will be rejected much less often than it should be. On the other hand, the second approach may result in a high false-positive error.

There is no information about appropriate levels of significance for the tests of these two approaches in order to maintain the 10% overall false-positive rate or about the

exact consequences of the two approaches in the second practice of dealing with dual controls with differences in tumor incidence or mortality. Unless all three tests yield consistent results—that is, all are significant or all are not significant from the statistical perspective—the most prudent way of interpreting the test results under this circumstance may be either to regard the study as providing equivocal evidence of carcinogenicity or to consider the study as inadequate for meaningful evaluation.^[38] Alternatively, from the biological perspective, the dual-control data may be viewed as equivalent to contemporary historical data. In such a biological evaluation of the data, consideration of other appropriate historical control data is essential.

Recently, the statisticians and pharmacologists/toxicologists reviewing carcinogenicity studies of pharmaceuticals in the FDA's Center for Drug Evaluation and Research (CDER) have reached an agreement that data of the dual identical control groups should always be pooled in the statistical analysis of the tumor data. However, the incidence rates of tumors of the two dual identical control groups should be tabulated separately with those of the treated groups.

INTERPRETATION OF RESULTS OF STATISTICAL TESTS

Interpreting the results of carcinogenicity experiments is a complex process, and there are risks of both false-negative and false-positive results. The relatively small number of animals used and the low tumor incidence rates can cause a carcinogenic drug not to be detected (i.e., a false-negative error is committed). Because of the large number of comparisons involved (usually two species, two sexes, 20–30 tissues examined), there is also a great potential for finding statistically significant positive trends or differences in some tumor types that are due to chance alone (i.e., a false-positive error is committed). Therefore, it is important that an overall evaluation of the carcinogenic potential of a drug takes into account the multiplicity of statistical significance of positive trends and differences and that it makes use of historical information and other information related to biological relevance (e.g., positive findings at the same site in the other sex and/or in the other species, and evidence of increased preneoplastic lesions at the target organs/tissues).

Adjustment for the Effect of Multiple Tests (Control over False-Positive Error)

It is well known that trend tests are more powerful (i.e., with higher probability) than pairwise comparisons for detecting a true carcinogen. Also, it is difficult to justify the use of data from only the control and the high-dose groups in a study having four or five treatment groups. Therefore,

it is believed that tests for trend, either Peto tests or poly-*k* tests, instead of pairwise comparison tests between control and high-dose groups, should be the primary tests in the evaluation of drug-related increases in tumor rate.

Statistical and nonstatistical procedures have been proposed for controlling the overall false-positive rate. Surveys of some of those procedures can be found in Lin and Ali,^[11] Rahman and Lin^[15] and Fairweather et al.^[24] Here only the statistical decision rules for controlling the overall false-positive rates associated with trend tests and pairwise comparisons in interpreting the final results of carcinogenicity studies are discussed.

In the past, statisticians used the statistical decision rule developed by Haseman^[39] in tests for significance of trends in tumor incidence. The decision rule was originally developed for pairwise comparison tests in tumor incidence between the control and the high-dose groups and was derived from results of carcinogenicity studies conducted in the NTP. The decision rule tests the significance differences in tumor incidence between the control and the dose groups at the 0.05 level for rare tumors and at the 0.01 level for common tumors. A tumor type with a background rate of 1% or less is classified as rare by Haseman and as common otherwise. Haseman's original study and a second study using more recent data with higher tumor rates show that the use of this decision rule in the control-high pairwise comparison tests would result in an overall false-positive rate between 7% and 8% and between 10% and 11%, respectively.^[39,40]

There has been some concern about the possibility of an excessive overall false-positive error when the Haseman decision rule is applied to the trend tests because data from all treatment groups are used in the tests. Unpublished results from recent studies within FDA show that this concern is valid. Based on those studies, the overall false-positive error associated with trend tests with the use of the Haseman decision rule just mentioned is about twice as large as that associated with control-high pairwise comparison tests.

Based on recent studies using real historical control data of CD-1 mice and CD rats from Charles River Laboratory and on simulation studies conducted by the author and in collaboration with the NTP, a new statistical decision rule for tests for positive trend in tumor incidence has been developed. The new decision rule tests, again either Peto tests or poly-*k* tests, for the positive trend in incidence rates in rare and common tumors at the 0.025 and 0.005 levels of significance, respectively. The new decision rule achieves an overall false-positive rate of around 10% in a standard two-species and two-sex study.^[28,41] The 10% overall false-positive rate is seen by regulatory statisticians as appropriate in a new-drug regulatory setting.

Statistical literature concentrates on the methodology of tests for positive trend in tumor rate for the two reasons already mentioned. However, there are situations

in which pairwise comparisons between control and individual treated groups may be more appropriate than trend tests based on the following pharmacological consideration: Trend tests assume that a response of the biological system (carcinogenic effect) is related to doses or systemic exposure weights or ranks. The assumption may be true for simple, direct-acting carcinogens in studies not complicated by excessive toxicity. However, there are many cases in which the response to a drug metabolite is mediated through receptor (or enzyme) effects that may be saturated even at the low dose, or is compounded by dose-related toxicity, or is complicated by a myriad of other nonlinear forms. Under those situations, the original decision rule by Haseman^[39] should be used in interpreting the results of the pairwise comparison tests. It is suggested that both trend tests and pairwise comparison tests be conducted.

As mentioned previously, the ICH made provision for drug sponsors either to continue to conduct two long-term rodent carcinogenicity studies or to use the alternative approach of conducting only one long-term rodent carcinogenicity study and a short- or medium-term transgenic rodent test system to evaluate the carcinogenic potential of their drug products. The current strategy agreed to by the regulatory agencies and the associations of pharmaceutical manufacturers of the ICH is to encourage drug sponsors to use the alternative approach to test the carcinogenicity of their products. Unpublished results from a separate study by the author using historical control data of CD rats and CD mice^[42,43] showed that the use of significance levels 0.05 and 0.01 in tests for positive trend in incidence rates of rare tumors and common tumors, respectively, will result in an overall false-positive rate of around 10% in a study in which only one 2-year rodent bioassay is conducted. The decision rule for testing differences in incidence rates between the control and individual treated groups in a study in which only one 2-year rodent bioassay is conducted is under development and is not available at the moment.

The decision rules for testing positive trend or differences between control and individual treated groups in incidence rates for standard studies using two species and two sexes, and studies following the ICH guideline and using only one 2-year rodent bioassay, are summarized in Table 16.

Use of Historical Control Data

The concurrent control group is always the one that is most appropriate and important in testing drug-related increases in tumor rates in a carcinogenicity experiment.^[44] However, if used appropriately, historical control data can be very valuable in the final interpretation of the study results. Due to the large study-to-study variations caused by differences in nomenclature, laboratory, time period, and pathologists, it is extremely important that the historical control data chosen should be comparable with the current study. Appropriate historical control data is useful in classifying and assessing the significance of rare tumors. Because statistically significant findings in rare tumors are less likely to be due to false-positive error, they provide strong evidence of carcinogenic potential. For this reason, rare tumors can be tested with less stringent statistical decision rules. However, as mentioned previously, an overall evaluation of the carcinogenic potential of a drug should take into account the multiplicity of statistical significance of positive trends and differences and make use of historical information and other information related to biological relevance. In cases of marginally significant trends, historical control data can be used to help investigators evaluate if those findings are biologically significant. If the tumor rates in the treated groups in the experiment are within the normal ranges of the historical control data, then the marginally significant findings can be discounted. Historical control data can also be used as a quality control mechanism for a carcinogenicity experiment by assessing the reasonableness of the spontaneous tumor rates in the concurrent control group^[44] and for evaluation of disparate findings in dual concurrent controls.

Often the informal use of historical control data in the interpretation of statistical testing results just mentioned above is not satisfactory because the normal ranges of the historical control data are too wide. There are proposed formal statistical procedures that allow the incorporation of appropriate historical control data in tests for trend in tumor rate having been proposed. Tarone,^[45] Hoel,^[46] Hoel and Yanagawa,^[47] and Tamura and Young^[48,49] proposed some empirical procedures using the beta-binomial distribution to model historical control tumor rates and to derive approximate and exact tests for trend. The results from those studies show that the incorporation of the historical control data improves the power of the tests.

Table 16 Statistical Decision Rules Aiming at Controlling Overall False-Positive Rates Associated with Tests for Positive Trend or with Control-High Pairwise Comparisons in Tumor Incidences to Around 10% in Carcinogenicity Studies of Pharmaceuticals

	Tests for positive trend	Control-high pairwise comparisons
Standard studies with two species and two sexes	Common and rare tumors are tested at 0.005 and 0.025 significance levels, respectively	Common and rare tumors are tested at 0.01 and 0.05 significance levels, respectively (by Haseman's rule)
Alternative ICH studies with one species and two sexes	Common and rare tumors are tested at 0.01 and 0.05 significance levels, respectively	Under development and not available at the moment

The greatest improvement of power is shown in the tests for rare tumors. Dempster et al.^[50] proposed a Bayesian procedure to incorporate historical control data in statistical analysis. The procedure uses the assumption that the logits of the historical control tumor rates were normally distributed.

There are some technical issues in the use of these formal statistical procedures. They work well in the situations in which historical data from a large number of studies with relatively large control groups are available for obtaining reliable estimations of the parameters of the prior distributions. The maximum likelihood estimators (MLEs) of the prior parameters were shown to be unstable, and the distributions of the MLEs were skewed to the right, i.e., with bunching above the mean and a long tail below the mean. The skewness is severe for cases in which only historical data of a few small control groups were available. The skewness of the MLEs inflates the type I error of the tests. Also, these procedures were developed to incorporate historical control data into the Cochran-Armitage test for linear trend in tumor incidence. As the Cochran-Armitage test is a survival-unadjusted procedure, these procedures cannot be applied to studies with significant differences in survival among treatment groups.

The formal procedure, discussed in Tarone^[45] and below, has been used by statistical reviewers in the FDA's Center for Drug Evaluation and Research (CDER). The formal statistical procedure works well in situations in which historical data from a large number of studies with relatively large control groups are available to provide reliable estimations of the parameters of the prior distributions. The results from this study show that the incorporation of the historical control data improves the power of the tests. The greatest improvement of power is shown in the tests of rare tumors.

Consider an experiment involving $r + 1$ groups (one control and r treated groups) of animals with increasing doses $d_0 = 0 < d_1 < \dots < d_r$. The numbers of animals in the $r + 1$ are $n_0, n_1, \dots, n_r$. The numbers of animals developed a particular tumor in those groups are $x_0, x_1, \dots, x_r$. To test for an increase in the proportions $\hat{p}_i = X_i/n_i$ with increasing dose level, Cochran and Armitage suggested the test statistic

$$\chi^2 = \left(\sum x_i d_i - \hat{p} \sum n_i d_i \right)^2 / \left\{ \hat{p} \hat{q} \left[\sum n_i d_i^2 - \left(\sum x_i d_i \right)^2 / n \right] \right\}, \quad (1)$$

where $n = \sum n_i$, $x = \sum x_i$, $\hat{p} = x/n$, and $\hat{q} = 1 - \hat{p}$. The summation is from zero to r when there are $r + 1$ treatment groups in an experiment.

The test statistic χ^2 is distributed asymptotically as a chi-square random variable with one degree of freedom.

For a given experiment, the number of animals that develop the tumor in the control group follows the binomial distribution with parameter p .

$$f(x) = f(x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, \dots, n. \quad (2)$$

p is the true spontaneous rate of the tumor, n is the total animals in the control group, and x is the number of animals in the control group developed the tumor.

It is proposed in the paper that the following beta distribution be used to model the distribution of the spontaneous tumor rate p of the control group that varies from experiment to experiment.

$$g(p) = \{ \Gamma(\alpha + \beta) / \Gamma(\alpha) \Gamma(\beta) \} p^{\alpha-1} (1-p)^{\beta-1}, \text{ for } 0 < p < 1.$$

The mean and variance of the beta distribution are

$$E(p) = \alpha / (\alpha + \beta)$$

and

$$V(p) = (\alpha\beta) / \{ (\alpha + \beta)^2 (\alpha + \beta + 1) \}$$

The unknown parameters α and β are estimated by the method of moments, that is, by equating the population mean and variance with the sample mean and variance and solving the two equations to find the estimates of α and β , or by the maximum likelihood method.

The positive trend in tumor incidence with the incorporation of historical control data is tested by the following statistic

$$\tilde{\chi}^2 = \left(\sum x_i d_i - \tilde{p} \sum n_i d_i \right)^2 / \left\{ \tilde{p} \tilde{q} \left[\sum n_i d_i^2 - \left(\sum x_i d_i \right)^2 / \tilde{n} \right] \right\},$$

where $\tilde{n} = n + \hat{\alpha} + \hat{\beta}$, $\tilde{p} = (x + \hat{\alpha}) / \tilde{n}$, $\tilde{q} = 1 - \tilde{p}$, $n = \sum n_i$, $x = \sum x_i$, and $\hat{\alpha}$ and $\hat{\beta}$ are estimates of α and β . The summation is from zero to r when there are $r + 1$ treatment groups in an experiment.

The test statistic $\tilde{\chi}^2$ is distributed asymptotically as a chi-square random variable with one degree of freedom.

Table 17 contains the data of hepatocellular adenoma and hepatocellular adenoma+carcinoma in B6C3F1

Table 17 Tumor Incidence Rates of Hepatocellular Adenoma, Adenoma + Carcinoma in B6C3F1 Female Mice in a Carcinogenicity Study of a New Drug

Species/sex	Tumor	Control 1	Control 2	Low	Medium	High
Mouse/female	Hepatocellular adenoma	6	6	8	9	15
Mouse/female	Hepatocellular adenoma+carcinoma	7	8	10	12	18

Table 18 Historical Control Data of Hepatocellular Adenoma, Hepatocellular Carcinoma, and Hepatocellular Adenoma and Carcinoma Combined in Female B6C3F₁ Mice 17 Studies

Study No.	No. of Livers Examined	Hepatocellular Adenoma	Hepatocellular Carcinoma	Hepatocellular Adenoma + Carcinoma
1	60	6	1	7
2	50	5	1	6
3	27	0	0	0
4	50	20	12	27
5	50	11	13	20
6	50	19	14	24
7	51	15	8	22
8	50	11	2	13
9	51	10	3	13
10	50	8	0	8
11	50	6	1	7
12	50	3	1	4
13	60	17	6	22
14	50	13	7	17
15	50	8	4	11
16	50	12	2	14
17	50	20	11	29
Total	849	184	86	244

Table 19 Comparison of Results of the Cochran-Armitage Tests on Data of Hepatocellular Adenoma, Hepatocellular Carcinoma, and Hepatocellular Adenoma and Carcinoma Combined in Female B6C3F₁ Mice with and without the Use of Historical Control Data

Tumor Type	<i>p</i> -value without the Use of Historical Control Data	<i>p</i> -value with the Use of Historical Control Data
Hepatocellular adenoma	0.0048	0.0090
Hepatocellular adenoma + Carcinoma	0.0020	0.0077

female mice in a carcinogenicity study of a new drug, and Table 18 contains the historical control data of the two tumor types from 17 studies using the same strain of animals. The statistical method of trend test incorporating historical control data described above was applied to the concurrent control data and the historical control data of the 17 previous studies. The results of the Cochran-Armitage trend tests with and without the use of the historical control data are presented in Table 19. The results in the table show that the use of the historical control data in the trend tests significantly increases the *p*-values although they are all statistically significant at 0.05 significance level.

Evaluation of Validity of Designs of Negative Studies

In negative or equivocal studies, i.e., studies for which statistical tests did not detect any statistically significant positive trend or difference in tumor rate,^[51] the statistical reviewers should perform a further evaluation of the validity of the design of the experiment to see if there were sufficient numbers of animals living long enough to get adequate exposure to the chemical and to be at risk of forming late-developing tumors and to see if the doses used were adequate to present a reasonable tumor challenge to the tested animals.

As a rule of thumb, a 50% survival rate of the 50 initial animals in any treatment group between weeks 80 and 90 of a 2-year study may be considered a sufficient number and adequate exposure. However, the percentage can be lower or higher if the number of animals used in each treatment/sex group is larger or smaller than 50, but between 20 and 30 animals should still be alive during these weeks.

The adequacy of doses selected and of the animal tumor challenge in long-term or transgenic rodent carcinogenicity experiments is evaluated by pharmacologists and toxicologists based on the previously mentioned ICH approaches as well as the results of the long-term or transgenic rodent carcinogenicity experiments. To assist the evaluation, statistical reviewers in the FDA's Center for Drug Evaluation and Research (CDER) are often asked to provide analyses of body weight and mortality differences and, occasionally, differences of other data between treated and control groups.

REFERENCES

1. International Conference on Harmonization (ICH). *Guidance for Industry: Dose Selection for Carcinogenicity Studies of Pharmaceuticals*; ICH-SIC 1996.
2. International Conference on Harmonization (ICH). *Guidance for Industry: Guideline on Testing for Carcinogenicity of Pharmaceuticals*; ICH-SIB 1997.
3. Dinse, G.E. A comparison of tumor incidence analyses applicable in single-sacrifice animal experiments. *Stat. Med.* **1994**, *13*, 689–708.
4. Dinse, G.E.; Haseman, J.K. Logistic regression analysis of incidental-tumor data from animal carcinogenicity experiments. *Fundam. Appl. Toxicol.* **1986**, *6*, 751–770.
5. Dinse, G.E.; Lagakos, S.W. Regression analysis of tumor prevalence data. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1983**, *32*, 236–248.
6. Gart, J.J.; Krewski, D.; Lees, P.N.I.; Tarone, R.E.; Wahrendorf, J. *Statistical Methods in Cancer Research, Volume III—The Design and Analysis of Long-Term Animal Experiments*. International Agency for Research on Cancer, World Health Organization 1986.
7. Haseman, J.K. Statistical issues in the design, analysis and interpretation of animal carcinogenicity studies. *Environ. Health Perspect* **1984**, *58*, 385–392.
8. Hoel, D.; Walburg, H. Statistical analysis of survival experiments. *J. Natl. Cancer Inst.* **1972**, *49*, 361–372.
9. Lin, K.K. Peto prevalence method versus regression methods in analyzing incidental tumor data from animal carcinogenicity experiments: an empirical study. 1988; American Statistical Association Annual Meeting Proceedings (Biopharmaceutical Section), New Orleans, LA.
10. U.S. Department of Health and Human Services (2001), *Guidance for Industry, Statistical Aspects of the Design, Analysis, and Interpretation of Chronic Rodent Carcinogenicity Studies of Pharmaceuticals*, Center for Drug Evaluation and Research, Food and Drug Administration, U.S. Department of Health and Human Services, Rockville, MD, May, 2001.
11. Lin, K.K.; Ali, M.W. Statistical Review and Evaluation of Animal Carcinogenicity Studies of Pharmaceuticals. In *Statistics in the Pharmaceutical Industry*, 3rd Ed.; Buncher, C.R.; Tsay, J.Y.; Eds.; Chapman & Hall/CRC: New York, 2006.
12. Peto, R.; Pike, M.C.; Day, N.E.; Gray, R.G.; Lee, P.N.; Parish, S.; Peto, J.; Richards, S.; Wahrendorf, J. Guidelines for Simple, Sensitive Significance Tests for Carcinogenic Effects in Long-Term Animal Experiments. *Long-Term and Short-Term Screening Assays for Carcinogens: A Critical Appraisal*; World Health Organization **1980**, 323.
13. Thomas, D.G.; Breslow, N.; Gart, J.J. Trend and homogeneity analyses of proportions and life table data. *Comput. Biomed. Res.* **1977**, *10*, 373–381.
14. Westfall, P.H.; Young, S.S. *Resampling-Based Multiple Testing, Examples and Methods for P Value Adjustment*; Wiley: New York, 1993.
15. Rahman, A.M.; Lin, K.K. Design and Analysis of Chronic Carcinogenicity Studies of Pharmaceuticals in Rodents. In *Design and Analysis of Clinical Trials with Time-to-Event Endpoints*; Peace, K.E.; Ed.; Chapman & Hall/CRC, Taylor & Francis Group, LLC: Boca Raton, FL, 2009.
16. Thomson, S.F.; Lin, K.K. Design and Summarization, Analysis and Interpretation of Time-to-Tumor Response in Animal Studies: A Bayesian Perspective. In *Design and Analysis of Clinical Trials with Time-to-Event Endpoints*; Peace, K.E.; Ed.; Chapman & Hall/CRC, Taylor & Francis Group, LLC: Boca Raton, FL, 2009.
17. Cox, D.R. Regression models and life tables (with discussion). *J. R. Stat. Soc., Ser. B Stat. Methodol.* **1972**, *34*, 187–220.
18. Breslow, N. A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. *Biometrics* **1970**, *57*, 579–594.
19. Gehan, E.A. A generalized Wilcoxon test for comparing K samples subject to unequal patterns of censorship. *Biometrika* **1965**, *52*, 203–223.
20. Cox, D.R. The analysis of exponentially distributed life-times with two types of failures. *J. R. Stat. Soc., Ser. B Stat. Methodol.* **1959**, *21*, 4121–4421.
21. Tarone, R.E. Tests for trend in life table analysis. *Biometrika* **1975**, *62*, 679–682.
22. Mantel, N.; Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **1959**, *22*, 719–748.
23. Ali, M.W. Exact versus asymptotic tests of trend of tumor prevalence in tumorigenicity experiments: a comparison of P-values for small frequency of tumors. *Drug Inf. J.* **1990**, *24*, 727–737.
24. Fairweather, W.R.; Bhattacharyya, A.; Ceuppens, P.P.; Heimann, G.; Hothorn, L.A.; Kodell, R.L.; Lin, K.K.; Mager, H.; Middleton, B.J.; Slob, W.; Soper, K.A.; Stallard, N.; Ventre, J.; Wright, J. Biostatistical methodology in carcinogenicity studies. *Drug Inf. J.* **1998**, *32*, 401–421.
25. Bailer, A.; Portier, C. Effects of treatment-induced mortality on tests for carcinogenicity in small samples. *Biometrics* **1988**, *44*, 417–431.
26. Bieler, G.S.I.; Williams, R.L. Ratio estimates, the delta method, and quantal response tests for increased carcinogenicity. *Biometrics* **1993**, *49*, 793–801.
27. Chen, J.J.; Lin, K.K.; Huque, M.F.; Arani, R.B. Weighted p-value for animal carcinogenicity trend test. *Biometrics* **2000**, *56*, 586–592.
28. Rahman, A.M.; Lin, K.K. A comparison of false positive rates of Peto and poly-3 methods for long-term carcinogenicity data analysis using multiple comparison adjustment method suggested by Lin and Rahman. *J. Biopharm. Stat.* **2008**, *18* (5), 849–858.
29. Cochran, W. Some methods for strengthening the common χ^2 tests. *Biometrics* **1954**, *10*, 417–451.
30. Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics* **1955**, *11*, 375–386.
31. Armitage, P. *Statistical Methods in Medical Research*; Wiley: New York, 1971.
32. Woodruff, R.S. A simple method for approximating the variance of a complicated estimate. *J. Am. Stat. Assoc.* **1971**, *66*, 411–414.
33. ILSI/Health and Environmental Sciences Institute. Alternatives to carcinogenicity testing project. *Toxicol. Pathol.* **2001**, *64*, 1–198.
34. David, B.; Dunson, D.B.; Haseman, J.K.; van Birgelen, A.P.J.M.; Stasiewicz, S.; Tennant, R.W. Statistical analysis of skin tumor data from Tg.AC mouse bioassays. *Toxicol. Sci.* **2000**, *55*, 293–302.

35. Society for Toxicology. Animal data in hazard evaluation: paths and pitfalls. *Fundam. Appl. Toxicol* **1982**, 2, 101–107.
36. Haseman, J.K.; Winbush, J.S.; O'Donnell, M.W. Use of dual control groups to estimate false positive rates in laboratory animal carcinogenicity studies. *Fundam. Appl. Toxicol.* **1986**, 7, 573–584.
37. Dayan, A.D. Biological Assumptions in Analysis of the Bioassay. In *Carcinogenicity, the Design, Analysis and Interpretation of Long-Term Animal Studies*; Springer-Verlag: New York, 1988.
38. Haseman, J.K.; Hajian, G.; Crump, K.S.; Selwyn, M.R.; Peace, K.E. Dual Control Groups in Rodent Carcinogenicity Studies. In *Statistical Issues in Drug Research and Development*; Marcel Dekker: New York, 1990.
39. Haseman, J.K. A reexamination of false-positive rates for carcinogenesis studies. *Fundam. Appl. Toxicol* **1983**, 3, 334–339.
40. Haseman, J.K. *Personal Communication to Robert Temple, M.D.*; CDER, FDA 1991.
41. Lin, K.K.; Rahman, M.A. Overall false positive rates in tests for linear trend in tumor incidence in animal carcinogenicity studies of new drugs. *J. Biopharm. Stat.* **1998**, 8, 1–22.
42. Lin, K.K. Control of overall false positive rates in animal carcinogenicity studies of pharmaceuticals. Presented at the 1997 FDA Forum on Regulatory Sciences, Bethesda, MD, 1997.
43. Lin, K.K.; Rahman, M.A. false positive rates in tests for trend and differences in tumor incidence in animal carcinogenicity studies of pharmaceuticals under ICH Guideline S1B. Division of Biometrics 2, Center for Drug Evaluation and Research, Food and Drug Administration 1998—Unpublished report.
44. Haseman, J.K.; Huff, J.; Boorman, G.A. Use of historical control data in carcinogenicity studies in rodents. *Toxicol. Pathol.* **1984**, 12, 126–135.
45. Tarone, R.E. The use of historical control information in testing for a trend in proportions. *Biometrics* **1982**, 38, 215–220.
46. Hoel, D.G. Conditional two-sample tests with historical controls. *Contributions to Statistics*; North-Holland: Amsterdam 1983.
47. Hoel, D.G.; Yanagawa, T. Incorporating historical controls in testing for a trend in proportions. *J. Am. Stat. Assoc.* **1986**, 81, 1095–1099.
48. Tamura, R.N.; Young, S.S. The incorporation of historical information in tests of proportions: simulation study of Tarone's procedure. *Biometrics* **1986**, 42, 343–349.
49. Tamura, R.N.; Young, S.S. A stabilized moment estimator for the beta-binomial distribution. *Biometrics* **1987**, 43, 813–824.
50. Dempster, A.P.; Selwyn, M.R.; Weeks, B.J. Combining historical and randomized controls for assessing trends in proportions. *J. Am. Stat. Assoc.* **1983**, 78, 221–227.
51. Chu, K.C.; Cueto, C.; Ward, J.M. Factors in the evaluation of 200 national cancer institute carcinogen bioassays. *J. Toxicol. Environ. Health* **1981**, 8, 251–280.

Carry-Forward Analysis

Naitee Ting

Pfizer Global Research and Development, New London, Connecticut, U.S.A.

Abstract

Carry-forward analysis is a useful tool in analyzing longitudinal clinical data with dropout problems. This entry addresses some of the regulatory considerations regarding the issue of carry-forward analysis, provides statistical methods, and presents a numerical example of this type of analysis.

INTRODUCTION

In clinical studies, data are often collected over time. The data collected over a period of time from subjects participating in these studies are referred to as *longitudinal* data by statisticians and biometricians. In longitudinal studies, each subject contributes multiple observations to the analysis. These observations are assumed to be correlated within a subject. It is frequently the case in clinical research that subjects drop out of a study before completion. This leads to an analytical challenge usually referred to as the *missing-data problem*. A number of references discussed the issues of missing data (e.g., Ref. [1]), and dropouts (e.g., Ref. [2]), as well as how to handle these problems (e.g., Ref. [3]).

OVERVIEW

In a report of clinical study results, it is often desirable to perform an “all randomized subjects” analysis. Implementation of this principle becomes a major challenge in data analysis when subjects drop out of the study, and no subsequent follow-up information is available. Supposing this principle is interpreted as analyzing all of those subjects who were randomized to the study, imputation techniques can be an easy way to implement it; that is, for those subjects who did not complete the study, data are inputted so that these subjects can be included in the “all randomized subjects” analysis.

One popular imputation method used in the pharmaceutical industry is to analyze the last observation carry-forward (LOCF) data. The concept of carrying observations forward can best be illustrated by an example. Suppose in a 4-week clinical study, subjects are evaluated at baseline (week 0) and at weeks 1, 2, 3, and 4 after baseline and the clinical objective is to study the drug effect at week 4. Efficacy measurements are collected at these five time points (weeks 0, 1, 2, 3, and 4). In this study, the primary comparison of treatment effect

is made by comparing changes-of-efficacy measurements from baseline to week 4. Among all subjects randomized to this study, assume that N subjects contribute the baseline measurement and at least one postbaseline measurement. After the study was completed, there were n_1 subjects who dropped out before week 2, an additional n_2 subjects who dropped out before week 3, and finally, n_3 more subjects who dropped out before week 4. The week 4 analysis would contain only $N - n_1 - n_2 - n_3 = n$ subjects, if only observed data are analyzed. In order to analyze all N subjects, data collected from the $(N - n)$ subjects who dropped out must be used. One way is to take the last observed value prior to dropout from each of these subjects and treat them as the final data (i.e., carry forward the last observation from each subject who dropped out before the week 4 evaluation).

This concept is very useful in analyzing data at a fixed time point. In this example, if the objective is to compare the subjects' responses at week 4, then LOCF has been a convenient way to include all of the N subjects. In this case, the focus is to analyze data at one or a few selected time points, instead of evaluating the response for the overall time course. It can be viewed as a type of univariate analysis, rather than the repeated-measure analysis. Under LOCF, it is assumed that the distribution of responses after some subjects have dropped out of the study is not different from the distribution of responses at the last available time point.

The section on “Regulatory Requirements” discusses some of the regulatory considerations regarding this issue. The section on “Statistical Methods” describes some statistical methods for handling carry-forward analysis. A numerical example is presented in the section “A Numerical Example.” Results are discussed in the following section (“Discussion”).

REGULATORY REQUIREMENTS

In general, regulatory guidelines do not cover carry-forward analyses specifically. These types of discussions

usually appear under the “Intention-to-treat” or other related topics. Recently, the International Conference on Harmonization (ICH) proposed a set of documents in an attempt to set global standards for drug development and approval considerations. These documents are generally referred to as the ICH Guidelines. The current version of ICH E9 Guideline^[4] discusses analysis based on “all randomized subjects” and analysis based on “per protocol subjects.” In Section 5.2 of ICH E9, this document states:

The intention-to-treat principle implies that the primary analysis should include all randomized subjects. In practice this ideal may be difficult to achieve.... [A]nalysis sets referred to as “all randomized subjects” may not, in fact, include every subject. For example, it is common practice to exclude from the all randomized set any subjects who failed to take at least one dose of trial medication or any subject without data postrandomization. No analysis is complete unless the potential biases arising from these exclusions are addressed and can be reasonably dismissed.

In practice, clinical study results are analyzed following the intention-to-treat principle. It is obvious from the quoted paragraph that the analysis of “all randomized subjects” may not be easily achieved. Therefore, the idea is to define a group of subjects as close to the “all randomized subjects” as possible, and statistical analyses are performed on this group. When this happens to the case, imputation methods are often used to obtain such a group of subjects. Carry-forward analysis is a natural application of the imputation concept.

Using the example given in the “Introduction” section, the best representative group of intention-to-treat subjects is the group of N subjects who contribute the baseline measurement and at least one postbaseline measurement. It is important to note that this group of subjects can be less than the “all randomized subjects,” because there may be subjects who were randomized to the study but did not contribute either the baseline or any postbaseline measurements, and those subjects are not included. In the analysis of changes from baseline to the week 4 response for these “intention-to-treat” subjects, carry-forward methods can be used to implement this data set.

The ICH E9 (Section 5.2) further states:

The “per protocol” set of subjects, sometimes described as the “valid cases,” the “efficacy” sample or the “evaluable subjects” sample, defines a subset of the data used in the all randomized subjects analysis and is characterized by the following criteria:

- i. the completion of a certain per-specified minimal exposure to the treatment regimen;
- ii. the availability of measurements of the primary variable(s);
- iii. the absence of any major protocol violations including the violation of entry criteria where the nature of

and reasons for these protocol violations should be defined and documented before breaking the blind.

In the previous example, the primary objective is to study the treatment effects after subjects were treated with the study medication for 4 weeks. Hence the group of per protocol subjects is the group of n subjects who completed the 4-week treatment.

ICH E9 further indicates that consistent analytical results across various sets of subjects will support clinical conclusions. In this example, if the intention-to-treat (LOCF), the per protocol subjects (week 4 completer), and other possible analyses demonstrate similar results, then the conclusion would be more robust. In case these results are inconsistent, additional analyses will be needed to investigate the differences further.

In the United States, the Food and Drug Administration (FDA) Guidelines are prepared for diseases or therapeutic areas separately. If carry-forward analysis is relevant to a specific therapeutic area, then the corresponding FDA Guideline may mention it briefly. For example, in the treatment of rheumatoid arthritis, long-term clinical studies are typically carried out, and dropouts are very common in these studies. Hence the FDA Draft Guideline for Rheumatic Diseases^[5] covers this issue. This draft states:

Methods used to handle dropouts, such as the “LOCF” and “completers” analyses, are not fully satisfactory although they have often served as the basis for determining that adequate statistical evidence of efficacy has been provided. The LOCF method generally does not preserve the size of the test.

This indicates that clinical conclusions based only on LOCF and completer analysis may not be robust, so additional analyses may also be necessary. The advantages and disadvantages of imputation methods (in particular, LOCF analysis) are discussed in “Discussion.”

STATISTICAL METHODS

Statistical methods used to analyze the carry-forward data are the same as those used to analyze the original data. For example, if the proposed analytical method is to apply analysis of variance (ANOVA) on the original data, then ANOVA can be applied to analyze the carry-forward data.

In clinical trials, response measurements are often disease specific. Some of these measurements include blood pressure (for cardiovascular diseases), forced expiration volume at 1 sec (FEV_1 , for pulmonary diseases), hemoglobin A_{1c} (HbA_{1c} , for diabetes), number of swollen joints (for rheumatoid arthritis), and brief psychiatric reading scale (BPRS, for schizophrenia) are usually treated as continuous variables in data analysis. Hence the carry-forward data from these variables are also analyzed as continuous

variables. For these data, *t*-test, ANOVA, analysis of covariance (ANCOVA), other linear models, or nonlinear models are often used to perform the statistical analysis.

In many cases, categorical data such as physician assessment (no symptom, mild, moderate, severe, and very severe), pain relief (5-point scale, 0 indicates no pain relief, 4 indicates complete pain relief), number of emetic events in a given time interval, and patient response scale (markedly deteriorated, moderately deteriorated, deteriorated, no change, improved, moderately improved, and markedly improved) are used as response variables. Hence the carry-forward data are categorical. In these cases, the methods used to analyze categorical data can also be applied to analyze the carry-forward data. Under certain circumstances, nonparametric techniques are applied for the analysis of the original data; these nonparametric methods can then be applied to analyze the carry-forward data.

There are also other ways of carrying forward observations are applied in clinical trials. Examples include carrying forward the worst case, carrying forward the best case, and carrying forward baseline. In the case of carrying forward the worst case, it is assumed that the subject's condition deteriorated after dropping out, and the worst value the subject ever experienced during the study is used as the analysis value. When the best case is carried forward, it is assumed that the subject dropped out because the subject recovered. In another situation, baseline values may be carried forward for the analysis. Carrying forward baseline requires the assumption that the treatments have no effect on responses for those subjects who dropped out. It is not a common practice to carry forward baselines. One extreme type of imputation in a study comparing a test drug against placebo is to carry forward the best values of placebo-treated subjects and the worst values for test-drug-treated subjects.

Carry-forward analysis can also be viewed as carrying forward the weighted-average responses (see, e.g., Ref. [6]), carrying forward the linear trend, or using other modeling approaches to impute data for subjects who dropped out. One simple way of carrying forward weighted-average responses is to compute the mean of all the available observations after baseline for each subject, to treat this mean as the response measurement for the subject, and then to analyze these individual means. Carry-forward linear trend can be implemented by estimating the slope for each subject first and then analyzing these slopes as subject responses. Some of these methods can be illustrated by the numerical example in the next section.

A NUMERICAL EXAMPLE

Simulated data are used to demonstrate the idea of carry-forward analysis in this numerical example. Suppose that, in a 4-week clinical trial to study subject responses to a new drug, 80 subjects were randomized to be treated with either the new drug (treatment B) or placebo (treatment A). Each subject visited the clinic five times—baseline (before the subject was dosed with the randomized study medication), 1, 2, 3, and 4 weeks after baseline. For this particular study, the primary response variable is a symptom score, possible values can be between 0 (no symptom) and 20 (worst case), inclusive. Data from some of these subjects are given in Table 1; the number of subjects at each visit by treatment group is summarized in Table 2.

In Table 1, we see that subjects 1, 4, 6, 8, 10,..., 80 were randomized to be treated with placebo (treatment A), and subjects 2, 3, 5, 7, 9,... were treated with the new drug (treatment B). For subject 1, the baseline value (BAS) was 18; that is, this subject started with a symptom score of 18 before receiving any dose of the study medication

Table 1 Simulated Data from Some of These Subjects

Treatment	Subject	BAS	W1	W2	W3	W4	LOCF	WOCF
A	1	18	16	20	16	18	18	18
B	2	13	11	13	9	10	10	10
B	3	15	17	14	11	.	11	17
A	4	13	13	13	15	.	15	15
B	5	19	20	20	18	14	14	14
A	6	7	5	5	8	9	9	9
B	7	12	9	.	.	.	9	9
A	8	12	17	11	9	12	12	12
B	9	12	11	13	14	.	14	14
A	10	14	16	15	10	.	10	16
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
A	80	15	13	13	14	17	17	17

Table 2 Number of Subjects at Each Visit by Treatment Group

Treatment	Baseline	Week 1	Week 2	Week 3	Week 4
A	40	40	39	32	29
B	40	40	34	32	28

(placebo, for this subject). The symptom improved slightly after the first week of treatment week 1 (W1) value was 16, and then deteriorated to 20 at week 2 (W2), and became 16 at week 3. After 4 weeks of treatment, the symptom score of this subject returned to a value of 18 (W4). In this case, the last observation carried forward (LOCF) is 18, and the worst case carried forward (WOCF) was also 18. Subject 3 visited the clinic for the first 3 weeks of treatment but did not have the week 4 evaluation. Hence the LOCF is to carry forward the last available value: 11 (LOCF = 11), but the worst case this subject experienced was 17 (from week 1), and thus the WOCF is 17. In this example, it is assumed that most subjects dropped out of the study because of lack of disease improvement. Hence carry-forward best values are not included in this example. Baselines are not carried forward here, either.

In [Table 2](#), we see that 40 subjects were randomized to each treatment group ($N = 40$ for both A and B). No subject dropped out before week 1 visit. One and six subjects dropped out after week 1 from treatments A ($n_1 = 1$) and B ($n_1 = 6$), respectively. For treatment group A, $n_2 = 7$, $n_3 = 3$, and $n = 29$. For treatment group B, $n_2 = 2$, $n_3 = 4$, and $n = 28$.

The objective of this study is to demonstrate that the new drug improves the disease symptom more than the placebo does. Before entering the study, subjects with the disease under study were randomized at baseline. After 4 weeks of treatment, the improvement can be measured using the difference in symptom score from baseline to week 4. In this example, two-sample t -tests are used to make these comparisons. [Table 3](#) illustrates these analytical results.

Note in [Table 3](#) that the BAS result is the comparison of baseline symptom scores (measured before subjects took any study medication) between treatment groups A and B.

DW4 is the comparison of differences from baseline to the week 4 visit; that is, for each subject, DW4 is obtained by subtracting the baseline score from the corresponding week 4 score. Then the DW4 values are compared between the two treatment groups. DLOCF denotes the difference from baseline to LOCF; DWOCF denotes the difference from baseline to WOCF. In these differences, a negative value indicates an improvement. The results in [Table 3](#) are part of the SAS output from the PROC TTEST based on the assumption that treatments A and B have equal variance for all four variables.

In [Table 3](#), the comparison under BAS (baseline) demonstrates that 40 subjects are randomized to each of treatment A and B ($N = 40$ for TRT = A, and $N = 40$ for TRT = B). Subjects in treatment group A have a baseline mean score of 12.775 with a standard deviation of 3.18. Subjects in treatment group B have a mean score of 11.85 with a standard deviation of 3.33. When comparing these two treatment groups, the t -test value (under the equal-variance assumption) is 1.2696 with 78 degrees of freedom (DF), and the p -value is 0.2080, indicating no significant difference between treatments A and B at baseline.

Comparisons under DW4 indicate that treatment A has 29 subjects completing the week 4 visit ($N = 29$), with a mean improvement of 0.552 units in symptom score and a standard deviation of 2.028. Treatment B has 28 subjects with a mean of 1.786 and a standard deviation of 2.079. Under the equal-variance assumption, the t -test value is 2.2683, which is significant at 0.0273 ($p < 0.05$), with 55 DF . This indicates that when the 4-week completers (subjects who completed 4 weeks of treatment) are analyzed, the new drug improves disease symptom by 1.786 units, which is significantly more than with the placebo (0.552 units).

DLOCF provides 40 subjects per treatment group because with LOCF, every subject has an improvement score to be analyzed, regardless of whether a subject dropped out or not. Mean improvement from the new drug treatment (B) is 1.65 units, which is significantly more ($p = 0.0099$, assuming equal variance) than the 0.425 units from the placebo-treated subjects. Similarly, DWOCF

Table 3 Results Obtained from t -Tests

Variable	TRT	N	Mean	Standard deviation	Standard error	T	DF	Prob T
BAS	A	40	12.77500000	3.18238340	0.50317900	1.2696	78.0	0.2080
	B	40	11.85000000	3.33243578	0.52690436			
DW4	A	29	-0.55172414	2.02812734	0.37661379	2.2683	55.0	0.0273
	B	28	-1.78571429	2.07912273	0.39291726			
DLOCF	A	40	-0.42500000	1.98568596	0.31396452	2.6440	78.0	0.0099
	B	40	-1.65000000	2.15489901	0.34071945			
DWOCF	A	40	-0.25000000	1.91819894	0.30329388	2.3578	78.0	0.0209
	B	40	-1.32500000	2.15296456	0.34041359			

indicates a mean improvement from the new drug of 1.325, which is significantly more ($p = 0.0209$, assuming equal variance) than the 0.25 from the placebo treatment, based on 40 subjects per treatment group.

Based on the results in Table 3, the 80 subjects entered in the study with comparable baseline disease symptoms ($p = 0.2080$). After 4 weeks, 29 and 28 subjects completed the study treatment of the placebo and the new drug, respectively. By analyzing the differences from baseline to week 4 (DW4), we see that the new drug provides a greater mean improvement than the placebo treatment. Using last observation carry-forward (DLOCF) analysis, the mean improvements from both treatment groups are less than those from the week 4 visit (1.65 vs. 1.786 for treatment B, 0.425 vs. 0.552 for treatment A). This fact indicates that the subjects who dropped out before week 4 tended to experience less improvement than the subjects who completed the 4-week treatment. As expected, worst observation carry forward (DWOFC) provides the least improvement among these three comparisons ($1.325 < 1.65 < 1.786$ for treatment B, and $0.25 < 0.425 < 0.552$ for treatment A). All three analyses (DW4, DLOCF, and DWOFC) provide consistent results that the new drug helps improve the disease symptom more than the placebo does.

In addition to the preceding analyses, these simulated data were used to perform the average change (DAVG) and slope analyses. Average change is computed by taking the mean of all postbaseline measures for each subject first and then subtracting the corresponding baseline of the subject. Let us use subject 1 as an example: this subject has scored 16, 20, 16, and 18 in weeks 1, 2, 3, and 4, respectively. The average of these four values is 17.5. Subtracting baseline 18 from 17.5 gives a DAVG of -0.5 for subject 1. A two-sample t -test is then applied to analyze these DAVG scores between treatment groups A and B.

Slope is estimated from each subject using simple linear regression (using PROC REG in the SAS software). These individual slopes are then used as the response variable in a two-sample t -test. The results from DAVG and SLOPE are given in Table 4.

The DAVG result in Table 4 indicates that the mean of average improvements are 1.15 for treatment B and 0.381 for treatment A, from 40 subjects of each treatment group. A two-sample t -test under the equal-variance assumption demonstrates a significant difference ($p = 0.0025$)

in favor of treatment B. For the SLOPE comparison, an F -test indicates the variance for treatment A is different from the variance for treatment B ($p = 0.006$). Hence an unequal-variance t -test is applied to compare these slopes. From the t -test, the mean slope of -0.5475 for treatment B is significantly different ($p = 0.0145$) from the -0.075 for treatment A, indicating a more rapid improvement can be obtained from the new drug, as compared to the placebo. Both DAVG results and SLOPE results are consistent with the findings from Table 3.

DISCUSSION

As mentioned previously, carry-forward analysis is one of the imputation methods used in analyzing clinical data. The primary advantages of carry-forward analysis include the following: It is convenient to implement for the intention-to-treat population, it is easy to communicate to nonstatisticians, and the results obtained from these analyses can be clearly interpreted. When an experimental drug is filed for regulatory approval, usually a large number of studies using various types of variables are combined to formulate an overall conclusion. In this case, a method applied consistently across different variables and studies can be very useful. Another advantage of carry-forward analysis is that this method can be consistently applied to all types of response variables and to different study designs.

However, there are a few disadvantages in the carry-forward methods. One of them is that the estimates obtained from these results can be biased. When the distribution of responses from subjects who completed the study is different from the distribution from those who dropped out, carry-forward analysis may provide biased estimates. The bias can affect the mean estimate, the variance estimate, or both. Biased estimates from carry-forward analysis may lead to incorrect conclusions from the study. Another disadvantage is that the carry-forward analysis may reduce power. When this is the case, clinical conclusions may also be misleading. The section on "Statistical Methods" cited an extreme case-imputing the best values for placebo-treated subjects and imputing worst values for test drug-treated subjects. Apparently, this method may create more bias.

Table 4 Additional Results from t -Tests

Variable	TRT	N	Mean	Standard deviation	Standard error	T	DF	Prob > T
DAVG	A	40	-0.38125000	0.92736329	0.14662901	3.1271	78.0	0.0025
	B	40	-1.15000000	1.24796130	0.19732001			
SLOPE	A	40	-0.07500000	0.58474189	0.09245581	2.5152	61.6	0.0145
	B	40	-0.54750000	1.03428100	0.16353418			

For H_0 : variances are equal, $F' = 3.13$, $DF = (39, 39)$, $\text{Prob} > F' = 0.0006$.

Shao and Zhang^[7] provided a systematic study of merits and limitations of the LCOF based on the ANOVA test. More discussion can be found in Chow and Shao.^[8]

As discussed previously, in addition to carrying forward the last observation, the worst observation, the best observation, or the baseline, a few other types of imputation can also be considered. The carry-forward weighted-average response and carry-forward linear-trend (slope) methods are illustrated in the section on “A Numerical Example.” Analysis using slope as the response variable is based on the assumption that the rate of change for each subject does not change after dropout. These variables can be analyzed using parametric or nonparametric methods, depending on distributional properties.

In general, imputation methods are frequently used in dealing with missing-data problems. Most of the advantages and disadvantages for imputation methods are similar to those for carry-forward analysis. Because these methods are widely used in application, Rubin^[9] developed multiple-imputation methods to overcome some of the disadvantages from single imputations. The idea of multiple imputation is to impute more than one possible value to each missing data point and to obtain multiple “imputed data sets.” Statistical analyses are then performed on these data sets, and results from the multiple analyses are combined to formulate inferences. A similar idea is to perform sensitivity analysis by examining results from many carry-forward analyses. In the section on “A Numerical Example,” analyses performed on the completers (DW4), DLOCF, DWOFC, average response, and individual slope all demonstrate similar results. This is a case where the clinical conclusion (that the new drug provides more benefit to the subject than the placebo does) can be supported from these analyses. In case different analyses demonstrate different results, further investigation will be needed to identify the reasons for the inconsistencies.

Before choosing appropriate methods to analyze clinical data, the statistician needs to consider the disease under study, the type of design, the study objective, the primary and secondary response variables for analysis, the distributional properties of these response variables, and the clinical interpretation of the results. Subjects often drop out of the study before completion. Under this circumstance, carry-forward methods can be considered for the data analysis. Parametric or nonparametric methods

can be used. When parametric methods are used—and in particular, linear or nonlinear models are chosen to perform the data analysis—potential covariates should be considered for inclusion in these models. Sometimes nonparametric analysis may be preferred.

CONCLUSION

Overall, carry-forward analysis is a useful tool in analyzing longitudinal clinical data with dropout problems. However, caution must be taken before choosing these methods and interpreting the analytical results. Sensitivity analysis often will be needed to check the robustness of clinical conclusions made from these analyses.

REFERENCES

1. Rubin, D.B. Inference and missing data. *Biometrika*. **1976**, 63, 581–592.
2. Diggle, P.; Kenward, M.G. Informative drop-out in longitudinal data analysis. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **1994**, 43 (1), 49–93.
3. Little, R.J.A. Modeling the drop-out mechanism in repeated-measure studies. *J. Am. Stat. Assoc.* **1995**, 90, 1112–1121.
4. International Conference on Harmonization (ICH) of Technical Requirements for Registration of Pharmaceuticals for Human Use-Draft Consensus Guideline E9, Jan. 16, 1997.
5. *Draft Guidance for Industry—Clinical Development Programs for Drugs, Devices, and Biological Products for the Treatments of Rheumatoid Arthritis (RA)*; U.S. Department of Human Services, FDA, Jan. 6, 1997.
6. Frison, L.; Pocock, S.J. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Stat. Med.* **1992**, 11, 1685–1704.
7. Shao, J.; Zhong, B. Last observation carry-forward and intention-to-treat analysis, **2001**. Manuscript submitted for publication.
8. Chow, S.; Shao, J. *Statistics in Drug Research—Methodologies and Recent Developments*; Marcel Dekker, Inc.: New York, 2002.
9. Rubin, D.B. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J. Am. Stat. Assoc.* **1977**, 72, 538–543.

Case-Control Studies: Inference in

Gary King

Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts, U.S.A.

Langche Zeng

Department of Political Science, University of California, San Diego, California, U.S.A.

Abstract

This entry is focused on developing methods that allow valid inferences about all relevant quantities of interest from either type of case-control study when completely ignorant of or only partially knowledgeable about relevant auxiliary population information.

Keywords: Case-control Studies; Data interpretation; Hazard rate; Logistic models; Number needed to treat; Odds ratio; Rate difference; Rate ratio; Relative risk; Risk difference; Risk ratio; Robust Bayesian inference; Statistics.

INTRODUCTION

Classic (or “cumulative”) case-control sampling designs do not admit inferences about quantities of interest other than risk ratios and then only by making the rare events assumption. Probabilities, risk differences, number needed to treat, and other quantities cannot be computed without knowledge of the population incidence fraction. Similarly, density (or “risk set”) case-control sampling designs do not allow inferences about quantities other than the rate ratio. Rates, rate differences, cumulative rates, risks, and other quantities cannot be estimated unless auxiliary information about the underlying cohort such as the number of controls in each full risk set is available. Most scholars who have considered the issue recommend reporting more than just risk and rate ratios, but auxiliary population information needed to do this is not usually available. We address this problem by developing methods that allow valid inferences about all relevant quantities of interest from either type of case-control study when completely ignorant of or only partially knowledgeable about relevant auxiliary population information.

OVERVIEW

Moynihan et al.^[1] express the conclusions of nearly all who have written about the standards of statistical reporting for academic and general audiences:

This is a somewhat revised and extended version of King, G.; Zeng, L. Estimating risk and rate levels, ratios, and differences in case-control studies. *Stat. Med.* **2002**, *21*, 1409–1427.

In general, giving only the absolute or only the relative benefits does not tell the full story; it is more informative if both researchers and the media make data available in both absolute and relative terms. For individual decisions ... consumers need information to weigh the probability of benefit and harm; in such cases it [also] seems desirable for media stories to include actual event [probabilities] with and without treatment.

Unfortunately, existing methods make this consensus methodological advice impossible to follow in the most used research design in many areas of medical research: case-control studies. In practice, medical researchers have historically used classic (i.e., “cumulative”) case-control designs along with the rare events assumption (i.e., that the exposure and non-exposure incidence fractions approach zero) to estimate risk ratios—or they abandon these quantities of interest altogether and merely report odds ratios. Although virtually no one supports the publication of odds ratios alone, this remains the dominant practice in the field.

In recent years, medical researchers have been switching to density sampling, which requires no such assumption for estimating rate ratios. (We discuss the failure-time matched version of density sampling called risk-set sampling by biostatisticians.) Unfortunately, researchers rarely have the population information needed by existing methods to estimate almost any other quantity of interest, such as absolute risks and rates, risk and rate differences, attributable fractions, or numbers needed to treat. We provide a way out of this situation by developing methods of estimating all relevant quantities of interest under classic and density case-control sampling designs when completely ignorant of, or only partially knowledgeable about, the relevant population information.

We begin with theoretical work on cumulative case-control sampling by Manski,^[2,3] who shows that informative bounds on the risk ratio and difference are identified for this sampling design even when no auxiliary population information is available. We build on these results and improve them in several ways to make them more useful in practice. First, we provide a substantial simplification of Manski's risk difference bounds, which also makes estimation feasible. Second, we show how to provide meaningful bounds for a variety of quantities of interest in situations of partial ignorance. Third, we provide confidence intervals for all quantities and a "robust Bayesian" interpretation of our methods that work even for researchers who are completely ignorant of prior information. Fourth, through the reanalysis of the hypothetical example from Manski's work and a replication and extension of an epidemiological study of bacterial pneumonia in individuals infected with human immunodeficiency virus (HIV), we demonstrate that adding information in the way we suggest is quite powerful as it can substantially narrow the bounds on the quantities of interest. Fifth, we extend our methods to the density case-control sampling design and provide informative bounds for all quantities of interest when auxiliary information on the population data is not available or only partially available. Finally, we suggest new reporting standards for applied research and offer software in Stata and in Gauss that implements the methods developed in this paper (available at <http://GKing.Harvard.edu>).

QUANTITIES OF INTEREST

For subject i ($i = 1, \dots, n$), define the outcome variable $Y_{i,(t,t+\Delta t)}$ as 1 when one or more "events" (such as disease incidence) occur in interval $(t, t + \Delta t)$ (for $\Delta t > 0$) and 0 otherwise. The variable t usually indexes time but can denote any continuous variable. In etiological studies, we shall be interested in $Y_{it} \equiv \lim_{\Delta t \rightarrow 0} Y_{i,(t,t+\Delta t)}$. In other studies, such as of perinatal epidemiology, conditions with brief risk periods such as acute intoxication, and some prevalence data, scholars only measure, or only can measure, $Y_{i,(t,t+\Delta t)}$, which we refer to as Y_i , because the observation period in these studies is usually the same for all i .^[4] Define a k -vector of covariates and constant term as X_i .^a In addition, let X_0 and X_c each denote k -vectors of possibly hypothetical values of the explanatory variables (often chosen so that the treatment variable changes and the others remain constant at their means).

Quantities of interest that are generally a function of t include the *rate* (or "hazard rate" or "instantaneous rate"), $\lambda_i(t) = \lim_{\Delta t \rightarrow 0} \Pr(Y_{i,(t,t+\Delta t)} = 1 | Y_{is} = 0, \forall_s < t, X_i) / \Delta t$, and functions of the rate. For example, the *rate ratio*, $rr_i = \lambda_c(t) / \lambda_0(t)$, and the

rate difference, $rd_i = \lambda_c(t) - \lambda_0(t)$, indicate how rates differ as the explanatory variables change from values X_0 to X_c .

Quantities of interest that are cumulated over an interval of time include the *risk* (also called the "probability" or "conditional probability of disease"),

$$\pi_i = \Pr(Y = 1 | X_i) = 1 - e^{-H(T_i, X_i)} = 1 - e^{-\int_t^{t+\Delta t} \lambda_i(s) ds} \quad (1)$$

[where $H(T_i, X_i)$ is the *cumulative hazard rate* for individual i over time interval $(t, t + \Delta t)$], the *risk ratio*, $RR \equiv \Pr(Y = 1 | X_c) / \Pr(Y = 1 | X_0)$, and the *risk difference* (or "first difference," as it is called in political science, or the "attributable risk," as economists and some epidemiologists call it), $RD \equiv \Pr(Y = 1 | X_c) - \Pr(Y = 1 | X_0)$. The probability is evaluated at some values of the explanatory variables, such as X_0 or X_c . The quantity RD is the increase in probability, and rr is the factor by which the probability increases [an $(RR - 1) \times 100$ percent increase] when the explanatory variables change from X_0 to X_c .

The quantities $\lambda(t)$, RR , RD , π , RR , and RD are normally used to study incidence in etiological analyses, but they are sometimes used to study prevalence by changing the definition of an "event." Functions of these quantities, such as the proportionate increase in the risk or rate difference, the attributable fraction, or the expected number of people needed to treat to prevent one adverse event (or, when the "treatment" has an adverse effect, the expected number of people needed to harm to result in one adverse event), can also be computed as a function of these quantities with the methods described below. For example, to compute the number needed to treat (NNT), we merely calculate:

$$MNT = \frac{1}{|RD|} \quad (2)$$

$$= \frac{1}{|\Pr(Y = 1 | X_c) - \Pr(Y = 1 | X_0)|} \quad (3)$$

That is, we simply use the reciprocal of the size of the risk difference between the "treatment" and "control" groups. The number needed to harm is obtained in the same way.

The other quantity frequently discussed and presented in the epidemiological literature is the *odds ratio*

$$\begin{aligned} OR &= \frac{\Pr(Y = 1 | X_c) / \Pr(Y = 0 | X_c)}{\Pr(Y = 1 | X_0) / \Pr(Y = 0 | X_0)} \\ &= \frac{\Pr(X_c | Y = 1) \Pr(X_0 | Y = 0)}{\Pr(X_0 | Y = 1) \Pr(X_c | Y = 0)} \end{aligned} \quad (4)$$

where the second equality holds by Bayes theorem. The main advantage of the OR is that it has been easier to

^aWe assume for the purpose of this paper that the X 's are not time-dependent.

estimate than the other quantities. In logit and other multiplicative intercept models (but not generally), the OR also has the attractive feature of being invariant with respect to the values at which control variables are held constant. The disadvantage of the OR is understanding what it means, and when or is not the quantity of interest—which is almost all the time—then its “advantages” are not sufficient to recommend its use. Some statisticians seem comfortable with the OR as their ultimate quantity of interest, but this is not common. Even more unusual is to find anyone who feels more comfortable with the OR than the other quantities defined above; we have found no author who claims to be more comfortable communicating with the general public using an OR than one of the other quantities.^[5] The OR has been used “largely because it serves as a link between results obtainable from follow-up studies and those obtainable from case-control studies.”^[6] We provide this link with other quantities so that the OR is no longer the only choice available.

Concluding a controversy on this subject in the *British Medical Journal*, Davies et al.^[7] write “On one thing we are in clear agreement: odds ratios can lead to confusion and alternative measures should be used when these are available.”

We show here how to make all relevant alternative measures *always* available, even in case-control data. With the results in this article, scholars should never again feel forced into presenting odds ratios.^b

CASE-CONTROL SAMPLING DESIGNS

We now introduce the two key case-control sampling designs. With appropriate modifications, most of the methods we introduce also work with many variants of them.

Classic (or *cumulative*) *case-control* designs involve sampling (usually all) “cases” (subjects for which $Y_i = 1$) and a random sample of “controls” (subjects for which $Y_i = 0$) at the end of the study period. This design is most appropriate in studies of closed populations and when Y_i but not Y_{it} is observed. Statistical models for such data specify the risk $\Pr(Y_i = 1|X_i)$ as a function of the input X_i . The most commonly used model is the logistic:

$$\Pr(Y_i = 1 | X_i) = \frac{1}{1 + e^{-X_i\beta}} \quad (5)$$

Some examples in medical research using the cumulative case-control design are studies of congenital malformation, analyses of “chronic conditions with ill-defined onset times and limited effects on mortality, such as obesity and multiple sclerosis, and studies of health services utilization”^[8]

and research that has as its goal descriptive, rather than causal, inferences (such as ascertaining the types of people that now have lung cancer so we can better prepare health-care facilities in anticipation).

The newer *density case-control* sampling design involves, in the study of cohort data organized by risk sets, sampling “cases” (subjects for which $Y_{it} = 1$) at their failure times and a subset of “controls” (subjects for which $Y_{it} = 0$) from all individuals at risk at the time of each failure, possibly matched on a set of other control variables. So each *sampled* risk set R_j ($j = 1, \dots, M$, where M is the total number of cases in the data) is composed of one case (or more in the case of timing ties in their occurrence) and a random sample of controls at the same time (or other continuous index) t . A subject may appear in multiple risk sets. This administratively convenient data collection strategy can result in a substantial reduction in the resources required for the study and also has the advantage of controlling non-parametrically (i.e., without functional form assumptions) for all omitted variables, and unmeasured heterogeneity, related to t . Statistical models for such data specify the incidence (hazard) rate as a function of t and X . The most commonly used model is the Cox proportional hazard:

$$\lambda_i(t) = \lambda(t)r(X_i, \beta) \quad (6)$$

where, in most applications, $r(X_i, \beta) = e^{X_i\beta}$ and the baseline hazard $\lambda(t)$ does not vary over subjects.^c

INFERENCE WITHOUT POPULATION INFORMATION

Classic Case-Control

Without additional population information or assumptions, the literature provides no method for estimating any quantity of interest from classic case-control data. Historically, medical researchers have used the *rare events assumption* to estimate the risk ratio by the odds ratio. Denote the population fraction of incident cases by τ , which is the key piece of auxiliary information about the population not reflected in the sample [although other quantities such as $P(X)$, the possibly multivariate density of X , are also sufficient]. The rare events assumption states that τ is arbitrarily small [while $P(X)$ stays bounded away from zero, or instead that $\Pr(Y_i = 1|X_i) \rightarrow 0$ for $j = 0, \ell$]. This assumption is not merely that cases are “rare,” but that they occur, at the limit, with zero probability. The advantage of this assumption is that, when correct, OR is a good approximation to the risk ratio: $\lim_{\tau \rightarrow 0} \text{OR} = \text{RR}$.

^bUnfortunately, the term “relative risk” no longer seems useful because it has been co-opted to denote such diverse quantities as RR and OR.

^cThe proportionality assumption holds when X_i is time-invariant. When this is not appropriate, a model that allows time-dependent covariates (such as the so-called Cox regression model) can be used.

This limit result is attractive because the OR is often easy to estimate in case-control designs. For example, if X is a single binary variable, or could be estimated by replacing elements in the second line of Eq. 4 with their sample analogs [e.g., $\Pr(X_i | Y = 1)$ can be estimated by the fraction of cases for which $X_i = X_c$ among all observations where $Y_i = 1$]. For another example, in logistic regression Eq. 5, case-control sampling only biases the intercept term, which drops out in the odds ratio expression, $OR = e^{(x_t - x_0)\beta}$. The coefficients of the control variables also drop out (because the elements of $X_c - X_0$ corresponding to those parameters are zero).

As is well known (e.g., Ref. [8]), the rare events assumption is problematic when τ is not nearly zero, and as a result, the odds ratio overestimates RR (when both are above 1; OR underestimates RR otherwise). In addition, this assumption implies, implausibly for most applications, that the risk difference tends to zero ($\lim_{\tau \rightarrow 0} RD = 0$), no matter how strong the real effect. In practice, RD is therefore not estimated, even when it is of more interest than RR.

When τ is very small but not zero, the bias introduced by using or as if it were RR can be ignored without practical consequence. The rare events assumption is inappropriate in etiology studies with more commonly occurring events, such as in highly infectious diseases. The assumption is also often not plausible when studying diseases with nonabsorbing states or when studying prevalence.

Density Case-Control

In the commonly used proportional hazard model (6), the coefficients β can be consistently estimated without any auxiliary information on the population risk sets. The contribution of each sampled risk set to the likelihood function is the ex ante incidence probability for the individual who will actually acquire the disease, conditional on the total number of cases in the set being one:

$$\Pr(Y_{it_j} = 1 | R_j) = \frac{\Pr(Y_{it_j} = 1)}{\sum_{k \in R_j} \Pr(Y_{kt_j} = 1)} = \frac{e^{x_{it}\beta}}{\sum_{k \in R_j} e^{x_{kt}\beta}} \quad (7)$$

where i indexes the case in the risk set and the summation in each denominator is over all observations in the risk set, $k = 1, \dots, n_j$, where n_j is the number of observations, or the size, of R_j . The likelihood function then is the product of terms such as Eq. 7 over all M sampled risk sets (see, e.g., Ref. [9]). When a risk set includes multiple cases, because of timing ties, the conditional probability expression is more complicated, but the approach remains the same. This same estimator was independently proposed in econometrics by Chamberlain^[10] to estimate consistently the parameters in a fixed-effect logit model when the number of subjects per group remains fixed even as the total number of subjects grows. Note also that Eq. 7 takes the same

form as the probability expressions in the popular multinomial logit model (or McFadden's choice model as some call it), and indeed the same software can be used on the data organized in risk sets to estimate β . Eq. 7 also takes the same form as the conditional probability expression for matched cohort data (for which Greenland^[11] gives an alternative modeling strategy for estimating the risk ratio).

Eq. 7 takes the same form as the likelihood function contribution of a *full* risk set in a Cox regression model for full cohort data, only that now the *sampled* risk sets are used. Fortunately, this does not bias the first-order condition and preserves the consistency and asymptotic normality of the maximum likelihood estimator (see Refs. [12–14]). Moreover, values of n_j of only about 6 can result in nearly full efficiency.^[12,15] For a discussion of biased selection of the controls, see Refs. [16,17] and see Ref. [18] for a discussion of alternative sampling designs.

With β consistently estimated by the conditional logit approach, we can compute the rate ratio using the incidence model (6), although the baseline hazard rate $\lambda(t)$ is not estimated:

$$rr = \frac{\lambda_t(t)}{\lambda_0(t)} = \frac{\lambda(t)e^{x_{it}\beta}}{\lambda(t)e^{x_{0t}\beta}} = \frac{e^{x_{it}\beta}}{e^{x_{0t}\beta}} = e^{(x_{0t} - x_{it})\beta} \quad (8)$$

Thus a key advantage of density designs over classic case-control designs is that we can estimate rr without any additional (rare events or other) assumptions or any population information. However, without auxiliary information on the population, the literature provides no method for estimating any other quantity of interest, such as the rate, rate difference, the risk, risk difference, risk ratio, or number needed to treat.

INFERENCE UNDER FULL POPULATION INFORMATION

We now discuss the quantities of interest that can be estimated under each sampling design when required auxiliary population information is available.

Classic Case-Control

In the classic case-control design, the population fraction of cases, τ , can supply the required auxiliary information.^[19–21] For example, in the simple case of a single binary explanatory variable, the risk can be estimated as

$$\begin{aligned} \pi &= \Pr(Y = 1 | X, \tau) \\ &= \frac{P(X | Y = 1) \Pr(Y = 1)}{P(X)} \\ &= \frac{P(X | Y = 1) \tau}{P(X | Y = 1) \tau + P(X | Y = 0) (1 - \tau)} \end{aligned} \quad (9)$$

where $P(X|Y = 1)$ and $P(X|Y = 0)$ are replaced with the sample mean of X among subjects where $Y = 1$ and $Y = 0$, respectively. From this expression, we can compute $RR = \Pr(Y = 1|X_{\tau})/\Pr(Y = 1|X_0, \tau)$, $RD = \Pr(Y = 1|X_{\tau}) - \Pr(Y = 1|X_0, \tau)$, or $NNT = 1/|rd|$.

For another example, the commonly used logistic regression model (5) on case-control data yields consistent maximum likelihood estimates of the slope coefficients and only the estimated intercept β_0 requires a correction such that the quantity estimated should instead be

$$\beta_0 - \ln \left[\left(\frac{1-\tau}{\tau} \right) \left(\frac{\bar{y}}{1-\bar{y}} \right) \right] \quad (10)$$

where $\bar{y}$ is the mean of Y or the sampling probability of $Y = 1$. Thus with exact knowledge of τ , all logit parameters can be consistently estimated (see Refs. [22–26]; see Ref. [19] for a comprehensive review). Let β_{τ} denote the logit coefficient vector, with the first element corrected as in Eq. 10. Then the quantities of interest are

$$\begin{aligned} \pi &= \Pr(Y = 1 | X, \tau) = [1 + e^{-X\beta_{\tau}}]^{-1}, \\ RR &= \frac{[1 + e^{-X_{\tau}\beta_{\tau}}]^{-1}}{[1 + e^{-X_0\beta_{\tau}}]^{-1}}, \\ RD &= [1 + e^{-X_{\tau}\beta_{\tau}}]^{-1} - [1 + e^{-X_0\beta_{\tau}}]^{-1}, \\ NNT &= \frac{1}{|RD|} \end{aligned}$$

For estimation, we can plug in the maximum likelihood estimates for β or use improved methods with lower mean square error in the presence of rare events.^[26] Standard errors for any of these quantities can be computed easily by simulation.^[27]

The correction in Eq. 10 applies not only to logit models, but also to any multiplicative intercept model, even including many neural networks (such as the feed-forward perceptron with a logit output function^[28,29]). Statistical research in medicine still predominately uses the logit specification for parametric analysis, whereas numerous other scholarly fields have been steadily switching to more flexible functional forms such as neural networks. In other fields, scholars have recognized that the logit model makes specific assumptions that are normally without substantive justification and indeed in the vast majority of research in this field are not even addressed explicitly. This is acknowledged in medical research in its use of matching models, but not when it comes to parametric approaches. Unfortunately, when using highly restrictive parametric forms such as logit, our empirical answers are functions of our theoretical specification rather than our data. By recognizing that the correction above applies to a much broader class of models, and makes considerably less stringent assumptions, researchers using case-control

designs will have many new and far more valid parametric approaches to choose from.

Density Case-Control

To estimate quantities of interest other than rr under density case-control sampling, we require information that enables the estimation of the baseline hazard rate. We employ the baseline hazard rate estimator proposed by Ref. [30] where the key population information needed is $\tau_j = n_j/N_j$, the sampling fraction for each observed risk set R_j ($j = 1, \dots, M$). With this information, and denoting the maximum likelihood estimate of β from the conditional logit procedure as b , the baseline incidence rate λ at time t_j is estimated as:^d

$$\lambda(t_j) = \frac{1}{\sum_{k \in R_j} (e^{X_k b}) (1/\tau_j)} = \frac{1}{\sum_{k \in R_j} e^{X_k b - \ln(\tau_j)}} \quad (11)$$

and with this, we can estimate the rate, $\lambda_i(t)$, from Eq. 6:

$$\lambda_i(t_j) = \frac{e^{X_i b}}{\sum_{k \in R_j} e^{X_k b - \ln(\tau_j)}} \quad (12)$$

With the rate, we can further estimate the cumulative rate:

$$H(T_i, X_i) = \sum_{t_j \in T_i} \lambda_i(t_j) = \sum_{t_j \in T_i} \sum_{k \in R_j} \frac{e^{X_i b}}{e^{X_k b - \ln(\tau_j)}} \quad (13)$$

and hence the risk:

$$\begin{aligned} \pi_i &= \Pr(Y = 1 | X_i) = 1 - e^{-H(T_i, X_i)} \\ &= 1 - \exp \left(- \sum_{t_j \in T_i} \sum_{k \in R_j} \frac{e^{X_i b}}{e^{X_k b - \ln(\tau_j)}} \right) \end{aligned} \quad (14)$$

and any of the other quantities of interest. The factor $1/\tau_j$ serves as a weighting factor that enables us to weight up each risk set to the full risk set size.

INFERENCE WITH BAYESIAN PRIOR ASSUMPTIONS

“Inference Without Population Information” and “Inference under Full Population Information” discuss inference under conditions of ignorance and full knowledge of the

^dThe proportionality assumption holds when X_i is time-invariant. When this is not appropriate, a model that allows time-dependent covariates (such as the so-called Cox regression model) can be used.

relevant auxiliary population information (τ and τ_j). When this information is not known exactly, but some prior information exists, Bayesian methods are appropriate and straightforward (although we are not aware of their use in applications in this context). Ideally, this prior information would come from registry or survey data from the target population, but similar information from closely related populations would help form reasonable priors also. The procedure is simply to put a prior distribution on τ (in classic case-control designs) or τ_j (in density case-control designs) and draw inferences about the quantity of interest from the posterior. These inferences have all the desirable properties of Bayesian estimates. They are consistent, efficient, asymptotically normal, and so on, when averaging over the prior. We strongly support their use when prior information is available and known to be valid.

Bayesian estimates also share at least two disadvantages. First, they are not calculable when the analyst is completely ignorant of τ or τ_j in some or all parts of its range because the required prior density cannot be fully specified, and, in Bayesian inference, one must have a full prior. Second, Bayesian estimates yield biased inferences when prior information is biased. This second problem is especially severe in case-control studies because the data contain *no* information about τ and τ_j , and so the prior does not become dominated by the likelihood as the sample size grows. That is, even for very large samples, inferences from case-control data depend heavily on the prior.

These disadvantages combine in unfortunate ways sometimes when analysts are unsure of the validity of their prior information. In these cases, the classic Bayesian paradigm may, in practice, have the effect of encouraging researchers to guess values for their prior, hence introducing biased information into their analyses and adversely affecting their inferences. “Diffuse” priors with large variances are also no solution here because in general, increasing the variance of τ or τ_j will also make the mean tend toward 0.5, which of course is a statement of knowledge, not ignorance.

We therefore pursue “robust Bayesian” methods below that allow full or partial ignorance to be represented accurately without biasing inferences.

INFERENCES WITHOUT FULL POPULATION INFORMATION OR ASSUMPTIONS: CLASSIC CASE-CONTROL

We now discuss inferences under classic case-control designs about π , RR, and RD without the rare events assumption or population information about τ . We begin by extending Manski’s results in the situation of pure ignorance and then add methods for partial ignorance, confidence intervals, and our preferred robust Bayesian interpretation. We conclude the section with two examples.

Extending Manski’s “Ignorance” Results

As an alternative to full knowledge of τ or the rare events assumption, Manski^[2,3] studied what inferences we could make when the researcher had no knowledge of τ . It is widely known that under these circumstances, no information about risk is available [i.e., $\pi \in (0,1)$] and RR is bounded between 1 and the OR. That is, $RR \in [\min(1,OR), \max(1,OR)]$, apart from sampling error. That this expression depends on the odds ratio is useful because of how often it is easily estimable.

Manski’s advance is that he also shows that RD can be bounded, although it requires the following complicated expression. Let

$$\phi = \left[\frac{\Pr(X_0 | Y=1)\Pr(X_0 | Y=0)}{\Pr(X_1 | Y=1)\Pr(X_1 | Y=0)} \right]^{\frac{1}{2}} \quad (15)$$

and

$$\gamma = \frac{a}{a - [\phi \Pr(X_1 | Y=1) - \Pr(X_0 | Y=1)]}, \quad (16)$$

where $a = \phi \Pr(X_1 | Y=0) - \Pr(X_0 | Y=0)$, and

$$RD_\gamma = \frac{\Pr(X_1 | Y=1)\gamma}{\Pr(X_1 | Y=1)\gamma + \Pr(X_1 | Y=0)(1-\gamma)} - \frac{\Pr(X_0 | Y=1)\gamma}{\Pr(X_0 | Y=1)\gamma + \Pr(X_0 | Y=0)(1-\gamma)} \quad (17)$$

Then RD is bounded between 0 and RD_γ , apart from sampling error, where RD_γ is the value of the risk difference if τ were equal to γ .

Manski’s expression for the risk difference is useful but only when it can be estimated. Unfortunately, except for very simple cases, sophisticated non-parametric methods are required to estimate each of the component probabilities in Eqs. 15–17], and ϕ , γ , and RD_γ are not easy to estimate directly; to our knowledge, they have never been estimated in a real application. We remedy this situation by showing, in [Appendix A](#), that these equations can be simplified so that $RD_\gamma = (\sqrt{OR} - 1) / (\sqrt{OR} + 1)$ (which, surprisingly, is exactly Yule’s^[31] “coefficient of colligation,” sometimes called Yule’s Y). The bounds are thus a simple function of the odds ratio:

$$RD \in \left[\min \left(0, \frac{\sqrt{OR} - 1}{\sqrt{OR} + 1} \right), \max \left(0, \frac{\sqrt{OR} - 1}{\sqrt{OR} + 1} \right) \right] \quad (18)$$

The advantages of our expression in Eq. 18 are not only algebraic simplicity and the familiarity with the odds ratio among applied researchers, but also that OR can be estimated very easily without non-parametric methods in simple discrete cases, in logistic regression, and in a wide

variety of multiplicative intercept models such as neural network models. Because these neural network models have arbitrary approximation capabilities, Eq. 18 can effectively always be applied.

A Proposed “Available Information” Assumption

Applied researchers have been reluctant to adopt Manski’s “ignorance” assumption, perhaps in part because the knowledge they have about τ is discarded entirely, often resulting in very wide bounds on the quantities of interest. Particularly uncomfortable for researchers is that no matter how strong the empirical relationship among the variables is, the bounds on RR always include 1 and on RD always include 0, which in both cases denote no treatment effect.

Thus existing literature effectively requires researchers to choose among three extreme assumptions: τ is essentially zero, is known exactly (possibly apart from sampling error), or is completely unknown. Our alternative approach is to elicit from researchers a range of values into which they are willing to say that τ must fall (e.g., [0.001, 0.05]), which appears to be a better reflection of the nature of prior information available in applied research settings than the extremes of exact knowledge or complete ignorance. Our approach seems consistent with Manski’s^[2] goals for future research and, like his specific methods, does not require a fully Bayesian prior distribution. Our approach could also be applied to bring available information and probabilistic inference to the methods Manski^[2] has offered in other areas.

Let π_τ , RR_τ , RD_τ , and NNT_τ denote values of the probability, relative risk, risk difference, and number needed to treat, respectively, evaluated at τ . Suppose that τ is known only to fall within the range $[\tau_0, \tau_1]$, where $0 < \tau_0 < \tau_1 < 1$. (Because, by definition, choice-based samples include at least one example of a case and one of a control, τ_0 and τ_1 are known not to equal zero or one exactly.) Then, because π and rr are monotonic in τ , their bounds are simply

$$\pi \in [\pi_{\tau_0}, \pi_{\tau_1}] \quad (19)$$

and

$$RR \in [\min(RR_{\tau_0}, RR_{\tau_1}), \max(RR_{\tau_0}, RR_{\tau_1})] \quad (20)$$

The bounds for the risk difference are more complicated because RD_τ is a parabolic function of τ , and so the bounds differ in the monotonic and nonmonotonic regions. The relationship is monotonic in regions where τ_0 and τ_1 are both greater than or both less than the value of τ that corresponds to $RD_\tau = (\sqrt{OR} - 1)/(\sqrt{OR} + 1)$. This region corresponds to cases where the derivative of RD_τ with respect to τ , evaluated at τ_0 and τ_1 , have the same sign. (This derivative can easily be checked numerically by comparing the signs of $RD_{\tau_0+\epsilon} - RD_{\tau_0}$ and $RD_{\tau_1+\epsilon} - RD_{\tau_1}$

for a suitably small value of ϵ .) When the relationship is monotonic, the bounds are

$$RD \in [\min(RD_{\tau_0}, RD_{\tau_1}), \max(RD_{\tau_0}, RD_{\tau_1})] \quad (21)$$

and otherwise, they are

$$RD \in [\min(RD_{\tau_0}, RD_{\tau_1}, RD_\gamma), \max(RD_{\tau_0}, RD_{\tau_1}, RD_\gamma)] \quad (22)$$

The bounds for number needed to treat can be directly derived from those for the risk difference by using reciprocals of the latter.

Revisiting a Numerical Example

We illustrate our methods by extending the numerical example concerning smoking and heart disease given by Manski.^[2,3] For clarity, we follow Manski in ignoring uncertainty (i.e., equating sample fractions with sampling probabilities as if $n \rightarrow \infty$) in this section (only); in the next two sections, we show how to include estimation uncertainty and compute confidence intervals. This section also demonstrates the degree to which results under our approach are sensitive to assumptions about the interval $[\tau_0, \tau_1]$ while holding constant (at zero) estimation uncertainty. (The effect of estimation uncertainty, while holding constant the interval, follows standard sampling theory.)

In Manski’s example, X is a binary explanatory variable taking values 1 for smokers and 0 for non-smokers, and Y takes on the values 1 for coronary heart disease and 0 for healthy individuals. The assumptions in his example imply that $\Pr(X = 1|Y = 1) = 0.6$, $\Pr(X = 1|Y = 0) = 0.49$, $\Pr(X = 0|Y = 1) = 0.4$, and $\Pr(X = 0|Y = 0) = 0.51$. Hence using Eq. 9, we can write the probabilities as functions of τ :

$$\begin{aligned} \Pr(Y = 1 | X = 1, \tau) &= \frac{0.6\tau}{0.6\tau + 0.49(1-\tau)} \\ &= \frac{\tau}{0.82 + 0.18\tau} \\ \Pr(Y = 1 | X = 0, \tau) &= \frac{0.4\tau}{0.4\tau + 0.51(1-\tau)} \\ &= \frac{\tau}{1.28 - 0.28\tau} \end{aligned}$$

For each of the quantities of interest, we now compare the case where τ is unknown, as Manski does, to where it is known to lie in the interval [0.05, 0.15] (Manski’s example implies that $\tau = 0.1$, which he treats as not known). Without bounds on τ , the problem provides no information about any probability, whereas the additional information about τ gives much more informative bounds: the probability of heart disease among smokers is $\Pr(Y = 1|X = 1) \in [0.06, 0.18]$, whereas among non-smokers, it is $\Pr(Y = 1|X = 0) \in [0.04, 0.12]$. For risk ratio, Manski’s “ignorance”

assumption gives $RR \in [1., 1.57]$, whereas our alternative approach implies the very tight bounds of $RR \in [1.46, 1.53]$, indicating that smoking increases the risk of heart disease between 46% and 53%. For the risk difference, the bounds given no information on τ are $RD \in [0, 0.11]$, whereas our approach yields much narrower bounds of $RD \in [0.021, 0.056]$, the increase in probability due to smoking. Because smoking has an adverse effect, we can also obtain bounds for NNT, the number needed to harm, based on that for the risk difference. With no information on τ , Manski's approach would result in $NNT \in [9, \infty]$, which is hardly informative at all, while our method gives $NNT \in [18, 48]$, a range infinitely narrower than "between 9 and infinity" and of course more substantively meaningful.

Classical Confidence Intervals

We now provide a method of computing classical confidence intervals, saving our preferred robust Bayesian interpretation for the following section. Because each end of the bounds on π , RR , and RD are measured with error, upper and lower confidence intervals could be computed and reported for each. However, the inner bounds (the upper confidence limit on the lower bound and the lower confidence limit on the upper bound) are not of interest. Thus we recommend defining a confidence interval (CI) as the range between the outer confidence limits. The actual CI coverage of the resulting interval is always at least as great as the nominal coverage.

For all methods provided above, confidence intervals can easily be computed by simulation (the delta method is also possible but difficult because of the discontinuities caused by the minimum and maximum functions). The bounds are known functions of, and derive their sampling distributions from, the estimated model parameters. Therefore the distribution of the bounds can be simulated using random draws from the sampling distributions of the parameters.^[27] For example, in the logit model, the asymptotic distribution of the estimated parameters is normal with mean vector and covariance matrix estimated by the usual maximum likelihood procedures. Random draws from this distribution can then be converted into random draws from the distribution of the bounds through the relevant formulas relating the bounds and the model parameters. A 90% (for example) CI for the bounds can be obtained by sorting the m random draws of the bounds and taking the 5th and 95th percentile values as the lower and upper bounds, respectively. Our software implements these procedures. The choice of m reflects the tradeoff between accuracy and speed: larger values of m improve accuracy and reduce speed. The required m depends on the example; 1000 will often be enough, but it is easy to verify: rerun the simulation and if anything changes in as many significant digits as is needed, increase m and try again.

A Robust Bayesian Interpretation

Our estimation procedure is not strictly Bayesian in that choosing an interval for τ is not equivalent to imposing a uniform (or any version of a "non-informative") prior density within those bounds. However, our procedure can be thought of as a special case of "robust Bayesian analysis" (e.g., Refs. [32, and 33]) and one that happens to be easier to apply and gives results that are considerably easier for applied researchers to use than most examples in this literature.

From this robust Bayesian perspective, the interval chosen for τ can be thought of as narrowing the choice of a prior to only a class of densities rather than a (fully Bayesian) single density. In our case, the class of priors is defined to include all densities $P(\tau)$ subject to the constraint that $\int_{\tau_0}^{\tau_1} P(\tau) d\tau = 1$. The advantages of this approach are that prior elicitation is much easier, it does not force analysts to give priors when no prior information exists, and, more importantly, estimates depend only on real information and so are the same for any density within the class. If a classical Bayesian prior is inaccurate, classical Bayesian inferences will be incorrect. In contrast, our "robust Bayesian" approach will give valid inferences even if we can only narrow the prior to a class of densities rather than one particular density function. (Because data do not help in making inferences about τ , this model is an example of the type of analysis for which Berger^[32] argues that robust Bayesian analysis is required.)

The cost of this approach is that the information about our quantity of interest can only be narrowed to a class of posterior densities. Fortunately, in the present case, this class can be conveniently summarized as an inequality (rather than an equality) statement regarding the credible intervals: the probability that the actual quantity of interest is within the computed interval is always at least as great as the nominal coverage.*

Thus reporting results from our methods can be as simple as from existing methods used in this field. This is a considerable advantage over other applications of robust Bayesian methods because although considerable research

*One objection to our procedure is that assuming a zero prior probability for τ outside the interval $[\tau_0, \tau_1]$ may be unrealistic. Of course, one may simply enlarge the prior interval to include the nonzero density area, but a probabilistic version is easy to construct: first, elicit the interval endpoints, τ_0 and τ_1 , and also a fraction α (e.g., 0.05) such that $1 - \alpha$ fraction of the time τ falls within $[\tau_0, \tau_1]$, i.e., $\int_{\tau_0}^{\tau_1} P(\tau) d\tau = 1 - \alpha$. Then, outside this interval, use a portion of a (single) density to allocate the remaining probability to $[0, \tau_0)$ and $(\tau_1, 1]$, so that the sum of the integrals of the two add to α . In this way, every member of the class of densities on the full interval $[0, 1]$ is proper, and robust Bayesian analysis can proceed as before. We have found these changes inconsequential in the real applications we have studied, although our software offers the option to handle this situation.

in statistics has focused on these approaches, applications are still fairly rare. The reason is that the class of priors turns into a class of posteriors, and it is the entire class of posteriors that constitutes the research results. Summarizing a single multi-dimensional posterior is already fairly difficult; summarizing a class of posteriors in multi-dimensional space only creates more problems. Yet in our case, the class of posteriors can be reduced for presentation purposes to a credible interval, the only disadvantage of which is that it provides an inequality rather than equality statement.

An Empirical Example

We now reanalyze data provided in Tumbarello et al.,^[34] the largest case-control study ever conducted of the risk factors leading to bacterial pneumonia in HIV-infected patients. We focus on their univariate analysis of risk factors in 350 cases and 700 controls. The authors report prior knowledge of τ (the fraction of HIV-seropositive individuals in the general population who have an episode of bacterial pneumonia), based on previous studies, as falling in the interval [0.097, 0.29]. We interpret this to be a 99% prior interval and, for simplicity, assume the remaining $\alpha = 0.01$ mass to be uniform in [0, 0.097] and (0.29, 0.6]; our experiments (not shown) indicate that inferences change very little across many reasonable choices for the density outside the [0.097, 0.29] interval for τ .

The first row of Table 1 replicates the CI for the univariate odds ratio reported in Ref. [34] for each of four risk factors they considered. The second row of the table gives $\geq 95\%$ CIs for the risk ratio of each risk factor, using the robust Bayesian methods described in “A Robust Bayesian Interpretation” (with $m = 1000$ simulations). As the table shows, the intervals for RR indicate somewhat smaller effects than the OR, with the most noticeable effects for IV drug use and smoking.

Perhaps more interesting are the final three rows of the table, which offer information not reported in any form in the original article. For example, Table 1 shows that smoking increases the probability of bacterial pneumonia

between 0.05 and 0.27 ($\alpha \geq 95\%$ CI for the risk difference, RD). For another example, the $\geq 95\%$ CI for the base probability of an IV drug user contracting bacterial pneumonia is 0.10–0.38. These examples and the other information in Table 1 all seem like valuable information for researchers and others interested in the study and its results. The information existed in the data from this study, but they are revealed only by application of the methods offered here.

INFERENCE WITHOUT FULL POPULATION INFORMATION OR ADDITIONAL ASSUMPTIONS: DENSITY CASE-CONTROL

We now give analogous results for density case-control designs to those provided in “Inferences Without Full Population Information or Assumptions: Classic Case-Control” for classic case-control designs and provide informative bounds for π , RR, RD, $\lambda_i(t)$, and rd_i when no information or only partial information is available about the τ_j ’s in “Inference under Full Population Information” (we skip rr_i because it does not depend on τ_j and is estimable from the conditional logit procedure). Inference from density case-control samples is complicated by the fact that more than one piece of population information is involved: there is a τ_j associated with each of the M risk sets.^f We elicit the minimum and maximum values of each τ_j , which we denote $\underline{\tau}_j$ and $\bar{\tau}_j$, respectively. The interval $(\underline{\tau}_j, \bar{\tau}_j]$ can change with j or can be constant over risk sets; it can be specified to include 100% of the prior density, as in “A Proposed ‘Available Information’ Assumption,” or $1 - \alpha$ fraction of the density, as in “A Robust Bayesian Interpretation.” When we are completely ignorant over τ_j , the interval is (0, 1].

To simplify notation, let $r_k = e^{v_{kb}}$ and $r^j = \sum_{k \in R_j} r_k$. Clearly $r_k > 0$ and $r^j > 0 \forall k, j$. Then the estimators for the rate,

^fSee Ref. [35] for methods for estimating the risk using only the overall cohort disease rate. However, this estimator is less efficient than the one proposed by Langholz and Ornulf[30] and employed here. See also Ref. [6] on estimating rates using information on the crude incidence density.

Table 1 Replication and Extension of 95% CIs

Quantity of interest	Risk factor			
	IV drug use	Smoking	Pneumonia	Cirrhosis
OR	1.44–2.70	1.81–3.64	1.01–1.88	1.01–2.49
RR	1.31–2.45	1.52–3.13	1.01–1.73	1.03–2.17
RD	0.03–0.19	0.05–0.27	0.00–0.13	0.00–0.20
Pr($Y = 1 X = 0$)	0.05–0.26	0.05–0.26	0.08–0.31	0.08–0.31
Pr($Y = 1 X = 1$)	0.10–0.38	0.13–0.47	0.09–0.40	0.10–0.50

The OR row is an exact replication and the others are extensions using the methods developed here. The last two rows give the probability of contracting bacterial pneumonia given the absence and presence of the given risk factor, respectively. (From Ref. [34].)

cumulative rate, and risk in Eqs. 12–14, respectively, can be rewritten as:

$$\lambda_i(t_j) = \frac{r_i \tau_j}{r^j} \quad (23)$$

$$H(T_i, X_i) = \sum_{t_j \in T_i} \lambda_i(t_j) = \sum_{t_j \in T_i} \frac{r_i \tau_j}{r^j} \quad (24)$$

and

$$\pi_i = \Pr(Y = 1 | X_i) = 1 - e^{-H(T_i, X_i)} = 1 - \exp\left(-\sum_{t_j \in T_i} \frac{r_i \tau_j}{r^j}\right) \quad (25)$$

respectively. We now develop bounds for the quantities of interest as functions of τ_j and $\bar{\tau}_j$.

Risk

From Eq. 25, we have $\frac{\partial \pi_i}{\partial \tau_j} = \frac{e^{-H(T_i, X_i)} r_i}{r^j} > 0 \quad \forall_j$ hence the risk is a monotonically increasing function with respect to every τ_j . Denote $\underline{\pi}_i$ and $\bar{\pi}_i$ as the values of π_i with all τ_j set to τ_j and $\bar{\tau}_j$, respectively; then the bounds for π_i are simply $\pi_i \in [\underline{\pi}_i, \bar{\pi}_i]$. For example, when we are completely ignorant about τ_j and therefore $(\tau_j, \bar{\tau}_j)$ is $(0, 1]$, the bounds give $\pi_i \in (0, 1 - \exp(-\sum_{t_j \in T_i} (r_i/r^j))]$.

Risk Ratio

We now examine $RR = \pi_1/\pi_0$, where π_i is as in Eq. 25, $i = 0, 1$. We have

$$\frac{\partial RR}{\partial \tau_j} = \frac{r_1(1 - \pi_1)\pi_0 - r_0(1 - \pi_0)\pi_1}{\pi_0^2 r^j} \quad (26)$$

the sign of which is determined by that of the numerator. When the numerator is positive, i.e., when

$$rr = r_1/r_0 > \frac{\pi_1(1 - \pi_0)}{(1 - \pi_1)\pi_0} = OR \quad (27)$$

the partial derivative is positive and so RR increases with respect to τ_j . Otherwise, the derivative is negative and RR decreases with respect to τ_j . In [Appendix B](#), we show that Eq. 27 holds whenever $rr < 1$, independent of the values of τ_j . Similarly, the sign is reversed whenever $rr > 1$. Hence RR is either monotonically increasing or monotonically decreasing with respect to τ_j for all j .

Thus when $rr < 1$, rr is monotonically increasing with respect to all τ_j and is therefore bounded by $\underline{RR}$ and $\bar{RR}$, which are values of RR with all τ_j set to their minimum and maximum values, respectively. Otherwise, RR is monotonically decreasing and the bounds are $\bar{RR}$ and $\underline{RR}$. In short, RR is bounded by

$$RR \in [\min(\underline{RR}, \bar{RR}), \max(\underline{RR}, \bar{RR})] \quad (28)$$

It is easy to see that $\lim_{\tau_j \rightarrow 0} RR = r_1/r_0$, hence when no information is available for τ_j and therefore $(\tau_j, \bar{\tau}_j)$ is $(0, 1]$, the bounds become $RR \in (\min(r_1/r_0, \bar{RR}), \max(r_1/r_0, \bar{RR})]$ where $\bar{RR}$ is RR evaluated at $\tau_j = 1 \quad \forall_j$.

Risk Difference

The case of $RD = \pi_1 - \pi_0$ is more complicated because RD is not a monotonic function of the τ_j 's and $\partial RD/\partial \tau_j = [r_1 e^{-H(T_1, X_1)} - r_0 e^{-H(T_0, X_0)}]/r^j$ can change signs depending on the values of τ_j . Under the proportional hazards model where r_i and r^j are not functions of time, however, we can reduce the analytically difficult or even intractable problem of constrained optimization in multi-dimensional space to a simple one in which RD is a one-dimensional function of the cumulative baseline hazard, which is a monotone function of the τ_j 's.

Let $Q(\tau_j) = \sum_{t_j} M \tau_j / r^j$ denote the cumulative baseline hazard rate, and note that $\partial Q(\tau_j)/\partial \tau_j = 1/r^j > 0$ for all j , so $Q(\tau_j)$ is monotonically increasing in all τ_j 's and therefore bounded between $\underline{Q} = Q(\tau_j)$ and $\bar{Q} = Q(\bar{\tau}_j)$. Now rewrite RD in terms of Q . From Eq. 24, we have $H(T_k, X_k) = r_k Q$ for $k = 0, 1$; hence

$$RD = (1 - e^{-r_1 Q}) - (1 - e^{-r_0 Q}) = e^{-r_0 Q} - e^{-r_1 Q} \quad (29)$$

and

$$\frac{\partial RD}{\partial \tau_j} = \frac{r_1 e^{-r_1 Q} - r_0 e^{-r_0 Q}}{r^j} \quad (30)$$

RD is not monotone in Q , but Q is a scalar and we know its bounds, which brings us to a situation mathematically similar to analyzing RD in classic case-control designs.

Let Q^* be the solution to the first-order condition $\partial RD/\partial \tau_j = 0$. From Eq. 30, we can solve for $Q^* = [1/(r_1 - r_0)] \ln(r_1/r_0)$. Then from Eq. 29, we have

$$RD(Q^*) = (r_0/r_1)^{r_0/(r_1-r_0)} - (r_0/r_1)^{r_1/(r_1-r_0)} \quad (31)$$

To obtain the bounds for RD, we first see whether $\underline{Q}, \bar{Q}$ contains Q^* . If it does, then

$$RD \in [\min(\underline{RD}, \bar{RD}, RD^*), \max(\underline{RD}, \bar{RD}, RD^*)] \quad (32)$$

where $\underline{RD} = RD(\underline{Q})$, $\bar{RD} = RD(\bar{Q})$, and $RD^* = RD(Q^*)$. Otherwise, the bounds are

$$RD \in [\min(\underline{RD}, \bar{RD}), \max(\underline{RD}, \bar{RD})] \quad (33)$$

When no information is available for τ_j , the bounds become

$$RD \in [\min(0, RD^*), \max(0, RD^*)] \quad (34)$$

Rate

From Eq. 23, we see that $\partial\lambda_i(t_j)/\partial\tau_j = r_i/r^j > 0$; hence $\lambda_i(t_j)$ is monotonically increasing in τ_j . It is therefore bounded in $(\lambda_i(t_j), \overline{\lambda_i(t_j)})$, where $\lambda_i(t_j)$ is $\lambda_i(t_j)$ evaluated at $\underline{\tau}_j$ and $\overline{\lambda_i(t_j)}$ is $\lambda_i(t_j)$ evaluated at $\overline{\tau}_j$. When we are ignorant with respect to the τ_j 's, the rate is bounded as $(0, r_i/r^j)$.

Rate Difference

Because $\partial rd/\partial\tau_j = (r_1 - r_0)/r^j$, rd is monotonically increasing in τ_j if $r_1 > r_0$ and decreasing otherwise. Hence the bound on the rate difference is

$$rd \in \left[\min \left[rd(\underline{\tau}_j), rd(\overline{\tau}_j) \right], \max \left[rd(\underline{\tau}_j), rd(\overline{\tau}_j) \right] \right] \quad (35)$$

and when ignorant of all information on τ_j , the bounds are

$$rd \in \left[\min \left[0, (r_1 - r_0)/r^j \right], \max \left[0, (r_1 - r_0)/r^j \right] \right]$$

CONCLUSION

As is increasingly recognized, the quantity of interest in most case-control studies is not the odds ratio, but rather some version or function of a probability, risk ratio, risk difference, rate, rate ratio, or rate difference, depending on context.^[6,7,36–41] We provide the methods to estimate each of these quantities from case-control studies, even if no auxiliary information, or only limited auxiliary information, is available.

Unless the odds ratio happens to approximate a parameter of central substantive interest, which is quite rare, we suggest that it should not be reported any more frequently than any other intermediate quantity in statistical calculations. We suggest instead that researchers justify their assumption regarding bounds on τ (in classic case-control studies) or τ_j (in risk set case-control studies) in the data or methods section of their work. Then, they can substitute the confidence interval (CI) now reported for the odds ratio with the CI for their chosen quantity (or quantities) of interest. For example, instead of an uninformative but presently common reporting style:

the effect of smoking on lung cancer is positive OR = 1:38
(95% CI 1.30–1.46)

researchers could give the much more interesting:

smoking increases the risk of contracting lung cancer by a factor of between 2.5 and 3.1 ($\alpha; \geq 95\%$ CI)

or

smoking increases the probability of contracting lung cancer between 0.022 and 0.051 ($\alpha; \geq 95\%$ CI)

If uncertainty exists over the appropriate bounds for the unknown quantities, we suggest using the widest bounds,

conducting sensitivity analyses by showing how the CI depends on different assumptions or setting α to a value other than zero.

The methods discussed here are meant to improve presentation and increase the amount of information that can be extracted from existing models and data collections. They do not enable scholars to ignore the usual threats to inference (measurement error, selection bias, confounding, etc.) that must be avoided in any study.

ACKNOWLEDGMENTS

We thank Sander Greenland for his generosity, insight, and wisdom about the epidemiological literature, Norm Breslow and Ken Rothman for many helpful explanations, and Chuck Manski for his suggestions, provocative econometric work, and other discussions. Thanks also to Neal Beck, Rebecca Betensky, Josué Guzmán, Bryan Langholz, Meghan Murray, Adrain Rafferty, Ted Thompson, Jon Wakefield, Clarice Weinberg, and David Williamson for helpful discussions, Ethan Katz for research assistance, and Mike Tomz for spotting an error in an earlier version. Thanks to the National Science Foundation (IIS-9874747), the Centers for Disease Control and Prevention (Division of Diabetes Translation), the National Institutes of Aging (P01 AG17625-01), the World Health Organization, and the Center for Basic Research in the Social Sciences for research support. Software to implement the methods in this paper is available for R, Stata, and Gauss from <http://GKing.Harvard.Edu>.

APPENDIX A: SIMPLIFYING MANSKI'S BOUNDS ON THE RISK DIFFERENCE

Proving Eq. 18 requires algebra only. For simplicity, let $P_{ab} = \Pr(X_a|Y = b)$, so that OR = $(P_{11}P_{00})/(P_{01}P_{10})$. Then, omitting tedious but straightforward algebra at several stages, $\phi = (ORP_{01}^2/P_{11}^2)^{1/2} = \sqrt{OR} P_{01}/P_{11}$, and $\gamma = \sqrt{OR}/(\sqrt{OR} + P_{11}/P_{10})$. Then the components of RD γ are

$$P_{11}\gamma = \frac{\sqrt{OR}P_{10}P_{11}}{\sqrt{OR}P_{10} + P_{11}}, \quad P_{01}\gamma = \frac{\sqrt{OR}P_{01}P_{10}}{\sqrt{OR}P_{10} + P_{11}}$$

$$P_{10}(1-\gamma) = \frac{P_{10}P_{11}}{\sqrt{OR}P_{10} + P_{11}}, \quad P_{00}(1-\gamma) = \frac{P_{11}P_{00}}{\sqrt{OR}P_{10} + P_{11}}$$

and so putting the terms together yields

$$RD_\gamma = \sqrt{OR}/(1 + \sqrt{OR}) - 1/(1 + \sqrt{OR})$$

$$= (\sqrt{OR} - 1)/(\sqrt{OR} + 1).$$

APPENDIX B: MONOTONICITY OF RISK RATIO UNDER DENSITY CASE-CONTROL DESIGNS

We show here that if $rr < 1$, then $rr > [\pi_1(1 - \pi_0)]/[(1 - \pi_1)\pi_0]$ (the $r_1/r_0 > 1$ case is similar). Let $H_k = H(T_k, X_k)$, $k = 0, 1$. From the definition of π_1 and π_0 , $[\pi_1(1 - \pi_0)]/[(1 - \pi_1)\pi_0]$ can be simplified to $(e^{H_1} - 1)/(e^{H_0} - 1)$. Because $rr = H_1/H_0$, we only need to show that if $H_1/H_0 < 1$, then $H_1/H_0 > (e^{H_1} - 1)/(e^{H_0} - 1)$ or, equivalently,

$$H_1(e^{H_0} - 1) > H_0(e^{H_1} - 1).$$

The e^{H_1} Taylor series expansions of $e^{H_1} - 1$ and $e^{H_0} - 1$ at 0 give

$$e^{H_1} - 1 = H_1 + (1/2)H_1^2 + (1/3!)H_1^3 + \dots \quad (36)$$

$$e^{H_0} - 1 = H_0 + (1/2)H_0^2 + (1/3!)H_0^3 + \dots \quad (37)$$

and hence

$$H_1(e^{H_0} - 1) = H_1H_0 + (1/2)H_0^2H_1 + (1/3!)H_0^3H_1 + \dots \quad (38)$$

$$H_0(e^{H_1} - 1) = H_0H_1 + (1/2)H_1^2H_0 + (1/3!)H_1^3H_0 + \dots \quad (39)$$

The first terms in Eqs. 38 and 39 are equal, and when $H_1/H_0 < 1$, hence $H_1 < H_0$, all other terms in Eq. 38 are greater than the corresponding terms in Eq. 39 (because both $H_0 > 0$ and $H_1 > 0$ always). Thus when $H_1/H_0 < 1$, $H_1(e^{H_0} - 1) > H_0(e^{H_1} - 1)$.

REFERENCES

- Moynihan, R.; Bero, L.; Ross-Degnan, D.; Henry, D.; Lee, K.; Watkins, J.; Mah, C.; Soumerai, S.B. Coverage by the news media of the benefits and risks of medications. *N. Engl. J. Med.* **2000**, *342* (22), 1645–1650.
- Manski, C.F. *Identification Problems in the Social Sciences*; Harvard University Press: Cambridge, MA, 1995; 31.
- Manski, C.F. *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya. Nonparametric Identification under Response-Based Sampling*; Cambridge University Press: Cambridge, 1999.
- Greenland, S. On the need for the rare disease assumption in case-control studies. *Am. J. Epidemiol.* **1982**, *116* (3), 547–553.
- Deeks, J.; Sackett, D.; Altman, D. Down with odds ratios. *Evid. Based Med.* **1996**, *1* (6), 164–166.
- Greenland, S. Interpretation and choice of effect measures in epidemiologic analysis. *Am. J. Epidemiol.* **1987**, *125* (5), 761–768.
- Davies, H. T. O.; Manouche, T.; Iain, C.K. When can odds ratio mislead. *Br. Med. J.* **1998**, *31*, 989–991.
- Rothman, K.J.; Greenland, S. *Modern Epidemiology*, 2nd; Ed.; Lippincott-Raven: Philadelphia, PA, 1998; 113–245.
- Prentice, R.L.; Breslow, N.E. Retrospective studies and failure-time models. *Biometrika* **1978**, *65*, 153–155.
- Chamberlain, G. Analysis of covariance with qualitative data. *Rev. Econ. Stud.* **1980**, *XLVII*, 225–238.
- Greenland, S. Modeling risk ratios from matched cohort data: An estimating equation approach. *Appl. Stat.* **1994**, *43* (1), 223–232.
- Goldstein, L.; Langholz, B. Asymptotic theory for nested case-control sampling in the cox regression model. *Ann. Stat.* **1992**, *20* (4), 1903–1928.
- Borgan, Ø.; Langholz, B.; Goldstein, L. Methods for the analysis of sampled cohort data in the cox proportional hazard model. *Ann. Stat.* **1995**, *23*, 1749–1778.
- Langholz, B.; Goldstein, L. Risk set sampling in epidemiologic cohort studies. *Stat. Sci.* **1996**, *11* (1), 35–53.
- Langholz, B.; Thomas, D.C. Efficiency of cohort sampling designs: Some surprising results. *Biometrics* **1991**, *47*, 1563–1571.
- Lubin, J.H.; Gail, M.H. Sampling strategies in nested case-control studies. *Environ. Health Perspect.* **1994**, *102* (Suppl. 8), 47–51.
- Robins, J.M.; Gail, M.H.; Lubin, J.H. More on biased selection of controls for case-control analyses of cohort studies. *Biometrics* **1986**, *42*, 293–299.
- Prentice, R.L. A case-cohort design for epidemiological studies and disease prevention trials. *Biometrika* **1986**, *73*, 1–11.
- Greenland, S. Multivariate estimation of exposure-specific incidence from case-control studies. *J. Chronic. Dis.* **1981**, *34*, 445–453.
- Neutra, R.R.; Drolette, M.E. Estimating exposure-specific disease rates from case-control studies using Bayes theorem. *Am. J. Epidemiol.* **1978**, *108* (3), 214–222.
- Cornfield, J. A method of estimating comparative rates from clinical data: Application to cancer of the lung, breast and cervix. *J. Natl. Cancer Inst.* **1951**, *11*, 1269–1275.
- Anderson, J.A. Separate-sample logistic discrimination. *Biometrika* **1972**, *59*, 19–35.
- Prentice, R.L.; Pyke, R. Logistic disease incidence models and case-control studies. *Biometrika* **1979**, *63*, 403–411.
- Mantel, N. Synthetic retrospective studies and related topics. *Biometrics* **1973**, *29*, 479–486.
- Manski, C.F. The estimation of choice probabilities from choice based samples. *Econometrics* **1977**, *45* (8), 1977–1988.
- King, G.; Zeng, L. Logistic regression in rare events data. *Polit. Anal.* **2001**, *9* (2), 137–163.
- King, G.; Tomz, M.; Wittenberg, J. Making the most of statistical analyses: Improving interpretation and presentation. *Am. J. Polit. Sci.* **2000**, *44* (2), 341–355.
- Bishop, C.M. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, 1995.
- Beck, N.; King, G.; Zeng, L. Improving quantitative studies of international conflict: A conjecture. *Am. Polit. Sci. Rev.* **1999**, *94* (1), 21–36.
- Langholz, B.; Ørnulf, B. Estimation of absolute risk from nested case-control data. *Biometrics* **1997**, *53*, 767–774.
- Yule, G.U. On the methods of measuring the association between two attributes. *J. R. Stat. Soc.* **1912**, *75*, 579–642.
- Berger, J. An overview of robust Bayesian analysis (with discussion). *Test* **1994**, *3*, 5–124.
- Insua, D.R.; Ruggeri, F. *Bayesian Analysis*; Springer Verlag: New York, 2000.

34. Tumbarello, M.; Tacconelli, E.; de Gaetano, K.; Ardit, F.; Pirroni, T.; Claudia, R.; Ortona, L. Bacterial pneumonia in HIV-infected patients: Analysis of risk factors and prognostic indicators. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* **1998**, *18*, 39–45.
35. Benichou, J.; Gail, M. Methods of inference for estimates of absolute risk derived from population-based case-control studies. *Biometrics* **1995**, *51*, 182–194.
36. Nurminen, M. To use or not to use the odds ratio in epidemiologic analysis. *Eur. J. Epidemiol.* **1995**, *11*, 365–371.
37. Zhang, J.; Yu, F. Kai What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *N. Engl. J. Med.* **1998**, *280* (19), 1690–1691.
38. Davies, H.T. O.; Manouche, T.; Iain, C. Kinloch Authors reply. *Br. Med. J.* **1998**, *317*, 1156–1157.
39. Deeks, J. Odds ratio should be used only in case-control studies and logistic regression analyses. *Br. Med. J.* **1998**, *317*, 1155–1156.
40. Michael, B.; Bracken, J.C. Avoidable systematic error in estimating treatment effects must not be tolerated. *Br. Med. J.* **1998**, *317*, 11–56.
41. Altman, D.G.; Deeks, J.J.; Sackett, D.L. Odds ratios should be avoided when events are common. *Br. Med. J.* **1998**, *317*, 1318.

Center Weighting: Multicenter Trials

Paul Gallo

Biostatistics, Novartis Pharmaceuticals Corporation, East Hanover, New Jersey, U.S.A.

Abstract

The primary motivation for using multiple centers is to enable enrollment of a sufficient number of clinical trial patients to investigate a hypothesis in a timely manner. This entry reviews the advantages of center weighting in multicenter trials, looks at the construction of composite centers, and discusses the regulatory guidelines of such trials.

Keywords: Fixed effect model; Multicenter trial; Random effect model; Treatment-by-center interaction.

INTRODUCTION

Clinical trials comparing treatment therapies are commonly conducted using patients from a number of different sites. The primary motivation for using multiple centers is to enable enrollment of a sufficient number of patients to investigate a hypothesis in a timely manner. A beneficial consequence is that this can provide evidence that the trial results are not strongly setting dependent; hence, it may be reasonable to interpret that these results apply to a broader population of clinical sites or patient populations. Some statistical analysis methods as applied to data from multicenter clinical trials can be formulated as combinations of treatment effect estimates or test statistics determined within the individual centers. There can thus be some ambiguity in how these within-center estimates should be combined; this issue may be described in terms of weights associated with the individual centers. For example, inference may be based on an overall treatment effect estimate that simply averages the within-center estimates, so that we might view the resultant analysis as using equal center weighting. Alternatively, since differing within-center sample sizes may lead to different precisions of the center-specific estimates, we might combine the within-center estimates according to some measure of their precision (i.e., more precise estimates—generally those from larger centers—receive more weight). Other formulations can be considered as well. Later sections describe various issues relating to center weighting in analysis of clinical trial data, and consider statistical, logistic, and regulatory aspects and implications.

STATISTICAL FRAMEWORK

A feature of multicenter trials with important implications for statistical analysis is that it is a natural consequence

of the manner in which such trials are conducted that the within-center sample sizes can be quite different from each other. Trials in which the largest center contributes 4 or 5 times as many patients as the smallest are quite common, and larger imbalances are often encountered. Attempting to enforce similarity of center sample sizes can substantially lengthen trial duration and increase potential for subtle selection biases as well.^[1]

Typically, statistical analysis models for clinical trial data take into account the site at which a patient was enrolled, i.e., analysis models are used that adjust expected response for center. Depending on the data structure and analysis method used, center might be included as a covariate, or perhaps might define strata in a stratified analysis. One main reason for such adjustment is that for logistical reasons, it is common for randomization schemes to be stratified within center, and it is considered good statistical practice for analysis models to reflect restrictions on randomization resulting from stratification.^[2] Additionally, the center at which a patient enrolled may well be predictive of response because of differences in investigator skills, local medical practices, patient populations, etc. (though when such effects are really due to other identifiable and measurable covariates, it will often be advisable to consider adjusting for those also—accounting for center in analysis models should not be viewed as a justification for not having adjusted for other variables expected to be meaningful). At another level, the consistency of treatment differences across centers and other key subgroups will generally be of value in fully interpreting trial results; hence treatment-by-center interaction will usually be of some interest to investigate. The rationale from a regulatory perspective for interest in center consistency and its relevance to the interpretation of the trial results as a whole is well described in Ref. [3].

Questions regarding how analysis models should account for centers are closely related to issues of “weights” that centers receive in the analysis. A partial list of articles on multicenter trial analysis with some relationship to this issue includes.^[3–16] Much of this literature discussion has played out in the specific case of analyses that utilize fixed-effect analysis of variance (ANOVA) models for normally distributed data, in which the principles and mechanics are most easily illustrated. This issue is closely related to the definition of different “types” of ANOVA sums of squares produced by the procedure PROC GLM in the SAS statistical software system,^[17] and also to the question of whether or not a term for treatment-by-center interaction is included in the statistical model. This framework is our main focus here, though these issues certainly have relevance for other data types (e.g., binary or categorical outcomes, time-to-event data), and later we discuss this issue a bit further. Within the ANOVA framework, we also discuss later random-effect models, in which center main effects or interactions with treatment are considered to be random variables; these present an intriguing alternative to fixed effect models and are very relevant to the general issue of center weighting.

For ease of illustration, we restrict consideration to a comparison between two treatment groups in a number of centers denoted by c . For further simplicity, assume that there are no covariates predictive of outcome under consideration other than treatment and center; the basic ideas extend readily to more complex formulations. We let R_{ijk} denote a response variable for a particular patient k randomized to treatment group i at center j , $i = 1, 2; j = 1, \dots, c$. We let n_{ij} denote the number of patients in center j assigned to treatment group i ; also, $n_j = n_{1j} + n_{2j}$, $n = \sum n_j$. Let D_j represent the true mean difference in the response variable between treatments 1 and 2 in center j , with d_j denoting an estimate of this quantity; in our simple situation, d_j is the difference of the mean responses in the two treatment groups for patients in center j .

Sensible measures of the overall treatment difference might be viewed as linear combinations of estimates computed from the individual centers; hence, we consider the following family of weighted averages of the within-center estimates:

$$d_w = \sum w_j d_j$$

with

$$\sum w_j = 1.$$

Two specific choices of weights would seem to be most immediately obvious, and historically have been most utilized and the focus of most of the literature discussion on this topic. In the first of these, $w_j \equiv 1/c$, so that the overall treatment difference estimate is the numerical average of the within-center estimates:

$$d_U = \sum d_j / c$$

In the second approach, weights for each center are chosen according to the precision of the within-center estimate, i.e., each w_j is the inverse of the variance of that estimate under an assumption of common variance for the responses, scaled to sum to 1:

$$w_j = (n_{1j}^{-1} + n_{2j}^{-1})^{-1} / \sum (n_{1j}^{-1} + n_{2j}^{-1})^{-1} \quad (1)$$

We refer to d_U as the *unweighted* estimator and d_w [with weights defined as in Eq. (1)] as the *weighted* estimator, and inferences based upon them as unweighted and weighted inferences, respectively.

It should be noted that if within each center the two treatment arms have the same number of patients (a type of balance nearly achieved in many clinical trials because of the randomization stratification mentioned previously), then the weighted estimator simplifies to the difference of the response averages computed over all patients within each treatment arm. In this sense, our choice of terminology might be considered a misnomer from a certain perspective: weighted inference essentially gives equal weight to responses from individual *patients*, while in unweighted inference, patient responses are in effect weighted relative to the sizes of their centers (a patient from a small center having more weight). If in addition the center sizes are the same, then the two parameters we have defined are identical. (Also, the forms of the estimators illustrated here of course lose these simple representations as soon as we deviate from the simple structure we have defined, e.g., if we adjust response for additional covariates, but the key concepts remain largely similar.)

Given the common usage within the pharmaceutical industry of the SAS statistical package,^[17] this issue as it applies to multicenter clinical trials is quite well known to relate to different analysis options in the SAS procedure PROC GLM. It can be noted that an unweighted analysis is obtained using type III sums of squares in PROC GLM when specifying an analysis model—a *computational model* in the sense described by Permutt^[13]—which contains treatment, center, and treatment-by-center interaction, say:

$$y_{ijk} = \mu + T_i + C_j + I_{ij} + e_{ijk} \quad (2)$$

Thus, a t -test constructed from d_U and its estimated standard error provides identical inference to the F -test obtained from the SAS type III ANOVA. A weighted analysis, i.e., one based upon d_w , can be obtained in the same computational model using SAS type II sums of squares. Additionally, inference from a computational ANOVA model that does not contain the interaction term is also based on d_w , albeit with a differently defined error term.

Other approaches to determining weights have been utilized or considered as well. In an approach that is sort of a hybrid of the two most well-known weighting types, it may be specified that a SAS type III analysis will be initially performed, but if the interaction term fails to reach a specified significance threshold then it will be dropped from the model. Thus, this approach can lead to an unweighted or weighted analysis, depending on the results of this interaction pretest (this approach seems to have been considered more in the past, but is rarely used currently). As a further example of a different type of weighting scheme, weights for centers corresponding to expected representation of patients from those centers in a broader population of interest are discussed in Ref. [12].

COMPARISONS OF WEIGHTING SCHEMES

Certainly, we are not restricted to considering only the results from a single analysis of trial data, nor to selecting a single center-weighting scheme. Examining consistency or differences of results across different analyses can be very important in properly interpreting trial results, as we discuss later. Nevertheless, it is conventional in the pharmaceutical industry to prespecify the details of a *primary analysis*, based upon prestudy expectations and assumptions concerning the most appropriate manner of investigating the main trial hypothesis. Though the data themselves may provide valuable clues concerning how they might most appropriately be analyzed, and a number of analyses will generally be performed as part of a thorough examination of the data, results of the primary analysis generally receive most emphasis in initial consideration and interpretation of trial results. Arguments have been put forth in support of the two weighting schemes described earlier for use in primary analyses; these are primarily based upon comparisons of the parameters on which the inferences are based, and upon the precision of the estimates of those parameters.

With σ^2 denoting the common variance of the responses, the variances of the unweighted and weighted estimators are

$$V_U = \sigma^2 \sum \sum n_{ij}^{-1} / c^2$$

$$V_W = \sigma^2 / \left(\sum \left(n_{1j}^{-1} + n_{2j}^{-1} \right)^{-1} \right),$$

respectively.

V_U is always at least as large as V_W , with equality only when all treatment-center combination sample sizes are identical. Illustrations for various levels of center-size imbalance are described in Ref. [9]; it seems fair to say that for magnitudes of imbalance often encountered in clinical trials, the difference between these variances can be substantial. To cite one specific example for illustration,

if most centers in a study have a common sample size, but a quarter of the centers have one-third as many patients as the others, V_U will be 25% larger than V_W .

Variance is of course not the sole basis on which the estimators can be compared, since they generally estimate different quantities (say, D_U and D_W , defined analogously to d_U and d_W), and inferences could thus be viewed as testing different hypotheses. An argument in support of equal center weighting is that in the presence of true treatment-by-center interaction—i.e., an interaction term appears in the *probability model*, again using the terminology of Permutt;^[13] equivalently, the D_j are not all equal—then unweighted inference tests a hypothesis that a priori is clearly defined, namely whether or not the average treatment difference across all centers is zero (or in a slightly broader sense, this quantity at least becomes clearly defined at whatever point we identify the centers that will participate in the study). On the other hand, weighted inference is based on a parameter that is not clearly defined in this sense, i.e., it is not known before the study since it depends on the within-center sample sizes that will be obtained:

$$D_W = \sum \left(n_{1j}^{-1} + n_{2j}^{-1} \right)^{-1} D_j / \sum \left(n_{1j}^{-1} + n_{2j}^{-1} \right)^{-1}$$

In this view, the deviation of this parameter from its unweighted counterpart is considered to be bias.^[11,12] Others have contended that this is not bias and that a weighted approach estimates a validly interpretable quantity.^[5,8,13,14] Senn^[1] has argued that in a practical sense, the parameters on which the two approaches are based will often be quite similar and a true treatment effect will tend to be reflected in either; hence inference based on the more precise of these (weighted) should generally be preferable. Gallo^[9] described how even if interaction were present but center effects were not systematically related to center size, the two approaches would still be based on essentially the same parameter, so that precision would be a sensible basis for decision (and this would of course support weighted inference). Ganju and Mehrotra^[14] discuss that the center sample sizes, and hence the weights in a weighted analysis, are random variables, and that conventional analyses fail to take this into account.

REGULATORY GUIDANCE

Some current regulatory guidance documents have reflected and influenced the evolution of opinions regarding primary analysis models and center-weighting issues. These recommend that in primary analyses, centers should usually be size weighted. Specifically, this is achieved by using analysis models that do not include terms for treatment-by-center interaction; the existence of such interaction is felt to be exploratory in most situations, and the efficiency of the weighted approach would often be

expected to provide the better initial evidence regarding treatment effectiveness. It is acknowledged that reasonable consistency of results across centers and consistency of estimated overall effects based upon different weighting schemes are highly desirable; hence, this should be investigated further in the presence of a positive result from the primary analysis.

Some specific relevant statements from the ICH-E9 document, *Guidance on Statistical Principles for Clinical Trials*,^[18] include:

- “Trials that avoid excessive variation in the numbers of subjects per center and trials that avoid a few very small centers have advantages if it is later found necessary to take into account the heterogeneity of the treatment effect from center to center, because they reduce the differences between different weighted estimates of the treatment effect.”
- “The main treatment effect may be investigated first using a model that allows for center differences, but does not include a term for treatment-by-center interaction. If the treatment effect is homogeneous across centers, the routine inclusion of interaction terms in the model reduces the efficiency of the test for the main effects.”
- “If positive treatment effects are found in a trial with an appreciable number of patients per center, there should generally be an exploration of the heterogeneity of treatment effects across centers, as this may affect the generalizability of the conclusions.”

The CPMP document *Points to Consider on Adjustment for Baseline Covariates*^[2] states:

- “When multicenter trials are stratified by centre, then centre should be adjusted for in the primary analysis regardless of its prognostic value.”
- “The primary model should include only the covariates prespecified in the protocol and no interaction terms. However, treatment by covariate interactions should be explored, as recommended in the ICH-E9 guideline.”

A useful summary of U.S. regulatory perspectives, consistent with the regulatory guidances quoted above, is provided by Anello et al.^[3]

SAMPLE SIZE CONSIDERATIONS

When linear model analyses are used in clinical trials with desired equal allocation to two treatment groups, the total sample size n is conventionally determined using the following formula or some variant:

$$n = 4 \left(z_{1-\alpha/2} + z_{1-\beta} \right)^2 / \Delta^2 \quad (3)$$

where z_p is the p th percentile of a standard normal distribution, and Δ is a standardized treatment difference desired to be detected with power $1 - \beta$ using a one-sided level $\alpha/2$ test (refinements to this formula might be made, for example, to use t -distributions or incorporate other covariates; here, we illustrate only the simple large-sample normal approximation). This is essentially the approach utilized in many commercial sample size software packages.

It has occasionally been stated in the literature that this calculation “assumes that there is no interaction.” This is not really accurate: the validity of the formula is not fully determined by assumed underlying probability models, but rather by the analysis method that will be used and the mechanics of the computations involved. Specifically, this sample size formula depends upon the assumption that the variance of the overall treatment difference estimator is

$$V = 4\sigma^2/n.$$

This in fact is the minimum possible variance based upon n patients. It is the variance of the weighted estimator when there is within-center balance and of the unweighted estimator only if additionally the center sizes are identical (in which case of course the estimators coincide). Thus, it would be more accurate to state that this formula assumes that:

- weighted analysis is being used, or
- center sizes are equal.

If an unweighted analysis is being used, then provided that center sizes are not the same, this sample size calculation will understate the number of patients needed to achieve the desired power. This does not depend on whatever under-lying model is assumed, but rather on the fact that the computational model allows the possibility of interaction. It seems to have been the case historically that studies in which there was no expectation that center sizes would be similar, and in which unweighted primary analyses were specified, were often designed without taking into account the inefficiency that would be introduced by those unequal center sizes, and were thus underpowered. To correct for this, it would be necessary to adjust the sample size calculation according to some measure of center-size imbalance, but of course this is almost always not known in advance. Approaches to approximate the necessary adjustment based on expected levels of center imbalance are described in Refs. [9,19]. This is directly related to the relationship between the unweighted and weighted variances described in a previous section. For example, in the same simple illustrative case we considered earlier, if most centers would be expected to have a common sample size but a quarter of them might be one-third as large, then the result from the conventional formula should be increased by 25% in order to achieve the desired power in an unweighted analysis.

For weighted analyses, this issue is very much minimized: the actual variance and that assumed in Eq. (3) are identical when there is within-center balance, and slight deviation from such balance will generally have little impact. (One caveat involves the possibility of many very small centers. For example, in a center with 20 patients, a treatment differential of 11 vs. 9 will be trivially less efficient than the assumed equal allocation; while in a center with four patients, a 3 vs. 1 treatment split would introduce more relative inefficiency compared to equal allocation. Small within-center randomization block sizes will tend to minimize this concern. Nevertheless, in a trial expected to have many very small centers, it might be considered whether or not the conventional sample size calculation needs to be modified.)

CONSTRUCTION OF COMPOSITE CENTERS

When some centers are very small, a practice of defining artificial “pooled” or “composite” centers is sometimes employed; that is, data from different centers are treated in the analysis as if they came from the same center. A number of small centers may be combined, or one or more small centers may be combined with a larger center. Composites may be constructed to the extent of eliminating empty cells to ensure that treatment effects are estimable in models that contain interaction terms, or so that all observations contribute to treatment effect estimates in models that include center as a main effect. More commonly, this is done to achieve some minimum cell size felt to appropriately limit the influence of individual observations or remove some instability from the analysis. Algorithms for determining the composites are generally specified before unblinding of the trial for final analysis.

In describing the practice of using composite centers, we distinguish this from consideration of covariates that reflect “broader” categories of grouped centers that might be expected to affect response. For example, geographic region might explain allergic or asthmatic responses, centers in less developed areas might enroll patients with a poorer prognosis, experienced staff might have better success using treatment regimens involving complex medical procedures, etc. Accounting for such effects in analysis models can be part of a sensible approach to model determination. The discussion below refers basically to numerical compositing algorithms based upon within-center sample sizes.

The practice of using composite centers seems to have evolved in order to minimize the large variance inflation and data instability that can arise in unweighted analyses when some centers are very small. We have previously described some simple illustrations in which some centers are one-third as large as others; in reality, in many clinical trials the relative imbalance between the largest and smallest centers can be far greater. There may be several

large centers, and then some centers contributing only a handful of patients; allowing estimates from the small centers to have the same weight as the large ones can cause the treatment effect estimates and inferences to be highly sensitive to even moderately unusual response values in the small centers. This instability is reflected in imprecision of the overall estimates. As described by Senn^[1,10] and others, defining composites for unweighted analyses shows concessions toward a weighted approach, that is, this approach moves in a “weighted direction” to decrease the level of imbalance among the (pooled) centers and to lessen the potential influence of very small centers. When using an unweighted analysis with composite centers, both the treatment estimate and its variance generally lie between those of weighted and unweighted analyses based on the original center definitions; see Ref. [11]

There are potential implications of this practice that should be understood. One immediate concern is that, to the extent that any center main effects exist, defining composites will tend to produce an overestimate of the error variance. This effect can be explored in a sensible manner by comparing the error mean square from an analysis using the original center definitions to that from one in which the composite centers are used. Questions of bias introduced by this practice are more subtle and are difficult to summarize briefly, though there would seem to be some potential for bias. This practice certainly changes the weights that centers receive in unweighted analyses: inclusion in a composite decreases the weight given to a center and its patients. If an unweighted approach were advocated on the philosophical grounds that centers must be treated equally and the parameter on which inference is based must be clearly defined, it should be realized that this can no longer be claimed when composites are utilized. In the presence of interaction, the parameter that is the basis for inference depends on center sizes in the same manner as weighted analyses, albeit to a lesser extent (unweighted analysis using composites has been described as weighted *within* composites, but unweighted *across* them^[11,12]). Unlike weighted analyses, it may not even be possible to claim that centers of similar sizes are weighted comparably in unweighted analyses using composites; for example, one center may be right at the maximum size limit for inclusion in a composite and thus substantially downweighted, while a center with a single additional patient might stand alone and receive full weight.

It seems to be the case that composites are sometimes used for primary weighted analyses as well (i.e., models without the interaction term). There is likely some residual usage of this practice left over from the days when unweighted primary analyses were more common, though there is far less motivation for it. The issues are much different for weighted analyses, in which composites are generally not needed because the influence of small centers is directly limited by their sizes. The CPMP document^[2] discourages use of the practice: “... pooling small centers to

form one center of size comparable to that of other centers has little or no scientific justification.”

A special case that might be considered further involves “empty cells” resulting from centers in which there is a treatment to which no patients have been assigned. In the ANOVA model we used, such centers will not contribute to the treatment difference estimate (they may affect other aspects of the analysis, e.g., the error term). Perhaps they should not contribute, if we really believe center to be strongly predictive in a main-effect sense. Nevertheless, there is understandable motivation to have all primary endpoint data contribute to the treatment effect estimate; pooling can be one way to do this, and a small amount of pooling should generally not be expected to impair weighted analyses. If many centers contributed empty cells, then the analysis model likely could not accommodate main-effect adjustment for center, and the interpretation of the analysis would depend on the reasonableness of this assumption.

CENTER WEIGHTING IN RANDOM-EFFECT MODELS

As mentioned previously, random-effect models, or mixed models, present an intriguing alternative to fixed-effect models for multicenter trial analyses. In these models, treatment-by-center interaction effects, and perhaps center effects as well, are considered to be random variables. Indeed, there are aspects of this situation that often seem much more logically consistent with a random-effect approach. For example, trial results are normally viewed as having relevance in a broader population of settings than those actually used in the trial (though perhaps not with effects identical to those in the centers selected for clinical trials). Also, inconsistency of effect across centers may tend to make one less certain that a treatment effect has been characterized accurately, since it would then be reasonable to consider whether the results might have been different had different centers been used, and random-effect models handle this in a very natural manner.

These models have received some attention in the literature (e.g., Refs. [1,10,15,16,20–22]), and can be considered an additional option in the pharmaceutical industry for primary analyses, but they have by no means superseded the use of fixed-effect models. Random-effects models are considered to be a “good fit” more often for trials in which there are a large number of centers each contributing only a small proportion of the total number of patients in the trial. It has frequently been stated, and is certainly true, that centers used in clinical trials are not selected by any type of random mechanism from some population of potential sites. On the contrary, they are often selected precisely because of attributes in which they differ from other sites; e.g., the investigators possess some special expertise, the center has experience in participating in clinical

trials, it can enroll the desired number of patients in a timely manner, etc. Thus, while presumably there is some continuity principle operating so that results from clinical trials would provide valuable information as to how treatments will work in other settings, it is usually difficult to describe the population from which the centers used in a clinical trial could be expected to behave as a random sample. While there is currently an increased body of opinion that random-effect models would provide more appropriate inferences in multicenter clinical trials, until relatively recently the historical consensus in the pharmaceutical industry, at least as evidenced by selection of primary analysis models, seems to have been much as described by Fleiss,^[20] namely that while neither fixed nor random approaches seem uniformly a perfect fit, clinical trial sites seem “more fixed than random.”

Though there are fundamental differences between the assumptions made in fixed and random-effect models that prevent simple comparisons, we can address the weights that centers receive in random-effect analyses, and it is instructive to compare these to the fixed-effect weights described in previous sections. For a model of the form of Eq. (2), we now view C_j and I_{ij} as independent zero-mean normal random variables with variances, σ_C^2 and σ_I^2 , respectively (though under equal within-center treatment allocation, or something close to it, the choice of C_j as fixed or random will usually have minimal impact on the analysis).

A mixed model analysis (for example, restricted maximum likelihood, the default method in SAS PROC MIXED) can be viewed as determining weights that minimize the variance of the treatment difference estimate, using variance components that are simultaneously estimated. Further details are provided in Ref. [9], but in summary, these center weights satisfy

$$w_j^* \sim n_j / (Rn_j + 1 - R)$$

(i.e., w_j^* is proportional to this quantity, and the $\{w_j^*\}$ sum to 1), where R is the following ratio:

$$R = s_I^2 / (s_I^2 + s^2)$$

with s_I^2 and s^2 denoting the estimates of σ_I^2 and σ^2 , respectively.

Weights given to centers in this mixed model analysis will always lie between those they would have received in weighted and unweighted fixed-effect analyses. Note that as R gets small, the weights w_j^* approach those of a weighted analysis, which intuitively seems sensible, since if the data suggest that interaction is negligible, then there is no ambiguity in the main-effect parameter being estimated, and size-weighted estimates are most efficient. As R approaches 1 (i.e., s^2 is small relative to s_I^2), we approach equal weighting. Though perhaps less intuitive, and certainly less realistic, if “patient” essentially provides no

information beyond its center and treatment, then “center” is the only experimental unit, and it should then seem sensible to average equally over levels of this factor.

One might compute these weights to try to validate in some sense the choice between the common types of fixed-effect weightings—there is nothing that prevents one from applying these random-effect type weights in a fixed-effect model (in fact, an interesting discussion of attributes of this weighting scheme in an underlying fixed-effect framework is presented in Ref. [13]). Directly comparing fixed and random-effect model inferences is more problematic, but there are some tendencies we can note. The variance of the mixed model estimate d_M determined using weights $\{w_j^*\}$ can be seen to have two components:

$$\text{var}(d_M) = 2\sigma_i^2 \sum w_j^{*2} + 2\sigma^2 \sum w_j^{*2} / n_j$$

Perhaps not surprisingly, the first component will tend to be inversely related to the number of centers, and the second to the number of patients. The coefficient of the second component is always between the coefficients of σ^2 in weighted and unweighted fixed-effect models; hence, the variance of the mixed model estimate will normally be at least as large as in a weighted analysis. The first term inflates the variance further for any apparent inconsistency of effect across centers, which as alluded to previously, should seem reasonable: one should have more confidence that a suggested effect is real if it is consistent across centers, because then the outcome would seem less likely to have been influenced by the selection of centers that actually participated in the study.

If the interaction variance component is estimated to be zero, then the first term in the estimated variance disappears. The center weights will exactly be those of a weighted fixed-effect analysis, and inferences will be nearly identical—the estimate will be the same, though the error term and d.f. might differ slightly (if our assumption of within-center balance is violated, the weighted and mixed model estimates will tend to be very slightly different). Roughly, this will occur when the conventional ANOVA F -ratio for interaction is < 1 , corresponding to an interaction p -value of at least 0.50–0.60, with the exact threshold depending on the size of the study. If the interaction estimate is nonzero (again, roughly when the ANOVA p -value for interaction < 0.50), the mixed model variance will be larger than the weighted fixed-effect variance, and possibly much larger, depending on the magnitude of the interaction signal and the number of centers.

Rightly or wrongly, this may hinder broader use of random-effect analyses in the pharmaceutical industry, since current conventions accept weighted fixed-effects analyses, and the tendency when using random-effect models will be toward less significant results, even for interactions that do not approach conventional standards of significance. Use of a large number of centers would seem

to be imperative in order to stabilize the corresponding variance component if there were a signal of interaction. On the other hand, this approach could decrease current concerns about thresholds for significant interaction that might be interpreted as invalidating or shedding doubt on overall treatment comparisons: in mixed models, interaction is handled in a systematic manner, with lower or higher evidence of interaction translating naturally into a more or less significant outcome.

One other practical problem with implementing mixed models would seem to involve trial design: sample size estimates will often be highly dependent on hypothesized interaction variances, for which it will often be difficult to determine in advance a dependable value. Also, it follows directly from the discussion above that sample sizes needed to achieve a particular power will usually be larger for random-effect models that assume a positive interaction variance compared to conventional fixed-effect sample size calculations. In fact, power will not be solely a function of the number of patients, but of the number of centers and distribution of patients across centers as well.

EXTENSIONS TO NON-NORMAL DATA

The discussion here has focused on ANOVA models for normally distributed data. The issues are of course relevant for other data structures as well, though these have not received the same level of detailed attention in the literature. Binary or categorical outcomes, longitudinal data, or time-to-event data are just a few examples that are analyzed using different methods to which the principles and tendencies illustrated above apply to some extent. (A thorough investigation of these issues for multicenter trials with binary responses is provided in Ref. [23].)

Primary analyses that can be viewed as being based on equal-center weighting (analogous to the SAS type III approach) have probably been considered and performed far less frequently for alternative data structures. In most cases, there is no technical reason why they could not be; for example, a time-to-event analysis utilizing a proportional hazard approach could estimate the log hazard ratio within each center and combine these using Type III weighting. More frequently than for normal data, other methods may not even account for center at all, particularly when most centers are quite small and it is plausible that center does not strongly predict outcome (see e.g., ICH-E9:[18] “In some trials, for example, some large mortality trials with very few subjects per center, there may be no reason to expect the centers to have any influence on the primary or secondary variables because they are unlikely to represent influences of clinical importance. In other trials, it may be recognized from the start that the limited number of subjects per center will make it impracticable to include the center effects in the statistical model. In these

cases, it is not considered appropriate to include a term for center in the model.”). So long as within-center treatment allocation is similar across centers, analyses that ignore center can generally be interpreted as providing similar weight to each patient, thus giving weight to centers that is related to their size.

Some analysis methods allow centers to define strata. For example, binary data might be analyzed using a Cochran-Mantel-Haenszel analysis that is stratified by center; time-to-event data might be analyzed using stratified logrank tests or stratified Cox proportional hazard models. In such stratified analyses, it can again be viewed that appropriately scaled treatment effects are computed within each stratum and then combined based upon their (inverse) precision—in other words, size-weighted analyses. It seems as if equal center weighting in primary analysis models has most often been considered for normal data methods. For other methods and data structures, random-effect models would also have some potential to modify center weights similarly as was illustrated above.

CONCLUSIONS

In recent years, it has become more standard practice for primary analyses that use fixed-effects models in multicenter clinical trials to utilize methods that can be viewed as combining center-specific estimates according to their precision, i.e., more precise within-center estimates receive more weight. Precision will often be directly related to the number of patients enrolled at a center. This is reflected in current regulatory guidance documents. The advantage of this weighted approach is that it is in general a highly efficient framework within which one can combine the within-center estimates, similar to the way many statistical methods in various contexts combine information on an inverse precision-related basis. In normal linear models analyses, this weighting is usually achieved by using an ANOVA model which does not contain a term for treatment-by-center interaction. It is true that the parameter on which inference is based is not necessarily the same as that in an analysis in which within-center estimates are weighted equally. Nevertheless, the main question of practical interest—whether there is evidence of treatment effectiveness—will generally be reflected in any reasonable weighting choice, and the efficiency of the size-weighted approach will often be best able to provide a good initial signal as to whether such evidence exists.

This by no means implies that analyses that use other weightings should not also be produced and considered in full investigation of trial results. Whether or not effects are similar across centers can be important in judging whether the magnitude of a treatment effect has been adequately characterized, and whether there are other population characteristics or treatment

administration details that are relevant to understanding how well patients respond to the treatment. ICH-E9^[18] recommends: “If positive treatment effects are found in a trial with appreciable numbers of subjects per center, there should generally be an exploration of the heterogeneity of treatment effects across centers, as this may affect the generalizability of the conclusions.” Such interaction could cause estimated effects based upon different weightings to “aim” at noticeably different targets if the treatment effects are systematically related to center size. For example, if a size-weighted analysis suggests a larger treatment effect than an unweighted analysis, then there must be some apparent positive correlation between effect and center size, i.e., a tendency toward larger effects in larger centers, which may point in a direction for further investigation. In general, examining how analyses seem to be affected by different center weightings can be a useful diagnostic tool assisting in most fully interpreting and understanding trial results.

REFERENCES

1. Senn, S. *Statistical Issues in Drug Development*, 2nd Ed.; Wiley: Chichester, 2007.
2. Committee for Proprietary Medicinal Products (CPMP). *Points to Consider on Adjustment for Baseline Covariates*. CPMP/EWP/283/99 **2003**. Available on: <http://www.emea.europa.eu/pdfs/human/ewp/286399en.pdf>.
3. Anello, C.; O'Neill, R.O.; Dubey, S. Multicentre trials: a US regulatory perspective. *Stat. Methods Med. Res.* **2005**, *14*, 303–318.
4. Lewis, J.A. Statistical issues in the regulation of medicines. *Stat. Med.* **1995**, *14*, 127–136.
5. Källén, A. Treatment-by-center interaction: what is the issue? *Drug Inform J.* **1997**, *31*, 927–936.
6. Snapinn, S. Interpreting interaction: the classical approach. *Drug Inform J.* **1998**, *32*, 433–438.
7. Jones, B.; Teather, D.; Wang, J.; Lewis, J.A. A comparison of various estimators of a treatment difference for a multi-centre clinical trial. *Stat. Med.* **1998**, *17*, 1767–1777.
8. Lin, Z. An issue of statistical analysis in controlled multi-centre studies. How shall we weight the centres? *Stat. Med.* **1999**, *18*, 365–373.
9. Gallo, P. Center-weighting issues in multicenter clinical trials. *J. Biopharm. Stat.* **2000**, *10* (2), 145–163.
10. Senn, S. The many modes of meta. *Drug Inform. J.* **2000**, *34*, 535–549.
11. Schwemer, G. General linear models for multicenter clinical trials. *Control Clin. Trials* **2000**, *21*, 21–29.
12. Dragalin, V.; Fedorov, V.; Jones, B.; Rockhold, F. Estimation of the combined response to treatment in multicenter trials. *J. Biopharm. Stat.* **2001**, *11* (4), 275–295.
13. Permutt, T. Probability models and computational models for ANOVA in multicenter clinical trials. *J. Biopharm. Stat.* **2003**, *13* (3), 495–505.
14. Ganju, J.; Mehrotra, D.V. Stratified experiments reexamined with emphasis on multicenter trials. *Control Clin. Trials* **2003**, *24*, 167–181.

15. Fedorov, V.; Jones, B. The design of multicenter trials. *Stat. Methods Med. Res.* **2005**, *14*, 205–248.
16. Gould, A.L. Bayesian analysis of multicenter trial outcomes. *Stat. Methods Med. Res.* **2005**, *14*, 249–280.
17. SAS/STAT User's Guide; Version 9.1, Vol. 1; SAS Institute, Inc.: Cary, NC, 2004.
18. International Conference on Harmonisation. *Guidance on Statistical Principles for Clinical Trials (ICH-E9)*; Food and Drug Administration, DHHS, 1998.
19. Ruvuna, F. Unequal center sizes, sample size, and power in multicenter clinical trials. *Drug Inform. J.* **2004**, *38*, 387–394.
20. Fleiss, J.L. Analysis of data from multiclinic trials. *Control Clin. Trials* **1986**, *7*, 267–275.
21. Boos, D.D.; Brownee, C. A rank-based mixed model approach to multisite clinical trials. *Biometrics* **1992**, *48*, 61–72.
22. Patel, H.I. Robust analysis of a fixed-effect model for a multicenter clinical trial. *J. Biopharm. Stat.* **2002**, *12* (1), 21–37.
23. Agresti, A.; Hartzel, J. Strategies for comparing treatments on a binary response with multi-centre data. *Stat. Med.* **2000**, *19*, 1115–1139.

Clinical Data Management

Terri Madison and Marianne Plaunt

STATPROBE, Ann Arbor, Michigan, U.S.A.

Abstract

Clinical data management (CDM) is the process of collecting and validating clinical information with the goal of converting it into an electronic format to answer research questions and to preserve it for future scientific investigation. The objective of this entry is to present key data management concepts and to provide a brief overview of CDM processes.

INTRODUCTION

Clinical data management (CDM) is the process of collecting and validating clinical information with the goal of converting it into an electronic format to answer research questions and to preserve it for future scientific investigation. As computerized data increasingly become the basis for scientific research, CDM continues to grow as a discipline. Reporting accurate and reliable study results is dependent on the ability to efficiently access and accurately manipulate information stored in computer files. The collection of files can quickly become unmanageable without a systematic approach, and the CDM discipline provides a practical framework for creating well-documented data of consistent quality, permitting research projects to proceed quickly and easily.^[1]

The objective of this article is to present key data management concepts and to provide a brief overview of CDM processes. The article begins with the identification of current regulations and guidance documents with relevance to CDM, highlighting the specific parts of these regulations that directly apply to CDM practices. This section is followed by a discussion of CDM processes, beginning with case report form (CRF) design and ending with database closure. In this section, key principles for each process are provided, followed by a process overview. The article concludes with an overview of CDM systems.

REGULATIONS AND GUIDANCES RELEVANT TO CLINICAL DATA MANAGEMENT

Most CDM practices are not specifically addressed in existing regulations and guidance documents. The regulations and guidance documents currently in place that apply to CDM are identified in this section. The objective of this section is to describe the purpose and scope of each relevant regulation or guidance, highlighting the specific parts of each regulation or guidance that directly address CDM practices.

21 CFR Part 11—Electronic Records, Electronic Signatures

These regulations describe the criteria under which the Food and Drug Administration (FDA) will consider electronic records and signatures to be generally equivalent to paper records and handwritten signatures, and applies to any records required by FDA regulation or submitted to FDA under agency regulations.^[2] The requirements for audit trail systems are also covered in 21 CFR Part 11. A large portion of 21 CFR Part 11 has direct relevance to CDM. Persons who create, modify, maintain, or transmit electronic records are required to have procedures in place that ensure the authenticity and integrity of these records. Systems must ensure electronic records are accurately and reliably retained, must be able to discern invalid or altered records, and must track revisions to electronic records such that previously recorded information is not obscured. Closed systems must ensure access only by authorized individuals, and open systems must use additional measures such as document encryption to ensure authenticity and integrity. Systems must also have the ability to generate documentation suitable for agency inspection to verify that these requirements have been met. Persons responsible for the records must have the necessary education, training, and experience to perform the assigned tasks. Written procedures must exist for actions relating to electronic records and electronic signatures. Electronic signatures must contain associated information that indicate authenticity, must be certified to the agency as a legally binding equivalent of a handwritten signature, and must be linked to their respective electronic records to ensure signatures cannot be transferred to falsify an electronic record.

21 CFR Part 312—Investigational New Drug Application

Part 312 contains the procedures and requirements that govern the use of investigational new drugs, including the

procedures and requirements for submitting an investigational new drug (IND) to the agency.^[2] Part 312.62(b) has direct relevance to CDM, and covers requirements for investigator recordkeeping and record retention. Specifically, investigators are required to keep adequate case histories that record all data pertinent to the investigation on each individual administered an investigational drug or used as a control in the investigation. These case histories include the CRFs and supporting data.

21 CFR Part 314—Applications for Food and Drug Administration Approval to Market a New Drug

This part establishes procedures and requirements for the submission of applications and abbreviated applications to the agency, and for the agency to review the application.^[2] It also applies to application amendments, supplements, and post-marketing reports. Parts 314.50(f)(1) and (f)(2) pertain to the inclusion of CRFs and tabulations in an application, and therefore directly apply to CDM. Full case report tabulations are required from each adequate and well-controlled study (Phase 2 and Phase 3 studies) as well as from clinical pharmacology studies (Phase 1 studies), unless prior agreement was obtained from the agency. Safety data tabulations are required from other clinical studies. Applications must also contain copies of individual CRFs for each patient who died during a clinical study or who withdrew prematurely from the study due to an adverse event.

Bioresearch Monitoring—Compliance Program Guidance Manual

This guidance provides specific and uniform instruction for agency inspections of clinical investigators, sponsors, laboratories, and institutional review boards as part of the agency compliance program to assure the quality and integrity of data submitted to the FDA.^[3] Program 7348.810 is directly applicable to CDM, and governs sponsors, contract research organizations, and monitors. The guidance provides for establishment inspections of organizations responsible for data collection and handling, and includes the review of data tabulations for each subject, the review of the sponsor's data collection and handling standard operations procedures to assure the integrity of the data collected, and the verification that procedures were followed. The guidance also covers automated entry of clinical data to determine compliance with 21 CFR Part 11.

Guidance for Industry—Computerized Systems Used in Clinical Trials

This guidance provides additional detail regarding the principles to be followed when computerized systems are

used to create, modify, maintain, archive, retrieve, or transmit clinical data intended for submission to the agency.^[4] The guidance also addresses requirements similar to those in 21 CFR Part 11. The primary focus of this guidance is computerized systems used at clinical sites to collect data; however, the principles may also be appropriate for computerized systems at sponsors, contract research organizations, and CDM centers. Computerized medical devices and diagnostic laboratory instruments are outside the scope of this guidance. As this guidance contains many key principles with direct relevance to CDM, specific principles set forth in this guidance are highlighted in the section “CDM Processes” of this article.

International Conference on Harmonization—Consolidated Guideline for Good Clinical Practice

This document sets forth a tripartite standard for the conduct of clinical trials, and covers preparation, monitoring, reporting, and archiving of clinical trials.^[5] Section 5.5 pertains to trial management, data handling, and recordkeeping, and requires that: (1) appropriately qualified individuals handle and verify the data, (2) electronic data-handling systems used are validated, have corresponding standard operating procedures (SOPs), maintain an audit trail, prevent unauthorized access, maintain adequate data backup, and safeguard the blinding, and (3) original data are maintained in situations where data are transformed during processing. Section 8 sets forth the requirements for signed, dated, and completed CRFs and documentation of CRF corrections as essential trial documents that must be maintained.

CLINICAL DATA MANAGEMENT PROCESSES

Case Report Form Design

21 CFR Part 12 requires the investigator to record all data pertinent to the investigation for each individual administered an investigational drug or used as a control in the investigation.^[2] The study protocol delineates the observations required to demonstrate the safety and efficacy of an investigational product; the CRF is the instrument developed to systematically record these observations.

Many factors need to be considered in designing the CRF. Some of the key principles to consider are:

1. The CRF should be designed to capture all data required per the protocol.
2. The CRF should be designed to collect data elements in standardized format.^[6] Standardization facilitates the aggregation of data across clinical investigations in a drug development program, promotes consistency in data reported to regulatory agencies, and

decreases the potential for bias from recording data using dissimilar collection methods.

3. Data elements should be captured on the CRF in a fashion that ensures the data are suitable for summarization and analysis. Numeric data or predefined checkboxes of possible outcomes are more useful than open text. If necessary, text may be recorded in a way that facilitates the assignment of codes to similar responses so that textual data can be summarized.
4. Data elements planned to be transcribed to the CRF from source documents should be organized and formatted on the CRF to reduce the possibility of transcription error and to facilitate subsequent comparison to source documents. Units, format, and level of precision on the CRF should be similar to those of the source document.
5. Ease of completion for the investigator and study coordinator is key to accurate and timely CRF completion. Instructions for filling out the CRF should be prominently displayed. Related data elements should be clustered together and should flow similarly to patterns of medical record documentation. Checkboxes and areas to record responses should provide adequate space to record the anticipated results. Predefined response choices should be used to minimize unanticipated responses or ambiguous data. Adherence to these principles will also facilitate the review performed by clinical research associates (CRAs) and data coordinators of the completed CRF.
6. Redundant data elements within the CRF should be avoided, and unnecessary data should not be collected.

The CRF design process begins with a careful review of the protocol and selection of applicable CRF standards, if available.^[6] A protocol-specific draft CRF is developed, often in consultation with the study manager to ensure efficacy and that the safety data elements will accurately capture the intended data. The draft CRF is typically reviewed by the CRA, data coordinator, biostatistician, and study manager prior to finalization. After the CRF and instructions have been finalized, the CRF and associated instructions are assembled in patient-specific booklets or binders, printed, and distributed to investigators. The CRF is typically printed on two- part or three-part NCR paper, with the original (and first NCR copy, if present) for the sponsor and the last copy to be retained by the investigator. The CRF should be made available to investigators prior to enrollment of the first subject.

Electronic data collection system, such as a web-based electronic data capture system or an interactive voice response system (IVRS), are increasingly being used in clinical investigations. Data may be collected directly into the electronic system, may be recorded on source documents such as medical record progress notes and later

input into the electronic system, or may be recorded on protocol-specific worksheets developed specifically for the electronic data capture setting to be subsequently entered into the electronic system. In all of these situations, a CRF is not developed; however, most of the CRF design principles described in this section apply to the design of worksheets or data entry screens in an electronic data collection setting.

A more thorough discussion of CRF design principles and processes can be found in Ref. [6].

Database Design

The storage of data from clinical investigations must preserve the integrity of the data while ensuring efficiency throughout various data-handling processes. This section discusses key principles for developing and validating protocol-specific databases. A discussion of various commercially available types of data management software and a comparison of the advantages and disadvantages of each are provided in the section "Clinical Data Management Systems."

The database structure is driven by the data elements collected on the CRF. The use of standard templates, where available, to build protocol-specific database modules and format libraries (data dictionaries) enhances the efficiency of the database development process and facilitates subsequent aggregation of the data. Codes should be assigned to nonnumeric data to allow efficient data entry and retrieval into standard data tabulations and summary reports.^[7] If data are transformed, the original data recorded on the CRF must also be stored in the database.

The steps in developing a protocol-specific database are illustrated in Fig. 1. The process begins with the identification of applicable standard templates. Once a draft structure is available, team members such as the programmer and biostatistician review the structure to ensure the storage format of data elements supports data tabulation and summarization. Test data should be developed to validate the database and associated structures. The validation process should ensure that all variables are storing data as intended, that derivations are working accurately, and that the audit trail captures data revisions as required by 21 CFR Part 11. Additionally, test data should attempt to "stress" the system to ensure data storage works in situations of unanticipated responses or larger than expected amounts of data. After the database has been thoroughly validated and approved by the project team, the database is put into production and is ready to accept actual data from the clinical investigation.

Data-Handling Standards

Data-handling standards are created to ensure that data are reliable, complete, and accurate. Data handling includes receiving, entering, cleaning, coding, reconciliation, and

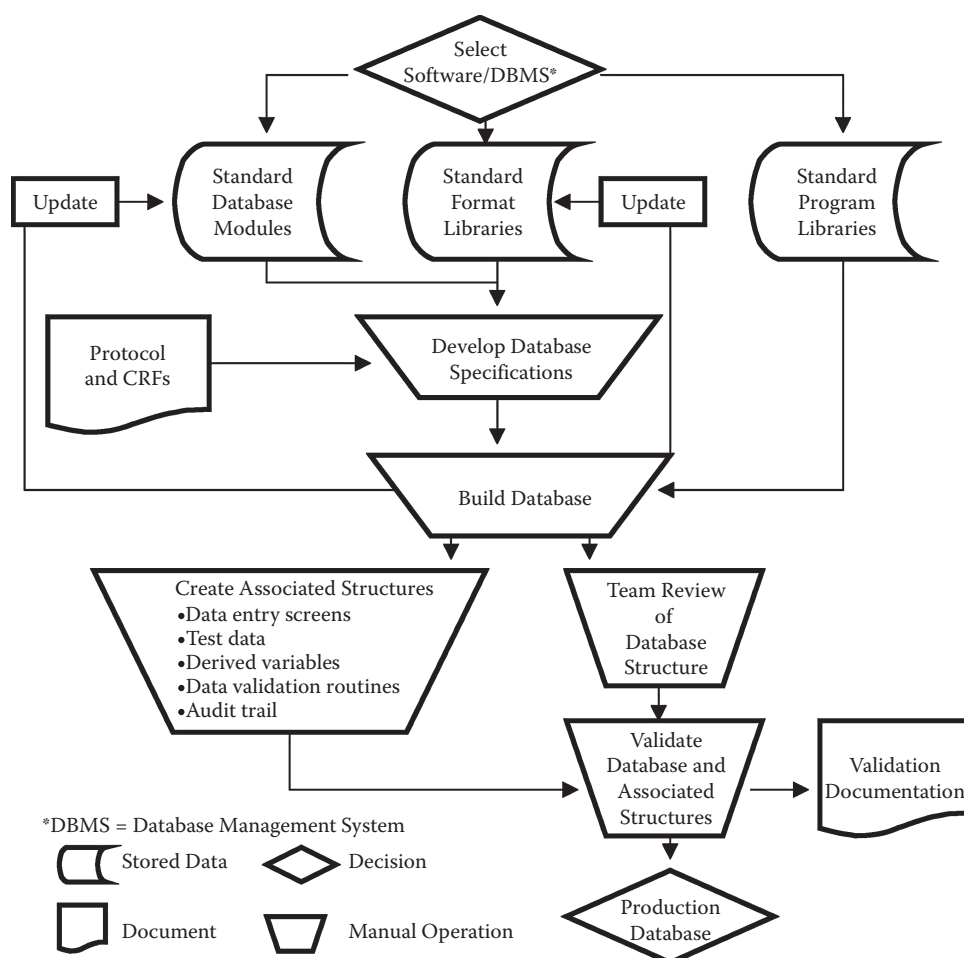


Fig. 1 Database development process

transferring of data. The data-handling process is illustrated in Fig. 2.

Prior to entering data, written procedures that describe data processing steps and the required quality level should be developed. A data management plan (DMP) is developed to document the data processing guidelines to be followed for a specific study. The DMP should describe CRF receipt and tracking processes, CRF review and query guidelines, specifications for computer edit check programs, data entry guidelines, quality assurance methods/audit plan for the study, coding processes and the coding dictionaries to be used, data collection and handling processes for electronic data from external sources, and serious adverse event (SAE) reconciliation processes. The DMP should provide enough specificity to reproduce the database from source documentation.

The data entry process utilized should address the data quality needs of the investigation. FDA and ICH regulations do not stipulate a specific data entry process. The skill level of the resource, the time available, the overall cost-effectiveness, and the impact of data errors on study conclusions should all contribute to the choice of a data entry process. Double data entry is helpful where data are

subject to frequent, random, keystroke errors or where a random error would be likely to impact the analyses. There are two main types of double data entry: (1) independent double data entry with a third person compare, where two persons independently enter the same data, and a third person resolves any discrepancies between the two entries and (2) double data entry with blind or interactive verification, where two persons enter the data independently but sequentially, and discrepancies are resolved during the second entry pass. A single data entry process, with or without visual verification, may also be appropriate for certain investigations. Imaging and optical character recognition (OCR) processes may be used to populate a database. This process automates the first entry, and is followed by manual verification of specific fields that did not meet the established confidence criteria.

In studies where electronic data collection systems are used, a typical process entails a single entry of the data into a centralized database through a web browser or utilizing a client/server model by site personnel, with periodic transmissions to the central database. The data are subsequently verified using source documents. Other forms of electronic data capture include IVRS and direct data capture

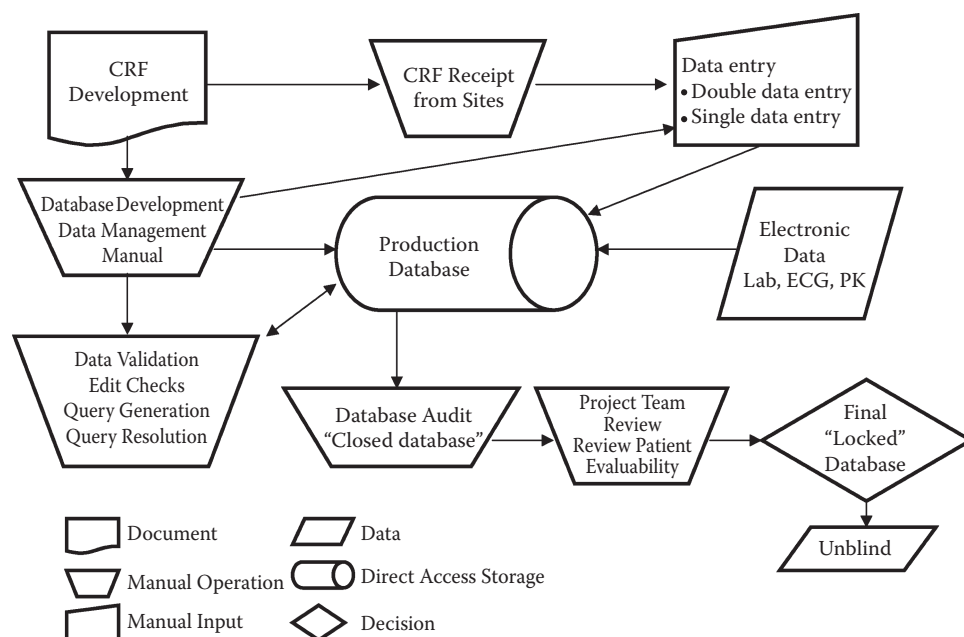


Fig. 2 Data management process traditional data collection

devices such as electronic patient diaries. If electronic data collection methods are utilized, validation of the data collection application is especially critical. In electronic data collection, data directly input into the system may be the only source data available, and thus the system must be demonstrated to perform reliably in collecting the data.

Incorporation of electronically collected data into the protocol-specific database presents several unique steps to the data-handling process. Key variables (data that uniquely identify each record) must be identified and included in the data transfer, as key variables are used to merge external data with CRF data (if applicable) in the protocol-specific database. Variables that will be transmitted from electronic data sources to be loaded into the protocol-specific database should be identified, and variable attributes such as length, type, level of precision, and the use of special characters should be defined. Variables to be transferred, record format, and file format should be agreed upon in writing prior to the first data transmission. A file naming convention and inclusion of information such as file origin, date created, date sent, and number of records should be followed to ensure proper identification and version control for each transmission. Ideally, test data from electronic data capture systems should be made available to use during database development and validation. Successful generation, transmittal, receipt, loading, and screening of the test data validate the data transmittal process.

Data Validation

Data validation is the collection of activities used to ensure the validity and accuracy of the data. Activities for data validation may include manual review of the data; however,

programmatic validation provides high consistency and lower error rates. Procedures should be neutral and should not suggest bias or lead responses. The objectives of data validation are:

- Ensuring data entered into a computer-readable file are consistent with the source data.
- Checking that character and numeric variables contain only valid values.
- Checking for and eliminating duplicate data entries.
- Checking for missing values for variables where complete data are necessary.
- Checking for uniqueness of certain values, such as subject identification.
- Checking for invalid date values and invalid date sequences.
- Verifying that complex multiframe (or cross panel) rules have been followed.^[9]

Data may be changed as a result of data validation procedures. Both the investigator site and the data center must retain a record of all changes to the data. Data changes must be recorded and documented by a fax or original site signature acknowledging the new data; this is usually accomplished using a query or data clarification form. Under certain circumstances, data conventions may specify data that can be modified without a site's acknowledgement. These are typically referred to as "obvious changes" or "level 1" changes, and may include correction of obvious spelling errors, unit conversions, and completion of missing identifiers. As required in 21 CFR Part 11, an audit trail should reflect all deletions, updates, or additions to data after initial entry. The audit trail should also

reflect the date and time of the change, the user making the change, and the reason for the change.

Data validation processes in an electronic data capture setting differ from those in a paper-based setting. Many of the objectives of data validation in a traditional paper-based system require postentry processes; an advantage of electronic data capture systems, if designed well, is that most data validation objectives can be accomplished within the system at the point of data collection. Data validation procedures for electronically captured data need to include conventions for handling partial data transfers and procedures for checking for duplicate demographic details and results. Additionally, processes for identification and resolution of discrepancies between electronically captured data and CRF data, where redundancies in some of the data captured from each source may be common, should be defined in the DMP.

Closing a Database

At the end of the data management process, after all data have been received, entered, or transferred into the database, and validated, the database is ready to be closed and prepared for analysis. Closing a database implies no further additions or revisions to the data will be made, and access will be restricted to “read-only” status. Some CDM centers further differentiate between a “closed” and a “locked” database, where “closed” implies that all CDM tasks have been completed. Project team review and additional derivations such as patient evaluability are subsequently completed prior to “locking” the database. Unblinding of treatment assignment is typically performed after the database has been locked,^[8] although exceptions where unblinding occurs just prior to lock have been noted.

Prior to closing a database, the following steps should be completed:

- All data have been received and entered (or transferred) into the clinical database.
- All data validation procedures have been performed, and resulting data discrepancies have either been resolved or deemed to be not resolvable.
- Derived variables have all been successfully populated, including variables derived from coding thesauri such as MedDRA^[10] or WHODrug.^[11]
- An audit of the data has been conducted that certifies the overall quality of the database to meet a preestablished quantitative metric.

Developing a checklist of steps to complete prior to database closure helps ensure a systematic approach and completion of necessary steps. Each step on the database closure checklist is systematically checked to ensure completion, and the process culminates with the database audit. The audit, performed by persons independent of the CDM project team, typically involves the selection of a statistically appropriate random sample of the data to estimate overall quality, with verification of the stored data against the source document and assurance that data-handling procedures were followed for data included in the random sample. If the database audit demonstrates that quality acceptance criteria have been met, the database is closed.

Currently, federal regulations and guidelines do not stipulate an acceptance level for the quality of clinical investigation data. The Institute of Medicine defines data quality as “data that support conclusions and interpretations equivalent to those derived from error-free data.”^[12] Acceptance rates in current use in the industry vary widely;

Table 1 Data Management Software Applications

Software	Research uses	Advantages	Disadvantages
Spreadsheet • MS Excel • Lotus 123	Data entry into a matrix of columns and rows; simple transformation, descriptions, sorting and reporting	• Easy entry of raw data • Derived values may be programmed • Summary descriptive statistics and frequency distributions • Facilitates logic checks through sorting and calculated fields	• Missing value codes cannot be ignored when calculating descriptive statistics • Slow speed when there are many observations • Absence of functions for editing, accessing, and analyzing the data
Database management • Oracle • MS Access • Sybase • SQL Server	Data entry with interactive editing and greater ability to manipulate, access and report the data	• Suitable for studies with large numbers of variables and observations • Interactive editing • Automatic logic checks • Allows duplicate data entry	• Require experienced programmer for system customization • Modifications due to changes in data collection forms may be difficult
Statistical/graphics • BMDP • SAS-PC • SPSS-PC • S-plus • STATA	In addition to some of the above features: statistical analysis and graphics	• Data entry directly into a statistical analysis program • Bypass the need to transfer data from spreadsheet or database management system for statistical analysis	• Limited editing and data management functions

however, current practice recommends an acceptance error rate between 0.01% and 1%, and best practice recommends an acceptance error rate < 0.01% (excluding comments).^[8]

CLINICAL DATA MANAGEMENT SYSTEMS

A CDM system is defined as a set of systematic CDM activities and procedures related to acquiring, manipulating, documenting, and storing research data in a computerized environment. A CDM system is composed of interrelated CDM processes, such as data collection, data entry, data validation, and backup.

One of three main classes of software is usually chosen for data management: spreadsheet programs, database management systems (relational databases), and statistical/graphics programs.^[7] Research applications of these software systems are shown in Table 1. The choice of software is in part driven by the nature of the investigation. An efficient approach is to choose the simplest alternative, considering the resources available for the project such as personnel, time, and computing equipment.

As required by 21 CFR Part 11 and ICH GCP guidelines, CDM systems are required to prevent unauthorized access to the data, and must ensure the authenticity of the stored data. Regulatory requirements for CDM systems are described in detail in the section "Regulations and Guidances Relevant to Clinical Data Management."

CONCLUSION

CDM is a broad and growing discipline that requires knowledge and expertise in information technology and systems, programming practices, regulations, and quality practices, as well as familiarity with medical terminology and clinical care to ensure data are appropriately collected, reviewed for consistency, and resolved within the context of clinical appropriateness. The tremendous growth in the scale and number of research investigations has created a need for knowledgeable clinical data coordinators. As the biopharmaceutical and medical device industry and regulatory bodies increasingly rely on computerized clinical trial data for critical decision making, it is essential that clinical data coordinators pursue additional training in this discipline to ensure the highest accuracy and integrity of these data. The development and implementation of a standardized CDM training program would greatly facilitate this objective.

ACKNOWLEDGMENTS

The authors would like to acknowledge Dave Thurston, M.A., Director of Quality Assurance at STATPROBE,

Inc., for his assistance in providing information regarding regulations and guidance documents pertinent to CDM. The authors also acknowledge the Society for Clinical Data Management and their publication on "Good Clinical Data Management Practices,"^[13] which provided valuable insight regarding industry CDM practices.

REFERENCES

1. Calvert, W.S.; Ma, M.J. *Concepts and Case Studies in Data Management*; SAS Institute: Cary, NC, 1996.
2. Food and Drug Administration. *Code of Federal Regulations, Title 21, Chapter 1, Parts 11, 312, and 314*. U.S. Department of Health and Human Services, Food and Drug Administration: Rockville, MD, 1997, 1998.
3. Food and Drug Administration. *Bioresearch Monitoring—Compliance Program Guidance Manual, Chapter 48*; U.S. Department of Health and Human Services, Food and Drug Administration: Rockville, MD, 1998, 2001.
4. Food and Drug Administration. *Guidance for Industry—Computerized Systems Used in Clinical Trials*; U.S. Department of Health and Human Services, Food and Drug Administration: Rockville, MD, April 1999.
5. International Conference on Harmonization, *Guideline for Good Clinical Practice, Federal Register* 62(90); U.S. Department of Health and Human Services, Food and Drug Administration: Rockville, MD, May 1997.
6. Spilker, B.; Schoenfelder, J. *Data Collection Forms for Clinical Trials*; Raven Press: New York, 1984.
7. Feigal, D.; Black, D.; Grady, D.; Hearst, N.; Fox, C.; Newman, T.B.; Hulley, S.B. Planning for Data Management and Analysis. In *Designing Clinical Research*; Hulley, S.B.; Cummings, S.R., Eds.; Williams and Wilkins: Baltimore, MD, 1988, 159–171.
8. Helms, R.W.; Fitzmartin, R.; Fillo, P.; Miller, S.; Talley, L.; Murphy, R. Metrics and best practices in clinical data management: Conclusions of a DIA roundtable workshop. *Drug Inf. J.* **2001**, 35 (3), 681–694.
9. Cody, R. *Cody's Data Cleaning Techniques Using SAS Software*; SAS Institute Inc.: Cary, NC, 1999.
10. *Medical Dictionary for Regulatory Activities Terminology (MedDRA), Version 3.3*; Maintenance and Support Services Organization (comprised of TRW, Cyntergy, Quintiles, and Ernst & Young), December 2000.
11. *WHO Drug Dictionary*; World Health Organization: Geneva, Switzerland, 1999. Maintained by the Uppsala Monitoring Center, Uppsala, Sweden
12. *Assuring Data Quality and Validity in Clinical Trials for Regulatory Decision Making, Roundtable on Research and Development of Drugs, Biologics, and Medical Devices*; Davis, J.R. Nolan, V.P. Woodcock, J., Estabrook, R.W., Eds., Division of Health Sciences Policy, Institute of Medicine, National Academy Press: Washington, DC, 1999.
13. *Good Clinical Data Management Practices, Version 2*; Society for Clinical Data Management: January 2002. <http://www.scdm.org/members/images/GCDMP%2010-15.PDF>

Clinical Endpoint

Sofia Paul

Eli Lilly and Company, Carmel, Indiana, U.S.A.

Abstract

Clinical endpoint is a clear, hard, clinically relevant outcome measure that is affected by the intervention under investigation and that can be objectively assessed. This entry works to illuminate several types of endpoints in clinical trials and addresses key medical and statistical issues.

INTRODUCTION

A clinical trial is an experiment performed by a health care organization or professional to evaluate the effect of an intervention or treatment against a control in a clinical environment. It is a prospective study to identify outcome measures that are influenced by the intervention. A clinical trial is designed to maintain health, prevent diseases, or treat diseased subjects. The safety, efficacy, pharmacological, pharmacokinetic, quality-of-life, health economics, or biochemical effects are measured in a clinical trial.

There are two different types of clinical trials, confirmatory and exploratory trials.^[1] In an exploratory trial, the choice of hypothesis is data dependent, even though this study may have clear objectives. These trials explore the doses of subsequent studies and provide a basis for a confirmatory study design. A confirmatory trial is a well-controlled study in which the hypothesis of interest is predefined and is intended to provide hard evidence in support of claims that have clinical benefits. A confirmatory trial is less vulnerable to bias and more robust. Atherosclerotic cardiovascular disease (myocardial infarction and cardiovascular death) is the leading cause of death in women in the United States and in most developed countries worldwide. An example of a confirmatory cardiovascular clinical trial would be a multicenter, double-blind, placebo-controlled, randomized, parallel study of 10,000 patients to test whether a new treatment compared with placebo reduces the incidence of the combined endpoint of coronary death and nonfatal myocardial infarction in postmenopausal women at risk for coronary events.

CLINICAL TRIAL

Objective

A clinical trial generally consists of two different sets of objectives: primary and secondary. The primary objective

is the most important question or the hypothesis an investigator desires to answer definitively at the end of the study and that drives the size and the power of the trial. The hypothesis of interest in a confirmatory trial is directly related to the primary objective. The secondary objective(s) answer other related or unrelated questions based on the efficacy or safety of the intervention. In our earlier example, the primary objective is to compare the effect of the treatment with placebo on the coronary death and nonfatal myocardial infarction. There could be several secondary objectives, e.g., to test the effect of the treatment on myocardial revascularization, stroke, all-cause mortality, all-cause hospitalization, breast cancer, fractures, and long-term safety.

Outcome Measures

An outcome measure is a direct or indirect measurement of a clinical effect in a single, adequate, well-controlled clinical trial (it can be double blind or open label) to obtain an effective claim regarding the intervention under investigation. The goal of the clinical trial is to assess the effect of the treatment on these outcome measures. In the primary objective there is generally one primary outcome, which is used to test whether the intervention is superior to a comparative agent (placebo control or active) or that it is not clinically inferior (equivalence) to a comparative agent (placebo control or active). Outcome measures should be selected prospectively.

Primary Endpoint

The primary endpoint is an outcome measure that is related to, and selected based on, the primary objective of the study. The size of the trial is based on the primary objective of the study and the primary endpoints. The primary endpoint in our earlier example is the combined endpoint of coronary death and nonfatal myocardial infarction.

Secondary Endpoints

Secondary endpoints are other outcome measures that may or may not be related to the primary objective of the study. Secondary variables are related to secondary objectives. They support the primary variables of the study. Secondary variables can be a “time to an event,” “incidence rate of an event,” or continuous variables related to efficacy or safety. In the previous example, the secondary endpoints are myocardial revascularization, stroke, all-cause mortality, all-cause hospitalization, breast cancer, fractures, and safety laboratory parameters.

Primary and secondary endpoints are intimately related to the primary and secondary hypotheses. These endpoints can be clinical, surrogate, economic, global, etc. To achieve the study objective, the design should be chosen adequately to reflect the objectives, appropriate primary, secondary, and global variables will be considered that will answer the objectives.

Clinical Endpoint

A clinical trial endpoint is defined as a measure that allows us to decide whether the null hypothesis of clinical trial should be accepted or rejected. Thus, a clinical endpoint is a clear, hard, appropriate outcome that can be objectively assessed, not depending on the judgment of an individual, especially the investigator. Nonfatal myocardial infarction is an example of an objective endpoint, in contrast to a subjective endpoint, e.g., relief of symptoms or severity of symptoms.

Surrogate Endpoint

An endpoint that provides an indirect measurement of a clinical effect when measuring outcome directly is not feasible or practical (because of the requirement of large sample size, long duration, and cost) is called a surrogate endpoint.^[2] A surrogate endpoint is a measure of effect of a certain treatment that may correlate with a clinical endpoint. Changes induced by a treatment on surrogate variables are expected to reflect changes in a clinical endpoint. It should be noted that NIH defines surrogate endpoint as a biomarker intended to substitute for a clinical endpoint. Generally, these endpoints are continuous response variables that can be substituted for the clinical outcomes. For example, surrogate variables might include biochemical markers of cardiovascular disease such as low-density lipoprotein cholesterol, total cholesterol instead of myocardial infarction; spine bone mineral density instead of incidence of vertebral fractures, and CD4 count and viral load effects for the effect on HIV infection.

Economic Endpoint

In recent clinical trials, the evaluation of a subject's health status data has become increasingly important. Several

measurements regarding a subject's use of health care, cost of hospitalization, etc. are considered as economic endpoints. The quality-of-life scores based on a subject's performance, daily activity, mood, etc. (defined by the World Health Organization) are subjective measures of health status. The overall score from these measurements can be analyzed for treatment comparison.

Global Assessment Variables

Global assessment variables measure the overall safety and efficacy of an intervention. These are generally based on rating scales. An idea of the risk–benefit profile of an intervention can be assessed from these variables. It helps the investigative physician to balance the safety and efficacy of the intervention and decide on treating subjects by weighing its risk and benefit outcomes.

CLINICAL ENDPOINT

Clinical endpoint is a clear, hard, clinically relevant outcome measure that is affected by the intervention under investigation and that can be objectively assessed, not depending on the investigator's judgment. A clinical endpoint is to be selected that can be reasonably and reliably assessed and can answer the primary objective. Sometimes it is difficult to achieve both aims, so in such cases clinical judgment is required. Any definitions used to characterize the primary outcome measure should be explained clearly. Sometimes two different clinically meaningful endpoints can cross-substantiate a claim for the effect of each outcome.

Medical Issues

The outcome measures chosen to evaluate the efficacy or the primary objective depends on a number of factors, including:

- Knowledge of adverse effects of closely related drugs.
- Information from nonclinical or earlier clinical trials.
- The types of subjects to be enrolled.
- Pharmacodynamic or pharmacokinetic properties of the treatment.

The endpoints should be capable of unbiased assessment. To have unbiased assessment of the endpoints, studies should be blinded. In some studies, the intervention is administered by one investigator and the measurements and evaluation of the endpoints is performed by an independent investigator not involved with the study. This is to maintain the trial integrity and to minimize bias introduced by any difference in investigators' judgments. The instrument to assess the primary variable should be reliable and accurate and adequately sensitive to detect any real change

for a subject. Clinical endpoints generally provide strong scientific evidence regarding efficacy in a targeted population. Clinical endpoints should be such that they can be assessed on all subjects. The response variable should be measured the same way for all subjects. The assessments of the response variables should be selected as standard, widely used, and generally recognized.

Statistical Issues

The sample size and power calculation for a study should be based on the primary endpoint. The endpoint could be incidence rate of an event or time to an event. Because recurrences of the primary event can occur, distribution of the number of events or frequency of episodes can be of interest in this case. Sometimes subject participation ends as soon as the primary event occurs, but sometimes the investigator wants to follow a subject for a subsequent primary variable or other secondary response variables. However, if a secondary variable occurs first, the subject must be followed, because he or she is still at risk for the primary variable. If the secondary variable is a mortality endpoint, then competing risks analyses may be appropriate.

For studies that have long-term duration, it is very important to have complete information on the subjects because if the subjects are lost to follow-up, sometimes loss of information occurs in both the treatment and the control groups. All the statistical analyses for primary or secondary variables are usually based on the intent-to-treat principle for all the randomized subjects.^[3] Sometimes post hoc analyses based on a subset of the randomized population are examined—for example, the effect of an intervention for reducing the rate of occurrence of myocardial infarction in the diabetic population, to test whether the treatment is more effective in the diabetics subset or, irrespective of the population, the treatment effect is the same. Binary data can be analyzed to obtain odds ratio or difference in proportions, and time-to-an-event data can be analyzed using proportional-hazard models, Kaplan–Meier estimates. There are additional statistical issues associated with several clinical endpoints, as shown in the next sections.

Combined Endpoints

Sometimes the primary outcome measure occurs very infrequently, so a large sample size or a long study duration is needed to have adequate power to detect a treatment difference. In this situation, combined endpoints (e.g., fatal and nonfatal myocardial infarction) are considered, to enable the detection of treatment differences with a smaller sample size or shorter study duration. This combined endpoint should have a meaningful clinical interpretation; in our earlier case it is a measure of coronary heart disease. Sometimes it is useful to combine

more than one measurement related to primary objective, e.g., rating scales in psychiatric disorders. In this case of combined endpoints, multiplicity adjustment to the type I error is not needed. Different variables of the combined endpoint can also be analyzed separately but may require multiplicity adjustment (see the next section). It is important to remember when answering a question relating to the combined endpoint that only one event per subject should be counted, and also a hierarchy of each component should be placed (for instance, the fatal myocardial infarction will be counted over the nonfatal myocardial infarction). In addition, a combined endpoint could be considered in order to achieve more power to perform subgroup analyses.

Multiple Endpoints

Multiple endpoints are different clinical events that reflect a common mechanism of action because of the intervention. Sometimes it is important to have more than one primary variable when the investigator cannot state which of several variables addresses the primary objective of the study. Use of multiple endpoints will result in an increased probability of having false-positive results. If more than one response variable is chosen, a nominal p value for each variable will be computed. If one of the multiple variables is of most importance and is most impacted by the intervention, then Bonferroni^[4] or similar adjustment methods can be used. On the other hand, if the intervention affects all the variables the same way, then adjustment is not necessary, although the effect on the type II error and the sample size should be evaluated. The O'Brien test^[5] of treatment equivalence as a global hypothesis can also be used in that case. Another way to deal with multiple correlated endpoints is to use a weighted average of the variables as a global index. The weight of each variable is the sum of the inverse variance and the covariances of that variable with the others.^[6] Thus, endpoints that are less correlated with other endpoints have more weights. When all the endpoints result in statistical significance, the study becomes more effective and assertive. For example, the clinical trial for prevention of exacerbations in multiple sclerosis using beta-interferon was based on a study where two completely different endpoints—a rate of exacerbations and an MRI-demonstrated disease activity—were used. Both endpoints were significantly decreased, which resulted in the approval.

CLINICAL VS. SURROGATE ENDPOINTS

Clinical trials with a clinical endpoint are usually longer in duration, to determine an extended exposure of the treatment. Since the clinical endpoint drives the calculation of the sample size, these trials involve a large number of subjects. On the other hand, surrogate endpoints can replace the true clinical endpoint, because it will often

result in a shorter duration and the smaller sample size of the trial. The sample size will be smaller because of the continuous nature of the data. Also, the duration of the trial will be reduced because the surrogate variables tend to change earlier in the study and the measurements can be taken at several time points.

While choosing the surrogate variables, there are certain important things to consider:

- Select the variables that have previously proven in literature to be highly correlated with the clinical endpoint.
- Choose the ones that will be accepted by the regulatory authorities and the medical community.
- Select the variables that can be reliably and accurately measured.
- Many invasive procedures should be avoided because they may result in high discontinuation rates.

According to Prentice,^[7] a valid surrogate variable should be correlated with the clinical endpoint and should capture fully the treatment's aggregate effect (accounting for all the mechanisms of action) on the clinical endpoint.

Statistical Considerations of Surrogate Endpoints

An example of a clinical endpoint in a cardiovascular trial is myocardial infarction; a surrogate variable is LDL cholesterol. Comparisons from the beginning and the end of the study or a certain interval using change or percentage change can be of interest. Comparison of slopes throughout the interval and comparison of treatments based on repeated-measures data are also analyzed.

Responder Variable

When a continuous response variable is transformed to a binary or a categorical variable it can be interpreted as a responder variable. For example, the original continuous variable percentage change in LDL cholesterol can be modified to a binary variable that takes two values: 1,

if the percentage change in LDL is less than 0; 0, if the percentage change in LDL is greater than or equal to 0. The disadvantage in creating this responder variable is that not all the information is used in the analyses, which results in loss of power.

ACKNOWLEDGEMENTS

Minor updates have been made by the Editor-in-Chief in order to reflect developments since publication of the *Encyclopedia of Biopharmaceutical Statistics, Third Edition*.

REFERENCES

1. ICH Steering committee. ICH Harmonized Tripartite Guideline, Statistical Principles for Clinical Trials, 1998.
2. Armitage, P.; Colton, T. *Encyclopedia of Biostatistics*, Eds.; Wiley: Chichester, 1998, Vol. 1–6.
3. Gillings, D.; Koch, G. The application of the principle of intention-to-treat to the analysis of clinical trials. *Drug Inf. J.* **1991**, 25, 411–424.
4. Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **1988**, 75, 800–802.
5. O'Brien, P.C. Procedures for comparing samples with multiple endpoints. *Biometrics* **1984**, 40, 1079–1087.
6. Pocock, S.J.; Geller, N.L.; Tsiatis, A.A. The analysis of multiple endpoints in clinical trials. *Biometrics* **1987**, 43, 487–498.
7. Prentice, R.L. Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat. Med.* **1989**, 8, 431–440.

Suggested Readings

1. Friedman, L.M.; Furberg, C.D.; Demets, D.L. *Fundamentals of Clinical Trials*, 2nd Ed.; PSG: Littleton, MA, 1985.
2. ICH Steering committee. ICH Harmonized Tripartite Guideline, General Considerations for Clinical Trials, 1997.
3. Pocock, S.J. *Clinical Trials: A Practical Approach*; Wiley: New York, 1983.

Clinical Endpoint BE Studies: Generic Drugs

Stella Grosser and Misook Park

Office of Biostatistics, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Sam G. Raney

Office of Research and Standards, Office of Generic Drugs, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Elena Rantou

Office of Biostatistics, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

We describe some of the statistical approaches used in clinical endpoint studies of bioequivalence for generic drugs. We outline a conventional design and statistical analysis for such studies, including the equivalence criteria. We exemplify this approach with a more detailed discussion of studies of nasal spray products for allergic rhinitis, pointing out some of the statistical issues particular to this area, and briefly describe clinical endpoint study design and analysis recommended for equivalence determination of dry-powder inhaled products. We also discuss a case with a topical dermatological drug product to illustrate how equivalence is assessed based upon clinical endpoints, and we discuss the future statistical challenges and opportunities associated with new approaches that are being developed to evaluate equivalence for locally-acting generic topical drug products.

Keywords: Bioequivalence; Clinical endpoint studies; Locally-acting generic drugs; Two one-sided tests; Equivalence limits.

INTRODUCTION

The 1984 Drug Price Competition and Patent Term Restoration Act, known as the “Hatch-Waxman Act,” established the modern system of generic drugs in the United States. Under this system, generic drug products are considered therapeutically equivalent to the innovator reference listed drug (RLD) product if they meet the regulatory criteria of pharmaceutical equivalence and bioequivalence. Pharmaceutically equivalent drug products contain the same amount of active ingredient in the same dosage form and meet the applicable standards defined by the United States Pharmacopeia. Independently, bioequivalence (BE) is defined as “the absence of a significant difference in the rate and extent to which the active ingredient or active moiety in pharmaceutical equivalents or pharmaceutical alternatives becomes available at the site of drug action when administered at the same molar dose under similar conditions in an appropriately designed study.” The U.S. Food and Drug Administration (FDA) may accept evidence from various approaches for determining the bioavailability (BA) or BE of a drug product, including in vivo pharmacokinetic (PK) studies, certain in vitro studies that are correlated with and predictive of human in vivo BA, and well-controlled clinical studies.^[1–3]

For most oral drug products, plasma PK studies provide useful information. Levels of the drug found in the systemic circulation are directly related to bioavailability at the site of action, and parameters such as AUC (the area under the drug concentration-time PK profile in the plasma) and C_{max} (the maximum drug concentration in the plasma) are used to compare the generic and brand name (innovator) products and determine bioequivalence. For locally acting products such as topical dermatological products or those nasal sprays and inhalation products that are intended for local action, traditional plasma PK studies can be useful to evaluate systemic exposure and safety but may not necessarily be relevant for evaluating BE because the drug levels in the systemic circulation may not necessarily be representative of local bioavailability of the drug at or near the site of action. In some situations, comparative clinical endpoint studies have been used to determine BE. In this article, we describe some of the statistical approaches used in clinical endpoint studies of bioequivalence. We outline a conventional design and statistical analysis for such studies, including the equivalence criteria. We exemplify this approach with a more detailed discussion of studies of nasal spray products for allergic rhinitis, pointing out some of the statistical issues particular to this area; and briefly describe clinical endpoint study

design and analysis recommended for equivalence determination of dry-powder inhaled products. We also discuss a case with a topical dermatological drug product to illustrate how equivalence is assessed based upon clinical endpoints, and we discuss the future statistical challenges and opportunities associated with new approaches that are being developed to evaluate equivalence for locally-acting generic topical drug products.

GENERAL CONSIDERATIONS

Traditionally, in PK studies, bioequivalence has been assessed using two one-sided tests^[4] applied to the ratio of average log-transformed responses. In these cases, the generic or ‘test’ product and the innovator or ‘reference’ product (RLD) are declared bioequivalent if the 90% confidence interval (CI) of the ratio of the population geometric means is contained within the interval [0.80, 1.25]^[5]. Use of the 90% CI ensures that the total Type I error is controlled at the 0.05 level. With the equivalence limits of [0.80, 1.25], the BE ‘margin’ corresponds to a value of 1.25, and, more generally, a value of *m*, when the interval [(1/*m*), *m*] is used. These limits are philosophically related to non-inferiority margins, but bi-directional, in the sense that superiority is ruled out as well as inferiority.

Clinical endpoint studies routinely utilize this as the default approach to a statistical analysis of BE. For continuous endpoints, the FDA recommends^[6] that the following compound hypotheses are tested using the per protocol population:

$$H_0 : \mu_T / \mu_R < \theta_1 \text{ or } \mu_T / \mu_R > \theta_2 \text{ versus}$$
$$H_A : \theta_1 \leq \mu_T / \mu_R \leq \theta_2$$

where μ_T = mean of the primary endpoint for the test group (generic), and μ_R = mean of the primary endpoint of the reference (brand name) group

The null hypothesis, H_0 , is rejected with a type I error (α) of 0.05 (two one-sided tests) if the estimated 90% confidence interval for the ratio of the means between test (T) and reference (R) products (μ_T / μ_R) is contained within the interval [θ_1 , θ_2], where $\theta_1 = 0.80$ and $\theta_2 = 1.25$. Rejection of the null hypothesis supports the conclusion of equivalence of the two products.

For binary endpoints, (e.g., cure rate), the difference in proportions should be within $\pm 20\%$. The FDA

recommends^[5] that the following compound hypotheses are tested using the per protocol population:

$$H_0 : p_T - p_R < -0.20 \text{ or } p_T - p_R > 0.20 \text{ versus}$$
$$H_A : -0.20 \leq p_T - p_R \leq 0.20$$

where p_T = cure rate of test treatment and p_R = cure rate of reference treatment.

Rejection of the null hypothesis H_0 supports the conclusion of equivalence of the two products.

The 90% confidence interval for the difference in proportions between the test and reference treatments should be contained within -0.20 to 0.20 in order to conclude equivalence.

To establish sensitivity within the study for either a dichotomous or continuous primary endpoint, the test and reference product should both be significantly superior to the placebo. For this reason, whenever ethically permissible, a placebo third arm is included in a clinical endpoint study.

We note that the primary clinical endpoint may be, but is not necessarily the same as, that used in the studies submitted in the NDA for the RLD. For example, the registration study may evaluate the success of the trial at thirty days, while the Abbreviated New Drug Application (ANDA) study may recommend the same measure evaluated at 15 days.

Some examples of clinical endpoint studies for locally-acting products are given below.

Anti-fungal cream

In the case of topical dermatological dosage forms (such as creams, ointments and gels) the drug is intended to act locally near the site of application on the skin, and typically cannot be monitored in the systemic circulation. Clinical endpoint studies have been used as a standard approach to evaluate BE.

As an example of such a situation, the FDA draft guidance^[7] recommends a clinical endpoint study for demonstrating equivalence to the RLD for ciclopirox cream, an antifungal agent commonly used in the treatment of several dermal infections. Signs and symptoms are to be graded on a four-point scale (Table 1).

A case study of a clinical endpoint trial in an ANDA for this product can be found in Peters’ chapter, “Clinical Endpoint Bioequivalence Study,” in Yu and Li, editors.^[5] This was a multicenter, parallel group study for the treatment of tinea pedis, with 462 patients randomized in a

Table 1 Skin grading scale

Signs (physician assessment)	Symptoms (patient assessment)	Scoring Scale
Erythema	Pruritus	0 = none (complete absence of any signs or symptoms)
Scaling	Burning/Stinging	1 = mild (slight)
Fissuring/Cracking		2 = moderate (definitely present)
Masceration		3 = severe (Marked, intense)

Table 2 Per protocol results - cure rates

	Generic	RLD	Difference in rates (90% CI)	Is confidence interval within + 20%?	Pass equivalence?
Therapeutic cure	50%	46%	4% (-5%, 14%)	Yes	Yes
Clinical cure	64%	63%	1% (-9%, 10%)	Yes	N/A
Mycological cure	78%	73%	5% (-2%, 15%)	Yes	N/A

2:2:1 ratio to receive one of three treatments (ciclopirox, the RLD Loprox® or the placebo). Patients returned for clinical evaluations at Days 15, 29, and 43.

The primary endpoint of the clinical endpoint BE study was the proportion of subjects with therapeutic cure at week 6, defined as both clinical cure and mycological cure. Clinical cure was defined as a total clinical score (sum of severity scores on signs and symptoms) of 2 or less, with a score of no more than 1 for any of the clinical parameters at Visit 3 (Day 43). Mycological cure was defined as a negative culture at visit 3.

These proportions (cure rates) were calculated and compared for week 6 (Table 2).

Equivalence was evaluated by comparing rates of therapeutic cure for the per-protocol population. The data and results from this clinical endpoint study were deemed to support equivalence.

Nasal sprays

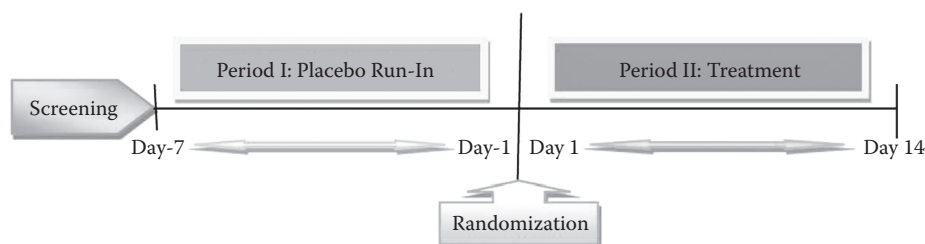
The evaluation of BE for nasal spray drug products that are intended to act locally demonstrates the need for a clinical endpoint study, and also illustrates the issues that arise in the design and analysis of such studies. These drug products are used, for example, to treat seasonal allergic rhinitis (SAR). Use of plasma PK studies alone is currently considered insufficient to establish equivalence because these products do not rely on systemic circulation for delivery to the site of action in the nasal passages.^[3] Beyond the potentially limited relevance of plasma PK, demonstrating BE in this situation through clinical endpoint studies poses particular statistical problems in the design and analysis. With the increasing number of applications for generic nasal spray drug products in the U.S., these problems are of increasing interest. This section describes current approaches to the design and analysis of clinical studies for determining equivalence for nasal spray products.

The FDA recommends conducting a clinical endpoint BE study for approval of generic nasal sprays indicated to treat SAR. According to the 2003 FDA Draft Guidance on Bioavailability and Bioequivalence Studies for Nasal Aerosols and Nasal Sprays for Local Action,^[8] the study should consist of 2 periods: an initial 7-day single-blind, placebo run-in period (Study Days -7 to -1) to establish a baseline and to identify placebo responders, followed by a 14-day treatment period (Study Days 1 to 14). Assessments are made twice daily, AM (before Noon) and PM (after Noon). A typical study design for a nasal spray study is illustrated in Fig. 1.

In general, patients who have a reflective total nasal symptom score (rTNSS; defined below) of at least 6 out of the total of 12 at screening (24 hours prior to run-in) with a history of SAR and a positive test for relevant specific allergens are allowed to enter into a single-blind placebo run-in period, during which all subjects are to receive the placebo vehicle administered daily for 7 days.

It is expected that a placebo will provide some clinical benefit, and the draft guidance recommendation allows the first four days of the baseline period for the participants to achieve their placebo symptom level. Placebo responders should be identified during the placebo run-in period. Under this assumption, the draft guidance recommendation is to use only assessments from days -3, -2, and -1 of the baseline period (plus the AM assessment from day 1) to determine baseline level.

The draft guidance also recommends that placebo responders should be excluded from the study to increase the ability to show a significant difference between active and placebo treatments, and to increase sensitivity to detect potential differences between Test and Reference products. Placebo responders should be defined a priori. Subjects are considered potential placebo responders if the average baseline reflective total nasal symptom (rTNSS) is less than six out of the maximum of 12.

**Fig. 1** Nasal spray study design

The treatment period itself should be a randomized, double-blind, placebo-controlled, parallel group trial. All subjects who qualify after the placebo run-in period are to be randomized to receive the generic nasal spray, the RLD or placebo vehicle during the treatment period. During the treatment period, the study drug is to be administered daily for 14 days. The draft guidance recommendation is to use the PM assessment on day 1 and the AM and PM assessments from days 2 to 14 of treatment period.

The total nasal symptom score (TNSS) is defined as the sum of subject-rated severity scores for runny nose, nasal congestion, nasal itching, and sneezing at each assessment time. The evaluation of TNSS is made twice daily (AM and PM) approximately 12 hours apart throughout the 7 day run-in period and the 14-day randomized treatment period. The severity score for each symptom is based on a 4-point scale:

- 0 = none
- 1 = mild
- 2 = moderate
- 3 = severe

There are two scores assessed at each time: the reflective TNSS (rTNSS), defined as the evaluation of symptoms during the period of time since the last assessment, and the instantaneous TNSS (iTNSS), which is evaluation of symptoms at that moment.

Outcome variables are based on averaged TNSS over the baseline period and over the treated period. The response of interest is the change from baseline, defined as (baseline mean TNSS) – (mean TNSS under treatment)

The primary endpoint is the change from baseline mean rTNSS to the treatment mean rTNSS.

A general linear model including the treatment and the site as factors and including the baseline as a covariate is used to analyze the efficacy (superiority of active

treatments over placebo) and equivalence. The model is based on empirical experience: in historical data from nasal spray clinical endpoint studies, we have observed a highly statistically significant linear relationship between the response variable (change from baseline in rTNSS) and baseline value. The inclusion of the baseline covariate in the statistical model results in an important reduction in the residual variance, at least for the efficacy analyses.

For the equivalence analyses, the same ANCOVA model is used to calculate adjusted least squares means for the test and reference groups. Ninety percent (90%) confidence intervals for the ratio of these least squares means are calculated using Fieller's method.^[9]

There is another twist to evaluating statistical equivalence for nasal spray products. In the assumed ANCOVA statistical model with the baseline as a covariate, the mean response (mean change-from-baseline) for a given product is a linear function of the baseline. Since, in our experience, the estimated slope is positive (which we would expect because higher baseline values allow for more of a chance for the score to decrease) we assume that the slope relating the response to the baseline is positive. That is, the mean response is higher for higher values of the baseline.

Under the assumed linear model, the difference between the mean response to the test product and the mean response to the reference product is constant and independent of the baseline. But the ratio of the test product mean to the reference product mean depends on the baseline. Therefore, for higher values of the baseline, the ratio of the test and reference means is necessarily closer to one. The possibility exists that the products could be inequivalent (have a ratio of means less than 0.80 or greater than 1.25) for a given value of the baseline, but equivalent for a higher value of the baseline.

The graph below illustrates typical behavior of CI and point estimates as a function of baseline values for the

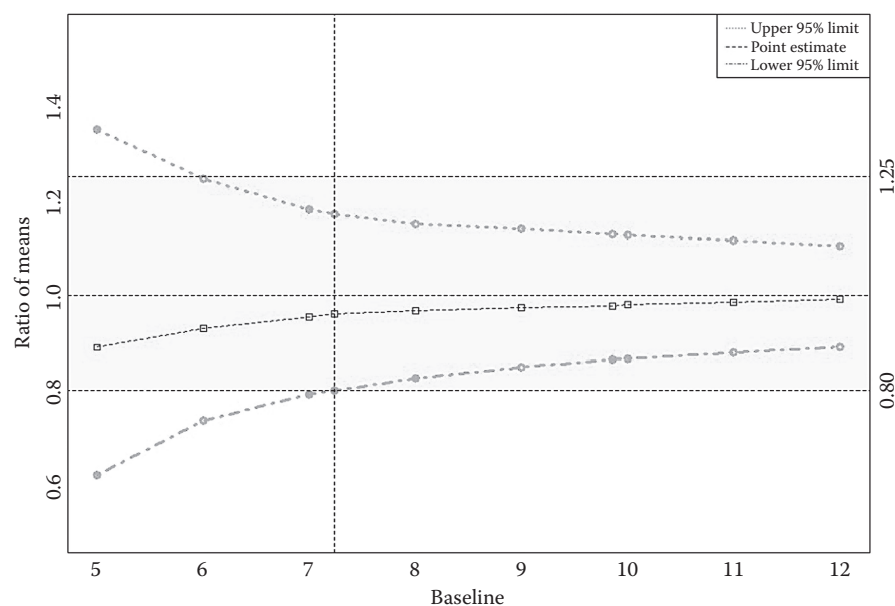


Fig. 2 90% confidence Intervals and point estimates for the ratio of the Test mean over the Reference mean (primary endpoint)

ratio of means (Fig. 2). Note that the curve connecting the point estimates approaches 1 as the baseline increases. The dashed vertical line shows the minimum value of the baseline for which the products pass the usual equivalence test for the mean change from the baseline.

In practice, we use the model to calculate the 90% CI at the average of the baseline values for the study population and check that the CI falls within the equivalence limits, in effect, checking that the study average baseline is larger than the dashed vertical line seen on a study-specific graph similar to Fig. 2.

Dry powder inhalers

The FDA issued a draft Guidance for Industry in September 2013 outlining recommendations for studies to be included in ANDA submissions to establish bioequivalence of the test and reference dry powder inhalers (DPIs) containing fluticasone propionate 100 mcg and salmeterol xinafoate 50 mcg (100/50 strength).^[10] This product is indicated to treat asthma and chronic obstructive pulmonary disease (COPD).

The draft guidance recommends conducting a clinical endpoint BE study along with in vitro studies and an in vivo PK study for approval of a generic DPI. The clinical endpoint study should consist of 2 periods, a single-blind, placebo run-in period of at least two weeks duration to wash out any pre-study corticosteroids or long acting bronchodilators and to establish baseline values of forced expiratory volume in 1 second (FEV₁); followed by a 4-week treatment period with the placebo, test or reference product. In general, subjects ≥ 12 years of age diagnosed with asthma as defined by the National Asthma Education and Prevention Program (NAEPP)^[11] at least 12 weeks prior to screening are eligible for the study.

During the study period, subjects take one inhalation of study medication twice a day (morning and evening). Assessments are made on the first day of treatment (Day 1) to determine FEV₁ at 0, 0.5, 1, 2, 3, 4, 6, 8, 10 and 12 hours post-dose as well as on the last day of a 4-week treatment to measure FEV₁ in the morning prior to the dosing of inhaled medications.

BE study endpoints are based on (i) Area under the serial FEV₁-time curve calculated from time zero to 12 hours (AUC_{0-12h}) on the first day of the treatment, and (ii) FEV₁ measured in the morning prior to the dosing of inhaled medications on the last day of a 4-week treatment.

The above two primary endpoints should be baseline adjusted through using the change from baseline. A FEV₁ baseline is defined as the average of pre-dose FEV₁ values of at least two time points measured in the morning of the first day of a 4 week treatment period. Sampling is recommended to correspond to the same time of day as used on the last day of a 4-week treatment.

The current statistical analysis approach is similar to that for a nasal spray study. Analysis of covariance (ANCOVA)

is used to analyze the superiority of active treatments over placebo and determine equivalence between active products; a general linear model is fit, including the treatment and the site as factors and the baseline as a covariate.

To ensure adequate study sensitivity, the T and R products should both be statistically superior to placebo ($p < 0.05$) in regards to the BE study primary endpoints.

Equivalence is based on the T/R ratio for the primary endpoints. The 90% CIs for the T/R ratios for the primary endpoints should fall within the limits of 80%–125%. The same ANCOVA model as the one mentioned above is used to calculate adjusted least squares means for the test and reference groups for the equivalence analyses. Ninety percent (90%) confidence intervals for the ratio of these least squares means are calculated by Fieller's method.

In addition, all adverse events (AEs) should be reported, whether or not they are considered to be related to the treatment. The report of AEs should include date of onset, description of the AE, severity, relation to study medication, action taken, outcome and date of resolution. This information is needed to determine if the incidence and severity of adverse reactions is different between the T and R products.

DISCUSSION

Clinical endpoint studies and nasal spray products

There is continued discussion in the literature about the appropriateness of the ANCOVA model in various situations, especially when we use baseline and post treatment measures in the statistical model for treatment comparisons in clinical trials. In practice, it is known that by incorporating the baseline information in the analysis, we can effectively reduce the variance of the estimated treatment effect. Use of the baseline as a covariate and analysis of difference of pre- and post-treatment measures are two ways to accomplish this. Senn^[25] has shown that the use of analysis of covariance as a means of adjusting for baseline measurements is appropriate in randomized clinical trials. He also makes the point that it is not a necessary condition for groups to be equal at baseline for ANCOVA to provide unbiased estimates of treatment effects. Even in analyses in which the change from the baseline is the response and the baseline is used as a covariate, the treatment efficacy comparisons are identical to those derived using the post-treatment measures as the response in ANCOVA.^[25,26]

This identity will not hold in equivalence tests where we are using the ratio of means as our test statistic. The change from baseline is necessarily smaller in magnitude than the post-treatment value and so which of these two quantities is used as the response value will affect the outcome of the equivalence test, and not consistently. In practice, the

change from the baseline has been established as the quantity of interest in both efficacy and equivalence testing; more so than for new drugs, precedent is followed in generic drug trials unless there is a persuasive reason to ignore it.

In determining the equivalence of nasal spray products for marketing approval, the estimate of the CI of the ratio calculated by Fieller's method provides the means for bioequivalence testing. In general, ratios of two random variables do not have good distributional properties and it is not straightforward to construct the corresponding CI. In practice, Fieller's method is often used for jointly normally distributed data to construct closed-form confidence regions of ratios (μ_T/μ_R). In our experience, the distribution of response values (change from baseline) tends to be moderately skewed. Luxburg and Franz have shown that Fieller's method behaves reasonably well even for asymmetric distributions.^[27]

There are other approaches that have been introduced in the literature in an attempt to relax the assumption of normality when estimating the confidence sets for ratios. But these methods should be used with caution and may not be appropriate for bioequivalence assessment. Luxburg and Franz introduced a geometric approach to construct confidence sets for ratios for more general classes of distributions.^[27] This approach will lead to conservative confidence sets for the ratio, with coverage greater than the nominal level and a higher chance of failure for the generic. An alternative is to consider numerical simulation methods such as the bootstrap. These methods, however, may be problematic in regulatory settings since, in addition to theoretical problems, the results may not be reproducible due to the randomness inherent in the calculations. In addition, it would be challenging to incorporate aspects of the nasal spray data, for example the positive linear relationship between the response and the baseline, in a bootstrap calculation or other nonparametric approach.

Topical dermatological products and IVPT

Clinical endpoint studies have been used as a standard approach to evaluate BE, and can play an important role in assessing the BE of locally-acting drug products, but these studies are not necessarily considered to be the most accurate, sensitive, and reproducible for evaluating BE.^[2] This is, in part, because the permeability of diseased skin to the drug can be altered depending upon the severity of the disease state, which can influence the BA of the drug in a heterogeneous way that introduces variability into the study; placebo effects can be very high for a variety of reasons; clinical endpoint evaluations may be based upon subjective assessments; patient compliance to the therapeutic regimen can be an issue, particularly because of the ability of the patient to continually observe the condition of their skin (those who suspect themselves to be in a placebo treatment group may revert to using other treatments to reduce social embarrassments associated with their

skin disease); and the natural cycles of exacerbation and resolution of the disease can lead to improvements in the endpoint over time, that may be unrelated to the effect of the treatment, and which can reduce the sensitivity of the clinical endpoint evaluations which are often performed several weeks following the initiation of treatment. Due to the inherent variability in results, the relatively high placebo effect, and the relative insensitivity of the assessment endpoint(s), these studies may require a very large number of subjects in order to be adequately powered.^[12–15]

Research is underway to explore the feasibility of potentially more efficient, sensitive and discriminating approaches to evaluate BE based upon the collective weight of evidence from extensive comparative characterizations of the RLD and generic product, supplemented by biologically relevant evidence of comparative BA from an in vitro permeation test (IVPT), which evaluates the rate and extent of drug delivery into and through excised human skin. However, in order to assess the equivalence of comparative BA based upon IVPT results, it is necessary to develop statistical methodologies that are appropriate to the inherently replicate design of these studies^[3] and to the highly nature of the resulting data.^[24]

CONCLUSION

During the decades since the Hatch-Waxman generic drug legislation was enacted, high quality generic versions of systemically acting oral drug products have become almost ubiquitous, and this has had a tremendously positive social impact in the United States. Current research is focused on developing suitable approaches to facilitate the evaluation of BE for drug products that act locally. While the scientific merits of such approaches are being investigated, it is essential to ensure that appropriate statistical methodologies are also developed to support the use of those alternative study designs to evaluate BE. It is hoped that clinical trial designs and analyses such as those described in this article as well as applications of novel statistical methodologies can aid in the efforts to make high quality generic medicines increasingly available to the American public.

ACKNOWLEDGEMENTS

We would like to thank Donald J. Schuirmann for valuable discussions and insights.

DISCLAIMER

This article reflects the views of the authors and should not be construed to represent the U.S. Food and Drug Administration's views or policies.

REFERENCES

1. Raney, S.G.; Franz, T.J.; Lehman, P.A.; Lionberger, R.; Chen M.L. Pharmacokinetics-based approaches for bioequivalence evaluation of topical dermatological drug products. *Clinical Pharmacokinetics* **2015**, *54* (11):1095–106.
2. United States Code of Federal Regulations, Title 21, § 320.24 accessible via <http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=320.24>. Last Accessed March 17, 2015.
3. Grosser, S.; Park, M.; Raney, S.G.; Rantou, E. Determining Equivalence for Generic Locally Acting Drug Products. *Statistics in Biopharmaceutical Research* **2015**, *7*(4):337–45.
4. Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* **1987**, *15*(6):657–80.
5. Lawrence, X.Y.; Li, B.V., editors. *FDA Bioequivalence Standards*. Springer; 2014 Sep 5.
6. These recommendations have been incorporated into FDA guidances for industry. See, for example, <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM191960.pdf>; <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM281452.pdf>
7. U.S. FDA Draft Guidance on Ciclopirox recommended Feb 2010. Available online at: <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM199627.pdf>. Last Accessed October 5, 2016.
8. Guidance for Industry: Bioavailability and Bioequivalence Studies for Nasal Aerosols and Nasal Sprays for Local Action (April 2003). <http://www.fda.gov/ohrms/dockets/ac/00/backgrd/3609b11.pdf>
9. Fieller EC. Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* **1954**:175–85.
10. U.S. Food and Drug Administration Draft Guidance on Fluticasone Propionate; Salmeterol Xinafoate (September 2013) <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM367643.pdf> Last Accessed October 12, 2016.
11. Guidelines for the Diagnosis and Management of Asthma: Expert Panel Report 3. National Asthma Education and Prevention Program; National Institute of Health; National Heart, Lung, and Blood Institute. 2007, Publication No. 07-4051.
12. Carroll CL, Feldman SR, Camacho FT, Manuel JC, Balkrishnan R. Adherence to topical therapy decreases during the course of an 8-week psoriasis clinical trial: commonly used methods of measuring adherence to topical therapy overestimate actual use. *Journal of the American Academy of Dermatology* **2004**, *51*(2):212–6.
13. Krejci-Manwaring J, Tusa MG, Carroll C, Camacho F, Kaur M, Carr D, Fleischer AB, Balkrishnan R, Feldman SR. Stealth monitoring of adherence to topical medication: adherence is very poor in children with atopic dermatitis. *Journal of the American Academy of Dermatology* **2007**, *56*(2):211–6.
14. Hick J, Feldman SR. Eligibility creep: a cause for placebo group improvement in controlled trials of psoriasis treatments. *Journal of the American Academy of Dermatology* **2007**, *57*(6):972–6.
15. Shah VP. Topical dermatological drug product NDAs and ANDAs—in vivo bioavailability, bioequivalence, in vitro release and associated studies. *US Dept. of Health and Human Services, Rockville* **1998**: 1–9.
16. N'Dri-Stempfer B, Navidi WC, Guy RH, Bunge AL. Improved bioequivalence assessment of topical dermatological drug products using dermatopharmacokinetics. *Pharmaceutical Research* **2009**, *26*(2):316–28.
17. Parfitt NR, Skinner MF, Bon C, Kanfer I. Bioequivalence of topical clotrimazole formulations: an improved tape stripping method. *Journal of Pharmacy & Pharmaceutical Sciences* **2011**, *14*(3):347–57.
18. Tettey-Amlalo RN, Kanfer I, Skinner MF, Benfeldt E, Verbeeck RK. Application of dermal microdialysis for the evaluation of bioequivalence of a ketoprofen topical gel. *European Journal of Pharmaceutical Sciences* **2009**, *36*(2):219–25.
19. García Ortiz P, Hansen SH, Shah VP, Sonne J, Benfeldt E. Are marketed topical metronidazole creams bioequivalent? Evaluation by in vivo microdialysis sampling and tape stripping methodology. *Skin Pharmacology and Physiology* **2010**, *24*(1):44–53.
20. Franz TJ, Lehman PA, Raney SG. Use of excised human skin to assess the bioequivalence of topical products. *Skin Pharmacology and Physiology* **2009**, *22*(5):276–86.
21. U.S. Food and Drug Administration Draft Guidance on Acyclovir; Ointment; Topical accessible via <http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm296733.pdf>. Last Accessed March 17, 2015.
22. U.S. Food and Drug Administration Guidance for Industry, Nonsterile Semisolid Dosage Forms, Scale-Up and Postapproval Changes: Chemistry, Manufacturing, and Controls; In Vitro Release Testing and In Vivo Bioequivalence Documentation, May 1997 <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM070930.pdf> Last Accessed APRIL 28, 2015.
23. Lehman PA, Raney SG, Franz TJ. Percutaneous absorption in man: in vitro-in vivo correlation. *Skin Pharmacology and Physiology* **2011**, *24*(4):224–30.
24. Haidar SH, Makhoul F, Schuirmann DJ, Hyslop T, Davit B, Conner D, Lawrence XY. Evaluation of a scaling approach for the bioequivalence of highly variable drugs. *The AAPS Journal* **2008**, *10*(3):450–4.
25. Senn S. Change from baseline and analysis of covariance revisited. *Statistics in Medicine* **2006**, *(24)*:4334–44.
26. Laird N. Further comparative analyses of pretest-posttest research designs. *The American Statistician* **1983**, *37*(4a):329–30.
27. Von Luxburg U, Franz VH. A geometric approach to confidence sets for ratios: Fieller's theorem, generalizations and bootstrap. *Statistica Sinica* **2009**:1095–117.

Clinical Inspection: Statistical Process

Fuyu Song

Center for Food and Drug Inspection, China Food and Drug Administration, Beijing, China

Minghuang Hong

Sun Yat-sen University Cancer Centre, Guangzhou, China

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

Regulatory inspection of clinical trial is necessary in order to (1) assess compliance with statutory requirements and regulatory requirement governing the conduct of clinical trials and (2) verify the accuracy and reliability of clinical trial data submitted to regulatory agencies such as the United States Food and Drug Administration (FDA) in support of research or marketing applications. This article provides an overview of clinical inspection process and issues that are commonly encountered during the conduct of clinical trials. In addition, a couple of sampling plans for clinical inspection of relatively large trials are proposed. The proposed statistical process for clinical inspection that will achieve a desired degree of inspection accuracy and reliability is useful when the resources of inspectors are limited. The proposed two-stage sampling plan can be applied to clinical inspection for multicenter and/or multinational multicenter clinical trials.

Keywords: Acceptance quality level; FDA 483 observations; Inspection accuracy and reliability; Two-stage sampling plan.

INTRODUCTION

ICH E6 guidance for *Good Clinical Practice* (GCP) defines (regulatory) inspection as the act by a regulatory authority of conducting an official review of documents, facilities, records, and any other resources that are deemed by the authority to be related to the clinical trial and that may be located at the site of the trial, at the sponsor's and/or contract research organization's (CROs) facilities, or at other establishments deemed appropriate by the regulatory authority.^[1] To assist the sponsors of clinical trials, the United States Food and Drug Administration (FDA) developed a bioresearch monitoring (BIMO) program to ensure the protection of the rights, safety, and welfare of human research subjects involved in FDA-regulated clinical trials. The purpose is not only to assess compliance with statutory requirements and FDA regulations governing the conduct of clinical trials but also to verify the accuracy and reliability of clinical trial data submitted to the FDA in support of research or marketing applications. For this purpose, in 2010, the FDA published a guidance "Information Sheet Guidance for IRBs, Clinical Investigators, and Sponsors—FDA Inspections of Clinical Investigators."^[2] The FDA conducts clinical investigator inspections of clinical investigators under the FDA's BIMO program is to determine if the clinical investigators are conducting

clinical studies in compliance with applicable statutory and regulatory requirements.

Regulatory inspections are important to evaluate the integrity of the data submitted to health authorities, the presence of infrastructure to conduct clinical research, measures implemented to protect patient's interest and safety, adequacy of site/sponsor quality systems, and verification of compliance with the principles of ICH-GCP as well as local regulations.^[3] In practice, although inspections can occur at any time, they generally occur after submission of data for marketing approval of an investigational drug. The inspectors generally conduct a thorough review of site data including all aspects of ICH-GCP, site infrastructure, and quality control procedures and system before inspections. Findings are discussed during the closeout meeting and a detailed inspection report issued afterward. The sponsors usually respond within 15–30 days with effective corrective and preventive action (CAPA) plan. In clinical trials, protocol noncompliance, inadequate/inaccurate records, inadequate drug accountability, informed consent issues, and adverse event reporting were some of the most common deficiencies observed during recent FDA inspections.^[3]

Clinical inspection of a relatively large clinical trial or multicenter or multinational multicenter clinical trial could be labor intensive and time consuming. It may not be feasible if there are limited resources available. Alternatively, one

may consider adopting a statistical process by inspecting a randomly selected proportion of subjects such as informed consents and case report forms (CRFs) for achieving a pre-specified inspection accuracy and reliability. A valid statistical process consists of sampling plan, acceptance criteria, and testing procedure. Under the statistical process for clinical inspection, the inspector will prespecify an acceptance quality level (AQL) or a desired inspection accuracy and/or reliability and then conduct inspection according to the sampling plan for achieving the desired inspection accuracy and reliability.

In this article, process and some general principles for clinical trial inspection are briefly outlined. FDA Form 483 observations and actions that need to be taken are discussed. Examples of top 483 observations issued by the FDA between 2006 and 2014 are also included. Statistical process including sampling plan, acceptance criteria, and testing procedure for clinical trial inspection is described and discussed. Some concluding remarks are given in the last section of this article.

INSPECTION PROCESS

Scope of Inspection

For clinical inspection, all site personnel who have a significant role in the conduct of clinical trial should be present during the inspection. These personnel usually include, but are not limited to, principal investigator (PI), sub-investigators (SIs) or co-PI, study coordinator (SC), laboratory personnel, and pharmacist. In many cases, trial sponsor representatives may also be present at the site to facilitate the process. It, however, should be noted that the sponsor representatives should not make any attempts to interfere site inspection.

Clinical inspection generally involves a facility tour for evaluation of site infrastructure, site's standard operation procedures (SOPs), and process of clinical trial operation including data collection/entry, data verification/validation, and data management for quality assurance to make sure the conduct of clinical trials is in compliance with GCP. Clinical inspection may include inspection of laboratories if trial-related laboratory tests are done locally. Laboratory inspection includes review of laboratory equipment, process, calibration, quality assurance/control method, and related documents for Good Laboratory Practice such as SOPs for training and handling of study specific laboratory testing procedures; archival storage of test samples; and stock checks, labeling, and accountability logs.

Note that clinical inspection may also include insurance coverage for subjects and a disaster plan in case of an emergency such as fire or flood, among others.

Before Inspection

Before inspection, it is suggested that a designated person (usually the research coordinator for the study) be identified

to serve as institutional monitor who will escort and oversee the inspection. The identified individual should complete the FDA audit checklist and identify records and/or documents that the FDA inspectors are likely to audit. The records and/or documents that the FDA are likely to inspect include, but are not limited to, (1) all informed consents forms and subject enrollment and screening log, and (2) all or selected CRFs or case record books (CRBs) and all supportive source documentation. All or selected CRFs (or CRBs) depends upon sampling plan adopted for the inspection, which will be discussed at later section of this article.

The inspection starts with an opening meeting. The inspector will present a *Notice of Inspection* (482) to the PI. This notice authorizes the inspection, and its presentation officially begins the inspection. The inspector will explain the intended purpose and scope of the inspection, then ask the PI to summarize the study. The PI/site representative will then make a general presentation about the site and study, and address initial questions from inspectors. It is an appropriate time to discuss the inspection schedule, availability of site personnel, logistic arrangements, etc., to ensure smooth conduct of inspection with minimum interference in routine site schedule. If there were important deviations from GCP, it might be wise to be up front with the inspectors and explain why they happened and how measures were taken to prevent them from happening again.

Note that the FDA investigator must not be permitted free access to areas where files are kept, and the escort serves as an institutional monitor as well as guide and general study contact person.

The Inspection

The inspection focus is usually on the investigator role in the study, delegation of duties (with documentation), qualification of site staff, and ethical issues (e.g., consent process and ethic committee review). The inspection also involves team member interviews—PI, SI, SC, lab personnel, and others. Sponsor/monitor may also be interviewed to evaluate their supervision, protocol knowledge, and adequacy to identify and resolve issues.

The inspector usually starts with the general study materials (or master files) including the regulatory documents/binders, then all signed informed consent forms, followed by a sampling of specific patient records. The data review generally includes subject history, informed consent, subject eligibility, investigational drug administration, compliance, safety reporting, compliance to study procedures, and other study-specific issues. There is a frequent interaction between inspectors and site personnel to clarify points and issues, and to provide supporting documents.

Only documents specifically requested by the inspector shall be provided for review. The escort may need to obtain patient records from the hospital or clinic records to supplement or corroborate the research records. The

inspector must never have access to any site records not specifically provided by the host. Standard procedure is for the inspector to request files for review.

Note that the inspector is not entitled to review or copy financial, personnel (except for training/qualification records), and internal audits (section 704(a) FDC Act).

Exit Interview

The FDA will usually hold an exit interview (or closeout meeting) at the conclusion of the inspection. This also provides opportunity for site personnel to clarify any points concerning the findings or to provide better context. The designated institutional monitor, PI, and other individuals as appropriate should be notified of the time and place and expected to attend. Note that the sponsor may attend the closeout meeting only if the PI agrees. At the exit interview, the PI will usually seek to correct any errors in the findings. Both the FDA and the PI will make sure everything is clear and understood. The institutional monitor should take noted on observations, comments, and commitments discussed at the closeout meeting.

During this exchange, if serious deficiencies have been found during the inspection, an FDA Form 483 (Inspectional Observations) will follow from the regional office, listing the deficiencies. If no deficiencies are found, or the inspector has comments that she or he believes are not serious enough to warrant a 483, no form will be issued. At the closeout meeting, FDA Form 483s are discussed with a company’s management at the conclusion of the inspection. Each observation is read and discussed so that there is a full understanding of what the observations are and what they mean.

FDA 483 OBSERVATIONS AND ACTIONS

FDA 483 Observations

The FDA is authorized to perform inspections under the Federal Food, Drug, and Cosmetic Act (FD&C Act), Sec. 704

(21 USC §374). FDA Form 483 (Inspectional Observations) is a form used by the FDA to document and communicate concerns discovered during these inspections (see also [4]). The FDA Form 483 notifies the company’s management of objectionable conditions. At the conclusion of an inspection, the FDA Form 483 is presented and discussed with the company’s senior management. Companies are encouraged to respond to the FDA Form 483 in writing with their corrective action plan and then implement that plan expeditiously. Note that the FDA inspector will file an establishment inspection report within approximately 30 days.

The FDA Form 483 is a report that does not include all observations of questionable or unknown significance at the time of the inspection. There may be other objectionable conditions that exist at the sponsor that are not cited on the FDA Form 483. FDA investigators are instructed to note only what they saw during the course of the inspection. Companies are responsible for taking corrective action to address the cited objectionable conditions and any related non-cited objectionable conditions that might exist.

It should also be noted that the FDA Form 483 does not constitute a final agency determination of whether any condition is in violation of the FD&C Act or any of its relevant regulations. The FDA Form 483 is considered, along with a written report called an Establishment Inspection Report, all evidence or documentation collected on site and any responses made by the company. The agency considers all of this information and then determines what further action, if any, is appropriate to protect public health.

As an example, Tables 1 and 2 give a list of top 10 FDA Form 483 observations for pharmaceuticals and devices, respectively. The top three 483 observations for pharmaceuticals are (1) procedures not in writing, fully followed (169); (2) investigations of discrepancies and failures (119); and (3) absence of written procedures (116). The top three 483 observations for devices are (1) lack of or inadequate procedures (372), (2) lack of or inadequate complaint procedure (259), and (3) lack of written MDR procedures (140). On the other hand, Fig. 1 plots total 483s issued by

Table 1 List of top 10 FDA 483 observations for pharmaceuticals

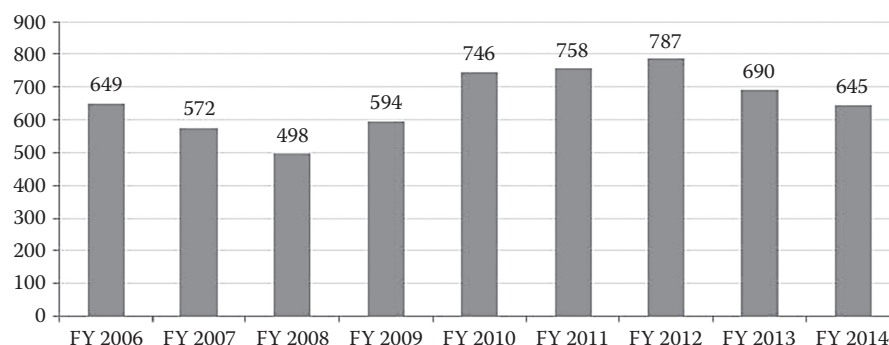
Top 10	CFR reference	Frequency	Description
1	21 CFR 211.22(d)	169	Procedures not in writing, fully followed
2	21 CFR 211.192	119	Investigations of discrepancies and failures
3	21 CFR 211.100(a)	116	Absence of written procedures
4	21 CFR 211.160(b)	115	Scientifically sound laboratory controls
5	21 CFR 211.110(a)	89	Control procedures to monitor and validate performance
6	21 CFR 211.67(b)	73	Written procedures not established/followed
7	21 CFR 211.68(a)	69	Calibration/inspection/checking not done
8	21 CFR 211.25(a)	65	Training-operations, GMPs, written procedures
9	21 CFR 211.67(a)	65	Cleaning/sanitizing/maintenance
10	21 CFR 211.100(b)	64	SOPs not followed/documented

Source: <http://compliance-insight.com/fda-483-warning-letters/top-ten-reasons-for-fda-483/>.

Table 2 List of top 10 FDA 483 observations for devices

Top 10	CFR reference	Frequency	Description
1	21 CFR 820.100(a)	372	Lack of or inadequate procedures
2	21 CFR 820.198(a)	259	Lack of or inadequate complaint procedures
3	21 CFR 803.17	140	Lack of written MDR procedures
4	21 CFR 820.50	126	Purchasing controls, lack of or inadequate procedures
5	21 CFR 820.100(b)	115	Documentation
6	21 CFR 820.90(a)	110	Nonconforming product control procedures
7	21 CFR 820.75(a)	102	Lack of or inadequate process validation
8	21 CFR 820.30(i)	101	Design changes—lack of or inadequate procedures
9	21 CFR 820.198(c)	96	Investigation of device failures
10	21 CFR 820.22	91	Quality audits—lack of or inadequate procedures

Source: <http://compliance-insight.com/fda-483-warning-letters/top-ten-reasons-for-fda-483/>.

**Fig. 1** Total 483s issued by FDA between 2006 and 2014

Source: <http://www.raps.org/Regulatory-Focus/News/2014/12/08/20928/The-Top-15-Pharmaceutical-Deficiencies-Cited-by-FDA-in-2014/>.^[5]

the FDA between 2006 and 2014. As it can be seen, the number of 473s issued (645) marked the second consecutive year thw FDA has issued fewer pharmaceutical 483s than the year prior, and the 2014 total (645) is 19% less than its 2012 high water mark (787 inspections).

Actions

The PI or designated personnel is responsible for drafting a written response to the FDA. The written response should be able to address each particular observation or finding point by point and delineate corrective actions, including justification of why the proposed response will remediate the issue and a realistic timeline for correction. The PI should also determine if a finding was an oversight/one-time occurrence or systemic, where a change of procedure is indicated. If the PI disagrees with an observation, respond factually, providing clear and verifiable evidence. It should be noted that the response should be sent within 2 weeks and a signed copy should be kept in the master file of the clinical trial.

In practice, the following steps for formulating a response to dealing with each FDA 483 observation are often considered (see Table 3).

Note that steps 1 through 7 are directly related to formulating a response to the observation. It is critical to

understand that steps 6 and 7 directly relate to the response letter in that these actions determine the time frame in which actions can be accomplished.

STATISTICAL PROCEDURE

For clinical trial inspection, the inspector will conduct inspection on site. The scope of inspection includes facility/site infrastructure, SOPs, and process of clinical trial operation including informed consents and clinical data capture. The primary focus of clinical inspection is on quality and validity of clinical data captured by CRFs or CRBs. For a large clinical trial, a complete clinical inspection could be very time consuming, especially when there are limited resources available. In this case, some sampling plans with prespecified acceptance criteria for clinical trial inspection could be useful.

Hypergeometric Distribution

Let N be the population size (i.e., number of CRBs or size of the clinical trial) and K be the CRBs with deficiencies (or failures). Then $P = K/N$ is the true proportion of failures (deficiencies). Also let $X = (X_1, X_2, \dots, X_n)$ be a

Table 3 An approach for formulating a response to each observation

Step	Description	Comment
1	Evaluate the current state of compliance in light of the audit observation	If possible, indicate what is compliant
2	Identify the root cause of the issue as appropriate	Understand how the firm will resolve the issue systemically
3	Review prior commitments	Repeated observations also greatly increase the likelihood of further regulatory actions
4	Identify the root cause	This is a fundamental principle in the FDA's view of systemic audits
5	Relate each observation to the appropriate quality system	Understand the area of impact of the issue in relation to the functions impacted by that system
6	Develop a corrective action plan	Around the entire scope of the issue(s) noted
7	Secure the required resources	Need to complete the job in the time frame indicated
8	Verify that responsibilities are assigned to key people	Make them accountable

random sample drawn from the population of size N without replacement. Thus, $S = \sum_{i=1}^n X_i$ follows a hypergeometric distribution with the probability density as follows:

$$P\{S = r\} = \frac{\binom{NP}{r} \binom{N-NP}{n-r}}{\binom{N}{n}} \quad (1)$$

The mean of S is nK/N and the variance is

$$n \frac{K}{N} \frac{(N-K)}{N} \left(\frac{N-n}{n-1} \right) = nP(1-P) \left(\frac{N-n}{n-1} \right)$$

Note that if N is extremely large, it is possible that even without replacement, the results may approximate the binomial distribution (sampling with replacement). Alternatively, we may consider normal approximation when N is sufficiently large. Denote $p = S/n$. Then, the corresponding z is given by

$$z = \frac{p - P}{s}$$

where

$$s = \sqrt{\frac{P(1-P) \frac{N-n}{N-1}}{n}}$$

Denote by δ the accuracy of the proportion P . Thus, for achieving this accuracy with certain probability (say $1 - \alpha$), the sample size required satisfies the following inequality:

$$P\{|p - P| \leq \delta\} \geq 1 - \alpha$$

As a result, we have

$$n \geq \frac{N^2 z_{\alpha/2}^2 P(1-P)}{\delta^2 (N-1) + z_{\alpha/2}^2 P(1-P)} \quad (2)$$

where $z_{\alpha/2}^2$ is the $\alpha/2$ quantile of standard normal distribution.

Single-Stage Sampling Plan

Let P_0 be the acceptable quality level (AQL) that is defined as the minimum level of quality or the maximum fraction of error rates that is acceptable to the inspector. In this case, if $P < P_0$, we claim that the study site meets AQL. Thus, for a clinical trial inspection, the following one-sided hypothesis is commonly considered for testing whether the study site meets a prespecified AQL P_0 :

$$H_0 : P = P_0 \quad \text{vs.} \quad H_a : P = P_1 \quad (3)$$

where $P_1 = P_0 - \delta$, $0 < \delta < P_0$, which is referred to as a desirable AQL. In practice, we would reject the null hypothesis and conclude the alternative hypothesis that the study site has met desirable AQL with an appropriate choice of δ . To test hypotheses in Eq. 3, various tests such as exact test or normal-based method can be considered.

Exact test—Given sample size n , for any integer C that $0 \leq C \leq NP$, we have

$$P\{S \leq C | P\} = \sum_{s=0}^C \frac{\binom{NP}{s} \binom{N-NP}{n-s}}{\binom{N}{n}}.$$

Given α , the level of significance, we choose the largest C that satisfies

$$P\{S \leq C | P = P_0\} \leq \alpha$$

If we denote this C by $C_{n, \alpha}$, power is given by

$$P\{S \leq C_{n, \alpha} | P = P_1 = P_0 - \delta\}$$

Thus, for a given β , we can choose the smallest n satisfying

$$P\{S \leq C_{n,\alpha} \mid P = P_1 = P_0 - \delta\} \geq 1 - \beta$$

Although this test is based on the exact hypergeometric distribution, there exists no closed form for the smallest sample size n that satisfies the error rate constraints. In practice, the interval searching for $C_{n,\alpha}$ can be obtained from $\max\{0, n - (N - NP_0)\}$ to $\min\{NP_0, n\}$.

Normality-based test—Under the null hypothesis of Eq. 3, the following test is useful:

$$T_p = \sqrt{\frac{n}{P(1-P)^{\frac{N-n}{N-1}}}} (p - P) \quad (4)$$

where $p = S/n$. For large sample size, T_p is approximately normal distributed. Thus, we have

$$P\{T_{p_0} \leq -z_\alpha \mid P = P_0\} \approx \alpha$$

where z_α is the α th quantile of standard normal distribution. Thus, we take the set of $\{T_{p_0} \leq -z_\alpha\}$ as the rejection set. To achieve the power of $1 - \beta$, we choose the smallest n satisfying

$$P\{T_{p_0} \leq -z_\alpha \mid P = P_1\} \geq 1 - \beta$$

Under the approximate normality, we have

$$\begin{aligned} & P\{T_{p_0} \leq -z_\alpha \mid P = P_1\} \\ &= P\left\{\sqrt{\frac{n}{P_1(1-P_1)^{\frac{N-n}{N-1}}}} (p - P_1) \leq -z_\alpha \sqrt{\frac{P_0(1-P_0)}{P_1(1-P_1)}} \right. \\ &\quad \left. + \delta \sqrt{\frac{n}{P_1(1-P_1)^{\frac{N-n}{N-1}}}} \mid P = P_1 = P_0 - \delta\right\} \\ &\approx \Phi\left(-z_\alpha \sqrt{\frac{P_0(1-P_0)}{P_1(1-P_1)}} + \delta \sqrt{\frac{n}{P_1(1-P_1)^{\frac{N-n}{N-1}}}}\right) \end{aligned}$$

Thus, we have

$$n = \frac{N}{\frac{(N-1)\delta^2}{P_1(1-P_1)\left(z_\beta + z_\alpha \sqrt{\frac{P_0(1-P_0)}{P_1(1-P_1)}}\right)^2 + 1}} \quad (5)$$

where z_β is the β th quantile of standard normal distribution.

As a result, the single-stage sampling plan for clinical inspection can be summarized as the following steps:

Step 1: Set up the AQL P_0 (or $P_0 - \delta$).

Step 2: Determine the sample size required n (or the number of acceptable deficiencies c , where $c = nP_0$ for achieving a desired statistical inference).

Step 3: Conduct clinical inspection of the selected n subjects (or CRBs).

Step 4: Determine whether the study site has passed or failed the clinical inspection based on whether they have met the prespecified AQL.

Two-Stage Sampling Plan

The idea of single-stage sampling plan can be extended to a two-stage sampling plan. In other words, we randomly select n_1 subjects at the first stage. If the n_1 subjects meet a prespecified AQL at the first stage, we claim that the site has passed clinical inspection; otherwise, proceed to the second stage of inspection. At the second stage, additional n_2 subjects will be randomly selected from the population. If the $n_1 + n_2$ subjects meet a prespecified AQL at the second stage, we conclude that the site has passed clinical inspection; otherwise, the site has failed clinical inspection.

Let $(X_1, X_2, \dots, X_{n_1})$ be a random sample at stage 1. Denote $S_1 = \sum_{i=1}^{n_1} X_i$. We then reject the null hypothesis of Eq. 3 and stop clinical inspection if $S_1 \leq C_1$. On the other hand, we will accept the null hypothesis and stop clinical inspection if $S_1 > C_2$, where C_1 and C_2 are some constant integers, where $C_1 < C_2$. If $C_1 < S_1 \leq C_2$, then proceed to stage 2 and randomly draw additional n_2 subjects, i.e., $(X_{n_1+1}, X_{n_1+2}, \dots, X_{n_1+n_2})$. Let $S_2 = \sum_{i=n_1+1}^{n_1+n_2} X_i$. We then reject the null hypothesis when $S_1 + S_2 \leq C_2$.

For this two-stage sampling plan, the overall sample size is n_1 if the null hypothesis is rejected at stage 1; otherwise, the overall sample size is $n_1 + n_2$. As a result, the expected sample size of this two-stage sampling plan is given by

$$E(n \mid P) = n_1 + n_2(1 - PET_p) \quad (6)$$

where PET_p is the probability of early termination of the two-stage sampling plan when the true value of the proportion is P . Now, given n_1, n_2, C_1, C_2 , and P , we have

$$\begin{aligned} PET_p &= H_1(\min\{n_1, NP\} \mid P, n_1) \\ &\quad - H_1(\min\{C_2, n_1, NP\} \mid P, n_1) + H_1(C_1 \mid P, n_1) \end{aligned} \quad (7)$$

and

$$\begin{aligned} P(\text{reject} \mid P, n_1, n_2, C_1, C_2) &= H_1(C_1 \mid P, n_1) \\ &\quad + \sum_{s_1=C_1+1}^{\min\{n_1, C_2, NP\}} h(s_1 \mid P, n_1) H_2(C_2 \mid P, n_1, n_2, s_1) \end{aligned} \quad (8)$$

where

$$\begin{aligned}
 h(s_1 | P, n_1) &= \frac{\binom{NP}{s_1} \binom{N-NP}{n_1-s_1}}{\binom{N}{n_1}} \\
 H_1(C_1 | P, n_1) &= \sum_{s_1=\max\{0, n_1-(N-NP)\}}^{\min\{C_1, n_1, P\}} h(s_1 | P, n_1) \\
 H_2(C_2 | P, n_1, n_2, s_1) &= \sum_{i=\max\{0, (n_1+n_2)-(N-NP)-s_1\}}^{\min\{C_2, NP, (n_1+n_2)\}-s_1} \frac{\binom{NP-s_1}{i} \binom{N-NP-(n_1-s_1)}{n_2-i}}{\binom{N-n_1}{n_2}} \quad (9)
 \end{aligned}$$

Given α and β , since there exist no closed forms for n_1 , n_2 , C_1 , and C_2 that satisfy the following constraints:

$$\begin{aligned}
 P\{\text{reject} | P_0, n_1, n_2, C_1, C_2\} &\leq \alpha \\
 \text{and} \\
 P\{\text{reject} | P_1, n_1, n_2, C_1, C_2\} &\geq 1 - \beta
 \end{aligned} \quad (10)$$

where $0 \leq C_1 + 1 \leq C_2$, $n_1 - (N - NP_0) \leq C_1 \leq \min\{NP_0, n_1\}$ and $(n_1 + n_2) - (N - NP_0) \leq C_2 \leq \min\{NP_0, n_1 + n_2\}$, an optimal sampling plan (n_1, n_2, C_1, C_2) can be obtained by an exhaustive search for all possible sampling plans.

As a result, the two-stage sampling plan for clinical inspection can be summarized as the following steps:

- Step 1: Given the size of a site subject to clinical inspection, N , set up the AQLs for the two stages, i.e., prespecified C_1 and C_2 that satisfy Eq. 10.
- Step 2: Determine the sample sizes required n_1 and n_2 according to C_1 and C_2 .
- Step 3: Conduct clinical inspection. At the first stage, n_1 subjects (e.g., CRBs) are randomly selected for inspection. If the error (or deficiency) rate is less than C_1/n_1 , then stop the inspection and conclude that the site has passed the inspection; otherwise, proceed to the next stage.
- Step 4: At the second stage, additional n_2 subjects will be randomly selected from the remaining $N - n_1$ subjects for inspection. If the error

(or deficiency) rate is less than $(C_1 + C_2)/(n_1 + n_2)$, the site is considered passing the clinical inspection; otherwise, we conclude that the site has failed clinical inspection.

Note that similarly, the concept of two-stage sampling plan can be extended to a multiple-stage (more than two stages) design for multinational, multicenter clinical trials, especially when there are limited resources available for inspection.

CONCLUDING REMARKS

Clinical inspection is necessary to ensure quality and integrity of the clinical trial conducted for the evaluation of safety and efficacy of a test treatment under investigation. Clinical inspection could be labor intensive and/or time consuming especially when the clinical trial to be inspected is relatively large and involving a number of study sites. In this case, it is suggested that a valid statistical process including sampling plan, acceptance criteria, and testing procedure be adopted for clinical inspection especially when there are limited resources available.

In this article, a simple single-stage sampling plan and a two-stage sampling plan are proposed for clinical inspection. Under the proposed sampling plans, the inspector can prespecify an AQL or a desired inspection accuracy/reliability and then conduct inspection according to the sampling plan for achieving the desired inspection accuracy and reliability. The proposed sampling plans can be applied to clinical inspection for multicenter and/or multinational multicenter clinical trials.

REFERENCES

1. ICH E6. *Guidance for Industry E6 Good Clinical Practice: Consolidated Guidance*; Geneva, Switzerland, 1996.
2. FDA. *Information Sheet Guidance for IRBs, Clinical Inspectors, and Sponsors—FDA Inspections of Clinical Inspectors*. The United States Food and Drug Administration: Rockville, MD, June, 2010.
3. Marwah, R.; Van de Voorde, K.; Parchman, J. Good clinical practice regulatory inspections: Lessons for Indian investigator sites. *Perspect. Clin. Res.* **2010**, *1* (4), 151–155. doi:10.4103/2229-3485.71776.
4. Goebel, P.W.; Whalen, M.D.; Khin-Maung-Gyi, F. *Applied Clinical Trials Online*; “What a Form 483 Really Means”, September 1, 2001.
5. FDA. The Top 15 Pharmaceutical Deficiencies Cited by FDA in 2014. <http://www.raps.org/Regulatory-Focus/News/2014/12/08/20928/The-Top-15-Pharmaceutical-Deficiencies-Cited-by-FDA-in-2014/>, 2014.

Clinical Pharmacology

Brian Smith

Eli Lilly and Company, Indianapolis, Indiana, U.S.A.

Abstract

Clinical pharmacology is, essentially, the field of medicine whose main interest is the effects of drugs in humans. This entry is focused on breaking down the various areas of clinical pharmacology, which is a field that encompasses a multitude of disciplines.

INTRODUCTION

This area of study is fairly broad. The following excerpt from the introduction to *The Pharmacological Basis of Therapeutics*^[1] gives a glimpse at this field of study: “The clinician is understandably interested mainly in the effects of drugs in man. This emphasis on *clinical pharmacology* is justified, because the effects of drugs are often characterized by significant interspecies variation and because they may be further modified by disease.” From this quote one may define *clinical pharmacology* as the field of medicine whose main interest is the effects of drugs in humans. Within this definition, the field encompasses a multitude of disciplines.

OVERVIEW

A different approach for getting at the heart of clinical pharmacology comes from asking what would be the core curriculum for a training program in this field. In a 1997 issue of *Clinical Pharmacology and Therapeutics*,^[2] knowledge of the following is proposed as a core curriculum in clinical pharmacology and therapeutics for primary care residents:

1. Clinical pharmacokinetics (including drug absorption, distribution, redistribution, clearance and half-life) and its application to patient-based clinical scenarios;
2. Therapeutic drug monitoring;
3. Adverse drug reactions;
4. Drug allergy;
5. Drug–drug interactions;
6. Pharmacokinetics;
7. Prescribing for elderly patients;
8. Prescribing for pediatric patients;
9. Prescribing for pregnant and nursing women;
10. Prescribing for patients with renal disease;
11. Toxicology;

12. Drug regulations applicable to primary care;
13. Bioethical issues and principles related to prescribing;
14. Therapeutic management of common clinical problems, including both drug and nondrug therapies.

This list helps identify the areas that need to be studied by clinical pharmacologists before a new chemical entity can be approved as a drug. Clinical pharmacology is often concerned with the safety, pharmacokinetics, and pharmacodynamics of a new compound. Clinical pharmacology studies are not synonymous with phase I studies. Phase I studies are usually clinical pharmacology studies; however, clinical pharmacology information can often be obtained from studies in other phases of development.

To further classify clinical pharmacology as a discipline, it would be useful to consider the types of clinical pharmacology studies that are performed when a new chemical entity is selected for development. The following list of studies is not meant to be inclusive or a checklist of studies that need to be carried out. It is meant only to give a flavor of clinical pharmacology research.

First Human Dose (Single-Dose Escalation) Studies

These studies are usually open-label, nonrandomized studies performed on healthy volunteers. (If a compound is known to be highly toxic and it is being developed for a life-threatening disease, then often all clinical pharmacology studies are carried out on patients who are not healthy volunteers.) A “no-effect dose,” in units of drug per body weight, is determined from toxicology studies. Then a fraction—say, one-tenth—of this dose becomes the highest that will be allowed in the clinical study. The first dose, which is much less than the maximal dose, is administered to subjects. If no adverse effect is observed, on a later date an increased dose is administered to the subject(s). The increased dose is typically restricted according to some accepted ethics-based algorithm. This is continued for each subject until either a drug-related adverse

event occurs, a drug-related abnormal laboratory value is observed, or the maximal dose is achieved. The primary objective of these studies is safety; however, blood samples and sometimes urine samples are collected to obtain a first glimpse at the compound's pharmacokinetic properties. If a surrogate dynamic or efficacy marker is known, then appropriate tests are carried out to obtain a first glimpse of whether the compound is doing what it is supposed to do.

Multiple-Dose Escalation Studies

These studies are usually open-label, nonrandomized studies performed on healthy volunteers. A maximal dose is chosen from the first study. Then a lower dose is given once or twice a day to each subject for an extended time, which should be sufficient to achieve steady-state systemic concentrations. If no effect is observed, then an increased dose is administered (usually after a washout period). This is continued for each subject until either a drug-related adverse event occurs, a drug-related elevated laboratory value is observed, or the maximal dose is achieved. Again, safety is the primary reason for the study; however, pharmacokinetic and pharmacodynamic information is often obtained.

Fed/Fasted Studies

These studies seek to determine if the pharmacokinetic properties of the compound are affected by food. They are usually a randomized two-period crossover design using healthy volunteers. (If a compound has a long half-life, than parallel designed trials may be more appropriate.)

Drug Interaction Studies

The goal of these studies is to determine if the pharmacokinetics and the pharmacodynamic properties of the compound are affected by another drug. Similarly, they determine if the pharmacokinetic and pharmacodynamic properties of the other drug are affected by the new compound. Randomized three-period crossover designs with healthy volunteers are the norm.

Bioequivalence Studies

These studies seek to determine if the bioavailability of two formulations of the same compound are similar. Again, randomized two-period crossover designs on healthy volunteers are the norm.

Special Population Studies

The goal of these studies is to determine if the pharmacokinetic properties and sometimes pharmacodynamic properties are the same for different groups of people. Population traits studied include gender, race, and disease

types. Special population studies have interpretation problems similar to those of observational studies. This is because, for instance, one cannot randomize an individual to a gender classification.

Absolute Bioavailability Studies

These studies compare the extent of absorption of an oral formulation to the intravenous formulation of a compound. Randomized two-period crossover designs with healthy volunteers are the norm.

Dose Proportionality Studies

A range of doses, on separate occasions, is given to each individual to determine if the compound exhibits "linear" pharmacokinetic properties. Some pharmacokinetic variables (for example, the concentration maximum and the area under the concentration time curve) are related to dose in a proportional manner if the compound exhibits "linear" pharmacokinetic properties. Other pharmacokinetic variables (for example, the plasma clearance and the apparent volume of distribution) are independent of dose if the compound exhibits "linear" pharmacokinetic properties. If the rate of elimination and absorption of a compound into the compartments of a pharmacokinetic model are proportional to the amount in the source compartment, then it is said that the compound exhibits linear pharmacokinetic properties. *Linear*, in this context, is not to be confused with linear models in a statistical context.

Population Pharmacokinetic and Pharmacodynamic Work

In phase II and phase III studies, sparse blood samples are collected from many individuals. With the help of nonlinear mixed-effect modeling, the pharmacokinetic and pharmacodynamic properties of a compound can be explored. One thing often carried out is to relate a pharmacokinetic or pharmacodynamic variable to covariates such as age, disease status, and gender. This work has the same kind of interpretation problems as observational studies.

STATISTICAL DESIGN, METHODS, AND ANALYSIS

Although clinical pharmacology studies are broad in range and scope, many of these studies have enough similarities for common statistical themes.

The designs are often sequential or randomized crossover designs. Period and crossover effects may come into play. It should be pointed out, however, that carryover effects should be easy to eliminate in pharmacokinetic studies. If one knows the half-life of a drug, then one should be able to forecast the amount of time necessary to clear

almost all of the drug, known as the *washout period*. It is not as easy to eliminate carryover effects if the outcome is pharmacodynamic or safety in nature. Period effects, such as carryover effects, are less likely with pharmacokinetic endpoints, but may pose a problem with pharmacodynamic and safety endpoints.

It is not uncommon for the nonrandomized studies discussed to have multiple treatments per subject and repeated measures for each treatment-by-subject combination. The analysis of that data often involves complex mixed-effect models.

The sample sizes are typically small. Sometimes they are chosen for a formal statistical goal; but often they are chosen for nonstatistical reasons, such as cost, or ethical reasons. The small sample size often precludes extensive statistical inferences, particularly where nonsignificant differences exist.

The work is often exploratory in nature. This is especially true for phase I studies. Any clues that could expedite drug development are examined. Undoubtedly, out of the many relationships examined, some will show “statistical significance.” These findings should only be considered hints and not conclusive. Also, the traditional 0.05 significance level should be used judiciously, because of the small sample size. A *p*-value of, say, 0.2 may sometimes be giving evidence of an important effect and should not be ignored. Lastly, the goal of a phase I study often is not hypothesis testing but simply the estimation of a pharmacokinetic or pharmacodynamic variable.

In most clinical pharmacology studies, the null hypothesis of no differences, such as “no safety concern” or “no difference in pharmacokinetic properties,” is the working hypothesis. Rejecting the no-difference hypothesis may be interpreted negatively for the compound. Yet it has been mentioned that these studies are often not formally sized. This may encourage undersizing of the studies. There are three options that could be used to fix this undersizing problem.

The first option is to do a proper power analysis, in which the minimal clinically significant effect is elicited from the clinical pharmacologist or pharmacokineticist.

One problem with this is that if the drug is in early development, notions of variability may be completely unknown. Another problem is that often there will be multiple significance tests performed. Thus there is not always an obvious variable to select for sizing purposes.

The second option can be called a field goal approach: Define an interval of clinically irrelevant differences. Construct a confidence interval for the variable. If the confidence interval (football) falls inside the interval (goalposts), then the hypothesis of no difference holds. This is the type of analysis that is currently performed with bioequivalence studies.

The third approach is similar to the field goal approach but depends on the Bayesian paradigm: Use a noninformative prior for the analysis. Set up an interval of clinically irrelevant differences. Calculate the probability that your statistic falls in that interval. If that probability is sufficiently high (higher than a prespecified threshold), then declare that no difference exists.

The difficulty with all three of these methods is inducing the clinical pharmacologist or the pharmacokineticist to define clinically irrelevant differences a priori.

RECENT DEVELOPMENTS

Two developments in the field are the recently proposed individual and population bioequivalence methods and the recent expectations that population pharmacokinetic and pharmacodynamic work is needed. Discussions of all of these topics can be found as entries in this volume.

REFERENCES

1. Benet, L.Z.; Sheiner, S.B. *The Pharmacological Basis of Therapeutics*, 7th Ed.; Gilman, A.C., Goodman, L.S., Rall, T.W., Murad, F., Eds.; Macmillan: New York, 1985, 1–2.
2. Gray, J.; Lewis, L.; Nierenberg, D. Clinical pharmacology education in primary care residency programs. *Clin. Pharmacol. Ther.* **1997**, *62* (3), 237–240.

Clinical Research: Benefit and Risk Assessment

Susan Talbot

Global Biostatistical Sciences, Amgen Limited, Uxbridge, U.K.

Shahrul Mt-Isa

Biostatistics and Research Decision Sciences (BARDS) Europe, MSD Research Laboratories, MSD, London, U.K.

Qi Jiang and Chunlei Ke

Global Biostatistical Sciences, Amgen Inc., Thousand Oaks, California, U.S.A.

Abstract

This article will discuss the concept of the benefit–risk assessment of medicines by decision makers and will provide an overview of key methodology to assess the benefit–risk balance. In addition, it will discuss how health agency and industry regulations are evolving to increase consistency and transparency of the decision-making process.

Keywords: Benefit–risk assessment; Multi-criteria decision analysis; NNT/NNH; Regulatory review; Weight determination.

INTRODUCTION

Regulatory authorities ensure drugs that were granted marketing authorization are high quality, safe, and effective for use in the intended population. As a core part of their drug review process, the regulators assess the benefit–risk (B–R) balance of drugs based on the evidence submitted by the sponsor on efficacy and safety. Regulators determine if the benefits outweigh the risks, in the context of the disease area, other available therapies, and additional factors that can affect the safety and efficacy of the drug.

B–R assessments (BRAs) are complex and involve the evaluation of a large amount of data. The assessment may involve both quantitative analyses and qualitative weighting of the available evidence to access the overall B–R balance of drugs. Over recent years, regulatory authorities including the European Medicines Agency (EMA) and the US Food and Drug Administration (FDA) have reviewed their approaches to BRA and moved to a more structured approach in order to increase the consistency and transparency as well as the communication of the decision-making process.^[1,2]

BRA is not limited to the regulatory review process. For example, sponsor companies continually assess the B–R balance of drugs during the drug development process, and health technology assessment (HTA) authorities may assess B–R balance in the context of a subpopulation, or review different aspects from the regulatory authorities such as patient needs.

In this article, the concept of the BRA of medicines by decision makers is discussed, key initiatives and their outputs are reviewed, structured frameworks and quantitative methodologies to assess the B–R balance are introduced, and finally challenges in the area are discussed.

DEFINITIONS

A “benefit” is the magnitude of a favorable effect (of the drugs) without reference to its probability of occurrence.^[3] The International Council for Harmonisation (ICH) M4E(R2) Guideline emphasizes “key benefits,” which are “favorable effects generally assessed by primary and other clinically important endpoints across the studies in a development program.”^[4] Important product characteristics such as “convenience” dose regimen or administration route that may lead to improved patient compliance and indirect favorable effects such as “vaccine herd immunity” that also affects those other than the patients may be considered as benefits.^[4]

On the other hand, “risk” (in relation to the use of a drug) is a multidimensional concept of the probability of occurrence and the magnitude of any unfavorable effects of the drugs.^[3] Risks relate to the quality, safety, or efficacy of the drugs on patients’ or public health, as well as any undesirable effects on the environment (European Commission (EC) Directive 2001/83/EC).^[5] The ICH M4E(R2) Guideline emphasizes “key risks,” which are “unfavorable effects that are important from a clinical and/or public health perspective in terms of their frequency and/or severity.”^[4]

A BRA is a process comprising many aspects and requires the use of various techniques to establish the B–R balance. The general principles of the BRA of medicines should be based on the available tests and clinical trials carried out on the product designed to test the efficacy and safety, and accounting for the potential impact on the evaluation of benefits and risks through the pharmacovigilance activities (EC Directive 2001/83/EC),^[5] not including economic and other considerations such as “cost-effectiveness” (EC Community Law Regulation 726/2004).^[6] It is noteworthy to acknowledge that there are alternative definitions of BRA outside the regulatory world, such as for HTA.

DECISION MAKERS

Decisions based on the B–R balance of drugs are made by multiple stakeholders throughout the product development process and eventual clinical use (Fig. 1). Specific examples of stakeholders include the following:

- Sponsor internal decision-making portals: decision to develop a drug, to progress from one phase to the next, or to submit for regulatory review
- Regulatory review: initial review, periodic safety updates, and emerging safety issue
- HTA/reimbursement authorities/payers: comparative effectiveness, assessment of subpopulations
- Health-care providers: clinical guidance, resource utilization
- Patient/clinician joint decision-making: ultimately patients make the decision on whether they want to receive a treatment

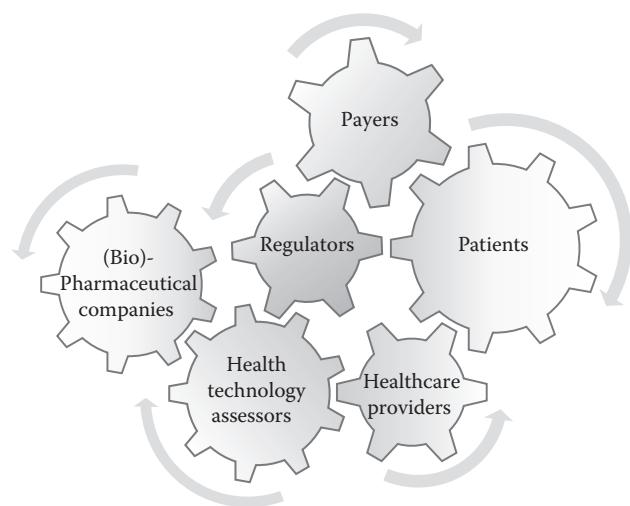


Fig. 1 Examples of decision makers

Source: Adapted from IMI PROTECT (<http://protectbenefitrisk.eu/index.html>)

These stakeholder groups or individuals may have differing perspectives on the benefits of a medicine or differing tolerance levels for risks. Consequently, these may lead to different views on the B–R balance of a medicine. Increasingly, sponsors, regulators, reimbursement/health-care providers, and clinicians are aiming to ensure patient perspectives on benefits and risks are appropriately considered during the decision-making process.^[7–10]

KEY INITIATIVES

Over recent years, the pharmaceutical industry, government regulatory agencies, HTA authorities, academia, and patient advocacy groups have increased their focus on B–R management. Several initiatives were formed aiming to standardize BRA, to improve transparency, reproducibility, quality, and communication of those assessments. Key initiatives and regulatory activities are described in this section.

Food and Drug Administration

Starting in 2009, the FDA Center for Drug Evaluation and Research (CDER) identified the need for a structured BRA framework to facilitate discussions and improve the communication of the reasoning behind their decision on the B–R balance. The fifth reauthorization of the Prescription Drug User Fees Act 2013–2017 (PDUFA V) in the Food and Drug Administration Safety and Innovation Act further committed the FDA to implement a structured B–R (sB/R) framework in the FDA’s review templates.^[1] (See FDA BRAF framework.)

The FDA CDER developed a qualitative sB/R framework through an extensive review and analysis of previous and ongoing regulatory decisions. They planned a staged approach to the implementation of the sB/R framework in order to allow opportunity for continual refinement. In 2015, the FDA implemented the framework for new molecular entity and original biologics license application (BLA) with the eventual aim of expanding its use to all new drug applications by fiscal year 2017.^[1]

The current FDA sB/R framework does not explicitly incorporate patient perspectives on benefit and risks. The 21st Century Cures Act, enacted into US law in December of 2016, importantly considers patient-focused drug development and requires the FDA to establish a structured BRA framework where patient experience data may be considered.^[8] Furthermore, the sixth reauthorization of the PDUFA (PDUFA VI, 2018–2022) aims to strengthen efforts to incorporate patient perspectives into the drug development and review process, and advances the use of a structured approach to inform regulatory decisions on whether the benefits of a medicine outweigh its risks.^[11]

Since 2012, the FDA Centers for Devices and Radiological Health (CDRH) have examined ways to incorporate

the patient perspectives into the B–R decision-making process on medical devices. Their guidance document, issued in 2016, “Factors to Consider When Making Benefit-Risk Determinations in Medical Device Premarket Approvals and De Novo Classifications” highlighted the importance of considering the patient perspectives on what constitutes a meaningful benefit and their tolerance levels to risk within the BRA of devices.^[7] A companion guidance document on patient preference information (PPI) was issued in parallel—explaining the concepts that sponsors and other stakeholders should consider when choosing to collect PPI, which may inform FDA’s B–R determinations.^[12] The CDRH is a member of the Medical Device Innovation Consortium (MDIC), a nonprofit public–private partnership. In 2015, the MDIC Patient Centered Benefit-Risk (PCBR) issued a Framework Report. The Framework Report includes a catalog, which provides an overview of a range of available patient preference methods.^[12] These patient preference methods are defined in the PCBR Framework Report as “Patient preference methods are methods for collecting and analyzing data that allow quantitative assessments of the relative desirability or acceptability to patients of attributes that differ among alternative medical treatment approaches.”^[13]

European Medicines Agency

Started in 2009, the 3-year EMA B–R project aimed to “enhance the transparency of the B–R decision-making process, and facilitate the communication of the rationale for each decision, both within the regulatory system and to the public.”^[14] The project focused initially on researching the current practice of BRA in regulatory agencies in the European Union, developing and testing tools for BRA, with the final phase of the project being training and implementation phase (started in 2012).^[2] One of the recommendations of the project was the use of the Effects Table as a tool to summarize the key benefits and risks.^[14] The Effects Table subsequently was implemented by the EMA in the Rapporteur’s Day 80 assessment report template and is updated throughout the assessment until the Day 210 report and eventual inclusion in the European Public Assessment Report (EPAR). (For an example, see the “EMA Effects Table” section.)

International Council for Harmonisation

The ICH recently updated two guidance documents relating to BRA, intending to set a common standard among the ICH regions. The ICH guidance E2C(R2) Periodic Benefit-Risk Evaluation Report (PBRER), adopted from November 2012, expanded the concept of the PSUR (Periodic Safety Update Report) from an interval safety report to a cumulative B–R report. Section 3.18 of the PBRER should provide an “integrated and critical analysis of the key information” in the BRA.^[15]

With the emphasis on providing greater structure for BRA in regulatory decision-making, ICH revised their guidance on the Common Technical Document (CTD) Section 2.5.6 “Benefits and Risks Conclusions” in M4E(R2), adopted in June 2016. Section 2.5.6 of the CTD should “provide a succinct, integrated, and clearly explained benefit-risk assessment of the medicinal product for its intended use.”^[4]

CIRS UMBRA Initiative

In 2012, the Centre for Innovation in Regulatory Science (CIRS) formed the “Unified Methodologies for Benefit-Risk Assessment” (UMBRA) initiative, consolidating various activities in the field of BRA with the aim to provide “a single platform for the development, assessment, implementation and ongoing refinement of an internationally acceptable, structured, systematic, standardized approach to the benefit-risk assessment of medicines.”^[16] Initiatives consolidated under the UMBRA initiative included Pharmaceutical Research and Manufacturers of America (PhRMA) Benefit-Risk Action Team (BRAT), which initiated in 2006 and developed the PhRMA-BRAT framework,^[17,18] which was later renamed as CIRS-BRAT (see the “Framework” section).

Innovative Medicines Initiative and European Federation of Pharmaceutical Industries and Associations

The IMI-PROTECT is the Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium (PROTECT) program funded by the Innovative Medicines Initiative (IMI) Joint Undertaking between the European Union and the pharmaceutical industry association European Federation of Pharmaceutical Industries and Associations (EFPIA), coordinated by the EMA. The IMI-PROTECT Benefit-Risk group specifically focused on structured methodologies for B–R integration and representation (<http://protectbenefitrisk.eu>).^[19] They conducted a systematic review of BRA methodologies^[20] and proposed candidate methodologies for testing in several case studies.^[21] The final report provides recommendations for BRA methods and visual representation including a five-stage guidance “roadmap.”^[22,23] Preliminary research was also conducted in “patient and public involvement” in BRA, with a focus on testing preference elicitation methods, assessing the preference and understanding of visual display formats, and ways to communicate the B–R balance.^[24]

IMI ADVANCE (Accelerated Development of Vaccine Benefit-Risk Collaboration in Europe) project was launched in December 2013 to build a pan-European framework for rapidly assessing and communicating the benefits and risks of vaccines. Several white papers and code of conduct for B–R monitoring and assessment of vaccines have been published, including near real-time BRA using multinational health-care databases.^[25]

IMI PREFER (Patient Preferences in Benefit-Risk Assessments during the Drug Life Cycle) officially started in October 2016. It is part of the second phase of IMI including partners representing patient organizations and clinical research, HTA bodies, and industries. The aim of the project is to establish recommendations to support the development of guidelines on how and when to include patient perspectives on benefits and risks of medicinal products.^[26]

ISPOR Risk-Benefit Management Working Group

The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) established the Risk-Benefit Management Working Group in 2005 with a focus on the methodology for BRA. They subsequently published a comprehensive systematic literature review to identify quantitative methodologies for BRA.^[27] The ISPOR Good Practices for Outcomes Research reports (www.ispor.org/workpaper/practices_index.asp) also include useful guidance recommendations for conducting research relevant for BRA (including multi-criteria decision analysis and conjoint analysis).^[28–30]

EFSPI Benefit-Risk Special Interest Group and QSPI Benefit-Risk Working Group

Due to the technical nature of the structured quantitative BRA, it is inevitable that many statisticians are at the forefront of the scientific advancement of the area. Two B–R statistical special interest/working groups (BRSIG/BRWG) were established under the auspices of the European Federation of Statisticians in the Pharmaceutical Industry (EFSPI), and the Society of Clinical Trials/Quantitative Sciences in the Pharmaceutical Industry (SCT/QSPI) in the United States, respectively. Although the ultimate goals of these two B–R working groups are similar,^[31,32] i.e., to promote and enhance quantitative BRA methodologies from statistical standpoints, EFSPi BRSIG focuses on the Europe efforts, whilst QSPI BRWG focuses on the US efforts. Both groups exchange the knowledge on a regular basis to keep each other up to date. The outputs from the EFSPi BRSIG and QSPI BRWG include newsletters, review articles,^[33] a book,^[34] and an internet blog (www.benefit-risk-assessment.com/).^[35]

STRUCTURING BRAS

Frameworks

Frameworks are structured step-by-step guides as to how to conduct BRAs. The frameworks aim to aid the decision-making process by establishing a standardized approach, and thereby increasing consistency, transparency in the decision-making process, and subsequently facilitating the communication of the decision.

Five frameworks emerging from the key initiatives are described in this section. Although the frameworks differ in the number of steps and the details contained within, many contain common elements: clearly define the decision problem including the target population and treatment alternatives, identify the key benefits and risks, identify the sources of data (including the clinical trials results), assess the (relative) importance of the key benefits and risks, evaluate uncertainty, summarize the data, and interpret the results.

Nixon^[36] provides a side-by-side comparison of the ProACT-URL and BRAT frameworks and concludes, “Although the two frameworks BRAT and ProACT-URL differ in their details, the common roots of the methods lead them to being essentially the same.”

ProACT-URL

ProACT-URL was first introduced by John Hammond, Ralph Keeney, and Howard Raiffa as a stepwise approach to making effective decisions.^[37] The essence is to divide and conquer—even the most complex—decision problems in eight steps. The ProACT steps remind decision makers that best decisions are the ones made proactively, and the remaining steps help to reflect and clarify the decisions in volatile and evolving environments.^[37] The URL steps reflects the uncertainty and risk faced by decision makers. A feature of ProACT-URL is the use of an *Effects Table* to represent the key drivers of the BRA, which has subsequently been recommended by the EMA (EMA/90842/2015 Section 5. Benefit risk assessment—Effects Table).^[38]

The EMA Benefit-Risk Working Group developed an adaptation of Hammond’s ProACT-URL approach to improve the EMA’s assessment of benefits and risks of medicines. They expanded ProACT-URL to be more specific to the Agency’s needs.^[39] The eight steps of the ProACT-URL for use in drug BRA are summarized in Fig. 2.

EMA Effects Table

Following the EMA Benefit-Risk Working Group adaptation of ProACT-URL for drug BRA, they subsequently recommended the use of an Effects Table for Day 80 critical assessment report.^[38] An Effects Table is a compact and consistent tabular presentation of the key benefits and risks data considered in the decision-making process, which should be used for new marketing applications and important extensions of indication applications. Figure 3 shows the proposed structure of the EMA Effects Table, which is a living document that would be included in the Regulatory Rapporteur’s Day 80 report and is updated for Day 120 List of Questions and for the Committee on Human Medicinal Products Day 210 report, and subsequently included in the EPAR. Further discussions on Effects Table are available on www.benefit-risk-assessment.com/.^[35]

Problem	Determine the nature of the problem and its context
	Frame the problem
Objective	Establish objectives that indicate the overall purposes to be achieved
	Identify criteria for (a) favourable effects, and (b) unfavourable effects
Alternatives	Identify the options to be evaluated against the criteria
Consequences	Describe how the alternatives perform for each criterion, i.e., the magnitudes of effects, and their desirability or severity, and incidence
Trade-off	Assess the balance between favourable and unfavourable effects
Uncertainty	Report the uncertainty associated with the favourable and unfavourable effects
	Consider how the balance between favourable and unfavourable effects is affected by uncertainty
Risk tolerance	Judge the relative importance of the decision maker's risk attitude for the product
	Report how this affected the benefit-risk balance
Linked decisions	Consider the consistency of this decision with similar past decisions, and assess whether taking this decision could impact future decisions

Fig. 2 PrOACT-URL for drug BRA
Source: Adapted from <http://www.protectbenefitrisk.eu/PrOACT-URL.html>.^[19]

Effect	Short description	Unit	Treatment	Control	Uncertainties/ Strength of evidence	References
Favourable effects (key benefits)						
Unfavourable effects (key risks)						

Fig. 3 EMA Effects Table
Source: From <http://www.benefit-risk-assessment.com/>.^[35]

BRAT Framework

The BRAT framework, originally developed by PhRMA, comprises processes and tools to improve BRA by enhancing the transparency, reproducibility, and communication of the B–R balance.^[17,18] In 2012, the work of PhRMA-BRAT was transitioned to the CIRS and renamed CIRS-BRAT. The CIRS-BRAT tool and User Guides are available for download at www.cirs-brat.org and include a tutorial of B–R analysis within the Framework User Guide.

The framework comprises a six-step process: (i) define the decision context, (ii) select benefit and risk outcomes,

(iii) identify and extract source data, (vi) customize the framework, (v) assess outcome importance, and (vi) display and interpret key B–R measures.^[17,18]

FDA sB/R Framework

The sB/R framework is a 5 × 3 grid that FDA’s reviewers populate with a clear description of key B–R determinations, including an analysis of the condition, current treatment options, benefits, risks, and measures to minimize potential safety concerns. The framework aims to clarify how FDA reaches the final regulatory decision in a

succinct summary. Like the EMA Effects Table, the FDA sB/R Framework is also a living document. Figure 4 shows a blank template for FDA.

Quantitative Methodologies

Quantitative methodologies can be useful to aid decision makers when exploring the B–R balance, particularly where there are multiple criteria to be considered or conflicting viewpoints on the importance or relevance of these criteria from different stakeholders.^[39] Quantitative methodology may also make explicit the use of views from stakeholder groups or individuals by formally incorporating data on preferences in order to assess the B–R balance. It is important to note that the aim of quantitative methodology in BRA is not to replace the qualitative expert judgment but to aid the decision-making process.^[40]

The ISPOR Risk-Benefit Management Working Group, EMA Benefit-Risk methodology project, and IMI-PROTECT performed literature reviews of the available methodologies and provided the details of several quantitative methodologies with potential application in BRA, including MCDA and stochastic multi-attribute acceptability analysis (SMAA), and quantitative measures including number needed to treat (NNT), number needed to harm (NNH), impact numbers, incremental net health benefit, and benefit:risk ratio. For further reading on the available methodology, the final review papers from the three initiatives are recommended.^[20,27,41]

Here, the quantitative measures NNT/NNH and the quantitative methodology MCDA are described in the context of BRA.

Number Needed to Treat/Number Needed to Harm

NNT refers to the average number of patients who need to be treated with a given treatment to prevent one additional event compared with a control therapy or no treatment.^[42] Let p_1 and p_2 denote the proportion of patients with an

event in the control and treatment group, respectively, and the NNT is defined as

$$\text{NNT} = 1/(p_1 - p_2).$$

Preventing the event used to evaluate the efficacy of a treatment is a benefit event. NNT is considered a useful measure of clinical benefit, being intuitive and interpretable.^[42–44]

Similarly, this measurement can be used to evaluate the risk and usually is termed as the number needed to harm (NNH). Specifically, the NNH is calculated as

$$\text{NNH} = 1/(q_2 - q_1)$$

where q_1 and q_2 are the incidence rates of an adverse event (AE) in the control and treatment groups, respectively. NNH is interpreted as the number of patients who need to be treated to cause one more AE compared with a control therapy or no treatment.

In the case where a single benefit and a risk can be identified in BRA, NNT and NNH can be the useful and simple methods to provide an overall balancing of benefit and risk. For example, AE-adjusted NNT (AE-NNT) estimates the number of patients needed to treat to prevent one additional event without inducing treatment-related AEs compared with the control.^[45] AE-NNT is useful for a situation where there is only one AE of interest, and the benefit event and AE are independent.

The overall BRA can be performed through a comparison of NNH and NNT using the ratio $r = \text{NNH}/\text{NNT}$. This ratio refers to the number of benefit events prevented for each AE caused by treatment on average. A favorable B–R profile may be concluded if the ratio is >1 . One weakness of this approach is that the relative importance of the AE to the benefit is not incorporated, which may imply that the benefit and risk are of the same importance. However, a threshold C_r can be set for the ratio to claim that the benefit outweighs the risk when $r > C_r$, where the value of C_r depends on the relative clinical importance of the benefit and AE. The relative importance or relative weight of the benefit and risk can also be incorporated into the calculation of NNH and NNT as an adjusted ratio.^[46]

To evaluate the uncertainties in the ratio of NNH and NNT, a confidence interval (CI) can be constructed. The ratio can be expressed as

$$\hat{r} = \frac{\hat{q}_2 - \hat{q}_1}{\hat{p}_1 - \hat{p}_2} = \frac{\Delta \hat{q}}{\Delta \hat{p}}.$$

A delta method can be applied to derive the variance estimate for $\hat{r}$:

$$\hat{v}(\hat{r}) = \hat{r}^2 \left[\frac{\hat{v}(\Delta \hat{q})}{\Delta \hat{q}^2} + \frac{\hat{v}(\Delta \hat{p})}{\Delta \hat{p}^2} - 2 \frac{\text{cov}(\Delta \hat{q}, \Delta \hat{p})}{\Delta \hat{q} \Delta \hat{p}} \right]$$

Benefit-risk integrated assessment		
Benefit-risk dimensions		
Dimensions	Evidence and uncertainties	Conclusions and reasons
Analysis of condition		
Current treatment options		
Benefit		
Risk		
Risk management		

Fig. 4 FDA sB/R framework template

Source: From Draft PDUFA V Implementation Plan.^[1]

Assuming the normal approximation, a Wald-type CI for r can be constructed. However, in general, the normal approximation may not always work well and the ratio could be heavy-tailed or bimodal.^[47] This CI may also be problematic when either NNH or NNT is not significantly greater than zero. As an alternative, a bootstrapping approach can be used to construct the CI. The issue with the construction of CIs is similar to that for NNT. Altman^[48] provides a way to interpret the CI for NNT that crosses zero. Refer to the work of Ke and Jiang^[49] for more discussions on the CI for the NNH-to-NNT ratio.

NNT has been extended to a survival endpoint.^[49–52] The ratio of NNH and NNT can be evaluated in a similar fashion. Ke and Jiang^[49] proposed a general semiparametric additive hazards model to estimate NNT and the ratio of NNH and NNT for use in BRA.

As an example, He^[53] described the use of NNT and NNH in evaluating the B–R balance of vorapaxar, an oral antiplatelet agent, versus a placebo on the background of standard of care in the secondary prevention of atherothrombotic events in patients who had a history of myocardial infarction, ischemic stroke, or peripheral arterial disease. CV death (efficacy) and fatal bleeding (safety) are the important benefit and risk endpoints. The exposure-adjusted NNT and NNH for vorapaxar was 918 patient years and 2971 patient years compared with placebo for CV death and fatal bleeding, respectively. Therefore, treatment with vorapaxar plus standard of care in 918 patient years may prevent one CV death, whereas treatment with vorapaxar plus standard of care in 2971 patients may cause one fatal bleeding. The ratio of NNH to NNT is 17, indicating that one fatal bleeding would be caused for every 17 CV deaths prevented by vorapaxar. Based on the results, and a favorable balance observed during the analysis of other key benefits and risks, it was concluded that vorapaxar plus standard of care demonstrated a favorable B–R profile compared to the standard of care alone.

Multi-Decision Criteria Analysis

MCDA originates from decision theory and allows for the consideration of multi-attributed decision outcomes,^[54] thereby fitting for use in the assessment of the B–R balance

of a medicine, where multiple benefit and risk criteria may be considered.^[55]

The MCDA was defined by Keeney and Raiffa as “an extension of decision theory that covers any decision with multiple objectives. A methodology for appraising alternatives on individual, often conflicting criteria, and combining them into one overall appraisal....”^[54] Other broader definitions exist such as the approach taken by ISPOR MCDA Good Practices Task Force, which includes “partial” MCDA techniques where “...including in our consideration of MCDA methods that help deliberative discussions using explicitly defined criteria, but without quantitative modeling.”^[28] Here, we consider Keeney–Raiffa’s approach, using MCDA to provide an overall quantitative assessment of B–R balance and thereby enabling simultaneous comparison of multiple outcomes for two or more treatment options. We describe the weighted sum model (WSM) approach to MCDA using the value function method, which is a common approach used.^[39]

MCDA follows a structured stepwise approach, with the initial steps (Fig. 5) essentially in line with the broader decision-making frameworks previously described (see the “Frameworks” section). At a high level, the steps involve breaking the problem into smaller components once the decision context has been set: identifying the B–R criteria to evaluate the treatment alternatives, transforming the data into a standardized unit (using value functions), scoring the data, directly comparing the criteria (weighting), and finally aggregating the individual criteria into an overall assessment.

The criteria are often structured visually into a hierarchical value tree (see the “Visualization” section), to group criteria and aid selection of the final criteria that reflect the value of each treatment option (i.e., the key benefits and risks).

The next steps of the MCDA (Fig. 6) bring together the multiple criteria using a common unit: scoring the data (or performance), assessing the relative importance of the criteria (weighting), and finally calculating the overall score for each treatment option. The scoring and weighting steps involve subjective judgment by the relevant stakeholders (or decision makers).

Score Performance: Partial value functions are constructed for each B–R criterion and used to obtain

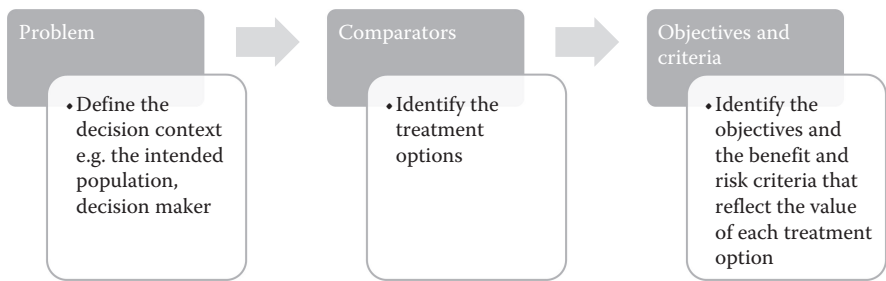


Fig. 5 Initial steps of MCDA: In line with broader decision-making framework

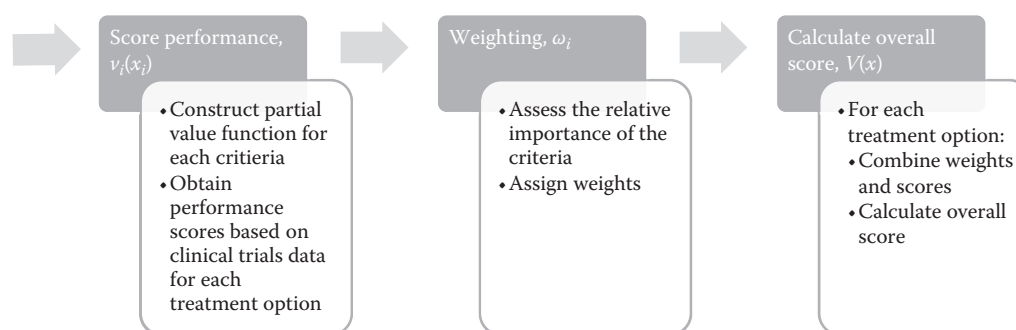


Fig. 6 Next steps of MCDA: Bring together multiple criteria

performance scores based on the clinical trials data (see Fig. 7). They represent a stakeholder's judged value in the performance of a treatment option across the plausible levels within a criterion. These partial value functions are essentially standardized scales across all criteria. Linear partial value functions (e.g., Fig. 7a) that assumes equal judged value per unit change in the level within a criterion are often used, although nonlinear functions (e.g., Fig. 7b) such as step or exponential may also be considered.

The partial value functions are often bounded by a lower value that represents the worst plausible outcome and an upper value that represents the best plausible outcome. The choice of the plausible values may be the range of possible observed values for an endpoint within the clinical trial data (e.g., systolic blood pressure: a minimum plausible value of 70 mmHg and a maximum plausible value of 270 mmHg). When bounded by 0 and 100, the value “100” represents the most preferred score and “0” represents the least preferred score, based on the mapped plausible outcome values. The partial value functions represent not only the ordering of preferences but also the strength of preferences for changes within each criterion. For further reading on value functions, see the work of Belton and Stewart.^[39]

Weighting: This step assesses the relative importance *between* the criteria. Swing weighting method can be used to define the trade-off that the stakeholder will make

between criteria (for details, see the work of Mussen^[56]). The stakeholder judges the weights to be allocated to each criterion, with the weights ω_i representing the relative importance of the “swing” from the worst plausible value to the best plausible value in the i th criterion. Weights are assigned based on the criteria alone, independent of treatment. The weights can then be normalized across criteria so the total sum of all weights equals 100, but preserving the ratios between the weights.

Methodology for constructing partial value functions and preference weights is not described in detail here, including the elicitation methods. For further reading, see the works of Keeney and Raiffa^[54] and Belton and Stewart,^[39] and also Hockley^[26] for recommendations on elicitation methods for the patient and public involvement specific for BRA of medicines.

Calculate overall score: In the final step, the partial value functions are combined into an overall score, taking into account the relative importance (weights) of the criteria. Assuming an additive model, the overall score is calculated using the weighted sum approach. Weighted score can be calculated separately for the benefits and for the risks, and overall for the benefits and risks combined.

For a given treatment a , let $v_i(a_i)$ represent its performance score for criterion i , and ω_i represent the weight for each criterion i . Then, for n criteria, the overall weighted B–R score is given by

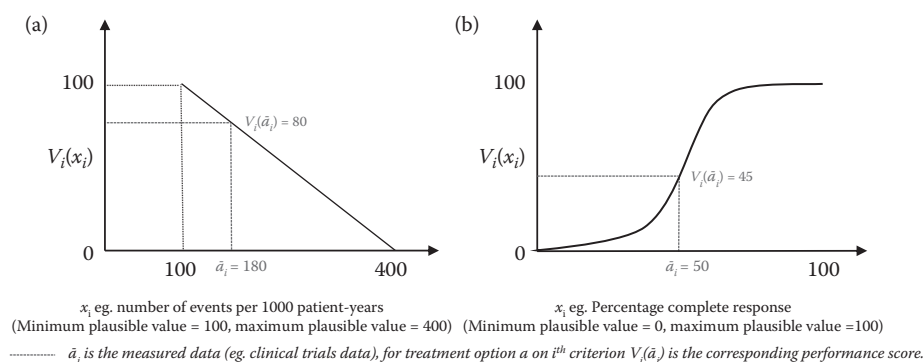


Fig. 7 Examples of partial value functions

$$V(a_i) = \sum_{i=1}^n \omega_i v_i(a_i)$$

where $\sum_{i=1}^n \omega_i = 100$

The overall B–R score and the contribution of each individual criteria can be displayed graphically, e.g., using stacked bar charts (see the “Visualization” section), which can aid in the reporting and interpretation of results.

The WSM MCDA approach described here requires several assumptions regarding the criteria, and their weights, including additive independence, preference independence of criteria, value independence, and the need for the weights to satisfy the trade-off requirements.

Sensitivity analysis is required to assess the robustness of the overall B–R score to changes in the shape of the value function, scores, weights, and input data, and to understand the effects of all sources of uncertainty. Broekhuizen^[57] performed a literature review of how uncertainty can be explicitly taken into account in MCDA and concluded that deterministic sensitivity analysis is most likely suitable for most health-care policy decisions.

For an example of MCDA used in the regulatory setting, see the work of Marcelon,^[58] for details of the development of the MCDA model, and EMA/415775/2014^[59] for the EMA Assessment Report. For a literature review of health-care MCDAs, see the work of Marsh,^[60] and for a detailed discussion of application of MCDA for HTA, see the work of Thokala.^[61]

General Methods to Deal with Uncertainties

With much increased emphasis from regulatory agencies, industry, and the medical community on the efforts to further enhance structured BRA and improving transparency, consistency, and communication, uncertainty has been recognized as a key issue in BRA. The FDA has highlighted the importance of evaluating the uncertainty related to BRA through its structured B–R framework,^[1] including how to translate the BRA from premarket to postmarket. EMA indicated that a quantitative model could be useful to explore the impact of uncertainty related to BRA, while judgment plays a key role in BRA.^[55]

Identifying and addressing uncertainties is critical for BRA. Uncertainty could come from many aspects of BRA, such as sources, timing, and nature of information available throughout a drug’s life cycle.^[62] Some uncertainties could be considered known such as efficacy estimates from clinical trials; some uncertainties are unknown unknowns such as AEs with low frequency that may not be well understood in clinical trials; and some uncertainties are unknown unknowns such as a very rare safety event, which may only emerge in the postmarketing setting. Our understanding is that there are three major sources for uncertainties, which could be associated with the level of evidence

for benefits and risks in clinical development, respectively; associated with the relative importance of benefits versus risks; or associated with translation of BRA from a premarket to a postmarket setting.^[63,64]

Not all of the uncertainties can be quantified and addressed. To quantify or account for the uncertainties, we identify the source for the uncertainties first and then come up with ways to address these uncertainties.^[63,64] CIs and visualization tools could be useful to highlight the uncertainties related to the level of evidence for benefits and risks. Sensitivity analyses of weight selection are critical to evaluate the uncertainties related to the relative importance of benefits versus risks. Weights can be fixed or random, and the B–R decision can be based on the methods that incorporate the uncertainty of weighting, e.g., SMAA.^[65] In addition, continuous ongoing BRA, risk management plan, postmarket safety and efficacy studies, etc., are important to mitigate the uncertainties associated with the translation of BRA from a premarket to a postmarket setting.^[66]

In addition to the sources of uncertainties described above, there are a few other sources. For example, if a clinical event is included in both the efficacy and safety endpoints, this endpoint definition could lead to these events being counted twice in BRA. Double counting could also occur if a serious adverse reaction is identified as an important safety endpoint and malignancy is cited as a separate safety endpoint for BRA. One approach is to conduct sensitivity analysis using endpoints that do not overlap. Evaluation of uncertainties related to BRA is critical and challenging. The decision-making could be complicated. While qualitative thinking is the focus, quantitative approaches could be helpful. It is important to consider all of the relevant aspects including context matters, sources of uncertainties (clinical, methodological, and statistical) matters, and operational challenges.^[62]

Visualization

Appropriate use of visual representations not only succinctly summarize the information on B–R balance but may also contribute to better perception and may trigger advantageous emotional response.^[67] Regulators may have different perception of B–R balance depending on their gender, length of experience, and cultural background.^[68] By capitalizing on the structured quantitative BRA methodologies and effective visual displays, decision makers may control these affective factors to assure that conflicting perceptions and emotions do not dominate decisions.^[69]

However, the use of visual displays in regulatory decision-making is still lacking, but the paradigm has shifted toward being more inclusive of visual displays when communicating B–R information. A key review commissioned by the FDA acknowledged that it is not necessary to prescribe specific visual display formats for all circumstances and concluded that “no single specific format, structure, or

graphical approach emerged as consistently superior.”^[70] Additionally, numeric B–R presentation had a positive impact on several outcomes when compared to non-numeric presentation. Consequently, visual displays that present both graphical and numeric B–R information would be of better value—so long as the visual displays are not exhaustive and well designed.^[71,72] The common visual representation of benefits and risks include effects tables, tree diagrams, forest plots, and bar graphs. More examples of visual representations in BRA are available in the literature.^[23,70]

Effects Table

There are various proposals on how to create a table to present the technical information in a BRA of drugs in a regulatory setting. An exemplary use of such a table is by the EMA, which is known as an Effects Table^[38] (Fig. 3)—also see the “EMA Effects Table” section. A structured table like an Effects Table provides an increased level of clarity, consistency, and transparency to facilitate the assessment of key benefits and key risks that contribute to the current regulatory decision.

Tree Diagrams

A tree diagram is a hierarchical structure that looks like a “tree,” commonly inverted on its side, as shown in Fig. 8. It is an explicit visual map that groups associated outcomes or criteria together to assist decision makers to be more open about the selection and nonselection of any particular outcome for the current decision problem.^[69] There is not a consensus on how best to construct a tree diagram, but it is seen as an iterative process of *growing* and *pruning* the tree until all outcomes that matter have been included.

In BRA, particularly in relation to the MCDA (see the “Multi-Decision Criteria Analysis” section), tree diagrams are commonly referred to as *value tree* or *effects tree diagrams*. They are borne out of a specific context and/or a methodology^[69] and are also customizable.^[17] Value/effects tree diagrams are particularly useful at the beginning of a

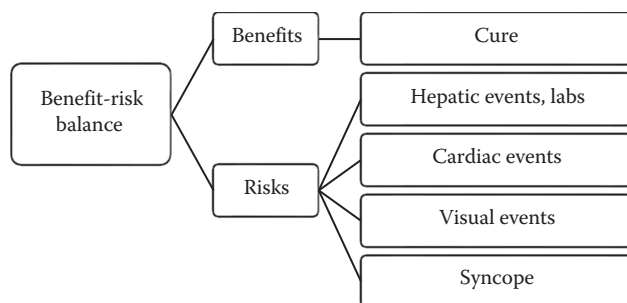


Fig. 8 An example of a tree diagram assessing the B–R balance of treatment options for acute exacerbation chronic bronchitis
 Source: Data from IMI PROTECT telithromycin case study. Quartey, G. 2012.^[21]

BRA during the planning stage, and then fine-tuned with respect to the available evidence.^[22]

Forest Plots

A forest plot (or an interval plot) can be useful to display several benefit and risk outcomes in one graphical summary to allow for a “side-by-side” comparison. An example of a forest/interval plot is shown in Fig. 9.

A forest or interval plot contains the point estimate and the associated uncertainty of several outcomes. In the example, the point estimate is the rate difference of two treatments and depicted as dots for each benefit and risk outcome. The bars represent the associated 95% CIs, with additional color-coded as the benefit or risk criteria. The plot is flexible to accommodate other measures such as odds ratios and means. However, because of a single axis, every outcome on a forest/interval plot must be presented in the same unit.

Bar Graphs

A bar graph, or a bar chart, is a common visual representation method for categorical data. The height (or length in the case of a horizontal bar graph) of the bars represents the quantitative information. For example, Fig. 10 shows the stacked bars of different risk outcomes by treatment. Each colored bar represents the rates per 1000 patients for each risk outcome, e.g., 5/1000 patients of cardiac events for treatment B, measured from the bottom to the top of the lilac-colored bar on the right. The color scheme is chosen from the Color Advice for Cartography website (<http://colorbrewer2.org/>).

There are many various other forms of bar graph,^[23] including the histogram that carries information in both the height and width of the bars. Stacked bar graphs are often found useful in a BRA to represent multiple B–R outcomes being considered. In particular, a single stacked bar graph can easily accommodate quantitative benefit and risk data measured in the same units using the value-based B–R methods such as MCDA and SMAA. The overall height of the bar would subsequently represent the total B–R value score of that treatment option that can be used to rank the overall B–R balance to support decision-making.

CHALLENGES

Challenges with BRA have been well recognized (see the work of Guo^[27] and EMA/297405/2012^[55]). It involves multifactorial sources of evidence and is informed by science, medicine, and policy (FDA PDUFA V^[11]). Human judgment, which is unavoidably subjective, plays an essential role in BRA as data and models do not make a decision. Despite many efforts from regulatory,

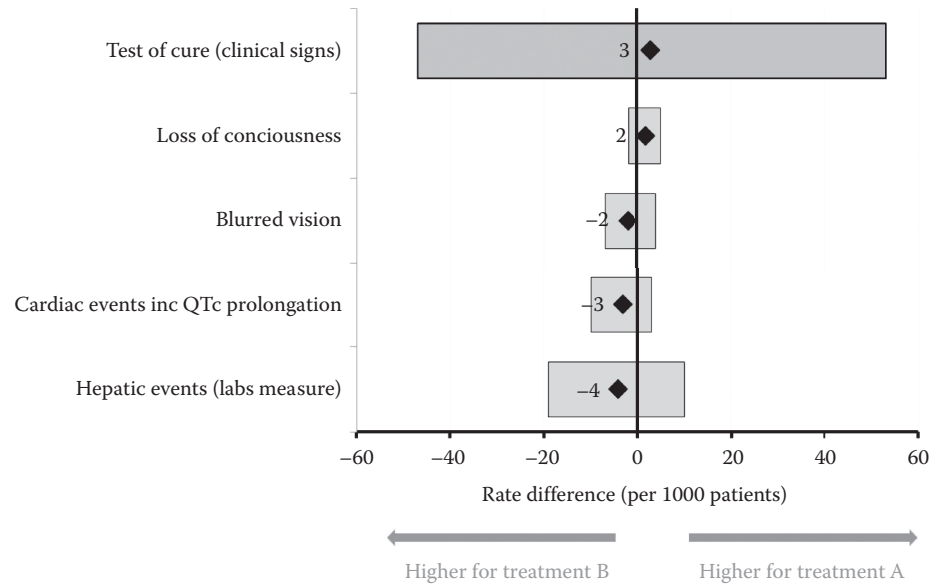


Fig. 9 A forest/interval plot comparing the rate difference between two treatments for acute exacerbation chronic bronchitis
Source: Data from IMI PROTECT telithromycin case study. Quartey, G. 2012.^[21] Reproduced using CIRS-BRAT tool (<http://www.cirsci.org/brat/>).

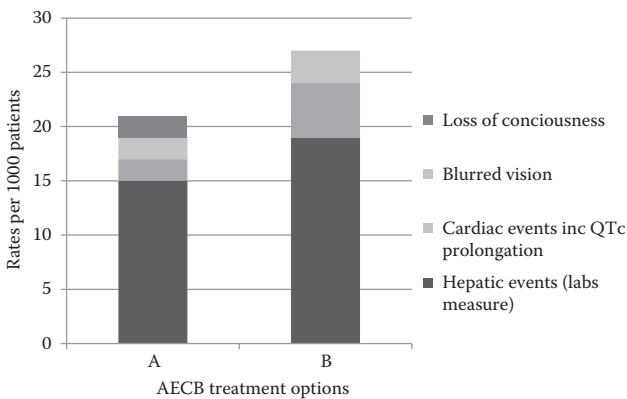


Fig. 10 A stacked (vertical) bar chart showing the rates of unfavorable outcomes (risks) for two treatments for acute exacerbation chronic bronchitis
Source: Data from IMI PROTECT telithromycin case study. Quartey, G. 2012.^[21]

In addition, there may not be sufficient data or evidence for a BRA. The drug development process has been centered around the evidence for efficacy, and thus, safety information in the premarket BRA is not specifically targeted with potentially short follow-up duration to understand long-term risk or risk for rare events. On the other hand, during postmarket BRA, no or little efficacy evidence continued to be collected, and large amount of safety data are usually collected in an observational fashion. A drug development strategy targeting evidence generation for BRA should be recommended. There is also limitation of access to data for competing treatment.

CONCLUSION

BRA is receiving increased attention across all stages of clinical research including product development, regulatory assessment, and HTA. Several initiatives have looked and continue to look at how to improve the transparency and consistency of the decision-making process, as well as how to best capture and incorporate patient preferences in the decision-making process. Methodologies are emerging for use in clinical research to aid the decision-making process—including framework and quantitative methods such as NNT/NNH and MCDA described here.

Challenges exist in the implementation and the acceptability of quantitative methodology, but increased focus on quantifying patient opinion may lend itself to more quantitative methodology. Alongside the framework, structured quantitative methods can aid discussions and increase transparency of the BRA with well-designed visual representations and appropriate sensitivity analysis.

academia, and industry, there is no widely accepted approach to BRA.

To apply quantitative methods to BRA, it is necessary to incorporate values of benefits and risks, e.g., through weighting. Whose values should be included is not clear. Different stakeholders, including sponsors, regulators, payers, and individual physicians, patients, and caregivers, may all have different views on the values. While patients actually use the medical products, their perspectives are not typically included in the BRA. There have been some efforts quantifying the values of patients or experts, but this has not been done systematically and is usually done outside the clinical trials.

REFERENCES

- Food and Drug Administration (FDA). Structured Approach to Benefit-Risk Assessment in Drug Regulatory Decision-Making: Draft PDUFA V Implementation Plan—February 2013. <https://www.fda.gov/downloads/ForIndustry/UserFees/PrescriptionDrugUserFee/UCM329758.pdf> (accessed June 2017).
- European Medicines Agency (EMA). Benefit-Risk Methodology Project, London 12 March 2009. Doc. Ref EMEA/108979/2009. http://www.ema.europa.eu/docs/en_GB/document_library/Report/2011/07/WC500109477.pdf (accessed June 2017).
- Weijer, C. The ethical analysis of risks and potential benefits in human subjects research: History, theory, and implications for U.S. regulation. In *Ethical and Policy Issues in Research Involving Human Participants. Volume II*, Commissioned Papers and Staff Analysis, National Bioethics Advisory Commission (NBAC), Eds.; NBAC: Bethesda, MD, 2001, 1–29.
- International Conference on Harmonisation. M4E (R2)—Common Technical Document for the Registration of Pharmaceuticals for Human Use—Efficacy. EMA/CPMP/ICH/2887/1999; http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2016/07/WC500211309.pdf (accessed June 2017).
- The European Parliament and of the Council of the European Directive 2001/83/EC of the European Parliament and of the Council of 6 November 2001 on the Community code relating to medicinal products for human use. *Off. J. Eur. Communities* **2001**, *311*, 67–128.
- The European Parliament and the Council of the European Regulation (EC) No 726/2004 of the European Parliament and of the Council of 31 March 2004 laying down Community procedures for the authorisation and supervision of medicinal products for human and veterinary use and establishing a European Medicines Agency. *Off. J. Eur. Communities* **2004**, *136*, 1–33.
- FDA. Center for Devices and Radiological Health (CDRH). Guidance for Industry and Food and Drug Administration Staff. Factors to Consider When Making Benefit-Risk Determinations in Medical Device Premarket Approval and De Novo Classifications. Issued on August 26, 2016.
- 21st Century Cures Act. United States Public Law No: 114–255. 12/13/2016. Sec. 3002.
- ASCO Value Framework Update. Press Release May 31, 2016. <https://www.asco.org/about-asco/press-center/news-releases/asco-value-framework-update> (accessed June 2017).
- National Institute for Health and Care Excellence (NICE). Patient and Public Involvement Program. <https://www.nice.org.uk/media/default/About/NICE-Communities/Public-involvement/Patient-and-public-involvement-policy/Patient-and-public-involvement-policy-November-2013.pdf> (accessed June 2017).
- FDA. PDUFA VI. PDUFA Reauthorization performance goals and procedures fiscal years 2018 through 2022. <https://www.fda.gov/downloads/forindustry/userfees/prescriptiondruguserfee/ucm511438.pdf> (accessed June 2017).
- FDA. Center for Devices and Radiological Health (CDRH). Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders. Patient Preference Information—Voluntary Submission, Review in Premarket Approval Applications, Humanitarian Device Exemption Applications, and De Novo Requests, and Inclusion in Decision Summaries and Device Labeling. Issued on August 26, 2016.
- Medical Device Innovation Consortium (MDIC) Patient Centered Benefit-Risk Project Report: A Framework for Incorporating Information on Patient Preferences Regarding Benefit and Risk into Regulatory Assessments of New Medical Technology. 2015. http://mdic.org/wp-content/uploads/2015/05/MDIC_PCBR_Framework_Web1.pdf (accessed June 2017).
- EMA Benefit-Risk Methodology Project update on Work Package 5: Effects Table Pilot (Phase I). London: European Medicines Agency, EMA/74168/2014. http://www.ema.europa.eu/docs/en_GB/document_library/Report/2014/02/WC500162036.pdf (accessed June 2017).
- ICH guidance E2C (R2) on Periodic Benefit-Risk Evaluation Report (PBRER). EMA/CHMP/ICH/544553/1998. http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2012/12/WC500136402.pdf (accessed June 2017).
- Centre for Innovation in Regulatory Science (CIRS). Unified Methodologies for Benefit-Risk Assessment (UMBRA) initiative. <http://www.cirsci.org/decision-making-frameworks/umbra-initiative/> (accessed June 2017).
- Coplan, P.M.; Noel, R.A.; Levitan, B.S.; Ferguson, J.; Mussen, F. Development of a framework for enhancing the transparency, reproducibility and communication of the benefit-risk balance of medicines. *Clin. Pharmacol. Ther.* **2011**, *89* (2), 312–315.
- Levitan, B.S.; Andrews, E.B.; Gilsenan, A.; Ferguson, J.; Noel, R.A.; Coplan, P.M.; Mussen, F. Application of the BRAT framework to case studies: Observations and insights. *Clin. Pharmacol. Ther.* **2011**, *89* (2), 217–224.
- IMI Benefit-Risk website. Available from <http://protect-benefitrisk.eu>. Also directly available from <http://imi-protect-eu.cc.ic.ac.uk/> (accessed June 2017).
- Mt-Isa, S.; Hallgreen, C.E.; Wang, N.; Callreus, T.; Genov, G.; Hirsch, L.; Hobbiger, S.F.; Hockley, K.S.; Luciani, D.; Phillips, L.D.; Quartey, G.; Sarac, S. B.; Stoeckert, I.; Tzoulaki, I.; Micaleff, A.; Ashby, D.; and on behalf of the IMI-PROTECT Benefit-Risk participants. Balancing benefit and risk of medicines: A systematic review and classification of available methodologies. *Pharmacoepidemiol. Drug Saf.* **2014**, *23* (7), 667–678.
- Micaleff, A. 2013; Quartey, G. 2012; Nixon, R. 2013; Juhaeri, J. 2012; Phillips, L. 2013; Nixon, R. 2013; Hallgreen, C. 2013. IMI PROTECT Case Studies. <http://protectbenefitrisk.eu/Ing.html> (accessed June 2017).
- Hughes, D.; Waddingham, E.; Mt-Isa, S.; Goginsky, A.; Chan, E.; Downey, G.F.; Hallgreen, C.E.; Hockley, K.S.; Juhaeri, J.; Lieftucht, A.; Metcalf, M.A.; Noel, R.A.; Phillips, L.D.; Ashby, D.; Micaleff, A.; and on behalf of the PROTECT Benefit-Risk Group. Recommendations for benefit-risk assessment methodologies and visual representations. *Pharmacoepidemiol. Drug Saf.* **2016**, *25*, 251–262.
- Hallgreen, C.E.; Mt-Isa, S.; Lieftucht, A.; Phillips, L.D.; Hughes, D.; Talbot, S.; Asimwe, A.; Downey, G.; Genov, G.; Hermann, R.; Noel, R.; Peters, R.; Micaleff, A.

- Tzoulaki, I.; Ashby, D.; and on behalf of the PROTECT Benefit-Risk group. Literature review of visual representation of the results of benefit-risk assessments of medicinal products. *Pharmacoepidemiol. Drug Saf.* **2015**, *25*, 238–250.
24. Hockley, K.S.; and the IMI PROTECT Work Package Five Patient and Public Involvement Workstream. Recommendations for Patient and Public Involvement in the Assessment of Benefit and Risk of Medicines. 2015; <http://protectbenefitrisk.eu/documents/ppi150413v1.pdf> (accessed June 2017).
 25. ADVANCE—Accelerated Development of Vaccine Benefit-Risk Collaboration in Europe; <http://www.advance-vaccines.eu/> (accessed June 2017).
 26. PREFER—Patient Preferences in Benefit-Risk Assessments during the Drug Life Cycle; <http://www.imi-prefer.eu/> (accessed June 2017).
 27. Guo, J.J.; Pandey, S.; Doyle, J.; Bian, B.; Lis, Y.; Raisch, D.W. A review of quantitative risk-benefit methodologies for assessing drug safety and efficacy—Report of the ISPOR Risk-Benefit Management Working Group. *Value Health* **2010**, *13* (5), 657–666.
 28. Thokala, P.; Devlin, N.; Marsh, K.; Baltussen, R.; Boysen, M.; Kaló, Z.; Lönnngren, T.; Mussen, F.; Peacock, S.; Watkins, J.; Ijzerman, M. Multiple criteria decision analysis for health care decision making—An introduction: Report 1 of the ISPOR MCDA emerging good practices task force. *Value Health* **2016**, *19*, 1–13.
 29. Marsh, K.; Ijzerman, M.; Thokala, P.; Baltussen, R.; Boysen, M.; Kaló, Z.; Lönnngren, T.; Mussen, F.; Peacock, S.; Watkins, J.; Devlin, N. Multiple criteria decision analysis for health care decision making—Emerging good practices: Report 2 of the ISPOR MCDA emerging good practices task force. *Value Health* **2016**, *19*, 125–137.
 30. Hauber, A.B.; González, J.M.; Groothuis-Oudshoorn, C.G.M.; Prior, T.; Marshall, D.A.; Cunningham, C.; Ijzerman, M.J.; Bridges, J.F.P. Statistical methods for the analysis of discrete choice experiments: A report of the ISPOR conjoint analysis good research practices task force. *Value Health* **2016**, *19* (4), 300–315.
 31. European Federation of Statisticians in the Pharmaceutical Industry (EFSPI). Benefit-Risk Special Interest Group. http://www.efspi.org/EFSPi/SIGs/Benefit_Risk_SIG/EFSPi/Special_Interest_Groups/Benefit-Risk_SIG.aspx (accessed June 2017).
 32. QSPI Benefit Risk Work Group. <http://www.phusewiki.org/wiki/index.php?title=QSPI> (accessed June 2017).
 33. Mt-Isa, S.; Ouwens, M.; Robert, V.; Gebel, M.; Schacht, A.; Hirsch, I. Structured Benefit-risk assessment: A review of key publications and initiatives on frameworks and methodologies. *Pharm. Stat.* **2016**, *15*, 324–332.
 34. *Benefit-Risk Assessment Methods in Medical Product Development: Bridging Qualitative and Quantitative Assessments*, 1st Ed.; Jiang, Q., He, W., Eds.; Chapman & Hall/CRC series; Taylor & Francis Group, LLC: New York, NY, 2016.
 35. EFSPI. Benefit-Risk Assessment: The Blog around Structured Decision Making in Medicine. <http://www.benefit-risk-assessment.com/> (accessed June 2017).
 36. Nixon, R.; Dierig, C.; Mt-Isa, S.; Stöckert, I.; Tong, T.; Kuhls, S.; Hodgson, G.; Pears, J.; Waddingham, E.; Hockley, K.; Thomson, A. A case study using the ProACT-URL and BRAT frameworks for structured benefit risk assessment. *Biom. J.* **2016**, *58*, 8–27.
 37. Hammond, J.S.; Keeney, R.L.; Raiffa, H. *Smart Choices. A Practical Guide to Making Better Life Decisions*. Broadway Books: New York, NY, 2002.
 38. EMA. Guidance Document on the Content of the <co> Rapporteur Day 80 Critical Assessment Report. EMA/90842/2015. http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2009/10/WC500004800.pdf (accessed June 2017).
 39. Belton, V.; Stewart, T. *Multiple Criteria Decision Analysis: An Integrated Approach*. Springer Science & Business Media: Norwell, MA and Dordrecht, The Netherlands, 2002.
 40. European Medicines Agency. Benefit-risk Methodology Project Work Package 3 Report: Field Tests. London: European Medicines Agency EMA/718294/2011. http://www.ema.europa.eu/docs/en_GB/document_library/Report/2011/09/WC500112088.pdf (accessed June 2017).
 41. EMA. Benefit-Risk Methodology Project. Work Package 2 Report: Applicability of Current Tools and Processes for Regulatory Benefit-Risk Assessment. London. 2010. http://www.ema.europa.eu/docs/en_GB/document_library/Report/2010/10/WC500097750.pdf (accessed June 2017).
 42. Laupacis, A.; Sackett, D.L.; Roberts, R.S. An assessment of clinically useful measures of the consequences of treatment. *N. Engl. J. Med.* **1988**, *318* (26), 1728–1733.
 43. Cook, R.J.; Sackett, D.L. The number needed to treat: A clinically useful measure of treatment effect. *Br. Med. J.* **1995**, *310*, 452–454.
 44. Holden, W.L.; Juhaeri, J.; Dai, W. Benefit-risk analysis: A proposal using quantitative methods. *Pharmacoepidemiol. Drug Saf.* **2003**, *12*, 611–616.
 45. Schulzer, M.; Mancini, G. B. “Unqualified success” and “unmitigated failure”: Number-needed-to-treat-related concepts for assessing treatment efficacy in the presence of treatment-induced adverse events. *Int. J. Epidemiol.* **1996**, *25*, 704–712.
 46. Guyatt, G.H.; Sinclair, J.; Cook, D.J.; Glasziou, P. Users’ guides to the medical literature: XVI. How to use a treatment recommendation. Evidence-Based Medicine Working Group and the Cochrane Applicability Methods Working Group. *JAMA, J. Am. Med. Assoc.* **1999**, *281* (19), 1836–1843.
 47. Marsaglia, G. Ratios of normal variables and ratios of sums of uniform variables. *JAMA, J. Am. Med. Assoc.* **1965**, *60*, 193–204.
 48. Altman, D.G. Confidence intervals for the number needed to treat. *Br. Med. J.* **1998**, *317*, 1309–1312.
 49. Ke, C.; Jiang, Q.; Snapinn, S. *Risk-benefit assessment approaches*, Chapter 15 in *Quantitative Evaluation of Safety in Drug Development: Design, Analysis and Reporting*, 1st Ed.; Jiang, Q., Xia, H.A. Eds.; Chapman & Hall/CRC Biostatistics Series; Taylor & Francis Group, LLC: New York, NY, 2014, 267–288.
 50. Altman, D.G.; Andersen, P.K. Calculating the number needed to treat for trials where the outcome is time to an event. *Br. Med. J.* **1999**, *319*, 1492–1495.

51. Lubsen, J.; Hoes, A.; Grobbee, D. Implications of trial results: The potentially misleading notations of number needed to treat and average duration life gained. *Lancet* **2000**, *356*, 1757–1759.
52. Mayne, T.J.; Whalen, E.; Vu, A. Annualized was found better than absolute risk reduction in the calculation of number needed to treat in chronic conditions. *J. Clin. Epidemiol.* **2006**, *59*, 217–223.
53. He, W.; Bloomfield, D.; Mai, Y.; Evans, S. *Practical considerations for benefit–risk assessment and implementation: Vorapaxar TRA-2°P TIMI 50 case study, Chapter 11 in Benefit-Risk Assessment Methods in Medical Product Development: Bridging Qualitative and Quantitative Assessments*, 1st Ed.; Jiang, Q., He, W., Eds.; Chapman & Hall/CRC Biostatistics Series; Taylor & Francis Group, LLC: New York, NY, 2016, 235–249.
54. Keeney, R.L.; Raiffa, H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, John Wiley: New York, NY, 1976, reprinted, Cambridge University Press, 1993.
55. EMA Benefit-Risk Methodology Project Work Package 4 Report: Benefit-Risk Tools and Processes EMA/297405/2012. http://www.ema.europa.eu/docs/en_GB/document_library/Report/2012/03/WC500123819.pdf (accessed June 2017).
56. Mussen, F.; Salek, S.; Walker, S. *Benefit-Risk Appraisal of Medicines: A Systematic Approach to Decision-making*. John Wiley & Sons, Chichester, 2009.
57. Broekhuizen, H.; Groothuis-Oudshoorn, C.G.; van Til, J.A.; Hummel, J.M.; IJzerman, M.J. A review and classification of approaches for dealing with uncertainty in multi-criteria decision analysis for healthcare decisions. *Pharmacoeconomics* **2015**, *33*, 445–455.
58. Marcelon, L.; Verstraeten, T.; Dominiak-Felden, G.; Simonon, F. Quantitative benefit-risk assessment by MCDA of the quadrivalent HPV vaccine for preventing anal cancer in males. *Exp. Rev. Vaccines* **2016**, *15* (1), 139–148.
59. EMA CHMP Assessment Report for HPV EMA/415775/2014. http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Assessment_Report_-_Variation/human/000703/WC500170695.pdf (accessed June 2017).
60. Marsh, K.; Lanitis, T.; Neasham, D.; Orfanos, P.; Caro, J. Assessing the value of healthcare interventions using multi-criteria decision analysis: A review of the literature. *Pharmacoeconomics* **2014**, *32* (4), 345–365.
61. Thokala, P. NICE decision support group (DSU). Multiple criteria decision analysis for health technology assessment. *Value Health* **2012**, *15* (8), 1172–1181.
62. Hammad, T.A.; Neyarapally, G.A.; Iyasu, S.; Staffa, J.A.; Dal Pan, G. The future of population-based postmarket drug risk assessment: A regulator's perspective. *Clin. Pharmacol. Ther.* **2013**, *94* (3), 349–358.
63. Jiang, Q.; Ma, H.; Chuang-Stein, C.; Evans, S.; He, W.; Quartey, G.; Scott, J.; Wen, S.; Arani, R. *Understanding and evaluating uncertainties in the assessment of benefit-Risk in pharmaceutical drug development*. In *Benefit-Risk Assessment Methods in Medical Product Development: Bridging Qualitative and Quantitative Assessments*, 1st Ed.; Jiang, Q., He, W., Eds.; Chapman & Hall/CRC Biostatistics Series; Taylor & Francis Group, LLC: New York, NY, 2016, 71–83.
64. Ma, H.; Jiang, Q.; Chuang-Stein, C.; Evans, S.; He, W.; Quartey, G.; Scott, J.; Wen, S.; Arani, R. Consideration on endpoint selection, weighting determination, and uncertainty evaluation in the benefit-risk assessment of medical product. *Stat. Biopharm. Res.* **2016**, *8*, 417–425.
65. Tervonen, T.; van Valkenhoef, G.; Buskens, E.; Hillege, H.L.; Postmus, D. A stochastic multicriteria model for evidence-based decision making in drug benefit–risk analysis. *Stat. Med.* **2011**, *30*, 1419–1428.
66. DiSantostefano, R.L.; Berlin, J.A.; Chuang-Stein, C.; Quartey, G.; Eichenbaum, G.; Levitan, B. Selecting and integrating data sources in benefit–risk assessment: Considerations and future directions. *Stat. Biopharm. Res.* **2016**, *8* (4), 394–403.
67. Slovic, P.; Weber, E.U. *Perception of Risk Posed by Extreme Events*. Risk Management strategies in an Uncertain World Conference. Palisades: New York, NY, 2002.
68. Beyer, A.R.; Fasolo, B.; Phillips, L.D.; de Graeff, P.A.; Hillege, H.L. Risk perception of prescription drugs: Results of a survey among experts in the European regulatory network. *Med. Decis. Making* **2013**, *33* (4), 579–592.
69. Mt-Isa, S.; Hallgreen, C.E.; Asiimwe, A.; Downey, G.; Genov, G.; Hermann, R.; Hughes, D.; Liefucht, A.; Noel, R.; Peters, R.; Phillips, L.D.; Shepherd, S.; Micaleff, A.; Ashby, D.; Tzoulaki, I.; and on behalf of PROTECT Work Package 5 participants. Reviews of visual representation methods for benefit risk assessment of medicine. Visual representation review—Part two. April 2013; <http://protectbenefitrisk.eu> (accessed June 2017).
70. West, S.L.; Squiers, L.; McCormack, L.; Southwell, B.; Brouwer, E.; Ashok, M.; Lux, L.; Boudewyns, V.; Tindell, G.; Tant, E.; Peinado, S.; Voisin, C. Quantitative summary of the benefits and risks of prescription drugs: A literature review. Food and Drug Administration, US Department of Health and Human Services: Silver Spring, MD, August 15, 2011. Report No.: 0212305.010.001.
71. Mt-Isa, S.; Peters, R.; Phillips, L.D.; Chan, K.; Hockley, K.S.; Wang, N.; Ashby, D.; Tzoulaki, I. On behalf of PROTECT Work Package 5 participants. Reviews of visual representation methods for benefit risk assessment of medicine. Visual representation review—Part one. February 2013; <http://protectbenefitrisk.eu> (accessed June 2017).
72. Wickens, C.D.; Lee, J.; Liu, Y.D.; Gordon-Becker, S. *An Introduction to Human Factors Engineering*. 2nd Ed., Pearson Prentice-Hall: Upper Saddle River, NJ, 2004.

Clinical Research: Comparing Variabilities

Yonghee Lee

Forest Research Institute, Jersey City, New Jersey, U.S.A.

Hansheng Wang

Peking University, Beijing, People's Republic of China

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

In most clinical trials comparing a test drug and a control treatment effect is usually established by comparing mean response change from baseline of some primary study endpoints assuming that their corresponding variabilities are comparable. This entry gives several methods of comparing variabilities and explores how to analyze the results.

INTRODUCTION

In most clinical trials comparing a test drug and a control (e.g., a placebo control or an active control), treatment effect is usually established by comparing mean response change from baseline of some primary study endpoints assuming that their corresponding variabilities are comparable. In practice, however, variabilities associated with the test drug and the control could be very different. When the variability of the test drug is much larger than that of the reference drug, safety of the test drug could be a concern. Thus, in addition to comparing mean responses between treatments, it is also of interest to compare the variabilities associated with the responses between treatments.

In practice, the variabilities associated with the response are usually classified into two categories, namely, the intrasubject (or within-subject) variability and the intersubject (or between-subjects) variability. Intrasubject variability refers to the variability observed from repeated measurements from the same subject under the same experimental condition. On the other hand, intersubject variability is the variability due to the heterogeneity among subjects. The total variability is simply the sum of the intra- and intersubject variabilities. In practice, it is of interest to test for equality, noninferiority/superiority, and similarity (or equivalence) between treatments in terms of the intrasubject, intersubject, and/or total variabilities. Under a replicated parallel design, the usual F -test can be performed to compare intrasubject variabilities. Under a replicated crossover design, Chinchilli and Esinhart^[1] proposed the F -test using an orthogonal transformation to test for difference in intrasubject variability between treatments. Under replicated parallel designs, however, intersubject and total variabilities between two treatments cannot be compared by the usual F -test because the distribution of

estimators does not follow a chi-square distribution. In these cases, the modified large sample (MLS) method is suggested (see, e.g., Refs.[2–5]). The MLS method is shown to be superior to the other approximation methods in practice. Under a replicated crossover design, the MLS method is not appropriate to compare intersubject and total variabilities because estimators of variance components are not independent. Lee et al.^[6] extend the MLS method when estimators of variance components are dependent. In addition, Lee et al.^[7] study a test for comparing intersubject and total variabilities under replicated crossover designs based on the extension of the MLS method.

In practice, instead of comparing intrasubject variabilities, it may be of interest to compare intrasubject coefficient of variation (CV). In recent years, the use of intrasubject CV has become popular when assessing the reproducibility, similarity, and stability of a drug product when the drug product is repeatedly administered over a period of time. In this entry, we will discuss two statistical methods by Chow and Tse^[8] and Quan and Shih.^[9]

OVERVIEW

In the next four sections, statistical tests for equality, superiority/noninferiority, and similarity under a crossover design or a parallel design in clinical trials are derived for comparing intrasubject variability, intrasubject CV, intersubject variabilities, and total variabilities, respectively. The section “Sample Size Calculation” provides formulas for sample size calculation corresponding to respective statistical tests. Some concluding remarks are given in the last section. Throughout this article, for the sake of convenience, we will denote test formulation (drug) and reference formulation (drug) by T and R, respectively.

COMPARING INTRASUBJECT VARIABILITIES

To assess intrasubject variability, replicates from the same subject are necessarily obtained. For this purpose, replicated crossover designs or parallel group designs with replicates are commonly employed. In what follows, statistical tests for comparing intrasubject variabilities under a parallel design with replicates and a replicated crossover design (e.g., a $2 \times 2m$ replicated crossover design) will be studied.

Parallel Design with Replicates

Let x_{ijk} be the observation of the k th replicate ($k = 1, \dots, m$) of the j th subject ($j = 1, \dots, n_i$) from the i th treatment ($i = T, R$). The following linear mixed effects model is usually considered:

$$x_{ijk} = \mu_i + S_{ij} + e_{ijk} \quad (1)$$

where μ_i is the treatment effect, S_{ij} is the random effect due to the j th subject in the i th treatment group, and e_{ijk} is the intrasubject variability under the i th treatment. It is assumed that for a fixed i , S_{ij} is independent and identically distributed as a normal random variable with mean 0 and variance $\sigma_{B_i}^2$, and e_{ijk} , $k = 1, \dots, m$ is independent and identically distributed as a normal random variable with mean 0 and variance $\sigma_{w_i}^2$. Under this model, an unbiased estimator for intrasubject variance $\sigma_{w_i}^2$ is given by

$$\hat{\sigma}_{w_i}^2 = \frac{1}{n_i(m-1)} \sum_{j=1}^{n_i} \sum_{k=1}^m (x_{ijk} - \bar{x}_{ij.})^2 \quad (2)$$

where

$$\bar{x}_{ij.} = \frac{1}{m} \sum_{k=1}^m x_{ijk} \quad (3)$$

It can be seen that $n_i(m-1)\hat{\sigma}_{w_i}^2/\sigma_{w_i}^2$ is distributed as a χ^2 -distribution with $n_i(m-1)$ degrees of freedom. Note that $\sigma_{w_T}^2$ and $\sigma_{w_R}^2$ are intrasubject variances for test and reference formulation, respectively. Also, $\sigma_{B_T}^2$ and $\sigma_{B_R}^2$ are the intersubject variances for test and reference formulation, respectively.

Test for Equality

In practice, it is often of interest to test whether two drug products have the same intrasubject variability. The following hypotheses are then of interest:

$$H_0: \frac{\sigma_{w_T}^2}{\sigma_{w_R}^2} = 1 \quad \text{vs.} \quad H_a: \frac{\sigma_{w_T}^2}{\sigma_{w_R}^2} \neq 1 \quad (4)$$

A commonly used test statistic for testing the above hypotheses is given by

$$T = \frac{\hat{\sigma}_{w_T}^2}{\hat{\sigma}_{w_R}^2}$$

Under the null hypothesis, T is distributed as an F distribution with $n_T(m-1)$ and $n_R(m-1)$ degrees of freedom. Hence, we reject the null hypothesis at the α level of significance if

$$T > F_{\alpha/2, n_T(m-1), n_R(m-1)}$$

or

$$T < F_{1-\alpha/2, n_T(m-1), n_R(m-1)}$$

where $F_{\alpha/2, n_T(m-1), n_R(m-1)}$ is the upper $(\alpha/2) \times 100\%$ quantile of an F distribution with $n_T(m-1)$ and $n_R(m-1)$ degrees of freedom.

Test for Noninferiority/Superiority

The problem of testing noninferiority and superiority can be unified by the following hypotheses

$$H_0: \frac{\sigma_{w_T}^2}{\sigma_{w_R}^2} \geq \delta \quad \text{vs.} \quad H_a: \frac{\sigma_{w_T}^2}{\sigma_{w_R}^2} < \delta \quad (5)$$

When $\delta < 1$, the rejection of the null hypothesis indicates the superiority of the test product over the reference in terms of the intrasubject variability. When $\delta > 1$, the rejection of the null hypothesis indicates the noninferiority of the test product over the reference. The test statistic is given by

$$T = \frac{\hat{\sigma}_{w_T}^2}{\delta \hat{\sigma}_{w_R}^2}$$

Under the null hypothesis, T is distributed as an F random variable with $n_T(m-1)$ and $n_R(m-1)$ degrees of freedom. Hence, we reject the null hypothesis at the α level of significance if

$$T < F_{1-\alpha, n_T(m-1), n_R(m-1)}$$

Test for Similarity

For testing similarity, the following hypotheses are usually considered:

$$H_0: \frac{\sigma_{w_T}^2}{\sigma_{w_R}^2} \notin \left(\frac{1}{\delta}, \delta \right) \quad \text{vs.} \quad H_a: \frac{\sigma_{w_T}^2}{\sigma_{w_R}^2} \in \left(\frac{1}{\delta}, \delta \right) \quad (6)$$

where $\delta > 1$ is the similarity limit. The above hypotheses can be decomposed into the following two one-sided hypotheses:

$$\begin{aligned} H_{01}: \frac{\sigma_{w_T}^2}{\sigma_{w_R}^2} \geq \delta \quad \text{vs.} \quad H_{a1}: \frac{\sigma_{w_T}^2}{\sigma_{w_R}^2} < \delta \\ H_{02}: \frac{\sigma_{w_T}^2}{\sigma_{w_R}^2} \leq \frac{1}{\delta} \quad \text{vs.} \quad H_{a2}: \frac{\sigma_{w_T}^2}{\sigma_{w_R}^2} > \frac{1}{\delta} \end{aligned} \quad (7)$$

These two one-sided hypotheses can be tested by the following two test statistics

$$T_1 = \frac{\hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2} \quad \text{and} \quad T_2 = \frac{\delta \hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2}$$

We then reject the null hypothesis and conclude similarity at the α level of significance if

$$T_1 < F_{1-\alpha, n_T(m-1), n_R(m-1)} \quad \text{and} \quad T_2 > F_{\alpha, n_T(m-1), n_R(m-1)}$$

Replicated Crossover Design

Compared with the parallel design with replicates, the merit of a crossover design is the ability to make comparisons of treatment effect within subjects. In this section, without loss of generality, consider a $2 \times 2m$ replicated crossover design comparing two treatments. Under a $2 \times 2m$ replicated crossover design, in each sequence, each subject receives the test formulations m times and the reference formulations m times at different dosing periods. When $m = 1$, the $2 \times 2m$ replicated crossover design reduces to the standard two-sequence, two-period (2×2) crossover design. On the other hand, when $m = 2$, the $2 \times 2m$ replicated crossover design becomes a 2×4 crossover design, which is recommended by the United States Food and Drug Administration (FDA) for assessment of population/individual bioequivalence.^[10]

Suppose that n_1 subjects are assigned to the first sequence and n_2 subjects are assigned to the second sequence. Let x_{ijkl} be the observation from the j th subject ($j = 1, \dots, n_i$) in the i th sequence ($i = 1, 2$) under the l th replicate ($l = 1, \dots, m$) of the k th treatment ($k = T, R$). As indicated in Ref. [1], the following linear mixed effects model can be considered to describe a $2 \times 2m$ replicated crossover design:

$$x_{ijkl} = \mu_k + \gamma_{ikl} + S_{ijk} + \epsilon_{ijkl} \quad (8)$$

where μ_k is the treatment effect for formulation k , γ_{ikl} is the fixed effect of the l th replicate on treatment k in the i th sequence with constraint

$$\sum_{i=1}^2 \sum_{l=1}^m \gamma_{ikl} = 0$$

$(S_{ijT}, S_{ijR})'$ are the random effects of the j th subject in the i th sequence, which are independent and identically distributed as a bivariate normal random vector with mean $(0, 0)'$ and covariance matrix

$$\Sigma_B = \begin{pmatrix} \sigma_{BT}^2 & \rho \sigma_{BT} \sigma_{BR} \\ \rho \sigma_{BT} \sigma_{BR} & \sigma_{BR}^2 \end{pmatrix}$$

Note that σ_{BT}^2 and σ_{BR}^2 are intersubject variances under the test formulation and the reference formulation, respectively.

ϵ_{ijkl} s are independent random variables from the normal distribution with mean 0 and variance σ_{WT}^2 or σ_{WR}^2 , which are intrasubject variabilities under the test formulation and the reference formulation, respectively. It is assumed that $(S_{ijT}, S_{ijR})'$ and ϵ_{ijkl} are independent.

To obtain estimators of intrasubject variances, it is a common practice to use an orthogonal transformation, which is considered by Chinchilli and Esinhart.^[1] A new random variable z_{ijkl} can be obtained by using the orthogonal transformation

$$\mathbf{z}_{ijk} = \mathbf{P}' \mathbf{x}_{ijk} \quad (9)$$

where

$$\mathbf{x}'_{ijk} = (x_{ijk1}, x_{ijk2}, \dots, x_{ijkn})', \quad \mathbf{z}'_{ijk} = (z_{ijk1}, z_{ijk2}, \dots, z_{ijkn})'$$

and $\mathbf{P}$ is an $m \times m$ orthogonal transformation under which $\mathbf{P}'\mathbf{P}$ is an $m \times m$ diagonal matrix. The first column of $\mathbf{P}$ is usually defined by the vector $1/m(1, 1, \dots, 1)'$ to obtain $z_{ijk1} = \bar{x}_{ijk}$ and the other columns can be defined to satisfy the orthogonality of $\mathbf{P}$ and $\text{Var}(z_{ijkl}) = \sigma_{Wi}^2$ for $l = 2, \dots, m$. For example, in the 2×4 crossover design, the new random variable z_{ijkl} can be defined as

$$z_{ijk1} = \frac{x_{ijk1} + x_{ijk2}}{2} \equiv \bar{x}_{ijk}, \quad \text{and} \quad z_{ijk2} = \frac{x_{ijk1} - x_{ijk2}}{\sqrt{2}}$$

Now, the estimator of intrasubject variance can be defined as

$$\begin{aligned} \hat{\sigma}_{WT}^2 &= \frac{1}{(n_1 + n_2 - 2)(m - 1)} \sum_{i=1}^2 \sum_{j=1}^{n_i} \sum_{l=2}^m (z_{ijTl} - \bar{z}_{i.Tl})^2 \\ \hat{\sigma}_{WR}^2 &= \frac{1}{(n_1 + n_2 - 2)(m - 1)} \sum_{i=1}^2 \sum_{j=1}^{n_i} \sum_{l=2}^m (z_{ijRl} - \bar{z}_{i.Rl})^2 \end{aligned} \quad (10)$$

where

$$\bar{z}_{i.kl} = \frac{1}{n_i} \sum_{j=1}^{n_i} z_{ijkl}$$

It should be noted that $\hat{\sigma}_{WT}^2$ and $\hat{\sigma}_{WR}^2$ are independent and unbiased estimators for σ_{WT}^2 and σ_{WR}^2 , respectively.

Test for Equality

Under the null hypothesis of Eq. 4, a test statistic

$$T = \frac{\hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2}$$

is distributed as an F random variable with d and d degrees of freedom, where $d = (n_1 + n_2 - 2)(m - 1)$. Hence, we reject the null hypothesis at the α level of significance if

$$T > F_{\alpha/2, d, d} \quad \text{or} \quad T < F_{1-\alpha/2, d, d}$$

Test for Noninferiority/Superiority

Under the null hypothesis of Eq. 5, $T = \hat{\sigma}_{WT}^2 / \delta \hat{\sigma}_{WT}^2$ is distributed as an F random variable with d and d degrees of freedom. Hence, we reject the null hypothesis at the α level of significance if

$$T < F_{1-\alpha, d, d}$$

Test for Similarity

Under the two one-sided hypotheses as discussed in Eq. 7, the following two test statistics

$$T_1 = \frac{\hat{\sigma}_{WT}^2}{\delta \hat{\sigma}_{WR}^2} \quad \text{and} \quad T_2 = \frac{\delta \hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2}$$

are used. We then reject the null hypothesis and conclude similarity at the α level of significance if

$$T_1 < F_{1-\alpha, d, d} \quad \text{and} \quad T_2 > F_{\alpha, d, d}$$

COMPARING INTRASUBJECT COEFFICIENTS OF VARIATION

In addition to comparing intrasubject variabilities, it is often of interest to compare intrasubject CV, which is a relative standard deviation adjusted by the mean. In recent years, the use of intrasubject CV has become increasingly popular. For example, the FDA defines highly variable drug products based on their intrasubject CVs. A drug product is said to be a highly variable drug if its intrasubject CV is greater than 30%. The intrasubject CV is also used as a measure for reproducibility of blood levels (or blood concentration–time curves) of a given formulation when the formulation is repeatedly administered at different dosing periods. In addition, the information regarding the intrasubject CV of a reference product is usually used for performing power analysis for sample size calculation in bioavailability and bioequivalence studies. In practice, two methods are commonly used for comparing intrasubject CVs. One is proposed by Chow and Tse,^[8] which is referred to as conditional random effects model. The other one is suggested by Quan and Shih,^[9] which is a simple one-way random effects model. In what follows, we will briefly introduce these two methods.

Simple Random Effects Model

Quan and Shih^[9] developed a method to estimate the intrasubject CV based on the simple mixed effects model Eq. 1 under which the intrasubject variability is a constant for all subjects in the same treatment. An intuitive unbiased estimator for μ_i is then given by

$$\hat{\mu}_i = \frac{1}{n_i m} \sum_{j=1}^{n_i} \sum_{k=1}^m x_{ijk}$$

Hence, an estimator of the intrasubject CV can be obtained by, for $i = T, R$,

$$\widehat{CV}_i = \frac{\hat{\sigma}_{Wi}}{\hat{\mu}_i}$$

By Taylor's expansion, it follows that

$$\begin{aligned} \widehat{CV}_i - CV_i &= \frac{\hat{\sigma}_{Wi}}{\hat{\mu}_i} - \frac{\sigma_{Wi}}{\mu_i} \\ &\approx \frac{1}{2\mu_i\sigma_{Wi}}(\hat{\sigma}_{Wi}^2 - \sigma_{Wi}^2) - \frac{\sigma_{Wi}}{\mu_i^2}(\hat{\mu}_i - \mu_i) \end{aligned}$$

Hence, by central limit theorem (CLT), $\widehat{CV}_i$ is asymptotically distributed as a normal random variable with mean CV_i and variance σ_i^{*2}/n_i , where

$$\begin{aligned} \sigma_i^{*2} &= \frac{\sigma_{Wi}^2}{2m\mu_i^2} + \frac{\sigma_{Wi}^4}{\mu_i^4} \\ &= \frac{1}{2m}CV_i^2 + CV_i^4 \end{aligned}$$

Thus, an intuitive estimator of σ_i^{*2} is given by

$$\hat{\sigma}_i^{*2} = \frac{1}{2m}\widehat{CV}_i^2 + \widehat{CV}_i^4$$

Test for Equality

The following hypotheses are usually considered for testing equality in intrasubject CVs

$$H_0: CV_T = CV_R \quad \text{vs.} \quad H_a: CV_T \neq CV_R \quad (11)$$

Under the null hypothesis, the test statistic

$$T = \frac{\widehat{CV}_T - \widehat{CV}_R}{\sqrt{\hat{\sigma}_T^{*2}/n_T + \hat{\sigma}_R^{*2}/n_R}}$$

is asymptotically distributed as a standard normal random variable. Hence, we reject the null hypothesis at the α level of significance if $|T| > z_{\alpha/2}$.

Test for Noninferiority/Superiority

Similarly, the problem of testing noninferiority and superiority can be unified by the following hypotheses:

$$H_0: CV_T - CV_R \geq \delta \quad \text{vs.} \quad H_a: CV_T - CV_R < \delta \quad (12)$$

where δ is the noninferiority/superiority margin. When $\delta > 0$, the rejection of the null hypothesis indicates the noninferiority of the test formulation over the reference formulation. When $\delta < 0$, the rejection of the null hypothesis indicates the superiority of the test formulation over the reference formulation.

Under the null hypothesis, the test statistic

$$T = \frac{\widehat{CV}_T - \widehat{CV}_R - \delta}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}}$$

is asymptotically distributed as a standard normal random variable. Hence, we reject the null hypothesis at the α level of significance if $T < -z_\alpha$.

Test for Similarity

For testing similarity, the following hypotheses are usually considered:

$$H_0: |\widehat{CV}_T - \widehat{CV}_R| \geq \delta \quad \text{vs.} \quad H_a: |\widehat{CV}_T - \widehat{CV}_R| < \delta \quad (13)$$

The two drug products are concluded similar to each other if the null hypothesis is rejected at a given significance level. The null hypothesis is rejected at the α level of significance if

$$\frac{\widehat{CV}_T - \widehat{CV}_R + \delta}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}} > z_\alpha \quad \text{and} \quad \frac{\widehat{CV}_T - \widehat{CV}_R - \delta}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}} < -z_\alpha$$

Conditional Random Effects Model

In practice, the variability of the observed response often increases as the mean increases. In many cases, the standard deviation of the intrasubject variability is approximately proportional to the mean value. To best describe this type of data, Chow and Tse^[8] proposed the following conditional random effects model.

$$x_{ijk} = A_{ij} + A_{ij} e_{ijk} \quad (14)$$

where x_{ijk} is the observation from the k th replicate ($k = 1, \dots, m$) of the j th subject ($j = 1, \dots, n_i$) from the i th treatment ($i = T, R$) and A_{ij} is the random effect due to the j th subject in the i th treatment. It is assumed that A_{ij} is normally distributed as a normal random variable with mean μ_i and variance σ_{Bi}^2 and e_{ijk} is normally distributed as a normal random variable with mean 0 and variance σ_{wi}^2 . For a given subject with a fixed A_{ij} , x_{ijk} is normally distributed as a normal random variable with mean A_{ij} and variance $A_{ij}^2 \sigma_{wi}^2$. Hence, the CV for this subject is given by

$$CV = \frac{|A_{ij} \sigma_{wi}|}{|A_{ij}|} = \sigma_{wi}$$

As can be seen, the conditional random effects model assumes the CV is a constant across subjects.

Define

$$\begin{aligned} \bar{x}_{i..} &= \frac{1}{nm} \sum_{j=1}^{n_i} \sum_{k=1}^m x_{ijk} \\ M_{i1} &= \frac{m}{n_i - 1} \sum_{j=1}^{n_i} (\bar{x}_{ij.} - \bar{x}_{i..})^2 \\ M_{i2} &= \frac{1}{n_i(m_i - 1)} \sum_{j=1}^{n_i} \sum_{k=1}^m (x_{ijk} - \bar{x}_{ij.})^2 \end{aligned}$$

It can be verified that

$$\begin{aligned} E(\bar{x}_{i..}) &= \mu_i \\ E(M_{i1}) &= (\mu_i^2 + \sigma_{BT}^2) \sigma_{wi}^2 + m \sigma_{Bi}^2 = \tau_{i1}^2 \\ E(M_{i2}) &= (\mu_i^2 + \sigma_{BT}^2) \sigma_{wi}^2 = \tau_{i2}^2 \end{aligned}$$

It follows that

$$CV_i = \sigma_{wi} = \sqrt{\frac{E(M_{i2})}{E^2(\bar{x}_{i..}) + (M_{i1} - M_{i2})}}$$

Hence, an estimator for CV_i is given by

$$\widehat{CV}_i = \sqrt{\frac{M_{i2}}{\bar{x}_{i..}^2 + (M_{i1} - M_{i2})/m}}$$

By Taylor's expansion,

$$\begin{aligned} \widehat{CV}_i - CV_i &\approx \frac{1}{\tau_{i1} \tau_{i2}} (M_{i2} - \tau_{i2}^2) \\ &\quad - \frac{\mu_i \tau_{i2}}{\tau_{i1}^3} (\bar{x}_{i..} - \mu_i) - \frac{\tau_{i2}}{2m \tau_{i1}^3} (M_{i1} - M_{i2} - (\tau_{i1}^2 - \tau_{i2}^2)) \\ &= k_0 (\bar{x}_{i..} - \mu_i) + k_1 (M_{i1} - \tau_{i1}^2) + k_2 (M_{i2} - \tau_{i2}^2) \end{aligned}$$

where

$$\begin{aligned} k_0 &= -\frac{\mu_i \tau_{i2}}{\tau_{i1}^3} \\ k_1 &= -\frac{\tau_{i2}}{2m \tau_{i1}^3} \\ k_2 &= \left(\frac{1}{\tau_{i1} \tau_{i2}} + \frac{\tau_{i2}}{2m \tau_{i1}^3} \right) \end{aligned}$$

As a result, $\widehat{CV}_i$'s distribution can be approximated by a normal random variable with mean CV_i and variance σ_i^{*2}/n_i where

$$\sigma_i^{*2} = \text{var} \left[k_0 \bar{x}_{ij.} + m k_1 (\bar{x}_{ij.} - \bar{x}_{i..})^2 + \frac{k_2}{m-1} \sum_{k=1}^m (x_{ijk} - \bar{x}_{ij.})^2 \right]$$

An intuitive estimator for σ_i^{*2} is the sample variance, denoted by $\hat{\sigma}_i^{*2}$, of

$$\begin{aligned} k_0 \bar{x}_{ij.} + m k_1 (\bar{x}_{ij.} - \bar{x}_{i..})^2 + \frac{k_2}{m-1} \sum_{k=1}^m (x_{ijk} - \bar{x}_{ij.})^2 \\ j = 1, \dots, n_i \end{aligned}$$

Test for Equality

Under the null hypothesis in Eq. 11, the test statistic

$$T = \frac{\widehat{CV}_T - \widehat{CV}_R}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}}$$

is asymptotically distributed as a standard normal random variable. Hence, we reject the null hypothesis at the α level of significance if $|T| > z_{\alpha/2}$.

Test for Noninferiority/Superiority

Under the null hypothesis of Eq. 12, the test statistic

$$T = \frac{\widehat{CV}_T - \widehat{CV}_R - \delta}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}}$$

is asymptotically distributed as a standard normal random variable. Hence, we reject the null hypothesis at the α level of significance if $T < -z_\alpha$.

Test for Similarity

For testing similarity, consider the hypotheses

$$H_0: |\widehat{CV}_T - \widehat{CV}_R| \geq \delta \quad \text{vs.} \quad H_a: |\widehat{CV}_T - \widehat{CV}_R| < \delta$$

The two drug products are concluded similar to each other if the null hypothesis is rejected at a given significance level. The null hypothesis is rejected at α level of significance if

$$\begin{aligned} \frac{\widehat{CV}_T - \widehat{CV}_R + \delta}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}} &> z_\alpha \quad \text{vs.} \\ \frac{\widehat{CV}_T - \widehat{CV}_R - \delta}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}} &< -z_\alpha \end{aligned}$$

COMPARING INTERSUBJECT VARIABILITIES

In addition to comparing intrasubject variabilities or intrasubject CVs, it is also of interest to compare intersubject variabilities. In practice, it is not uncommon that clinical results may not be reproducible from subject to subject within the target population or from subjects within the target population to subjects within a similar but slightly different population due to the intersubject variability. How to test a difference in intersubject and total variability between two treatments is a challenging problem to clinical scientists, especially biostatisticians, because of the following reasons. First, unbiased estimators of the intersubject and total variabilities are

usually not chi-square distributed under both parallel and crossover design with replicates. Second, the estimators for the intersubject and total variabilities under different treatments are usually not independent under a crossover design. As a result, unlike tests for comparing intrasubject variabilities, the standard F -test is not applicable. Tests for comparing intersubject variabilities under a parallel design can be performed by using the MLS method (see Refs. [2–5]). As indicated earlier, the MLS method is superior to many other approximation methods. Under crossover designs, however, the MLS method cannot be directly applied because estimators of variance components are not independent. Lee et al.^[6] proposed an extension of the MLS method when estimators of variance components are not independent. In addition, tests for comparing intersubject and total variabilities under crossover designs are studied by Lee et al.^[7] Note that the MLS method by Hyslop et al.^[5] is recommended as a statistical test for individual bioequivalence by the FDA.^[10]

Parallel Design with Replicates

Under model Eq. 1, for $i = T, R$, define

$$s_{Bi}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\bar{x}_{ij} - \bar{x}_{i..})^2 \quad (15)$$

where $\bar{x}_{ij}$ is given in Eq. 3 and

$$\bar{x}_{i..} = \frac{1}{n_i} \sum_{j=1}^{n_i} \bar{x}_{ij}$$

Note that $E(s_{Bi}^2) = \sigma_{Bi}^2 + \sigma_{wi}^2/m$. Therefore, the unbiased estimators for the intersubject variances are given by

$$\hat{\sigma}_{Bi}^2 = s_{Bi}^2 - \frac{1}{m} \hat{\sigma}_{wi}^2$$

where $\hat{\sigma}_{wi}^2$ is defined in Eq. 2.

Test for Equality

For testing equality in intersubject variability, the following hypotheses are usually considered

$$H_0: \frac{\sigma_{BT}^2}{\sigma_{BR}^2} = 1 \quad \text{vs.} \quad H_a: \frac{\sigma_{BT}^2}{\sigma_{BR}^2} \neq 1 \quad (16)$$

Testing the above hypotheses is equivalent to testing the following hypotheses

$$H_0: \sigma_{BT}^2 - \sigma_{BR}^2 = 0 \quad \text{vs.} \quad H_a: \sigma_{BT}^2 - \sigma_{BR}^2 \neq 0 \quad (17)$$

Let $\eta = \sigma_{BT}^2 - \sigma_{BR}^2$. An intuitive estimator of η is given by

$$\hat{\eta} = \hat{\sigma}_{BT}^2 - \hat{\sigma}_{BR}^2$$

It follows that

$$\begin{aligned}\hat{\eta} &= \hat{\sigma}_{BR}^2 - \hat{\sigma}_{BT}^2 \\ &= s_{BT}^2 - s_{BR}^2 - \hat{\sigma}_{WT}^2/m + \hat{\sigma}_{WR}^2/m\end{aligned}$$

A $(1 - \alpha) \times 100\%$ MLS confidence interval of η is given by

$$(\hat{\eta}_L, \hat{\eta}_U)$$

where

$$\hat{\eta}_L = \hat{\eta} - \sqrt{\Delta_L}, \quad \hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U}$$

and

$$\begin{aligned}\Delta_L &= s_{BT}^4 \left(1 - \frac{n_T - 1}{\chi_{\alpha/2, n_T - 1}^2}\right)^2 + s_{BR}^4 \left(1 - \frac{n_R - 1}{\chi_{1-\alpha/2, n_R - 1}^2}\right)^2 \\ &\quad + \frac{\hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_T(m-1)}{\chi_{1-\alpha/2, n_T(m-1)}^2}\right)^2 + \frac{\hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_R(m-1)}{\chi_{\alpha/2, n_R(m-1)}^2}\right)^2 \\ \Delta_U &= s_{BT}^4 \left(1 - \frac{n_T - 1}{\chi_{1-\alpha/2, n_T - 1}^2}\right)^2 + s_{BR}^4 \left(1 - \frac{n_R - 1}{\chi_{\alpha/2, n_R - 1}^2}\right)^2 \\ &\quad + \frac{\hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_T(m-1)}{\chi_{\alpha/2, n_T(m-1)}^2}\right)^2 + \frac{\hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_R(m-1)}{\chi_{1-\alpha/2, n_R(m-1)}^2}\right)^2\end{aligned}$$

We reject the null hypothesis at the α level of significance if $0 \notin (\hat{\eta}_L, \hat{\eta}_U)$.

Test for Noninferiority/Superiority

Similar to testing intrasubject variabilities, the problem of testing noninferiority/superiority can be unified by the following hypotheses

$$H_0: \frac{\sigma_{BT}^2}{\sigma_{BR}^2} \geq \delta \quad \text{vs.} \quad H_a: \frac{\sigma_{BT}^2}{\sigma_{BR}^2} < \delta \quad (18)$$

Testing the above hypotheses is equivalent to testing the following hypotheses

$$H_0: \sigma_{BT}^2 - \delta\sigma_{BR}^2 \geq 0 \quad \text{vs.} \quad H_1: \sigma_{BT}^2 - \delta\sigma_{BR}^2 < 0 \quad (19)$$

Define

$$\eta = \sigma_{BT}^2 - \delta\sigma_{BR}^2$$

For a given significance level α , similarly, the $(1 - \alpha) \times 100\%$ th MLS upper confidence bound of η can be constructed as

$$\hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U}$$

where Δ_U is given by

$$\begin{aligned}\Delta_U &= s_{BT}^4 \left(1 - \frac{n_T - 1}{\chi_{1-\alpha, n_T - 1}^2}\right)^2 + \delta^2 s_{BR}^4 \left(1 - \frac{n_R - 1}{\chi_{\alpha, n_R - 1}^2}\right)^2 \\ &\quad + \frac{\hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_T(m-1)}{\chi_{\alpha, n_T(m-1)}^2}\right)^2 \\ &\quad + \frac{\delta^2 \hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_R(m-1)}{\chi_{1-\alpha, n_R(m-1)}^2}\right)^2\end{aligned}$$

We then reject the null hypothesis at the α level of significance if $\hat{\eta}_U < 0$.

Test for Similarity

Similarly, consider the following hypotheses for establishment of similarity in variability:

$$H_0: \frac{\sigma_{BT}^2}{\sigma_{BR}^2} \notin \left(\frac{1}{\delta}, \delta\right) \quad \text{vs.} \quad H_a: \frac{\sigma_{BT}^2}{\sigma_{BR}^2} \in \left(\frac{1}{\delta}, \delta\right) \quad (20)$$

where $\delta > 1$ is the similarity limit. The above hypotheses can be decomposed into the following two one-sided hypotheses:

$$\begin{aligned}H_{01}: \sigma_{BT}^2 - \delta\sigma_{BR}^2 \geq 0 \quad \text{vs.} \quad H_{a1}: \sigma_{BT}^2 - \delta\sigma_{BR}^2 < 0 \\ H_{02}: \delta\sigma_{BT}^2 - \sigma_{BR}^2 \leq 0 \quad \text{vs.} \quad H_{a2}: \delta\sigma_{BT}^2 - \sigma_{BR}^2 > 0\end{aligned} \quad (21)$$

Similarity in intersubject variability between two treatments can be established if both of the above two hypotheses are rejected at the α level of significance. The test can be performed by constructing the $(1 - \alpha) \times 100\%$ upper confidence bound of $\eta_1 = \sigma_{BT}^2 - \delta\sigma_{BR}^2$ and the $(1 - \alpha) \times 100\%$ lower confidence bound of $\eta_2 = \delta\sigma_{BT}^2 - \sigma_{BR}^2$ by the MLS method. For example, the $(1 - \alpha) \times 100\%$ lower confidence bound of η_2 is given by

$$\hat{\eta}_{2L} = \hat{\eta}_2 - \sqrt{\Delta_L}$$

where

$$\begin{aligned}\Delta_L &= \delta^2 s_{BT}^4 \left(1 - \frac{n_T - 1}{\chi_{\alpha, n_T - 1}^2}\right)^2 + s_{BR}^4 \left(1 - \frac{n_R - 1}{\chi_{1-\alpha, n_R - 1}^2}\right)^2 \\ &\quad + \frac{\delta^2 \hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_T(m-1)}{\chi_{1-\alpha, n_T(m-1)}^2}\right)^2 + \frac{\hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_R(m-1)}{\chi_{\alpha, n_R(m-1)}^2}\right)^2\end{aligned}$$

Replicated Crossover Design

Under model Eq. 8 and new random variable $z_{ijk1} \equiv \bar{x}_{ijk}$ in Eq. 9, the intersubject sums of squares are defined by

$$\begin{aligned} s_{BT}^2 &= \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^2 \sum_{j=1}^{n_i} (\bar{x}_{ijT} - \bar{x}_{i.T})^2 \\ s_{BR}^2 &= \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^2 \sum_{j=1}^{n_i} (\bar{x}_{ijR} - \bar{x}_{i.R})^2 \end{aligned} \quad (22)$$

where

$$\bar{x}_{i.k.} = \frac{1}{n_i} \sum_{j=1}^{n_i} \bar{x}_{ijk}.$$

Note that $E(s_{Bk}^2) = \sigma_{Bk}^2 + \sigma_{wk}^2/m$ for $k = T, R$. Therefore, the unbiased estimators for the intersubject variance are given by

$$\hat{\sigma}_{BT}^2 = s_{BT}^2 - \frac{1}{m} \hat{\sigma}_{WT}^2 \quad (23)$$

$$\hat{\sigma}_{BR}^2 = s_{BR}^2 - \frac{1}{m} \hat{\sigma}_{WR}^2 \quad (24)$$

where $\hat{\sigma}_{WT}^2$ and $\hat{\sigma}_{WR}^2$ are defined in Eq. 10.

Test for Equality

For testing the equality in intersubject variability, the hypotheses in Eq. 17 are considered. Let $\eta = \sigma_{BT}^2 - \sigma_{BR}^2$. An intuitive estimator of η is given by

$$\hat{\eta} = \hat{\sigma}_{BT}^2 - \hat{\sigma}_{BR}^2$$

where $\hat{\sigma}_{BT}^2$ and $\hat{\sigma}_{BR}^2$ are given in Eqs. 23 and 24, respectively. It follows that

$$\begin{aligned} \hat{\eta} &= \hat{\sigma}_{BT}^2 - \hat{\sigma}_{BR}^2 \\ &= s_{BT}^2 - s_{BR}^2 - \frac{1}{m} \hat{\sigma}_{WT}^2 + \frac{1}{m} \hat{\sigma}_{WR}^2 \end{aligned}$$

Mean vector for the j th subject in i th sequence $(\bar{x}_{ijT}, \bar{x}_{ijR})'$ has a bivariate normal distribution with covariance matrix given by

$$\Omega_B = \begin{pmatrix} \sigma_{BT}^2 + \sigma_{WT}^2/m & \rho \sigma_{BT} \sigma_{BR} \\ \rho \sigma_{BT} \sigma_{BR} & \sigma_{BR}^2 + \sigma_{WR}^2/m \end{pmatrix} \quad (25)$$

The unbiased estimator of covariance matrix Ω_B can be obtained as

$$\hat{\Omega}_B = \begin{pmatrix} s_{BT}^2 & s_{BTR}^2 \\ s_{BTR}^2 & s_{BR}^2 \end{pmatrix} \quad (26)$$

where

$$s_{BTR}^2 = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^2 \sum_{j=1}^{n_i} (\bar{x}_{ijT} - \bar{x}_{i.T})(\bar{x}_{ijR} - \bar{x}_{i.R})$$

which is the sample covariance between $\bar{x}_{ijT}$ and $\bar{x}_{ijR}$. Let $\lambda_i, i = 1, 2$ be the two eigenvalues of the matrix $\Theta \hat{\Omega}_B$ where

$$\Theta = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad (27)$$

Hence, the population eigenvalues $\lambda_i, i = 1, 2$ can be estimated by the sample eigenvalues of $\Theta \hat{\Omega}_B$ and the explicit form of estimators is given as

$$\hat{\lambda}_i = \frac{s_{BT}^2 - s_{BR}^2 \pm \sqrt{(s_{BT}^2 + s_{BR}^2) - 4s_{BTR}^2}}{2} \quad \text{for } i = 1, 2$$

Without loss of generality, it can be assumed that $\hat{\lambda}_1$. By the extension of the MLS method in Ref. [7], a $(1 - \alpha) \times 100\%$ MLS confidence interval of η is given by

$$(\hat{\eta}_L, \hat{\eta}_U)$$

where

$$\hat{\eta}_L = \hat{\eta} - \sqrt{\Delta_L}, \quad \hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U}$$

and

$$\begin{aligned} \Delta_L &= \hat{\lambda}_1^2 \left(1 - \frac{n_s - 1}{\chi_{\alpha/2, n_s - 1}^2} \right)^2 + \hat{\lambda}_2^2 \left(1 - \frac{n_s - 1}{\chi_{1-\alpha/2, n_s - 1}^2} \right)^2 \\ &\quad + \frac{\hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_s(m-1)}{\chi_{\alpha/2, n_s(m-1)}^2} \right)^2 + \frac{\hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_s(m-1)}{\chi_{1-\alpha/2, n_s(m-1)}^2} \right)^2 \\ \Delta_U &= \hat{\lambda}_1^2 \left(1 - \frac{n_s - 1}{\chi_{1-\alpha/2, n_s - 1}^2} \right)^2 + \hat{\lambda}_2^2 \left(1 - \frac{n_s - 1}{\chi_{\alpha/2, n_s - 1}^2} \right)^2 \\ &\quad + \frac{\hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_s(m-1)}{\chi_{1-\alpha/2, n_s(m-1)}^2} \right)^2 + \frac{\hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_s(m-1)}{\chi_{\alpha/2, n_s(m-1)}^2} \right)^2 \end{aligned}$$

where $n_s = n_1 + n_2 - 2$. Then, we reject the null hypothesis at the α level of significance if $0 \notin (\hat{\eta}_L, \hat{\eta}_U)$.

Test for Noninferiority/Superiority

Similar to testing intrasubject variabilities, the problem of testing noninferiority/superiority can be unified by the hypotheses described in Eq. 19. For a given significance level of α , the $(1 - \alpha) \times 100\%$ MLS upper confidence bound of $\eta = \sigma_{BT}^2 - \delta \sigma_{BR}^2$ can be constructed as

$$\hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U}$$

where Δ_U is given by

$$\Delta_U = \hat{\lambda}_1^2 \left(1 - \frac{n_s - 1}{\chi_{1-\alpha, n_s-1}^2} \right)^2 + \hat{\lambda}_2^2 \left(1 - \frac{n_s - 1}{\chi_{\alpha, n_s-1}^2} \right)^2 + \frac{\hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_s(m-1)}{\chi_{1-\alpha, n_s(m-1)}^2} \right)^2 + \frac{\delta^2 \hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_s(m-1)}{\chi_{\alpha, n_s(m-1)}^2} \right)^2$$

where $n_s = n_1 + n_2 - 2$ and $\hat{\lambda}_1 < 0 < \hat{\lambda}_2$ are the eigenvalues of $\Theta \hat{\Omega}_B$ with

$$\Theta = \begin{pmatrix} 1 & 0 \\ 0 & -\delta \end{pmatrix} \quad (28)$$

Test for Similarity

Consider the hypotheses in Eq. 21 for establishment of similarity in intersubject variability. Similarity between two treatments can be established if both of the two hypotheses in Eq. 21 are rejected at the α level of significance. The test can be performed by constructing the $(1 - \alpha) \times 100\%$ upper confidence bound of $\eta_1 = \sigma_{BT}^2 - \delta \sigma_{BR}^2$ and the $(1 - \alpha) \times 100\%$ lower confidence bound of $\eta_2 = \delta \sigma_{BT}^2 - \sigma_{BR}^2$ by the MLS method. For example, the $(1 - \alpha) \times 100\%$ lower confidence bound of η_2 is given by

$$\hat{\eta}_{2L} = \hat{\eta}_2 - \sqrt{\Delta_L}$$

where

$$\Delta_L = \hat{\lambda}_1^2 \left(1 - \frac{n_s - 1}{\chi_{\alpha, n_s-1}^2} \right)^2 + \hat{\lambda}_2^2 \left(1 - \frac{n_s - 1}{\chi_{1-\alpha, n_s-1}^2} \right)^2 + \frac{\delta^2 \hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_s(m-1)}{\chi_{\alpha, n_s(m-1)}^2} \right)^2 + \frac{\hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_s(m-1)}{\chi_{1-\alpha, n_s(m-1)}^2} \right)^2$$

where $n_s = n_1 + n_2 - 2$ and $\hat{\lambda}_1 < 0 < \hat{\lambda}_2$ are the eigenvalues of $\Theta \hat{\Omega}_B$ with

$$\Theta = \begin{pmatrix} \delta & 0 \\ 0 & -1 \end{pmatrix} \quad (29)$$

COMPARING TOTAL VARIABILITIES

In practice, it may be also of interest to compare total variabilities between drug products. Total variability is related to the drug's prescribability,^[10] which is the main concept for population bioequivalence. Because the total variability can be estimated under parallel/crossover designs both with and without replicates, we will consider both designs with and without replicates.

Parallel Designs Without Replicates

The following model can be used for a parallel design without replicates

$$x_{ij} = \mu_i + \epsilon_{ij}$$

where x_{ij} is the observation from the j th subject in the i th treatment group. Also we assume that the random variable ϵ_{ij} has normal distribution with mean 0 and variance σ_{Ti}^2 for $i = T, R$. Hence, the total variability can be estimated by

$$\hat{\sigma}_{Ti}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2$$

where

$$\bar{x}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

Test for Equality

For testing equality in total variability, the following hypotheses are considered:

$$H_0: \frac{\sigma_{TT}^2}{\sigma_{TR}^2} = 1 \quad \text{vs.} \quad H_a: \frac{\sigma_{TT}^2}{\sigma_{TR}^2} \neq 1 \quad (30)$$

Under the null hypothesis, the test statistic

$$T = \frac{\hat{\sigma}_{TT}^2}{\hat{\sigma}_{TR}^2}$$

is distributed as an F random variable with $n_T - 1$ and $n_R - 1$ degrees of freedom. Hence, we reject the null hypothesis at the α level of significance if

$$T > F_{\alpha/2, n_T, n_R}$$

or

$$T < F_{1-\alpha/2, n_T, n_R}$$

Test for Noninferiority/Superiority

Similarly, the problem of testing noninferiority and superiority can be unified by the following hypotheses

$$H_0: \frac{\sigma_{TT}^2}{\sigma_{TR}^2} \geq \delta \quad \text{vs.} \quad H_a: \frac{\sigma_{TT}^2}{\sigma_{TR}^2} < \delta \quad (31)$$

The test statistic is given by

$$T = \frac{s_{TT}^2}{\delta s_{TR}^2}$$

Under the null hypothesis, T is distributed as an F random variable with $n_T - 1$ and $n_R - 1$ degrees of freedom. Hence, we reject the null hypothesis at the α level of significance if

$$T < F_{1-\alpha, n_R, n_T}$$

Test for Similarity

For testing similarity, the hypotheses of interest are given by

$$H_0: \frac{\sigma_{TT}^2}{\sigma_{TR}^2} \notin \left(\frac{1}{\delta}, \delta\right) \quad \text{vs.} \quad H_a: \frac{\sigma_{TT}^2}{\sigma_{TR}^2} \in \left(\frac{1}{\delta}, \delta\right) \quad (32)$$

where $\delta > 1$ is the similarity limit. The above hypotheses can be decomposed into the following two one-sided hypotheses

$$\begin{aligned} H_{01}: \frac{\sigma_{TT}^2}{\sigma_{TR}^2} \geq \delta \quad \text{vs.} \quad H_{a1}: \frac{\sigma_{TT}^2}{\sigma_{TR}^2} < \delta \\ H_{02}: \frac{\sigma_{TT}^2}{\sigma_{TR}^2} \leq \frac{1}{\delta} \quad \text{vs.} \quad H_{a2}: \frac{\sigma_{TT}^2}{\sigma_{TR}^2} > \frac{1}{\delta} \end{aligned} \quad (33)$$

These two hypotheses can be tested by the following two test statistics

$$T_1 = \frac{s_{TT}^2}{\delta s_{TR}^2} \quad \text{and} \quad T_2 = \frac{\delta s_{TT}^2}{s_{TR}^2}$$

We then reject the null hypothesis and conclude similarity at the α level of significance if

$$T_1 < F_{1-\alpha, n_T, n_R} \quad \text{and} \quad T_2 > F_{\alpha, n_T, n_R}$$

Parallel Design with Replicates

In practice, a parallel design with replicates can also be used to assess total variability. Although total variabilities can be assessed by the design without replicates, there is a merit of the design with replicates. Because intersubject and intrasubject variabilities can be assessed under a parallel design with replicates, it can serve more than just one purpose. If we consider model Eq. 1, an unbiased estimator for total variabilities is given by

$$\hat{\sigma}_{Ti}^2 = s_{Bi}^2 + \frac{m-1}{m} \hat{\sigma}_{wi}^2$$

where s_{Bi}^2 is defined in Eq. 15 and $\hat{\sigma}_{wi}^2$ is defined in Eq. (2). Let $\eta = \sigma_{TT}^2 - \sigma_{TR}^2$. Then, a natural estimator for η is given by

$$\hat{\eta} = \hat{\sigma}_{TT}^2 - \hat{\sigma}_{TR}^2$$

Test for Equality

For testing the equality in total variability, consider the hypotheses in Eq. 30, which is equivalent to testing the following hypotheses:

$$H_0: \sigma_{TT}^2 - \sigma_{TR}^2 = 0 \quad \text{vs.} \quad H_a: \sigma_{TT}^2 - \sigma_{TR}^2 \neq 0 \quad (34)$$

A $(1 - \alpha) \times 100\%$ MLS confidence interval of η is given by

$$(\hat{\eta}_L, \hat{\eta}_U)$$

where

$$\hat{\eta}_L = \hat{\eta} - \sqrt{\Delta_L}, \quad \hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U}$$

and

$$\begin{aligned} \Delta_L &= s_{BT}^4 \left(1 - \frac{n_T - 1}{\chi_{\alpha/2, n_T-1}^2}\right)^2 + s_{BR}^4 \left(1 - \frac{n_R - 1}{\chi_{1-\alpha/2, n_R-1}^2}\right)^2 \\ &\quad + \frac{(m-1)^2 \hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_T(m-1)}{\chi_{1-\alpha/2, n_T(m-1)}^2}\right)^2 \\ &\quad + \frac{(m-1)^2 \hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_R(m-1)}{\chi_{\alpha/2, n_R(m-1)}^2}\right)^2 \\ \Delta_U &= s_{BT}^4 \left(1 - \frac{n_T - 1}{\chi_{1-\alpha/2, n_T-1}^2}\right)^2 + s_{BR}^4 \left(1 - \frac{n_R - 1}{\chi_{\alpha/2, n_R-1}^2}\right)^2 \\ &\quad + \frac{(m-1)^2 \hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_T(m-1)}{\chi_{\alpha/2, n_T(m-1)}^2}\right)^2 \\ &\quad + \frac{(m-1)^2 \hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_R(m-1)}{\chi_{1-\alpha/2, n_R(m-1)}^2}\right)^2 \end{aligned}$$

We reject the null hypothesis at the α level of significance if $0 \notin (\hat{\eta}_L, \hat{\eta}_U)$.

Test for Noninferiority/Superiority

Similar to testing intrasubject variabilities, the problem of testing noninferiority/superiority can be unified by the hypotheses described in Eq. 31. Testing the hypotheses in Eq. 31 is equivalent to testing the following hypotheses:

$$H_0: \sigma_{TT}^2 - \delta \sigma_{TR}^2 \geq 0 \quad \text{vs.} \quad H_a: \sigma_{TT}^2 - \delta \sigma_{TR}^2 < 0 \quad (35)$$

Let $\eta = \sigma_{TT}^2 - \delta \sigma_{TR}^2$. For a given significance level of α , similarly, the $(1 - \alpha)100\%$ upper confidence bound of η can be constructed as

$$\hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U}$$

where $\hat{\eta} = \hat{\sigma}_{\text{TT}}^2 - \hat{\sigma}_{\text{TR}}^2$ and Δ_U is given by

$$\begin{aligned}\Delta_U = & s_{\text{BT}}^4 \left(1 - \frac{n_{\text{T}} - 1}{\chi_{1-\alpha, n_{\text{T}}-1}^2} \right)^2 + \delta^2 s_{\text{BR}}^4 \left(1 - \frac{n_{\text{R}} - 1}{\chi_{\alpha/2, n_{\text{R}}-1}^2} \right)^2 \\ & + \frac{(m-1)^2 \hat{\sigma}_{\text{WT}}^4}{m^2} \left(1 - \frac{n_{\text{T}}(m-1)}{\chi_{\alpha, n_{\text{T}}(m-1)}^2} \right)^2 \\ & + \frac{\delta^2 (m-1)^2 \hat{\sigma}_{\text{WR}}^4}{m^2} \left(1 - \frac{n_{\text{R}}(m-1)}{\chi_{1-\alpha, n_{\text{R}}(m-1)}^2} \right)^2\end{aligned}$$

We then reject the null hypothesis at the α level of significance if $\hat{\eta}_U < 0$.

Test for Similarity

For testing similarity, the hypotheses of interest are described in Eq. 33. We then reject the null hypothesis and conclude similarity at the α level of significance if both hypotheses in Eq. 33 are rejected with significance level α . The test can be performed by constructing the $(1 - \alpha) \times 100\%$ upper confidence bound of $\eta_1 = \sigma_{\text{TT}}^2 - \delta \sigma_{\text{TR}}^2$ and the $(1 - \alpha) \times 100\%$ lower confidence bound of $\eta_2 = \sigma_{\text{TT}}^2 - \delta \sigma_{\text{TR}}^2$ by the MLS method. For example, the $(1 - \alpha) \times 100\%$ lower confidence bound of η_2 is given by

$$\hat{\eta}_{2\text{L}} = \hat{\eta}_2 - \sqrt{\Delta_{\text{L}}}$$

Where

$$\begin{aligned}\Delta_{\text{L}} = & \delta^2 s_{\text{BT}}^4 \left(1 - \frac{n_{\text{T}} - 1}{\chi_{\alpha, n_{\text{T}}-1}^2} \right)^2 + s_{\text{BR}}^4 \left(1 - \frac{n_{\text{R}} - 1}{\chi_{1-\alpha, n_{\text{R}}-1}^2} \right)^2 \\ & + \frac{(m-1)^2 \delta^2 \hat{\sigma}_{\text{WT}}^4}{m^2} \left(1 - \frac{n_{\text{T}}(m-1)}{\chi_{1-\alpha, n_{\text{T}}(m-1)}^2} \right)^2 \\ & + \frac{(m-1)^2 \hat{\sigma}_{\text{WR}}^4}{m^2} \left(1 - \frac{n_{\text{R}}(m-1)}{\chi_{\alpha, n_{\text{R}}(m-1)}^2} \right)^2\end{aligned}$$

The Standard 2×2 Crossover Design

For the standard 2×2 crossover design, model Eq. 8 is still useful if the index for replicates l is omitted (i.e., $m = 1$). Unbiased estimators for the total variabilities are given by

$$\begin{aligned}\hat{\sigma}_{\text{TT}}^2 = s_{\text{BT}}^2 &= \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^2 \sum_{j=1}^{n_i} (x_{ij\text{T}} - \bar{x}_{i\text{T}})^2 \\ \hat{\sigma}_{\text{TR}}^2 = s_{\text{BR}}^2 &= \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^2 \sum_{j=1}^{n_i} (x_{ij\text{R}} - \bar{x}_{i\text{R}})^2\end{aligned}$$

where

$$\bar{x}_{i\text{T}} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij\text{T}} \quad \text{and} \quad \bar{x}_{i\text{R}} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij\text{R}}$$

Test for Equality

For testing the equality in total variability, consider the hypotheses in 34. Let $\eta = \sigma_{\text{TT}}^2 - \sigma_{\text{TR}}^2$. An intuitive estimator of η is given by

$$\hat{\eta} = \hat{\sigma}_{\text{TT}}^2 - \hat{\sigma}_{\text{TR}}^2$$

Let

$$s_{\text{BTR}}^2 = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^2 \sum_{j=1}^{n_i} (x_{ij\text{T}} - \bar{x}_{i\text{T}})(x_{ij\text{R}} - \bar{x}_{i\text{R}})$$

Let $\hat{\lambda}_1 < 0 < \hat{\lambda}_2$ be the two eigenvalues of the matrix $\Theta \hat{\Omega}_{\text{B}}$ where $\hat{\Omega}_{\text{B}}$ is defined in Eq. 26 and Θ is given in Eq. 27. A $(1 - \alpha) \times 100\%$ MLS confidence interval of η is given by

$$(\hat{\eta}_{\text{L}}, \hat{\eta}_{\text{U}})$$

where

$$\hat{\eta}_{\text{L}} = \hat{\eta} - \sqrt{\Delta_{\text{L}}}, \quad \hat{\eta}_{\text{U}} = \hat{\eta} + \sqrt{\Delta_{\text{U}}}$$

and

$$\begin{aligned}\Delta_{\text{L}} &= \hat{\lambda}_1^2 \left(1 - \frac{n_1 + n_2 - 2}{\chi_{1-\alpha/2, n_1+n_2-2}^2} \right)^2 + \hat{\lambda}_2^2 \left(1 - \frac{n_1 + n_2 - 2}{\chi_{\alpha/2, n_1+n_2-2}^2} \right)^2 \\ \Delta_{\text{U}} &= \hat{\lambda}_1^2 \left(1 - \frac{n_1 + n_2 - 2}{\chi_{\alpha/2, n_1+n_2-2}^2} \right)^2 + \hat{\lambda}_2^2 \left(1 - \frac{n_1 + n_2 - 2}{\chi_{1-\alpha/2, n_1+n_2-2}^2} \right)^2\end{aligned}$$

Test for Noninferiority/Superiority

Similar to testing intrasubject variabilities, consider the hypotheses in Eq. 35. For a given significance level of α , similarly, the $(1 - \alpha) \times 100\%$ upper confidence bound of $\eta = \sigma_{\text{TT}}^2 - \sigma_{\text{TR}}^2$ can be constructed as

$$\hat{\eta}_{\text{U}} = \hat{\eta} + \sqrt{\Delta_{\text{U}}}$$

where Δ_{U} is given by

$$\Delta_{\text{U}} = \hat{\lambda}_1^2 \left(\frac{n_1 + n_2 - 2}{\chi_{\alpha, n_1+n_2-2}^2} - 1 \right)^2 + \hat{\lambda}_2^2 \left(\frac{n_1 + n_2 - 2}{\chi_{1-\alpha, n_1+n_2-2}^2} - 1 \right)^2$$

and $\hat{\lambda}_1 < 0 < \hat{\lambda}_2$ be the two eigenvalues of the matrix $\Theta \hat{\Omega}_{\text{B}}$ where Θ is given in Eq. 28. We then reject the null hypothesis at the α level of significance if $\eta_{\text{U}} < 0$.

Test for Similarity

For testing similarity, the hypotheses of interest are described in Eq. 33. We then reject the null hypothesis and conclude similarity at the α level of significance if both hypotheses in Eq. 33 are rejected with significance level α .

The test can be performed by constructing the $(1 - \alpha) \times 100\%$ upper confidence bound of $\eta_1 = \sigma_{TT}^2 - \delta\sigma_{TR}^2$ and the $(1 - \alpha) \times 100\%$ lower confidence bound of $\eta_2 = \sigma_{TT}^2 - \delta\sigma_{TR}^2$ by the MLS method. For example, the $(1 - \alpha) \times 100\%$ lower confidence bound of η_2 is given by

$$\hat{\eta}_{2L} = \hat{\eta}_2 + \sqrt{\Delta_L}$$

where

$$\Delta_L = \hat{\lambda}_1^2 \left(1 - \frac{n_1 + n_2 - 2}{\chi_{1-\alpha, n_1+n_2-2}^2} \right)^2 + \hat{\lambda}_2^2 \left(1 - \frac{n_1 + n_2 - 2}{\chi_{\alpha, n_1+n_2-2}^2} \right)^2$$

where $\hat{\lambda}_1 < 0 < \hat{\lambda}_2$ are the two eigenvalues of the matrix $\Theta \hat{\Omega}_B$ where Θ is given in Eq. 29.

Replicated 2 x 2m Crossover Design

We can use a similar argument for the test of intersubject variabilities under model Eq. 8 with the estimators, for $k = T, R$,

$$\hat{\sigma}_{Tk}^2 = s_{Bk}^2 + \frac{m-1}{m} \hat{\sigma}_{wk}^2$$

where $\hat{s}_{Bk}^2$ and $\hat{\sigma}_{wk}^2$ are defined in Eq. 22 and Eq. 10, respectively.

Test for Equality

For testing the equality in total variability, consider the hypotheses in Eq. 34. An intuitive estimator of $\eta = \sigma_{TT}^2 - \sigma_{TR}^2$ is given by $\hat{\eta} = \hat{\sigma}_{TT}^2 - \hat{\sigma}_{TR}^2$. Let $\hat{\lambda}_1 < 0 < \hat{\lambda}_2$ be the two eigenvalues of the matrix $\Theta \hat{\Omega}_B$ where Θ is given in Eq. 27. By Lee et al.^[7], a $(1 - \alpha) \times 100\%$ confidence interval of η is given by

$$(\hat{\eta}_L, \hat{\eta}_U)$$

where

$$\hat{\eta}_L = \hat{\eta} - \sqrt{\Delta_L}, \quad \hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U}$$

and

$$\begin{aligned} \Delta_L = & \hat{\lambda}_1^2 \left(1 - \frac{n_s - 1}{\chi_{1-\alpha/2, n_s-1}^2} \right)^2 + \hat{\lambda}_2^2 \left(1 - \frac{n_s - 1}{\chi_{\alpha/2, n_s-1}^2} \right)^2 \\ & + \frac{(m-1)^2 \hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_s(m-1)}{\chi_{\alpha/2, n_s(m-1)}^2} \right)^2 \\ & + \frac{(m-1)^2 \hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_s(m-1)}{\chi_{1-\alpha/2, n_s(m-1)}^2} \right)^2 \end{aligned}$$

$$\begin{aligned} \Delta_U = & \hat{\lambda}_1^2 \left(1 - \frac{n_s - 1}{\chi_{\alpha/2, n_s-1}^2} \right)^2 + \hat{\lambda}_2^2 \left(1 - \frac{n_s - 1}{\chi_{1-\alpha/2, n_s-1}^2} \right)^2 \\ & + \frac{(m-1)^2 \hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_s(m-1)}{\chi_{1-\alpha/2, n_s(m-1)}^2} \right)^2 \\ & + \frac{(m-1)^2 \hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_s(m-1)}{\chi_{\alpha/2, n_s(m-1)}^2} \right)^2 \end{aligned}$$

We reject the null hypothesis at the α level of significance if $0 \notin (\hat{\eta}_L, \hat{\eta}_U)$.

Test for Noninferiority/Superiority

Consider a test of the hypotheses in Eq. 35. Let $\eta_2 = \sigma_{TT}^2 - \delta\sigma_{TR}^2$. For a given significance level of α , similarly, the $(1 - \alpha) \times 100\%$ upper confidence bound of η can be constructed as

$$\hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U}$$

where $\hat{\eta} = \hat{\sigma}_{TT}^2 - \delta\sigma_{TR}^2$ and Δ_U is given by

$$\begin{aligned} \Delta_U = & \hat{\lambda}_1^2 \left(1 - \frac{n_s - 1}{\chi_{\alpha, n_s-1}^2} \right)^2 + \hat{\lambda}_2^2 \left(1 - \frac{n_s - 1}{\chi_{1-\alpha, n_s-1}^2} \right)^2 \\ & + \frac{(m-1)^2 \hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_s(m-1)}{\chi_{1-\alpha, n_s(m-1)}^2} \right)^2 \\ & + \frac{(m-1)^2 \hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_s(m-1)}{\chi_{\alpha, n_s(m-1)}^2} \right)^2 \end{aligned}$$

where $n_s = n_1 + n_2 - 2$ and $\hat{\lambda}_1 < 0 < \hat{\lambda}_2$ are the eigenvalues of $\Theta \hat{\Omega}_B$ with Θ given in Eq. 28. We then reject the null hypothesis at the α level of significance if $\hat{\eta}_U$.

Test for Similarity

For testing similarity, the hypotheses of interest are described in Eq. 33. We then reject the null hypothesis and conclude similarity at the α level of significance if both hypotheses in Eq. 33 are rejected with significance level α . The test can be performed by constructing the $(1 - \alpha) \times 100\%$ upper confidence bound of $\eta_1 = \sigma_{TT}^2 - \delta\sigma_{TR}^2$ and the $(1 - \alpha) \times 100\%$ lower confidence bound of $\eta_2 = \sigma_{TT}^2 - \delta\sigma_{TR}^2$ by the MLS method. For example, the $(1 - \alpha) \times 100\%$ lower confidence bound of η_2 is given by

$$\hat{\eta}_{2L} = \hat{\eta}_2 - \sqrt{\Delta_L}$$

where

$$\Delta_L = \hat{\lambda}_1^2 \left(1 - \frac{n_s - 1}{\chi_{1-\alpha, n_s-1}^2} \right)^2 + \hat{\lambda}_2^2 \left(1 - \frac{n_s - 1}{\chi_{\alpha, n_s-1}^2} \right)^2 \\ + \frac{(m-1)^2 \hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_s(m-1)}{\chi_{\alpha, n_s(m-1)}^2} \right)^2 \\ + \frac{(m-1)^2 \hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_s(m-1)}{\chi_{1-\alpha, n_s(m-1)}^2} \right)^2$$

where $\hat{\lambda}_1 < 0 < \hat{\lambda}_2$ be the two eigenvalues of the matrix $\Theta \hat{\Omega}_B$ where Θ is given in Eq. 29.

SAMPLE SIZE CALCULATION

For clinical scientists/biostatisticians, a great concern is how many subjects should be recruited in the clinical trial to achieve the desired power of statistical test, which is used to show equality, superiority, or similarity of the test formulation. If an explicit power function of a test is not available or difficult to obtain, sample size calculation depends on asymptotic power, which is usually obtained by using the central limit theorem. Also, some assumption about unknown parameters should be made to calculate sample size. These assumptions may come from previous or related clinical studies that can provide reasonable estimates for unknown parameters. If information about parameters is not available, it is common to assume the worst-case scenario so that sample size is big enough to assure that a clinical study will not fail to achieve its objective. In practice, however, it is difficult to obtain a large sample size because a clinical study is restricted by time and budget. Here, we introduce some statistical techniques to calculate sample size for comparing variabilities to achieve a desired power under some assumption about unknown parameters.

Sample Size for *F*-Test: Comparing Intrasubject Variabilities

If a test is based on *F* distribution, the explicit power function can be obtained so that fairly accurate sample size can be provided. Also, for most cases, the target sample size is available as an implicit solution of the equation so that no explicit formula for sample size is available. However, the target sample size can be obtained by numerical methods such as the grid search method or Newton–Raphson algorithm.

First, consider the hypotheses in Eq. 4 for test for equality of intrasubject variabilities under a parallel design with replicates in “Parallel Design with Replicates” under the section “Comparing Intrasubject Variabilities.” Under the alternative hypothesis, without loss of generality, we assume that $\sigma_{WT}^2 < \sigma_{WR}^2$. The power of the test can then be approximated by

$$1 - \beta = P(T < F_{1-\alpha/2, n_T(m-1), n_R(m-1)}) \\ = P(1/T > F_{\alpha/2, n_R(m-1), n_T(m-1)}) \\ = P\left(\frac{\hat{\sigma}_{WR}^2 / \sigma_{WR}^2}{\hat{\sigma}_{WT}^2 / \sigma_{WT}^2} > \frac{\sigma_{WT}^2}{\sigma_{WR}^2} F_{\alpha/2, n_R(m-1), n_T(m-1)}\right) \\ = P\left(F_{(n_R(m-1), n_T(m-1))} > \frac{\sigma_{WT}^2}{\sigma_{WR}^2} F_{\alpha/2, n_R(m-1), n_T(m-1)}\right)$$

where $F_{(a,b)}$ denotes an *F* random variable with *a* and *b* degrees of freedom. Under the assumption that $n = n_R = n_T$ and with a fixed value for *m*, σ_{WT}^2 , and σ_{WR}^2 , the sample size needed to achieve a desired power of $1 - \beta$ can be obtained by solving the following equation for *n*:

$$\frac{\sigma_{WT}^2}{\sigma_{WR}^2} = \frac{F_{1-\beta, n(m-1), n(m-1)}}{F_{\alpha/2, n(m-1), n(m-1)}}$$

Next, consider the hypotheses in Eq. 6 for the test for similarity of intrasubject variabilities under a parallel design with replicates. Assuming that $n = n_T = n_R$, under the alternative hypothesis that $\sigma_{WT}^2 = \sigma_{WR}^2$, the power of the test can be approximated by

$$1 - \beta = P\left(\frac{F_{\alpha, n(m-1), n(m-1)}}{\delta} < \frac{\hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2} < \delta F_{1-\alpha, n(m-1), n(m-1)}\right) \\ = P\left(\frac{1}{F_{\alpha, n(m-1), n(m-1)}} < \frac{\hat{\sigma}_{WT}^2}{\delta \hat{\sigma}_{WR}^2} < \delta F_{1-\alpha, n(m-1), n(m-1)}\right) \\ = 1 - 2P\left(\frac{\hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2} > \delta F_{1-\alpha, n(m-1), n(m-1)}\right) \\ = 1 - 2P\left(F_{(n(m-1), n(m-1))} > \delta F_{1-\alpha, n(m-1), n(m-1)}\right)$$

Thus, the sample size required for achieving a desired power of $1 - \beta$ can be obtained by solving the following equation for *n*

$$\delta = \frac{F_{\beta/2, n(m-1), n(m-1)}}{F_{1-\alpha, n(m-1), n(m-1)}}$$

Note that under the alternative hypothesis, it is also possible that $\sigma_{WT}^2 \neq \sigma_{WR}^2$. Without loss of generality, we assume that $\sigma_{WT}^2 / \sigma_{WR}^2 = r \in (1, \delta)$. In this case, the power of the above test can be approximated by

$$1 - \beta = P\left(\frac{F_{\alpha, n(m-1), n(m-1)}}{\delta} < \frac{\hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2} < \delta F_{1-\alpha, n(m-1), n(m-1)}\right) \\ \approx P\left(\frac{\hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2} < \delta F_{1-\alpha, n(m-1), n(m-1)}\right) \\ = P\left(\frac{\hat{\sigma}_{WT}^2}{r \hat{\sigma}_{WR}^2} < \frac{\delta}{r} F_{1-\alpha, n(m-1), n(m-1)}\right) \\ = P\left(F_{(n(m-1), n(m-1))} < \frac{\delta}{r} F_{1-\alpha, n(m-1), n(m-1)}\right)$$

Hence, the sample size needed to achieve the desired power of $1 - \beta$ can be obtained by solving the following equation for n :

$$\frac{\delta}{r} = \frac{F_{\beta, n(m-1), n(m-1)}}{F_{1-\alpha, n(m-1), n(m-1)}}$$

Sample Size for Z-Test: Comparing Intrasubject Coefficients of Variation

If a test is already based on the normal approximation of the test statistic, then sample size can be easily obtained by simple calculation of power function, which is based on the normal distribution. In this case, the explicit form of target sample size is usually available.

Here, we consider the test for comparing intrasubject CVs under the section “Simple Random Effects Model.” First, consider the hypothesis in Eq. 11 for the test of equality of intrasubject CVs. Under the alternative hypothesis, without loss of generality, it is assumed that $CV_T > CV_R$. The distribution of test statistics can be approximated by a normal distribution with unit variance and mean

$$\frac{CV_T - CV_R}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}}$$

Thus, the power is given by

$$\begin{aligned} 1 - \beta &= P(|T| > z_{\alpha/2}) \approx P(T > z_{\alpha/2}) \\ &= P\left(N(0,1) > z_{\alpha/2} - \frac{CV_T - CV_R}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}}\right) \end{aligned}$$

Under the assumption that $n = n_1 = n_2$, the sample size needed to have a power of $1 - \beta$ can be obtained by solving the following equation

$$z_{\alpha/2} - \frac{CV_T - CV_R}{\sqrt{\sigma_T^{*2}/n + \sigma_R^{*2}/n}} = -z_\beta$$

This leads to

$$n = \frac{(\sigma_T^{*2} + \sigma_R^{*2})(z_{\alpha/2} + z_\beta)^2}{(CV_T - CV_R)^2}$$

Next, consider the hypothesis in Eq. 13 for the test of similarity of intrasubject CVs under simple random effects model. Under the alternative hypothesis that $CV_T = CV_R$, the power of the test procedure is given by

$$\begin{aligned} 1 - \beta &= P\left(z_\alpha - \frac{\delta}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}}\right. \\ &\quad \left.< N(0,1) < \frac{\delta}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}} - z_\alpha\right) \\ &= 1 - 2P\left(N(0,1) > \frac{\delta}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}} - z_\alpha\right) \end{aligned}$$

Hence, under the assumption that $n = n_1 = n_2$, the sample size needed to achieve $1 - \beta$ power at the α level of significance can be obtained by solving the following equation

$$\frac{\delta}{\sqrt{\sigma_T^{*2}/n + \sigma_R^{*2}/n}} - z_\alpha = z_{\beta/2}$$

This leads to

$$n = \frac{(z_\alpha + z_{\beta/2})^2(\sigma_T^{*2} + \sigma_R^{*2})}{\delta^2}$$

Under the alternative, without loss of generality, we assume that $0 < CV_T - CV_R < \delta$. Then, the power of the test procedure can be approximated by

$$\begin{aligned} 1 - \beta &= P\left(z_\alpha - \frac{\delta + |CV_T - CV_R|}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}} < N(0,1)\right. \\ &\quad \left.< \frac{\delta - |CV_T - CV_R|}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}} - z_\alpha\right) \\ &\approx P\left(N(0,1) < \frac{\delta - |CV_T - CV_R|}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}} - z_\alpha\right) \end{aligned}$$

Hence, under the assumption that $n = n_1 = n_2$, the sample size needed to achieve $1 - \beta$ power at the α level of significance can be obtained by solving the following equation

$$\frac{\delta - |CV_T - CV_R|}{\sqrt{\sigma_T^{*2}/n_T + \sigma_R^{*2}/n_R}} - z_\alpha = z_\beta$$

This gives

$$n = \frac{(z_\alpha + z_\beta)^2(\sigma_T^{*2} + \sigma_R^{*2})}{(\delta - |CV_T - CV_R|)^2}$$

Sample Size for Modified Large Sample Method: Comparing Intersubject Variabilities

If a test is based on the MLS method, the explicit power function is difficult to calculate so that approximated power function should be obtained by using CLT to find the target sample size. The crucial step for normal approximation is the calculation of the asymptotic variance of a test statistic. The asymptotic variance of test statistics based on the MLS method can be easily calculated because estimators of variance components are usually distributed as chi-square distribution. When estimators of variance components are not independent, the asymptotic variance can be easily obtained via the eigenvalues of $\Theta\Omega_B$ where Ω_B is a 2×2 intersubject covariance matrix and Θ is some 2×2 matrix.

We will consider the test of intersubject variabilities in “Parallel Design Replicates” under the section “Comparing Intersubject Variabilities.” Consider the hypothesis in Eq. 11 for the test of equality of intersubject variabilities. Under the alternative hypothesis, without loss of generality, we assume that $\sigma_{BR}^2 > \sigma_{BT}^2$ and $n = n_T = n_R$. Thus, the power of the test procedure can be approximated by

$$P\left(N\left(\frac{\sqrt{n}(\sigma_{BR}^2 - \sigma_{BT}^2)}{\sigma^*}, 1\right) > z_{\alpha/2}\right)$$

where asymptotic variance σ^* of the test statistic is given by

$$\sigma^{*2} = 2\left[\left(\sigma_{BT}^2 + \frac{\sigma_{WT}^2}{m}\right)^2 + \left(\sigma_{BR}^2 + \frac{\sigma_{WR}^2}{m}\right)^2 + \frac{\sigma_{WT}^4}{m^2(m-1)} + \frac{\sigma_{WR}^4}{m^2(m-1)}\right]$$

Because a parallel design is used, the estimators of variance components are independent so that asymptotic variance σ^{*2} can be obtained by using variance of chi-square random variable. As a result, the sample size needed to achieve the desired power of $1 - \beta$ at the α level of significance can be obtained by solving the following equation

$$z_{\alpha/2} - \frac{\sqrt{n}(\sigma_{BT}^2 - \sigma_{BR}^2)}{\sigma^*} = -z_{\beta}$$

This leads to

$$n = \frac{\sigma^{*2}(z_{\alpha/2} + z_{\beta})^2}{(\sigma_{BT}^2 - \sigma_{BR}^2)^2}$$

Now consider the test of equality of total variabilities under the section “Standard 2×2 Crossover Design.” Here, the test is based on the MLS method and estimators of variance components are not independent. Under the alternative hypothesis and letting $n = n_1 + n_2 - 2$, without loss of generality, we assume $\sigma_{TR}^2 > \sigma_{TT}^2$. The power of the above test can be determined by

$$P\left(N\left(\frac{\sqrt{n}(\sigma_{TT}^2 - \sigma_{TR}^2)}{\sigma^*}, 1\right) > z_{\alpha/2}\right)$$

where

$$\sigma^{*2} = 2(\lambda_1^2 + \lambda_2^2) \equiv 2(\sigma_{TT}^4 + \sigma_{TR}^4 - 2\rho^2\sigma_{BT}^2\sigma_{BR}^2)$$

where let $\lambda_1 < 0 < \lambda_2$ are the two eigenvalues of the matrix $\Theta\Omega_B$ where Ω_B is defined in Eq. 25 and Θ is given in Eq. 27.

Hence, the sample size needed to achieve the power of $1 - \beta$ at the α level of significance can be obtained by solving the following equation

$$z_{\alpha/2} = -\frac{\sqrt{n}(\sigma_{TT}^2 - \sigma_{TR}^2)}{\sigma^*} = -z_{\beta}$$

which implies that

$$n = \frac{\sigma^{*2}(z_{\alpha/2} + z_{\beta})^2}{(\sigma_{TT}^2 - \sigma_{TR}^2)^2}$$

CONCLUSION

For comparing intrasubject variabilities and/or intrasubject CVs between treatment groups, replicates from the same subject are essential regardless of the study design. In clinical research, data are often log-transformed before the analysis. It should be noted that the intrasubject standard deviation of log-transformed data is approximately equal to the intrasubject CV of the untransformed (raw) data. As a result, it is suggested that intrasubject variability be used when analyzing log-transformed data and the intrasubject CV be considered when analyzing untransformed data.

In recent years, the assessment of reproducibility in terms of intrasubject variability or intrasubject CV in clinical research has received much attention. Chow^[11] defined reproducibility of a study drug as a collective term that encompasses consistency, similarity, and stability (control) within the therapeutic index (or window) of a subject's clinical status (e.g., clinical response of some primary study endpoint, blood levels, or blood concentration–time curve) when the study drug is repeatedly administered at different dosing periods under the same experimental condition. Reproducibility of clinical results observed from a clinical study can be quantitated through the evaluation of the so-called reproducibility probability, which will be briefly introduced.^[12]

For assessment of intersubject variability and/or total variability, Chow and Tse^[13] indicated that the usual analysis of variance model could lead to negative estimates of the variance components, especially the intersubject variance component. In addition, the sum of the best estimates of the intrasubject variance and the intersubject variance may not lead to the best estimate for the total variance. Chow and Shao^[14] proposed an estimation procedure for variance components, which cannot only avoid negative estimates but can also provide a better estimate than the maximum likelihood estimates. For estimation of total variance, Chow and Tse^[13] proposed a method as an alternative to the sum of estimates of individual variance components. These ideas could be applied to provide a better estimate of sample sizes for studies intended for comparing variabilities between treatment groups.

REFERENCES

1. Chinchilli, V.M.; Esinhart, J.D. Design and analysis of intra-subject variability in cross-over design. *Stat. Med.* **1996**, *15*, 1619–1634.
2. Howe, W.G. Approximate confidence limits on the mean of $x + y$ where x and y are two tabled independent random variables. *J. Am. Stat. Assoc.* **1974**, *69*, 789–794.

3. Graybill, F.A.; Wang, C.-M. Confidence intervals on non-negative linear combinations of variances. *J. Am. Stat. Assoc.* **1980**, *75*, 869–873.
4. Ting, N.; Burdick, R.K.; Graybill, F.A.; Jeyaratnam, S.; Lu, T.-F.C. Confidence intervals on linear combinations of variance components that are unrestricted in sign. *J. Stat. Comput. Simul.* **1990**, *35*, 135–143.
5. Hyslop, T.; Hsuan, F.; Holder, D.J. A small sample confidence interval approach to assess individual bioequivalence. *Stat. Med.* **2000**, *19*, 2885–2897.
6. Lee, Y.; Shao, J.; Chow, S.-C. Confidence intervals for Linear Combinations of Variance Components when Estimators of Variance Components are Dependent: An Extension of the Modified Large Sample Method with Application. *J. Amer. Stat. Assoc.* **2002**.
7. Lee, Y.; Shao, J.; Chow, S.-C.; Wang, H. Test for intersubject and total variabilities under crossover design. *J. Biopharm. Stat.* **2002**, *12*.
8. Chow, S.-C.; Tse, S.-K. A related problem in bioavailability/ bioequivalence studies—Estimation of the intrasubject variability with a common Cv. *Biom. J. J. Math. Methods Biosci.* **1990**, *32*, 597–607.
9. Quan, H.; Shih, W.J. Assessing reproducibility by the within-subject coefficient of variation with random effects models. *Biometrics.* **1996**, *52*, 1195–1203.
10. FDA. *Guidance for Industry: Statistical Approaches to Establishing Bioequivalence*; Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, MD, 2001.
11. Chow, S.C. *Statistical Methods for Assessment of Reproducibility*; Presented at Critical Assessment Review Advisory Board Meeting, Pfizer-Aventis: New York, 1995.
12. Chow, S.-C.; Shao, J. *Statistics in Drug Research—Methodology and Recent Development*; Marcel Dekker, Inc.: New York, 2001.
13. Chow, S.-C.; Tse, S.-K. On the estimation of total variability in assay validation. *Stat. Med.* **1991**, *10*, 1543–1553.
14. Chow, S.-C.; Shao, J. A new procedure for the estimation of variance components. *Stat. Probab. Lett.* **1988**, *6*, 349–355.

Clinical Research: Outlier Detection

Siu-Keung Tse

City University of Hong Kong, Hong Kong, China

Liming Xiang

Nanyang Technology University, Singapore

Abstract

This entry discusses some of the commonly used outlier detection techniques in clinical research, discussing methods of detection in normally distributed data, in generalized linear mixed models, and in crossover models.

INTRODUCTION

In clinical studies, detection of potential outlying observations is a critical step in the data analysis process before reaching any final conclusion of the study. The presence of outlying observations can affect the results of an analysis. In particular, it is possible that a totally opposite conclusion can be reached in marginal cases depending on the inclusion or exclusion of outlying observations. A bioavailability analysis that includes the possible outlying observations may lead to the rejection of bioequivalence when, in fact, the formulations are bioequivalent. Furthermore, outlying observations are often presented in the form of extremely large and/or small values. The inclusion of extreme values would inevitably increase the estimate of the variability. Large estimate of variability would affect the sensitivity of the statistical methodology in the detection of clinical difference. It is necessary to examine critically the presence of outlying observations in a data set. Thus, the study of outlier detection has received much attention in the literature. Many techniques have been proposed for the detection of outlying observations under different settings. This entry discusses some of the commonly used outlier detection techniques in clinical research.

Section “Detection of Outlier in Normally Distributed Data” reviews the statistical tests for the detection of outliers when data are normally distributed. The discussion is then extended to methods which are applicable to the regression settings such that the responses are dependent on one or several explanatory variables. Section “Detection of Outlier in Crossover Model” is focused on crossover models which are common designs in clinical study for assessing bioequivalence among several formulations. Statistical tests for detecting outlying subjects and observations within a given formulation will be outlined. Section “Outlier Detection Under Generalized Linear Mixed Models” discusses the detection of outliers in generalized linear mixed models (GLMM). GLMM is an extension of

the linear mixed models (LMM) and generalized linear models (GLM). It provides more general structure to study correlated data in a clinical study where data are collected from the same cluster or repeated measures resulting from a longitudinal study. An example is given in the section “An Example” to demonstrate the proposed detection procedure. Some concluding remarks are provided in the last section of this article.

DETECTION OF OUTLIER IN NORMALLY DISTRIBUTED DATA

Test for Extreme Values

Data from clinical studies are often assumed to be approximately normally distributed. The validity of the analysis is in doubt if the extreme observation(s), either too large or too small, is (are) included in the analysis. A statistical test proposed by Dixon^[1] can be used to determine whether the extreme value(s) should be deleted in the analysis.^[1] Suppose that $Y_1, Y_2, \dots, Y_n$ are n observations collected from a study. These observations are sorted ascendingly according to their magnitude and are denoted as $Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$, where $Y_{(1)}$ is the smallest value and $Y_{(n)}$ is the largest. For different values of sample size n , Dixon proposed criteria for the detection of outlying values and tabulated the corresponding critical values at significance level $\alpha = 5\%$ and 1% . For example, if the sample size n is between 14 and 25, the ratio r , defined in the following, is considered

$$r = \begin{cases} \frac{Y_{(3)} - Y_{(1)}}{Y_{(n-2)} - Y_{(1)}} & \text{if the smallest value is suspected} \\ & \text{to be outlier} \\ \frac{Y_{(n)} - Y_{(n-2)}}{Y_{(n)} - Y_{(3)}} & \text{if the largest value is suspected} \\ & \text{to be outlier} \end{cases} \quad (1)$$

This ratio r , which depends on the sample size n , is compared to appropriate tabulated values and the corresponding suspected value is considered to be an outlier if the ratio r is greater than or equal to the tabulated value. In particular, the expression (1) for r is different for different ranges of values of the sample size n . Details can be found in Dixon.^[1]

An alternative method to detect outliers for normally distributed data is the so-called T method, which is based on the standardized distance of the extreme observation from the mean of the sample. In particular, the standardized distance d is defined as $d = \frac{Y_n - \bar{Y}}{S}$ where Y_n is the extreme value, $\bar{Y}$ is the mean and S is the standard deviation of the n observations. For a given sample size n and significance level α , the extreme value is identified as an outlier if d is equal to or exceeds the corresponding critical value. Critical values for various n and $\alpha = 5\%$ are tabulated in Bolton.^[2]

Test for Residuals in Regression Models

Data collected from a clinical study can often be described by a regression model where the response Y is dependent on a set of p covariates $x_{1j}, x_{2j}, \dots, x_{pj}, j = 1, 2, \dots, n$. Thus, a multiple linear regression model $Y = X\beta + \varepsilon$ is considered, where Y is an $n \times 1$ vector of responses, X is an $n \times (p + 1)$ matrix representing the p covariates, β is an $(p + 1) \times 1$ vector of unknown parameters, and ε is an $n \times 1$ vector of random errors. In particular, the random errors are independently and normally distributed with mean 0 and variance σ^2 . Residuals, which are the differences between the fitted responses and the observed responses, resulting from the least squares estimation of the unknown parameters are given by $e = Y - X(X'X)^{-1}X'Y$. Many techniques proposed for the detection of outliers in the linear regression model are based on the examination of these residuals or the corresponding Studentized residuals, which is the ratio of the residual to its standard error. Reviews of these techniques can be found in Cook and Weisberg, Chatterjee and Hadi.^[3,4] In particular, Lund^[5] proposed an outlier detection method based on the Studentized residuals.^[5] The idea is to use a second order Bonferroni inequality to obtain upper and lower bounds for the probability of the maximum Studentized residuals exceeding a critical value. Tables of critical values for test at levels $\alpha = 0.01, 0.05$, and 0.1 are given in Lund.^[5] The U.S. Food and Drug Administration (FDA) has recommended using these results to identify outliers if data are fitted in a multiple linear regression model.

However, the methods discussed above are useful only in a multiple linear regression setting that requires statistical independence of all pharmacokinetic responses. In many clinical studies, it is quite common to adopt a crossover design for assessing bioequivalence among several formulations. Lund's method is not appropriate for a design in which the responses from the same subject are correlated.

Discussion in the following sections is focused on methods that are applicable to cases when data are correlated.

DETECTION OF OUTLIER IN CROSSEVER MODEL

In many clinical studies, a crossover method is adopted to study the bioequivalence of several formulations of a drug. Observations from the same subject are correlated and do not fit into the framework of a linear regression model. Therefore, it is necessary to consider other techniques for the detection of outlying observations under this correlated setting. Consider the following crossover model:

$$Y_{ijk} = \mu + S_i + F_j + P_k + \varepsilon_{ijk} \quad (2)$$

$$j, k = 1, \dots, n; i = 1, \dots, m$$

where Y_{ijk} is the response variable on the i th subject in the k th period under the j th formulation, μ is the overall mean; S_i is the random effect of the i th subject; F_j is the fixed effect of the j th formulation; P_k is the fixed effect of the k th period; and ε_{ijk} is the error term. In particular, we assume that the fixed effects F_j and P_k satisfy the conditions that $\sum_{j=1}^n F_j = 0$ and $\sum_{k=1}^n P_k = 0$; the random subject effects S_i and the errors ε_{ijk} are independently and normally distributed with mean 0 and variances σ_a^2 and σ_e^2 , respectively. In Eq. (2), the sequence effect is pooled together with the subject effect, i.e., S_i comprises the “between sequences” and “subjects within sequence” effects. In a bioavailability/bioequivalence study, the distribution of the response variable of primary interest is often skewed. Thus, these responses are often applied with a log transformation to remove the skewness and then the transformed data are analyzed using Eq. (2).

Likelihood Distance Test

In bioavailability studies, some subjects may exhibit extremely high or low bioavailability with respect to the reference formulation. These extreme observations indicate that the variability of subject responses to a formulation is not homogeneous and they may induce dramatic effects on the bioequivalence test. Chow and Tse^[6] have proposed a procedure based on the idea of likelihood distance to detect potential outlying subject.^[6] To highlight the essence of their approach, the discussion is focused on a simplified model of (2) with the assumption that there are no period and formulation effects. Then, Eq. (2) reduces to

$$Y_{ij} = \mu + S_i + \varepsilon_{ij} \quad j = 1, 2, \dots, n; \quad i = 1, 2, \dots, m. \quad (3)$$

The parameters associated with this reduced model are μ , σ_a^2 , and σ_e^2 . The log-likelihood function $L(\mu, \sigma_a^2, \sigma_e^2)$ is

$$L(\mu, \sigma_a^2, \sigma_e^2) = -\frac{mn}{2} \log 2\pi - \frac{m}{2} \log(\sigma_e^2 + n\sigma_a^2)(\sigma_e^2)^{n-1} \\ - \frac{1}{2\sigma_e^2} \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \mu)^2 \quad (4) \\ - \frac{n}{2} \left(\frac{1}{\sigma_e^2 + n\sigma_a^2} - \frac{1}{\sigma_e^2} \right) \sum_{i=1}^m (\bar{Y}_i - \mu)^2.$$

The maximum likelihood estimators (MLE) of μ , σ_a^2 and σ_e^2 can be found by maximizing (4) with respect to the parameters and are given by

$$\hat{\mu} = \bar{Y} = \frac{1}{nm} \sum_{i=1}^m \sum_{j=1}^n Y_{ij} \\ \hat{\sigma}_e^2 = \frac{1}{m(n-1)} \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2 \text{ with } \bar{Y}_i = \frac{1}{n} \sum_{j=1}^n Y_{ij} \\ \hat{\sigma}_a^2 = \frac{1}{n} \left\{ \frac{n}{m} \sum_{i=1}^m (\bar{Y}_i - \bar{Y})^2 - \frac{1}{m(n-1)} \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2 \right\}$$

If $\frac{n}{m} \sum_{i=1}^m (\bar{Y}_i - \bar{Y})^2 < \frac{1}{m(n-1)} \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2$, then the maximum likelihood estimators of σ_a^2 and σ_e^2 are derived by maximizing $L(\mu, \sigma_a^2, \sigma_e^2)$ under the condition that $\sigma_a^2 \geq 0$. Thus, the resulting MLEs are modified to $\hat{\mu} = \bar{Y}$, $\hat{\sigma}_e^2 = \frac{1}{nm} \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \bar{Y})^2$ and $\hat{\sigma}_a^2 = 0$.

The likelihood distance test is based on Cook and Weisberg's idea of likelihood distance to identify an outlying subject. Let $\theta = (\mu, \sigma_e^2, \sigma_a^2)$ be the parameter vector. Because the log-likelihood function $L(\theta)$ provides a summary of the information concerning for a given data set, we calculate the influence of the i th case based on the difference of the log-likelihood between $\hat{\theta}$ and $\hat{\theta}_{(-i)}$, where $\hat{\theta}_{(-i)}$ denotes the MLE of θ with deletion of the i th subject. Thus, the difference $L(\hat{\theta}) - L(\hat{\theta}_{(-i)})$ measures the influence of the i th subject of the given data set. If the difference is large, we consider the i th subject as an outlier. Furthermore, the statistic, which is based on the log-likelihood distance,

$$LD_i(\hat{\theta}) = 2[L(\hat{\theta}) - L(\hat{\theta}_{(-i)})] \quad (5)$$

is asymptotically chi-square distributed with 3 degrees of freedom as the number of subjects $m \rightarrow \infty$. Thus, we consider the i th subject as an outlier if

$$LD_i(\hat{\theta}) > \chi_3^2(\alpha) \quad (6)$$

where $\chi_3^2(\alpha)$ is the upper α percentage point of χ_3^2 .

Though the discussion above is based on the assumption that the period effects P_k and the formulation effects F_j are not presented in the model, the idea can be easily

generalized to cases including these effects. Under Eq. (2), the MLE of the period effects and the formulation effects are given by $\hat{P}_k = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n Y_{ijk} - \bar{Y}$ and $\hat{F}_j = \frac{1}{mn} \sum_{i=1}^m \sum_{k=1}^n Y_{ijk} - \bar{Y}$, where $\bar{Y} = \frac{1}{nm^2} \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^n Y_{ijk}$. The above estimation process is repeated by deleting one subject at a time to obtain $\hat{P}_{k(-i)}$ and $\hat{F}_{j(-i)}$. For the likelihood distance method, we can find the likelihood distance LD_i by taking the difference between the likelihood functions evaluated with and without the i th subject. The resulting difference LD_i is asymptotically chi-square distributed with $3 + 2(n-1)$ degrees of freedom, where the degree of freedom is equal to the number of independent parameters estimated in the model. However, note that the distribution of likelihood distance is based on asymptotic reasoning. Thus, it would be a concern if the number of subjects, m , is relatively small. Further discussion on the distribution properties of the likelihood distance can be found in Wang, Chow, and Wei.^[7]

Test Based on Hotelling T^2 Statistic

An alternative approach for outlier detection under Eq. (3) was proposed by Liu and Weng.^[8] Their idea was based on the order statistics of the two-sample Hotelling T^2 statistics. Let $Y'_i = (Y'_{i1}, \dots, Y'_{in})$ be the vector of responses. Then Y'_i is multivariate normally distributed with mean vector $\mu' = (\mu + S_1, \mu + S_2, \dots, \mu + S_m) = (\mu_1, \dots, \mu_m)$ and covariance matrix Σ with diagonal elements $\sigma_a^2 + \sigma_e^2$ and off-diagonal elements σ_a^2 .

Suppose that the i th subject is suspected to have a shift in the location. The hypotheses for the detection of this outlying subject are given as

$$H_0: Y'_i \sim N(\mu, \Sigma), \quad \text{for all } i = 1, 2, \dots, m \\ H_1: Y'_i \sim N(\mu + \Delta_i, \Sigma), \quad \text{for at least one } i \quad (7)$$

Equivalently, these hypotheses can be decomposed into m subhypotheses:

$$H_{0i}: Y'_i \sim N(\mu, \Sigma) \text{ vs } H_{1i}: Y'_i \sim N(\mu + \Delta_i, \Sigma), \\ \text{where } i = 1, 2, \dots, m. \quad (8)$$

The definition of (7) and (8) implies that $H_0 = \bigcap_{i=1}^m H_{0i}$ and $H_1 = \bigcup_{i=1}^m H_{1i}$. Denote T_i^2 to be the Hotelling T^2 statistic for the i th subject, then

$$T_i^2 = \frac{(m-1)(m-2)}{m} (Y'_i - \bar{Y}_{(-i)})' A_{(-i)}^{-1} (Y'_i - \bar{Y}_{(-i)}) \\ = \frac{(m-2)(Y'_i - \bar{Y})' A^{-1} (Y'_i - \bar{Y})}{\left[\frac{m-1}{m} - (Y'_i - \bar{Y})' A^{-1} (Y'_i - \bar{Y}) \right]} \quad (9)$$

where $\bar{Y}$ and A are the sample mean vector and matrix of sums of squares and cross products computed from the

responses $Y_i, \dots, Y_m$ respectively; and $\bar{Y}_{(-i)}$ and $A_{(-i)}$ are the similar statistics computed without the i th subject. Then, hypotheses (8) can be tested using test statistic T_i^2 given in (9) with the sequential step-down testing procedure to detect possible multiple outlying subjects. Details of this test procedure can be found in Liu and Weng.^[8] However, one difficulty associated with using this approach to detect outlying subjects is that the sampling distribution of the test statistic $T_{(i)}^2$ under the null hypothesis $H_{0(i)}$ is analytically intractable. Alternatively, Liu and Weng^[8] used Monte Carlo simulation to give the 5% and 1% upper quantiles of the distribution of $T_{(i)}^2$ for $n = 2, 3$; $m = 10(1)20$ and $20(5)50$.^[8] For other combinations of n and m values encountered in practice, one has to rely on simulation to evaluate the corresponding sample distribution of the test statistic.

The results of the above two methods, based either on the likelihood distance or the Hotelling T^2 statistic, are for the detection of outlying subjects. Liu and Weng^[8] have discussed how to generalize their idea for the detection of outlying observations within a given formulation.^[8] Further details and examples can be found in Chow and Liu.^[9]

OUTLIER DETECTION UNDER GENERALIZED LINEAR MIXED MODELS

The linear multiple regression models discussed in section II are employed based on an assumption of independent responses. In many clinical studies, this assumption may be in doubt for data collected in clusters due to the design of the study. For example, responses of subjects from the same hospital are anticipated to be correlated, though the responses may be independent between hospitals. Thus, each hospital is considered to be a cluster and the analysis has to take into account of the correlation of data from the same cluster. Besides, individual subject may be measured repeatedly over time after a treatment is applied. For a given subject, the measurements observed in such a longitudinal study are correlated across time but independent between subjects. Hence, the data can be regarded as clustered data for individual subject. The development of the GLMM was motivated by the need of fitting clustered data described in the above situations, in which the correlation within clusters is incorporated by the random effects in the model. Unlike ordinary linear regression, the GLMM can fit data from a broad distribution family, namely, the exponential distribution family, which includes normal distribution as a special case.

Formulation of GLMM

The GLMM for N clustered responses Y from the exponential family conditional on random effects u is defined by

$$Y | u \sim f_{y|u}(Y | u, \beta, \phi) = \exp\{\phi[Y'\eta - b(\eta) - c(Y)] - s(Y, \phi)\}. \quad (10)$$

Here the random effects u are assumed to be multivariate normally distributed, $u \sim N(0, \sigma^2 D)$ where $\sigma^2 (> 0)$ is an unknown parameter and D is a known non-negative definite matrix. The conditional expectation of Y given u is connected to a linear predictor η through a link function $g(\cdot)$, that is, $g(E(Y | u)) = \eta$ with $\eta = X\beta + Zu$. Here, X and Z are known design matrices corresponding to fixed effects β and random effects u , respectively. Each u_j has n_j components ($N = \sum_{j=1}^q n_j$), distributed as $N(0, \sigma^2 D_j)$ with a known covariance matrix D_j , and is independent of each other for $j = 1, \dots, q$. Hence, D may be expressed as $D = \text{diag}(D_1, D_2, \dots, D_q)$. The best linear unbiased predictor (BLUP) type joint log-likelihood of parameters β, u, σ^2 and ϕ is then defined as $l = l_1 + l_2$ with

$$l_1 = \ln f(Y; \beta | u) = \phi[Y'\eta - b(\eta) - c(Y)] - s(Y, \phi)$$

$$l_2 = \ln \phi(u | D) = -\frac{1}{2} \sum_{j=1}^q [n_j \log(2\pi\sigma^2) + \log |D_j| + u_j' D_j^{-1} u_j], \quad (11)$$

The approximate maximum likelihood estimates $\hat{\beta}, \hat{u}, \hat{\sigma}^2$, and $\hat{\phi}$ are obtained by maximizing l . Details can be found in McGilchrist.^[10]

Detection of Outlying Observations

Various approaches have been proposed to detect outlying observation(s) that may impact on model fitting and/or statistical inferences for the assumed statistical model, including case-deletion and mean-shift outlier models; see, e.g., Cook and Weisberg,^[3] Chatterjee and Hadi,^[4] and Atkinson.^[11] Among them, case-deletion is the most popular approach to identify such anomalous observations due to its intuition and simplicity. If an observation is deleted and the model refitted, the parameter estimates will change. The case-deletion approach, known as deletion diagnostics, uses formulae to judge such changes of the inference. Traditionally, the focus of deletion diagnostics is on the effects of removing an individual observation or a case. Since GLMM is appropriate for analyzing subgroups of correlated responses, it is natural to consider deletion in the context of a cluster. Therefore, the effects of outlying clusters and/or observations on parameter estimates are of interest in the GLMM setting.

In particular, the importance of the deletion of the k th cluster is judged by Cook's distance, defined generally by

$$CD_k = (\hat{\beta} - \hat{\beta}_{(k)})' \text{Cov}^{-1}(\hat{\beta})(\hat{\beta} - \hat{\beta}_{(k)}) \quad (12)$$

where $\hat{\beta}$ and $\hat{\beta}_{(k)}$ denote the estimates of β with and without the k th cluster, respectively. Cook's distance is a measure the difference between $\hat{\beta}$ and $\hat{\beta}_{(k)}$. Large values of Cook's distance correspond to "outlying" clusters.

To investigate the change of deleting a cluster of correlated observations on the estimator $\hat{\beta}$, there is no exact expression for such change provided by CD_k in the GLMM setting. Therefore, a first-order approximation to $\hat{\beta}$ with the k th cluster deleted is considered which would facilitate the derivation of an approximate Cook's distance CD_k . This approximate CD_k is computationally efficient because it avoids repeated refitting of the model after the deletion of the k th cluster. Pregibon^[12] has investigated this idea of using the first-order approximation in the GLM case.^[12]

Denote $\hat{\beta}_{(k)}$, $\hat{u}_{j(k)}$, and $\hat{\sigma}_{(k)}^2$ as the respective maximum likelihood estimators of β , u_j , and σ^2 after removing observations of the k th cluster. Since u_j 's are assumed to be mutually independent, deleting the k th cluster should not affect the estimator of u_j for $j \neq k$. The focus is thus on $\hat{\beta}_{(k)}$, treating u_j , σ^2 and ϕ as nuisance parameters. Let $\dot{l}(\hat{\beta})$ and $\ddot{l}(\hat{\beta})$ represent respectively the first-order and second-order derivatives of the log-likelihood l , defined in (11), with respect to β evaluated at $\beta = \hat{\beta}$. If $l_{1(k)}$ denotes the log-likelihood l_1 based on data without the k th cluster, then $\dot{l}_{1(k)}(\hat{\beta}_{(k)}) = 0$. Consider the first-order Taylor expansion of $\dot{l}_{1(k)}(\hat{\beta}_{(k)})$ at $\hat{\beta}$, then

$$\dot{l}_{1(k)}(\hat{\beta}) + \ddot{l}_{1(k)}(\hat{\beta})(\hat{\beta} - \hat{\beta}_{(k)}) \approx 0. \quad (13)$$

To simplify presentation, let $E = Y - \mu$ and $B = -a(\phi) \frac{\partial^2 l}{\partial \eta \partial \eta'} = \frac{\partial^2 c(\eta)}{\partial \eta \partial \eta'}$. Since the clusters are assumed to be mutually independent, $B = \text{diag}(B_1, B_2, \dots, B_q)$ is a block diagonal matrix with $n_j \times n_j$ component matrix B_j . Then,

$$\begin{aligned} \dot{l}_{1(k)}(\hat{\beta}) &= \left(\frac{\partial \eta_{1(k)}}{\partial \beta'} \right)' \cdot \frac{\partial l_{1(k)}}{\partial \eta_{(k)}} = a^{-1}(\phi) X'_{(k)} E_{(k)}, \\ \ddot{l}_{1(k)}(\hat{\beta}) &= \left(\frac{\partial \eta_{1(k)}}{\partial \beta'} \right)' \left(\frac{\partial^2 l_{1(k)}}{\partial \eta_{(k)} \partial \eta'_{(k)}} \right) \left(\frac{\partial \eta_{1(k)}}{\partial \beta'} \right) + \left(\frac{\partial l_{1(k)}}{\partial \eta_{(k)}} \right)' \left(\frac{\partial^2 \eta_{(k)}}{\partial \beta \partial \beta'} \right) \\ &= X'_{(k)} (-a^{-1}(\phi) B_{(k)}) X_{(k)} \end{aligned}$$

where the derivatives are evaluated at $\hat{\beta}$, $\hat{u}$ and $\hat{\phi}$. Rewrite X and E in partitioned matrix form as $X = (X'_1, \dots, X'_q)'$ and $E = (E'_1, \dots, E'_q)'$, with $n_j \times p$ component matrix X_j and $n_j \times 1$ component vector E_j corresponding to the j th cluster, for $j = 1, \dots, q$. From (13), after removing the k th cluster, the first-order approximate updating formula for $\hat{\beta}_{(k)}$ is

$$\hat{\beta}_{(k)} \approx \hat{\beta} - (\tilde{X}'\tilde{X})^{-1} \tilde{X}'(I - \tilde{H}_k)^{-1} \tilde{E}_k \quad (14)$$

where $\tilde{X} = B^{-\frac{1}{2}}X = (\tilde{X}'_1, \dots, \tilde{X}'_q)'$, $\tilde{X}_k = B_k^{-\frac{1}{2}}X_k$, $\tilde{E}_k = B_k^{-\frac{1}{2}}E_k$ and $\tilde{H}_k = \tilde{X}'_k(\tilde{X}'\tilde{X})^{-1}\tilde{X}_k$ for $k = 1, \dots, q$.

Note that the Cook's distance $CD_k(\hat{\beta})$ for assessing influence of the k th cluster on $\hat{\beta}$ in GLMM can be defined as $CD_k(\hat{\beta}) = (\hat{\beta} - \hat{\beta}_{(k)})'(-\ddot{l}(\hat{\beta}))(\hat{\beta} - \hat{\beta}_{(k)}) / (p \cdot a^{-1}(\hat{\phi}))$.

To facilitate computation, a first-order approximation to $CD_k(\hat{\beta})$ based on (14) is recommended by Xiang et al.:^[13]

$$CD_k(\hat{\beta}) \approx \tilde{E}'_k(I - \tilde{H}_k)^{-1} \tilde{H}_k(I - \tilde{H}_k)^{-1} \tilde{E}_k / p. \quad (15)$$

Results (14) and (15) are analogous to those for the linear regression model as discussed in Cook and Weisberg.^[3] Here $\tilde{E}_k$ is the residual vector associated with the k th cluster, and $\tilde{H}_k$ has a structure similar to that of the hat matrix in regression. Furthermore, if a subset of parameters is of interest, $CD_k(\hat{\beta})$ given in (15) can be modified accordingly through the corresponding partitioned matrix. In particular, to assess influence of the k th cluster on individual coefficient, the corresponding Cook's distance measure is

$$\begin{aligned} CD_k(\hat{\beta}_j) &= (\hat{\beta}_j - \hat{\beta}_{j(k)})'(-d'_j \ddot{l}^{-1}(\hat{\beta}) d_j)^{-1} (\hat{\beta}_j - \hat{\beta}_{j(k)}) / (a^{-1}(\hat{\phi})) \\ &\approx \tilde{E}'_k(I - \tilde{H}_k)^{-1} \tilde{X}_k \left[(\tilde{X}'\tilde{X})^{-1} - \begin{pmatrix} 0 & 0 \\ 0 & (\tilde{X}'_{(j)}\tilde{X}_{(j)})^{-1} \end{pmatrix} \right] \\ &\quad \times \tilde{X}'_k(I - \tilde{H}_k)^{-1} \tilde{E}_k \end{aligned} \quad (16)$$

with $\tilde{X} = (\tilde{x}_j, \tilde{X}_{(j)})$ and $\beta = (\beta_j, \beta'_{(j)})'$, where $\beta_{(j)}$ is a $(p-1) \times 1$ vector.

Diagnostics for Individual Observations

The idea discussed in the previous section can be generalized to detect outlying observations within a cluster. Let i denote the index for the individual observation that is deleted from data, and the quantity with subscript $[i]$ is evaluated by the dataset without the i th observation. Without loss of generality, X is partitioned as $X = (x_i, X_{[i]})'$, where x_i is a vector of the i th row of X ; and correspondingly,

$$B \text{ and } B^{-1} \text{ as } B = \begin{pmatrix} b_{ii} & b_{i1} \\ b_{1i} & B_{[i]} \end{pmatrix} \text{ and } B^{-1} = \begin{pmatrix} c_{ii} & c_{i1} \\ c_{1i} & C_{[i]} \end{pmatrix}.$$

Then, the first order approximation to the updating formula for $\hat{\beta}_{[i]}$ is

$$\hat{\beta}_{[i]} \approx \hat{\beta} - (\tilde{X}'\tilde{X})^{-1} \tilde{x}_i \bar{e}_i / (b_{ii}^{-1} - s_{ii}) \quad (17)$$

and the Cook's distance statistic for the i th observation on $\hat{\beta}$ has the following form

$$\begin{aligned} CD_{[i]}(\hat{\beta}) &= (\hat{\beta} - \hat{\beta}_{[i]})'(-\ddot{l}(\hat{\beta}))(\hat{\beta} - \hat{\beta}_{[i]}) / (p \cdot a^{-1}(\hat{\phi})) \\ &\approx \bar{e}_i^2 s_{ii} b_{ii}^2 / [p(1 - s_{ii} b_{ii}^2)] \end{aligned} \quad (18)$$

where $\tilde{x}_i = x_i - c_{1i} C_{[i]}^{-1} X_{[i]}$, $\bar{e}_i = e_i - c_{1i} C_{[i]}^{-1} E_{[i]}$, e_i is the i th component of the residual vector E , b_{ii} is the i th diagonal element of the matrix B and $s_{ii} = \tilde{x}_i'(\tilde{X}'\tilde{X})^{-1} \tilde{x}_i$, $i = 1, \dots, n$, $n = \sum_{j=1}^q n_j$.

Similar to the case of cluster deletion diagnostics, the Cook's distance (18) can be modified correspondingly

through the partition matrix for a subset of parameters. In particular, to assess influence of the i th individual observation on the j th component of the coefficient, the assessment $CD_{[i]}(\hat{\beta}_j)$ is then given by

$$CD_{[i]}(\hat{\beta}_j) = (\hat{\beta}_j - \hat{\beta}_{j[i]})' (-d_j' \tilde{I}^{-1}(\hat{\beta}) d_j)^{-1} (\hat{\beta}_j - \hat{\beta}_{j[i]}) / (\sigma^2(\hat{\phi})) \\ \approx \bar{e}_i^2 (s_{ii} - s_{ij})^2 / (1 - s_{ii} b_{ii})^2, \quad (19)$$

where d_j is a $p \times 1$ vector with the j th component being one and zero elsewhere, for $j = 1, \dots, p$; $s_{ii} = \bar{x}_i' (\tilde{X}' \tilde{X})^{-1} \bar{x}_i$, $\bar{x}_i = (\bar{x}_{ij}, \bar{x}_{i(j)})'$ and $s_{ij} = \bar{x}_{ij}' (\tilde{X}' \tilde{X})^{-1} \bar{x}_{i(j)}$.

In practice, a two-step detection procedure is recommended to detect outlying observations. In other words, the first step is to examine the diagnostics plot of $CD_k(\hat{\beta})$ against cluster index k to identify clusters with anomalous large influence on β . Then, the plot of $CD_{[i]}(\hat{\beta})$ against observation index i is used to identify influential observations. Results from both plots can identify whether there is any relationship between both deletion diagnostics. An example is discussed in details in the next section to illustrate the application of this detection procedure.

AN EXAMPLE

Inpatient maternity length of stay (LOS) is a readily available indicator of hospital activity. It is also considered to be a reasonable proxy of resource consumption. Obstetrical delivery is one of the most frequent causes of hospitalization in many countries. However, the appropriate LOS after delivery is a complex issue. Recent advances in medical technology and clinical practice have generally enabled obstetrical services to be provided more effectively and to improve pregnancy outcomes. In particular, streamlined delivery procedures and improved pharmaceuticals have contributed to the decline in maternity LOS.

Based on the 1998/1999 data from $N = 1,869$ women hospitalized for caesarean delivery without complicating diagnosis, the outcome variable LOS is defined to be the number of (whole) days from admission to discharge. The sample LOS observations are clustered within $q = 28$ public hospitals. In addition, concomitant information on the number of diagnoses, number of procedures, admission type (0 = elective, 1 = emergency), and payment classification (0 = public, 1 = private) are also available. In general, it is reasonable to assume that the LOS counts are independent between hospitals while certain within-hospital correlation is anticipated. Considering the histogram of responses LOS and fitted Poisson density line, data are fitted reasonably by the Poisson mixed regression model. The random effects are intended to capture discrepancies in medical expertise, obstetrical care, and other unmeasurable characteristics of the population in each hospital's catchment area. Results of which are presented in Table 1a.

First, the cluster-specified Cook's distance is considered. The exact CD_k and its first-order approximation are plotted against hospital index k in Fig. 1. The first-order approximation appears to be biased downwards compared to the exact Cook's distance. However, it clearly shows that Hospital 15 ($n_{15} = 370$) is influential on the parameter estimates. It is interesting to note that Hospital 6 ($n_6 = 111$) has the smallest average LOS among all hospitals. Results of the fit after removing Hospital 15 are summarized in Table 1b. It is noted that the patient's admission type now becomes significant, but conversely for number of procedures. After removing Hospital 15, analysis on the re-computed CD_k values based on the remaining hospitals showed that Hospital 1 is also a potential outlying hospital. This phenomenon is due to the masking effect on Hospital 1 by Hospital 15. To reinforce the indications of the diagnostics, we compare the estimation results of the Poisson linear mixed model fitted without Hospital 1 (Table 1c) and without both Hospitals 1 and 15 (Table 1d). The results clearly demonstrate the presence of masking effect in this data set.

Table 1 Parameter Estimates of the Poisson Linear Mixed Model for Maternity LOS

	(a) Full data		(b) Hospital 15 deleted		(c) Hospital 1 deleted		(d) Hospitals 1 and 15 deleted	
	$\hat{\beta}$	p -value	$\hat{\beta}$	p -value	$\hat{\beta}$	p -value	$\hat{\beta}$	p -value
Constant	1.533	0.000	1.528	0.000	1.528	0.000	1.523	0.000
No. of diagnoses	0.019	0.005	0.026	0.004	0.019	0.007	0.026	0.006
No. of procedures	0.021	0.038	0.017	0.118	0.021	0.035	0.018	0.109
Emergency admission	-0.080	0.057	-0.111	0.023	-0.049	0.254	-0.070	0.156
Private payment	0.189	0.000	0.172	0.017	0.169	0.003	0.131	0.105
$\hat{\sigma}$	0.122		0.121		0.125		0.125	

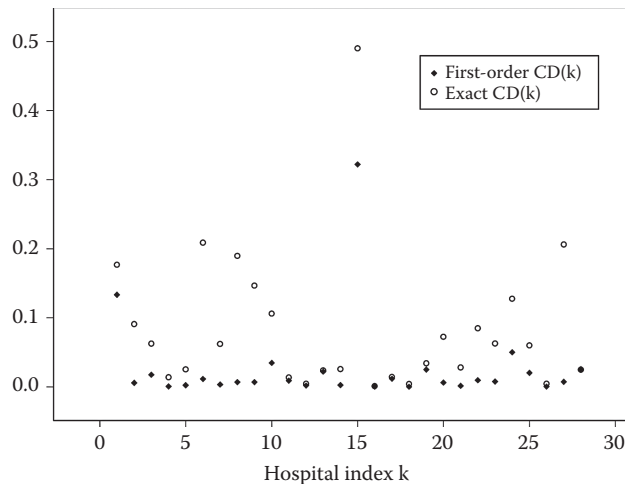


Fig. 1 Exact Cook's distance and its first-order approximation

Besides, standardized random effect predictions $\frac{\hat{u}_j}{\hat{\sigma}}$ corresponding to the four situations in Table 1 are presented in Fig. 2. Despite the substantial influence exerted by Hospitals 1 and 15 on the regression coefficients, their omissions apparently have little impact on the random effect estimates. It can be seen that Hospital numbers 6 and 15 are relatively efficient in terms of shortening the duration of hospitalization, after adjusting for the relevant risk factors. Incidentally, Hospital number 15 corresponds to a large teaching hospital while Hospital 1 is a small rural hospital with $n_1 = 62$. Hospital 6 is a regional maternity hospital that reported the shortest average LOS of 3.7 days and none of its $n_6 = 111$ patients have private insurance cover. These findings have important implications in terms of hospital resource allocation and efficiency assessment. Regarding the influence of each hospital on each regression coefficient, Fig. 3 shows that only Hospital 15 is influential.

Finally, regarding the issue of detecting outlying observations, the diagnostic plot of $CD_{[i]}$ versus individual index

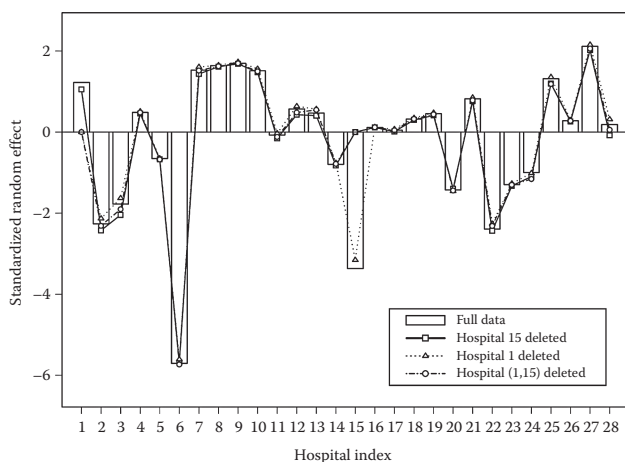


Fig. 2 Estimated hospital effects from fitting Poisson linear mixed model

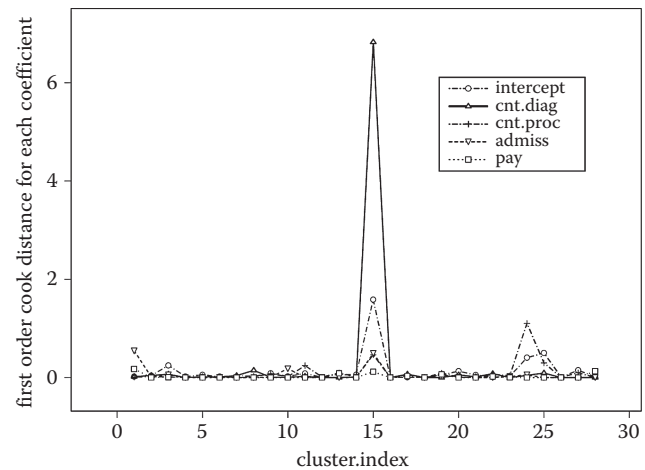


Fig. 3 Approximate Cook's distance for each regression coefficient

i is given in Fig. 4. It can be found that the observation of patient No. 1777 has the greatest $CD_{[i]}$, which is obviously much greater than the second largest one. In fact, the LOS of this patient is 14 days, which is the second largest response value. However, this individual patient is from the twenty-third hospital, which is not detected to be an influential cluster according to the previous analysis. It indicates that this influential individual does not cause the hospital where he or she stayed to be influential. On the other hand, no individual observations from influential cluster 15 or cluster 1 show obvious evidence of individual influence on the estimate of the parameter. In particular, after deleting the individual 1777 from the complete dataset, the estimates of β and σ are $\hat{\beta} = (1.5324, 0.0196, 0.0205, -0.0926, 0.1664)'$ and $\hat{\sigma} = 0.1221$ respectively, which are not much different from parameter estimates based on the full data as shown in Table 1a. Hence, only regarding the magnitude of the estimate may not identify the anomalous individual observation out of a large data set.

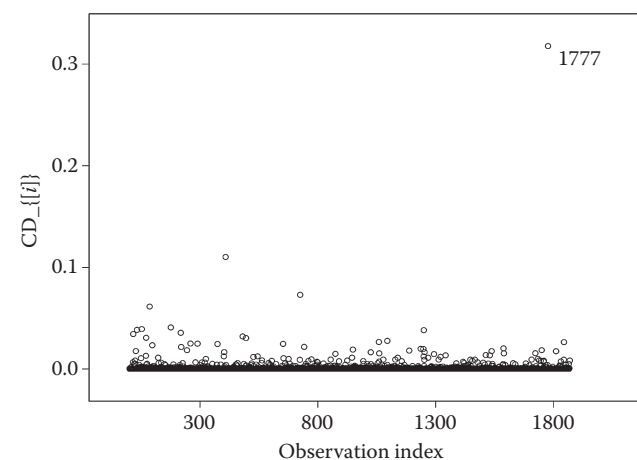


Fig. 4 First order approximate Cook's distance for individual observations

One final remark before concluding this section is on the assessment of the Cook's values: There has been much debate about how these values should be used. Some suggested comparing these to an absolute cutoff value based on some approximate distribution of the Cook's distance statistic. For example, in the linear regression case, values of CD_k and $CD_{[i]}$ can be approximately evaluated with reference to the χ^2_{p-1} distribution to some extent as pointed out by Cook and Weisberg.^[3] However, this criterion is always overly conservative in practice. The situation is anticipated to be much more unclear in GLMM because of its complicated model structure. Alternatively, it would be more effective to identify those clusters or observations with unusually large CD values; and compare the fitting results with and without the identified observations.

CONCLUSION

All the techniques discussed above are focused on the detection of a single outlying observation or cluster based on the assumption that the proposed model is valid. The problem of identifying a single outlying observation or cluster is relatively simple from both analytical and computational points of view. If the data set contains more than one outlying observation or cluster, which is not unlikely for many data sets, it would be much more difficult to identify these outliers due to the masking and swamping effects. The former is concerned with the impact of a particular observation that may not be apparent until another observation has been deleted. The latter refers to observations that are jointly but not individually influential. In fact, these effects are presented in the LOS data set discussed in section "An Example". The masking effects of deletion diagnostics may be considered in terms of joint influence and conditional influence. Hadi and Simonoff^[14] consider the problem of identifying and testing multiple outliers in linear models.^[14] Generalized cluster-wise joint and conditional deletion measures can be developed for investigating effects of masking in GLMM setting similar to the approach suggested by Lawrance.^[15]

The objective of detecting outliers from a clinical study is to improve the validity of the statistical procedure and draw proper conclusions. After some data are identified as potential outliers in a data set by appropriate statistical techniques, statisticians are often faced with the dilemma

in how to deal with these outliers. However, there is not any standard answer to the question of whether or not to include the data. Experimenters should review the process critically to identify a cause for these outlying observations. If they are caused by errors committed in the research process, then the data can be discarded from the study without inducing any bias to the conclusion. In general, the decision to discard any outlier(s) from the study should be made cautiously. It should be done only after thorough review of the properties of the data and the insight of the experimenter.

REFERENCES

1. Dixon, W.J.; Massey, Jr., F.J. Processing data for outlier. *Biometrics* **1953**, *9*, 74–89.
2. Bolton, S. *Pharmaceutical Statistics: Practical and Clinical Application*, 3rd Ed.; Marcel Dekker: New York, 1997.
3. Cook, R.D.; Weisberg, S. *Residuals and Influence in Regression*; Chapman and Hall: New York, 1982.
4. Chatterjee, S.; Hadi, A.S. *Sensitivity Analysis in Linear Regression*; Wiley: New York, 1988.
5. Lund, R.E. Tables for an approximate test for outliers in linear models. *Technometrics* **1975**, *17*, 473–476.
6. Chow, S.C.; Tse, S.K. Outlier detection in bioavailability/bioequivalence studies. *Stat. Med.* **1990**, *9*, 549–558.
7. Wang, W.P.; Chow, S.C.; Wei, W. On likelihood distance for outliers detection. *J. Biopharm. Stat.* **1995**, *5*, 307–322.
8. Liu, J.P.; Weng, C.S. Detection of outlying data in bioavailability/bioequivalence studies. *Stat. Med.* **1991**, *10*, 1375–1389.
9. Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*, 2nd Ed.; Marcel Dekker: New York, 2000.
10. McGilchrist, C.A. Estimation in generalized mixed models. *J. Roy. Stat. Soc. B* **1994**, *56*, 61–69.
11. Atkinson, A.C. *Plots, Transformations and Regression*; Clarendon: Oxford, 1985.
12. Pregibon, D. Logistic regression diagnostics. *Ann. Stat.* **1981**, *9*, 705–724.
13. Xiang, L.; Tse, S.K.; Lee, A.H. Deletion diagnostics for generalized linear mixed models, applications to clustered data. *Comput. Stat. Data Anal.* **2002**, *40*, 759–774.
14. Hadi, A.S.; Simonoff, J.S. Procedures for the identification of multiple outliers in linear models. *J. Am. Stat. Assoc.* **1993**, *88*, 1264–1272.
15. Lawrance, A.J. Deletion influence and masking in regression. *J. Roy. Stat. Soc. B* **1995**, *57*, 181–189.

Clinical Research: Reproducibility Probability

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

This entry reviews three approaches to reproducibility probability and provides examples and applications for this clinical research tool.

INTRODUCTION

For marketing approval of a new drug product, the United States Food and Drug Administration (FDA) requires that at least two adequate and well-controlled clinical trials be conducted to provide substantial evidence regarding the effectiveness and safety of the drug product under investigation.^[1] Under certain circumstances, however, the FDA Modernization Act (FDAMA) of 1997 includes a provision (Section 115 of FDAMA) to allow data from one adequate and well-controlled clinical trial investigation and confirmatory evidence to establish effectiveness for risk/benefit assessment of drug and biological candidates for approval. The purpose of requiring at least two clinical studies is to assure that clinical results are reproducible. *Reproducibility*, in clinical research, is then referred to as whether the clinical results are reproducible from a location (e.g., a study site or center) to another location within the same region (e.g., the United States, Europe, or Asian Pacific Region) or from region to region with respect to the same patient population.

In practice, it is often of interest to determine whether a clinical trial that produced positive clinical results provides substantial evidence to assure reproducibility of the clinical results. In this entry, the reproducibility of a positive clinical result is studied by evaluating the probability of observing a positive result in a future clinical study with the same study protocol, given that a positive clinical result has been observed.

In the section “Reproducibility Probability,” the concept of reproducibility probability of a positive clinical result is introduced. Three different approaches of evaluating the reproducibility probability, namely, the estimated power approach, the confidence bound approach, and the Bayesian approach, are discussed in the sections “The Estimated Power Approach” and “The Confidence Bound Approach,” respectively. Applications and extensions of the derived results are discussed in the section “The Bayesian Approach.” For example, if the reproducibility probability is high (e.g., 95%) with some statistical assurance based on the results from the first clinical trial, then it may not be necessary to require the sponsor to conduct the second clinical trial.

REPRODUCIBILITY PROBABILITY

Let H_0 be the null hypothesis that the mean response of the drug product is the same as the mean response of a control (e.g., placebo). When H_0 is concluded, the drug product is not considered to be effective. Let T be the test statistic based on the responses from a clinical trial. The clinical trial produces a positive clinical result if the observed value of T leads to the rejection of the null hypothesis H_0 .

Let H_1 be the alternative hypothesis which states that H_0 is not true. Suppose that the null hypothesis H_0 is rejected if and only if $|T| > c$, where c is a positive known constant. This is usually related to a two-sided alternative hypothesis. The discussion for one-sided alternative hypotheses is similar. In statistical theory, the probability of observing a positive clinical result when H_1 is indeed true, which is commonly denoted by

$$P(|T| > c | H_1) \quad (1)$$

is referred to as the power of the test procedure. Typically, the power is a function of unknown parameters.

Consider two independent clinical trials of the same study protocol and sample size. The power for each trial is given by Eq. 1, whereas the probability that *both* trials yield positive results when H_1 is true is

$$[P(|T| > c | H_1)]^2$$

because the two trials are independent. For example, if the power for a single study is 80%, then the probability of observing two positive clinical results when H_1 is true is $(80\%)^2 = 64\%$.

Suppose now that one clinical trial was conducted and the result is positive. What is the probability that the second trial will produce a positive result, i.e., the positive result from the first trial is reproducible? Mathematically, because the two trials are independent, the probability of observing a positive result in the second trial when H_1 is true is still given by Eq. 1, regardless of whether the result from the first trial is positive or not. However, information from the first

clinical trial should be useful in the evaluation of the probability of observing a positive result in the second trial. This leads to the concept of reproducibility probability, which is different from the power defined by Eq. 1.

In general, the reproducibility probability is a person's subjective probability of observing a positive clinical result from the second trial, when he/she observes a positive result from the first trial. For example, the reproducibility probability can be defined as the probability in Eq. 1 with all unknown parameters replaced by their estimates from the first trial.^[2] In other words, the reproducibility probability is defined to be an estimated power of the second trial using the data from the first trial. More discussions of this approach will be given in the subsequent sections.

Perhaps a more sensible definition of reproducibility probability can be obtained by using the Bayesian approach.^[3] Under the Bayesian approach, the unknown parameter under H_1 , denoted by θ is a random vector with a prior distribution $\pi(\theta)$ assumed to be known. Thus

$$P(|T| > c | H_1) = P(|T| > c | \theta)$$

and the reproducibility probability can be defined as the conditional probability of $|T| > c$ in the second trial, given the data set x observed from the first trial, i.e.,

$$P(|T| > c | x) = \int P(|T| > c | \theta) \pi(\theta | x) d\theta \quad (2)$$

where $T = T(y)$ is based on the data set y from the second trial and $\pi(\theta | x)$ is the posterior density of θ , given x .

THE ESTIMATED POWER APPROACH

To study the reproducibility probability, we need to specify the test procedure, i.e., the form of the test statistic T . We first consider the simplest design of two parallel groups with equal variances. The case of unequal variances is considered next. The parallel-group design is discussed at the end of this section.

Two Samples with Equal Variances

Suppose that a total of $n = n_1 + n_2$ patients are randomly assigned to two groups, a treatment group and a control group. In the treatment group, n_1 patients receive the treatment (or a test drug) and produce responses $x_{11}, \dots, x_{1n_1}$. In the control group, n_2 patients receive the placebo (or a reference drug) and produce responses $x_{21}, \dots, x_{2n_2}$. This design is a typical two-group parallel design in clinical trials. We assume that x_{ij} 's are independent and normally distributed with means $\mu_i, i = 1, 2$, and a common variance σ^2 . Suppose that the hypotheses of interest are

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_1: \mu_1 - \mu_2 \neq 0 \quad (3)$$

The discussion for a one-sided H_1 is similar.

When σ^2 is known, we reject H_0 at the 5% level of significance if and only if

$$|T| > z_{0.975} \quad (4)$$

where $z_{0.975}$ is the 97.5th percentile of the standard normal distribution,

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (5)$$

and $\bar{x}_i$ is the sample mean based on the data in the i th group.

Consider two independent clinical trials of the same design. Let x denote the data observed from the first trial and y be the data to be observed from the second trial. The test statistics for two trials are denoted by $T(x)$ and $T(y)$, respectively. For T given by Eq. 5, the power of T for the second trial is

$$P(|T(y)| > z_{0.975}) = \Phi(\theta - z_{0.975}) + \Phi(-\theta - z_{0.975}) \quad (6)$$

where Φ is the standard normal distribution function and

$$\theta = \frac{\mu_1 - \mu_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (7)$$

If we replace θ in Eq. 6 by its estimate based on the data from the first trial, $T(x)$, then an estimated power, which can be defined as the reproducibility probability, is given by

$$\hat{P} = \Phi(T(x) - z_{0.975}) + \Phi(-T(x) - z_{0.975}) \quad (8)$$

Note that $\hat{P}$ is a function of $|T(x)|$. When $|T(x)| > z_{0.975}$ (i.e., the first trial yields a positive clinical result),

$$\hat{P} \approx \Phi(|T(x)| - z_{0.975})$$

which is the formula for reproducibility probability given by Goodman.^[2] Values of in Eq. 8 for several different values of $|T(x)|$ (and the related p -values) are listed in the third column of Table 1. As can be seen in Table 1, if the P -value observed from the first trial is less than 0.001, there is more than 90% chance that the clinical result will be reproducible in the second trial under the same protocol and clinical trial environments.

In practice, σ^2 is usually unknown and, thus, the test given by Eqs. 4 and 5 needs to be replaced by the commonly used two sample t -test which rejects H_0 if and only if

$$|T| > t_{0.975; n-2}$$

Table 1 Reproducibility Probability $\hat{P}$

$ T(x) $	Known σ^2		Unknown σ^2 ($n = 30$)	
	p-Value	$\hat{P}$	p-Value	$\hat{P}$
1.96	0.050	0.500	0.060	0.473
2.05	0.040	0.536	0.050	0.508
2.17	0.030	0.583	0.039	0.554
2.33	0.020	0.644	0.027	0.614
2.58	0.010	0.732	0.015	0.702
2.81	0.005	0.802	0.009	0.774
3.30	0.001	0.910	0.003	0.890

where $t_{0.975;n-2}$ is the 97.5th percentile of the t-distribution with $n - 2$ degrees of freedom,

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (9)$$

and s_i^2 is the sample variance based on the data from the i th group. Then the power of T for the second trial is

$$p(\theta) = P(|T(y)| > t_{0.975;n-2}) \\ = 1 - T_{n-2}(t_{0.975;n-2}|\theta) + T_{n-2}(-t_{0.975;n-2}|\theta) \quad (10)$$

where θ is given by Eq. 7 and $T_{n-2}(\cdot|\theta)$ denotes the distribution function of the noncentral t-distribution with $n - 2$ degrees of freedom and the noncentrality parameter θ . Note that $p(\theta) = p(|\theta|)$. Values of $p(\theta)$ as a function of $|\theta|$ is provided in Table 2. Using the same idea of replacing θ by its estimate $T(x)$, where T is now defined by Eq. 9, we obtain the following reproducibility probability:

$$\hat{p} = 1 - T_{n-2}(t_{0.975;n-2}|T(x)) + T_{n-2}(-t_{0.975;n-2}|T(x)) \quad (11)$$

which is a function of $|T(x)|$. Again, when $|T(x)| > t_{0.975;n-2}$,

$$\hat{P} \approx \begin{cases} 1 - T_{n-2}(t_{0.975;n-2}|T(x)) & \text{if } T(x) > 0 \\ T_{n-2}(-t_{0.975;n-2}|T(x)) & \text{if } T(x) < 0 \end{cases}$$

For comparison, values of $\hat{P}$ in Eq. 11 for $n = 30$ corresponding to $\hat{P}$ in Eq. 8 are listed in the fifth column of Table 1. It can be seen that with the same value of $T(x)$, the reproducibility probability for the case of unknown σ^2 is lower than that for the case of known σ^2 .

Table 2 can be used to find the reproducibility probability $\hat{P}$ in Eq. 11 with a fixed sample size n . For example, if $T(x) = 2.9$ was observed in a clinical trial with $n = n_1 + n_2 = 40$, then the reproducibility probability is 0.807. If $|T(x)| = 2.9$ was observed in a clinical trial with $n = 36$, then an extrapolation of the results in Table 2 (for $n = 30$ and 40) leads to a reproducibility probability of 0.803.

Two Samples with Unequal Variances

Consider the problem of testing hypotheses (Eq. 3) under the two-group parallel design without the assumption of equal variances. That is, x_{ij} 's are independently distributed as $N(\mu_i, \sigma_i^2)$, $i = 1, 2$.

When $\sigma_1^2 \neq \sigma_2^2$, there exists no exact testing procedure for the hypotheses in Eq. 3. When both n_1 and n_2 are large, an approximate 5% level test rejects H_0 when $|T| > z_{0.975}$, where

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (12)$$

Because T is approximately distributed as $N(\theta, 1)$ with

$$\theta = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (13)$$

the reproducibility probability obtained by using the estimated power approach is given by Eq. 8 with T defined by Eq. 12.

When the variances under different treatments are different and the sample sizes are not large, a different study design, such as a matched-pair parallel design or a 2×2 crossover design is recommended. A matched-pair parallel design involves m pairs of matched patients. One patient in each pair is assigned to the treatment group and the other is assigned to the control group. Let x_{ij} be the observation from the j th pair and the i th group. It is assumed that the differences $x_{1j} - x_{2j}$, $j = 1, \dots, m$, are independent and identically distributed as $N(\mu_1 - \mu_2, \sigma_D^2)$. Then, the null hypothesis H_0 is rejected at the 5% level of significance if $|T| > t_{0.975;m-1}$, where

$$T = \frac{\sqrt{m}(\bar{x}_1 - \bar{x}_2)}{\hat{\sigma}_D} \quad (14)$$

and $\hat{\sigma}_D^2$ is the sample variance based on the differences $x_{1j} - x_{2j}$, $j = 1, \dots, m$. Note that T has the noncentral t -distribution with $n - 1$ degrees of freedom and the noncentrality parameter

$$\theta = \frac{\sqrt{m}(\mu_1 - \mu_2)}{\sigma_D} \quad (15)$$

Consequently, the reproducibility probability obtained by using the estimated power approach is given by Eq. 11 with T defined by Eq. 16 and $n - 2$ replaced by $m - 1$.

Suppose that the study design is a 2×2 crossover design in which n_1 patients receive the treatment at the first period and the placebo at the second period and n_2 patients

Table 2 Values of the Power Function $p(\theta)$

$ \theta $	n							
	10	20	30	40	50	60	100	∞
1.96	0.407	0.458	0.473	0.480	0.484	0.487	0.492	0.500
1.98	0.414	0.466	0.481	0.488	0.492	0.495	0.500	0.508
2.00	0.421	0.473	0.489	0.496	0.500	0.503	0.508	0.516
2.02	0.429	0.481	0.496	0.504	0.508	0.511	0.516	0.524
2.04	0.435	0.488	0.504	0.511	0.516	0.518	0.524	0.532
2.06	0.442	0.496	0.512	0.519	0.523	0.526	0.532	0.540
2.08	0.448	0.503	0.519	0.527	0.531	0.534	0.540	0.548
2.10	0.455	0.511	0.527	0.535	0.539	0.542	0.548	0.556
2.12	0.462	0.519	0.535	0.542	0.547	0.550	0.555	0.563
2.14	0.469	0.526	0.542	0.550	0.555	0.557	0.563	0.571
2.16	0.476	0.534	0.550	0.558	0.562	0.565	0.571	0.579
2.18	0.483	0.541	0.558	0.565	0.570	0.573	0.579	0.587
2.20	0.490	0.549	0.565	0.573	0.578	0.581	0.586	0.594
2.22	0.497	0.556	0.573	0.581	0.585	0.588	0.594	0.602
2.24	0.504	0.563	0.580	0.588	0.593	0.596	0.602	0.610
2.26	0.511	0.571	0.588	0.596	0.601	0.604	0.609	0.618
2.28	0.518	0.578	0.595	0.603	0.608	0.611	0.617	0.625
2.30	0.525	0.586	0.603	0.611	0.616	0.619	0.625	0.633
2.32	0.532	0.593	0.610	0.618	0.623	0.626	0.632	0.640
2.34	0.538	0.600	0.618	0.626	0.630	0.634	0.639	0.648
2.36	0.545	0.608	0.625	0.633	0.638	0.641	0.647	0.655
2.38	0.552	0.615	0.632	0.640	0.645	0.648	0.654	0.662
2.40	0.559	0.622	0.640	0.648	0.652	0.655	0.661	0.670
2.42	0.566	0.629	0.647	0.655	0.660	0.663	0.669	0.677
2.44	0.573	0.636	0.654	0.662	0.667	0.670	0.676	0.684
2.46	0.580	0.643	0.661	0.669	0.674	0.677	0.683	0.691
2.48	0.586	0.650	0.668	0.676	0.681	0.684	0.690	0.698
2.50	0.593	0.657	0.675	0.683	0.688	0.691	0.697	0.705
2.52	0.600	0.664	0.682	0.690	0.695	0.698	0.704	0.712
2.54	0.606	0.671	0.689	0.697	0.702	0.705	0.711	0.719
2.56	0.613	0.678	0.695	0.704	0.708	0.711	0.717	0.725
2.58	0.620	0.685	0.702	0.710	0.715	0.718	0.724	0.732
2.60	0.626	0.691	0.709	0.717	0.722	0.725	0.731	0.739
2.62	0.632	0.698	0.715	0.724	0.728	0.731	0.737	0.745
2.64	0.639	0.704	0.722	0.730	0.735	0.738	0.743	0.751
2.66	0.646	0.711	0.728	0.736	0.741	0.744	0.750	0.758
2.68	0.652	0.717	0.735	0.743	0.747	0.750	0.756	0.764
2.70	0.659	0.724	0.741	0.749	0.754	0.757	0.762	0.770
2.72	0.665	0.730	0.747	0.755	0.760	0.763	0.768	0.776
2.74	0.671	0.736	0.753	0.763	0.766	0.769	0.774	0.782
2.76	0.678	0.742	0.759	0.767	0.772	0.775	0.780	0.788
2.78	0.684	0.748	0.765	0.773	0.778	0.780	0.786	0.794
2.80	0.690	0.754	0.771	0.779	0.783	0.786	0.792	0.799
2.82	0.696	0.760	0.777	0.785	0.789	0.792	0.797	0.804

(Continued)

Table 2 Values of the Power Function $p(\theta)$ (Continued)

$ \theta $	n							
	10	20	30	40	50	60	100	∞
2.84	0.702	0.766	0.783	0.790	0.795	0.798	0.803	0.810
2.86	0.708	0.772	0.788	0.796	0.800	0.803	0.808	0.815
2.88	0.714	0.777	0.794	0.801	0.806	0.808	0.814	0.820
2.90	0.720	0.783	0.799	0.807	0.811	0.814	0.819	0.825
2.92	0.725	0.789	0.805	0.812	0.816	0.819	0.824	0.830
2.94	0.731	0.794	0.810	0.817	0.821	0.824	0.829	0.833
2.96	0.737	0.799	0.815	0.822	0.826	0.829	0.834	0.840
2.98	0.742	0.805	0.820	0.827	0.831	0.834	0.839	0.845
3.00	0.748	0.810	0.825	0.832	0.836	0.839	0.844	0.850
3.02	0.753	0.815	0.830	0.837	0.841	0.844	0.849	0.854
3.04	0.759	0.820	0.835	0.842	0.846	0.848	0.853	0.860
3.06	0.764	0.825	0.840	0.847	0.850	0.853	0.858	0.864
3.08	0.770	0.830	0.844	0.851	0.855	0.857	0.862	0.867
3.10	0.775	0.834	0.849	0.856	0.859	0.862	0.866	0.872
3.12	0.780	0.839	0.853	0.860	0.864	0.866	0.871	0.876
3.14	0.785	0.843	0.858	0.864	0.868	0.870	0.875	0.880
3.16	0.790	0.848	0.862	0.868	0.872	0.874	0.879	0.884
3.18	0.795	0.852	0.866	0.873	0.876	0.878	0.883	0.888
3.20	0.800	0.857	0.870	0.877	0.880	0.882	0.887	0.891
3.22	0.805	0.861	0.874	0.881	0.884	0.886	0.890	0.895
3.24	0.809	0.865	0.878	0.884	0.888	0.890	0.894	0.899
3.26	0.814	0.869	0.882	0.888	0.891	0.894	0.898	0.902
3.28	0.819	0.873	0.886	0.892	0.895	0.897	0.901	0.906
3.30	0.823	0.877	0.890	0.895	0.899	0.901	0.904	0.910
3.32	0.827	0.881	0.893	0.899	0.902	0.904	0.908	0.913
3.34	0.832	0.884	0.897	0.902	0.905	0.907	0.911	0.916
3.36	0.836	0.888	0.900	0.906	0.909	0.911	0.914	0.919
3.38	0.840	0.892	0.904	0.909	0.912	0.914	0.917	0.923
3.40	0.844	0.895	0.907	0.912	0.915	0.917	0.920	0.925
3.42	0.849	0.898	0.910	0.915	0.918	0.920	0.923	0.928
3.44	0.853	0.902	0.913	0.918	0.921	0.923	0.926	0.930
3.46	0.856	0.905	0.916	0.921	0.924	0.925	0.929	0.932
3.48	0.860	0.908	0.919	0.924	0.926	0.928	0.931	0.935
3.50	0.864	0.911	0.922	0.927	0.929	0.931	0.934	0.937
3.52	0.868	0.914	0.925	0.929	0.932	0.933	0.936	0.940
3.54	0.871	0.917	0.927	0.932	0.934	0.936	0.939	0.942
3.56	0.875	0.920	0.930	0.934	0.937	0.938	0.941	0.944
3.58	0.879	0.923	0.933	0.937	0.939	0.941	0.943	0.947
3.60	0.882	0.925	0.935	0.939	0.941	0.943	0.946	0.949
3.62	0.885	0.928	0.937	0.941	0.944	0.945	0.948	0.951
3.64	0.889	0.931	0.940	0.944	0.946	0.947	0.950	0.953
3.66	0.892	0.933	0.942	0.946	0.948	0.949	0.952	0.955
3.68	0.895	0.935	0.944	0.948	0.950	0.951	0.954	0.957
3.70	0.898	0.938	0.946	0.950	0.952	0.953	0.956	0.959

(Continued)

Table 2 Values of the Power Function $p(\theta)$ (Continued)

θ	n							
	10	20	30	40	50	60	100	∞
3.72	0.901	0.940	0.948	0.952	0.954	0.955	0.958	0.961
3.74	0.904	0.942	0.950	0.954	0.956	0.957	0.959	0.963
3.76	0.907	0.944	0.952	0.956	0.958	0.959	0.961	0.965
3.78	0.910	0.946	0.954	0.958	0.959	0.961	0.963	0.966
3.80	0.913	0.949	0.956	0.959	0.961	0.962	0.964	0.968
3.82	0.915	0.950	0.958	0.961	0.963	0.964	0.966	0.969
3.84	0.918	0.952	0.960	0.963	0.964	0.965	0.967	0.971
3.86	0.920	0.954	0.961	0.964	0.966	0.967	0.969	0.972
3.88	0.923	0.956	0.963	0.966	0.967	0.968	0.970	0.973
3.90	0.925	0.958	0.964	0.967	0.969	0.970	0.971	0.975
3.92	0.928	0.959	0.966	0.968	0.970	0.971	0.973	0.976
3.94	0.930	0.961	0.967	0.970	0.971	0.972	0.974	0.977
3.96	0.932	0.963	0.969	0.971	0.973	0.973	0.975	0.978
3.98	0.935	0.964	0.970	0.972	0.974	0.975	0.976	0.979
4.00	0.937	0.966	0.971	0.974	0.975	0.976	0.977	0.980
4.02	0.939	0.967	0.972	0.975	0.976	0.977	0.978	0.981
4.04	0.941	0.968	0.974	0.976	0.977	0.978	0.979	0.982
4.06	0.943	0.970	0.975	0.977	0.978	0.979	0.980	0.983
4.08	0.945	0.971	0.976	0.978	0.979	0.980	0.981	0.984

receive the placebo at the first period and the treatment at the second period. Let x_{ij} be the normally distributed observation from the j th patient at the i th period and l th sequence. Then the treatment effect μ_D can be unbiasedly estimated by

$$\hat{\mu}_D = \frac{\bar{x}_{11} - \bar{x}_{12} - \bar{x}_{21} + \bar{x}_{22}}{2} \sim N\left(\mu_D, \frac{\sigma_{1,1}^2}{4} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

where $\bar{x}_{ij}$ is the sample mean based on x_{ij} , $j = 1, \dots, n_i$, and $\sigma_{1,1}^2$ is a function of variance components, where $\sigma_{1,1}^2 = \sigma_D^2 + \sigma_{WT}^2 + \sigma_{WR}^2$ in which σ_{WT}^2 and σ_{WR}^2 are intrasubject variances for the test product and the reference product, respectively. An unbiased estimator of $\sigma_{1,1}^2$ is

$$\hat{\sigma}_{1,1}^2 = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^2 \sum_{j=1}^{n_j} (x_{i1j} - x_{i2j} - \bar{x}_{i1} + \bar{x}_{i2})^2$$

which is independent of $\hat{\mu}_D$ and distributed as $\sigma_{1,1}^2 / (n_1 + n_2 - 2)$ times the χ^2 distribution with $n_1 + n_2 - 2$ degrees of freedom. Thus the null hypothesis $H_0: \mu_D = 0$ is rejected at the 5% level of significance if $|T| > t_{0.975, n-2}$, where $n = n_1 + n_2$ and

$$T = \frac{\hat{\mu}_D}{\frac{\hat{\sigma}_{1,1}}{2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (16)$$

Note that T has the noncentral t -distribution with $n - 2$ degrees of freedom and the noncentrality parameter

$$\theta = \frac{\mu_D}{\frac{\sigma_{1,1}}{2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (17)$$

Consequently, the reproducibility probability obtained by using the estimated power approach is given by Eq. 11 with T defined by Eq. 16.

Parallel-Group Designs

Parallel-group designs are often employed in clinical trials to compare more than one treatment with a placebo control or to compare one treatment, one placebo control, and one active control. Let $a \geq 3$ be the number of groups and x_{ij} be the observation from the j th patient in the i th group, $j = 1, \dots, n_i$, $i = 1, \dots, a$. Assume that x_{ij} 's are independently distributed as $N(\mu_i, \sigma^2)$. The null hypothesis H_0 is then

$$H_0: \mu_1 = \mu_2 = \dots = \mu_a$$

which is rejected at the 5% level of significance if $T > F_{0.95; a-1, n-a}$, where $F_{0.95; a-1, n-a}$ is the 95th percentile of the F -distribution with $a-1$ and $n-a$ degrees of freedom, $n = n_1 + \dots + n_a$,

$$T = \frac{SST/(a-1)}{SSE/(n-a)}, \quad SST = \sum_{i=1}^a n_i (\bar{x}_i - \bar{x})^2, \quad (18)$$

$$= SSE = \sum_{i=1}^a \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

$\bar{x}_i$ is the sample mean based on the data in the i th group, and is the overall sample mean.

Note that T has the noncentral F -distribution with $a-1$ and $n-a$ degrees of freedom and the noncentrality parameter

$$\theta = \sum_{i=1}^a \frac{n_i (\mu_i - \bar{\mu})^2}{\sigma^2} \quad (19)$$

where $\bar{\mu} = \sum_{i=1}^a n_i \mu_i / n$. Let $F_{a-1, n-a}(\cdot | \theta)$ be the distribution function of T . Then, the power of the second clinical trial is

$$P(T(y) > F_{0.95; a-1, n-a}) = 1 - F_{a-1, n-a}(F_{0.95; a-1, n-a} | \theta)$$

Thus the reproducibility probability obtained by using the estimated power approach is

$$\hat{P} = 1 - F_{a-1, n-a}(F_{0.95; a-1, n-a} | T(x))$$

where $T(x)$ is the observed T based on the data x from the first clinical trial.

THE CONFIDENCE BOUND APPROACH

Because $\hat{P}$ in Eq. 8 or Eq. 11 is an estimated power, it provides a rather optimistic result. Alternatively, we may consider a more conservative approach, which considers a 95% lower confidence bound of the power as the reproducibility probability.^[3]

Consider first the case of the two-group parallel design with a known common variance σ^2 . Note that a 95% lower confidence bound for $|\theta|$, where θ is given by Eq. 7, is

$$|T(x)| - z_{0.975}$$

where T is given by Eq. 5. Because

$$P(|T(y)| > z_{0.975}) \geq \Phi(|\theta| - z_{0.975})$$

and Φ is a strictly increasing function, a 95% lower confidence bound for the power is

$$\hat{P}_- = \Phi(|T(x)| - 2z_{0.975}) \quad (20)$$

Note that the lower confidence bound in Eq. 20 is not useful unless $|T(x)| \geq 2z_{0.975}$, i.e., the positive clinical result from the first trial is highly significant.

When σ^2 is unknown, it is more difficult to obtain a confidence bound for θ in Eq. 7. Note that $T(x)$ defined by Eq. 9 has the noncentral t -distribution with $n-2$ degrees of freedom and the noncentrality parameter θ . Let $\mathcal{T}_{n-2}(\cdot | \theta)$ be the distribution function of $T(x)$ for any given θ . It can be shown that $\mathcal{T}_{n-2}(t | \theta)$ is a strictly decreasing function of θ for any fixed t . Consequently, a 95% confidence interval for θ is $(\hat{\theta}_-, \hat{\theta}_+)$, where $\hat{\theta}_\pm$ is the unique solution of

$$\mathcal{T}_{n-2}(T(x) | \theta) = 0.975$$

and $\hat{\theta}_+$ is the unique solution of

$$\mathcal{T}_{n-2}(T(x) | \theta) = 0.025$$

(see, e.g., Theorem 7.1 in Ref. [4]). Then, a 95% lower confidence bound for $|\theta|$ is

$$|\hat{\theta}|_- = \begin{cases} \hat{\theta}_- & \text{if } \hat{\theta}_- > 0 \\ -\hat{\theta}_+ & \text{if } \hat{\theta}_+ < 0 \\ 0 & \text{if } \hat{\theta}_- \leq 0 \leq \hat{\theta}_+ \end{cases} \quad (21)$$

and a 95% lower confidence bound for the power $p(\theta)$ in Eq. 10 is

$$\hat{P}_- = 1 - \mathcal{T}_{n-2}(t_{0.975; n-2} | |\hat{\theta}|_-) + \mathcal{T}_{n-2}(-t_{0.975; n-2} | |\hat{\theta}|_-) \quad (22)$$

if $|\hat{\theta}|_- > 0$ and $\hat{P}_- = 0$ if $|\hat{\theta}|_- = 0$. Again, the lower confidence bound in Eq. 22 is useful when the clinical result from the first trial is highly significant.

Table 3 contains values of the lower confidence bound $|\hat{\theta}|_-$ corresponding to $|T(x)|$ values ranging from 4.0 to 6.5. If $4 \leq |T(x)| \leq 6.5$ and the value of $|\hat{\theta}|_-$ is found in Table 3, the reproducibility probability $\hat{P}_-$ in Eq. 22 can be obtained in Table 2. For example, suppose that $|T(x)| = 5$ was observed from a clinical trial with $n = 30$. From Table 3, $|\hat{\theta}|_- = 2.6$. Then, by Table 2, $\hat{P}_- = 0.709$.

While $\hat{P}$ in Eq. 11 is rather optimistic, $\hat{P}_-$ in Eq. 22 may be too conservative. One compromise is to reduce the significance level of the lower confidence bound $|\hat{\theta}|_-$ to 90%. Table 4 contains values of the 90% lower confidence bound for $|\theta|$. For $n = 30$ and $|T(x)| = 5$, $|\hat{\theta}|_-$ is 2.98, which leads to $\hat{P}_- = 0.82$.

Consider next the two-group parallel design with unequal variances σ_1^2 and σ_2^2 . When both n_1 and n_2 are large, T given by Eq. 12 is approximately distributed as $N(\theta, 1)$ with θ given by Eq. 13. Hence the reproducibility probability obtained by using the lower confidence bound approach is given by Eq. 20 with T defined by Eq. 12.

Table 3 95% Lower Confidence Bound $|\hat{\theta}|$

$ T(x) $	n							
	10	20	30	40	50	60	100	∞
4.5	1.51	2.01	2.18	2.26	2.32	2.35	2.42	2.54
4.6	1.57	2.09	2.26	2.35	2.41	2.44	2.52	2.64
4.7	1.64	2.17	2.35	2.44	2.50	2.54	2.61	2.74
4.8	1.70	2.25	2.43	2.53	2.59	2.63	2.71	2.84
4.9	1.76	2.33	2.52	2.62	2.68	2.72	2.80	2.94
5.0	1.83	2.41	2.60	2.71	2.77	2.81	2.90	3.04
5.1	1.89	2.48	2.69	2.80	2.86	2.91	2.99	3.14
5.2	1.95	2.56	2.77	2.88	2.95	3.00	3.09	3.24
5.3	2.02	2.64	2.86	2.97	3.04	3.09	3.18	3.34
5.4	2.08	2.72	2.95	3.06	3.13	3.18	3.28	3.44
5.5	2.14	2.80	3.03	3.15	3.22	3.27	3.37	3.54
5.6	2.20	2.88	3.11	3.24	3.31	3.36	3.47	3.64
5.7	2.26	2.95	3.20	3.32	3.40	3.45	3.56	3.74
5.8	2.32	3.03	3.28	3.41	3.49	3.55	3.66	3.84
5.9	2.39	3.11	3.37	3.50	3.58	3.64	3.75	3.94
6.0	2.45	3.19	3.45	3.59	3.67	3.73	3.85	4.04
6.1	2.51	3.26	3.53	3.67	3.76	3.82	3.94	4.14
6.2	2.57	3.34	3.62	3.76	3.85	3.91	4.03	4.24
6.3	2.63	3.42	3.70	3.85	3.94	4.00	4.13	4.34
6.4	2.69	3.49	3.78	3.93	4.03	4.09	4.22	4.44
6.5	2.75	3.57	3.86	4.02	4.12	4.18	4.32	4.54

For the matched-pair parallel design described above, T given by Eq. 14 has the noncentral t -distribution with $m - 1$ degrees of freedom and the noncentrality parameter θ given by Eq. 15. Hence the reproducibility probability obtained by using the lower confidence bound approach is given by Eq. 22 with T defined by Eq. 14 and $n - 2$ replaced by $m - 1$.

Suppose now that the study design is the 2×2 crossover design. Because T defined by Eq. 16 has the noncentral t -distribution with $n - 2$ degrees of freedom and the noncentrality parameter θ given by Eq. 17, the reproducibility probability obtained by using the lower confidence bound approach is given by Eq. 22 with T defined by Eq. 16.

Finally, consider the parallel-group design described above. Because T in Eq. 18 has the noncentral F -distribution with $a - 1$ and $n - a$ degrees of freedom and the noncentrality parameter θ given by Eq. 19 and $F_{a-1, n-a}(t | \theta)$ is a strictly decreasing function of θ , the reproducibility probability obtained by using the lower confidence bound approach is

$$\hat{P}_- = 1 - F_{a-1, n-a}(F_{0.95; a-1, n-a} | \hat{\theta}_-)$$

where $\hat{\theta}_-$ is the solution of

$$F_{a-1, n-a}(T(x) | \theta) = 0.95$$

THE BAYESIAN APPROACH

As discussed earlier, the reproducibility probability can be viewed as the posterior mean of the power function $p(\theta) = P(|T| > c | \theta)$ for the second trial. Thus under the Bayesian approach, it is essential to construct the posterior density $\pi(\theta | x)$ in formula 2, given the dataset x observed from the first trial.^[3]

Two Samples with Equal Variances

Consider the two-group parallel design described above with equal variances, i.e., x_{ij} 's are independent and normally distributed with means μ_1 and μ_2 and a common variance σ^2 .

If σ^2 is known, then the power for testing hypotheses in Eq. 3 is given by Eq. 6 with θ defined by Eq. 7. A commonly used prior for (μ_1, μ_2) is the noninformative prior $\pi(\mu_1, \mu_2) \equiv 1$. Consequently, the posterior density for θ is $N(T(x), 1)$, where T is given by Eq. 5, and the posterior mean given by Eq. 2 is

$$\begin{aligned} & \int [\Phi(\theta - z_{0.975}) + \Phi(-\theta - z_{0.975})] \pi(\theta | x) d\theta \\ &= \Phi\left(\frac{T(x) - z_{0.975}}{\sqrt{2}}\right) + \Phi\left(\frac{-T(x) - z_{0.975}}{\sqrt{2}}\right) \end{aligned}$$

Table 4 90% Lower Confidence Bound $|\hat{\theta}|_L$

$ T(x) $	N							
	10	20	30	40	50	60	100	∞
4.5	1.95	2.40	2.54	2.62	2.67	2.70	2.76	2.86
4.6	2.02	2.48	2.63	2.71	2.76	2.79	2.85	2.96
4.7	2.09	2.56	2.72	2.80	2.85	2.88	2.95	3.06
4.8	2.16	2.64	2.81	2.89	2.94	2.98	3.05	3.16
4.9	2.22	2.73	2.90	2.98	3.03	3.07	3.14	3.26
5.0	2.29	2.81	2.98	3.07	3.13	3.16	3.24	3.36
5.1	2.36	2.89	3.07	3.16	3.22	3.26	3.33	3.46
5.2	2.43	2.97	3.16	3.25	3.31	3.35	3.43	3.56
5.3	2.49	3.05	3.24	3.34	3.40	3.44	3.52	3.66
5.4	2.56	3.13	3.33	3.43	3.49	3.54	3.62	3.76
5.5	2.63	3.21	3.42	3.52	3.59	3.63	3.72	3.86
5.6	2.69	3.30	3.50	3.61	3.68	3.72	3.81	3.96
5.7	2.76	3.38	3.59	3.70	3.77	3.81	3.91	4.06
5.8	2.83	3.46	3.68	3.79	3.86	3.91	4.00	4.16
5.9	2.89	3.54	3.76	3.88	3.95	3.00	4.10	4.26
6.0	2.96	3.62	3.85	3.97	4.04	4.09	4.19	4.36
6.1	3.02	3.70	3.93	4.06	4.13	4.18	4.29	4.46
6.2	3.09	3.78	4.02	4.15	4.22	4.28	4.38	4.56
6.3	3.15	3.86	4.11	4.23	4.31	4.37	4.48	4.66
6.4	3.22	3.94	4.19	4.32	4.40	4.46	4.57	4.76
6.5	3.28	4.02	4.28	4.41	4.49	4.55	4.67	4.86

When $|T(x)| > z_{0.975}$, this probability is nearby the same as

$$\Phi\left(\frac{|T(x)| - z_{0.975}}{\sqrt{2}}\right) \quad (23)$$

which is exactly the same as that in formula 1 in Goodman.^[2]

For the more realistic situation where σ^2 is unknown, we need a prior for σ^2 . A commonly used noninformative prior for σ^2 is the Lebesgue (improper) density $\pi(\sigma^2) = \sigma^{-2}$. Assume that the priors for μ_1 , μ_2 , and σ^2 are independent. The posterior density for (δ, u^2) is

$$\pi(\delta|u^2, x)\pi(u^2|x)$$

where

$$\delta = \frac{\mu_1 - \mu_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}},$$

$$u^2 = \frac{(n - 2)\sigma^2}{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}, \quad \pi(\delta|u^2, x) = \frac{1}{u} \phi\left(\frac{\delta - T(x)}{u}\right)$$

ϕ is the density function of the standard normal distribution, T is given by Eq. 9, and $\pi(u^2|x) = f(u)$ with

$$f(u) = \left[\Gamma\left(\frac{n-2}{2}\right) \right]^{-1} \left(\frac{n-2}{2}\right)^{(n-2)/2} u^{-n} e^{-(n-2)/(2u^2)}$$

Because θ in Eq. 7 is equal to δ/u , the posterior mean of $p(\theta)$ in Eq. 10 is

$$\hat{P} = \int_0^\infty \left[\int_{-\infty}^\infty p\left(\frac{\delta}{u}\right) \phi\left(\frac{\delta - T(x)}{u}\right) d\delta \right] 2f(u) du \quad (24)$$

which is the reproducibility probability under the Bayesian approach. It is clear that depends on the data x through the function $T(x)$.

The probability $\hat{P}$ in Eq. 24 can be evaluated numerically. A Monte Carlo method can be applied as follows. First, generate a random variate γ_j from the gamma distribution with the shape parameter $(n-2)/2$ and the scale parameter $2/(n-2)$, and generate a random variate δ_j from $N(T(x), u_j^2)$, where $u_j^2 = \gamma_j^{-1}$. Repeat this process independently N times to obtain (δ_j, u_j^2) , $j = 1, \dots, N$. Then in Eq. 24 can be approximated by

$$\hat{P}_N = 1 - \frac{1}{N} \sum_{j=1}^N \left[T_{n-2} \left(t_{0.975; n-2} \left| \frac{\delta_j}{u_j} \right| \right) - T_{n-2} \left(-t_{0.975; n-2} \left| \frac{\delta_j}{u_j} \right| \right) \right] \quad (25)$$

Values of $\hat{P}_N$ for $N = 10,000$ and some selected values of $T(x)$ and n are given by Table 5. It can be seen that in assessing reproducibility, the Bayesian approach is more conservative than the estimated power approach, but less conservative than the confidence bound approach.

Two Samples with Unequal Variances

Consider first the two-group parallel design with unequal variance and large n_i 's. The approximate power for the second trial is

$$p(\theta) = \Phi(\theta - z_{0.975}) + \Phi(-\theta - z_{0.975}),$$

$$\theta = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Suppose that we use the following noninformative prior density for $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$:

$$\pi(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \sigma_1^{-2} \sigma_2^{-2}, \quad \sigma_1^2 > 0, \sigma_2^2 > 0$$

Let $\tau_i^2 = \sigma_i^{-2}, i = 1, 2$, and $\zeta^2 = 1/n_1 \tau_1^2 + 1/n_2 \tau_2^2$. Then, the posterior density $\pi(\mu_1 - \mu_2 | \tau_1^2, \tau_2^2, x)$ is the normal density with mean $1 - 2$ and variance ζ^2 and the posterior density

$$\pi(\tau_1^2, \tau_2^2 | x) = \pi(\tau_1^2 | x) \pi(\tau_2^2 | x)$$

where $\pi(\tau_i^2 | x)$ is the gamma density with the shape parameter $(n_i - 1)/2$ and the scale parameter $2/[(n_i - 1) s_i^2], i = 1, 2$. Consequently, the posterior mean of $p(\theta)$ is

$$\begin{aligned} & \iiint \Phi(\theta - z_{0.975}) \pi(\mu_1 - \mu_2 | \tau_1^2, \tau_2^2, x) d(\mu_1 - \mu_2) \\ & \times \pi(\tau_1^2, \tau_2^2 | x) d\tau_1^2 d\tau_2^2 + \iiint \Phi(-\theta - z_{0.975}) \\ & \times \pi(\mu_1 - \mu_2 | \tau_1^2, \tau_2^2, x) d(\mu_1 - \mu_2) \pi(\tau_1^2, \tau_2^2 | x) d\tau_1^2 d\tau_2^2 \end{aligned}$$

Note that

$$\begin{aligned} & \int \Phi(\theta - z_{0.975}) \pi(\mu_1 - \mu_2 | \tau_1^2, \tau_2^2, x) d(\mu_1, \mu_2) \\ & = \Phi\left(\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{2}\zeta} - \frac{z_{0.975}}{\sqrt{2}}\right) \end{aligned}$$

and

$$\begin{aligned} & \int \Phi(-\theta - z_{0.975}) \pi(\mu_1 - \mu_2 | \tau_1^2, \tau_2^2, x) d(\mu_1, \mu_2) \\ & = \Phi\left(-\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{2}\zeta} - \frac{z_{0.975}}{\sqrt{2}}\right) \end{aligned}$$

Hence the reproducibility probability based on the Bayesian approach is

$$\hat{P} = \int \left[\Phi\left(\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{2}\zeta} - \frac{z_{0.975}}{\sqrt{2}}\right) + \Phi\left(-\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{2}\zeta} - \frac{z_{0.975}}{\sqrt{2}}\right) \right] \times \pi(\zeta | x) d\zeta$$

where $\pi(\zeta | x)$ is the posterior density of ζ constructed using $\pi(\tau_i^2 | x), i = 1, 2$.

The following Monte Carlo method can be applied to evaluate. Generate independent random variates τ_{1j}^2 and τ_{2j}^2 , respectively, from the gamma densities $\pi(\tau_1^2 | x)$ and $\pi(\tau_2^2 | x)$ as previously specified. Repeat this process independently N times to obtain $(\tau_{1j}^2, \tau_{2j}^2), j = 1, \dots, N$. Then $\hat{P}$ can be approximated by

$$\hat{P}_N = \frac{1}{N} \sum_{j=1}^N \left[\Phi\left(\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{2}\zeta_j} - \frac{z_{0.975}}{\sqrt{2}}\right) + \Phi\left(-\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{2}\zeta_j} - \frac{z_{0.975}}{\sqrt{2}}\right) \right]$$

where

$$\zeta_j = \frac{1}{n_1 \tau_{1j}^2} + \frac{1}{n_2 \tau_{2j}^2}, \quad j = 1, \dots, N$$

For the matched-pairs parallel design, the power function is

$$p(\theta) = 1 - T_{m-1}(t_{0.975; m-1} | \theta) + T_{m-1}(-t_{0.975; m-1} | \theta)$$

with θ given by Eq. 15. Under the noninformative prior $\pi(\mu, \sigma_D^2) = \sigma_D^{-1}$, the reproducibility probability, the posterior mean of $p(\theta)$, is given by Eq. 24 with T given by Eq. 14, $n - 2$ replaced by $m - 1$, and δ and u^2 defined by

$$\delta = \frac{\sqrt{n}(\mu_1 - \mu_2)}{\sigma_D} \quad \text{and} \quad \mu^2 = \frac{\sigma_D^2}{\hat{\sigma}_D^2}$$

respectively.

For the 2×2 crossover design, the power function is $p(\theta) = 1 - T_{n-2}(t_{0.975; n-2} | \theta) + T_{n-2}(-t_{0.975; n-2} | \theta)$ with θ given by Eq. 17. If we use the noninformative prior $\pi(\mu, \sigma_{1,1}^2) = \sigma_{1,1}^{-1}$, then the reproducibility probability is still given by Eq. 24 with T given by Eq. 16, δ and u^2 defined by

Table 5 Reproducibility Probability $\hat{P}_N$ Given by Eq. 25

$ T(x) $	N							
	10	20	30	40	50	60	100	∞
2.02	0.435	0.482	0.495	0.503	0.504	0.508	0.517	0.519
2.08	0.447	0.496	0.512	0.515	0.519	0.523	0.532	0.536
2.14	0.466	0.509	0.528	0.530	0.535	0.543	0.549	0.553
2.20	0.478	0.529	0.540	0.547	0.553	0.556	0.565	0.569
2.26	0.487	0.547	0.560	0.564	0.567	0.571	0.577	0.585
2.32	0.505	0.558	0.577	0.580	0.581	0.587	0.590	0.602
2.38	0.519	0.576	0.590	0.597	0.603	0.604	0.610	0.618
2.44	0.530	0.585	0.610	0.611	0.613	0.617	0.627	0.634
2.50	0.546	0.609	0.624	0.631	0.634	0.636	0.640	0.650
2.56	0.556	0.618	0.638	0.647	0.648	0.650	0.658	0.665
2.62	0.575	0.632	0.654	0.655	0.657	0.664	0.675	0.680
2.68	0.591	0.647	0.665	0.674	0.675	0.677	0.687	0.695
2.74	0.600	0.660	0.679	0.685	0.686	0.694	0.703	0.710
2.80	0.608	0.675	0.690	0.702	0.705	0.712	0.714	0.724
2.86	0.629	0.691	0.706	0.716	0.722	0.723	0.729	0.738
2.92	0.636	0.702	0.718	0.730	0.733	0.738	0.742	0.752
2.98	0.649	0.716	0.735	0.742	0.744	0.748	0.756	0.765
3.04	0.663	0.726	0.745	0.753	0.756	0.759	0.765	0.778
3.10	0.679	0.738	0.754	0.766	0.771	0.776	0.779	0.790
3.16	0.690	0.754	0.767	0.776	0.781	0.786	0.792	0.802
3.22	0.701	0.762	0.777	0.790	0.792	0.794	0.804	0.814
3.28	0.708	0.773	0.793	0.804	0.806	0.809	0.820	0.825
3.34	0.715	0.784	0.803	0.809	0.812	0.818	0.828	0.836
3.40	0.729	0.793	0.815	0.819	0.829	0.830	0.838	0.846
3.46	0.736	0.806	0.826	0.832	0.837	0.839	0.847	0.856
3.52	0.745	0.816	0.834	0.843	0.845	0.846	0.855	0.865
3.58	0.755	0.828	0.841	0.849	0.857	0.859	0.867	0.874
3.64	0.771	0.833	0.854	0.859	0.863	0.865	0.872	0.883
3.70	0.778	0.839	0.861	0.867	0.870	0.874	0.884	0.891
3.76	0.785	0.847	0.867	0.874	0.882	0.883	0.890	0.898
3.82	0.795	0.857	0.878	0.883	0.889	0.891	0.898	0.906
3.88	0.800	0.869	0.881	0.891	0.896	0.899	0.904	0.913
3.94	0.806	0.873	0.890	0.897	0.904	0.907	0.910	0.919
4.00	0.812	0.883	0.896	0.905	0.908	0.911	0.916	0.925
4.06	0.822	0.888	0.902	0.910	0.915	0.917	0.924	0.931

$$\delta = \frac{\mu_D}{\frac{\sigma_{1,1}^2}{2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{and} \quad u^2 = \frac{\sigma_{1,1}^2}{\hat{\sigma}_{1,1}^2}$$

respectively.

Parallel-Group Designs

Consider the a -group parallel design, where the power is given by

$$p(\theta) = 1 - F_{a-1, n-a}(F_{0.95; a-1, n-a} | \theta)$$

with θ given by Eq. 19. Under the noninformative prior

$$\pi(\mu_1, \dots, \mu_a, \sigma^2) = \sigma^{-2}, \sigma^2 > 0$$

the posterior density $\pi(\theta | \tau^2, x)$, where $\tau^2 = \text{SSE} / [(n - a)\sigma^2]$, is the density of the noncentral χ^2 distribution with $a - 1$ degrees of freedom and the noncentrality parameter $\tau^2(a - 1)T(x)$. The posterior density $\pi(\tau^2 | x)$ is the gamma

distribution with the shape parameter $(n - a)/2$ and the scale parameter $2/(n - a)$. Consequently, the reproducibility probability under the Bayesian approach is

$$\hat{P} = \int_0^\infty \left[\int_{-\infty}^\infty p(\theta) \pi(\theta | \tau^2, x) d\theta \right] \pi(\tau^2 | x) d\tau^2$$

A similar Monte Carlo method can be applied to compute $\hat{P}$.

APPLICATION AND EXTENSIONS

In this section, we discuss some applications and extensions of the results obtained in the previous sections. More details can be found in Chow and Shao^[5] and Chow and Liu.^[6]

Substantial Evidence with a single Trial

An important application of the results derived in the previous sections is to address the following question: is it necessary to conduct a second clinical trial when the first trial produces a relatively strong positive clinical result (e.g., a relatively small p -value is observed)? As indicated earlier, the FDAMA of 1997 includes a provision stating that under certain circumstances, data from one adequate and well-controlled clinical investigation and confirmatory evidence may be sufficient to establish effectiveness for risk/benefit assessment of drug and biological candidates for approval. This provision essentially codified an FDA policy that had existed for several years but whose application had been limited to some biological products approved by the Center for Biologic Evaluation and Research (CBER) of the FDA and a few pharmaceuticals, especially orphan drugs such as zidovudine and lamotrigine. As can be observed in Table 1, a relatively strong significant result observed from a single clinical trial (say, the p -value is less than 0.001) would have about 90% chance of reproducing the result in future clinical trials. Consequently, a single clinical trial is sufficient to provide substantial evidence for demonstration of efficacy and safety of the medication under study. In 1998, FDA published a guidance which shed the light on this approach despite that the FDA has recognized that advances in sciences and practice of drug development may permit an expanded role for the single controlled trial in contemporary clinical development.^[1]

Suppose that it is agreed that the second trial is not needed if the probability for reproducing a positive clinical result in the second trial is equal to or higher than 90%. If a positive clinical result is observed in the first trial and the confidence bound $\hat{P}_-$ is equal to or higher than 90%, then we have 95% statistical assurance that, with a probability of at least 90%, the positive result will be reproduced in the second trial. For example, under the two-group parallel design with a common unknown variance and $n = 40$, the

95% lower confidence bound $\hat{P}_-$ given by Eq. 20 is equal to or higher than 90% if and only if $|T(x)| \geq 5.7$, i.e., the clinical result in the first trial is highly significant. Alternatively, if the Bayesian approach is applied to the same situation, the reproducibility probability in Eq. 23 is equal to or higher than 90% if and only if $|T(x)| \geq 3.96$.

On the other hand, if the reproducibility probability is very low (e.g., $\hat{P}$ in Eq. 11 is less than 50%), it may also indicate that a clinically meaningful difference between the placebo and the drug product under study cannot be detected through the intended trials and, thus, there is no need to run the second trial.

Reproducibility Probabilities for K Trials

Consider the situation where a number of clinical trials are conducted in a sequential manner. Our interest is the reproducibility probability for the k th trial, given that all previous $k - 1$ trials produce positive clinical results.

The extension of the results under the Bayesian approach is straightforward. Let $\pi_0(\theta | x_0) = \pi(\theta)$ be the prior for θ (which is usually a noninformative prior) and $\pi_k(\theta | x_k)$ be the posterior after observing the data from the k th trial, $k = 1, 2, \dots$. Then $\pi_k(\theta | x_k)$ can be used as the prior for the $(k + 1)$ th trial. The reproducibility probability for the $(k + 1)$ th trial is

$$P(|T| > c | x_k) = \int P(|T| > c | \theta) \pi_k(\theta | x_k) d\theta, \\ k = 0, 1, \dots$$

where x_k is the data set containing all data up to the k th trial,

For the non-Bayesian approach, the simplest way of constructing a reproducibility probability for the $(k + 1)$ th trial is to apply the procedures given in the sections “Reproducibility Probability” and “The Estimated Power Approach” with $x = x_k$, assuming that all trials have the same design.

Binary Data

In some clinical trials data are binary and μ_i' are proportions. Consider a two-group parallel design. For binary data, sample sizes n_1 and n_2 are usually large so that the Central Limit Theorem can be applied to sample proportions $\bar{x}_1$ and $\bar{x}_2$, i.e., $\hat{x}_i$ is approximately $N(\mu_i, \mu_i(1 - \mu_i)/n_i)$. Thus the results in the sections “The Estimated Power Approach,” “The Confidence Bound Approach,” and “The Bayesian Approach” for the case of unequal variances with large n can be applied.

CONCLUSION

In clinical research, it is of interest to determine whether the observed clinical results are reproducible. The assessment of reproducibility probability using the estimated power approach, confidence bound approach, or Bayesian

approach is helpful. The reproducibility probability could be a useful indication for regulatory review/approval of the drug product under investigation.

REFERENCES

1. FDA. *Guideline for Formal and Content of the Clinical and Statistical Sections of New Drug Applications*; Center for Drug Evaluation and Research; The United States Food and Drug Administration: Rockville, MD, USA, 1988.
2. Goodman, S.N. A comment on replication, p -values and evidence. *Stat. Med.* **1992**, *11*, 875–879.
3. Shao, J.; Chow, S.C. Reproducibility probability in clinical trials. *Stat. Med.* **2002**, *21*, 1727–1742.
4. Shao, J. *Mathematical Statistics*; Springer: New York, 1999.
5. Chow, S.C.; Shao, J. *Statistics in Drug Research-Methodologies and Recent Development*; Marcel Dekker, Inc.: New York, 2002.
6. Chow, S.C.; Li, J.P. *Design and Analysis of Clinical Trials-Revised and Expanded*, 2nd Ed.; John Wiley & Sons: New York, 2003.

Clinical Trial Design: Statistical Assurance

Shuyen Ho

Global Data Operations, PAREXEL International, Durham, North Carolina, U.S.A.

Abstract

The frequentist statistical power is the probability of rejecting the null hypothesis with a prespecified true clinical treatment effect. As such, power is a conditional probability conditioned on the true but actually unknown effect. The Bayesian statistical assurance is an unconditional probability of rejecting the null hypothesis, given a prior probability distribution of the unknown treatment effect. In this entry, we provide a brief description of assurance and illustrate its computation by a simple two-sample normal example.

Keywords: Statistical power; Assurance; Probability of success; Bayesian prior distribution; Conditional and unconditional probability; Bayesian clinical trial simulation.

INTRODUCTION

Traditional (frequentist) clinical trial design uses statistical power to indicate how likely a trial is going to be successful. However, more often than not, the power is rarely a precise indicator of trial success because power is simply a conditional probability that is given a fixed parameter value. In the Bayesian framework, a parameter is characterized by a probability distribution, therefore similar to the expected value of a random variable being the integral of the random variable with respect to its probability distribution; assurance is defined as an expected power with respect to a probability distribution of the unknown parameter. Assurance in clinical trial design can be useful for internal decisions.

BACKGROUND

The concept of assurance (O'Hagan and Stevens^[1]), also called “expected power” or “average power,” appeared earlier in the studies by Spiegelhalter and Freedman^[2] and Gillett^[3] and “averaged success probability” later in the study by Chuang-Stein.^[4] Also a similar concept of “reproducibility probability” using priors in assessing the success probability before replicating a clinical trial appeared in the study by Shao and Chow.^[5]

Assurance in clinical trial design was first referred by O'Hagan, Stevens, and Campbell^[6] where they “argue that assurance is an important measure of the practical utility of a proposed trial, and indeed that it will often be appropriate to choose the size of the sample (and perhaps other aspects of the design) to achieve a desired assurance, rather than to achieve a desired power conditional on an assumed treatment effect.”

A BASIC EXAMPLE

We illustrate the assurance concept by the basic two-sample hypothesis testing example from the study by O'Hagan, Stevens, and Campbell.^[6]

Suppose that in a two-treatment clinical trial with n_i patients randomized to treatment i ($i = 1, 2$), the continuous outcome x_{ij} from j^{th} patient is normally distributed as $x_{ij} \sim N(\mu_i, \sigma_i^2)$. Assuming that σ_i^2 is known, we estimate the population means with the sample means as $\bar{x}_i \sim N(\mu_i, \sigma_i^2 / n_i)$. Then, the treatment difference $\delta = \mu_1 - \mu_2$ can be estimated by $\hat{\delta} = \bar{x}_1 - \bar{x}_2$ and the standard deviation can be calculated as:

$$\tau = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \text{ i.e., } \hat{\delta} = \bar{x}_1 - \bar{x}_2 \sim N(\delta, \tau^2) \quad (1)$$

The statistical power is then calculated based on this distribution that is a conditional distribution on the unknown treatment difference δ . In a two-sided superiority trial to test the null hypothesis $H_0: \delta = 0$ against the two-sided alternative $H_a: \delta \neq 0$, the null hypothesis is rejected if $|\bar{x}_1 - \bar{x}_2| > \tau Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the upper $\alpha/2$ significance point of the standard normal distribution.

The power is:

$$\begin{aligned} &P(\text{null hypothesis is rejected}) \\ &= P\left(\bar{x}_1 - \bar{x}_2 > \tau Z_{\alpha/2}\right) + P\left(\bar{x}_1 - \bar{x}_2 < -\tau Z_{\alpha/2}\right) \\ &= \Phi\left(-Z_{\alpha/2} + \delta / \tau\right) + \Phi\left(-Z_{\alpha/2} - \delta / \tau\right) \end{aligned} \quad (2)$$

The assurance is then defined based on the expected power with a prior distribution on this unknown parameter δ . Using

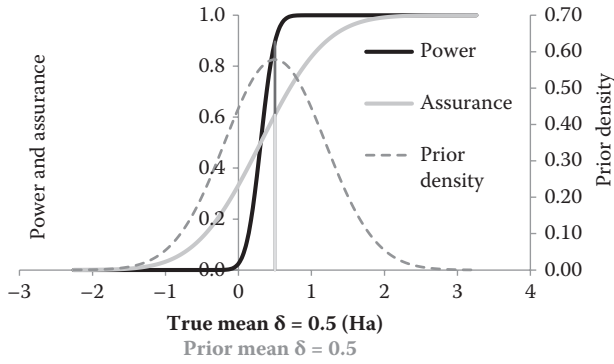


Fig. 1 Case 1: Given $\delta = 0.5$, $\tau = 0.65$, $n_1 = n_2 = 35$, power = 90%; $\delta \sim N(0.5, 0.69^2)$, assurance = 61%

a commonly used conjugated prior normal distribution from previous clinical trials as $\delta \sim N(m, v)$, the unconditional distribution can be obtained as $\bar{x}_1 - \bar{x}_2 \sim N(m, \tau^2 + v)$. The assurance that the null hypothesis is rejected can be formulated based on the unconditional distribution of $\bar{x}_1 - \bar{x}_2 \sim N(m, \tau^2 + v)$ as:

$$\begin{aligned} \gamma &= P(\text{null hypothesis is rejected}) \\ &= P\left(\bar{x}_1 - \bar{x}_2 > \tau Z_{\alpha/2}\right) + P\left(\bar{x}_1 - \bar{x}_2 < -\tau Z_{\alpha/2}\right) \\ &= \Phi\left(\frac{-\tau Z_{\alpha/2} + m}{\sqrt{\tau^2 + v}}\right) + \Phi\left(\frac{-\tau Z_{\alpha/2} - m}{\sqrt{\tau^2 + v}}\right) \end{aligned} \quad (3)$$

Now let us look at a couple of specific cases.

Case 1: If the true treatment effect $\delta = 0.5$, $\tau = 0.65$, sample size = 35 per group, and $\alpha = 0.05$, then the power is 0.9 (90%). If the prior distribution for $\delta \sim N(0.5, 0.69^2)$, then the assurance is 0.61 (61%), as illustrated in Fig. 1.

Case 2: Same as case 1, but change the prior mean to 0.2 so $\delta \sim N(0.2, 0.69^2)$, then the assurance becomes 0.44 (44%), as illustrated in Fig. 2.

As can be seen from these two cases, the prior of δ has a major impact on assurance. This is similar to other Bayesian

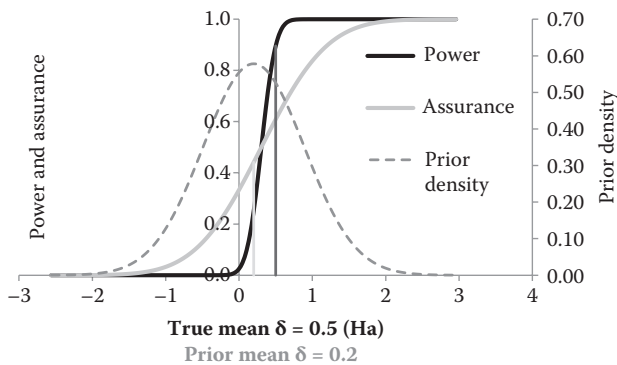


Fig. 2 Case 2: Given $\delta = 0.5$, $\tau = 0.65$, $n_1 = n_2 = 35$, power = 90%; $\delta \sim N(0.2, 0.69^2)$, assurance = 44%

approaches using informative priors. Therefore, it is crucial to obtain a relevant and defensible prior distribution for the parameter(s); otherwise, it can be misleading and hinder informed and appropriate clinical trial decisions.

There are generally two ways of obtaining the most relevant priors, one is by data-driven, for example, as in the study by Novick, Ho, and Best,^[7] and the other is using prior elicitation as in the studies by Garthwaite, Kadane, and O'Hagan^[8] and Oakley and O'Hagan.^[9] Generally speaking, data-driven methods rely on historical data and analytical tools, whereas the prior elicitation would include experts' perspectives on historical data, as well as their subjective personal beliefs. Because different priors produce different assurances, when there is limited and consistent historical data, data-driven methods can be recommended. However, if historical data are not consistent and/or abundant, then prior elicitation or some other forms of reconciliation will be recommended.

One of the major differences between power and assurance is on their asymptotic properties. For power, due to its conditional on a fixed parameter value, when sample size increases to infinity, the power approaches to 100%. This is undesirable because it undermines the clinical meaningfulness in which one can always try to increase sample size in order to achieve statistical significant test results, regardless of whether the effect is practically meaningful or not. On the other hand, the assurance is unconditional and based on a prior distribution of the parameter(s); therefore, it is bounded by the probability of positive outcomes. This can be seen from the assurance formula, Eq. 3. As study sample sizes n_1 and n_2 approach infinity, the standard error τ approaches zero and then the assurance approaches:

$$\gamma = 2\Phi\left(\frac{m}{\sqrt{v}}\right) \quad (4)$$

This is the prior probability of rejecting the null hypothesis, i.e., prior probability of positive outcomes. This nice property tells that in reality, there should always be some probability for non-positive outcomes, and therefore, it is impossible to be 100% assured of positive outcomes, unless the prior probability distribution for the parameter(s) does not cover any non-positive outcomes. In this extreme case, there will be little point to conduct a new clinical trial because the success has already been perceived and reflected in the prior distribution of the parameter(s).

ASSURANCE BY SIMULATION

For priors that are not conjugate or obtained from prior elicitation, or for priors that involve multiple parameters, there are usually no closed forms for their corresponding posterior distributions. In these cases, assurance can be obtained by simulations such as Bayesian Clinical Trial Simulation by O'Hagan, Stevens, and Campbell^[6] or Monte-Carlo Simulation-Based simulations by Chen and Ho.^[10]

APPLICATION OF ASSURANCE

The assurance concept can be applied to various designs such as non-inferiority, equivalence, and cost-effectiveness where trial success is more complicatedly defined. Assurance can also be applied to any meaningful outcome of a trial and need not have any direct connection to power. For these generalizations, the study by O'Hagan, Stevens, and Campbell^[6] is referred.

CONCLUSION

In the real world of clinical trials, the assurance can provide a more realistic and robust measure of probability of success than the traditional power can. Assurance does depend on the prior distribution of treatment effects; therefore, it is crucial to have appropriate priors that closely reflect historical data and/or experts' expertise. Assurance typically is used for internal decisions by the pharmaceutical industry sponsors. Good decision-making is important for clinical development, and therefore, methods such as assurance can be beneficial to patients, sponsors, and regulators as well as society in general.

ACKNOWLEDGMENT

The author would like to thank Dr. Yin Yin for help with the figures.

REFERENCES

1. O'Hagan, A.; Stevens, J.W. Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Med. Decis. Making* **2001**, *21*, 219–230.
2. Spiegelhalter, D.J.; Freedman, L.S. A predictive approach to selecting the size of a clinical trial based on subjective clinical opinion. *Stat. Med.* **1986**, *5* (1), 1–13.
3. Gillett, R. An average power criterion for sample size estimation. *Statistician* **1994**, *43* (3), 389–394.
4. Chuang-Stein, C. Sample size and the probability of a successful trial. *Pharm. Stat.* **2006**, *5* (4), 305–309.
5. Shao, J.; Chow, S. Reproducibility probability in clinical trials. *Stat. Med.* **2002**, *21* (12), 1727–1742.
6. O'Hagan, A.; Stevens, J.W.; Campbell, M. Assurance in clinical trial design. *Pharm. Stat.* **2005**, *4* (3), 187–201.
7. Novick, S.; Ho, S.; Best, N. Data-driven prior distributions for a phase-2 COPD dose-finding clinical trial. submitted for publication 2017.
8. Garthwaite, P.H.; Kadane, J.B.; O'Hagan, A. Statistical methods for eliciting probability distributions. *J. Am. Stat. Assoc.* **2005**, *100* (470), 680–701.
9. Oakley, J.E.; O'Hagan, A. SHELF: The Sheffield Elicitation Framework (version 3.0), School of Mathematics and Statistics University of Sheffield, 2010, Accessed at <http://tonyohagan.co.uk/shelf> (accessed March 4, 2017).
10. Chen, D.; Ho, S. From statistical power to statistical assurance: It's time for the paradigm change in clinical trial design. Accessed at [http://www.tandfonline.com/doi/full/10.1080/03610918.2016.1259476?scroll=top&needAccess=true&"\);](http://www.tandfonline.com/doi/full/10.1080/03610918.2016.1259476?scroll=top&needAccess=true&) to appear in *Communications in Statistics*, 2017.

Clinical Trial Designs with Prospective Patient Population Enrichment by Response

Anastasia Ivanova

*Department of Biostatistics, The University of North Carolina at Chapel Hill,
Chapel Hill, North Carolina, U.S.A.*

Marc DeSomer

PPD, Wilmington, North Carolina, U.S.A.

Jurgen Hummel

PPD, Bellshill, U.K.

Abstract

We discuss four clinical trial designs using prospective patient population enrichment—placebo lead-in, randomized withdrawal, sequential parallel comparison design (SPCD), and two-way enriched design (TED). All designs are two-stage designs. In the placebo run-in and randomized withdrawal, the first stage is usually shorter and data from the first stage is not used in the primary efficacy analysis. In the SPCD and TED, both stages are randomized, double-blind, and placebo-controlled, and data from both stages are used in the primary efficacy analysis.

Keywords: Placebo response; Placebo run-in; Placebo lead-in; Randomized withdrawal; SPCD; TED.

INTRODUCTION

The randomized, double-blind, placebo-controlled clinical trial is the gold standard of evidence-based medicine, informing individual patient and prescriber about therapeutic choices. In a randomized trial, randomization and blinding remove treatment allocation bias, provide the foundation for clinical and statistical inference, and promote the balance of known and unknown covariates between intervention arms. By design, clinical trials target a sample of patients with a specific health condition, who are able and willing to participate in the evaluation of the investigational intervention (drug, device, other medical, or non-medical treatment). The appropriate target patient population for a clinical trial is defined using demographic and diagnostic criteria, severity and time course of disease, response to past treatment attempts, co-morbidity, and co-medication. For evident ethical, medical, scientific, and operational reasons, the clinical trial is not a random sample of a defined general source population, and all patient populations enrolled in clinical trials are implicitly “enriched.” The clinical trial designs described in this entry represent prospective enrichment of the clinical trial patient population based on outcomes observed during the trial.

The U.S. Food and Drug Administration (FDA) defines enrichment in the context of clinical trial design as “the prospective use of any patient characteristic to select a study population in which detection of a drug effect (if one

is in fact present) is more likely than it would be in an unselected population.” Demographical, pathophysiological, historical, genetic or proteomic, clinical, and psychological characteristics have been used for enrichment.^[1]

Enrichment strategies and objectives can include the following:

1. To reduce interpatient variability, by restricting the baseline characteristics of the study population to a narrow range (thus reducing the impact of potential confounding variables) and excluding patients who are at high risk of non-compliance.
2. To reduce inpatient variability, by excluding patients who exhibit marked fluctuations and variability in disease-related symptoms and parameters at baseline.
3. To reduce the potential impact of missing data, by excluding patients who are at high risk of missing study visits or early drop out (whether due to poor tolerability or other factors).
4. To select a subgroup of patients who are more likely to respond to the intervention than the general population as a whole (“predictive enrichment”), based on the following selection criteria:
 - a. presence of a physiologic or disease characteristic that indicates potential sensitivity to the intervention, for example, the presence of a biomarker, proteomic marker, presence of a sensitive microorganism for an anti-infective, etc.

- b. prospective or historic documented response to the intervention under study or to another intervention with a similar mechanism of action.
5. To reduce the potential adverse impact of a high placebo response rate on efficacy signal detection, by excluding patients who exhibit a response to prospective treatment with placebo.
6. To maximize the end point event rate in the study, by enrolling patients with prognostic factors that indicate a high probability of an end point event occurring while on study ("prognostic enrichment").

The potential benefits of enrichment strategies can thus be summarized as follows:

1. Enhancement of internal validity, achieved by good control over confounding variables and the incidence of missing data.
2. Decrease in end point data variability, resulting in an increase in statistical power and a decrease in sample size required.
3. The avoidance of exposing patients to experimental interventions to which they are unlikely to respond and may not tolerate well.

The potential disadvantage of enrichment is the difficulty in generalizing or extrapolating data generated in an enriched target patient population to a broader target patient population. In essence, the gains in internal validity conferred by enrichment are often achieved at the expense of losses in external validity.

Four different enrichment designs are discussed herein (placebo lead-in, sequential parallel comparison design [SPCD], two-way enriched design [TED], and randomized withdrawal). The first three designs involve enrichment for patients who have been shown not to respond to placebo. These are of value in situations where a high placebo response in the general patient population is likely to compromise signal detection. Randomized withdrawal features enrichment with patients who have already exhibited a response to active treatment during an open-label active lead-in. This design is used typically for longer term clinical studies to demonstrate persistence of efficacy and long-term safety.

ENRICHMENT ON THE BASIS OF PLACEBO NON-RESPONSE

Placebo is the optimal clinical trial reference group to evaluate a therapeutic response, where acceptable from an ethical standpoint. A placebo control group enables the evaluation of the absolute effect of drug treatment (from both efficacy and safety perspectives), with intrinsic assurance of assay sensitivity and minimization of expectation bias.^[2]

However, the steady rise of placebo response and patient heterogeneity, since the 1980s, to the current high levels observed in many clinical trials across a wide range of therapeutics and disease indication, has undermined assay sensitivity in many conventional single stage randomized, double-blind, placebo-controlled trials. High failure rates of therapeutic research and development threaten innovation and the economics at the heart of modern pharmaceutical, medical device, surgical, and non-medical health care. The result has also been an inability to replicate prior results in studies of well-established, efficacious therapeutic interventions, approved by regulatory authorities.^[3] The rise of placebo response has coincided with rapidly rising access to health care, health-related information, therapeutic interventions, and co-medication as well as profound changes in behavior, perceptions, and expectations among patients and investigators.

The very decision to enroll and stay in a clinical trial, and the change in exposure to health care and clinical trial procedures, can profoundly impact patient expectations, behavior and perceptions resulting in high placebo response and increased variability. This is true across many therapeutic areas, even in the presence of objective, quantitative diagnostic and therapeutic end points, such as in diabetes and arterial hypertension. Changes in behavior, expectations, and perceptions can influence the observed treatment response, assessed through objective as well as subjective end points, reported by patients, investigators, independent assessors, adjudicators, or dedicated service providers, including central laboratories, imaging, or signal information processing facilities. High placebo response has been observed in psychiatry, neurology, cardiovascular, pulmonary, gastrointestinal, hepatobiliary, musculoskeletal, endocrinological, urogenital, inflammatory, immunological, and other disease indications.

Attempts to address rising placebo response and decreasing separation between response to active treatment and placebo include compensation by sample size increases in fixed and adaptive trial design, enrichment on the basis of placebo non-response, and other measures designed to increase the accuracy and precision of patient diagnosis and end point assessment.

Study designs featuring enrichment on the basis of placebo non-response include the placebo lead-in design, the SPCD, and the TED options, as discussed below. The rationale for placebo non-responder enrichment is that patients who have failed to respond during a prospective period of placebo treatment may be less likely to exhibit a placebo response in a subsequent randomized, double-blind period of active treatment versus placebo. It is anticipated that a lower placebo response in the enriched population should enhance the ability of the study to detect an efficacy signal. However, the ability to demonstrate a separation between active and placebo in this enriched population will depend on the degree to which placebo non-responders can be separated from placebo responders during the lead-in.

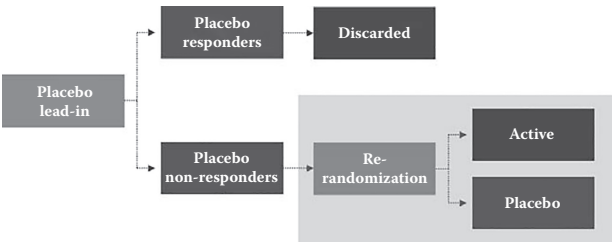


Fig. 1 Placebo lead-in design. Primary efficacy analysis set is in gray

PLACEBO LEAD-IN DESIGN

The primary objective of a placebo lead-in design (Fig. 1) is to improve trial assay sensitivity by reducing the excessive background noise generated by placebo response.

A placebo lead-in trial (also referred to as placebo run-in) consists of an initial stage in which all consented patients are exposed to placebo during a fixed or variable open-label, single- or double-blind period. At the end of the lead-in, placebo non-responders are admitted to the single randomized, double-blind controlled stage. Placebo responders do not enter the randomized stage of the trial. Placebo lead-ins serve the following objectives:^[4]

- weeding out non-compliers before randomization
- ensuring that patients are stable
- washing out previous treatment
- provide a period for baseline measurement, and
- eliminating placebo non-responders

There is uncertainty as to whether or not placebo lead-in design is successful in reducing placebo response and increasing the observed treatment. Chen, Yang, and Hung^[5] reported that trials with placebo lead-in yielded slightly lower placebo response in the randomized phase compared

to trials without a lead-in. Trivedi and Rush^[6] conducted a meta-analysis of a number of single-blind placebo lead-in trials in major depressive disorder and showed that placebo lead-ins did neither decrease placebo response nor increased the separation between active and placebo. Among the possible explanations is the patients’ and investigators’ awareness of the lead-in stage, even single- or double-blind, which may unconsciously or consciously influence their behavior, expectations, adherence, co-intervention/co-medication usage, observation, or reporting.

SPCD

The SPCD was proposed by Fava et al.^[7] It consists of two consecutive randomized, double-blind, placebo-controlled stages, often of equal duration, with continuity of exposure, observation, and blinding for all patients and investigators. Stage 1 is conducted in a non-enriched patient population, whereas stage 2 provides data from an enriched population of placebo non-responders identified in stage 1. At the end of stage 1, placebo non-responders are re-randomized in equal proportions to receive active treatment or placebo in stage 2 (Fig. 2). Placebo responders are often re-randomized to receive active treatment or placebo in stage 2. This can provide additional insights into the nature and persistence of the placebo response and its impact upon signal detection. Patients randomized to active treatment in stage 1 may continue on active treatment in stage 2. This provides the opportunity to evaluate the efficacy and safety of the active treatment over a longer exposure period. Consider a SPCD trial where two-third of patients receive placebo in stage 1, all patients receiving placebo are re-randomized between active and placebo in stage 2, and patients who received active treatment in stage 1 continue on active in stage 2. In this trial, one-third of all patients will receive placebo during both stages of SPCD

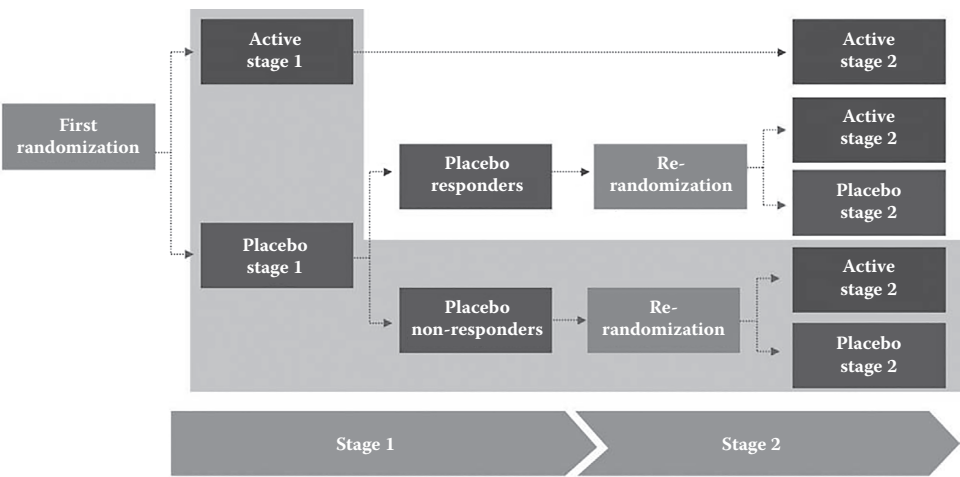


Fig. 2 Sequential parallel comparison design. Primary efficacy analysis set is in gray

and one-third of the patients will receive active treatment during both stages, providing data for active placebo comparison over a period of time that is twice as long as the length of a single stage.

As an alternative to re-randomization, treatment allocations for both stages can be determined in the initial randomization at the start of the study. This option is simpler from a trial, randomization and clinical supply management perspective but, unlike re-randomization at the end of stage 1, it does not provide assurance of balanced groups in stage 2.

The continuity of investigators' and patients' blinding is maintained throughout both SPCD stages through the use of an appropriately masked protocol, and patient informed consent disclosure mentioning the fact of randomization, but concealing the specific details of stages, their respective admission criteria, as well as times of randomization and end point assessment. Boessen, Knol, and Groenwold^[8] proposed to use SPCD with variable stage duration to further facilitate masking of patients and investigators.

The primary analysis set includes all-comer patient data from stage 1 as well as stage 2 data from stage 1 placebo non-responders (Fig. 2). The null hypothesis is that there is no treatment effect in the stage 1 population of all-comers and there is no treatment effect in the population of placebo non-responders. The combined treatment effect across both stages is defined as the weighted sum of effect estimates from stages 1 and 2, using a prespecified weight. Dividing this weighted sum of estimates by its standard error leads to an asymptotically normally distributed test statistic, which is used to test the null hypothesis and calculate the associated p value. Methods of analyzing the SPCD data with continuous end point have been discussed by Tamura and Huang;^[9] Chen, Yang, and Hung;^[5] Doros, Pencina, and Rybin;^[10] and Chi et al.^[11] Methods to combine data from the two stages of SPCD have been proposed for binary outcomes.^[7,12,13] Wang and Ivanova^[14] investigated the possibility of changing the allocation proportion or dropping arms in a SPCD trial with multiple arms and Silverman and Ivanova^[15] investigated sample size re-estimation strategies and other adaptations in a SPCD trial.

SPCD is suited for clinical trials in which it is ethically and medically acceptable and feasible to expose and observe patients during two consecutive stages. SPCD avoids the issue of carryover effect which biases the evaluation of many crossover trials, as no patient crosses over from receiving active treatment in this design. Baer and Ivanova^[16] reviewed results of several completed SPCD studies. Results from multiple SPCD trials have been reported in peer-reviewed journals, including five studies in major depressive disorder^[17–20] and one study of agitation and aggression in Alzheimer's dementia patients.^[21] In most studies, the placebo response was reduced in stage 2 compared to stage 1, thus demonstrating the effectiveness of SPCD stage 1 in identifying placebo non-responders

as well as the continued lower placebo response among stage 1 placebo non-responders in stage 2.

THE RANDOMIZED WITHDRAWAL TRIAL DESIGN

Responders to be enrolled in a randomized withdrawal study are identified in an open-label single arm treatment period or the active arm of a controlled clinical study. Patients are randomized to continue receiving active treatment or to switch to placebo.^[22] The primary efficacy end point is typically time to relapse/recurrence (return of symptoms). Patients experiencing relapse/recurrence are removed from the study immediately and receive appropriate standard of care.^[2]

The randomized withdrawal design is useful in providing information on the persistence and durability of efficacy effects, optimal duration of treatment, and safety during longer term treatment. In addition, randomized withdrawal studies with multiple dose arms can be used to provide information on the optimal maintenance dose. From an ethical perspective, the randomized withdrawal design may offer some benefits, compared to conventional long-term placebo-controlled studies, because relapse/recurrence leads automatically to immediate withdrawal and transfer to appropriate standard of care (which entails only a short period of poor response associated with placebo treatment).

When designing a randomized withdrawal study, the potential impact of selection and attrition bias, the development of tolerance during the open-label active lead-in stage, the potential for carryover effects from the lead-in to the randomized stage, as well as withdrawal and rebound in the second, randomized stage have to be considered carefully. Patients who develop tolerance may derive little or no ongoing benefit from continuing active treatment during the open-label lead-in or the randomized stage. Patients may also experience withdrawal symptoms and/or rebound disease exacerbation when active treatment is discontinued in stage 2, upon switching to placebo. These effects can lead to differential adherence and retention as well as safety, tolerability, or efficacy signs and symptoms potentially unblinding patients or investigators and biasing observation and reporting. These sources of bias may result in an erroneous evaluation of the therapeutic benefit–risk. If an intervention is known to exhibit tolerance, withdrawal, and/or rebound phenomena, the potential impact upon the study results may be mitigated by slowly tapering off of the dose in those patients randomized to placebo.

Randomized withdrawal studies are inherently enriched with active responders and exclude patients who are unable to tolerate the intervention. Hence, the efficacy and safety profiles obtained in such studies are expected to be more favorable than would be the case in a general, unselected population.

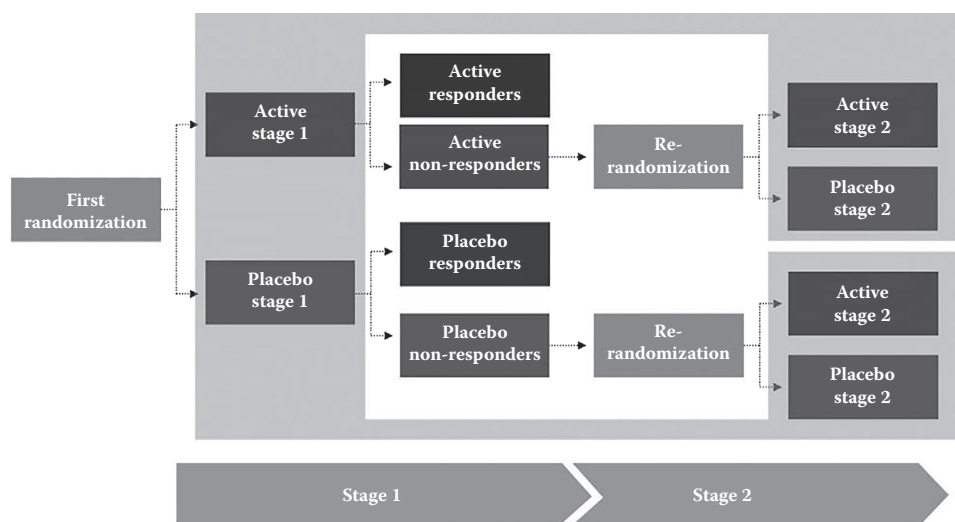


Fig. 3 TED. Primary efficacy analysis set is in gray

TED

The TED^[23] can be viewed as a combination of the parallel design, placebo run-in, and randomized withdrawal (Fig. 3). It is an extension of the SPCD that, in addition, re-randomizes stage 1 active arm responders in the second stage to active and placebo exposure and utilizes information from this patient cohort versus the placebo non-responders in the overall active placebo comparison. Patients who respond to active treatment are examined to see whether the maintenance of response is different between those patients switched to placebo versus those who remain on active.

Ivanova and Tamura^[23] describe the analysis methodology for TED trials with binary outcomes. An overall test of the treatment effect in all-comers, placebo non-responders, and active treatment responders is presented.

CONCLUSION

Enrichment trial designs are increasingly important to detect an accurate and reliable efficacy signal in the appropriate target patient population, especially, in the presence of placebo response and high heterogeneity among patients. All trials are enriched in the process of selecting and admitting patients, according to diagnostic, prognostic, predictive, and inclusion/exclusion criteria. We define enrichment trials as those in which an explicit and prospective selection occurs during the trial, based on the patient's response to the investigational intervention. All designs described are two-stage designs. In the placebo lead-in and randomized withdrawal designs, the first stage is usually shorter, non-randomized, and uncontrolled, and only the data from the second stage contribute to the primary efficacy analysis. In SPCD and TED, both stages are

randomized, double-blind, and placebo-controlled. The two stages are often, but not necessarily, of the same duration, and data from both stages contribute to the primary efficacy analysis. The relative power of these designs for a given sample size depends on the treatment responses and their variability in all-comers, placebo non-responders, and, in TED, active responders, as well as on the proportion of placebo non-responders and active responders. Generally, SPCD and TED have higher power, given equal assumptions, as some of the patients contribute two data points to the primary efficacy analysis, and a larger treatment effect is expected in the second stage.

REFERENCES

1. U.S. Department of Health and Human Services, Food and Drug Administration. Guidance for Industry. In *Enrichment Strategies for Clinical Studies to Support Approval of Human Drugs and Biological Products. Draft Guidance*; 2012.
2. ICH. *Harmonised Tripartite Guideline: Choice of Control Group and Related Issues In Clinical Trials (Topic E10)*. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals For Human Use (ICH): Geneva, 2000.
3. Laughren, T.P. The scientific and ethical basis for placebo-controlled studies in depression and schizophrenia: An FDA perspective. *Eur. Psychiat.* **2001**, *16* (7), 418–423.
4. Senn, S. Are placebo run ins justified? *Brit. Med. J.* **1997**, *314*, 1191.
5. Chen, Y.F.; Yang, Y.; Hung, H.M.J. Evaluation of performance of some enrichment designs dealing with high placebo response in psychiatric clinical trials. *Contemp. Clin. Trials.* **2011**, *32* (4), 592–604.
6. Trivedi, M.H.; Rush, A.J. Does a placebo run-in or placebo treatment cell affect the efficacy of antidepressant medications? *Neuropsychopharmacology.* **1994**, *11*, 33–43.

7. Fava, M.; Evins, A.E.; Dorer, D.J.; Schoenfeld, D.A. The problem of the placebo response in clinical trials for psychiatric disorders: Culprits, possible remedies, and a novel study design approach. *Psychother. Psychosom.* **2003**, *72*, 115–127.
8. Boessen, R.; Knol, M.J.; Groenwold, R.H.H. Increasing study efficiency by early reallocation of placebo non-responders in sequential parallel comparison designs: Application to antidepressant trials. *Clin. Trials.* **2012**, *9* (5), 578–587.
9. Tamura, R.N.; Huang, X. An examination of the efficiency of the sequential parallel design in psychiatric clinical trials. *Clin. Trials.* **2007**, *4* (4), 309–317.
10. Doros, G.; Pencina, M.; Rybin, D. A repeated measures model for analysis of continuous outcomes in sequential parallel comparison design studies. *Stat. Med.* **2013**, *32*, 2767–2789.
11. Chi, G.Y.H.; Li, Y.; Liu, Y.; Lewin, D.; Lim, P. On clinical trials with high placebo response rate. *Contemp. Clin. Trials Commun.* **2016**, *2*, 34–53.
12. Huang, X.; Tamura, R. Comparison of test statistics for the sequential parallel design. *Stat. Biopharm. Res.* **2010**, *2* (1), 42–50.
13. Ivanova, A.; Qaqish, B.; Schoenfeld, D.A. Optimality, sample size, and power calculations for the sequential parallel comparison design. *Stat. Med.* **2011**, *30*, 2793–2803.
14. Wang, J.J.; Ivanova, A. Dose finding with the sequential parallel comparison design. *J. Biopharm. Stat.* **2014**, *24* (5), 1091–1101.
15. Silverman, R.K.; Ivanova, A. Sample size re-estimation and other midcourse adjustments with sequential parallel comparison design. *J. Biopharm. Stat.* **2017**, *27* (3).
16. Baer, L.; Ivanova, A. When should the sequential parallel comparison design be used in clinical trials? *Clin. Invest.* **2013**, *3* (9), 823–833.
17. Fava, M.; Mischoulon, D.; Iosifescu, D.; Witte, J.; Pencina, M.; Flynn, M.; Harper, L.; Levy, M.; Rickels, K.; Pollack, M. A double-blind, placebo-controlled study of aripiprazole adjunctive to antidepressant therapy among depressed outpatients with inadequate response to prior antidepressant therapy (ADAPT-A Study). *Psychother. Psychosom.* **2012**, *81* (2), 87–97.
18. Papakostas, G.I.; Shelton, R.C.; Zajecka, J.M.; Etamad, B.; Rickels, K.; Clain, A.; Baer, L.; Dalton, E.D.; Sacco, G.R.; Schoenfeld, D.; Pencina, M.; Meisner, A.; Bottiglieri, T.; Nelson, E.; Mischoulon, D.; Alpert, J.E.; Barbee, J.G.; Zisook, S.; Fava, M. L-methylfolate as adjunctive therapy for SSRI-resistant major depression: Results of two randomized, double-blind, parallel-sequential trials. *Am. J. Psychiat.* **2012**, *169* (12), 1267–1274.
19. Papakostas, G.I.; Vitolo, O.V.; Ishak, W.W.; Rapaport, M.H.; Zajecka, J.M.; Kinrys, G.; Mischoulon, D.; Lipkin, S.H.; Hails, K.A.; Abrams, J.; Ward, S.G.; Meisner, A.; Schoenfeld, D.A.; Shelton, R.C.; Winokur, A.; Okasha, M.S.; Bari, M.A.; Fava, M. A 12-week, randomized, double-blind, placebo-controlled, sequential parallel comparison trial of ziprasidone as monotherapy for major depressive disorder. *J. Clin. Psychiat.* **2012**, *73* (12), 1541–1547.
20. Fava, M.; Memisoglu, A.; Thase, M.E.; Bodkin, J.A.; Trivedi, M.H.; de Somer, M.; Du, Y.; Leigh-Pemberton, R.; DiPetrillo, L.; Silverman, B.; Ehrlich, E. Opioid modulation with buprenorphine/samidorphan as adjunctive treatment for inadequate response to antidepressants: A randomized double-blind placebo-controlled trial. *Am. J. Psychiat.* **2016**, *173* (5), 499–508.
21. Cummings, J.L.; Lyketsos, C.G.; Peskind, E.R.; Porsteinsson, A.P.; Mintzer, J.E.; Scharre, D.W.; De La Gandara, J.E.; Agronin, M.; Davis, C.S.; Nguyen, U.; Shin, P.; Tariot, P.N.; Siffert, J. Effect of dextromethorphan-quinidine on agitation in patients with Alzheimer disease dementia: A randomized clinical trial. *JAMA.* **2015**, *314* (12), 1242–1254.
22. Amery, W.; Dony, J. A clinical trial design avoiding undue placebo treatment. *J. Clin. Pharmacol.* **1975**, *15*, 674–679.
23. Ivanova, A.; Tamura, R.N. A two-way enriched clinical trial design: Combining advantages of placebo lead-in and randomized withdrawal. *Stat. Methods Med. Res.* **2015**, *24* (6), 871–890.

Clinical Trial Process

Paul F. Kramer

Bristol-Myers Squibb Company, Plainsboro, New Jersey, U.S.A.

Abstract

This entry outlines the importance of monitoring the effects of implementing changes in work processes based on the conclusions from the information gathered and explores the importance of continually looking for answers, acting on what was learned, and monitoring the effects of process changes.

INTRODUCTION

Regardless of the setting—whether a research group in an academic or healthcare institution, or a research and development organization in a small venture capital-based or large multinational pharmaceutical company—there are forces that demand these organizations, and their members, to respond successfully to the following question: How can more be done with the same resources in shorter time periods with no significant reduction in the quality of the output of the activity? The challenge this question poses to clinical research is now a routine part of the activities in most pharmaceutical companies. With aggressive commitments from the corporate offices to increase dramatically the number of new compounds entering and completing the clinical research and development process, all members of the research team are being affected.

This entry will (1) outline the types of information necessary for finding answers to this vexing question; (2) outline assessments of what was learned during information gathering; (3) outline the importance of monitoring the effects of implementing changes in work processes based on the conclusions from the information gathered; and (4) explore the importance of continually looking for answers, acting on what was learned, and monitoring the effects of process changes.

OVERVIEW

At a basic level, finding out how more can be done with the same resources involves using the steps in the scientific method that are familiar to many readers. That is, gather facts, data, observations, and related information concerning the topic under investigation; articulate a problem based on what is known; develop a hypothesis surrounding the problem; design an appropriate methodology that will test the hypothesis; execute the methodology to generate data; analyze and evaluate the data to test the hypothesis;

and draw appropriate conclusions. As will be discussed later, another parallel to the scientific method is that the critical endpoint is not only the conclusion, but also the successful application of the conclusion. The application of conclusions may be to provide a practical solution to the problem that was articulated or to use the new data and insights to define the problem better, to develop and test a more refined hypothesis, and to derive added conclusions to follow along a path of continuous improvement.

UNDERSTANDING WORK PROCESSES

In order to answer the question introduced in the preceding section, it is critical to develop a context by defining what tasks need to be accomplished and what the output is from those tasks. The thoroughness of describing the process, the output(s), and the tasks needed to accomplish them will determine the success of the information-gathering step.

Gathering Information

Understanding the work processes is not the endpoint but just the beginning. Some of the information needed to understand a work process include the following.

Define the Starting Point for the Process

What is the starting point? What are the decisions or events that trigger the initiation of the process? Who makes the decision to start, and how is it communicated? Who is responsible for the process or who owns it? What material or information is required to begin the process?

Define the End of the Process

What is the ultimate, final output of the completed process? Who is the customer for the final output? How is

the customer organizationally related to the person(s) or group(s) responsible for the process? How is the final output delivered? How do you know that the output is ready for delivery? How do you measure and report on the success, or lack of success, of the final output? How is the feedback concerning the output received from the customer?

Define the Tasks in a Process

Identify and be able to describe each task required to complete a process. Identify who is responsible for a task or who owns it. Is the owner inside or outside of the group responsible for completing the task? What event(s), task(s), information, material(s), or decision(s) are required to begin each task? How are the necessary information, material(s), or decision(s) received? What resources—e.g., persons, equipment, and systems—are required to complete each task? For intermediate outputs, to whom is the output delivered? For intermediate outputs, how is the product of the task delivered to the person or group responsible for the next task? Which tasks need to be completed before the next task can begin? Which tasks can be completed in parallel with or independent of other tasks in the process?

Define Intermediate Outputs or Milestones

Are there any intermediate outputs that must be achieved prior to the final output? What are the intermediate outputs? Who is the customer for (who receives) each intermediate output? How is the customer organizationally related to the persons or groups responsible for the tasks to achieve each output? How is the feedback concerning the adequacy of the output received? How do you measure and report on the success, or lack thereof, of each intermediate output? What milestones are meaningful to the management?

Assessing the Information

In order to understand the work processes, the information and data gathered must be assessed to define or refine the statement of the problem and to develop a statement of the hypothesis. The following topics should be considered during work process assessment.

Assess the Starting Point for the Process

Evaluate the decisions or events that trigger the initiation of the process to determine if there is a predictable pattern. Describe the adequacy, or lack thereof, of the systems for transmitting the decision to start and providing the material or information that is required to begin. Define what information the person(s) or group(s) responsible for the process actually need to start the process. Describe what happens if the decision to start or the material or

information that is required to begin is not adequately delivered or communicated. What control or influence, if any, can be exercised by the process owner at the starting point? How can those items that are within the area of authority of the process owner be controlled, influenced, or managed by that person? How can the resources within the area of authority of the process owner be affected by decisions or activities outside the owner's area of authority? For those areas of authority that cannot be controlled or influenced by the process owner, what steps can be taken to manage the impact when those areas do not support, or prove inadequate for, the process?

Assess the End of the Process

Describe the characteristics of the output that will satisfy the customers' requirements. Describe the adequacy, or lack thereof, of the systems for delivering the output to the customer. Describe the monitoring systems, if any, that can predict if the output will meet the required quality and the delivery schedule. Describe what happens if the output is delivered sooner than scheduled, if the delivery is delayed, and if the quality is unsatisfactory to the customer or it does not meet the needs of the customer. What control or influence, if any, can be exercised by the process owner at the end of the process? Does the current process work to achieve the desired output, within the desired time schedule, with the systems and resources currently available?

Assess the Tasks Involved in the Process

Categorize the importance of each task to the completion of the process. Is each moderately or significantly important task well understood? What is needed to start the task? Who provides what is needed, and how? Who executes the task? What resources are needed? How and to whom is the product of the task delivered? Define the tasks on the critical path in the process. What are the dependencies between tasks, e.g., which task(s) have to be completed before another specific task can begin or can be completed in parallel with or independent of the other tasks in the process? Describe the adequacy, or lack thereof, of the systems for transmitting the material or information that is required to begin the task. Define what information the person(s) responsible for the task actually need to start the task. Describe what happens if the material or information that is required to begin is not adequately delivered or communicated. What control or influence, if any, can be exercised on the task by the process or task owner? How can those items that are within the area of authority of the task owner be controlled, influenced, or managed by that person to facilitate task completion? How can the resources within the area of authority of the task owner be affected by the decisions or activities outside the owner's area of authority? For those

areas of authority that cannot be controlled or influenced by the task owner, what steps can be taken to manage the impact when those areas do not support, or prove inadequate for, the process?

Assess the Intermediate Outputs or Milestones

Describe the requirements the output must achieve. Describe the systems, if any, that can monitor whether the output will meet the required quality and delivery schedule. Describe what happens if the output is delivered sooner than scheduled, if the delivery is delayed, or if the quality is unsatisfactory. Do the current tasks work to achieve each of the desired intermediate outputs, within the desired time schedule, with the systems and resources currently available?

Prepare a Statement of Problems

Based on a clear understanding of the work process, output(s), and tasks, the next step would be to identify specific problems that could result from work inefficiencies, poor resource utilization, unnecessary or redundant tasks, or low work capacity. Because the objective of these steps is to determine how to do more with the same, avoid focusing on inadequate resources, especially human resources, as a primary problem. Attempt to look beyond this often simplistic statement of the problem. Inadequate resources are a valid problem if you have metrics and benchmarking information to substantiate your conclusions or if the focus of the problem involves equipment or technology that enables increased efficiency or increased work capacity for the available human resources.

The problem list must be prioritized to have the basis for decision making that will manage the limited resources available to explore and/or implement change in the process, output(s), or tasks. Prioritization can be based on one or more of the following parameters: whether the process or task(s) are within the area of authority of the owner (focus on change that is within the area of authority of the task or process owner); the magnitude of the resources and the period of time that will be required to implement the change successfully (look for problems with short-term, fairly easy solutions to build the confidence of the staff, management, and organization while working on other longer term solutions); the size of the benefit derived from the change to the strategic and/or operational goals of the organization (do not overlook the benefit derived by another part of the organization in addition to the portion implementing the change); objectives and goals established by the management (avoid undertaking work process change without having a well-defined, specific set of goals and objectives that have been agreed upon by the necessary levels of management).

From the list of problems, develop a potential solution for the priority problems chosen. The solution should

portray the proposed change and the resulting measurable benefit. This could resemble a statement of hypothesis; for example, by developing and implementing a common computer platform, duplicate entry of the same data into two different systems can be eliminated, thus reducing the processing time by *X* hr or allowing a *Y*% increase in the number of forms processed.

GATHERING DATA—MONITORING WORK PROCESSES

As with the scientific method, a crucial activity is to test the potential solution, or hypothesis, with a well-designed methodology and to capture data from the endpoints needed to test the solution. As already noted, when stating the solution, it is necessary to identify the measurement(s) that can be used to determine if the solution was successful (having preestablished goals for evaluating the solution is critical).

The term *metrics* is often used to refer to the data gathered while monitoring work processes or tasks. Metrics can include such items as the time elapsed for completing a task or process; the number of times a task must be repeated to achieve the standards required for the output; the rate at which the errors occur; and the objective feedback from the customer of the output.

As is true of clinical research, gathering data solely for the sake of having the data is a poor use of resources. There must be value that can be derived from the data. In the case of evaluating the work process, the value could include the following: providing the management with a real-time indication of critical conditions in a process or task; predicting whether the output will meet the standards or the delivery schedule; and identifying potential problems in a task or process to allow implementation of the required solution before there could be a significant delay or loss in quality. An issue that should be addressed when testing a potential solution is to assess the value of the metrics used to evaluate the process, output, or task. Some metrics may not be predictive of, or related to, the quality or efficiency of the output. Some metrics may be too complex to collect for use in routine assessment of a process.

CONTINUOUS WORK PROCESS IMPROVEMENT

The question introduced at the start of this entry of how more can be done with the same resources in shorter time periods with no significant reduction in the quality of the output of the activity includes an element of continuity. As potential solutions to work process problems are tested, a recurring set of events should be set in place. These events include the ongoing collection of metrics to describe and assess the work process, to identify previously undetected

or new problems, to refine insight into previously identified problems, to develop new or alternate solutions and the associated benefits, to implement and test the solutions, and to continue monitoring the process through meaningful metrics. Thus what may have started as a one-time exercise in doing more with the same, or even doing more with less, needs to become a routine part of the culture in which clinical research takes place.

CONCLUSION

Clinical trial process plays an important role in clinical research and development. It ensures not only the quality

of data collected from the trial, but also maintains the integrity of the trial within available resources and timeline. In addition, it enables upper management to make a critical go/no go decision during, or at the end of the process. In practice, it is suggested that tests or procedures employed at various critical stages of the process should be in compliance with good clinical practice (GCP) as specified in the International Conference on Harmonization (ICH) guideline.^[1]

REFERENCE

1. ICH Secretariat. ICH Harmonized Tripartite Guideline: Guideline for Good Clinical Practice; 1996; Geneva.

Clinical Trial Simulation

Hung-Ir Li and Pan-Yu Lai

Allergan, Inc., Irvine, California, U.S.A.

Abstract

Clinical trial simulation utilizes the computer to simulate virtual patients in the clinical research context. This entry discusses how clinical trial simulation has been applied to better understand the physiology, pharmacology, properties, and performance of statistical and data analysis methods, and the possible trial outcomes to optimize clinical study design and clinical planning.

Keywords: Computer assisted study design; Model-based drug development; Modeling and simulation.

INTRODUCTION

In the new era of high technology, the traditional process for medical research and drug development has also evolved to become more time- and cost-effective. Clinical trial simulation, which utilizes the computer to simulate virtual patients in the clinical research context, has been applied to better understand the physiology, pharmacology, properties, and performance of statistical and data analysis methods, and the possible trial outcomes to optimize clinical study design and clinical planning. The concept of simulating clinical trials started in the late 1970s following the successful applications of simulation in the auto and aerospace industry. However, the application in the pharmaceutical industry to help the clinical development did not occur until the mid-1990s when computing and user-friendly trial simulation tools were developed. As of November 2000, more than 100 clinical trial simulations in more than seven therapeutic areas have been performed by more than 17 pharmaceutical companies.^[1] Applications include selecting the optimal dose, evaluating the effect of changes in formulation, dose regimen, or dosing route on changes in pharmacodynamic response, and so forth. Vigorous collaboration among regulatory agencies, academia, and industries continues through the twenty-first century, particularly in the United States. FDA Critical Path Initiative was launched in 2003 which promotes the innovation of the drug development and particularly encourages the best use of clinical trial simulation via the End-of-Phase-2A (EOP2A) pilot program beginning in 2004. The related Guidance for Industry: End-of-Phase-2A Meetings released in 2008 details the objectives and recommended utilities of clinical trial simulation in the exploration of trial design alternatives to increase the likelihood of future trial success.^[2–19]

WHAT IS A CLINICAL TRIAL SIMULATION?

Clinical trial simulation is a process that uses computers to mimic the conduct of a clinical trial by creating virtual patients and generating the outcomes for each virtual patient based on the prespecified models, preferably probabilistic but quite often deterministic models. These models include mathematical structures and estimates of the parameters or the corresponding variation or both. The simulated data are then analyzed to evaluate the properties of the clinical trial setting. The primary objective of a clinical trial simulation is to extrapolate or predict the clinical outcomes beyond the scope of previous studies from which the existing models were derived using the modeling techniques. It can also be used to verify or confirm the models depicting the relationships between the inputs such as dose, dosing time, patient characteristics, disease stage, and so forth, and the outcomes such as changes in the signs and symptoms, or time to events within the study domain. By evaluating the summarized virtual trial outcomes under various design scenarios, the impact of changes in formulation, dose regimen, or dosing route can be assessed. Initially, simulations use the distribution of patient characteristics to create virtual patients and the quantitative relationship between the patient characteristics, dose, and response to generate the outcome. Over the past decade, simulations have evolved to the current practice, which integrates the whole understanding of the disease physiology and pharmacological mechanism into a complex mathematical model to mimic human body responses to the treatment intervention. The simulation based on this integrated system of models will predict the potential clinical outcomes under different assumptions and various design scenarios, enabling a better selection or planning of the actual clinical trials.^[2–11] The resultant study design is often called the computer-assisted trial design (CATD). It improves

the cost-effectiveness and the timeliness of the clinical development process and expedites the availability of new therapies to the waiting patients. For example, to study an anti-obesity agent, one can build a model that includes the dosing regimen, the patient's daily activity, the caloric intake (food consumption including the percentage of fat), the insulin effect on the obese patients, the duration of the study, and the pharmacological mechanism of the agent, lipolysis, and energy expenditure. A clinical trial simulation can then be conducted to address the following questions: Does increased lipolysis lead to weight loss? Does exercise enhance drug efficacy? Which biomarker (such as fasting plasma free fatty acids levels or the respiratory quotient) is a more reliable measure of lipolysis? How does the plasma concentration of the investigational drug relate to tissue concentration and to the weight reduction? Does the type of food consumption influence the efficacy? How does the dosing regimen such as frequency, timing, and dose level affect the efficacy? What is the efficacy and safety profile compared to other marketed compounds?

APPLICATIONS AND SOME TOOLS

According to presentations at the U.S. Food and Drug Administration (FDA) Pharmaceutical Science Advisory Committee meeting in November 2000, more than 100 clinical trial simulations in more than seven therapeutic areas have been performed by more than 17 of the pharmaceutical companies.^[1] These applications include selecting the optimal dose; evaluating the effect of changes in formulation, dose regimen, or dosing route on changes in pharmacodynamic response; and integrating preclinical/clinical pharmacology and biopharmaceutics study results into late-phase clinical trials to ensure safe and effective study design, or unbiased, powered, and robust studies to maximize the treatment benefit/risk ratio. Further, the clinical trial simulation has become major part of the adaptive design, particularly the seamless phase II–phase III adaptive design, due to the complex nature of the adaptive design and the necessity of controlling the type I error rate.^[12] Therefore depending on the objectives, the required components for a clinical trial simulation can be as simple as a binomial variable, Dose/PK/PD modeling, or as complicated as an integration of all appropriate body compartments with the mechanism of actions of the compound. Sources for building these models can include (but are not limited to) all available data from the compound and literature review and the SBA (summary basis of approval) from the FDA of compounds in the same class (including animal data, pharmacokinetic/pharmacodynamic data, and biomarker/surrogate endpoints). When data are not available, models based on established theories and assumptions may be used and then updated later when data become available. Simulation is useful and helpful only when it is coupled with appropriate models. A negative

conclusion of a trial simulation can be the result of the lack of action of the compound, the improper specification of the model, or the improper study design strategy. The models used in the simulation need to be validated before and after the simulation is performed. There are more than 15 software packages for model-building and clinical trial simulations. These include NONMEM® by GloboMax LCC / NONMEM7™ by ICON Development Solution; WinNonlin®, WinNonMix®, and Pharsight Trial Simulator™ by Pharsight Corporation; S-PLUS® by Insightful; SAS® by SAS Institute; PhysioLab® by Entelos®; CellML™ and PathwayPrism™ by Physiome Sciences; and East® 5/SiZ® by Cytel Inc.

CLINICAL TRIAL SIMULATION IN REGULATORY DECISIONS

The regulatory agencies worldwide have recognized the promises of the clinical trial simulation in the drug development and been supportive in various venues since the twentieth century. The FDA Modernization Act of 1997 set the stage for the regulatory agency's proactive interaction with the industry toward common goals. The Critical Path Initiative was launched in 2003 which seeks not only changes in the science and technology of drug development but also the way FDA and Industry interact and share information. The FDA Guidance for Industry: Exposure-Response Relationships—Study Design, Data Analysis, and Regulatory Application and subsequently the FDA Guidance for Industry: End-of-Phase 2A Meeting (EOP2A) provide the scientific base and recommendations for clinical trial simulation and procedural details of FDA initiated forum in helping industry to streamline the drug development using clinical trial simulation.^[13,14] A pilot EOP2A program was initiated in 2004, which included various applications of clinical trial simulation in regulatory decisions such as basis for approval, labeling claims, designing trials, and setting policy. Examples include pediatric approval of trileptal without additional clinical trials, subgroup approval of zoledronic acid, and risk management of busulfan in pediatric patients.^[15–19] Similar regulatory interests in the application of clinical trial simulation have been observed in European Union and Japan.^[20,21]

CONCLUSION

Clinical trial simulation is a useful process to mimic human body responses to the treatment intervention by simulating clinical trials based on some pre-specified models before the clinical trials are launched. It is particularly useful in selecting the optimal dose, evaluating the effects of changes in formulation, dose regimen, dosing route on changes in pharmacodynamic response, and optimizing trial designs in clinical research and development.

It not only bridges the previous data to new trials using models but also checks the feasibility of the design of new trials to ensure an efficient and successful drug development. With the current high level of collaborations among academia, regulatory agencies, and industries, the clinical trial simulation has become the core of the model-based drug development in meeting the patient needs.

REFERENCES

1. Lee, P.; Machado, S.; Lesko, L. A Framework of Modeling and Simulation in Regulatory Decisions. FDA Advisory Committee Meeting of Pharmaceutical Science, 2003, <http://www.fda.gov/ohrms/dockets/ac/cder03.html#PharmaceuticalScience>.
2. Blatant, L.P.; Gex-Fabry, M. Modeling during drug development. *Eur. J. Pharm. Biopharm.* **2000**, *50*, 13–26.
3. Bland, J.M.; Altman, D.G. Measuring agreement in methods comparison studies. *Stat. Methods Med. Res.* **1999**, *8*, 135–160.
4. FDA. Antiviral Advisory Committee Meeting on PK/PD, 2000, <http://www.fda.gov/ohrms/dockets/ac/cder00.htm#Anti-Viral> (accessed August 2009).
5. Gieschke, R.; Reigner, B.G.; Steimer, J.-L. Exploring clinical study design by computer simulation based on pharmacokinetic/pharmacodynamic modeling. *Int. J. Clin. Pharmacol. Ther.* **1997**, *35*, 469–474.
6. Gustafson, P. On measuring sensitivity to parametric model misspecification. *J. R. Stat. Soc. B Part 1* **2001**, *63*, 81–94.
7. Holford, N.H.G.; Hale, M.; Ko, H.C.; Steimer, J.-L.; Sheiner, L.B. Simulation in Drug Development: Good Practices, 1999, <http://bts.ucsf.edu/cdds/research/sddgpreport.php> (accessed August 25 2009).
8. Hale, M.D. *Modeling and Simulation in Drug Development*; FDA Advisory Committee Meeting of Pharmaceutical Science, 2000; http://www.fda.gov/ohrms/dockets/ac/00/slides/3657s2_05.ppt (accessed August 25 2009).
9. Holford, N.H.G.; Kimko, H.C.; Monteleone, J.P.R.; Peck, C.C. Simulation of clinical trials. *Annu. Rev. Pharmacol. Toxicol.* **2000**, *40*, 209–234.
10. Sheiner, L.B.; Steimer, J.L. PK/PD modeling in drug development. *Annu. Rev. Pharmacol. Toxicol.* **2000**, *40*, 67–95.
11. Kimko, H. Duffull, S.B. *Simulation for Designing Clinical Trials: A Pharmacokinetic-Pharmacodynamic Modeling Perspective*; Informa Healthcare: New York, 2003.
12. Chow, S.C.; Chang, M. *Adaptive Design Methods in Clinical Trials*; Chapman & Hall/CRC: Boca Raton, FL, 2007.
13. FDA Guidance for Industry: Exposure-Response Relationships – Study design, Data Analysis, and Regulatory Applications, March 2003.
14. FDA Guidance for Industry: End-of-Phase 2A Meetings, September 2008.
15. Yim, D.S.; Zhou, H.; Buckwater, M.; Nestoro, I.; Peck, C.C.; Lee, H. Population pharmacokinetic analysis and simulation of the time-concentration profile of etanercept in pediatric patients with juvenile rheumatoid arthritis. *J. Clin. Pharmacol.* **2005**, *45*, 246–256.
16. Lesko, L.J. Paving the critical path: how can clinical pharmacology help achieve the vision? *Clin. Pharmacol. Ther.* **2007**, *81*, 170–177.
17. Bhattaram, V.A., et al. Impact of pharmacometrics on drug approval and labeling decisions: a survey of 42 new drug applications. *AAPS J.* **2005**, *7* (3), Article 51 (<http://www.aapsj.org>).
18. Lesko, L.J. *Clinical Trial Simulation in The Critical Path Initiative and regulatory Decision-Making*; Annual Meeting of American Association of Pharmaceutical Science: San Diego, CA, 2007.
19. Bhattaram, V.A.; et al. Impact of pharmacometric reviews on new drug approval and labeling decisions—a survey of 31 new drug applications submitted between 2005 and 2006. *Clin. Pharmacol. Ther.* **2007**, *81*, 213–221.
20. Aarons, L.; Karlsson, M.O.; Mentre, F.; Rombout, F.; Steimer, J.L.; van Peer, A. and invited COST B15 Experts. Role of modeling and simulation in phase I drug development. *Eur. J. Pharm. Sci.* **2001**, *13*, 115–122.
21. Ozawa, K.; Minami, H.; Sato, H. Clinical trial simulations for dosage optimization of docetaxel in patients with liver dysfunction based on a log-binominal regression for febrile neutropenia. *Yakugaku Zasshi* **2009**, *129* (6), 749–757.

Clinical Trial Simulations: Earlier Development Phases

Mark Chang

Veristat, Southborough, Massachusetts, U.S.A.

Abstract

This entry discusses the simulation in early phases of clinical trials, specifically through oncology dose-escalation trials to discuss some theoretical and practical issues in trial simulations.

INTRODUCTION

Clinical trial simulation (CTS) is a powerful tool for designing, monitoring, and analyzing clinical trials. In recent years, simulation has been used in pharmacokinetic and pharmacodynamic modeling, assessing cardiologic abnormalities, and optimizing dose-finding strategies.^[1] In this section, we will study the simulation in early phases of clinical trials, specifically through oncology dose-escalation trials to discuss some theoretical and practical issues in trial simulations. We have chosen an oncology study, because of the following reasons.

1. Only a very small number of patients are allowed due to potential high toxicity of the test drug, and precise estimation of the dose-toxicity relationship is challenging.
2. The traditional escalation methods are an ad hoc approach.
3. There are many new approaches proposed using simulations.^[2–5]

Our intention is not to review or compare all the methods and make a recommendation, but to use several typical methods as vehicle to illustrate CTS.

OVERVIEW OF DOSE-ESCALATION TRIAL

Objectives of a Phase I Clinical Trial

The goal of the phase I trial is to characterize the safety profile of a new treatment in humans to set the basis for later investigations of efficacy and superiority, for example, in a cancer study, beginning treatment at a low dose very likely to be safe; so small cohorts of patients are treated at progressively higher doses until drug-related toxicity reaches a predetermined level. The primary objectives are to determine the maximum tolerated dose (MTD) of a drug for a specified regimen and to characterize the dose-limiting toxicity (DLT).

Population for Treatment

The phase I trial should define a standardized treatment regimen to be safely applied to humans and with some biological activities that are worth further investigation for efficacy. For non-life-threatening diseases, phase I trials are usually conducted with human volunteers, at least as long as the expected toxicity is mild and can be controlled without harm. In life-threatening diseases such as cancer or AIDS, phase I studies are conducted in patients because of the aggressiveness and possible harmfulness of treatments, possible systemic treatment effects, and the high interest in the new drug's efficacy in those patients.

Dose-Limited Toxicity and Maximum Tolerated Dose

Drug toxicity is considered tolerable if the toxicity is acceptable, manageable, and reversible. Common Terminology Criteria (CTC) established by the U.S. National Cancer Institute (NCI) is often used as the basis for defining a DLT. For example, in a cancer study, any AE from the CTC catalogue of grade 3 and higher related to treatment is considered a DLT. Often, a judgment of “possibly” and better is considered as drug-related toxicity. The MTD is defined as a dose level at which DLTs occurs with at least a certain frequency, e.g., 1/3.

Dose-Level Selection

The initial dose given to the first patients in a phase I study should be low enough to avoid severe toxicity. The commonly used starting dose is the dose at which 10% mortality (LD_{10}) occurs in mice. The subsequent dose levels are usually selected based on the following multiplicative set, $x_i = f_i \cdot x_{i-1}$ ($i = 1, 2, \dots, k$), where f_i is called the dose-escalation factor. Two popular sequences of dose-escalation factors are the so-called the Fibonacci numbers (2, 1.5, 1.67, 1.60, 1.63, 1.62, 1.62, ...) and modified Fibonacci numbers (2, 1.65, 1.52, 1.40, 1.33, 1.33, ...). The highest dose level in the trial should be selected such that it covers the biologically active dose.

Sample Size Per Dose Level

The number of patients to be treated per dose level should be small enough to limit any potential safety issues, but large enough to allow the investigators to determine the optimal dose levels. For oncology studies, recommendations vary between one and eight patients per dose level. Other suggestions include three to six patients per lower dose level or a minimum of three per dose level and a minimum of five near the MTD.

TRADITIONAL DOSE-ESCALATION DESIGN

There are usually 5 to 10 predetermined dose levels in an oncology dose-escalation study. The commonly employed dose-escalation rules are the traditional escalation rules (TER), which are also known as the “3 + 3” rule. The “3 + 3” rule can be described as follows:

1. Enter three patients at a new dose level.
2. Enter another three patients when one toxicity is observed; and
3. The assessment of the three or six patients will be performed to determine whether the trial should be stopped at that level or escalated to a higher dose level. Basically, there are two types of the “3 + 3” rules namely, TER and strict TER (or STER). TER do not allow dose deescalation, but STER do when two of three patients have DLTs (Fig. 1).

The “3 + 3” STER can be generalized to the $A + B$ TER and STER escalation rules. To introduce the traditional $A + B$ escalation rule, let A , B , C , D , and E be integers. The notation A/B indicates that there are A toxicity incidences out of B subjects and $>A/B$ means that there are more than A toxicity incidences out of B subjects. We assume that there are n predefined doses. In what follows, general $A + B$ designs with and without dose deescalation will be described. Lin and Shih^[6] derive close forms of some operating characteristics of general $A + B$ designs.

$A + B$ Escalation Without Dose Deescalation

The general $A + B$ designs without dose deescalation can be described as follows. Suppose that there are A patients at dose level i . If fewer than C/A patients have DLTs, then the dose is escalated to the next dose level $i + 1$. If more than D/A (where $D \geq C$) patients have DLTs, then the previous dose $i - 1$ will be considered the MTD. If no fewer than C/A but no more than D/A patients have DLTs, B more patients are treated at this dose level i . If no more than E (where $E \geq D$) of the total of $A + B$ patients have DLTs, then the dose is escalated. If more than E of the total of $A + B$ patients have DLT, then the previous dose $i - 1$ will be considered the MTD. It can be seen that the traditional “3 + 3”

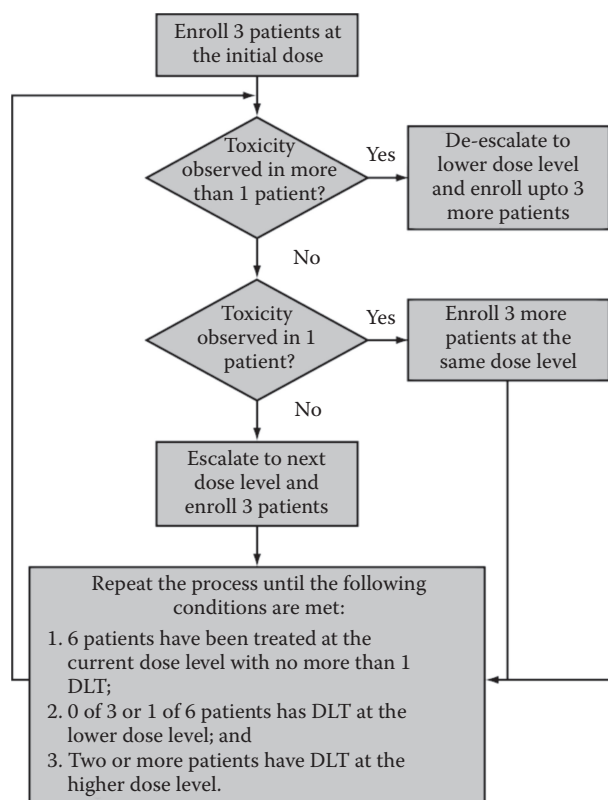


Fig. 1 “3 + 3” Strict traditional escalation rule. (View this art in color at www.informaworld.com.)

design without dose deescalation is a special case of the general $A + B$ design with $A = B = 3$ and $C = D = E = 1$.

$A + B$ Escalation with Dose Deescalation

The general $A + B$ design with dose deescalation is similar to the design without dose deescalation; however, it permits more patients to be treated at a lower dose (i.e., dose deescalation), when excessive DLT incidences occur at the current dose level. The dose deescalation occurs when more than D/A (where $D \geq C$) or more than $E/(A + B)$ patients have DLTs at dose level i . In this case, B more patients will be treated at dose level $i - 1$, provided that only A patients have been previously treated at this prior dose. If more than A patients have already been treated previously, then dose $i - 1$ is the MTD. The deescalation may continue to the next dose level $i - 2$ and so on if necessary.

TWO-STAGE DOSE-ESCALATION DESIGN

Because of safety concerns, the starting dose is usually very low with minimal toxicity, and MTDs are often underestimated using the traditional escalation “3 + 3” rule, especially when there are many dose levels. An improved method is to use a two-stage escalation algorithm (TSER), where in the first stage, only a single patient is treated at

each dose level. When a DLT is observed, the traditional “3 + 3” rule will kick in. In what follows, we will compare TER and TSER using computer simulations.

EVALUATION OF DOSE-ESCALATION ALGORITHMS

There are advantages and disadvantages of different dose-escalation schemes. For example, the traditional “3 + 3” escalation is easy to apply, but the MTD estimation is usually biased, especially when there are many dose levels. The commonly used criteria for evaluating an escalation scheme are as follows. The expected number of patients is used to measure the efficiency of a trial. The number of DLTs is used to measure the safety protection of patients in the trial as a whole. Controlling the number of patients dosed above the MTD is to minimize the risk of all individuals in the trial. Mean predicted MTDs can be used to assess the accuracy of an algorithm. The dispersion of the predicted MTDs from simulations measures the precision of an algorithm.

SIMULATION EXAMPLE USING TER AND TSER

To choose an appropriate dose-escalation algorithm for a phase I oncology study to determine the MTD for a new test drug, computer simulations were conducted to compare the traditional “3 + 3” escalation rule and the two-stage escalation rule mentioned earlier. Based on results from preclinical studies, it was estimated that the toxicity (DLT rate) is 1% for the starting dose of 30 mg/m² (1/10 of the lethal dose). The DLT rate at the MTD was defined as 17% and the MTD was estimated to be 300 mg/m². The true dose toxicity was assumed to follow a logistic model, i.e., $\text{Logit}(p) = -4.952 + 0.011 \text{ Dose}$, where p is the probability of DLT or DLT rate. There were seven planned dose levels with a constant dose increment factor of 1.5. Only one dose level deescalation was allowed. Twenty thousand simulations were conducted using ExpDesign Studio^[5] for each of the two algorithms. The key results are compared in Tables 1 through 3.

The true MTD is somewhere between dose level 6 and 7. From Tables 1 and 2, it is obvious that the two-stage design requires many fewer patients at the dose levels lower than MTD. From Table 3, we can see a large gain in the expected sample size with TSER, i.e., 17 compared 26 in TER. On average, both methods cause two DLTs each, and five patients are treated above the MTD. The TSER increases slightly in precision compared to TER (66 vs. 70 in dispersion of simulated MTDs). In the given scenario, both TER and TSER underestimate the true MTD (about 13%). This bias can be reduced using the so-called continual reassessment method^[4,7] to be discussed next.

Table 1 Simulation Results from TER

Dose level	1	2	3	4	5	6	7
Dose	30	45	68	101	151	228	342
DLT rate (%)	1	1.2	1.5	2.2	3.7	8.4	25
Mean n	3.1	3.1	3.1	3.2	3.5	5.0	5.0
MTD rate (%)	0	0	1	2	6	57	33

Table 2 Simulation Results from TSER

Dose	30	45	68	101	151	228	342
Mean n	1.1	1.1	1.2	1.3	1.5	5.9	5.2
MTD rate (%)	0	0	0	0	2	69	29

Table 3 Comparison of TER and TSER

Method	MTD	MTD	σ_{MTD}	N	DLTs	n
“3 + 3” TER	300	257	70	26	2	5
Two-stage	300	258	66	17	2	5.2

MTD, Simulated MTD; σ_{MTD} , dispersion of MTDs; N , expected sample size; n , patients treated above MTD.

CONTINUAL REASSESSMENT METHOD

The continual reassessment method (CRM)^[2,4,8] was initially developed for phase I oncology dose escalation trial. However, it can be used for many different trials. With CRM, the dose-response relationship is continually reassessed based on accumulative data collected from an ongoing trial. The next enrolled patient is then assigned to the current estimated MTD level. This approach is more efficient than that of the usual TER with respect to the allocation of the MTD; however, the efficiency of CRM could diminish when there are delayed responses and a constraint on dose jump in practice.^[2] In what follows, we will discuss computer simulation using CRM.

Dose-Toxicity Modeling

In most phase I dose-response trials, it is assumed that there is a monotonic relationship between dose and toxicity. This ideal relationship suggests that the biologically inactive dose is lower than the active dose, which is in turn lower than the toxic dose. To characterize this relationship, the choice of an appropriate dose-toxicity model is important. In practice, the logistic model is often utilized

$$p(x) = [1 + b \exp(-ax)]^{-1} \quad (1)$$

where $p(x)$ is the probability of toxicity associated with dose x , and a and b are positive parameters to be determined. Practically, $p(x)$ is equivalent to the toxicity rate or

DLT rate. Denote θ the probability of DLT (or DLT rate) at the MTD. Then, the MTD can be expressed as

$$\text{MTD} = \frac{1}{a} \ln \left(\frac{b\theta}{1-\theta} \right) \quad (2)$$

If we can estimate a and b or their posterior distributions (for the common logistic model, b is predetermined.), we will be able to determine the MTD from (2) or provide predictive probabilities for the MTD. The choice of the toxicity rate at the MTD, θ , depends on the nature of the DLT and the type of target tumor. For an aggressive tumor and a transient and non-life-threatening DLT, θ could be as high as 0.5. For persistent DLT and less-aggressive tumors, it could be as low as 0.1 to 0.25. A commonly used value is somewhere between 0 and $1/3 = 0.33$.^[3]

Reassessment of Model Parameters

The key is to estimate the parameter a in the response model [Eq. (1)]. An initial assumption or a prior distribution of the parameter is necessary to assign patients to the dose level based on the dose–toxicity relationship. This estimation of a is continually updated based on cumulative data observed from the trial. The estimation method can be a Bayesian or frequentist approach. The Bayesian approach leads to the posterior distribution of a . Frequentist approaches such as the maximum likelihood estimate or the least square estimate are straight forward. Note that the Bayesian approach requires specification of a prior distribution of parameter a . It provides posterior distribution of a and predictive probabilities of the MTD.

Prior Distribution of Parameter a

The Bayesian approach requires the specification of prior probability distribution of the unknown parameter a , i.e.,

$$a \sim g_0(a) \quad (3)$$

where $g_0(a)$ is the prior probability distribution.

Likelihood Function

The next step is to construct the likelihood function. Given n observations with y_i associated with dose x_{m_i} , the likelihood function can be written as

$$f_n(\mathbf{r} | a) = \prod_{i=1}^n [p(x_{m_i}, a)]^{r_i} [1 - p(x_{m_i}, a)]^{1-r_i} \quad (4)$$

where

$$r_i = \begin{cases} 1, & \text{if } y_i \geq \tau \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

Assessment of Parameter a

The assessment of the parameters in the model can be carried out in different ways using Bayesian and frequentist approaches. The Bayesian approach assesses the posterior probability distribution of the parameter, while a frequentist approach provides a point estimate of the parameter.

Bayesian approach For the Bayesian approach, the posterior probability can be obtained as follows

$$g(a | \mathbf{r}) = \frac{f_n(\mathbf{r} | a)g_0(a)}{\int f_n(\mathbf{r} | a)g_0(a)da} \quad (6)$$

We then update the probabilities of the outcome or the predictive probabilities

$$p = \int p(\mathbf{x}, a)g(a | \mathbf{r})da \quad (7)$$

Maximum likelihood approach The maximum likelihood estimate of the parameter is given by

$$a_{\text{MLE}} = \arg \max_a \{f_n(\mathbf{r} | a)\} \quad (8)$$

where a_{MLE} is the parameter vector.

Least square approach The least square approach minimizes the difference between predictive probability and the observed rate or probability, which is given below.

$$a_{\text{LSE}} = \arg \min_a \{L_j(a)\} \quad (9)$$

where

$$L(a) = \sum_{i=1}^k (p(x_i, a) - \hat{p}(x_i, a))^2$$

Assignment of Next Patient

The updated dose–toxicity model is used to determine the dose level for the next patient. In other words, the next patient enrolled in the trial is assigned to the current estimated MTD based on the dose–response model. Practically, this assignment is subject to safety constraints such as a limited dose jump.

SIMULATION EXAMPLE USING CRM

We simulate the previous dose-escalation trial using the Bayesian continual reassessment method. The logistic model is used to model the dose–toxicity relationship, i.e., $\Pr(x = 1) = (1 + 150e^{-ax})^{-1}$ where parameter a follows an

Table 4 Simulation Results with CRM

Dose level	1	2	3	4	5	6	7
Dose	30	45	68	101	151	228	342
P_{DLT} (%)	1	1.2	1.5	2.2	3.7	8.4	25
$\hat{P}_{\text{DLT}}$ (%)	0.9	1.1	1.4	2.2	4.1	10.3	30
σ_{DLT} (%)	1	1	0.3	0.6	1.5	5.2	15
n	1	1	1	1	1.24	9.24	6.5

P_{DLT} , DLT rate, $\hat{P}_{\text{DLT}}$ and σ_{DLT} are the predicted rate and its standard deviation. n = number of patients.

uniform prior distribution over a range of [0, 0.3]. The next patient is assigned to the dose level that has the predicted response rate closest to the target DLT rate of 0.17 as defined earlier for the MTD. No response delay is considered. Owing to safety concerns, no dose jump is allowed. Ten thousand simulations were conducted using *ExpDesign Studio*® version 2.0, where 21 subjects were used for each trial or simulation. The results are summarized in Table 4.

In this example, the CRM produces excellent predictions for the DLT rates. We can see that one of the advantages with CRM is that it produces the posterior parameter distribution and predicted probability of DLT for each dose level (any dose) and allows us to select an unplanned dose level as the MTD. In the current case, the true MTD is 300 mg/m² with a DLT rate $P_{\text{DLT}} = 0.17$, which is an unplanned dose level. As long as the dose–response model is appropriately selected, bias can be avoided. This is an advantage compared to TER and TSER discussed earlier. The simulations also indicate that the number of DLTs per trial is 2.5. The number of patients treated with dose higher than MTD is 6.5 per trial. These numbers are larger than those in TER and TSER, which can be viewed as a trade-off with a reduction in bias.

CONCLUSIONS

Simulations have many applications in clinical trials. In early phases of the clinical development with many uncertainties (variables), CTS can be used to assess the impacts of the variables. CTS with the Bayesian approach could produce certain desirable operating characteristics that can answer questions raised in the early development phase such as posterior distribution of toxicity. There are a few things we should caution about with regard to CTS. The quality of simulation results is very much dependent

on the quality of the pseudorandom number. We should be aware that most built-in random number generators in computer languages and software tools are poor in quality; therefore it should not be used directly if we are not sure about the algorithm.

Implementing a high quality of random-number generator is as simple as a dozen of lines of computer code^[9, 10] as implemented by *ExpDesign Studio*. The common steps for conducting a CTS are defining the objectives, analyzing the problem, assessing the scope of the work and proposing time and resources required to complete the task, examining the assumptions, obtaining and validating the data source if applicable, nailing down evaluation criteria for various trial designs/scenarios, selecting an appropriate software tool, outlining the computer algorithms for the simulation, implementing and validating the algorithms, proposing scenarios to simulate, conducting simulations, interpreting results and making recommendations, and, last but not least, addressing the limitations of the performed simulations.

REFERENCES

1. Kimko, H.C., Duffull, S.B., Eds.; *Simulation for Designing Clinical Trials*; Marcel Dekker, Inc.: New York, 2003.
2. Babb, J.S.; Rogatko, A. *Bayesian Methods for Cancer Phase I Clinical Trials, Advances in Clinical Trial Biostatistics*; Nancy, L. Geller, Ed.; Marcel Dekker, Inc.: New York, 2004.
3. Crowley, J. *Handbook of Statistics in Clinical Oncology*; Marcel Dekker, Inc.: New York, 2001.
4. O'Quigley, J.; Pepe, M.; Fisher, L. Continual reassessment method: a practical design for phase I clinical trial in cancer. *Biometrics* **1990**, *46*, 33–48.
5. CTriSoft Intl. *Clinical Trial Design with ExpDesign Studio*; www.ctrisoft.net.; CTriSoft Intl.: Lexington, MA, 2002.
6. Lin, Y.; Shih, W.J. Statistical properties of the traditional algorithm-based designs for phase I cancer clinical trials. *Biostatistics* **2001**, *2*, 203–215.
7. Chang, M.; Chow, S.C. A hybrid Bayesian adaptive design for dose response trials. Special Issue, *J. Biopharm. Stat.* **2005**.
8. O'Quigley, J.; Shen, L. Continual reassessment method: a likelihood approach. *Biometrics* **1996**, *52*, 673–684.
9. Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P. *Numerical Recipes in C++*, 2nd Ed.; Cambridge University Press: Cambridge, 2002.
10. Gentle, J.E. *Random Number Generator and Monte Carlo Methods*; Springer-Verlag, Inc.: New York, 1998.

Clinical Trial Simulations: Later Development Phases

Mark Chang

Veristat, Southborough, Massachusetts, U.S.A.

Abstract

Clinical trial simulation plays an important role in drug development and, in recent years, it has become increasingly popular in the pharmaceutical industry. The focus of this entry is the simulation for adaptive trial designs, which will be defined in detail through several statistical examples.

INTRODUCTION

The traditional approach to drug development is a mixture of statistical and ad hoc methods. In many cases, assumptions are often introduced to simplify the problem and fit it into the statistics paradigm. In other cases, ad hoc decisions have to be made to deal with more complicated situations. Clinical trial simulation (CTS) with computers has been used for several decades.^[1] However, it was not until recent years that CTS started to play an important role in drug development and became increasingly popular in the pharmaceutical industry.^[2] Computer simulation can be used to model very complicated situations (virtually any situation that we can specify) with minimum assumptions. It can be used to search for an optimal trial design or clinical development plan. CTS can also be used to visualize the dynamic trial process from recruiting patients, distributing drugs, and conducting treatments to observing pharmacokinetic processes and clinical responses (efficacy and safety). Trial simulation can be combined with analytical approaches to improve the performance or efficiency of the simulation. For instance, a biotech company planned to launch a phase IIb early ACS (early glycoprotein IIb–IIIa inhibition in non-ST-segment elevation acute coronary syndrome) study. However, the pivotal studies used for approval were done more than six years ago. Since then, there have been noticeable changes in the patient population composition and medical technology. Concerns were raised. What is the likelihood of success for the trial? What would be the optimal design? What treatment effect would likely be seen? What would be the right population for the trial? A set of stringent patient inclusion/exclusion criteria may increase the probability of success, but it could also shrink the patient population to exclude patients who may have gained potential benefit from the drug. On the other hand, if we target a larger patient population by loosening the inclusion/exclusion criteria, the average effect of the drug would likely diminish and so would the probability of success. Facing these complex issues, the company decided to launch CTS to assist in the decision making.

First, a huge data source was identified from relevant trials conducted internationally over the past eight years. Then, analytical approaches, such as logistic regression, were used to search for sensitive variables (baseline and procedure variables) that are both clinically and statistically significant. As a result, a dozen sensitive variables were identified among nearly 40 potential candidates. Next, a stochastic model was introduced for computer simulations. Meanwhile, the data source was reconstructed as a training set and a test/validation set. Hundreds of simulations were conducted under various scenarios with different combinations of the sensitive variables. The simulations were able to provide the probability of trial success, distribution of the treatment effect, and the associated cost for each design scenario. The company made the decision to launch the large clinical trial with the assistance of the valuable information from the computer simulations. Further discussion of the whole spectrum of the simulations is beyond the scope of this section and will not be pursued here. Instead, in what follows, we will focus on the simulation for adaptive trial designs, which will be defined shortly. There are several reasons to choose adaptive design for the simulation topic: (1) adaptive design is very attractive to both practitioners and researchers;^[1,3–23] (2) the statistical theory of adaptive design is very complicated with very limited analytical solutions available under some strong assumptions, and it is unlikely that there will be sufficient analytical solutions for complicated adaptive designs in the near future; and (3) computer simulation of an adaptive design is straight forward and easy to implement.

OVERVIEW OF ADAPTIVE DESIGN

Drug development is a sequence of complicated decision-making processes. Options are provided at each stage, and decisions are dependent on the prior information and the probabilistic consequence of each action (decision). This requires the trial design to be flexible such that it can be modified during the trial process. Adaptive design emerges

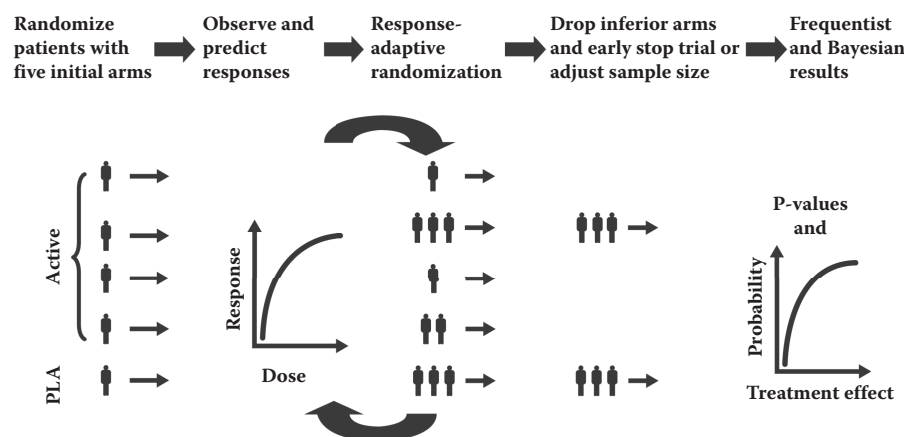


Fig. 1 Overview of adaptive design

for this reasoning and has become very attractive in the pharmaceutical industry. Adaptive design refers to a design that allows adaptations or modifications to some aspects of the trial after its initiation without undermining the validity and integrity of the trial. The adaptations may include, but are not limited to: (1) early stopping due to efficacy or futility; (2) sample size reestimation; (3) dropping inferior treatment groups; (4) adding new treatment groups; (5) response-adaptive randomization; (6) modification of the target population; (7) study endpoint change; (8) hypothesis change; (9) treatment switching (crossover); and (10) any combination of the above adaptations. An adaptive design usually consists of two or more stages; at each stage, data analyses are conducted, and adaptations are made based on updated information to maximize the chance of success.

Although methods using the Brownian motion^[15] and Fisher's combination of independent p -values^[4,5] have been used in group-sequential and adaptive designs, because of great uncertainty surrounding source information, mechanism of action, treatment effect and its variability, interaction with environment, etc., using a traditional mathematical or statistical approach makes it difficult to model the dynamic system of trials. CTS is a powerful tool for achieving the optimal design. An overall process of an adaptive design is depicted in Fig. 1.

UTILITY-BASED TRIAL OBJECTIVE

A clinical trial typically involves multiple endpoints such as efficacy, safety, and cost. Therefore, a single measure, i.e., utility index, that summarizes the effects of major endpoints is desirable. The trial objective then becomes to find the dose or treatment with the maximum response probability (rate). The response probability is defined as $\Pr(u \geq c)$, where u is the utility index and c is a threshold. The utility index is the weighted average of trial endpoints such as safety and efficacy. The weights and the threshold are often determined by experts in the relevant field.

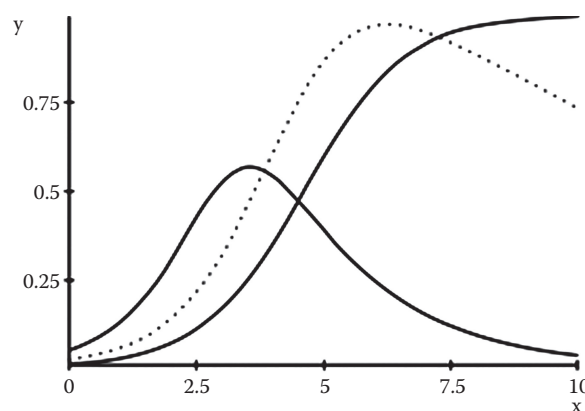


Fig. 2 Hyperlogistic function

DOSE-RESPONSE MODEL

The response of an ongoing trial can be modeled using a function. We find that the following so-called hyperlogistic function can be used to model many different response shapes (Fig. 2). The hyperlogistic function is defined by the probability of response

$$\Pr(x=1) = (a_1 \exp(a_2 x) + a_3 \exp(-a_4 x))^{-a_5}$$

The modeling can be on a continual basis, i.e., the model is updated when new response data become available. This approach refers to the continual reassessment method (CRM), which can be either a Bayesian- or frequentist-based method. If the observed responses are used as the basis for an adaptation instead of modeled or predicted response, we call it a null-model approach.

ADAPTATION RULES

Randomization Rules

It is desirable to randomize more patients to superior treatment groups. This can be accomplished by increasing

the probability of assigning a patient to the treatment group when the evidence of responsive rate increases in a group. The response-adaptive randomization rule can be the randomized-play-the-winner or Utility-offset model (UOM).

Response-adaptive randomization requires unblinding the data, which may not be feasible at real time. There is often a delayed response, i.e., randomizing the next patient before knowing responses of previous patients. Therefore, it is practical to unblind the data several times during the trial, i.e., group sequential response-adaptive randomization, instead of fully sequential adaptive randomization.

Early Stopping Rules

It is desirable to stop a trial when the efficacy or futility of the test drug becomes obvious during the trial. To stop a trial prematurely, we provide a threshold for the number of subjects randomized and at least one of the following:

1. *Utility rules:* The difference in response rate between the most responsive group and the control group exceeds a threshold, and the corresponding two-sided 95% naive confidence interval lower bound exceeds a threshold.
2. *Futility rules:* The difference in response rate between the most responsive group and the control is lower than a threshold, and the corresponding two-sided 90% naive confidence interval upper bound is lower a threshold.

Rules for Dropping Losers

In addition to the response-adaptive randomization, you can also improve the efficiency of a trial design by dropping some inferior groups (losers) during the trial. To drop a loser, we provide two thresholds for: (1) maximum difference in response rate between any two dose levels and (2) the corresponding two-sided 90% naive confidence lower bound. We may choose to retain all the treatment groups without dropping a loser and/or to retain the control group with a certain randomization rate for the purpose of statistical comparisons between the active groups and the control.

Sample Size Adjustment

Sample size determination requires anticipation of the expected treatment effect size defined as the expected treatment difference divided by its standard deviation. It is not uncommon that the initial estimation of the effect size turns out to be too large or small, which consequently leads to an underpowered or overpowered trial. Therefore,

it is desirable to adjust the sample size according to the effect size for an ongoing trial.

The sample size adjustment is determined by a power function of treatment effect size, i.e.,

$$N = N_0 \left(\frac{E_{0\max}}{E_{\max}} \right)^a \quad (1)$$

where N is the newly estimated sample size, N_0 , the initial sample size, and a is a constant. The effect size $E_{\max}$ is defined as

$$E_{\max} = \frac{p_{\max} - p_1}{\sigma^2}; \quad \sigma^2 = \bar{p}(1 - \bar{p});$$

$$\bar{p} = \frac{p_{\max} + p_1}{2}$$

$p_{\max}$ and p_1 are the maximum response rates and the control response rate, respectively, and $E_{0\max}$ is the initial estimation of $E_{\max}$.

RESPONSE-ADAPTIVE RANDOMIZATIONS

Conventional randomization refers to any randomization procedure with a constant treatment allocation probability such as simple randomization. Unlike the conventional randomization, response-adaptive randomization is a randomization in which the probability of allocating a patient to a treatment group is based on the response of the previous patients. The purpose is to improve the overall response rate in the trial. There are many different algorithms such as random-play-the-winner (RPW) (Fig. 3), the UOM and the maximum utility model (MUM).

RPW

The generalized RPW denoted by $\text{RPW}(n_1, n_2, \dots, n_k; m_1, m_2, \dots, m_k)$ can be described as follows:

1. Place n_i balls of the i th color (corresponding to the i th treatment) into an urn ($i = 1, 2, \dots, k$), where k is the number of treatment groups. There are initially $N = \sum n_i$ balls in the urn.
2. Randomly choose a ball from the urn. If it is the i th color, assign the next patient to the i th treatment group.
3. Add m_k balls of the i th color to the urn for each response observed in the i th treatment. This creates more chances for choosing the i th treatment.
4. Repeat steps 2) and 3).

When $n_i = n$ and $m_i = m$ for all i , we simply write $\text{RPW}(n, m)$ for $\text{RPW}(n_1, n_2, \dots, n_k; m_1, m_2, \dots, m_k)$.

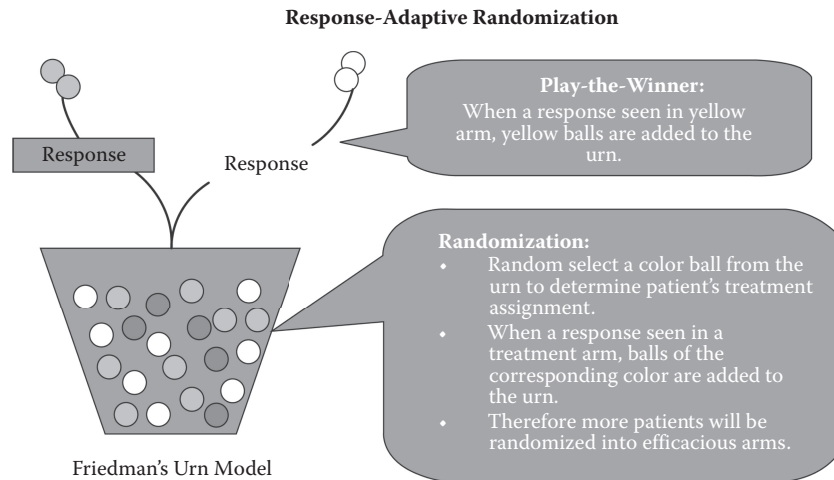


Fig. 3 Random-play-the-winner

UOM

To have a high probability of achieving target patient distribution among the treatment groups, the probability of assigning a patient to a group should be proportional to the corresponding predicted or observed response rate minus the proportion of patients that have been assigned to the group, i.e.,

$$\Pr(i) = \begin{cases} \frac{1}{k}, & S_R = 0 \\ \frac{1}{c} \max \left\{ 0, \left(\frac{R_i}{S_R} - \frac{n_i}{N} \right) \right\}, & S_R > 0 \end{cases} \quad (2)$$

where the normalization factor $c = \sum_i \left(\frac{R_i}{S_R} - \frac{n_i}{N} \right)$, n_i is the number of patients that have been randomized to the i th group, R_i = the observed response rate for the i th group, $S_R = \sum_{i=1}^k R_i$, k = the number of dose groups, and N = the total estimated number of patients in the trial.

MUM

The MUM for the adaptive randomization always assigns the next patient to the group that has the highest response rate based on current estimation of either the observed or model-based predicted response rate.

NULL MODEL VS. MODEL APPROACH

It is interesting to compare model and null-model approaches. When the sample size is larger than 20 per group, there is no obvious advantage of using the model-based method with respect to the precision and accuracy (Table 1). Therefore, a null-model approach will be used in the subsequent simulations.

Table 1 Comparisons of Simulations Results

Dose level	1	2	3	4	5
Target rate	0.02	0.07	0.37	0.73	0.52
Simulated rate	0.02	0.07	0.36	0.73	0.52
Predicted rate	0.02	0.07	0.40	0.65	0.41
Standard deviation	0.00	0.02	0.11	0.09	0.04
Number of subjects	1.02	2.48	12.6	20.5	13.4
Simulated rate	0.02	0.05	0.36	0.73	0.51
Predicted rate	0.02	0.07	0.40	0.63	0.40
Standard deviation	0.00	0.02	0.13	0.11	0.05
Number of subjects	1.00	1.62	7.50	11.9	8.00
Simulated rate	0.02	0.06	0.34	0.72	0.51
Predicted rate	0.02	0.07	0.37	0.58	0.38
Standard deviation	0.00	0.02	0.15	0.14	0.06
Number of subjects	1.00	1.03	4.68	7.60	5.68

TEST STATISTIC

It is very interesting to know that the choice of test statistics for hypothesis tests is very flexible if the analysis is carried out through computer simulations. The only requirement is that the test statistic should be a monotonic function of both the treatment effect δ and the sample size n , which can be, for example, $\sqrt{n}\delta$ the treatment difference or the effect size $\sqrt{n}\delta/\sigma$, where σ is the standard deviation of δ . Using computer simulation, it is easy to generate the distributions of the test statistic under the null hypothesis and alternative hypothesis or any other specified conditions for monitoring purposes.

α -ADJUSTMENT

The α -adjustment is required when: (1) there are multiple comparisons with more than two groups involved; (2) there

Table 2 α Inflation/Deflation

Randomization	Inflated FWE α	Adjusted α
RPW(1,0)	0.146	0.015
RPW(1,1)	0.143	0.007

Note: Five group with total $N = 100$. H_0 rate = 0.5.

are interim looks, i.e., early stopping for futility or efficacy; and (3) there is a response-dependent sampling procedure such as response-adaptive randomization and unblinded sample size reestimation. When samples or observations from the trial are not independent, the response data are no longer normally distributed. Therefore, the p -value from a normal distribution assumption should be adjusted, or equivalently the α should be adjusted if the p -value is not adjusted (Table 2). For the same reason, the other statistic estimates from normal assumption should also be adjusted.

The adjusted α can be found easily using simulation. Run simulations under the null hypothesis H_0 to find the adjusted α for the pairwise comparisons such that familywise error rate = 0.025. Then run simulations under the alternative hypothesis H_a using the adjusted α to find the power for various sample sizes and other scenarios.

BIAS DUE TO ADAPTIVE RANDOMIZATION RPW(1,1)

The commonly used estimators that are based on the assumption of independent samples are often biased in the case of adaptive design (Table 3).

SIMULATION EXAMPLES

To investigate the effect of the adaptations, we will compare the classic, group sequential, and adaptive designs with regard to their operating characteristics using computer simulations. In what follows, each example represents a different trial design.

Examples 1 through 5 will use the following scenario: Assume a phase II oncology trial with two treatment groups. The primary endpoint is tumor response

Table 3 Bias in Rate Due to Adaptive Randomization

Dose level	1	2	3	4	5
Target rate	0.5	0.1	0.2	0.7	0.55
Number of patients	130	14	23	362	171
Observed rate	0.47	0.08	0.16	0.7	0.53
Standard deviation	0.10	0.08	0.10	0.05	0.08
Bias in rate	0.03	0.02	0.04	0.00	0.02

$N = 700$; 100,000 simulations.

(PR and CR) and the estimated response rates for the two groups are 0.2 and 0.3, respectively. We use simulation to calculate the sample size required, given that one-sided $\alpha = 0.05$ and power = 80%.

Example 1: Conventional Design with Two Parallel Treatment Groups

A classic fixed sample size design with 600 subjects will have a power of 81.4% at one-sided $\alpha = 0.025$. The total number of responses per trial is 150 based on 10,000 simulations.

Example 2: Flexible Design with Sample Size Reestimation

The power of a trial is heavily dependent on the estimated effect size; therefore it is desirable to have a design that allows modification of the sample size at some point during the trial. Let us redesign the trial in Example 1 such that it allows a sample size reestimation and then study the robustness of the design.

To control the FWE at 0.025, the α must be adjusted to 0.023 which can be done by computer simulation under the null hypothesis. The average sample size is 960 under the null hypothesis. Using the algorithm for sample size reestimation (1), where $E_{0\max} = 0.1633$ and $a = 2$, the design has 92% power with an average sample size of 821.5.

Now assume that the initial effect sizes are not 0.2 vs. 0.3 for the two treatment groups. Instead, they are 0.2 and 0.28, respectively. We want to know what the power of the flexible design will have. Keep everything the same (also keep $E_{0\max}$ to be 0.1633), but change the response rates to 0.2 and 0.28 for the two dose levels and run the simulation again. It turns out that the design has 79.4% power with an average sample size of 855.

Given the two response rates 0.2 and 0.28, the design with a fixed sample size of 880 has a power of 79.4%. We can see that there is a saving of 25 patients by using the flexible design. If the response rates are 0.2 and 0.3, for 92.1% power, the required sample size is 828 with the fixed sample size design, which means that the flexible design saves six to seven subjects. A flexible design increases power when observed effect size is less than expected, while a traditional design with a fixed sample size either increases or decreases the power regardless of the observed effect size when the sample increases.

Example 3: Design with RPW Randomization

To randomize more patients to the superior treatment group using response-adaptive randomization, the same example with a sample size of 600 subjects is used. The commonly used response-adaptive randomization is RPW(1, 1), i.e., one initial ball for each group and one additional ball with the corresponding color for each response. The data will

Table 4 Group Sequential Design

Dose level	Control	Active
Number of patients	244	244
Response rate	0.2	0.3
Observed rate	0.198	0.303
Standard deviation	0.028	0.033

be unblinded for every 100 new patients. The adjusted α is found to be 0.02. The design has 77.3% power with an average sample size of 600. On average, there are 223 subjects in dose level 1 and 377 in dose level 2. The expected number of responses is 158. Comparing this to the classic design in Example 1, the RPW design gains eight responses but loses 4% power.

Example 4: Group Sequential Design with One Interim Analysis

Group sequential design is very popular in clinical trials. For the two arm trial, the adjusted α is found to be 0.024. The maximum number of subjects is 700. The trial will stop if 350 or more patients are randomized, and one of the following criteria is met: 1) The efficacy (utility) stopping criterion: The maximum difference in response rate between any dose and the control is larger than 0.1 with the lower bound of the two-sided 95% naive confidence interval greater than or equal to 0.0 or 2) The futility stopping criterion: The maximum difference in response rate between any dose and dose level 1 is less than 0.05 with the upper bound of the one-sided 95% naive confidence interval less than 0.1 (Table 4).

The average total number of subjects for each trial is 488. The total number of responses per trial is 122. The probability of correctly predicting the most responsive dose level is 0.988 based on observed rates. Under the alternative hypothesis, the probability of early stopping for efficacy is 0.505, and the probability of early stopping for futility is 0.104. The power for testing the treatment difference is 0.825.

Example 5: Adaptive Design Permitting Early Stopping and Sample Size Reestimation

Sometimes it is desirable to have a design permitting both early stopping and sample size modification (Table 5).

With an initial sample size of 700 subjects, a grouping size of 350, and a maximum sample size of 1000, the one-sided adjusted α is found to be 0.05. The simulation results are presented as follows:

The maximum sample size is 700. The trial will stop if 350 patients or more are randomized, and one of the following criteria is met: (1) the efficacy (utility) stopping criterion: the maximum difference in response rate between

Table 5 Design with Early Stopping and n -Reestimation

Dose level	Control	Active
Number of patients	271	272
Response rate	0.2	0.3
Observed rate	0.198	0.303
Standard deviation	0.028	0.032

any dose and the control is larger than 0.1 with the lower bound of the two-sided 95% naive confidence interval greater than or equal to 0.0 or (2) the futility stopping criterion: the maximum difference in response rate between any dose and the control is less than 0.05 with the upper bound of the one-sided 95% naive confidence interval less than 0.1. The sample size will be reestimated when 350 subjects have been randomized.

When the null hypothesis is true ($p_1 = p_2 = 0.2$), the average total number of subjects for each trial is 398.8. The probability of early stopping for efficacy is 0.0096. The probability of early stopping for futility is 0.9638.

When the alternative hypothesis is true ($p_1 = 0.2$, $p_2 = 0.3$), the average total number of subjects for each trial is 543.5. The total number of responses per trial is 136. The probability of correctly predicting the most responsive dose level is 0.985 based on observed rates. The probability of early stopping for efficacy is 0.6225. The probability of early stopping for futility is 0.1546. The power for testing the treatment difference is 0.842.

Examples 6 through 8 are for the same scenario as the six arm study with response rates of 0.5, 0.4, 0.5, 0.6, 0.7, and 0.55 for the six dose levels from doses 1 through 6, respectively.

Example 6: Conventional Design with Multiple Treatment Groups

With 800 subjects, a 0.5 response rate under H_0 , and a grouping size of 100, we found the one-sided adjusted α to be 0.0055. The total number of responses per trial is 433. The probability of correctly predicting the most responsive dose level is 0.951 based on observed rates. The power for testing the maximum effect comparing any dose level to the control is 80%. The powers for comparing each of the five dose levels to the control are 0, 0.008, 0.2, 0.796, and 0.048, respectively.

Example 7: Response-Adaptive Design with Multiple Treatment Groups

To further investigate the effect of RPW randomization RPW(1, 1), a design with 800 subjects, a grouping size of 100, and a response rate of 0.2 under the null hypothesis are simulated. The one-sided adjusted α is found to be 0.016. Using this adjusted α and response rates 0.5, 0.4, 0.5,

Table 6 Design with RPW(1, 1) Under H_0

Dose level	1	2	3	4	5	6
Response rate	0.5	0.5	0.5	0.5	0.5	0.5
Observed rate	0.50	0.49	0.49	0.49	0.49	0.49

Table 7 Design with RPW(1, 1) Under H_a

Dose level	1	2	3	4	5	6
Number of subjects	200	74	100	133	176	116
Response rate	0.5	0.4	0.5	0.6	0.7	0.55
Observed rate	0.50	0.39	0.49	0.59	0.7	0.54

0.6, 0.7, and 0.55 for dose levels 1 through 6, respectively, the simulation indicates that the designed trial has 86% power and 447 responders per trial on average. In comparison to 80% power and 433 responders for the design with simple randomization RPW(1, 0), the adaptive randomization is superior in both power and number of responders. The simulation results also indicate that there are biases in the estimated mean response rates in all dose levels except dose level 1, where a fixed randomization rate is used (Tables 6 and 7).

The average total number of subjects for each trial is 800. The total number of responses per trial is 446.8. The probability of correctly predicting the most responsive dose level is 0.957 based on observed rates. The power for testing the maximum effect comparing any dose level to the control (dose level 1) is 0.861 at a one-sided significance level (α) of 0.016. The powers for comparing each of the five dose levels to the control are 0, 0.008, 0.201, 0.853, and 0.051, respectively.

Example 8: Adaptive Design with Dropping Losers

Implementing the mechanism of dropping losers can also improve the efficiency of a design. With 800 subjects, a grouping size of 100, a response rate of 0.2 under the null hypothesis, and a fixed randomization rate in dose level 1 at 0.25, an inferior group (loser) will be dropped if the maximum difference in response between the most effective group and the least effective group (loser) is greater than 0 with the lower bound of the one-sided 95% naive confidence interval greater than or equal to 0. Using the simulation, the adjusted α is found to be 0.079. From the simulation results below, more biases can be observed with this design. The design has 90% power with 467 responders. The probability of correctly predicting the most responsive dose level is 0.965 based on observed rates. The powers for comparing each of the five dose levels to the control (dose level 1) are 0.001, 0.007, 0.205, 0.889, and 0.045, respectively. The design is superior to both RPW(1, 0) and RPW(1, 1) (Tables 8 and 9).

Table 8 Bias in Rate with Dropping Losers Under H_0

Dose level	1	2	3	4	5	6
Response rate	0.5	0.5	0.5	0.5	0.5	0.5
Observed rate	0.50	0.46	0.46	0.46	0.46	0.46

Table 9 Bias in Rate with Dropping Losers Under H_a

Dose level	1	2	3	4	5	6
Number of subjects	200	26	68	172	240	95
Response rate	0.5	0.4	0.5	0.6	0.7	0.55
Observed rate	0.50	0.37	0.46	0.57	0.69	0.51

Example 9: Dose Response Trial Design

The trial objective is to find the optimal dose with the highest response rate. There are five dose levels and 30 planned subjects in each simulation. The hyperlogistic model is defined with the parameters $a_3 \in [20, 100]$ and $a_4 \in [0, 0.05]$. The RPW(1, 1) is used for the randomization. The simulation results show that the probability of correctly predicting the most responsive dose level is 0.992 by the model and only 0.505 based on observed rates (Table 10).

DISCUSSION OF SIMULATION RESULTS

From classic design to group sequential design to adaptive design, each step forward increases in complexity and at the same time improves the efficiency of clinical trials as a whole. Adaptive designs can increase the number of responses in a trial and provide more benefits to the patient in comparison to the classic design. With sample size reestimation, an adaptive design can preserve the power even when the initial estimations of treatment effect and its variability are inaccurate. In the case of a multiple-arm trial, dropping inferior arms or response-adaptive randomization can improve the efficiency of a design dramatically. Finding analytic solutions for adaptive designs is theoretically challenging. However, computer simulation makes it easier to achieve an optimal adaptive design. It allows a wide range of test statistics as long as they are monotonic functions of treatment effects. Adjusted α s and

Table 10 Dose Response Trial Design

Dose level	1	2	3	4	5
Number of subjects	15	30	50	85	110
Response rate	0.2	0.3	0.6	0.7	0.5
Observed rate	0.193	0.294	0.593	0.691	0.489
Predicted Rate	0.098	0.181	0.406	0.802	0.379
σ_{obs}	0.185	0.209	0.221	0.204	0.226
σ_{prd}	0.074	0.073	0.104	0.151	0.056

p -values because of response-adaptive randomization and other adaptations with multiple comparisons can be determined easily using computer simulations. Unbias in point estimations with adaptive designs has not been completely resolved yet using computer simulations. However, the bias can be ignored in practice by using a proper grouping size (cluster) such that there are only a limited number of adaptations (<8).

CONCLUSIONS

We have demonstrated that CTS can be used for various complex designs. By comparing the operating characteristics of each design provided by CTS, we are able to choose an optimal design or development strategy. Computer simulation is commonly seen in statistics literature related to clinical trials, most of them for power calculations and data analyses. We will see more pharmaceutical companies using CTS for their clinical development planning to streamline the drug process, increase the probability of success, and reduce the cost and time-to-market. Ultimately, CTS will bring the best treatment to the patients. When conducting CTS for an adaptive design, the following is recommended to protect the integrity of the trial: Specify in the trial protocol: (1) the type of adaptive design; (2) details of the adaptations to be used; (3) the estimator for treatment effect; and (4) the hypotheses and the test statistics. Run the simulation under the null hypothesis and construct the distribution of the test statistic. To estimate the sample size for the adaptive design, run the simulation under the alternative hypothesis and construct the distribution of the test statistic. Sensitivity analyses are suggested to simulate potential major protocol deviations that would impact the validity of the trial simulations. To achieve an optimal design, a Bayesian or frequentist–Bayesian hybrid adaptive approach should be used. There are several simulation tools available on the web. For adaptive designs discussed in this section ExpDesign Studio® can be used.

REFERENCES

1. Maxwell, C.; Domemet, J.G. Instant experience in clinical trial: a novel aid to teaching by simulation. *J. Clin. Pharmacol* **1971**, *11* (5), 323–331.
2. Parmigiani, G. *Modeling in Medical Decision Making*; John Wiley & Sons, Ltd.: Chichester, 2002.
3. Babb, J.S.; Rogatko, A. Bayesian methods for cancer phase I clinical trials. In *Advances in Clinical Trial Biostatistics*; Geller, N.L., Ed.; Marcel Dekker, Inc.: New York, 2004.
4. Bauer, P.K. Evaluation of experiments with adaptive interim analysis. *Biometrics* **1994**, 1029–1041.
5. Bauer, P.; Rohmel, J. An adaptive method for establishing a dose-response relationship. *Stat. Med.* **1995**, *14*, 1595–1607.
6. Chang, M.; Chow, S.C. A hybrid Bayesian adaptive design for dose response trials. *J. Biopharmaceut. Stat.* **2005**, *15*, 667–691.
7. Chow, S.C.; Chang, M.; Pong, A. Statistical consideration of adaptive methods in clinical development. *J. Biopharmaceut. Stat.* **2005**, *15*, 557–591.
8. Cui, L.; Hung, H.M.J.; Wang, S.J. Modification of sample size in group sequential trials. *Biometrics* **1999**, *55*, 853–857.
9. Hommel, G. Adaptive modifications of hypotheses after an interim analysis. *Biomet. J.* **2001**, *43*, 581–589.
10. Jennison, C.; Turnbull, B.W. *Group Sequential Method with Applications to Clinical Trials*; Chapman & Hall/CRC: Boca Raton, FL, 2000.
11. Jennison, C.; Turnbull, B.W. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Stat. Med.* **2003**, *22*, 971–993.
12. Kieser, M.; Bauer, P.; Lehmacher, W. Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biomet. J.* **1999**, *41*, 261–277.
13. Kimko, H.C.; Duffull, S.B., Eds.; *Simulation for Designing Clinical Trials*; Marcel Dekker, Inc.: New York, 2003.
14. Lan, K.K.G.; Trost, D.C. Estimation of parameters and sample size reestimation. *Proc. Biopharm Sec.* **1997**, 48–51.
15. Lan, G.K.K. Problems and issues in adaptive clinical trial design—Presented at New Jersey Chapter of the American Statistical Association, Piscataway, NJ, June 4, 2002.
16. Li, G.; Shih, W.J.; Xie, T.; Lu, J. A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics* **2002**, *3*, 277–287.
17. Liu, Q.; Proschan, M.A.; Pledger, G.W. A unified theory of two-stage adaptive designs. *J. Am. Stat. Assoc.* **2002**, *97*, 1034–1041.
18. Mehta, C.R.; Tsiatis, A.A. Flexible sample size considerations using information based interim monitoring. *Drug Inform. J.* **2001**, *35*, 1095–1112.
19. Muller, H.H.; Schafer, H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and classical group sequential approaches. *Biometrics* **2001**, *57*, 886–891.
20. Proschan, M.A.; Hunsberger, S.A. Designed extension of studies based on conditional power. *Biometrics* **1995**, *51*, 1315–1324.
21. Shen, Y.; Fisher, L.D. Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* **1999**, *55*, 190–197.
22. Tsiatis, A.A.; Mehta, C.R. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* **2003**, *90*, 367–378.
23. Wittes, J.; Brittain, E. The role of internal pilot studies in increasing efficiency of clinical trials. *Stat. Med.* **1990**, *9*, 65–72.

Clinical Trial: Combination Drug Development

H. M. James Hung

Division of Biometrics I, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

Combination treatments are widely used in medicine. This entry discusses the many interesting yet difficult statistical design and inference problems in regards to combination treatment trials.

Keywords: Combination dose; Combination treatment; Fixed dose.

INTRODUCTION

Combination treatments are widely used in medicine. The U.S. Food and Drug Administration's general policy regarding fixed-dose, combination-dosage form prescription drugs for humans is that two or more drugs may be combined in a single-dosage form when each component makes a contribution to the claimed effects of the combination and the dosage of each component is such that the combination is safe and effective for a significant patient population requiring such concurrent therapy as defined in the labeling for the drug [21 CFR 300.50]. As described in Ref. [1], this general rule could be interpreted in several distinct contexts between the notion of contributing in the strict sense of superiority and the notion of additivity, synergy, and antagonism that scientists in the fields of pharmacology and toxicology have long considered. In clinical studies, the most prevalent interpretation of the rule is that the combination is superior to its components in terms of effectiveness or safety. The relationship between these notions has been explored by Laska et al.^[1]

In many diseases, for example, hypertension, drug treatments for the patients are administered according to a dose titration schedule. Dose-effect information is essential for drug labeling, particularly if one or both of the component drugs have dose-dependent side effects. Therefore, a multilevel factorial design trial for simultaneously studying multiple-dose combinations is often asked for during the development of combination drugs, as articulated in Refs. [2] and [3]. Such a factorial design trial is also needed for studying a combination drug proposed for use as an initial therapy.

STATISTICAL DESIGNS AND ANALYSIS METHODS

Fixed-Dose Combination

The clinical trial design for evaluation of a fixed-dose combination from two drugs is usually a full two-by-two

factorial design that exposes the four combinations composed from placebo and the specified dose for one drug and placebo and the specified dose for the other, as illustrated in the following layout.

P	B
A	AB

Here P labels the placebo-placebo combination, A is the combination of Drug A and the placebo of Drug B, B is the combination of Drug B and the placebo of Drug A, and AB is the combination of Drug A and Drug B. The placebo-placebo combination may sometimes be voided but usually it is preferred to include this combination, at least, for the purpose of checking the activity of each component drug. The subjects enrolled are randomly allocated into the three or four combination cells. The allocation can be in equal or unequal proportion.

The primary statistical hypothesis of concern involves two pairwise comparisons between the combination and the two components. Both of the comparisons must succeed in order to prove the notion of contributing. This is a well-known statistical testing problem studied in the literature.^[4,5] A natural joint test is the so-called min test^[6] that is generated by taking the minimum of the individual standardized tests for the respective pairwise comparisons. In parametric testing, such as testing for mean difference, the distribution of the min test involves a primary parameter and a nuisance parameter. The primary parameter is the minimum of the values of the parameters for the two respective comparisons. The nuisance is the difference of the values of the parameters for the two comparisons. If the hypothesis pertains to a single response, then the primary parameter measures the least gain of effect by use of the combination relative to its components whereas the nuisance is the mean difference of the components.

Snapinn^[7] constructed five tests based on the min test. Each of his first four tests ($T_1 - T_4$) employs a different estimate of the nuisance parameter that is derived from

the same set of the data used for testing the hypothesis of the primary parameter. As reported by Snapinn^[7] and Hung et al.,^[3] the tests could be biased in favor of the combination treatment. That is, if the true value of the nuisance parameter differs from the estimate, then the tests could lead one to falsely assert that the combination is superior to the components more often than it should. The maximum Type I error probability for each of these tests exceeds the target α level. Snapinn's fifth test (T_5) is the α -level min test, later presented by Laska and Meisner,^[6] which requires the individual test statistics to reject the respective null hypotheses at the level of at most α . Equivalently, the individual p values for all pairwise comparisons must not exceed α . The α -level min test has existed early in the literature,^[4] and its maximum Type I error probability is α . Laska and Meisner^[6] showed an optimality property of this test, the implication of which needs to be an important consideration in regulatory applications. The performance of the α -level min test as a function of the nuisance parameter was further explored in Refs. [3,8–10]. Others (e.g., Refs. [11–13]) also contemplated the possibility of improving the min test. It is clear that the α -level min test cannot be improved unless the true value of the nuisance parameter is known to be in a very narrow range or under a specific alternative.

The statistical power of the α -level min test depends on the primary parameter and the nuisance parameter discussed above. In case of testing a mean parameter, given any fixed value of the primary parameter, the power of the α -level min test increases as a function of the absolute value of the nuisance parameter. Therefore, given any value of the primary parameter, the α -level min test has the smallest power when the nuisance parameter is null. Ideally, the sample size for a fixed-dose combination trial should be planned to ensure a sufficient power for detecting an expected value of the primary parameter, assuming that the nuisance parameter is zero, i.e., the two components have an identical effect.

The problem of comparing a combination treatment with its component drugs on multiple-response variables has been extensively studied in Ref. [14]. The multivariate responses present more than one way of interpreting “making a contribution” depending on the claimed effects. The optimality of the α -level min test in the univariate response case also carries to the multiple-response case for showing that the combination is superior to the components on all response variables.

MULTIPLE-DOSE COMBINATIONS

A multilevel factorial design is a natural choice for studying multiple-dose combinations of two drugs simultaneously in one trial. A factorial design layout is illustrated as follows:

$$\begin{array}{ccccccc} A_0 B_0 & A_0 B_1 & A_0 B_2 & \dots & A_0 B_s \\ A_1 B_0 & A_1 B_1 & A_1 B_2 & \dots & A_1 B_s \\ A_2 B_0 & A_2 B_1 & A_2 B_2 & \dots & A_2 B_s \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_r B_0 & A_r B_1 & A_r B_2 & \dots & A_r B_s \end{array}$$

Here in the $(r+1) \times (s+1)$ factorial layout, $A_i B_j$ labels the combination of the i th dose of Drug A and the j th dose of Drug B, and A_0 and B_0 are the placebos of Drug A and Drug B, respectively. The subjects enrolled are randomly allocated to the cells. As articulated in Refs. [2] and [3], such a factorial design trial carries multiple objectives. For one, the study needs to provide confirmatory evidence for the assertion that the combination of the drugs is more beneficial than each component drug alone. Another objective is to obtain reliable and useful dose-effect information for writing instruction for use in drug labeling.

The definition for superiority of the combination drug over the component drugs in studying multiple-dose combination is not obvious. Hung et al.^[2] presented two distinct concepts of global superiority. In a weak sense of global superiority, the average of all the nonzero dose combinations is superior to both of the averages of the individual component doses. In a strong sense, there exist some nonzero dose combinations that are superior to the respective components in the dose region under study. The weak sense is less desirable because by showing it there is no guarantee that there is a single-dose combination superior to its components.

To demonstrate the strong sense of global superiority, the classical analysis of variance (ANOVA) using the additive model (i.e., no drug by drug interaction terms) can be a very powerful method. The additive model prescribes that the effect of each nonzero dose combination is the sum of the effects of its components. If the true effect of a dose combination is less than the sum of the component effects, the additive model will overestimate the effect of the combination. For such dose combinations, the probability of falsely asserting superiority over the components will be greater than it should. Thus, the assumption of no drug interaction is essential to the validity of the ANOVA using the additive model. The classical global F test for testing additivity does not guarantee that nonadditivity is always detectable at each dose combination. The patterns depicted by the estimates of the drug by drug interactions for the dose combinations may be more informative for detecting the possible presence of nonadditivity at the local level.

When the validity of the ANOVA is in doubt, some alternative useful methods for testing the strong-sense global superiority are α -level AVE and MAX tests developed by Hung et al.^[15] These global tests are based on the estimate of the least gain in effect resulting from use of each combination relative to its component. The AVE

test is an average of the min test statistics applied to the nonzero dose combinations under study whereas the MAX test takes the largest min test statistic. The α -level AVE test imposes a weak control on the familywise Type I error probability; that is, the probability of the Type I error associated with the global null hypothesis that none of the dose combinations under study are superior to their respective components is maintained at α . The α -level MAX test controls at the maximum probability of the familywise Type I error incurring in the pairwise comparisons between all nonzero dose combinations and their respective components (i.e., strong control). These tests have been extended to the incomplete factorial designs where some nonzero dose combinations are not studied and to the unbalanced designs where the cell sample sizes are not equal (see Refs. [16] and [17]).

The power functions of both the α -level AVE test and the α -level MAX test contain the nuisance parameters, each quantifying the difference of the respective component doses. The most conservative but perhaps impractical strategy of sample size planning for a multifactorial trial is one that assumes an identical effect for each dose level of both component drugs. A practical strategy should rely on Monte Carlo simulation studies based on a number of layouts for possible expected effect sizes at all local cells.

A more informative approach is to identify a dose combination that is superior to the highest dose of Drug A and the highest dose of Drug B under study. The most natural dose combination for testing is the combination of the highest doses of Drug A and Drug B. In general, the α -level min test as described above is the test method to use. Some other considerations may lead to testing some lower dose combinations against the highest dose of each single drug alone. Strong control of type I error rate associated with testing such multiple dose combinations is generally required to assert which dose combination(s) is superior to the highest dose of either drug.

Dose-effect relationship can be studied using a biological model or an empirical statistical model, such as a quadratic polynomial model via response surface or regression analyses. The surface can provide valuable information as to whether the effect increases as a function of dose, whether the effect increases start to level off at some dose combination, or whether there may be drug by drug interactions. If the response surface model fits the dose-effect data well, then one can construct confidence surfaces to identify desirable dose combinations that have greater effects than the individual components with a desired level of confidence.^[18] Other useful analysis approaches using response surface method can be found in Ref. [19].

THERAPEUTIC SYNERGISM

The concept of therapeutic synergism proposed by Vennitti et al.^[20] and discussed further by Mantel^[21] can be

useful in characterizing the activities of a combination drug in factorial design clinical trials. This concept has been extensively employed in preclinical research on combination cancer therapies. Literature is abundant, such as an early review paper by Carter and Wampler,^[23] for this type of applications. Carter et al.^[22] stated that therapeutic synergism is said to exist if a combination treatment provides improved therapy as compared to the best-component single-drug therapy. As stipulated in Ref. [23], with response surface methodology, the location of optimum dose combination that gives the maximal effect in the study dose range can be estimated and a confidence region around the estimated dose combination can be calculated. If the confidence region excludes a zero dose of each drug in the combination, then therapeutic synergism is established. This method can be used to identify nonzero dose combinations as potential candidates for an initial therapy.

CONCLUSION

Combination treatment trials present many interesting yet difficult statistical design and inference problems. For testing a single dose combination versus its components, it is a typical intersection-union testing problem that involves the difference between the effects of the two components as the nuisance or ancillary parameter. Without a complete knowledge of the value of this parameter, the α -level min test described above is the only solution to the testing problem. From a design perspective, a good sample size plan should ensure that this test has sufficient statistical power under the worst scenario that the effects of the two component drugs are equal.

The multi-level factorial design provides a viable avenue for studying multiple dose combinations with multiple objectives entertained above. Depending on the study objective, sample size planning requires careful attention. If the objective is to assert that at least one dose combination is superior to its respective components, then the sample sizes need to be carefully allocated to all dose combination cells under testing. If, however, the objective is to identify which dose combination(s) is superior to the highest dose of each drug, then the tested dose combinations need to be loaded with the major portion of the total sample size. Multiple comparison adjustments need to be incorporated in sample size planning. Consequently, the dose-response exploration which involves other dose combinations will be based on much smaller sample sizes.

REFERENCES

1. Laska, E.M.; Meisner, M.J.; Tang, D.I. Classification of the effectiveness of combination treatments. *Stat. Med.* **1997**, *16*, 2211–2228.

2. Hung, H.M.J.; Ng, T.H.; Chi, G.Y.H.; Lipicky, R.J. Response surface and factorial designs for combination antihypertensive drugs. *Drug Inf. J.* **1990**, *24*, 371–378.
3. Hung, H.M.J.; Chi, G.Y.H.; Lipicky, R.J. On some statistical methods for analysis of combination drug studies. *Commun. Stat. Theory Methods* **1994**, *A23* (2), 361–376.
4. Lehmann, E.L. Testing multiparameter hypotheses. *Ann. Math. Stat.* **1952**, *23*, 541–552.
5. Berger, R.L. Multiparameter hypothesis testing and acceptance sampling. *Technometrics* **1982**, *24*, 295–300.
6. Laska, E.M.; Meisner, M.J. Testing whether an identified treatment is best. *Biometrics* **1989**, *45*, 1139–1151.
7. Snapinn, S.M. Evaluating the efficacy of a combination therapy. *Stat. Med.* **1987**, *6*, 657–665.
8. Patel, H.I. Comparison of treatments in a combination therapy trial. *J. Biopharm. Stat.* **1991**, *1* (2), 171–183.
9. Wang, S.J.; Hung, H.M.J. Large sample tests for binary outcomes in fixed-dose combination drug studies. *Biometrics* **1997**, *53*, 498–503.
10. Hung, H.M.J. Two-stage tests for studying monotherapy and combination therapy in two-by-two factorial trials. *Stat. Med.* **1993**, *12*, 645–660.
11. Gibson, J.M.; Overall, J.E. The superiority of a drug combination over each of its components. *Stat. Med.* **1989**, *8*, 1479–1484.
12. Sarkar, S.K.; Snapinn, S.M.; Wang, W. On improving the min test for the analysis of combination drug trials. *Proceedings of the Biopharmaceutical Section of the American Statistical Association*; **1993**, 212–217.
13. Snapinn, S.M.; Sarkar, S.K. A Note on Assessing the Superiority of a Combination Drug with a Specific Alternative. *Proceedings of the Biopharmaceutical Section of the American Statistical Association*; **1995**, 20–25.
14. Laska, E.M.; Tang, D.I.; Meisner, M.J. Testing hypotheses about an identified treatment when there are multiple endpoints. *J. Am. Stat. Assoc.* **1992**, *87*, 825–831.
15. Hung, H.M.J.; Chi, G.Y.H.; Lipicky, R.J. Testing for the existence of a desirable dose combination. *Biometrics* **1993**, *49*, 85–94.
16. Hung, H.M.J. Global tests for combination drug studies in factorial trials. *Stat. Med.* **1996**, *15*, 233–247.
17. Hung, H.M.J. Evaluation of a combination drug with multiple doses in unbalanced factorial design clinical trials. *Stat. Med.* **2000**, *19*, 2079–2087.
18. Hung, H.M.J. On identifying a positive dose-response surface for combination agents. *Stat. Med.* **1992**, *11*, 703–711.
19. Phillips, J.A.; Cairns, V.; Koch, G.G. The analysis of a multiple-dose, combination-drug clinical trial using response surface methodology. *J. Biopharm. Stat.* **1992**, *2*, 49–67.
20. Venditti, J.M.; Humphreys, S.R.; Mantel, N.; Goldin, A. Combined treatment of advanced leukemia (L1210) in mice with amethopterin and 6-mercaptopurine. *J. Natl. Cancer Inst.* **1956**, *17*, 631–658.
21. Mantel, N. Therapeutic synergism. *Cancer Chemother. Rep. Part 2* **1974**, *4*, 147–149.
22. Carter, W.H.; Wampler, G.L.; Stablein, D.M.; Campbell, E.D. Drug activity and therapeutic synergism in cancer treatment. *Cancer Res.* **1982**, *42*, 2963–2971.
23. Carter, W.H.; Wampler, G.L. Review of the application of response surface methodology in the combination therapy of cancer. *Cancer Treat. Rep.* **1986**, *70* (1), 133–140.

Clinical Trial: Failure-Time Model

Chin-Shang Li

St. Jude Children's Research Hospital, Memphis, Tennessee, U.S.A.

Abstract

In many clinical follow-up studies, the primary endpoint is the time to some event of interest, e.g., time from initial treatment to local recurrence of tumor. This type of event is generally referred to as a “failure.” This entry provides a brief overview of the model and discusses estimation models.

INTRODUCTION

In many clinical follow-up studies, the primary endpoint is the time to some event of interest, e.g., time from initial treatment to local recurrence of tumor. Such events are generically called *failures*. The interval of interest, referred to as *failure time*, *survival time*, or *event time*, is often subject to right censoring, a form of incomplete observation in which the event of interest has not yet occurred. Other forms of censoring include left censoring and interval censoring. In left censoring, some observed times are greater than or equal to the actual failure times. Interval censoring means that some failures are known only to have occurred within some time interval.

One can describe the probability distribution of the failure time of an individual from a homogeneous population (i.e., no explanatory variables) in three ways: the survival function, the hazard function, and the probability density function. The survival and hazard functions are of particular interest in summarizing failure time data. Parametric models for failure time for a homogeneous population include the Weibull, gamma, log-normal, and log-logistic distributions. Nonparametric methods of estimating the survival function include Kaplan–Meier (product–limit) and Nelson–Aalen estimators. However, in follow-up trials, the failure times of individuals usually depend on characteristics that are also referred to as explanatory variables, covariates, or predictor variables. The explanatory variables may include demographic variables such as age, gender, and race, physiological variables such as weight, height, and blood pressure, and behavioral factors such as smoking history and dietary habits. The proportional hazards, accelerated failure time, and—proportional odds models can be used to model the hazard function, to determine which explanatory variables affect the hazard function, and to estimate the hazard function for an individual.

OVERVIEW

Let $T \geq 0$ be a random variable representing the failure time of an individual in a homogeneous population. The survival function is defined to be the probability that the failure time T is greater than or equal to time t , i.e., $S(t) = \Pr(T \geq t)$. The hazard function $h(t)$, expressing the risk or hazard of failure at time t given that failure has not occurred before time t , is expressed as

$$h(t) = \lim_{\Delta t \rightarrow 0^+} \frac{\Pr(T < t + \Delta t \mid t \leq T)}{\Delta t} = \frac{f(t)}{S(t)}$$

where $f(t) = \lim_{\Delta t \rightarrow 0^+} \Pr(t \leq T < t + \Delta t) = \Delta t = -dS(t)/dt$ is the probability density function of T . The hazard function is also called the *hazard rate*, the *instantaneous death rate*, the *intensity rate*, or the *force of mortality*. The survival function may be written as $S(t) = \exp(-H(t))$, where $H(t) = \int_0^t h(u)du$ is called an *integrated* or *cumulative hazard function*.

If the failure time T is assumed to have a Weibull distribution with scale parameter $\lambda > 0$ and shape parameter $\gamma > 0$, the survival and hazard functions are $S(t) = \exp(-\lambda t^\gamma)$ and $h(t) = \lambda \gamma t^{\gamma-1}$, which is monotonic decreasing ($\gamma < 1$) or increasing ($\gamma > 1$). The Weibull distribution has a variety of forms, depending on γ , and has simple formulas for survival and hazard functions, so it is widely used in the parametric analysis of failure time data. In general, gamma and log-normal distributions are less convenient to compute, but are still frequently used. The Weibull hazard function is a monotonic function of time; however, the log-logistic hazard function $h(t) = \exp(\theta)\kappa t^{\kappa-1}/(1 + \exp(\theta)t^\kappa)$ is monotonic decreasing if $0 < \kappa \leq 1$, and has a single mode if $\kappa > 1$. Furthermore, the log-logistic distribution has a simple algebraic expression for the survival function $S(t) = (1 + \exp(\theta)t^\kappa)^{-1}$, and it is therefore an alternative to the Weibull distribution in situations where the hazard function changes direction. It is also more convenient than the log-normal distribution for handling censored data.

Let $t_{(1)} < \dots < t_{(m)}$ be m ordered failure times with $t_{(m+1)} = \infty$, and let r_j and d_j be the number of individuals at risk of failure and the number of observed failures at time $t_{(j)}$, respectively. The Kaplan–Meier^[1] (product–limit) estimator of the survival function is most commonly used and is expressed as

$$\hat{S}_{\text{KM}}(t) = \prod_{j=1}^k \left(1 - \frac{d_j}{r_j} \right) \quad (1)$$

for $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, \dots, m$, with $\hat{S}_{\text{KM}}(t) = 1$ for $t < t_{(1)}$. Note that $\hat{S}_{\text{KM}}(t)$ is, strictly speaking, undefined for t greater than the largest observation that is a censored time; $\hat{S}_{\text{KM}}(t) = 0$ for t greater than the largest observation that is the failure time $t_{(m)}$, because $r_m = d_m$. In the absence of censoring, the Kaplan–Meier estimator is an empirical survival function, and therefore is a generalization of the empirical survival function to accommodate censored observations.

The Nelson–Aalen estimator,^[2,3] also called Altshuler's estimator,^[4] is an alternative estimator of the survival function, given by

$$\hat{S}_{\text{NA}}(t) = \prod_{j=1}^k \exp \left(-\frac{d_j}{r_j} \right) \quad (2)$$

which is obtained from an estimator of the cumulative hazard function $H(t) = -\log S(t)$, denoted by $\hat{H}_{\text{NA}}(t) = \sum_{j=1}^k d_j/r_j$ for $t_{(k)} \leq t < t_{(k+1)}$. These two estimates are asymptotically equivalent, but the Kaplan–Meier estimate is always less than the Nelson–Aalen estimate, because $\exp(-d_j/r_j) \geq 1 - d_j/r_j$. Colosimo et al.^[5] studied small-sample properties of these estimators by performing Monte Carlo simulations. Their results showed the Nelson–Aalen estimator to be slightly superior in survival fraction estimation. However, for percentile estimation, the Kaplan–Meier estimator showed better performance for decreasing failure rates while the Nelson–Aalen estimator outperformed the Kaplan–Meier estimator for increasing failure rates.

The standard error of the Kaplan–Meier estimator can be obtained by Greenwood's formula.^[6] Greenwood's formula can underestimate the variance of the estimated survival function when it is close to zero or one. Instead, one can use a more conservative standard error proposed by Peto et al.,^[7] which tends to be larger than it should be. However, the Greenwood estimate is recommended for general use.

PROPORTIONAL HAZARDS MODELS

Assume that $\mathbf{x} = (x_1, \dots, x_p)$ is a vector of covariates for an individual and that $h(t; \mathbf{x})$ is the hazard rate at time t for the individual with covariates $\mathbf{x}$. Cox^[8] assumed that the

covariates act multiplicatively on the hazard function and proposed the proportional hazards model (also called the Cox model) as follows:

$$h(t; \mathbf{x}) = h(t) \exp(\mathbf{x}\boldsymbol{\beta}) \quad (3)$$

where $h(t)$ is an arbitrary unspecified baseline hazard function at time t for an individual with covariates equal to zero, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is a vector of regression parameters for T denoting transposition. Or, in terms of survival function, we can express survival function of the individual with covariates $\mathbf{x}$ as

$$S(t; \mathbf{x}) = (S(t))^{\exp(\mathbf{x}\boldsymbol{\beta})} \quad (4)$$

where $S(t) = \exp(-\int_0^t h(u) du)$ is the survival function of an individual with covariates equal to 0. That is, in this model the survival function $S(t; \mathbf{x})$ of the individual with covariates $\mathbf{x}$ is the baseline survival function $S(t)$ to a power. The class of models is sometimes called the *Lehmann class*.^[9] Because no particular functional form of the baseline hazard function is assumed, the model is referred to as a *semiparametric regression model*. The proportional hazards regression model (4) can be reformulated as a linear transformation model

$$g(T) = -\mathbf{x}\boldsymbol{\beta} + \varepsilon \quad (5)$$

where $g(T) = \log(-\log(S(T)))$ is a strictly increasing function and $\varepsilon = \log(-\log(S(T; \mathbf{x})))$ has an extreme value (minimum) distribution. If the baseline hazard function is $h(t) = \lambda \gamma t^{\gamma-1}$, the proportional hazards regression model is a Weibull regression model, and the linear transformation model (5) can be written as $\log T = \boldsymbol{\beta}_0^* + \mathbf{x}\boldsymbol{\beta}^* + \sigma\varepsilon$, where $\sigma = \gamma^{-1}$, $\boldsymbol{\beta}_0^* = -\sigma \log \lambda$, and $\boldsymbol{\beta}^* = -\sigma\boldsymbol{\beta}$. Note that if $\gamma = 1$, the model reduces to an exponential regression model.

Estimation

Assume that the observed data consist of the triplets $(t_i, \delta_i, \mathbf{x}_i)$, $i = 1, 2, \dots, n$, where t_i is the observation time of individual i (either the failure time or the censoring time); δ_i is the censoring indicator, which is 1 if t_i is the failure time and 0 otherwise; and $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ is the covariate vector of individual i . When the baseline hazard function is specified, the unknown parameters of interest can be estimated by using the method of maximum likelihood.

When the baseline hazard function is unspecified, one can first estimate the regression parameter vector $\boldsymbol{\beta}$ and then construct an estimate of the baseline survival function (or cumulative hazard function) based on the estimate of $\boldsymbol{\beta}$. Assume that failures occur at distinct times, $t_{(1)} < t_{(2)} < \dots < t_{(m)}$. Let $\mathbf{x}_{(i)}$ be the covariate values of the individual that fails at time $t_{(i)}$. Cox^[8,10] proposed the following *partial likelihood* for estimating $\boldsymbol{\beta}$:

$$L(\beta) = \prod_{j=1}^m \frac{\exp(\mathbf{x}_{(j)}\beta)}{\sum_{t \in R(t_{(j)})} \exp(\mathbf{x}_t\beta)} \quad (6)$$

where $R(t_{(j)})$, called the *risk set* at time $t_{(j)}$, is the set of all r_j individuals at risk just prior to $t_{(j)}$. Kalbfleisch and Prentice^[11] showed that in the absence of censoring and time-dependent covariates, the partial likelihood (6) is the marginal likelihood of the ranks.

To accommodate tied observations, Kalbfleisch and Prentice^[12] derived an exact expression for the partial likelihood that has a very complex form. The computation of this exact partial likelihood is very time consuming if there are a large number of ties at any failure time. To overcome the computational burden, a number of approximation methods have been proposed, e.g., Refs. [8,13,14]. For a rigorous proof of the asymptotic properties of the maximum partial likelihood estimator, see, e.g., Refs. [15,16].

Kalbfleisch and Prentice^[11] used an approach based on the method of nonparametric maximum likelihood to derive an estimate of the baseline hazard function at time t and gave the maximum likelihood estimate $\hat{\alpha}_j$ of $\alpha_j = S(t_{(j)})/S(t_{(j-1)})$ as a solution to the following equation:

$$\sum_{t \in \mathcal{D}(t_{(j)})} \frac{\exp(\mathbf{x}_t \hat{\beta})}{1 - \hat{\alpha}_j^{\exp(\mathbf{x}_t \hat{\beta})}} = \sum_{t \in R(t_{(j)})} \exp(\mathbf{x}_t \hat{\beta}) \quad (7)$$

where $\mathcal{D}(t_{(j)})$ is the set of all d_j individuals failing at time $t_{(j)}$. If only a single failure occurs at $t_{(j)}$ (i.e., $d_j = 1$), the solution $\hat{\alpha}_j$ to Eq. 7 can be solved, and, hence, the Kalbfleisch–Prentice estimate of the baseline survival function is

$$\hat{S}_{\text{KP}}(t) = \prod_{j: t_{(j)} \leq t} \left(\frac{1 - \exp(\mathbf{x}_{(j)} \hat{\beta})}{\sum_{t \in R(t_{(j)})} \exp(\mathbf{x}_t \hat{\beta})} \right)^{\exp(-\mathbf{x}_{(j)} \hat{\beta})} \quad (8)$$

If $d_j > 1$, one can have $\hat{S}_{\text{KP}}(t)$ by numerically solving Eq. 7. As a result, the estimated survival function for an individual with covariates $\mathbf{x}$ is $\hat{S}_{\text{KP}}(t, \mathbf{x}) = \hat{S}_{\text{KP}}(t)^{\exp(\mathbf{x} \hat{\beta})}$. Alternatively, to avoid the laborious computation when solving Eq. 7, one can substitute $\hat{\alpha}_j^{\exp(\mathbf{x}_t \hat{\beta})} \approx 1 + \exp(\mathbf{x}_t \hat{\beta}) \log \tilde{\alpha}_j$ into Eq. 7 to obtain $\tilde{\alpha}_j$. With $\tilde{\alpha}_j$'s, one can obtain the estimate of the baseline survival function referred to as the Nelson–Aalen estimate or the Breslow estimate. Note that when there are no covariates, one can see from Eq. 7 that $\hat{S}_{\text{KP}}(t)$ is the Kaplan–Meier estimate as shown in Eq. 1.

ACCELERATED FAILURE-TIME MODELS

In the proportional hazards regression model, the effect of covariates is intuitively assumed to act multiplicatively

on the hazard function, but the model does not propose a direct relationship between failure time and covariates. In contrast, the accelerated failure time (accelerated life) model^[12,17] has an intuitive physical interpretation in which the effect of covariates $\mathbf{x}$ is assumed to act multiplicatively on the failure time T or additively on the log failure time, log T , expressed as:

$$\log T = \mathbf{x}\beta + \epsilon \quad (9)$$

with error density $f(e)$. The accelerated failure time model with unspecified error distribution is called a *semiparametric accelerated failure model* and can be considered a semiparametric alternative to the proportional hazards regression model. Let $h(t^*)$ be the hazard function of $T^* = \exp(\epsilon)$. The hazard function of T can be written in terms of this baseline hazard function $h(\cdot)$ as

$$h(t; \mathbf{x}) = h(t \exp(-\mathbf{x}\beta)) \exp(-\mathbf{x}\beta) \quad (10)$$

where $\exp(\mathbf{x}\beta)$ is referred to as the *accelerated factor*. The survival function of T is $S(t; \mathbf{x}) = S(t \exp(-\mathbf{x}\beta))$. In this model, the role of covariates is to accelerate or decelerate the time to failure. If $h(t) = \lambda \gamma t^{\gamma-1}$, $h(t; \mathbf{x}) = \lambda \exp(-\gamma \mathbf{x}\beta) \gamma t^{\gamma-1}$, and the accelerated failure time model (10) is a Weibull model with scale parameter $\lambda \exp(-\gamma \mathbf{x}\beta)$ and shape parameter γ . The Weibull distribution possesses the accelerated failure time property and is the only probability distribution that has both the proportional hazard property and the accelerated failure time property. Note that if $\gamma = 1$, the accelerated failure time model (10) reduces to an exponential model with scale parameter $\lambda \exp(-\mathbf{x}\beta)$.

If the baseline hazard function is $h(t) = \exp(\theta) \kappa t^{\kappa-1} = (1 + \exp(\theta) t^\kappa)$ for $\kappa > 0$, the hazard function of the individual with covariates $\mathbf{x}$ is $h(t; \mathbf{x}) = \exp(\theta - \kappa \mathbf{x}\beta) \kappa t^{\kappa-1} / (1 + \exp(\theta - \kappa \mathbf{x}\beta) t^\kappa)$, and the accelerated failure time model (10) is a log-logistic model with parameters $\theta - \kappa \mathbf{x}\beta$ and κ . Thus, the log-logistic distribution has the accelerated failure time property. Other distributions that have the accelerated failure time property include the log-normal, gamma, and inverse Gaussian distributions.

Estimation

A number of semiparametric estimators have been proposed for use in estimating the parameters of interest and handling the presence of censored observations when the error density is completely unspecified. For example, Miller^[18] provided the Kaplan–Meier least-squares estimator obtained by minimizing the weighted sum of squares of the residuals, with weights derived from the Kaplan–Meier estimator of the error distribution based on the residuals. Buckley and James^[19] proposed a modification of the least-squares estimator in which censored observations are replaced with an estimate of their expectation, but they gave no mathematical justification and did not provide

a reliable computational method for implementation. Koul et al.^[20] proposed an inverse censoring probability weighted modification of the least-squares estimator. Lai and Ying^[21] rigorously studied the asymptotic properties of the Buckley–James estimator.^[19] Louis^[22] proposed the log-rank test-based estimator in the two-sample problem and studied the asymptotic properties of the estimator, while Wei and Gail^[23] studied corresponding results for a broader class of censored data rank tests. Jones^[24] proposed a general class of nonparametric test statistics, including both linear and nonlinear rank tests for the model, which are inverted into equations for estimating the regression parameters β .

Prentice^[25] proposed rank estimators based on the *weighted log-rank* statistics. The asymptotic properties of the rank estimators were studied by, e.g., Tsiatis^[26] and Ying.^[27] The weighted log-rank estimating functions for β are expressed as:

$$U_{\phi}(\beta) = \sum_{i=1}^n \delta_i \phi(\beta; e_i(\beta)) (\mathbf{x}_i - \bar{\mathbf{x}}(\beta; e_i(\beta))) \quad (11)$$

where ϕ is a possibly data-dependent weight function and $e_i(\beta) = \log t_i - \mathbf{x}_i \beta$, $\bar{\mathbf{x}}(\beta, u) = S^{(1)}(\beta; u) / S^{(0)}(\beta; u)$ for $S^{(0)}(\beta; u) = n^{-1} \sum_{i=1}^n I(e_i(\beta) \geq u)$ and $S^{(1)}(\beta; u) = n^{-1} \sum_{i=1}^n \mathbf{x}_i I(e_i(\beta) \geq u)$, with $I(\cdot)$ being the indicator function. If $\phi(\beta; e_i(\beta)) = 1$, then $U_{\phi}(\beta)$ is the log-rank^[28] statistic. If $\phi(\beta; u) = \prod_{j: e_j(\beta) \leq u} n S^{(0)}(\beta; e_j(\beta)) / (n S^{(0)}(\beta; e_j(\beta)) + 1)$, then $U_{\phi}(\beta)$ is the generalized Wilcoxon statistic. When $\phi(\beta; u) = S^{(0)}(\beta; u)$, $U_{\phi}(\beta)$ is the Gehan^[29] statistic. $U_{\phi}(\beta)$ is a p -dimensional step function of β that need not be monotonic, so in general, no exact solutions to the system of equations $U_{\phi}(\beta) = 0$ can be obtained. Wei et al.^[30] suggested choosing $\hat{\beta}$ that minimizes $\|U_{\phi}(\beta)\|$, where $\|\cdot\|$ is the Euclidean norm. Fygenon and Ritov^[31] studied the monotonicity properties of the weighted log-rank estimating function $U_{\phi}(\beta)$ in Eq. 11 and characterized the class of weight functions for which the weighted log-rank estimating function 11 is a monotonic function of each component of β . They noted that the Gehan estimating function is in this class, but the log-rank and generalized Wilcoxon estimations are not.

The theoretical results of the semiparametric accelerated failure time models have been well developed, but the semiparametric methods have not been widely used in practice, because no efficient and reliable computational methods are available to address the problem of minimizing $\|U_{\phi}(\beta)\|$, which is discontinuous and may have multiple local minima. Because the problem of minimizing $\|U_{\phi}(\beta)\|$ cannot be solved by the conventional optimization algorithms, Lin and Geyer^[32] applied the simulated annealing method^[33] to calculate $\hat{\beta}_{\phi}$ when the dimension of β is too large to permit a direct grid search; however, as Jin et al.^[34] pointed out, their algorithm is not guaranteed to yield the true minimum. To solve the numerical problem, Jin et al.^[34] provided a simple and reliable method in which they introduced a

class of monotonic estimating functions to approximate the weighted log-rank estimating function $U_{\phi}(\beta)$ in Eq. 11 by

$$\begin{aligned} \tilde{U}_{\phi}(\beta; \hat{\beta}) &= \sum_{i=1}^n \delta_i \psi(\hat{\beta}, e_i(\beta) + \mathbf{x}_i(\beta - \hat{\beta})) S^{(0)}(\beta; e_i(\beta)) \\ &\quad \times (\mathbf{x}_i - \bar{\mathbf{x}}(\beta; e_i(\beta))) \\ &= n^{-1} \sum_{i=1}^n \sum_{j=1}^n \psi(\hat{\beta}; e_i(\hat{\beta})) \delta_i (\mathbf{x}_i - \mathbf{x}_j) I(e_i(\beta) \leq e_j(\beta)) \end{aligned}$$

where $\psi(b; u) = \phi(b; u) / S^{(0)}(b; u)$ and $\hat{\beta}$ is a preliminary consistent estimator of β , such as the Gehan estimator. $\tilde{U}_{\phi}(\beta; \hat{\beta})$ is monotonic in each component of β and is the gradient of

$$L_{\phi}(\beta; \hat{\beta}) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n \psi(\hat{\beta}; e_i(\hat{\beta})) \delta_i |e_i(\beta) - e_j(\beta)| I(e_i(\beta) - e_j(\beta) < 0) \quad (12)$$

Thus, $\hat{\beta}_{\phi}$ can be found by the iterative algorithm: $\hat{\beta}_{(0)} = \hat{\beta}$, $\hat{\beta}_{(k)} = \arg \min_{\beta} L_{\phi}(\beta; \hat{\beta}_{(k-1)})$ for $k \geq 1$, in which the minimization of $L_{\phi}(\beta; \hat{\beta}_{(k-1)})$ can be accomplished by linear programming. The resulting estimator is consistent, although the original estimating equation may contain inconsistent roots, and is asymptotically normal.^[34]

PROPORTIONAL ODDS MODELS

Bennett^[35] generalized the proportional odds model to the survival analysis context and fit the model to the data set from the Veterans Administration lung cancer trial.^[36] The model assumes that the covariates $\mathbf{x}$ have a multiplicative effect on the odds against survival beyond time t , expressed as

$$\frac{1 - S(t; \mathbf{x})}{S(t; \mathbf{x})} = \frac{1 - S(t)}{S(t)} \exp(\mathbf{x}\beta) \quad (13)$$

where $(1 - S(t; \mathbf{x})) / S(t; \mathbf{x})$ is the odds of the occurrence of an event of interest by time t for the individual with covariates $\mathbf{x}$, and $R(t) = (1 - S(t)) / S(t)$ is an unknown baseline odds for the occurrence of the event of interest by time t for an individual with covariates equal to 0. Model (13) is called a *semiparametric proportional odds model*, because the form of the baseline odds $R(t)$ is unspecified.

From model (13), the hazard function of the individual with covariates $\mathbf{x}$ can be expressed as

$$h(t; \mathbf{x}) = \frac{1}{R(t) + \exp(-\mathbf{x}\beta)} \frac{dR(t)}{dt} \quad (14)$$

In terms of hazard ratio, the proportional odds regression model (13) can be written as $h(t; \mathbf{x}) / h(t) = (1 + (\exp(-\mathbf{x}\beta) - 1))$

$S(t)^{-1}$. It is structurally similar to the Cox proportional hazards model and may be used in similar situations. However, whereas the hazard ratio is constant in the Cox model, it converges monotonically from $\exp(\mathbf{x}\boldsymbol{\beta})$ to unity with time in the proportional odds model. This feature can be more useful than the constant hazard ratio when initial effects, such as differences in the stage of disease or in treatment, disappear with time. It may also be appropriate to demonstrate an effective cure, when the mortality rate of the disease group approaches that of a control population.^[35] One can reformulate the proportional odds regression model (13) as a linear transformation model

$$g(T) = -\mathbf{x}\boldsymbol{\beta} + \epsilon \quad (15)$$

where $g(T) = \log R(t)$ is a strictly increasing function and $\epsilon = \log((1-S(T;\mathbf{x}))/S(T;\mathbf{x}))$ has the standard logistic distribution. The proportional odds model can be considered to be a proportional hazards model with unobservable heterogeneity.^[37] If the baseline survival function is the survival function of a log-logistic distribution with unknown parameters θ and κ , the linear transformation model (15) is a log-logistic regression model expressed as $\log T = \beta_0^* + \mathbf{x}\boldsymbol{\beta}^* + \epsilon^*$, where $\boldsymbol{\beta}_0^* = \theta/\kappa$, $\boldsymbol{\beta}^* = -\boldsymbol{\beta}/\kappa$, and $\epsilon^* = \epsilon/\kappa$. The log-logistic distribution has the proportional odds property and is the only one distribution that possesses both the accelerated failure time and proportional odds properties.

Estimation

Methods of estimating $\boldsymbol{\beta}$ in model (13) have been proposed by a number of authors, including, e.g., Bennett,^[35] who used a *profile likelihood* approach; Pettitt,^[38] who applied a rank-based method; Cheng et al.,^[39] who used an estimating equation approach; and Rossini and Tsiatis,^[40] Huang and Rossini,^[41] and Shen,^[42] who used a *sieve maximum likelihood* approach for current status data, interval-censored data, and right-censored and interval-censored data, respectively.

In general, the approach of Cheng et al.^[39] requires nonparametric estimation of the conditional censoring distribution, given the covariates. Murphy et al.^[37] showed that the estimator based on the profile likelihood approach of Bennett^[35] is asymptotically efficient. The profile likelihood approach requires simultaneous estimation of the regression parameters and the unknown baseline odds function $R(t)$. The computation of the profile likelihood is complex.

Yang and Prentice^[45] derived three classes of estimators to avoid the complex calculation of the profile likelihood while retaining good efficiency properties and the strong assumptions about the censoring patterns or distributions of covariate, and to consider estimation problem in the general proportional odds regression model with censored data, rather than a two-sample case considered by

Dabrowska and Doksum^[43] and Wu.^[44] These estimators, including *pseudomaximum likelihood estimators*, *martingale residual-based estimators*, and *minimum distance estimators*, were derived via some *weighted empirical odds functions*. These estimators are computed easily and are asymptotically normally distributed, and the asymptotic covariance was estimated reliably in closed form. To describe the three classes of estimators, define $x_{i0} = 1$, $i = 1, \dots, n$. For $0 \leq t < \infty$, $j = 0, 1, \dots, p$, and fixed $\mathbf{b} = (b_1, \dots, b_p)^T$, we define the following *weighted empirical processes*: $\hat{W}_{nj}(t) = \sum_{i=1}^n \delta_i x_{ij} I(t_i \leq t)$, $\hat{W}_{nj}(t; \mathbf{b}) = \sum_{i=1}^n \delta_i x_{ij} \exp(-\mathbf{x}_i \mathbf{b}) I(t_i \leq t)$, and $\hat{K}_{nj}(t) = \sum_{i=1}^n x_{ij} I(t_i \geq t)$. For $t \leq t_{nj} = \sup\{s: \hat{K}_{nj}(s) > 0\}$, $j = 0, 1, \dots, p$, we define $\hat{H}_{nj}(t) = \int_0^t \frac{1}{\hat{K}_{nj}(u)} d\hat{W}_{nj}(u)$, $\hat{S}_{nj}(t) = \prod_{u \leq t} (1 - \Delta \hat{H}_{nj}(u))$, and $\hat{H}_{nj}(t; \mathbf{b}) = \int_0^t \frac{1}{\hat{K}_{nj}(u)} \hat{W}_{nj}(du; \mathbf{b})$, where $\Delta g(u) = g(u) - g_{-}(u)$ for any function g and the subscript “ $-$ ” denotes the left-continuous version of a function. Note that $\hat{H}_{n0}(t)$ and $\hat{S}_{n0}(t)$ are the usual Nelson–Aalen and Kaplan–Meier estimators. By multiplying both sides of Eq. 14 by the denominator of its right-hand side, we obtain the following first-order linear integral equation:

$$R(t) = \int_0^t \sum_{i=1}^n a_i(u) (R(u) + \exp(-\mathbf{x}_i \mathbf{b})) dH(u; \mathbf{x}_i) \quad (16)$$

where the $a_i(u)$ s are any negative functions such that $\sum_i a_i(u) = 1$, and $H(u; \mathbf{x}_i)$ is the cumulative hazard function of individual i with covariates $\mathbf{x}_i$. As in Eq. 16 with $a_i(u) = x_{ij} S(u; \mathbf{x}_i) (1 - G_i(u)) / (\sum_{k=1}^n x_{kj} S(u; \mathbf{x}_k) (1 - G_k(u)))$, with $S(u; \mathbf{x}_i)$ and $G_i(u)$ being the survival and censoring functions of individual i with covariates $\mathbf{x}_i$, for $t \leq t_{nj}$ and $j = 0, 1, \dots, p$, we can define the following weighted empirical odds functions

$$\hat{R}_{nj}(t; \mathbf{b}) = \frac{1}{\hat{S}_{nj}(t)} \int_{u \leq t} \hat{S}_{nj-}(u) \hat{H}_{nj}(ds; \mathbf{b}) \quad (17)$$

The first class of estimators is derived on the basis of the score function of $\boldsymbol{\beta}$ obtained by differentiating the likelihood function with respect to $\boldsymbol{\beta}$ by assuming that $R(t)$ is known. By replacing $R(t)$ by $\hat{R}_{n0}(t; \mathbf{b})$ in the score function of $\boldsymbol{\beta}$, Yang and Prentice^[45] defined a pseudomaximum likelihood estimator of $\boldsymbol{\beta}$ to be the solution to the system of estimating equations

$$\sum_{i=1}^n \mathbf{x}_i \frac{\delta_i \exp(-\mathbf{x}_i \mathbf{b}) - \hat{R}_{n0}(t_i; \mathbf{b})}{\exp(-\mathbf{x}_i \mathbf{b}) + \hat{R}_{n0}(t_i; \mathbf{b})} = 0 \quad (18)$$

which are asymptotically unbiased in the sense that replacing $\hat{R}_{n0}(u; \mathbf{b})$ by $R(u)$ results in the system of unbiased estimating equations.

The second class of estimators is derived from the martingale residuals $M_i(t) = \delta_i I(t_i \leq t) - \int_0^t I(t_i \geq s) \frac{dR(s)}{R(s) + \exp(-\mathbf{x}_i \boldsymbol{\beta})}$, and a martingale residual-based estimator of $\boldsymbol{\beta}$ is defined as the solution to the system of estimating equations

$$\sum_{i=1}^n \int_0^\infty J_i(t; \mathbf{b}) \hat{M}_i(dt; \mathbf{b}) = 0 \quad (19)$$

where the J_i 's are p -dimensional data-dependent functions and $M_i(t; \mathbf{b})$ is defined by replacing $R(t)$ with $R_{n0}(t; \mathbf{b})$ in $M_i(t)$. Note that for the Cox model, $J_i(t; \mathbf{b}) = \mathbf{x}_i$ gives the maximum partial likelihood estimator, and $J_i(t; \mathbf{b}) = \mathbf{x}_i \hat{S}_{n0}(t; \mathbf{b})$ gives the Peto–Prentice^[25,46] version of generalized Wilcoxon test statistic.

The third class of estimators is obtained by minimizing some discrepancy measures between the $\hat{R}_{nj}(t; \mathbf{b})$ functions, $j = 1, \dots, p$, and $\hat{R}_{n0}(t; \mathbf{b})$, or between the $\hat{F}_{nj}(t; \mathbf{b})$ functions and $\hat{F}_{n0}(t; \mathbf{b})$, where $\hat{F}_{nk}(t; \mathbf{b}) = \hat{R}_{nk}(t; \mathbf{b}) / (1 + \hat{R}_{nk}(t; \mathbf{b}))$ are weighted empirical distribution functions, $k = 0, 1, \dots, p$. In particular, Yang and Prentice^[45] defined a minimum L_2 distance estimator of $\boldsymbol{\beta}$ as the minimizer of the following distance

$$\int_0^\infty \sum_{j=1}^p (\hat{F}_{nj}(t; \mathbf{b}) - \hat{F}_{n0}(t; \mathbf{b}))^2 \Phi_n(dt; \mathbf{b}) \quad (20)$$

with some integrating measure Φ_n possibly dependent on the data. Their numerical studies showed that the minimum L_2 distance estimators with properly chosen integrating measures had the overall best performance among the three classes of estimators and were comparable to the estimator of Cheng et al.^[39] and the maximum profile likelihood estimator.

OTHER ISSUES

Cox^[8] included time-dependent covariates in a proportional hazards model and gave an appropriate partial likelihood for estimating the regression parameters. For a detailed account of the construction of the partial likelihood, see, e.g., Kalbfleisch and Prentice.^[12] Time-dependent covariates included in the Cox model can be used to examine the assumption of proportional hazards in the Cox model. Ataman and De Stavola^[47] discussed a number of practical problems encountered in survival analysis with time-dependent covariates.

The accelerated failure time model with time-dependent covariates was introduced by Cox and Oakes.^[17] Robins and Tsiatis^[48] derived a class of semiparametric rank estimators for the parameters for this model; Lin and Ying^[49] studied the asymptotic properties of the estimators obtained from the log-rank estimating

function and applied their methodology to the Stanford heart transplant data.

Yang and Prentice^[45] reported that the semiparametric proportional odds model (13) can be extended to incorporate time-dependent covariates, and they briefly stated that their corresponding self-consistency integral equations for the regression parameters could still be solved, and that the weighted odds function approach was still available.

In survival analysis, it is common to see that estimated survival curves level off at a nonzero value after a certain time, even when many individuals are followed beyond that time. The data typically have heavy censoring at the end of the follow-up period. The population can be viewed as consisting of two groups of individuals: those who are not susceptible to the event of interest, and those who will be susceptible if they are followed long enough. Therefore standard methods of survival analysis are inappropriate for this type of censored data, and they should be adapted. A number of parametric and semiparametric cure (or mixture) models have been proposed, see, e.g., Refs.^[50–58]. The identifiability of cure models was discussed by Li et al.^[59]

CONCLUSION

The numerical problem of estimating the parameters for the semiparametric accelerated failure time model is challenging, but over the past decade, the model has received a lot of attention, and much progress on solving the numerical problem has been made, such that the model is considered a practical alternative to the Cox model. The accelerated failure time model proposes a direct relationship between failure time and covariates, and this natural type of regression relationship led Sir David Cox to remark that “accelerated life time models are in many ways more appealing” than the proportional hazards model “because of their quite direct physical interpretation.”^[60]

The semiparametric proportional odds model with the property of providing a converging hazard is used as an alternative to the Cox model, but further investigation, especially of the model with time-dependent covariates, is needed to make the model a more useful alternative to the Cox model in survival analysis.

Survival analysis topics not discussed herein were surveyed by Oakes,^[61] with special emphasis on methodology reported in Biometrika.

ACKNOWLEDGMENTS

This work was partially supported by Cancer Center Support Grant CA21765 from the National Institutes of Health and by the American Lebanese Syrian Associated Charities (ALSAC).

REFERENCES

1. Kaplan, E.L.; Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **1958**, *53*, 457–481.
2. Nelson, W. Theory and applications of hazard plotting for censored failure data. *Technometrics* **1972**, *14*, 945–965.
3. Aalen, O. Nonparametric inference for a family of counting processes. *Ann. Stat.* **1978**, *6*, 701–726.
4. Altshuler, B. Theory for the measurement of competing risk in animal experiments. *Math. Biosci.* **1970**, *6*, 1–11.
5. Colosimo, E.A.; Ferreira, F.F.; Oliveira, M.D.; Sousa, C.B. Empirical comparisons between Kaplan–Meier and Nelson–Aalen survival function estimators. *J. Stat. Comput. Simul.* **2002**, *72*, 299–308.
6. Greenwood, M. The Errors of Sampling of the Survivorship Tables. In *Reports on Public Health and Statistical Subjects*; HMSO: London, 1926, Vol. 33. Appendix 1.
7. Peto, R.; Pike, M.C.; Armitage, P.; Breslow, N.E.; Cox, D.R.; Howard, S.V.; Mantel, N.; McPherson, K.; Peto, J.; Smith, P.G. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br. J. Cancer* **1977**, *35*, 1–39.
8. Cox, D.R. Regression models and life-tables (with discussion). *J. R. Stat. Soc., Ser. B* **1972**, *34*, 187–220.
9. Lehmann, E.L. The power of rank tests. *Ann. Math. Stat.* **1953**, *24*, 23–43.
10. Cox, D.R. Partial likelihood. *Biometrika* **1975**, *62*, 269–276.
11. Kalbfleisch, J.D.; Prentice, R.L. Marginal likelihoods based on Cox's regression and life model. *Biometrika* **1973**, *60*, 267–278.
12. Kalbfleisch, J.D.; Prentice, R.L. *The Statistical Analysis of Failure Time Data*, 2nd Ed.; John Wiley: New York, 2002; 218.
13. Breslow, N.E. Covariance analysis of censored survival data. *Biometric* **1974**, *30*, 89–99.
14. Efron, B. The efficiency of Cox's likelihood function for censored data. *J. Am. Stat. Assoc.* **1977**, *72*, 557–565.
15. Tsiatis, A.A. A large sample study of Cox's regression model. *Ann. Stat.* **1981**, *9*, 93–108.
16. Andersen, P.K.; Gill, R.D. Cox's regression model for counting processes: A large sample study. *Ann. Stat.* **1982**, *10*, 1100–1120.
17. Cox, D.R.; Oakes, D. *Analysis of Survival Data*; Chapman and Hall: London, 1984, 66.
18. Miller, R.G. Least squares regression with censored data. *Biometrika* **1976**, *63*, 449–464.
19. Buckley, J.; James, I. Linear regression with censored data. *Biometrika* **1979**, *66*, 429–436.
20. Koul, H.; Susarla, V.; Van Ryzin, J. Regression analysis with randomly right-censored data. *Ann. Stat.* **1981**, *9*, 1276–1288.
21. Lai, T.L.; Ying, Z. Large sample theory of a modified Buckley–James estimator for regression analysis with censored data. *Ann. Stat.* **1991**, *19*, 1370–1402.
22. Louis, T.A. Nonparametric analysis of an accelerated failure time model. *Biometrika* **1981**, *68*, 381–390.
23. Wei, L.J.; Gail, M.H. Nonparametric estimation for a scale-change with censored observations. *J. Am. Stat. Assoc.* **1983**, *78*, 382–388.
24. Jones, M.P. A class of semiparametric regression for the accelerated failure time model. *Biometrika* **1997**, *84*, 73–84.
25. Prentice, R.L. Linear rank tests with right censored data. *Biometrika* **1978**, *65*, 167–179.
26. Tsiatis, A.A. Estimating regression parameters using linear rank tests for censored data. *Ann. Stat.* **1990**, *18*, 354–372.
27. Ying, Z. A large sample study of rank estimation for censored regression data. *Ann. Stat.* **1993**, *21*, 76–99.
28. Mantel, N. Evaluation of survival data and two new rank order statistics arising in its considerations. *Cancer Chemother. Rep.* **1966**, *50*, 163–170.
29. Gehan, E.A. A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika* **1965**, *52*, 203–223.
30. Wei, L.J.; Ying, Z.; Lin, D.Y. Linear regression analysis of censored survival data based on rank tests. *Biometrika* **1990**, *77*, 845–851.
31. Fyngenson, M.; Ritov, Y. Monotone estimating equations for censored data. *Ann. Stat.* **1994**, *22*, 732–746.
32. Lin, D.Y.; Geyer, C.J. Computational methods for semiparametric linear regression with censored data. *J. Comput. Graph. Stat.* **1992**, *1*, 77–90.
33. Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P. Optimization by simulated annealing. *Science* **1983**, *220*, 671–680.
34. Jin, Z.; Lin, D.Y.; Wei, L.J.; Ying, Z. Rank-based inference for the accelerated failure time model. *Biometrika* **2003**, *90*, 341–353.
35. Bennett, S. Analysis of survival data by the proportional odds model. *Stat. Med.* **1983**, *2*, 273–277.
36. Prentice, R.L. Exponential survivals with censoring and explanatory variables. *Biometrika* **1973**, *60*, 279–288.
37. Murrphy, S.A.; Rossini, A.J.; van der Vaart, A.W. Maximum likelihood estimation in the proportional odds model. *J. Am. Stat. Assoc.* **1997**, *92*, 968–976.
38. Pettitt, A.N. Proportional odds models for survival data and estimates using ranks. *Appl. Stat.* **1984**, *33*, 169–175.
39. Cheng, S.C.; Wei, L.J.; Ying, Z. Analysis of transformation models with censored data. *Biometrika* **1995**, *82*, 835–845.
40. Rossini, A.J.; Tsiatis, A.A. A semiparametric proportional odds regression model for the analysis of current status data. *J. Am. Stat. Assoc.* **1996**, *91*, 713–721.
41. Huang, J.; Rossini, A.J. Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *J. Am. Stat. Assoc.* **1997**, *92*, 960–967.
42. Shen, X. Proportional odds regression and sieve maximum likelihood estimation. *Biometrika* **1998**, *85*, 165–177.
43. Dabrowska, D.M.; Doksum, K.A. Estimation and testing in a two-sample generalized odds-rate model. *J. Am. Stat. Assoc.* **1988**, *83*, 744–749.
44. Wu, C.O. Estimating the real parameter in a two-sample proportional odds model. *Ann. Stat.* **1995**, *23*, 376–395.
45. Yang, S.; Prentice, R.L. Semiparametric inference in the proportional odds regression model. *J. Am. Stat. Assoc.* **1999**, *94*, 125–136.
46. Peto, R.; Peto, J. Asymptotically efficient rank invariant test procedures (with discussion). *J. R. Stat. Soc., A* **1972**, *135*, 185–207.
47. Altman, D.G.; De Stavola, B.L. Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates. *Stat. Med.* **1994**, *13*, 301–341.

48. Robins, J.; Tsiatis, A.A. Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika* **1992**, *79*, 311–319.
49. Lin, D.Y.; Ying, Z. Semiparametric inference for the accelerated failure time model with time-dependent covariates. *J. Stat. Plan. Inference* **1995**, *44*, 47–63.
50. Farewell, V.T. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **1982**, *38*, 1041–1046.
51. Yamaguchi, K. Accelerated failure-time regression models with a regression model of surviving fraction: An application to the analysis of 'permanent employment' in Japan. *J. Am. Stat. Assoc.* **1992**, *87*, 284–292.
52. Kuk, A.Y.C.; Chen, C.H. A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **1992**, *79*, 531–541.
53. Taylor, J.M.G. Semi-parametric estimation in failure time mixture models. *Biometrics* **1995**, *51*, 899–907.
54. Chen, M.H.; Ibrahim, J.G.; Sinha, D. A new Bayesian model for survival data with a surviving fraction. *J. Am. Stat. Assoc.* **1999**, *94*, 909–919.
55. Sy, J.P.; Taylor, J.M.G. Estimation in a Cox proportional hazards cure model. *Biometrics* **2000**, *56*, 227–236.
56. Peng, Y.; Dear, K.B.G. A nonparametric mixture model for cure rate estimation. *Biometrics* **2000**, *56*, 237–243.
57. Li, C.S.; Taylor, J.M.G. Smoothing covariate effects in cure models. *Commun. Stat., Theory Methods* **2002**, *31*, 477–493.
58. Li, C.S.; Taylor, J.M.G. A semi-parametric accelerated failure time cure model. *Stat. Med.* **2002**, *21*, 3235–3247.
59. Li, C.S.; Taylor, J.M.G.; Sy, J.P. Identifiability of cure models. *Stat. Probab. Lett.* **2001**, *54*, 389–395.
60. Reid, N. A conversation with Sir David Cox. *Stat. Sci.* **1994**, *4*, 439–455.
61. Oakes, D. *Biometrika centenary: Survival analysis*. *Biometrika* **2001**, *88*, 99–142.

Clinical Trial: *N*-of-1 Design Analysis

Can Cui

Duke University, Durham, North Carolina, U.S.A.

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

The development of biosimilar products provides a more affordable alternative to general patient population, compared to the brand-name products. However, the availability of more and more biosimilar products in the marketplace also raises a critical issue, that is, whether these biosimilars can be used interchangeably with the brand-name products is unknown. To this end, a crossover design is necessarily employed. In this entry, we focus on the advantages of complete *N*-of-1 designs compared to the commonly used crossover designs. Specifically, the relative efficiency and power analysis are performed for the complete *N*-of-1 designs with two, three, and four dosing periods. As a result, complete *N*-of-1 designs are more efficient and require less sample size in clinical trials. Therefore, complete *N*-of-1 designs are superior than partial designs in general and can potentially work for the evaluation in rare diseases.

Keywords: Biosimilars; Clinical trials; Complete *N*-of-1 designs; Drug interchangeability; Power analysis; Rare diseases; Relative efficiency.

INTRODUCTION

Biological products as therapeutic agents are made from a variety of living cells or living organisms.^[1] The application of biologics ranges from treating disease and medical conditions to preventing and diagnosing diseases. In practice, due to the expensive cost to access to the original biologics, the alternative products (follow-on biologics or biosimilars) are often produced by biopharmaceutical and biotechnology companies. The popularity of biosimilars brings not only advantages but also several critical issues. For benefits, considering that patents of the early biological products will expire in the next few years, biosimilars with the reduction of the cost are more affordable to general patient population. However, biosimilar products need to be demonstrated to be equivalent to the originator product in terms of safety and efficacy before they can be released in the marketplace. According to the FDA, a biological product may be demonstrated to be biosimilar if data show that, among other things, the product is “highly similar” to an already-approved biological product.^[2] To this end, the relevant properties of the biosimilar products are compared with those of the original drugs. If a proposed biosimilar product has been demonstrated to be highly similar to the originator product, the proposed biosimilar product will be approved by the regulatory authority. An approved biosimilar product can be used as a substitute for the originator product for

treating patients with the diseases under study. In practice, as more and more biosimilar products become available in the marketplace, interesting questions have been raised.^[3] The questions are whether approved biosimilar products can be used interchangeably with the originator product and whether the approved biosimilar products can be used interchangeably with other approved biosimilar products.

As indicated in the Biologics Price Competition and Innovation (BPCI) Act passed by the US Congress in 2009, a proposed biosimilar product is considered interchangeable (1) if it is highly similar to the originator (reference) product and can be expected to produce the same clinical result as the reference product in any given patient, and (2) if the biological product is administered more than once to an individual, the risk in terms of safety or diminished efficacy of alternating or switching between the use of the biological product and the reference product is not greater than the risk of using the reference product without such alternation or switch. Thus, the assessment of drug interchangeability of biosimilar products includes the evaluation of (1) the risk in terms of safety and diminished efficacy of alternating or switching between the use of the biological product and the reference product, (2) the risk of using the reference product without such alternation or switch, and (3) the relative risk between switch/alternation and without switch/alternation. For this purpose, a crossover design is necessarily employed. In recent years, three types of hybrid parallel-crossover designs are commonly

used to address the drug interchangeability: the parallel plus 2×2 crossover design, the parallel plus 2×3 crossover design, and the parallel plus 2×4 crossover design. These designs evaluate the drug interchangeability of biosimilars to some extent, but with insufficient efficiency. On the contrary, the complete *N*-of-1 designs assess the drug interchangeability more efficiently than those common crossover designs, especially for the rare diseases.

Complete *N*-of-1 design with S periods is an $N^S \times S$ crossover design where N is the number of drugs. Under S dosing periods, the complete *N*-of-1 design gives all possible combinations of drugs in each of the S dosing periods, which results in N^S sequences. Compared to the commonly used crossover designs, the complete *N*-of-1 designs provide more information when conducting the clinical trials, which allow us to evaluate drug interchangeability. Due to their sensitivity, the complete *N*-of-1 designs potentially work for rare diseases. Considering the rareness of some specific diseases, the number of subjects that can enroll in the trial will be small. With limited participants under the study, the complete *N*-of-1 trial has more advantages to evaluate drug interchangeability than common designs. Actually, the popular designs mentioned above are motivated from a complete *N*-of-1 design with two, three, and four dosing periods, respectively.

Therefore, this article aims to demonstrate the superiority of complete *N*-of-1 designs compared to the popular designs with the corresponding dosing periods and to discuss the potential application of complete *N*-of-1 designs. In Section “Complete *N*-of-1 Design,” the complete *N*-of-1 designs with two, three, and four dosing periods will be introduced. In Section “Comparison between Complete *N*-of-1 Designs and Common Designs,” the comparison of complete *N*-of-1 designs will be discussed in terms of relative efficiency and sample size calculation (Fig. 1). Finally, brief conclusions will be marked according to the analysis results.

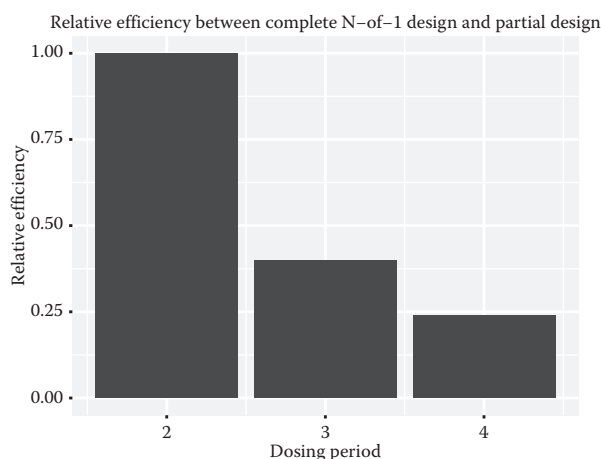


Fig. 1 Relative efficiency between complete *N*-of-1 design and partial design

COMPLETE N-OF-1 DESIGN

Complete *N*-of-1 Design with Two Periods

Complete *N*-of-1 design with two periods is a 4×2 crossover design with four sequences and two periods (see Table 1). It is also called Balaam's design, which is a popular design to evaluate the switching for biosimilars. With all possible arrangements of testing and reference drugs, the part of interchangeability, the switching, can be comprehensively assessed under this crossover design. For example, starting from the testing drug, the switch from T to T can be compared to the switch from T to R to evaluate the drug biosimilarity. When using the reference drug at first, the switch from R to R and R to T can also be compared by using group 2 and group 4. However, it cannot assess the alternation due to the limited number of periods in this study design, since alternation requires two switches for each subject.

Complete *N*-of-1 Design with Three Periods

Complete *N*-of-1 design with three periods is an 8×3 design which contains all the possible arrangements for the switching and alternation among testing drugs and reference drugs under three periods (see Table 2).

One of the popular designs to evaluate drug interchangeability is 2×3 dual crossover design, which is a partial design derived from the *N*-of-1 design (see Table 3). This design can evaluate both the switching and the alternation. To compare switching, the switch from T to R and the switch from R to T can be compared within the same cohort of subjects in group 1. Similarly, the switch from R to T and the switch from T to R can be compared within

Table 1 Complete *N*-of-1 design with two periods

Group	Period I	Period II
1	T	T
2	R	R
3	T	R
4	R	T

Table 2 Complete *N*-of-1 design with three periods

Group	Period I	Period II	Period III
1	R	R	R
2	R	T	R
3	T	T	R
4	T	R	R
5	R	R	T
6	R	T	T
7	T	R	T
8	T	T	T

Table 3 2×3 dual crossover design

Group	Period I	Period II	Period III
1	T	R	T
2	R	T	R

Table 4 2×4 replicated crossover design

Group	Period I	Period II	Period III	Period IV
1	T	R	T	R
2	R	T	R	T

Table 5 Complete *N*-of-1 design

Group	Period I	Period II	Period III	Period IV
1	R	R	R	R
2	R	T	R	R
3	T	T	R	R
4	T	R	R	R
5	R	R	T	R
6	R	T	T	T
7	T	R	T	R
8	T	T	T	T
9	R	R	R	T
10	R	R	T	T
11	R	T	R	T
12	R	T	T	R
13	T	R	R	T
14	T	R	T	T
15	T	T	R	T
16	T	T	T	R

group 2. What is more, the switch between T and R can be compared to a different group of subjects as well. However, this crossover design has the problem with the assessment of alternations within the same group of subjects. For example, the alternations of (T to R to T) and (R to T to R) can only be compared to a different group of patients.

Complete *N*-of-1 Design with Four Periods

The 2×4 replicated crossover design contains two sequences and four periods (see Table 4), which is suitable for evaluating both the switching and the alternation, especially the alternations within the same group of subjects. For instance, the alternations of (T to R to T) and (R to T to R) can be evaluated by comparing the drug effect from Period I to Period III to that from Period II to Period IV within the first group of patients.

The 2×4 replicated crossover design is not a complete *N*-of-1 design either. With four periods, qualified subjects will be randomly assigned to the 16 treatment sequences to evaluate the drug interchangeability (see Table 5).

Under the complete *N*-of-1 crossover design with four dosing periods, all possible switching and alternations can be assessed, and the results can be compared within the same group of patients and between different groups of patients.

COMPARISON BETWEEN COMPLETE *N*-OF-1 DESIGNS AND COMMON DESIGNS

Although the partial designs such as 2×3 dual design and 2×4 replicated design can assess drug interchangeability to some extent, complete *N*-of-1 designs are more suitable because they bring a broader framework for evaluation. For instance, outcomes under all possible switches and alternations can be utilized to demonstrate drug interchangeability for the complete *N*-of-1 design with four periods. Compared to the 2×4 replicated design, the complete *N*-of-1 design contains more information for testing and reference drugs, which provides the opportunity to comprehensively assess the drug biosimilarity and interchangeability.

As for the relative efficiency, the partial designs are less efficient than the complete *N*-of-1 design with a corresponding number of periods. To achieve the expected power for evaluating the biosimilarity/interchangeability under two common hypotheses, the partial design requires a larger sample size than the complete design. Because of the smaller sample size required in the complete *N*-of-1 design, conducting the drug trial under the complete design will save money, time, and be more efficient than the partial design.

The comparison of drug interchangeability between complete *N*-of-1 designs and commonly considered designs with the same periods will be discussed in detail. The evaluation between the two designs is based on the relative efficiency adjusting for the carryover effect. The sample size for testing different hypotheses under the two study designs are also calculated.

Relative Efficiency

As it can be seen, the commonly considered designs and complete *N*-of-1 trial designs discussed in Section “Complete *N*-of-1 Design” can be generally described as a $J \times K$ crossover design. Thus, in this section, the drug effect and the carryover effect will be assessed using the following statistical model under a general K -sequence and J -period crossover design comparing two drugs:

$$Y_{ijk} = \mu + G_k + S_{ik} + P_j + D_{d(j,k)} + C_{d(j-1,k)} + e_{ijk}$$

$$i = 1, 2, \dots, n_k; j = 1, 2, \dots, J; k = 1, 2, \dots, K; d = T \text{ or } R$$

where μ is the overall mean, G_k is the fixed k th sequence effect, S_{ik} is the random effect for the i th subject within the k th sequence with mean 0 and variance σ_s^2 , P_j is the

fixed effect for the j th period, $D_{d(j,k)}$ is the drug effect for the drug at the k th sequence in the j th period, $C_{d(j-1,k)}$ is the carryover effect, and e_{ijk} is the random error with mean 0 and variance σ_e^2 . Under the model, it is assumed that S_{ik} and e_{ijk} are mutually independent.

Complete N-of-1 Design with Four Periods vs. RRRR/RTRT Design

Taking the complete N-of-1 design with four periods as an example, the complete design will be compared with a recently popular 2×4 design, which is the RRRR/RTRT design. This partial design from the complete N-of-1 design is proposed and supported by many physicians for evaluating the switching between R and T and the relative risk of with/without alternation (see Table 6).

To compare the relative efficiency between the complete and partial designs, the statistics for testing biosimilarity/bioequivalence between biosimilar drug and reference drug under two study designs are derived.

Under the statistical model, the expected values of the sequence-by-period means for the RRRR/RTRT design are derived based on the above model, which contains for the first-order carryover effect (see Table 7), and the corresponding analysis of variance table is presented in Table 8.

Denote P as a parameter vector, referring to $(\mu, G_1, G_2, P_1, P_2, P_3, P_4, D_T, D_R, C_T, C_R)$, which contains all unknown parameters in the model. The aim is to construct a method of moments estimator in linear form of the observed cell means $\tilde{Y} = \beta' \bar{Y}$, where $\bar{Y}$ is the vector of observed cell means (Table 6).

$$\bar{Y} = (\bar{Y}_{.11}, \bar{Y}_{.21}, \bar{Y}_{.31}, \bar{Y}_{.41}, \bar{Y}_{.12}, \bar{Y}_{.22}, \bar{Y}_{.32}, \bar{Y}_{.42})', \bar{Y}_{.jk} = \frac{\sum_{i=1}^{n_k} Y_{ijk}}{n_k}$$

Then $E(\tilde{Y}) = E(\beta' \bar{Y}) = L'P$ which is the parameter of interest based on the linear contrast L .

In the (RRRR, RTRT) design, let $\omega = (\omega_{11}, \omega_{21}, \omega_{31}, \omega_{41}, \omega_{12}, \omega_{22}, \omega_{32}, \omega_{42})'$ be the vector of expected values for all the cells, where ω_{jk} is the expected value at the k th sequence and j th period, e.g., $\omega_{11} = \mu + G_1 + P_1 + D_R$.

Table 6 RRRR/RTRT design

Group	Period I	Period II	Period III	Period IV
1	R	R	R	R
2	R	T	R	T

Table 7 Expected values of the sequence-by-period means for RRRR/RTRT

Group	Period I	Period II	Period III	Period IV
1	$\mu + G_1 + P_1 + D_R$	$\mu + G_1 + P_2 + D_R + C_R$	$\mu + G_1 + P_3 + D_R + C_R$	$\mu + G_1 + P_4 + D_R + C_R$
2	$\mu + G_2 + P_1 + D_R$	$\mu + G_2 + P_2 + D_T + C_R$	$\mu + G_2 + P_3 + D_R + C_T$	$\mu + G_2 + P_4 + D_T + C_R$

Table 8 Analysis of variance table for RRRR/RTRT design

Source of variation	Degree of freedom
Intersubject	$n_1 + n_2 - 1$
Sequence	1
Residual	$n_1 + n_2 - 2$
Intra-subject	$n_1 + n_2$
Period	3
Formulation	1
Carryover	1
Residual	$n_1 + n_2 - 5$
Total	$2(n_1 + n_2) - 1$

According to Table 6, $\omega = X'P$ with X a design matrix given in the below.

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \end{pmatrix}, \beta = \begin{pmatrix} \beta_{11} \\ \beta_{21} \\ \beta_{31} \\ \beta_{41} \\ \beta_{12} \\ \beta_{22} \\ \beta_{32} \\ \beta_{42} \end{pmatrix}$$

Setting $E(\beta' \bar{Y}) = L'P$ implies that $\beta' \omega = L'P \Rightarrow \beta' X'P = L'P$. Thus, in order to have $E(\beta' \bar{Y}) = L'P$, we should set $\beta' X' = L' \Rightarrow L = X\beta$. Therefore, the method of moments estimate for $L'P$ is $\beta' \bar{Y}$ with $\beta = (X'X)^{-1} X'L$.

The linear contrast L can be chosen based on the parameter of interest. For testing the difference in drug effect $D = D_T - D_R$, we set $L = (0, 0, 0, 0, 0, 0, 1, -1, 0, 0)'$. For testing the difference in carryover effect $C = C_T - C_R$, we set $L = (0, 0, 0, 0, 0, 0, 0, 0, 1, -1)'$.

Table 9 provides the β 's needed to estimate D_R , D_T , and $D_T - D_R$, respectively, with three sets of contrasts, $L = (0, 0, 0, 0, 0, 0, 0, 1, 0, 0)'$, $L = (0, 0, 0, 0, 0, 0, 0, 0, 1, 0)'$, and $L = (0, 0, 0, 0, 0, 0, 1, -1, 0, 0)'$. Notice that considering the design matrix X is not a squared matrix, generalized inverse (the Moore–Penrose generalized inverse) is needed to obtain a unique solution.

Based on β 's for estimating $D_T - D_R$, we have the following unbiased estimators for drug effect:

Table 9 Coefficients for estimates of drugs in RRRR/RTRT design (adjusting for carryover effect)

Sequence	2β 's to estimate D_R				2β 's to estimate D_T				2β 's to estimate D			
	Period I	Period II	Period III	Period IV	Period I	Period II	Period III	Period IV	Period I	Period II	Period III	Period IV
1	-1	0.5	0	0.5	1	-0.5	0	-0.5	2	-1	0	-1
2	1	-0.5	0	-0.5	-1	0.5	0	0.5	-2	1	0	1

$$\tilde{D} = \frac{1}{2} \left[(2\bar{Y}_{.11} - \bar{Y}_{.21} - \bar{Y}_{.41}) - (2\bar{Y}_{.12} - \bar{Y}_{.22} - \bar{Y}_{.42}) \right]$$

$$E(\tilde{D}) = D_T - D_R, \text{Var}(\tilde{D}) = \frac{3}{2} \sigma_e^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right).$$

When $n_1 = n_2 = n$, $\text{Var}(\tilde{D}) = \frac{3\sigma_e^2}{n}$.

To test the bioequivalence based on average bioequivalence method,

$$H_0 : |D_T - D_R| > \theta \quad \text{vs.} \quad H_1 : |D_T - D_R| \leq \theta,$$

where θ is the prespecified margin. The null hypothesis will be rejected and the bioequivalence will be demonstrated when the statistic $T_D = \frac{\tilde{D} - \theta}{\hat{\sigma}_e^2 \sqrt{\frac{3}{n}}} > t \left[\frac{\alpha}{2}, n_1 + n_2 - 5 \right]$.

The corresponding confidence interval at α significance level is $\tilde{D} \pm t \left[\frac{\alpha}{2}, n_1 + n_2 - 5 \right] \hat{\sigma}_e^2 \sqrt{\frac{3}{2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$.

Table 10 Coefficients for estimates of drug in complete *N*-of-1 design (adjusting for carryover effect)

132β 's to estimate $D = D_T - D_R$				
Sequence	Period I	Period II	Period III	Period IV
1	3	-1	-1	-1
2	-3	-7	-7	17
3	-5	-9	15	-1
4	-11	-15	9	17
5	-5	15	-1	-9
6	-11	9	-7	9
7	-13	7	15	-9
8	-19	1	9	9
9	19	-1	-9	-9
10	13	-7	-15	9
11	11	-9	7	-9
12	5	-15	1	9
13	11	15	-9	-17
14	5	9	-15	1
15	3	7	7	-17
16	-3	1	1	1

Table 11 Analysis of variance table for complete *N*-of-1 design

Source of variation	Degree of freedom
Intersubject	$N - 1$
Sequence	15
Residual	$N - 16$
Intra-subject	N
Period	3
Formulation	1
Carryover	1
Residual	$N - 5$
Total	$2N - 1$

Note: N is the total number of patients enrolled in the study.

To evaluate the biosimilarity in the complete *N*-of-1 design, a similar approach is needed as well to derive the estimated coefficients for estimating the difference in drug effect (Table 10). The analysis of variance table is shown in Table 11.

The coefficients in Table 10 are needed to construct $\tilde{D}$, and we have

$$E(\tilde{D}) = D_T - D_R, \text{Var}(\tilde{D}) = \frac{\sigma_e^2}{11n},$$

where n is the number of subjects enrolled in each sequence (assuming equally allocated).

The null hypothesis will be rejected and the bioequivalence will be demonstrated when the statistic $T_D = \frac{\tilde{D} - \theta}{\hat{\sigma}_e^2 \sqrt{\frac{1}{11n}}} > t \left[\frac{\alpha}{2}, 16n - 5 \right]$. The corresponding confidence interval at α significance level is $\tilde{D} \pm t \left[\frac{\alpha}{2}, 16n - 5 \right] \hat{\sigma}_e^2 \sqrt{\frac{1}{11n}}$.

For a fixed sample size, e.g., $N=48$, the number of patients can be randomly assigned to each sequence in the RRRR/RTRT design is 24, whereas the number of patients can be randomly assigned to each sequence in the complete *N*-of-1 design with four dosing periods is 3. Correspondingly, the variances of drug effects for the (RRRR, RTRT) and complete *N*-of-1 designs when adjusting for the carryover effect are $\sigma_e^2/8$ and $\sigma_e^2/33$, respectively. The relative efficiency between two study designs is 24.24%, indicating that the efficiency of RRRR/RTRT design is 24.24% of the complete *N*-of-1 design. Therefore, the partial design (RRRR/RTRT) is less efficient than the

Table 12 Complete N-of-1 design with two periods

Group	Period I	Period II
1	R	R
2	R	T

Table 13 Coefficients for estimates of drugs in Balaam's design (adjusting for carryover effect)

Sequence	β 's to estimate $D = D_T - D_R$	
	Period I	Period II
1	-0.5	0.5
2	0.5	-0.5
3	0.5	-0.5
4	-0.5	0.5

complete design (N-of-1 design with four periods). Similar analysis can be performed for the comparison between the complete design and the partial design with two periods and three periods.

Complete N-of-1 Design with Two Periods vs. 2 × 2 Crossover Design

As introduced earlier, the complete N-of-1 design with two periods is a Balaam design. The comparison between the Balaam design and the 2 × 2 crossover design (see Table 12) is performed based on the similar procedure.

The estimated coefficients for estimating the difference in drug effect is derived for the complete design (Table 13).

The coefficients in Table 13 are needed to construct $\tilde{D}$, and we have

$$E(\tilde{D}) = D_T - D_R, \text{Var}(\tilde{D}) = \frac{2\sigma_e^2}{n},$$

where n is the number of subjects enrolled in each sequence (assuming equally allocated).

Following the similar procedure, the estimated coefficients for the estimated difference in drug effect is derived for the RR/RT design (Table 14).

$$\text{According to this table, } E(\tilde{D}) = D_T - D_R, \text{Var}(\tilde{D}) = \frac{4\sigma_e^2}{n}$$

when adjusting for the carryover effect.

When the total sample size is 48, 12 patients will be enrolled in each sequence for the Balaam design and 24 patients will be enrolled in each sequence for the 2 × 2 design. Correspondingly, the variances of drug effects for the (RR, RT) design and the complete N-of-1 design when adjusting for the carryover effect are both $\sigma_e^2/6$. The relative efficiency between the two study designs is 100%, indicating that the efficiency of RR/RT design is the same as that of the complete N-of-1 design with two periods.

Table 14 Coefficients for estimates of drugs in RR/RT design (adjusting for carryover effect)

Sequence	β 's to estimate $D = D_T - D_R$	
	Period I	Period II
1	1	-1
2	-1	1

Table 15 Coefficients for estimates of drug effect in complete N-of-1 design with three periods (adjusting for carryover effect)

Sequence	40β 's to estimate $D = D_T - D_R$		
	Period I	Period II	Period III
1	12	-4	-4
2	10	-6	-6
3	2	-14	-6
4	0	-16	-8
5	2	-6	18
6	0	-8	16
7	-8	-16	16
8	-10	-18	14

Table 16 Coefficients for estimates of drug effect in 2 × 3 dual design (adjusting for carryover effect)

Sequence	β 's to estimate $D = D_T - D_R$		
	Period I	Period II	Period III
1	1	-0.5	-0.5
2	-1	0.5	0.5

Complete N-of-1 Design with Three Periods vs. 2 × 3 Dual Design

Comparing the relative efficiency between the complete N-of-1 design with three periods and the 2 × 3 dual design, we then have the estimated coefficients of difference in drug effect for two designs (Tables 15 and 16).

For the complete design, we have $E(\tilde{D}) = D_T - D_R$, $\text{Var}(\tilde{D}) = \frac{3\sigma_e^2}{10n}$ according to Table 15. For the partial design, we have $E(\tilde{D}) = D_T - D_R$, $\text{Var}(\tilde{D}) = \frac{3\sigma_e^2}{n}$ according to Table 16.

For a sample size of 48, there would be 6 subjects in each sequence for the complete design and 24 subjects in each sequence for the partial design. Then the variances of drug effects for the 2 × 3 dual design and the complete N-of-1 design with three periods when adjusting for the carryover effect are $\sigma_e^2/8$ and $\sigma_e^2/20$, respectively. The relative efficiency between two study designs is 40%, indicating that the efficiency of the RR/RT design is 40% less than that of the complete N-of-1 design with two periods.

Summary

In general, complete *N*-of-1 designs show more efficiency than partial designs with the same number of periods. The relative efficiency has no difference when comparing the complete *N*-of-1 design with two periods to the RR/RT design. However, as the dosing period increases, the relative efficiency is decreasing for the comparison, indicating that the complete designs are superior than the partial designs in terms of relative efficiency. With four dosing periods, the relative efficiency is only 24% comparing the partial design with the complete *N*-of-1 design. Table 17 is a summary table when the sample size is 48 to show the relative efficiencies for comparison with two, three, and four dosing periods.

Power Analysis

The power analysis can be also performed to see the superiority of the complete *N*-of-1 design compared to the partial design. The sample size determination under the fixed power and significance level is derived based on the following hypothesis testing.

$$H_0: |D_T - D_R| > \theta \quad \text{vs} \quad H_1: |D_T - D_R| \leq \theta.$$

According to the $\pm 20\%$ rule, the bioequivalence is concluded if the average bioavailability of the test drug effect is within $\pm 20\%$ of that of the reference drug effect with a certain assurance. Therefore, θ is usually represented by $\nabla \mu_R$ where $\nabla = 20\%$, and the hypothesis testing can be rewritten as follows:

$$H_0: \mu_T - \mu_R \left\langle -\nabla \mu_R \text{ or } \mu_T - \mu_R \right\rangle \nabla \mu_R \quad \text{vs.} \\ H_a: -\nabla \mu_R \leq \mu_T - \mu_R \leq \nabla \mu_R.$$

The power function can be written as

$$P(\theta) = F_v \left(\left[\frac{\nabla - R}{CV \sqrt{b/n}} \right] - t(\alpha, v) \right) \\ - F_v \left(t(\alpha, v) - \left[\frac{\nabla + R}{CV \sqrt{b/n}} \right] \right),$$

Table 17 Summary table for the relative efficiency (sample size = 48)

Dosing periods	Relative efficiency (%)
Two	100
Three	40
Four	24

Table 18 Summary for additive model (with four periods)

	<i>R</i>	
	5%	10%
RRRR/RTRT design	136	300
Complete <i>N</i> -of-1 design	32	48

where $R = \frac{\mu_T - \mu_R}{\mu_R}$ is the relative change; $CV = \frac{S}{\mu_R}$; μ_T, μ_R are the average bioavailability of the test and reference formulations, respectively; S is the squared root of the mean square error from the analysis of variance table for each crossover design; $[-\nabla \mu_R, \nabla \mu_R]$ is the bioequivalence limit; $t(\alpha, v)$ is the upper α th quantile of a *t*-distribution with v degrees of freedom; F_v is the cumulative distribution function of the *t*-distribution; and b is the constant value for the variance of drug effect.

Accordingly, the exact sample size formula when $R = 0$ is $n \geq b \left[t(\alpha, v) + t\left(\frac{\beta}{2}, v\right) \right]^2 [CV/\nabla]^2$; the approximate sample size formula when $R > 0$ is $n \geq b \left[t(\alpha, v) + t(\beta, v) \right]^2 [CV/(\nabla - R)]^2$.

Another way to determine the sample size is based on testing the ratio of the drug effect between the bio-similar product and the reference product. Considering $\delta \in (0.8, 1.25)$ is the bioequivalence range of μ_T/μ_R , the hypothesis changes to

$$H_0: \frac{\mu_T}{\mu_R} \left\langle 0.8 \text{ or } \frac{\mu_T}{\mu_R} \right\rangle 1.25 \quad \text{vs.} \quad H_a: 0.8 \leq \frac{\mu_T}{\mu_R} \leq 1.25.$$

In case of the skewed distribution, the hypotheses are transformed to logarithmic scale:

$$H_0: \log \mu_T - \log \mu_R \left\langle \log(0.8) \text{ or } \log \mu_T - \log \mu_R \right\rangle \log(1.25) \\ \text{vs.} \quad H_a: \log(0.8) \leq \log \mu_T - \log \mu_R \leq \log(1.25)$$

Then, the sample size formulas for different δ 's are given below:

$$n \geq b \left[t(\alpha, v) + t\left(\frac{\beta}{2}, v\right) \right]^2 [CV/\ln 1.25]^2 \quad \text{if } \delta = 1,$$

$$n \geq b \left[t(\alpha, v) + t(\beta, v) \right]^2 [CV/(\ln 1.25 - \ln \delta)]^2 \\ \text{if } 1 < \delta < 1.25,$$

$$n \geq b \left[t(\alpha, v) + t(\beta, v) \right]^2 [CV/(\ln 0.8 - \ln \delta)]^2 \\ \text{if } 0.8 < \delta < 1.$$

To compare the sample sizes between two designs, for instance, we have a summary table (see Table 18) of the number of subjects for the additive models between the RRRR/RTRT design and the complete *N*-of-1 design with four periods under 80% power, $CV = 20\%$, and $\theta = 5\%$ and 10% .

Table 19 summarizes the number of subjects for the multiplicative models between the RRRR/RTRT design and the complete *N*-of-1 design with four periods under 80% power, $CV = 20\%$, and $\delta = 0.90$ and 1.00 .

Table 19 Summary for multiplicative model (with four periods)

	δ	
	0.90	1.00
RRRR/RTRT design	110	44
Complete <i>N</i> -of-1 design	32	16

CONCLUSION REMARKS

Although the drug interchangeability including switching and alternation can be assessed in both the partial design and the complete *N*-of-1 design (especially with four dosing periods), complete *N*-of-1 designs show better performance in terms of relative efficiency and power analysis. With more information, the complete *N*-of-1 design has the advantage of evaluating all possible switches and alternations within one clinical trial. Requiring less number of subjects enrolled in the trial, the complete *N*-of-1 design can be also potentially used in rare disease evaluation. Because of the difficulty in enrolling patients with rare disease, it is important to conduct the clinical trial with moderate sensitivity to detect biosimilarity with a limited

number of patients. Under such situations, complete *N*-of-1 designs are suitable for rare diseases as this kind of crossover designs is more informative and efficient and requires less number of subjects enrolled than partial designs. However, too many sequences in the complete design will cause much fewer subjects to be allocated to each sequence. In case of the withdrawal of subjects, the conduction of complete design may face a missing data issue, which will affect the evaluations more than that in the partial design.

REFERENCES

1. Chow, S.C. *Biosimilars: Design and Analysis of Follow-on Biologics*, 1st Ed.; Chapman & Hall/CRC: Boca Raton, FL, 2013.
2. FDA. *Guidance on Scientific Considerations in Demonstrating Biosimilarity to a Reference Product*; The United States Food and Drug Administration: Silver Spring, MD, 2015.
3. Chow, S.; Song, F.; Cui, C. On hybrid parallel–crossover designs for assessing drug interchangeability of biosimilar products. *J. Biopharm. Stat.* **2017**. doi:10.1080/10543406.2017.1275956.

Clinical Trials

Kenneth B. Schechtman

Washington University School of Medicine, St. Louis, Missouri, U.S.A.

Abstract

This entry provides a broad overview of some general principles that govern the design and conduct of clinical trials. Specifically discussed are operational approaches to conducting a randomized and blinded trial and reasons why the absence of randomization, unblinded studies, and loss to follow-up can be major sources of bias in clinical trials.

INTRODUCTION

The randomized clinical trial (RCT) is broadly considered to be the most effective approach to evaluating a new therapeutic modality and to testing an existing therapy in a new setting. In this article, we provide a broad overview of some general principles that govern the design and conduct of clinical trials. Specifically, we discuss operational approaches to conducting a randomized and blinded trial and reasons why the absence of randomization, unblinded studies, and loss to follow-up can be major sources of bias in clinical trials. In addition, we focus on the rationale for the prospective selection of a small number of primary endpoints, for including interim analyses in a trial, and for the use of intention-to-treat when analyzing clinical trial data. Brief discussions of procedural issues that are relevant to the recruitment of clinical trial subjects and to the conduct and structure of multicenter trials are also presented. Because they are discussed elsewhere in this volume, we omit such key issues as sample-size determination, statistical power, compliance to prescribed therapeutic regimens, and data analytic considerations.

DEFINITION

A *clinical trial* is a prospective study in which two or more treatments are compared with one another. All therapeutic arms may be active treatments, or, alternatively, a placebo may be used in one group. The most definitive kind of clinical trial is the randomized trial, which requires that group assignment be the result of some random process beyond the control of both the investigator and the subject. A well-designed clinical trial requires a prospectively written protocol that describes in detail the design of the proposed research. The protocol should include a small number of clearly defined hypotheses, specific aims, and outcome measures. It should also provide a detailed delineation of eligibility and exclusion criteria, a description of

the proposed interventions, discussions of recruitment and randomization processes, a detailed consideration of how the planned study will avoid potential sources of bias, a discussion of the quality control measures that will minimize the potential for errors in the collection and the computerization of data, a justification of the planned sample size, and a presentation of the statistical methods that will be used to evaluate the data.

RANDOMIZATION

The purposes of randomization are as follows. First, to reduce the likelihood that there will be selection bias, the bias that may result when the selected sample is unrepresentative of the target population. Second, to minimize the likelihood that subjective decisions about treatment assignment will yield between-group differences in subject characteristics at baseline. When selection bias yields an unrepresentative sample, it may be difficult to define the population to which study results can be generalized. If one group is either older, sicker, or, otherwise, less likely to benefit from treatment, the apparent success of the alternative therapy may reflect nothing more than the positive prognostic status of the sample that has been assigned to that treatment. While carefully conducted nonrandom treatment assignment may yield balance with respect to known risk factors such as age and disease status, there is no way that subjective processes can address the impact of unknown risk factors. Similarly, while covariate adjustment during the analysis stage may account for the effect of known predictors that differ between groups, you cannot adjust for factors you do not know about.

The theoretical benefits of randomization have been demonstrated in a number of studies that have compared results from randomized trials with those that have used nonrandom treatment allocation.^[1–5] The consistent message is that trials that use nonrandom assignment tend to produce biased overestimates of true therapeutic efficacy.

For example, Sacks et al.^[1] evaluated 50 RCTs and 56 trials that used historical controls (HCTs) and found that 44 of the 56 HCT reports (79%) concluded that the therapy was better than the control regimen, while only 10 of the 50 (20%) reports on RCTs reached this conclusion. Chalmers et al.^[3] used similar methods in comparing the conclusions of RCTs, nonrandomized trials that used concurrent controls (CCTs) and HCTs in evaluating the efficacy of anti-coagulants following acute myocardial infarction. Their conclusion was that efficacy estimates were smallest with RCTs, were second smallest with CCTs, and were largest with HCTs.

It has recently been argued that it is common for investigators to work actively toward a subversion of the randomization process through a variety of approaches to determining in advance the group to which subjects will be assigned.^[6] Careful attention to the mechanics of randomization is essential because inadequately concealed randomization can inflate the odds ratios that estimate treatment effects^[7] by 30–40% and can bias results in the direction of unwarranted significance.^[2] The most commonly used approaches to maximizing the integrity of the process are as follows: (1) the use of numbered opaque and sealed envelopes that contain randomization information and that are opened only after eligibility has been established, and the subject has given informed consent; (2) randomization by telephone or fax communication with a coordinating center (CC) or some other agent that is physically separated from the investigator; (3) randomization using password-protected Web-based systems.

The literature contains many discussions of the scientific approaches to group assignment.^[8–11] Regardless of the mechanics of the process itself, randomization is almost always accomplished using random-number generators to make the assignment. While equal allocation is most common, it is not at all unusual to employ randomization schemes that assign different numbers of patients to different groups. The most basic approach to determining the order of treatment assignment is to use simple randomization, a process whose defining characteristic is that each new treatment assignment is random and is fully independent of all preceding assignments. Two disadvantages of this approach are that the laws of chance mean that simple randomization may yield an imbalance in the number of patients assigned to the various study groups, and, alternatively, there may be temporal bias caused by too many subjects being assigned to one group early in the study and too few at later time points.

To address these problems, *blocked* randomization is often used. When blocking is employed, each consecutive group of k subjects is assigned in some fixed proportion to all study groups. In this way, temporal bias is avoided and the target randomization percentages can be guaranteed. When there is concern that investigators in unblinded studies may use information about block sizes to break randomization codes, it may be best to keep investigators

uninformed as to the size of blocks or to use varying block sizes. With or without blocking, it is common to use *stratified* randomization. Most frequently, the use of stratification reflects the desire to ensure that there will be between-group balance with respect to some known important risk factor. Thus, if gender is known to be associated with treatment success, stratification by gender can ensure gender balance by using separate blocked randomization schemes for male and female subjects. Other less commonly used randomization schemes include a variety of adaptive methods^[10,11] in which nonconstant randomization probabilities depend partly or fully on the degree of between-group balance that has been achieved in previously randomized subjects. With some adaptive procedures, such as the play-the-winner rule,^[12] randomization probabilities can depend on success rates with previous subjects.

BLINDING

The term *blinding* (or *masking*) refers to the decision to keep subjects and/or investigators uninformed as to treatment assignment as the study progresses. A *single-blind* study is one in which the subject is unaware of his treatment group assignment; a *double-blind* study is one in which neither the subject nor the investigator is provided with this information. In the less commonly used *triple-blind* study, the Data- and Safety-Monitoring Committee (DSMC) is also kept blinded when they evaluate data as the study progresses.

The purpose of blinding is to reduce the likelihood that study assessments will be biased because subject or investigator behavior has been influenced by the knowledge of treatment group assignment. Thus, a subject who knows he has been assigned to a usual care group may be more inclined to drop out or become noncompliant because he would prefer to be in the investigative arm of the trial. Similarly, an investigator who wants to establish the efficacy of a new drug may be influenced in subtle ways when she must make close decisions about ambiguous events if she knows which treatment the subject has received. As is the case with nonrandom or inadequately concealed treatment allocation, the literature suggests that there can be substantial differences between the results of blinded and nonblinded studies. Thus, in a review of 250 controlled clinical trials, Schulz et al.^[7] found that the absence of a double blind was associated with a 17% inflation of the odds ratio that describes the benefit of treatment. Similarly, Colditz et al.^[4] found that RCTs that did not use a double blind were significantly more likely to report a benefit on the new treatment than were double-blind RCTs.

A basic principle of clinical trial conduct is that while blinding is not always feasible, it should not be viewed as an all-or-nothing proposition. Thus, in a medication vs. surgery trial, it is obviously impossible to keep physicians

or subjects blinded as to treatment assignment. However, it may be quite reasonable to insist that the individual who interprets the laboratory test or the CT scan that determines therapeutic success is blinded as to the treatment received by the patient. In general, the goal of every trial should be to maximize the level of blinding within the framework of the limitations that are imposed by practical considerations. Whatever the level of blinding, ethical considerations require that there be a mechanism through which blinding codes can be broken when such action is necessitated by concerns about patient safety.

SELECTION AND ASSESSMENT OF ENDPOINTS

An endpoint in a clinical trial is a parameter that will be compared across the arms of the study in order to evaluate the success of the study treatment. Endpoints should be prospectively and unambiguously defined, quantifiable or categorical measures that correspond precisely to the scientific goals of the trial. Because they are the ultimate determinants of treatment success, every effort should be made to avoid bias in the evaluation of endpoints. Thus, following the suggestion in the “Blinding” section, if the blinding of investigators is not feasible in a given trial, procedures that ensure the blinded evaluation of endpoints should be implemented whenever it is practical to do so. When the endpoint is, e.g., cardiovascular mortality, committees are often set up to provide an unbiased determination of the true cause of death. When endpoints in multicenter trials involve the evaluation of blood samples or of electrocardiograms, it is common to use a centralized reading center for all clinical sites to ensure that determinations are made using the same procedures.

Endpoints in clinical trials are often categorized as primary, secondary, and, sometimes, exploratory. A basic principle of trial design is that there should be a small number of *primary* endpoints that are the defining measures of the success of the study treatment. The primary endpoints are the most important measures of treatment success. The emphasis on a small number of such endpoints reflects concern about the possible effect of multiple-comparison problems, the unavoidable reality that if you have many primary endpoints, the probability that one or more will be significant by chance is increased. If there are many primary endpoints, concerns about multiple comparisons may compromise the most basic conclusions of the study. With one or two primary endpoints, the multiple-comparison problem does not arise.

In many clinical trials, a decision must be made as to whether the primary endpoints should be “true” endpoints or “surrogate” endpoints. Generally speaking, a *true* endpoint is one that has direct clinical relevance to the patient. Examples include mortality, the remission of a cancer, and the successful treatment of a disease. *Surrogate* endpoints are used as a less rigorous alternative to the clinical

endpoint in settings where the clinical endpoint is too difficult, time consuming, or expensive to measure. Such endpoints can only be useful if they are highly associated with the true endpoint of interest. Thus, blood pressure is a potentially viable surrogate for stroke because hypertension is known to be a major risk factor for stroke. For similar reasons, CD4 levels may be a reasonable surrogate for mortality in AIDS patients.

The most important reason for using surrogate endpoints is the potentially large reduction in sample size, trial duration, and cost that they may yield when the true endpoint is infrequent or requires many years to observe. Thus, Wittes et al.^[13] suggest that a trial whose primary endpoint was a reduction in stroke rate might require 5 years of study and 25,000 patients, whereas the same trial using diastolic blood pressure as a surrogate endpoint could be completed in 1–2 years using 200 patients. Unfortunately, this huge sample-size advantage has a large potential downside because there are many predictors of stroke and reducing blood pressure may leave stroke rates unchanged. In the Cardiac Arrhythmia Suppression Trial (CAST^[14,15]), the effect on mortality of three FDA-approved antiarrhythmic drugs were tested in high-risk patients with a history of both myocardial infarction and ventricular ectopy. The logic of CAST was that as these drugs were known to reduce the rate of a major risk factor for cardiac death, they should also reduce death rates. Unfortunately, the exact opposite happened, and two arms of the study were stopped early because of a substantial increase in mortality. Similar problems with surrogate endpoints have been found by meta-analyses that have demonstrated that quinidine is associated with increased mortality although it can control atrial fibrillation^[16] and that lidocaine is associated with both a one-third reduction in ventricular tachycardia and a one-third increase in mortality.^[17,18]

RECRUITMENT PROCEDURES

A vital component of every clinical trial is the expeditious recruitment of a sufficient number of eligible subjects. Because of this, the protocol writing and planning stages of every trial must include a basic feasibility evaluation: Are we capable of recruiting the required number of subjects? The answer to this practical question can be a driving force in determining eligibility and exclusion criteria. In some cases, it can force a modification of the scientific goals of a study. In the worst case, it can lead to the conclusion that a proposed study should not be attempted. When recruitment is slower than anticipated, the cost of an already expensive process goes up, investigators may lose interest, study quality may deteriorate, and there is an incentive to randomize subjects whose eligibility may be questionable. If delays are sufficiently prolonged, the scientific questions that motivated the study may diminish in importance or become moot. Finally, it is always possible

that a study may have to be stopped early because of inadequate recruitment.

In spite of, and perhaps because of, its importance, there is a natural tendency to overestimate recruitment capabilities. This is especially true because no clinic will be invited to participate in a multicenter trial if it cannot justify its ability to enroll some minimum required number of subjects. The tendency toward overestimation is emphasized in a comprehensive review of the recruitment literature by Hunninghake et al.,^[19] who asserted that “it is extremely rare for a study to enroll the required number of participants within the time frame originally designated.” This observation is highlighted by an evaluation of recruitment success in seven multicenter clinical trials sponsored by the Veterans Administration.^[20] This review found that one trial was stopped early because of inadequate recruitment, while the remaining six required more time than was initially planned to reach recruitment targets.

The potential magnitude of the recruitment burden is illustrated by large multicenter trials such as the Lipid Research Clinics Primary Prevention Trial^[21] and the Systolic Hypertension in the Elderly Program.^[22] In both of these studies, it was necessary to screen several hundred thousand subjects because both trials were such that approximately 100 subjects had to be screened in order to find a single subject who was ultimately randomized. While screened to eligible ratios of 100 or more are not typical, these examples provide ample evidence that a realistic and prospective appraisal of recruitment capabilities is vital in every clinical trial. They also emphasize the fact that the mundane details of the recruitment process may sometimes deserve as much attention as do the more glamorous scientific concerns of a trial. The many standard approaches to recruiting subjects for clinical trials include direct contact at clinics; public screenings for conditions such as hypertension, diabetes, and hypercholesterolemia; mailings and phone calls to subjects who are known to have a given disease; referrals from private physicians and primary care clinics; retrospective chart reviews, an approach that may be the only option when diseases are rare; and television and radio advertisements. The strengths and weaknesses of these and other approaches to recruitment are discussed by Meinert.^[23]

LOSS TO FOLLOW-UP

A fundamental objective of every well-conducted RCT is to minimize the rate at which subjects are lost to follow-up. If loss to follow-up occurs in a purely random fashion and to the same degree in each arm of the study, it may be associated with a loss of statistical power because of the reduced sample size, but it is unlikely to be an important source of bias. However, even when the apparent reason for loss to follow-up is benign, the impact on study results may be substantial. For example, patients may terminate

their involvement in a study because they have moved out of town or because their level of frailty makes it difficult for them to get to the clinic. While these behaviors are likely to be balanced across groups in a randomized study, they may compromise the generalizability of results by excluding too many of the sickest patients or a preponderance of older patients who have moved away when they retired. More serious consequences may result if the reasons for loss to follow-up are associated with characteristics of the treatment. Thus, the Coronary Drug Project (CDP) Research Group^[24] found that loss to follow-up was greatest in the treatment arm that also had the largest number of drug-related adverse events. In the Lipid Research Clinics Coronary Primary Prevention Trial,^[25] the somewhat different problem was that treatment side effects led to worse compliance in the treatment group than in the placebo group.^[26] In a more recent RCT involving the treatment of AIDS patients,^[27] it was reported that loss to follow-up was significantly associated with several measures of disease progression. The authors concluded that “losses to follow-up probably decreased substantially the observed number of primary endpoints, curtailed the power of the trial to demonstrate any difference between immediate and deferred initiation of antiretroviral therapy, and may have introduced large bias in the estimated hazard ratio for the primary endpoint and its statistical significance.”

There are many standard approaches to encouraging clinical trial participants to remain in a study and to attend required follow-up visits. These include providing routine conveniences such as flexible appointment schedules, limiting waiting times, and easy clinic access; regular contacts with subjects between visits; acknowledging special occasions such as patient birthdays and Christmas; payment of parking and travel fees associated with the study and, sometimes, paying the subject for his willingness to participate; and keeping the subject informed of the details of the protocol and of the scientific importance of his contribution as a participant.

It is often the case that loss to follow-up is associated with subjects who relocate. When investigators are not directly informed of the new address, simple steps, such as contacting friends, relatives, and employers, or getting address change information from the post office, are likely to be helpful. When the new address is known and when the study is multicenter, it is often feasible to see the subject at another study clinic. Alternatively, subjects who relocate can sometimes be seen when they return for visits to their former city of residence by enlisting and paying colleagues who live in the city to which a subject moves, or by having study personnel see the subject when they are in the vicinity of a subject's new residence. If a subject is lost to follow-up in a study where the primary endpoint is mortality, the National Death Index can serve as a valuable resource for determining with a high degree of confidence whether or not the subject has died. The fact that these

somewhat extreme measures are not uncommon highlights the importance attached to maximizing follow-up rates.

MULTICENTER TRIALS

Multicenter clinical trials are trials that enrol patients from more than one clinical site. They are generally used because the required sample size is too large for the trial to be feasible using only one clinical site, or because the use of a single site would mean that it would take an excessive period of time to complete the study. The primary challenge of multicenter trials is found in the organizational complexity that is necessary to ensure that the protocol is followed in a uniform way at the geographically distant participating clinics. At the top of the trial organization is the study chair, who has the primary executive responsibility for the trial and who is chair of the executive committee. In addition to the study chair, executive committee members will generally include the principal investigator at all resource centers along with other individuals from within and without the trial who have expertise that will facilitate the scientific goals or practical conduct of the trial. The executive committee has overall day-to-day decision-making responsibility for the trial. Its specific responsibilities include maintaining the scientific integrity of the trial; preparing or overseeing the preparation of a manual of procedures and other study documents; appointing subcommittees that are responsible for tasks that include writing manuscripts, verifying endpoints, and codifying publication policies; overseeing the development of study forms; reviewing study manuscripts prior to submission; responding to concerns about clinics that may be having difficulty with enrolment or other trial activities; setting priorities for the study; and having the basic responsibility for the design and conduct of the trial.

The conduct of every multicenter trial is facilitated by one or more resource centers. The one such center that is integral to the functioning of every multicenter trial is the CC. The CC has a broad range of responsibilities that cover all phases of the trial. During the protocol development stage, the focus of the CC is on questions of sample size and statistical power, the scientific integrity of the proposed study design, a data analysis plan, and the development of data-processing and quality control procedures. When patients are being recruited and followed up, the CC is focused on such activities as the day-to-day processing and quality control of data; regular communication with the clinical sites so as to ensure the completeness and accuracy of the data they provide; discussing quality control problems with clinical sites and with the executive committee; and preparing reports that measure the progress of the trial through clinic-specific and studywide statistics on recruitment rates, follow-up rates, error rates on forms, missed patient visits, and other measures of study progress. In many multicenter trials, the CC will also be involved in

activities that include site visits to the clinics; designing, modifying, and disseminating study forms; and training clinic personnel in assessment procedures. When the enrolment and follow-up of subjects have been completed, the activities of the CC are redirected toward the analysis of data, extensive collaboration in the preparation of manuscripts, and archiving the data generated by the study.

In addition to a CC, most multicenter clinical trials have resource centers that include reading centers and central laboratories. These are centralized facilities whose purpose is to ensure the uniform interpretation of laboratory and clinical data. They tend to be used when the parameter of interest is critical to the trial and when the complexity or lack of standardization of the assessment suggests that local evaluations would be: (1) impossible because the expertise or equipment is not available at some sites, or (2) difficult to standardize because of the inherent characteristics of the assay.

An additional organizational component of most multicenter clinical trials is the *DSMC*. This committee contains outside experts who are recommended by the executive committee and, in general, appointed by the funding agency. The primary purpose of the DSMC is to provide independent oversight and advice regarding all major study activities. It meets periodically with the executive committee to evaluate interim data, to assess the progress of the study as measured in reports prepared by the CC and the reading centers, and to review the ethical conduct of the study. It is frequently responsible for providing external review of the resource center and for providing advice as to whether poorly functional clinical centers should be dropped from the study or placed on probation. Following its periodic reviews of interim data, the DSMC may recommend changes in the study protocol such as an increase in the sample size. In more extreme settings, it may suggest the termination of the study because of treatment side effects or because of inadequate response to study treatment.

INTERIM DATA ANALYSES

A common feature of many clinical trials is the interim data analysis. An *interim* analysis is an evaluation of the data while patient recruitment or follow-up is still progressing. The purpose of these analyses is to determine whether early trends in the data suggest the need to modify the study protocol. Such modifications may take many forms. For example, an interim analysis may suggest that a modification of the dosing schedule is necessary because of excessive toxicity. The Cardiac Arrhythmia Suppression Trial^[14,15] provides a more extreme illustration of the potential impact of an interim analysis. In this trial, the interim data resulted in the termination of the encainide and flecainide arms of the study because the data suggested increased mortality with these two drugs. A clinical trial (or an arm of the trial) may

also be terminated if an interim analysis suggests that the treatment is ineffective and that there is no hope of ever achieving statistical significance, or, as happened with the aspirin arm of the Physicians Health Study,^[28] that the treatment is so clearly effective that it would be unethical to continue to assign subjects to the placebo group. If the interim data suggest that between-group differences in primary outcome measures are smaller than was anticipated in the original sample-size computations, the conclusion may be that the target sample size should be increased. Because interim analyses are concerned with such basic issues as toxicity and the continued use of a placebo when existing data may have established the efficacy of the treatment, such analyses are viewed by many investigators as a being moral imperative.

In a single-center trial, interim analyses are ordinarily conducted by the study statistician. In multicenter trials, these analyses are conducted by the CC, often with active input from the DSMC, the committee responsible for making recommendations in response to the interim report. Results of the interim analysis are presented by the executive committee to the DSMC. However, because of concerns about blinding, it is sometimes the case that these results can be seen only by the DSMC, the principal investigator of the CC, and the representative of the funding agency. Thus, meetings about interim reports are sometimes conducted after the chair of the executive committee and the directors of the reading centers have been asked to leave the room. In single-center trials, the ideal committee structure may not exist, and it may be necessary to devise more ad hoc approaches to ensure that investigators who recruit or evaluate patients are not aware of the contents of the interim reports.

A statistical concern that arises naturally in trials that contain interim data analyses is the multiple-look problem. The *multiple-look* problem reflects the fact that the possible early termination of a study, because of results from interim looks at the data are significant, increases the likelihood that significant results will be achieved when the null hypothesis is true.^[29,30] Thus, if the type I error is supposed to be 0.05 and if you may stop the study early because an interim analysis produced a *p*-value of less than 0.05, the repeated opportunities to claim significance artificially inflate the type I error. That is, the true probability of claiming significance when the null hypothesis is true is greater than 0.05. In order to address this problem, a variety of “group sequential” testing procedures have been developed. While we do not discuss the details in this article, all of these group sequential procedures have a common theme. Using some formal statistical mechanism, the “nominal” *p*-value that is necessary to establish significance when the various data analyses are conducted is reduced to something less than 0.05. The precise nominal *p*-values are defined by the desire to ensure that the true type I error is 0.05. To accomplish this task, the sum of the probabilities of claiming significance at all analyses

combined must be exactly 0.05, when the null hypothesis is true. For details on specific group sequential testing procedures, we refer the reader to the methods developed by Pocock,^[31] O’Brien and Fleming,^[32] and Lan and DeMets.^[33] An excellent general review of these methods is provided by Geller and Pocock.^[34]

WHO GETS ANALYZED?

Protocol violations are an unavoidable feature of most clinical trials. Subjects may drop out of the study early, miss follow-up visits, cross over to a different arm of the study, or, otherwise, be noncompliant with the prescribed therapeutic regimen. The investigator may be the source of the protocol violation when assessment schedules or procedures are not followed and when randomized subjects are ineligible or, as sometimes happens, do not have the disease in question. When protocol violations occur, there can be a powerful incentive to exclude the relevant patients when data are analyzed. After all, the thinking goes, why evaluate a treatment using patients who may have received only a small portion of the prescribed medication?

While this logic has an intuitive appeal, there is ample evidence that in many circumstances, subjects who are excluded from data analyses because of protocol violations will be a source of biased results. The problem with excluding protocol violators is that subjects who violate the protocol may be fundamentally different from those who are fully compliant in the way they respond to treatment, or, in particular, the protocol violation may be partly or entirely a result of the study intervention itself. For these and other reasons, the broadly accepted view is that the intention-to-treat (ITT) principle should govern the primary analyses of clinical trials data unless there is a strong reason to believe that bias will not result from some alternative type of analysis.^[35-37] When ITT is used for the primary analysis, it is sometimes considered appropriate to perform some exploratory analyses using subgroups as long as the reader has been clearly cautioned about the potential limitations of violating the ITT principle.

There are many illustrations of the bias that can result from excluding patients from data analyses. For example, in trials aimed at reducing infarct size in patients with acute myocardial infarction, it is essential to include patients in whom an infarct was ruled out because the infarct size was measured as zero. If you exclude such patients, you may be excluding precisely the patients in whom the treatment was most effective, the ones in whom all myocardium was salvaged by the treatment. More typical are the examples of the potential impact of excluding poor compliers. In the CDP,^[38] the mortality rate in compliant clofibrate patients was 15.7% as compared to a much higher 24.6% in noncompliant clofibrate patients. The drug appears to be effective until you discover that compliance to placebo was associated with an even larger apparent “treatment”

benefit. The problem is that CDP compliers were fundamentally different from CDP noncompliers because the compliers were more likely to engage in heart-healthy behaviors or, otherwise, more likely to have an improved outcome. If the CDP investigators had excluded poor compliers, they would have been analyzing a cohort of patients fundamentally different from the patients who had been randomized in the first place.

GOOD CLINICAL PRACTICE

Good clinical practice (GCP) is usually referred to as a set of standards for clinical trials to achieve and maintain high-quality clinical research in a sensible and responsible manner.^[39] In essence, GCP deals with patients' protection and the quality of data used to prove the efficacy and safety of a drug product. GCP ensures that all data, information, and documents related to clinical trials can be confirmed as being properly generated, recorded, and reported through the institution by independent audits.

CONCLUSION

The randomized clinical trial is broadly viewed as the gold standard approach to clinical research aimed at comparing the efficacy of two or more interventions. The key advantage of randomization is that it is likely to yield between group balance with respect to both known and unknown risk factors, a circumstance that is far less likely using other approaches to clinical research. But the results of a randomized trial may be severely compromised by poor research practice. Thus, results can be biased by the absence of blinding, by excessive dropout and by different rates of dropout in the study groups, or by inappropriate data analytic decisions regarding the use of intention-to-treat. The interpretation of trial results can also be compromised by an inappropriate selection of primary endpoints and the loss of focus that comes from selecting too many endpoints, by inadequate recruitment procedures that yield too few subjects and uncertain conclusions, and by poor quality control in multicenter investigations. Careful attention to the details of the design and conduct of clinical trials is essential if results and conclusions are to be accepted by the research community.

REFERENCES

1. Sacks, H.; Chalmers, T.C.; Smith, H. Jr. Randomized versus historical controls for clinical trials. *Am. J. Med.* **1982**, *72*, 233–240.
2. Chalmers, T.C.; Celano, P.; Sacks, H.S.; Smith, H. Jr. Bias in treatment assignment in controlled clinical trials. *New Engl. J. Med.* **1983**, *309*, 1358–1361.
3. Chalmers, T.C.; Matta, R.J.; Smith, H., Jr.; Kunzler, A. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *New Engl. J. Med.* **1977**, *297*, 1091–1096.
4. Colditz, G.A.; Miller, J.N.; Mosteller, F. How study design affects outcomes in comparisons of therapy: I. Medical. *Stat. Med.* **1989**, *8*, 441–454.
5. Miller, J.N.; Colditz, G.A.; Mosteller, F. How study design affects outcomes in comparisons of therapy: II. Surgical. *Stat. Med.* **1989**, *8*, 455–466.
6. Schulz, K.F. Subverting randomization in controlled trials. *JAMA* **1995**, *274*, 1456–1458.
7. Schulz, K.F.; Chalmers, I.; Hayes, R.J.; Altman, D.G. Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* **1995**, *273*, 408–412.
8. Zelen, M. The randomization and stratification of patients to clinical trials. *J. Chronic Dis.* **1974**, *27*, 365–375.
9. Kalish, L.A.; Begg, C.B. Treatment allocation methods in clinical trials: A review. *Stat. Med.* **1985**, *4*, 129–144.
10. Hoel, D.G.; Sobel, M.; Weiss, G.H. A Survey of Adaptive Sampling for Clinical Trials. In *Perspectives in Biometrics*; Elashoff, R.M., Ed.; Academic Press: New York, 1975.
11. Signorini, D.F.; Leung, O.; Simes, R.J.; Beller, E.; Gebiski, V.J.; Callaghan, T. Dynamic balanced randomization for clinical trials. *Stat. Med.* **1993**, *12*, 2343–2350.
12. Zelen, M. Play-the-winner rule and the controlled clinical trial. *J. Am. Stat. Assoc.* **1969**, *64*, 131–146.
13. Wittes, J.; Lakatos, E.; Probstfield, J. Surrogate endpoints in clinical trials: Cardiovascular diseases. *Stat. Med.* **1989**, *8*, 415–425.
14. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *New Engl. J. Med.* **1989**, *321*, 406–612.
15. Echt, D.S.; Liebson, P.R.; Mitchell, L.B.; Peters, R.W.; Oblas-Manno, D.; Barker, A.H.; Arensberg, D.; Baker, A.; Friedman, L.; Greene, H.L. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. *New Engl. J. Med.* **1991**, *324*, 781–788.
16. Coplen, S.E.; Antman, E.M.; Berlin, J.A.; Hewitt, P.; Chalmers, T.C. Efficacy and safety of quinidine therapy for maintenance of sinus rhythm after cardioversion. A meta-analysis of randomized control trials. *Circulation* **1990**, *82*, 1106–1116.
17. Hine, L.K.; Laird, N.; Hewitt, P.; Chalmers, T.C. Meta-analytic evidence against prophylactic use of lidocaine in acute myocardial infarction. *Arch. Intern. Med.* **1989**, *149*, 2694–2698.
18. MacMahon, S.; Collins, R.; Peto, R.; Koster, R.W.; Yusef, S. Effects of prophylactic lidocaine in suspected acute myocardial infarction. An overview of results from the randomized controlled trials. *JAMA* **1988**, *260*, 1910–1916.
19. Hunninghake, D.B.; Darby, C.A.; Probstfield, J.L. Recruitment experience in clinical trials: Literature summary and annotated bibliography. *Control. Clin. Trials* **1987**, *8*, 6s–30s.
20. Collins, J.F.; Bingham, S.F.; Weiss, D.G.; Williford, W.O.; Kuhn, R.M. Some adaptive strategies for inadequate

- sample acquisition in Veterans Administration Cooperative Clinical Trials. *Control. Clin. Trials* **1980**, *1*, 227–248.
21. Lipid Research Clinics Program. *Recruitment for Clinical Trials: The Lipid Research Clinics Coronary Primary Prevention Trial Experience: Its Implication for Future Trials*; Agras, W.S., Bradford, R.H. Marshall, G.D., Eds.; Circulation 1982, *66* (Suppl. IV), IV1–IV78.
 22. Petrovitch, H.; Byington, R.; Bailey, G.; Borhani, P.; Carmody, S.; Goodwin, L.; Harrington, J.; Johnson, H.A.; Johnson, P.; Jones, M. Systolic Hypertension in the Elderly Program (SHEP): Part 2. Screening and recruitment. *Hypertension*. **1991**, *17* (Suppl.), II16–II23.
 23. Meinert, C.L. *Clinical Trials: Design, Conduct, and Analysis*; Oxford University Press: New York, 1986, 149–153.
 24. Coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. *JAMA*. **1975**, *231*, 360–381.
 25. Lipid Research Clinics. The Lipid Research Clinics coronary primary prevention trial results: II. The relationship of reduction in incidence of coronary heart disease to cholesterol lowering. *JAMA* **1984**, *251*, 365–374.
 26. Schechtman, K.B.; Gordon, M.E. The effect of poor compliance and treatment side effects on sample size requirements in randomized clinical trials. *J. Biopharm. Stat.* **1994**, *4*, 223–232.
 27. Ionnidis, J.P.A.; Bassett, R.; Hughes, M.D.; Volberding, P.A.; Sacks, H.S.; Lau, J. Predictors and impact of patients lost to follow-up in a long-term randomized trial of immediate versus deferred antiretroviral treatment. *J. Acquir. Immune Defic. Syndr. Human Retrovirol.* **1997**, *16*, 22–30.
 28. Preliminary report. Findings for the aspirin component of the ongoing Physicians' Health Study. *N. Engl. J. Med.* **1988**, *318*, 262–264.
 29. Armitage, P. *Statistical Methods in Medical Research*; Blackwell: Oxford, 1971.
 30. McPherson, K. Statistics: The problem of examining accumulating data more than once. *New Engl. J. Med.* **1974**, *290*, 501–502.
 31. Pocock, S.J. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. **1977**, *64*, 191–199.
 32. O'Brien, P.C.; Fleming, T.R. A multiple testing procedure for clinical trials. *Biometrics* **1979**, *35*, 549–556.
 33. Lan, K.K.G.; DeMets, D.L. Discrete sequential boundaries for clinical trials. *Biometrika* **1983**, *70*, 659–663.
 34. Geller, N.L.; Pocock, S.J. Design and Analysis of Clinical Trials with Group Sequential Stopping Rules. In *Biopharmaceutical Statistics for Drug Development*; Peace, K.E., Ed.; Marcel Dekker: New York, 1988, Chap. 11.
 35. Lewis, J.A.; Machin, D. Intention-to-treat—who should use ITT? *Br. J. Cancer* **1993**, *68*, 647–650.
 36. Peduzzi, P.; Detre, K.; Wittes, J.; Holford, T. Intention-to-treat analysis and the problem of crossovers. *J. Thorac. Cardiovasc. Surg.* **1991**, *101*, 481–487.
 37. Newell, D.J. Intention-to-treat analysis: Implications for quantitative and qualitative research. *Int. J. Epidemiol.* **1992**, *21*, 837–841.
 38. The Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project. *New Engl. J. Med.* **1980**, *303*, 1038–1041.
 39. Chow, S.C.; Liu, J.P. *Design and Analysis of Clinical Trials*; John Wiley and Sons: New York, 1998.

Clinical Trials: Controversial Issues

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Fuyu Song

Center for Food and Drug Inspection, China Food and Drug Administration, Beijing, China

Abstract

In clinical development of a test treatment under investigation, clinical trials are often conducted for the evaluation of safety and efficacy of the test treatment. To provide an accurate and reliable assessment of the test treatment under study, adequate and well-controlled clinical trials using valid study design are necessarily conducted for obtaining substantial evidence of the safety and efficacy of the test treatment under investigation. In practice, however, some debatable issues are commonly encountered regardless of the compliance of Good Statistics Practice (GSP) and Good Clinical Practice (GCP). These issues include, but are not limited to, (1) appropriateness of statistical hypotheses for clinical investigation, (2) instability of power calculation for sample size determination, (3) integrity of randomization and blinding, (4) clinical strategy for endpoint selection, (5) impact and sensitivity of protocol amendments, (6) controversial issue of multiplicity in clinical trials, (7) validity and power of missing data imputation, (8) flexibility and feasibility for the use of adaptive design methods, and (9) challenge of the independence of data monitoring committee (DMC). In this article, these issues are briefly described. The impact of these issues on the evaluation of the safety and efficacy of the test treatment under investigation is discussed with examples whenever applicable. Some recommendations regarding possible resolutions of these issues are also provided.

Keywords: Data safety monitoring committee; Endpoint selection; Integrity of blinding; Missing data imputation; Multiplicity; Protocol amendment; Two-stage adaptive designs.

INTRODUCTION

In clinical research and development of a test treatment, relevant clinical data are usually collected from subjects with the diseases under study in order to evaluate the safety and efficacy of the test treatment under investigation. To provide an accurate and reliable assessment, adequate well-controlled clinical trials using valid study designs are necessarily conducted for obtaining substantial evidence of the safety and efficacy of the test treatment under investigation. Clinical trial process, which consists of protocol development, trial conduct, data collection, statistical analysis/interpretation, and reporting, is a lengthy and costly process. This process is necessary to ensure a fair and reliable assessment of the test treatment under investigation. In practice, some controversial or debatable issues inevitably occur regardless of the compliance of Good Statistics Practice (GSP) and Good Clinical Practice (GCP). Controversial issues in clinical research are defined as debatable issues that are commonly encountered during the conduct of clinical trials (see, e.g., the works of Chow,^[1] Chow et al.,^[9] and Chow and Song^[10]). In practice, debatable issues could be raised from (1) compromises between theoretical and real

practices; (2) miscommunication, misunderstanding, and/or interpretation in perception among regulatory agencies, clinical scientists, and biostatisticians; and (3) disagreement, inconsistency, and errors in clinical practice.

In clinical research and development of a test treatment under investigation, commonly seen controversial issues include, but are not limited to, (1) appropriateness of traditional statistical hypotheses for clinical evaluation of *both* safety and efficacy, (2) correctness of power analysis for sample size calculation based on the information from a small-scale pilot study, (3) integrity of randomization and blinding for preventing potential biases, (4) post hoc endpoint selection (based on some derived endpoints), (5) impact of protocol amendments on the characteristics of the trial population, (6) multiplicity in clinical trials, (7) missing data imputation, (8) adaptive design methods, and (9) independence of the data monitoring committee (DMC).

In this article, we will review these debatable issues rather than provide resolutions. The impact of these issues on the evaluation of the safety and efficacy of the test treatment under investigation is discussed with examples whenever applicable. Recommendations regarding possible

resolutions of these issues are also provided whenever possible. It is our goal that medical/statistical reviewers from regulatory agencies such as the US Food and Drug Administration (FDA), clinical scientists, and biostatisticians will (1) pay attention to these issues, (2) identify the possible causes of these debatable issues, (3) resolve/correct the issues, and consequently (4) enhance GSP and GCP for achieving the study objectives of the intended clinical trials.

APPROPRIATE HYPOTHESES FOR CLINICAL INVESTIGATION

In clinical trials for the evaluation of safety and efficacy of a test treatment under investigation, a typical approach is to first test for the null hypothesis of no treatment difference in *efficacy* based on clinical data collected under a valid trial design. The rejection of the null hypothesis would lead to favor the alternative hypothesis that there is a difference. If there is a sufficient power for correctly detecting a clinically meaningful difference (improvement) when such a difference truly exists, we claim that the test treatment is efficacious. The test treatment will then be reviewed and approved by the regulatory agency such as the FDA if the test treatment is well tolerated and there appears no safety concern.

In practice, however, it is a concern whether the traditional approach based on hypothesis testing on efficacy alone (i.e., the study is powered based on the efficacy alone) for the evaluation of *both* safety and efficacy of a test treatment under investigation is appropriate. The test treatment may be approved by the regulatory agency based on the hypothesis testing on efficacy alone and subsequently may get withdrawn due to safety concerns. A typical example is the withdrawal of Vioxx. Vioxx is a COX-2 inhibitor drug intended for treating arthritis approved by the FDA in 1999, which was subsequently withdrawn from the market in 2004 due to the safety concern of increased risk of heart attack and stroke.

To overcome this problem, alternatively, Chow^[1] suggested that both safety and efficacy should be included in a composite hypothesis for testing the clinical benefit of the test treatment under investigation. Composite hypotheses, which take both safety and efficacy into consideration, are summarized in Table 1.

As it can be seen from the table, as an example, suppose that we are interested in demonstrating therapeutic superiority in efficacy and equivalence in safety. In this case, we may consider testing the composite null hypothesis that H_0 : not *SE*, where *S* denotes superiority in efficacy and *E* indicates equivalence in safety. Thus, we would reject the null hypothesis in favor of the alternative hypothesis that H_0 : *SE*. In other words, the test treatment is superior to the active control agent and its safety appears to be equivalent to the active control agent. To test the null hypothesis that

Table 1 Composite hypotheses for clinical investigation

Efficacy	Safety		
	N	S	E
N	NN	NS	NE
S	SN	SS	SE
E	EN	ES	EE

N = Non-inferiority; S = superiority; E = equivalence.

H_0 : not *SE*, appropriate statistical tests should be derived under the null hypothesis. Under the alternative hypothesis, the power of the test statistic can then be evaluated and appropriate sample size can be selected for achieving a desired power.

Switching from a single hypothesis testing (traditional approach based on efficacy alone) to a composite hypothesis testing (based on both efficacy and safety), an increase of the sample size is expected. In the interest of controlling the overall type I error rate at the α level, in composite hypothesis testing, appropriate α levels (say α_1 for efficacy and α_2 for safety) may be chosen.

INSTABILITY OF SAMPLE SIZE CALCULATION

In clinical trials, power analysis for sample size calculation (or power calculation) is necessarily performed to ensure that there is high probability of correctly detecting a clinically meaningful effect size if such an effect size truly exists. In practice, power calculation is often performed based on (1) the information obtained from previous studies or pilot studies, (2) the information from literature, or (3) the investigator's guess or belief with or without scientific justification. Since a pilot study is usually a small-scale study with limited number of subjects, the data obtained from the pilot study and/or the investigator's guess or belief could be deviated far from the truth, which will bias the assumptions made for power calculation for the sample size. Thus, it is a concern whether sample size calculation based on the information from the pilot study or the investigator's guess or belief is robust or stable.

Chow et al.^[8] studied the instability of power calculation for sample size in terms of its asymptotic bias. For simplicity and illustration purposes, consider the simple case for testing the equality of treatment effect where the primary endpoint is a continuous normal variable. In other words, consider testing the following null hypothesis:

$$H_0 : \mu_T - \mu_C = \delta = 0,$$

where μ_T and μ_C are the means for a test treatment and a (placebo) control, respectively, and δ is a clinically meaningful difference or effect size. Power calculation leads to the following formula for the required sample size per

arm, assuming that $N_T = N_C = N_0$ (see also the work of Chow et al.^[8]):

$$N_0 = \frac{2(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\delta^2},$$

where α and β are the probabilities of committing type I error and type II error, respectively, and σ is the standard deviation of the control. In practice, δ and σ^2 are often estimated by the difference in sample mean ($\hat{\delta} = \hat{\mu}_T - \hat{\mu}_C$) and sample variance (s^2). In other words, σ^2/δ^2 is estimated by $s^2/\hat{\delta}^2$. It can be verified that the leading asymptotic bias term of $E(\hat{\theta} = s^2/\hat{\delta}^2)$ is given by

$$E(\hat{\theta}) - \theta = \frac{3\theta^2}{N_0} \{1 + o(1)\}.$$

Table 2 provides the biases of $\hat{\theta} = s^2/\hat{\delta}^2$ in sample size calculation with various combinations of δ , σ , and θ . As can be seen from the table, the sample size calculation based on the estimates from a small-scale pilot study could be very misleading especially when there is a large variability associated with the observed data. Thus, it is a concern that the estimate of effect size and standard deviation based on a pilot study is often imprecise.

The above discussion justifies that there is a need for sample size reestimation or sample size adjustment during the conduct of clinical trials especially when the underlying assumptions are not met based on the review of accumulated data at interim. Sample size reestimation is often planned in clinical trials utilizing group sequential design with planned interim analyses. In practice, the following sample size adjustment based on the ratio of the initial estimated effect size (E_0) to the observed effect size (E) is usually considered:^[5]

$$N = \min \left\{ N_{\max}, \max \left(N_{\min}, \text{sign}(E_0 E) \left| \frac{E_0}{E} \right|^a N_0 \right) \right\},$$

where N is the sample size after adjustment; $N_{\max}$ and $N_{\min}$ are the maximum (e.g., due to financial and/or other constraints) and minimum (e.g., the sample size for the interim analysis) sample sizes; a is a constant (usually determined based on the review of the interim analysis results), and $\text{sign}(x) = 1$ for $x > 0$; otherwise $\text{sign}(x) = -1$.

Table 2 Biases of sample sizes

δ	σ	$\theta = \sigma^2/\delta^2$	Classic sample	Sample size	
			size N_0	Bias $3\theta^2/N_0$	with bias N
5	10	4	32	1.53	44
	20	16	126	6.12	174
	30	36	283	13.76	391
10	10	1	8	0.38	11
	20	4	32	1.53	44
	30	9	71	3.44	98

Note that the above sample size adjustment reduces to the method proposed by FDA statisticians for normal study endpoint with $a = 2$.

It, however, should be noted that the information obtained for sample size reestimation at interim is still an estimate of the treatment effect. Thus, the instability issue regarding the sample size remains unsolved because there is a variability associated with the observed (estimated) treatment effect at interim.

INTEGRITY OF RANDOMIZATION/BLINDING

In clinical trials, randomization and blinding (e.g., double-blind) are often employed to prevent or minimize the bias (e.g., operational bias) from the assessment of a test treatment under investigation. In randomized and double-blind clinical trials, due to human nature, both patients and the investigator may guess what treatment the patients are receiving. Karlowski et al.^[14] challenged the integrity of the use of randomization and blinding in a randomized, double-blind, placebo-controlled study conducted by the National Institutes of Health. The study was to evaluate the difference between the prophylactic and therapeutic effects of ascorbic acid for the common cold.^[6] After the completion of the study, a questionnaire regarding the knowledge of the treatment assignment was distributed to every subject enrolled in the study (a total of 190 subjects completed the study). Results from the 190 subjects are summarized in Table 3.

The table indicates that there is a high percentage of patients who correctly guessed the treatment assignment they received. Thus, there is a reasonable doubt that the blindness may not be preserved during the study. Thus, “How to test for the integrity of blinding in clinical trials?” is an interesting question.

To address this question, according to Table 3, Chow and Shao^[5] proposed a test for testing the integrity of blinding.^[7] Without loss of generality, consider a parallel design comparing $a \geq 2$ treatments conducted at a single study site. Let A_{ij} be the event that a patient in the j th treatment group who guesses that he/she is in the i th group, $i = 1, \dots, a, a + 1$, where $i = a + 1$ defines the event that a patient whose answer is “do not know” or “does not guess.” Consider the following null hypothesis:

$$H_0 : P(A_{ij}) = P(A_{i1}) \text{ for any } i \text{ and } j.$$

Table 3 Results of patients’ guesses

Patient’s guess	Actual treatment assignment	
	Treatment	Placebo
Treatment	40 (39.6%)	11 (12.4%)
Placebo	12 (11.9%)	39 (43.8%)
Did not respond	49 (48.5%)	39 (43.8%)
Total	101	89

If the above null hypothesis holds, then we claim that the blindness is preserved. Chow and Shao^[5] indicated that H_0 can be tested using the Pearson's chi-square test under the contingency tables constructed based on the observed counts given in Table 3. As a result, a simple calculation gives Pearson's chi-square statistic of 31.3. Thus, the null hypothesis of independence is rejected (p -value < 0.001). Thus, we conclude that the blindness is not preserved and the integrity of blinding is in doubt.

When the integrity of blinding is doubtful, it is suggested that an appropriate adjustment to statistical analysis should be made. In practice, one of the debatable issues is that whether a formal statistical test for the integrity of the blinding should be performed at the end of the clinical trial, although the results are positive or negative. If the results are positive, the sponsor would prefer not to perform the test. However, if the results are negative, the sponsor would argue to perform the test and hopefully to rescue the failed trial. In addition, what action should be taken if a positive clinical trial fails to pass the test for the integrity of the blinding?

CLINICAL STRATEGY FOR ENDPOINT SELECTION

In clinical trials, it is not uncommon to see that we may reach some clinical endpoints but fail to achieve other clinical endpoints. In this case, the selection of clinical endpoint plays an important role for achieving the study objectives with a desired power at a prespecified level of significance. For a given primary clinical endpoint, power calculation and statistical analysis are usually performed based on either the absolute change from baseline or the relative (or percentage) change from baseline. The absolute change from baseline and the relative change from baseline are referred to as derived study endpoints. Based on the original data obtained from the same target patient population, another derived endpoint based on the percentage of patients who show some improvement is often considered. A subject who shows some improvement is considered a responder. The definition of a responder, however, could be based on either the absolute change from baseline or the relative change from baseline of the primary study endpoint.

It should be noted that statistical analysis/interpretation, sample size calculation, and power for different derived

study endpoints are different. Thus, endpoint selection has become very controversial especially when a significant result is observed based on a derived endpoint but not on the other derived endpoints. For example, in a weight reduction study with obesity patient population, statistical analysis based on the absolute change from baseline is often different from that based on the relative (percentage) change from baseline. Besides, power analysis for sample size calculation based on the absolute and relative changes could be very different depending upon what difference is considered of clinical importance. For example, the sample size required in order to achieve the desired power based on the absolute change could be very different from that obtained based on the percentage change, or the percentage of patients who show an improvement based on the absolute change or relative change at α level of significance.

The issue could become more complicated if the intended trial is a non-inferiority trial for establishing non-inferiority of a test treatment to an active control agent. In this case, sample size calculation will also depend upon the selection of non-inferiority margin. Similar to the endpoint selection based on either the absolute change or the relative change, non-inferiority margin could be selected based on either the absolute change or the relative change. As a result, there are eight possible clinical strategies with different combinations of derived study endpoints (i.e., absolute change, relative change, responder analysis with absolute change, and responder analysis with relative change) and non-inferiority margin (absolute change and relative change) for the assessment of the treatment effect (Table 4). To ensure the success of the intended clinical trial, the sponsor will usually carefully evaluate the eight clinical strategies through extensive clinical trial simulation for selecting the most appropriate (derived) endpoint, clinically meaningful difference, and non-inferiority margin during the planning stage of protocol development.

In practice, some clinical strategies may be successful in achieving study objectives with desired power, whereas some strategies may not. These inconsistent results are debatable. The sponsors may choose the strategy to their best interest, whereas the FDA may challenge the sponsors regarding the inconsistent results. In other words, the FDA may ask the sponsors to address the questions that answer (1) which endpoint is telling the truth? and (2) can these study endpoints translate one another since they are derived based on the data collected from the same patient population?

Table 4 Clinical strategy for endpoint selection in non-inferiority trial

Study endpoint	Non-inferiority margin	
	Absolute difference (δ_1)	Relative difference (δ_2)
Absolute change (E_1)	$E_1\delta_1$	$E_1\delta_2$
Relative change (E_2)	$E_2\delta_1$	$E_2\delta_2$
Responder based on absolute change (E_3)	$E_3\delta_1$	$E_3\delta_2$
Responder based on relative change (E_4)	$E_4\delta_1$	$E_4\delta_2$

IMPACT AND SENSITIVITY OF PROTOCOL AMENDMENTS

In clinical trials, protocol amendments are commonly issued after the initiation of a clinical trial due to various reasons such as slow enrollment and/or safety concerns. In practice, before a protocol amendment can be issued, detailed description, rationales, and clinical/statistical justification regarding the changes must be provided to ensure the validity and integrity of the clinical trial. Statistically, it is often a concern that major or significant changes or modifications to study protocol could result in a similar but different patient population. For example, if major or significant changes are made to eligibility (inclusion/exclusion) criteria of the study, the original target patient population may have become a similar but different patient population. This raises the debatable issue regarding the validity and reliability of the statistical inference to be drawn based on the data collected before and after protocol amendment.

To evaluate whether major or significant changes made to the original target patient population have resulted in a similar but different target patient population after the implementation of protocol amendments, denote the original target patient population by (μ, σ) . Also, denote the resultant (actual) patient population after the implementation of a protocol amendment by (μ_1, σ_1) , where $\mu_1 = \mu + \varepsilon$ and $\sigma_1 = C\sigma$ ($C > 0$). The shift in treatment effect of the original target patient population can be characterized by

$$E_1 = \left| \frac{\mu_1}{\sigma_1} \right| = \left| \frac{\mu + \varepsilon}{C\sigma} \right| = |\Delta| \left| \frac{\mu}{\sigma} \right| = |\Delta| E,$$

where E and E_1 are the effect sizes before and after population shift, respectively, and $\Delta = (1 + \varepsilon/\mu)/C$, which is a sensitivity index measuring the change in effect size between the original target patient population and the actual patient population (see, e.g., the works of Chow et al.^[7] and Chow and Shao^[6]). Table 5 provides an evaluation of the impact of protocol amendment in terms of sensitivity index under various scenarios of location shift (i.e., change in ε) and scale shift (change in C). As it can be seen from the table, with the shifts in ε and C (inflation or deflation), the sensitivity index Δ varies from 0.667 to 1.500. It should also be noted that the shift in ε could be offset by the shift in C .

If there is evidence that the mean response is correlated to some covariates, Chow et al.^[3] proposed an alternative approach by considering a model that links the population means and the covariates. In many cases, however, such covariates may not be observable or may not even exist. In this case, Chow et al. approach is not applicable. Alternatively, it is suggested that the sensitivity index Δ be assessed by assuming that there are random shifts in both location and/or scale parameters.

In practice, it is not uncommon to have a number of protocol amendments after the initiation of a clinical trial.

Table 5 Evaluation of sensitivity index

ε/μ (%)	Increase in variability		Decrease in variability	
	C (%)	Δ	C (%)	Δ
-20	100	0.800	—	—
	120	0.667	80	1.000
-10	100	0.900	—	—
	120	0.750	80	1.125
-5	100	0.950	—	—
	120	0.792	80	1.188
0	100	1.000	—	—
	120	0.833	80	1.250
5	100	1.050	—	—
	120	0.875	80	1.313
10	100	1.100	—	—
	120	0.917	80	1.375
20	100	1.200	—	—
	120	1.000	80	1.500

Frequent issuance of protocol amendments may result in a shift in target patient population. Thus, regulatory guidance or regulations on (1) the levels of changes and (2) the number of protocol amendments that are allowed are necessarily developed for maintaining the validity and integrity of the intended study.

MULTIPLICITY IN CLINICAL TRIALS

In clinical trials, multiplicity in clinical trials is usually referred to as multiple inferences that are made simultaneously. The concept of multiplicity could include the comparison of (1) multiple doses (treatments), (2) multiple endpoints, (3) multiple time points, (4) multiple interim analyses, (5) multiple tests of the sample hypothesis, (6) variable/model selection, and (7) subgroup analyses. In practice, it is of interest to the investigators when adjustment for multiplicity should be performed for controlling the overall type I error rate at a prespecified level of significance.

To address this issue, the International Conference on Harmonization (ICH) published a guideline on “Statistical Principles in Clinical Trials” in 1998. This guideline indicates the multiplicity issue for providing substantial evidence in clinical trials. The ICH guideline suggests that data analysis of the clinical trial may necessarily adjust for controlling the overall type I error rate. Moreover, the guideline requires that any adjustment procedure or an explanation (justification) regarding why adjustment is not done be described in detail in the statistical analysis plan. Similarly, the issue of multiplicity is also addressed in the European Agency for the Evaluation of Medicinal

Products (EMA). In 2007, the Committee for Proprietary Medicinal Products published a draft guidance “Points to Consider on Multiplicity Issues in Clinical Trials.” This guidance points out that multiplicity can have a substantial influence on the false positive rate when there is an opportunity to select the most favorable results from two or more analyses. Both the EMA guideline and the ICH guideline recommend stating details of the multiple comparisons procedure in the statistical analysis plan.

In their review article, Westfall and Bretz^[19] indicated that the following are the most commonly seen controversial issues regarding multiplicity in clinical trials:

1. Penalizing for doing more
2. Adjusting α for all possible tests in the trial
3. Testing for family of hypotheses

Penalizing for doing a good job is referred to as the adjustment for multiplicity in dose-finding trials, which often involve several dose groups. For adjusting α for all possible tests, it is overkill to control the α at the pre-specified level because it is not the study objective to show that all of the observed differences (simultaneously) are not by chance alone. In clinical trials, it is debatable for choosing an appropriate family of hypotheses (e.g., primary and secondary study endpoints) for multiplicity adjustment of clinical investigation of the test treatment under study.

Although ICH and EMA did provide some guidance for adjustment of multiplicity, regulations regarding multiplicity adjustment are still not clear. Marcus et al.^[16] indicated that there is no need for multiplicity adjustment for closed testing procedure. Chow^[1] pointed out that one should always check the primary null hypotheses that the trial is powered to test before deciding whether there is a need for multiplicity adjustment.

VALIDITY OF MISSING DATA IMPUTATION

Missing data inevitably occur in clinical trials. When there are a few missing values, one of the approaches is to impute the missing values with their estimates under some valid and appropriate statistical models. Missing data imputation then becomes one of the most debatable issues in clinical trials. The following questions are often asked when there are missing values in clinical trials:

1. Why impute missing values?
2. What methods should be used if we are to impute the missing values?
3. What if when there are a high percentage of missing values?

For the first question, some clinical scientists criticize that missing data imputation actually makes up the data we do not observe. We should not make up data for missing

data as missing data imputation could bias the assessment of treatment effect, and hence missing data imputation does not add much value to the clinical research. For the second question, the method of last observation carried forward (LOCF) is widely used although it has been recognized that the validity of LOCF is questionable. In addition to the method of LOCF, other methods such as the mixed-effects model for repeated measures, generalized estimating equations, and complete case analysis of covariance are often employed in missing data imputation. When there is large proportion of missing values, it is suggested that missing data imputation not be applied. This, however, raises a debatable issue for the determination of the cutoff value for the proportion of missing values for preserving good statistical properties of the statistical inference derived based on the incomplete data set and imputed (complete) data set.

Statistically, one of the concerns for missing data imputation is the potential reduction of power. In clinical trials, it is recognized that missing data imputation may inflate variability due to additional variability associated with the imputed missing data. The inflation of variability will definitely decrease the power. Consequently, the intended clinical trial will not be able to address the scientific/clinical questions asked with desired power. This could be a major concern for regulatory review and approval.

Soon^[18] indicated that management of missing data involves missing data prevention and missing data analysis, which are equally important in handling of missing data. Missing data prevention can be achieved through the enforcement of GSP and GCP during clinical trial process including clinical operations personnel training for data collection. It should be noted that despite the effort, missing data cannot be totally avoided, which may occur due to factors beyond the control of patients, investigators, and clinical project team.

FLEXIBILITY AND FEASIBILITY OF TWO-STAGE ADAPTIVE DESIGN

A seamless trial design is a study design that can address study objectives within a single trial that are normally achieved through the conduct of separate independent trials. A seamless adaptive design is a seamless trial design that fully utilizes data collected from patients before and after the adaptation in the final analysis. A seamless trial design is called a two-stage seamless design if it combines two studies into a single study. Thus, a two-stage (seamless) adaptive design consists of two phases (stages; each stage contains one study), namely, learning or exploratory phase and confirmatory phase. It reduces the lead time between studies (i.e., the first study and the second study). Most importantly, data collected at the learning phase are combined with those obtained at the confirmatory phase for final analysis.

For a two-stage adaptive design, since it combines two independent trials into a single study, the study objectives

Table 6 Classifications of two-stage adaptive designs

Study objectives	Study endpoint	
	S	D
S	SS	SD
D	DS	DD

S = Same; D = different.

and study endpoints at different stages (studies) could be different. Depending upon study objectives and endpoints used, two-stage adaptive trial designs can be classified into four categories of designs as indicated in Table 6.

Thus, we have SS, SD, DS, and DD designs, where SS designs indicate study designs with same objectives and same endpoints at different stages and so on. In clinical trials, different study objectives could be dose finding or treatment selection at the first stage and efficacy confirmation at the second stage. Different study endpoints could include biomarkers, surrogate endpoints, and clinical endpoints with different (shorter) treatment durations at the first stage versus clinical endpoints at the second stage or the same clinical endpoints with different treatment durations.

SS designs are similar to typical group sequential designs with one planned interim analysis. Thus, standard methods for a typical group sequential design can be directly applied to the SS designs.

In this article, our emphasis will be placed on SD, DS, and DD designs. In practice, typical examples for SD, DS, and DD designs include a two-stage phase I/II adaptive design, which is often employed in early clinical development, and a two-stage phase II/III adaptive design, which is usually considered in the late phase of clinical development. For example, for the two-stage phase I/II adaptive design, the study objective at the first stage is for biomarker development and the study objective at the second stage is to establish early efficacy. For a two-stage phase II/III adaptive design, the study objective at the first stage could be for treatment selection, whereas the study objective at the second stage could be for efficacy confirmation.

One of the most debatable issues regarding the flexibility and feasibility of the use of two-stage adaptive design in the early phase and/or late phase of clinical development is the efficiency and effectiveness of the trial design compared to the traditional approach (i.e., conducting two separate trials). To address this issue, Table 7 provides a simple comparison in terms of significance level, power, lead time, and sample size required for achieving a desired power.

As it can be seen from the table, the traditional approach by conducting two separate trials does provide substantial evidence at 1/400 level of significance with a 64% power, whereas the two-stage adaptive design will achieve an 80% power at the 5% (1/20) level of significance. Besides, the use of two-stage adaptive design could reduce the lead time between studies and hence shorten the process of clinical development. In terms of the sample size required, the use of two-stage adaptive design may also reduce the sample

Table 7 Simple comparison

	Two separate trials	Two-stage adaptive design
Significance level	1/20 * 1/20	1/20
Power	0.8 * 0.8	0.8
Lead time	6 month–1 year	Reduced lead time
Sample size	$n = n_1 + n_2$	$m < n?$

Note: n_1 and n_2 are the sample sizes for the two separate trials and m is the sample size for the two-stage adaptive design.

size required for achieving the desired power depending upon the study objectives and the study endpoints used at different stages in the two-stage adaptive trial design. Note that sample size calculation and statistical analysis for SD, DS, and DD designs can be found in the work of Chow and Chang.^[2]

When applying a two-stage adaptive design in clinical trials, one of the most challenging and debatable questions often asked by the regulatory agency such as the FDA is concerned that the overall type I error rate may not be controlled at a prespecified level of significance when (1) O’Brien-Fleming type of boundaries such as Lan-DeMets boundary and (2) additional adaptations are applied.

CHALLENGE OF THE INDEPENDENCE OF DMC

In clinical trials, a DMC is usually established to monitor the validity and integrity of the intended clinical trial. The DMC is independent from the project team, which performs ongoing safety monitoring and/or interim analyses for efficacy. Typically, a DMC consists of experienced physicians and biostatisticians. A charter is necessarily developed not only to outline the activities and functions of the DMC, but also to describe the roles and responsibilities of DMC members.

One of the major concerns regarding an established DMC is the independence of the DMC. The DMC has the authority to perform a review of unblinded data though most DMCs prefer a blinded review of interim data. After the review of interim data, the DMC has the authority to stop the trial early due to safety, futility, and/or efficacy. In clinical trials, an independent DMC ensures the quality, validity, and integrity of the clinical trial. Some sponsors, however, will make every attempt to influence the function and activity of the DMC, which have challenged the independence of the DMC. The following is a summary of some observations that are commonly seen in DMCs across various therapeutic areas:

1. DMC members are selected and appointed by the sponsors.
2. DMC charter is usually developed by the sponsor. The charter is developed without consulting with the

DMC members. DMC members usually do not have the chance to review it until the organizational meeting. As result, the charter is usually approved in a hurry.

3. The trial may have begun to enroll patients prior to the DMC organizational meeting.
4. Those DMC members who have disagreements with the sponsor are replaced prior to the DMC meeting. To avoid selection bias, it is suggested that the reasons for replacing DMC members be documented.
5. DMC members and investigators are from the same organization with administrative reporting relationship.
6. There is no single voice from the DMC.
7. The sponsor issues protocol amendments or modifies randomization schedules without consulting with DMC members.
8. The project statistician and the unblinded (or DMC support) statistician are the same person.

Based on the above observations, it is doubtful that an independent DMC is really independent. In addition, the following two debatable issues have also been raised: First, should the DMC directly communicate with regulatory agencies for any wrong doing in the conduct of the intended clinical trial? Second, can the DMC perform well if a less well-understood adaptive design is used in the intended clinical trial?

CONCLUDING REMARKS

In this article, several commonly encountered statistical controversial issues in clinical research are discussed. In practice, many more debatable issues are still under tremendous discussion among regulatory agencies, academia, and the pharmaceutical industry. These debatable issues include the issue of placebo effect, the impact of baseline adjustment, the selection of non-inferiority margin in active control trials, the use of Bayesian methods in clinical research, the issues in bridging and/or multinational (multiregional) studies, and the potential misuse and abuse of adaptive trial designs (especially those less well-understood design as described in the FDA draft guidance on the adaptive trial design). Most recently, the scientific/statistical issues on the assessment of biosimilarity and interchangeability of biosimilar drug products have received much attention. These issues such as “how similar is considered highly similar?” and “A biosimilar product is expected to produce the same clinical result in *any* given patient.” are debatable from different perspectives of regulatory agencies, academia, and pharmaceutical/biotech industry.

As discussed above, debatable issues are likely encountered in clinical trials. Consequently, the accuracy and reliability of statistical inference on the treatment effect is a concern to the investigator. To address this issue, Shao

and Chow^[17] and Chow and Shao^[4,18] proposed the concept of reproducibility and generalizability of evaluation of the accuracy and reliability of the clinical trials. The reproducibility is defined as the probability of observing positive results (which have achieved statistical significance) of future clinical trials that are conducted under similar experimental conditions given the observed significant positive clinical results. Shao and Chow^[17] suggested considering the probability of reproducibility a monitoring tool for the performance of a test treatment under investigation for regulatory approval. The evaluation of reproducibility provides valuable information that protects patients from the unexpected risk of the test treatment. For example, for a given clinical trial with a relatively low probability of reproducibility, the observed significant positive clinical results may not be reproducible if the clinical shall be repeatedly conducted under similar experimental conditions. To ensure that there is a high reproducibility (say 95%), Chow and Shao^[4] indicated that the p -value for the observed positive results should be less than 0.001 (i.e., the study has to be highly significant).

REFERENCES

1. Chow, S.C. *Controversial Issues in Clinical Trials*; Chapman and Hall/CRC Press, Taylor & Francis: New York, NY, 2011.
2. Chow, S.C.; Chang, M. *Adaptive Design Methods in Clinical Trials*, 2nd Ed.; Chapman and Hall/CRC Press, Taylor & Francis: New York, NY, 2011.
3. Chow, S.C.; Chang, M.; Pong, A. Statistical consideration of adaptive methods in clinical development. *J. Biopharm. Stat.* **2005**, *15*, 575–591.
4. Chow, S.C.; Shao, J. *Statistics in Drug Research*. Marcel Dekker, Inc.: New York, NY, 2002.
5. Chow, S.C.; Shao, J. Analysis of clinical data with breached blindness. *Stat. Med.* **2004**, *23*, 1185–1193.
6. Chow, S.C.; Shao, J. Inference for clinical trials with some protocol amendments. *J. Biopharm. Stat.* **2005**, *15*, 659–666.
7. Chow, S.C.; Shao, J.; Hu, Y.P. Assessing sensitivity and similarity in bridging studies. *J. Biopharm. Stat.* **2002**, *12*, 385–400.
8. Chow, S.C.; Shao, J.; Wang, H. *Sample Size Calculation in Clinical Research*. Chapman and Hall/CRC Press, Taylor & Francis: New York, NY, 2007.
9. Chow, S.C.; Song, F.Y. On controversial issues in clinical research. *J. Clin. Trials* **2015**, *7*, 43–51.
10. Chow, S.C.; Yang, L.Y.; Lu, Y. Some controversial issues in clinical trials. *Drug Inf. J.* **2011**, *45*, 159–170.
11. Cui, L.; Hung, H.M.J.; Wang, S.J. Modification of sample size in group sequential trials. *Biometrics* **1999**, *55*, 853–857.
12. EMEA. Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. EMEA Doc. Ref. CHMP/EWP/2459/02, 2007, October 2007; <http://www.emea.europa.eu/pdfs/human/ewp/245902enadopted.pdf>.

13. ICH. *International Conference on Harmonization Guideline E9 on Statistical Principles for Clinical Trials*, Geneva, Switzerland, 1998.
14. Karlowski, T.R.; Chalmers, T.C.; Frenkel, L.D.; Kapikian, A.Z.; Lewis, T.L.; Lynch, J.M. Ascorbic acid for the common cold: A prophylactic and therapeutic trial. *J. Am. Med. Assoc.* **1975**, *231*, 1038–1042.
15. Lee, Y.; Wang, H.; Chow, S.C. A bootstrap-median approach for stable sample size determination when the specification parameters are estimated from a small pilot study. Unpublished manuscript, 2008.
16. Marcus, R.; Peritz, E.; Gabriel, K.B. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **1976**, *63*, 655–660.
17. Shao, J.; Chow, S.C. Reproducibility probability in clinical trials. *Stat. Med.* **2002**, *21*, 1727–1742.
18. Soon, G. Missing data—Prevention and analysis. *Special issue at J. Biopharm. Stat.* **2009**, *19* (6), 941–1161.
19. Westfall, P.; Bretz, F. Multiplicity in clinical trials. In *Encyclopedia of Biopharmaceutical Statistics*, 3rd Ed.; Chow, S.C., Ed.; Taylor & Francis: New York, NY, 2010, 889–896.

Encyclopedia of Biopharmaceutical Statistics, 4th Edition

Volume II

Clinical Trials (cont.)–Individual
Pages 589—1166

Clinical Trials
(cont.)–Clustered

Companion–Controlled

Correlated–Data
Mining

Data Monitoring–Dose

Dose–Ecologic

ED50–Factorial

False–Good Statistics

Good
Validation–Individual



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Clinical Trials: Interval-censored Failure Time Data

Jianguo Sun and Qingning Zhou

University of Missouri, Columbia, Missouri, U.S.A.

Ding-Geng (Din) Chen

School of Social Work, University of North Carolina, Chapel Hill, North Carolina, U.S.A.;

Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina, U.S.A.; Department of Statistics, University of Pretoria, Pretoria, South Africa

Abstract

Interval-censored failure time data are a general type of failure time or time-to-event data, where the failure time of interest is known or observed only to lie in an interval instead of being observed exactly. A common area where interval-censored data often occur is in medical or public health studies that entail periodic follow-ups such as clinical trials. In this situation, an individual due for the scheduled observations for a clinically observable change in disease status may miss some observations and return with a changed status, thus contributing to an interval-censored time of the occurrence of the change. This entry will overview and discuss some commonly used models and methods for the analysis of interval-censored failure time data as well as some advances in the area.

Keywords: Comparison of survival functions; Interval censoring; Non-parametric estimation; Semiparametric regression; Survival analysis.

INTRODUCTION

Interval-censored failure time or time-to-event data arise in many areas especially in cancer oncology and biopharmaceutics. Their analysis has attracted a great deal of attention. For example, two books have been published on this topic^[1,2] and also several review papers have been published including the study by Ma et al.^[3] and Sun and Li.^[4] In the context of failure time data, interval censoring means that the failure time variable of interest is observed or known only to lie within some intervals or windows instead of being observed exactly.^[1,5] If the observed interval includes only or reduces to a single time point, one obtains the exact failure time.

One field that often produces interval-censored failure time data is medical or public health studies that entail periodic follow-ups such as clinical trials. In these studies, an individual due for the prescheduled observations for a clinically observable change in disease or health status may miss some observations and return with a changed status. As a consequence, we only know that the true event time is greater than the last observation time at which the change has not occurred and less than or equal to the first observation time at which the change has been observed to occur. That is, we only have an interval that contains the real (but unobserved) time of

occurrence of the change. A representative and well-known example of interval-censored data is given by Finkelstein^[6] on comparing Radiotherapy Only treatment with Radiation + Chemotherapy treatment. The study consists of 94 patients with early breast cancer and they were supposed to be monitored every 4–6 months. Among them, 46 received radiation only and 48 received radiation plus chemotherapy. For the study, the event of interest is the time to the first appearance of moderate or severe breast retraction for which only interval-censored data are available. More discussion on this example will be given below. Another field that often produces similar data is the acquired immune deficiency syndrome (AIDS) trials. In this case, one maybe interested in times to AIDS for human immunodeficiency virus (HIV) infected subjects. The determination of AIDS onset is usually based on blood testing which can be performed only periodically but not continuously. In consequence, only interval-censored data are available for AIDS diagnosis times. A similar case is for studies on HIV infection times. If a patient is HIV positive at the beginning of a study, then the HIV infection time is usually determined by a retrospective analysis of his or her medical history. Therefore, we are only able to obtain an interval given by the last HIV negative test date and the first HIV positive test date for the HIV infection time.

In reality, interval censoring can occur in different forms, giving different types of interval-censored failure time data. Among them, an important type is the so-called current status data,^[1] which mean that each subject is observed only once for the occurrence of the failure event of interest. In other words, one does not directly observe the occurrence of the failure event but instead only knows the observation time and whether or not the event has occurred at the time. In consequence, the failure time is either left or right censored. One type of study that usually produce current status data is cross-sectional study on failure events and another type is tumorigenicity study in which the time to tumor onset is usually of interest but not directly observable. In these cases, one only knows or observes the exact value of the observation time, which is usually the death or sacrifice time of the subject. Note that for the first example, current status data occur due to the study design, while for the second case, they are often observed because of the inability of measuring the failure time variable directly and exactly. Sometimes we also refer to current status data as case I interval-censored data and the general case as case II interval-censored data.

Another type of interval-censored data is the so-called doubly censored data,^[7] which mean that the failure time of interest is defined as or represents the time between two related events and the observed data on the times to the occurrences of both events are interval censored. In contrast, the interval-censored data discussed above can be regarded as a special case of such doubly censored data in which one observes the times to the first event exactly and thus can treat them being zero for simplicity. An example of doubly censored data is provided by the AIDS studies discussed above where the variable of interest is AIDS incubation time,^[1] the time from HIV infection to AIDS diagnosis, with both HIV infection time and AIDS diagnosis time being right- or interval-censored.

This entry will discuss some commonly used models and methods for the analysis of interval-censored failure time data as well as some advances in the area. More specifically, after introducing some notation and assumptions, we will first discuss in some detail the analyses of univariate and multivariate interval-censored data, and then briefly review the analyses of some other types of interval-censored data. We will further introduce the available software for the analysis of interval-censored data and some final remarks will be given at last. We remark that this entry is a modified and updated version of Ma et al.^[3]

NOTATION, ASSUMPTIONS, AND LIKELIHOOD FUNCTION

We will first define some notation and describe the likelihood function along with some assumptions. Let T denote the failure time of interest. By saying that T is interval censored, we mean that the only available information for T is

an interval denoted by $I = (L, R]$ such that $T \in I$. Using the notation, we see that current status data correspond to the situation where either $L = 0$ or $R = \infty$. Interval-censored data reduce to right-censored data if $L = R$ or $R = \infty$ for all subjects in the study. Note that a more general way to describe interval-censored data is to assume that the observation on T is given by a group of intervals. However, we will not discuss this general representation as the resulting likelihood functions in both cases have essentially the same structure.

Suppose that there is a failure time study consisting of n independent subjects and let T_i and $I_i = (L_i, R_i]$ be defined as above but associated with subject i . Define $F(t) = P(T \leq t)$, the cumulative distribution function of T , and $S(t) = 1 - F(t)$, the survival function of T . Let $0 = t_0 < t_1 < \dots < t_m = t_{m+1} = \infty$ denote the unique ordered elements of $\{0, \{L_i\}_{i=1}^n, \{R_i\}_{i=1}^n, \infty\}$ and α_{ij} the indicator of the events $(t_{j-1}, t_j] \subseteq I_i$ and $p_j = S(t_{j-1}) - S(t_j)$. Then, for inference about S or about $p = (p_1, \dots, p_{m+1})'$, the likelihood function that is commonly used has the form:

$$L_s(p) = \prod_{i=1}^n [S(L_i) - S(R_i)] = \prod_{i=1}^n \sum_{j=1}^{m+1} \alpha_{ij} p_j \quad (1)$$

In reality, there may exist some covariates denoted by Z and in this case, the likelihood function above becomes:

$$L_s(p|Z_i's) = \prod_{i=1}^n [S(L_i|Z_i) - S(R_i|Z_i)] = \prod_{i=1}^n \sum_{j=1}^{m+1} \alpha_{ij} p_j(Z_i) \quad (2)$$

Of course, the likelihood functions given above come with some assumptions. The most fundamental and important one is perhaps the so-called non-informative interval censoring, which can be described by the following equality:^[1]

$$P(T \leq t | L = l, R = r, L < T \leq R) = P(T \leq t | l < T \leq r) \quad (3)$$

The assumption above essentially says that, except for the fact that T lies between l and r which are the realizations of L and R , the interval $(L, R]$ (or equivalently its endpoints L and R) does not provide any extra information for T . In other words, the probabilistic behavior of T remains the same except that the original sample space $T \geq 0$ is now reduced to $l = L < T \leq R = r$. In the presence of covariates, the assumption (3) becomes:

$$\begin{aligned} P(T \leq t | L = l, R = r, L < T \leq R, Z = z) \\ = P(T \leq t | l < T \leq r, Z = z) \end{aligned} \quad (4)$$

One could also employ different ways to characterize the non-informative interval censoring assumption. For example, one can use a stochastic process to describe the

underlying interval censoring mechanism by assuming that there exists a sequence of observation times or an observation process. Then, the non-informative assumption means that the process is independent of the failure time or process of interest.^[8] In practice, one question of interest is the conditions under which the assumption (3) or (4) holds and for this, the readers are referred to the discussion given in Oller, Gómez, and Calle^[9] among others. In the following, sections all discussion will be based on the assumption (3) or (4) unless specified otherwise.

STATISTICAL ANALYSIS OF UNIVARIATE INTERVAL-CENSORED DATA

In this section, we will consider statistical analysis of univariate interval-censored failure time data and mainly discuss three basic topics—non-parametric estimation of a survival function, non-parametric comparison of survival functions, and regression analysis with the focus on some of the new advances. In addition, some discussion about their applications will be provided.

For non-parametric estimation, one can easily see that given interval-censored data, the problem of finding the non-parametric maximum likelihood estimator (NPMLE) of the survival function S is equivalent to that of maximizing $L_s(p)$ given in Eq. (1) under the constraint that $\sum_{j=1}^{m+1} p_j = 1$ and $p_j \geq 0, j = 1, \dots, m+1$. Obviously, the likelihood function L_s depends on S only through the values $\{S(t_j)\}_{j=1}^m$. Thus, the NPMLE $\hat{S}$ of S can be uniquely determined only over the observed intervals $(t_{j-1}, t_j]$ and the behavior of S within these intervals will be unknown. Several methods have been proposed for maximizing $L_s(p)$ with respect to p . The first and simplest one is perhaps the self-consistency algorithm given by Turnbull.^[10] One drawback of the algorithm is that the convergence can be slow. To address this, Groeneboom and Wellner^[8] developed an iterative convex minorant (ICM) algorithm, which can be seen as an optimized version of the well-known pool-adjacent-violator algorithm for isotonic regression. Another faster algorithm is the EM iterative convex minorant (EMICM) algorithm, a hybrid one that combines the two approaches together. One can find some other procedures such as Bayesian methods and advances on this in the study by Ma et al.^[13]

It is worth to emphasize that all algorithms mentioned above are iterative and for general interval-censored data, the NPMLE $\hat{S}$ of S has no closed form, although for the special case, current status data, a closed form of $\hat{S}$ can be derived using max-min formula for the isotonic regression.^[11] Also, it should be noted that for general interval-censored data, the NPMLE may not be unique and the solutions derived from the algorithms mentioned above may not be the NPMLE. One sufficient condition for the uniqueness of the NPMLE is that the log likelihood is strictly concave. Another important but difficult issue with the NPMLE of S is the consistent variance estimation as

the standard Fisher information matrix approach does not work here. For this, one way is to employ the profile likelihood approach, which is often computationally intensive, and another method is the least-squares approach based on the efficient score function. For detailed discussion on these issues and the asymptotic properties as well as the differences between right-censored and interval-censored data, the readers are referred to the studies by Groeneboom and Wellner^[8] and Zhang and Sun,^[11] among others.

The comparison of different treatments or survival functions is another primary objective in most medical or clinical studies. To formalize the problem, suppose that there are K treatment arms in a clinical study and let $S^{(k)}(t)$ denote the survival function of the k^{th} arm with $k = 1, \dots, K$. Then the problem becomes testing the null hypothesis:

$$H_0: S^{(1)}(t) = S^{(2)}(t) = \dots = S^{(K)}(t) \text{ for all } t \quad (5)$$

In the case of right-censored data, many non-parametric test procedures have been developed and most of them can be classified into two categories—rank-based tests and survival-based tests. The fundamental difference between them is that the former relies on the differences between the estimated hazard functions, while the latter bases the comparison on the differences between the estimated survival functions. Among them, the log-rank test is perhaps the most widely used method and a few of them have been generalized to the case of interval-censored data.^[11] To give a representative of such procedures, in the following, we describe the one proposed by Zhao et al.^[12]

Consider a survival study consisting of n independent subjects with n_k subjects from treatment arm $k, k = 1, \dots, K$. Let the T_i 's and I_i 's be defined as before and D_{k1} and D_{k2} denote the sets of indices of the subjects in treatment arm k whose failure times are observed exactly and interval censored, respectively. Define $n_{k1} = |D_{k1}|$, $n_{k2} = |D_{k2}|$, $N_1 = \sum_{k=1}^K n_{k1}$, and $N_2 = \sum_{k=1}^K n_{k2}$, and let $\hat{S}$ denotes the NPMLE of the common survival function under the hypothesis H_0 . To test the hypothesis H_0 , Zhao et al.^[12] proposed to use the test statistic $U_\xi = (U_\xi^{(1)}, \dots, U_\xi^{(K)})'$, where:

$$U_\xi^{(k)} = \frac{N_1}{n_{k1}} \sum_{i \in D_{k1}} \frac{\xi(\hat{S}(T_i -)) - \xi(\hat{S}(T_i))}{\hat{S}(T_i -) - \hat{S}(T_i)} - \frac{N_2}{n_{k2}} \sum_{i \in D_{k2}} \frac{\xi(\hat{S}(L_i)) - \xi(\hat{S}(R_i))}{\hat{S}(L_i) - \hat{S}(R_i)} \quad (6)$$

and ξ is a positive known function over $(0,1)$. In practice, different ξ can be used and will yield different test statistics. The statistics above were motivated by that given in Peto and Peto,^[13] who discussed similar test statistics with $\xi(x) = x \log(x)$ for right-censored data. If $n_1 = 0$, that is, only interval-censored failure time data are available, the statistics U_ξ reduces to that proposed in the study by Sun, Zhao, and Zhao,^[14] a generalization of the log-rank

test discussed in the study by Peto and Peto.^[13] Under some regularity conditions and H_0 , Zhao et al.^[12] showed that as $n \rightarrow \infty$, $n_{k1}/n \rightarrow p_{k1}$, $n_{k2}/n \rightarrow p_{k2}$ with $0 < p_{kj} < 1$, $U_\xi/\sqrt{n}$ has an asymptotic normal distribution with mean zero. They also gave a consistent estimate of the asymptotic covariance matrix. Several other generalizations of both rank-based and survival-based non-parametric test procedures for right-censored data are available and can be found in Ma et al.^[3] and others.

One needs to perform regression analysis if there exist covariates and one is interested in, for example, quantifying the effect of covariates on the failure time of interest or predicting survival probabilities for new individuals. For this, the first step is usually to specify an appropriate regression model. For failure time data, the most commonly used model is perhaps the proportional hazards model.

$$\lambda(t|Z=z) = \lambda_0(t)e^{\beta'z} \quad (7)$$

with respect to the hazard function of T given covariates $Z = z$. Here $\lambda_0(t)$ denotes an unknown baseline hazard function and β the vector of unknown regression parameters. To fit model (7) to interval-censored data, in addition to the procedures discussed in Sun^[1] and Ma et al.,^[3] Wang et al.^[15] gave a spline-based maximum likelihood approach, where monotone B-splines were used to approximate the baseline cumulative hazard function. Another attractive semiparametric regression model often used in practice is the additive hazards model.

$$\lambda(t|Z=z) = \lambda_0(t) + \beta'z \quad (8)$$

again with respect to the hazard function of T given $Z = z$. It specifies that the effects of covariates are additive rather than multiplicative as in model (7). To fit model (8) to interval-censored data, some multiple imputation-based and estimating equation-based approaches have been developed.^[1,3]

In addition to the models (7) and (8), many other semiparametric models have been employed for regression analysis of interval-censored data. One such model is the proportional odds model expressed as:

$$\log\left(\frac{F(t|z)}{1-F(t|z)}\right) = h(t) + \beta'z \quad (9)$$

with respect to the conditional CDF $F(t|z)$ of the failure time T of interest given $Z = z$. Here, $h(t)$ is an unknown monotone-increasing function, also referred to as the baseline log odds, and β represents a vector of regression parameters as in models (7) and (8). The accelerated failure time model has also been considered for analyzing interval-censored data and it assumes that T and Z have the following relationship:

$$\log(T) = \beta'Z + \varepsilon \quad (10)$$

In the above, β is defined as before and ε is an error term whose distribution is usually unspecified.

It is easy to see that the four semiparametric models (7) to (10) are all specific models in terms of the functional form of the effects of covariates. Sometimes, one may prefer a model that gives more flexibility. One such model that has been investigated for interval-censored data is the linear transformation model that specifies the relationship between the failure time T and the covariate Z as:

$$h(T) = \beta'Z + \varepsilon \quad (11)$$

Here, $h: \mathbb{R}^+ \rightarrow \mathbb{R}$ ($\mathbb{R}$ denotes the real line and $\mathbb{R}^+$ the positive half real line) is an unknown strictly increasing function and the distribution of ε is assumed to be known. It is apparent that the model above gives different models depending on the specification of the distribution of ε and, especially, it includes models (7) and (9) as special cases. For inference about the model based on interval-censored data, some multiple imputation-based and empirical likelihood-based methods have been developed.^[3]

For a practical problem, one may also employ some parametric models such as piecewise exponential models,^[11] if there exists some prior information about the appropriateness of the model. A main advantage of adopting a parametric model is that one can readily apply the maximum likelihood approach for inference. On the other hand, it is well-known that parametric models usually may be too difficult to be verified and are less flexible than semiparametric models.

Besides the issues and models discussed above, some other issues or models regarding regression analysis of interval-censored data have been investigated. One such issue is to fit a cure model to interval-censored data in which it is assumed that there exists a cured subgroup. Another is to analyze interval-censored data with time-dependent covariates or using regression models with time-varying coefficients. Other regression models that have been considered in the literature but not mentioned above include the probit model, Cox-Aalen model, and non-proportional hazards models; and in particular, Zeng, Mao, and Lin^[16] studied a class of semiparametric transformation models that can accommodate time-dependent covariates. Some Bayesian approaches are also available for regression analysis of interval-censored data. Ma et al.^[3] gave more detailed discussions and references on these.

To give an illustration of the methods discussed above, we will briefly discuss the breast cancer study mentioned before. Recall that the objective of the study is to compare Radiotherapy Only treatment (46 patients) to Radiation + Chemotherapy treatment (48 patients) with respect to the time to the first appearance of moderate or severe breast retraction. For the problem, by applying the score test in Finkelstein,^[6] we can obtain a p value of 0.004 for comparing the two groups. If using the generalized log-rank test in Sun et al.,^[14] the p value would be 0.007. Both

methods suggested that adding chemotherapy to radiotherapy significantly increased the hazard of breast retraction.

STATISTICAL ANALYSIS OF MULTIVARIATE INTERVAL-CENSORED DATA

Multivariate interval-censored data arise if a failure time study involves several related failure time variables of interest and each of them suffers interval censoring. It is apparent that the analysis would be straightforward if the variables are independent, and when they are not independent of each other, as is usually the case, one needs to and should employ the inference procedures that can take into account the correlation among the failure time variables. Another difference between univariate and multivariate interval-censored data is that for the latter, a new and unique issue that does not exist for the former is to make inference about the association between the failure time variables. In the following, we will first discuss a couple of basic issues related to the analysis of multivariate interval-censored data and then some advances, with the focus on bivariate interval-censored data.

As with univariate interval-censored data, non-parametric estimation of a survival function is also one of the primary objectives in practice for multivariate interval-censored data. To discuss this, consider a survival study that involves n independent subjects from a homogeneous population with each subject giving rise to two failure times denoted by T_{1i} and T_{2i} , $i = 1, \dots, n$. Let $F(t_1, t_2) = P(T_{1i} \leq t_1, T_{2i} \leq t_2)$ denote their joint cumulative distribution function and suppose that only interval-censored failure time data in the form $\{U_i = (L_{1i}, R_{1i}] \times (L_{2i}, R_{2i}]\}$, $i = 1, \dots, n$ are available, where $(L_{1i}, R_{1i}]$ and $(L_{2i}, R_{2i}]$ represent the intervals to which T_{1i} and T_{2i} belong, respectively. It is easy to see that the observation on each subject could be a point, line segment, or a rectangle. If one treats points as rectangles that degenerate in both dimensions and line segments as rectangles that degenerate in one dimension, then the observed data consist of a collection of n rectangles.

For the determination of the NPMLE of F , note that as with univariate data, it will be a step function. Let $H = \{H_j = (r_{1j}, s_{1j}] \times (r_{2j}, s_{2j}]\}$, $j = 1, \dots, m$ denote the disjoint rectangles that constitute the regions of possible support of the NPMLE of F and define $\alpha_{ij} = I(H_j \cap (L_{1i}, R_{1i}] \times (L_{2i}, R_{2i}])$ and $p_j = F(H_j) = F(s_{1j}, s_{2j}) - F(r_{1j}, s_{2j}) - F(s_{1j}, r_{2j}) + F(r_{1j}, r_{2j})$, $i = 1, \dots, n$; $j = 1, \dots, m$. Then, the likelihood function has the form:

$$L_s(p) = \prod_{i=1}^n F(U_i) = \prod_{i=1}^n \sum_{j=1}^m \alpha_{ij} p_j \quad (12)$$

Again assuming non-informative interval censoring with $p = (p_1, \dots, p_m)'$, and the NPMLE of F can be determined by maximizing Eq. 12 over the p_j 's subject to $p_j \geq 0$ and $\sum_{j=1}^m p_j = 1$. One can easily see that the likelihood functions given Eqs. 1 and 12 actually have the same structure and

thus the algorithms developed for the maximization of Eq. 1 could be employed here for the maximization of Eq. 12 assuming that H is known. However, for a given data set, the determination of H is actually not straightforward, and for this, some algorithms have been developed and can be found in the study by Sun.^[1] It is apparent that the same holds for general multivariate interval-censored data.

As mentioned above, the estimation of the association between related failure time variables is often of interest for multivariate failure time data. To discuss this in the context of bivariate data, let $S_1(t)$ and $S_2(t)$ denote the marginal survival functions of possibly related T_1 and T_2 , respectively, and $S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$ their joint survival function. For the problem, one common way is to assume that $S(t_1, t_2)$ can be expressed by a copula model as $S(t_1, t_2) = C_\alpha(S_1(t_1), S_2(t_2))$, where C_α is a distribution function on the unit square and $\alpha \in \mathbb{R}$ is a global association parameter. One attractive feature of the above expression is its flexibility as it includes as special cases many useful bivariate failure time models such as the Archimedean copula family. Another attractive feature is that under this expression, the marginal distributions do not depend on the choice of the association structure and thus one can model the marginal distributions and the association separately.^[1,3] Of course, by using high-order copula models instead of the bivariate model, one can apply the same approach to general multivariate interval-censored data.

For regression analysis of multivariate interval-censored data, one can apply the same copula model approach as described above; and, for example, some efficient estimation procedures have been developed in the literature for the fitting of models (7) and (9) to bivariate current status data.^[3] Besides the copula model approach, another commonly used approach is to employ frailty or latent variable approaches in which some latent variables are used to characterize the correlation.^[1,3] Among others, one development was given by Zhou, Hu, and Sun,^[17] who considered regression analysis of general bivariate interval-censored data and presented an efficient sieve estimation procedure. A third approach that is often adopted for multivariate failure time data is the marginal model approach that leaves the correlation arbitrary, and several such estimation procedures can be found in the literature for inference about models (8), (9), and (11) based on general multivariate interval-censored data.^[3,4] For all three approaches, of course, one could put them in the Bayesian framework and find some existing work in the literature too.^[1]

It is well-known that multivariate failure time data can be seen as a special case of clustered failure time data and a key feature of the latter is that the cluster size, the number of correlated failure time variables, can differ from cluster to cluster. Thus, the existing methods for multivariate interval-censored data cannot be directly applied to clustered interval-censored data. Ma et al.^[3] described some

literature on the latter, and in particular, some inference procedures have been developed for the analysis under the proportional hazards model (7) with or without a cured subgroup. Also, Chen, Sun, and Xiong^[18] gave a multiple imputation-based estimation procedure under the additive hazards model. Compared to other types of interval-censored data discussed above, however, there exists very limited literature on clustered interval-censored data.

STATISTICAL ANALYSIS OF OTHER INTERVAL-CENSORED DATA

In this section, we will briefly discuss a few other types of interval-censored data that may occur in failure time studies, including competing risks interval-censored data, informatively interval-censored data, and doubly censored data.

Competing risks failure time data arise when there exist several related failure causes or types.^[5] The underlying structure behind them is actually similar to that behind multivariate failure time data. A key difference between the two types of data, also a key feature of the former, is that one observes only one failure type or one of several related underlying failure time variables. For a patient with several diseases, for example, one can observe the death time caused by one disease but not the times due to other diseases. Some work and references on interval-censored competing risk data can be found in Ma et al.,^[3] but as with clustered interval-censored data, there exists only limited literature on the topic.

In the previous sections, it has been assumed that the mechanism behind interval censoring is independent of the underlying failure time or process. In some applications, however, this assumption may not hold. For example, in medical or health follow-up studies, the patients with more serious conditions or diseases tend to see doctors more frequently. The situation where the censoring mechanism is related to the failure time is usually referred to as informative or dependent censoring.^[1] In addition to the literature described in Ma et al.,^[3] some more references on the topic include Ma, Hu, and Sun,^[19,20] who discussed the inference about the proportional hazards model, and Zhao et al.,^[21,22] who investigated the same estimation procedure but under the additive hazards model. Also Wang, Zhao, and Sun^[23] considered the situation where a sequence of observation times for each subject is available and the failure time of interest follows model (7).

As mentioned above, the interval-censored data discussed above can be regarded as a special case of doubly censored data, where the failure time of interest represents the elapsed time between two related events and the observations on both events suffer interval censoring. For the detailed discussion and general development on this type of failure time data, readers are referred to the study by Sun,^[1] who devoted one chapter for the topic, and the

study by Ma et al.,^[3] who also provided some discussion and literature. Furthermore, Wang et al.^[24,25] considered the situation where in addition to interval censoring, there may also exist left truncation and provided some inference procedures for the additive hazards model.

SOFTWARE FOR THE ANALYSIS OF INTERVAL-CENSORED DATA

It is well-known that software development is important for both the implementation of the developed methods and practitioners. Although there is no commercially available statistical software that is specifically designed for an extensive coverage of or covering most of the existing inference procedures for interval-censored data, there exist some packages or functions in R, SAS, or SPLUS that can be applied to the analysis of interval-censored failure time data.

In R, for example, the packages *interval* and *Icens* can be used to find the NPMLE of the survival and distribution function, and *interval* can also be used to perform two-sample weighted log-rank tests and k-sample tests. In addition, parametric methods are provided in the *survival* package using the *survreg* function and for semiparametric estimation, one can use *Icens* function in the *Epi* package. The *intcox* package can also be used for semiparametric estimation under the proportional hazards model. However, this package does not directly provide standard error estimation for the estimated regression parameters and the bootstrapping approach is one of the remedies for this drawback. Furthermore, the package *glrt* provides functions to perform four generalized log-rank tests including those given in the studies by Sun, Zhao, and Zhao^[14] and Zhao et al.^[12] as well as Finkelstein's score test. In addition, the packages *ICBayes*, *ICsurv*, *coxinterval*, and *dynsurv* were developed for semiparametric analysis of interval-censored data. *ICBayes* fits several Bayesian semiparametric regression models,^[26] *ICsurv* implements the expectation and maximization (EM) algorithm developed by Wang et al.^[15] for finding the maximum likelihood estimation (MLE) under the proportional hazards model, *coxinterval* provides the fit to the Cox-Aalen model,^[27] and *dynsurv* includes functions to fit Bayesian Cox model with time-independent, time-varying, or dynamic coefficients. One can find some details and examples on the use of R for the analysis of interval-censored data in the studies by Fay and Shaw^[28] and Gómez et al.,^[29] among others. In SPLUS, the function *kaplanMeier* can be used to compute the Turnbull estimator.

In SAS, the analysis of interval-censored data can be performed by, among others, the *LIFEREG* and *RELIABILITY* procedures, and the macros *%EMICM*, *%ICSTEST*, *%ICE*, *%INTCEN_\$SURV_P*, and *%TURNBULL_INT*. Some descriptions of them can be found in SAS website or in the study by Ma et al.^[3] the *ICLIFETEST* procedure

was specifically developed for analyzing interval-censored data and it offers a set of non-parametric statistical methods (e.g., EMICM, Turnbull's algorithm, and ICM) of estimating survival functions and statistical tests (e.g., weighted generalized log-rank tests with a variety of weight functions) for comparing survival functions of different groups. The methods used by the macros %EMICM, %ICSTEST, and %ICE have been incorporated into the ICLIFETEST procedure with minor modifications. Another SAS procedure is the ICPHREG procedure that can be used to fit the proportional hazards models with a variety of configurations in terms of the baseline hazard function (e.g., the piecewise constant model and the cubic spline model) by MLE based on the full likelihood.

CONCLUSION

This entry provided a relatively comprehensive overview of the issues and approaches in the analysis of interval-censored failure time data. It is apparent that given the page limit, there exist many issues that have been investigated but not discussed fully or not discussed at all, in the previous sections. These include the analysis of interval-censored data with missing covariates and/or truncation and the analysis of mixed interval-censored data. For more complete references, the readers are referred to the book by Sun^[1] and the reviews by Ma et al.^[3] and Sun.^[7] In addition, methodologically, there are many open questions in this area. Examples include but are not limited to model checking techniques and joint modeling of longitudinal and interval-censored data. Some of the methods discussed in the previous sections also need proper theoretical justifications. A major difficulty is the lack of basic tools as simple and elegant as the partial likelihood and the martingale theory for right-censored data. One can find some comprehensive theoretical work in the studies by Groeneboom and Wellner^[8] and Huang and Wellner,^[30] which mainly rely on complicated empirical processes and the optimization theory and are difficult to generalize.

REFERENCES

1. Sun, J. *The Statistical Analysis of Interval-censored Failure Time Data*; Springer: New York, 2006.
2. Chen, D.G.; Sun, J.; Peace, K.E. *Interval-Censored Time-to-Event Data: Methods and Applications*; Chapman and Hall/CRC: Boca Raton, FL, 2012.
3. Ma, L.; Feng, Y.; Chen, D.G.; Sun, J. Interval-censored time-to-event data and their applications in clinical trials. In *Clinical Trial Biostatistics and Biopharmaceutical Applications*; Young, W.R.; Chen, D.G., Eds.; Chapman and Hall/CRC: Boca Raton, FL, 2014; 307–333.
4. Sun, J.; Li, J. Interval censoring. In *Handbook of Survival Analysis*; Klein, J., van Houwelingen, H., Ibrahim, J., Scheike, T., Eds.; Chapman and Hall/CRC: Boca Raton, FL, 2013; 369–390.
5. Kalbfleisch, J.D.; Prentice, R.L. *The Statistical Analysis of Failure Time Data*; Wiley: New York, 2002.
6. Finkelstein, D.M. A proportional hazards model for interval-censored failure time data. *Biometrics*. **1986**, 42 (4), 845–854.
7. Sun, J. Statistical analysis of doubly interval-censored failure time data. In *Handbook of Statistics: Advances in Survival Analysis*; Balakrishnan, N., Rao, C.R., Eds.; Elsevier B. V.: Amsterdam, 2003; 23, 105–122.
8. Groeneboom, P.; Wellner, J.A. *Information Bounds and Nonparametric Maximum Likelihood Estimation*. DMV Seminar, Band 19; Birkhauser Basel, 1992.
9. Oller, R.; ómez, G.; Calle, M.L. Interval censoring: Identifiability and the constant-sum property. *Biometrika*. **2007**, 94 (1), 61–70.
10. Turnbull, B.W. The empirical distribution with arbitrarily grouped censored and truncated data. *J. R. Stat. Soc. B*. **1976**, 38 (3), 290–295.
11. Zhang, Z.; Sun, J. Interval censoring. *Stat. Methods Med. Res.* **2010**, 19 (1), 53–70.
12. Zhao, X.; Zhao, Q.; Sun, J.; Kim, J.S. Generalized log-rank tests for partly interval-censored failure time data. *Biomed. J.* **2008**, 50 (3), 375–385.
13. Peto, R.; Peto, J. Asymptotically efficient rank invariant test procedures. *J. R. Stat. Soc. A*. **1972**, 135 (2), 185–207.
14. Sun, J.; Zhao, Q.; Zhao, X. Generalized log-rank tests for interval-censored failure time data. *Scand. J. Stat.* **2005**, 32 (1), 49–57.
15. Wang, L.; McMahan, C.S.; Hudgens, M.G.; Qureshi, Z.P. A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics*. **2016**, 72 (1), 222–231.
16. Zeng, D.; Mao, L.; Lin, D.Y. Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika*. **2016**, 103 (2), 253–271.
17. Zhou, Q.; Hu, T.; Sun, J. A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *J. Am. Stat. Assoc.* **2016**, 112 (518), 664–672.
18. Chen, L.; Sun, J.; Xiong, C. A multiple imputation approach to the analysis of clustered interval-censored failure time data with the additive hazards model. *Comput. Stat. Data Anal.* **2016**, 103, 242–249.
19. Ma, L.; Hu, T.; Sun, J. Sieve maximum likelihood regression analysis of dependent current status data. *Biometrika*. **2015**, 102 (3), 731–738.
20. Ma, L.; Hu, T.; Sun, J. Cox regression analysis of dependent interval-censored failure time data. *Comput. Stat. Data Anal.* **2016**, 103, 79–90.
21. Zhao, S.; Hu, T.; Ma, L.; Wang, P.; Sun, J. Regression analysis of interval-censored failure time data with the additive hazards model in the presence of informative censoring. *Stat. Interface*. **2015**, 8 (3), 367–377.
22. Zhao, S.; Hu, T.; Ma, L.; Wang, P.; Sun, J. Regression analysis of informative current status data with additive hazards model. *Lifetime Data Anal.* **2015**, 21 (2), 241–258.
23. Wang, P.; Zhao, H.; Sun, J. Regression analysis of case K interval-censored failure time data in the presence of informative censoring. *Biometrics*. **2016**, 72 (4), 1103–1112.

24. Wang, P.; Tong, X.; Zhao, S.; Sun, J. *Efficient estimation for the additive hazards model in the presence of left-truncation and interval censoring*. Stat. Interface. **2015**, 8 (3), 391–402.
25. Wang, P.; Tong, X.; Zhao, S.; Sun, J. *Regression analysis of left-truncated and case I interval-censored data with the additive hazards model*. Commun. Stat. Theor. M. **2015**, 44 (8), 1537–1551.
26. Lin, X.; Cai, B.; Wang, L.; Zhang, Z. *A Bayesian proportional hazards model for general interval-censored data*. Lifetime Data Anal. **2015**, 21 (3), 470–490.
27. Boruvka, A.; Cook, R.J. *A Cox-Aalen model for interval-censored data*. Scand. J. Stat. **2015**, 42 (2), 414–426.
28. Fay, M.P.; Shaw, P.A. *Exact and asymptotic weighted log-rank tests for interval-censored data: The interval R package*. J. Stat. Softw. **2010**, 36 (2), 1–34.
29. Gómez, G.; Oller, R.; Calle, M.L.; Langohr, K. *Tutorial on methods for interval-censored data and their implementation in R*. Stat. Model. **2009**, 9 (4), 259–297.
30. Huang, J.; Wellner, J.A. *Interval-censored survival data: A review of recent progress*. In *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*; Lin, D.Y.; Fleming, T., Eds.; Springer: New York, 1997, 123–169.

Clinical Trials: Investigating Quality-of-Life

Patricia B. Cerrito

University of Louisville, Louisville, Kentucky, U.S.A.

Abstract

Medications that are potent enough to treat cancer can cause many serious adverse effects. This entry investigates the question of at what point is the toxicity of a treatment the cause of more suffering than life without it. This question is addressed by surveying patients about their quality of life while undergoing continuous treatment.

INTRODUCTION

In many cases, treatment for cancer cannot cure the disease but can only prolong life. However, treatment comes at a cost. Medications that are potent enough to treat cancer can cause many serious adverse effects. At what point is the toxicity of a drug sufficiently high enough to cause greater suffering than prolonged life? What is the risk of the toxic effect compared to the corresponding drug benefit?

To answer these questions, quality of life needs to be examined along with treatment effects in all drug studies. Because the assessment of the quality of life is subjective, patients need to be routinely consulted about their perception of treatment benefits. Measuring the patient's perception of the benefits of treatment will require the use of quality-of-life (QOL) surveys, either those that have been developed previously or one uniquely designed for a particular drug study. In this paper, the use of both types of surveys are discussed. Also, the use of QOL information to adjust patient outcomes is discussed.

EXISTING QUALITY-OF-LIFE INSTRUMENTS

One generalized QOL instrument, SF-36, was developed by the Rand Corporation.^[1,2] This instrument was designed to be relatively generic and relatively simple to administer and score. The SF-36 contains a total of 36 questions, and the Rand Corporation provides general permission for investigators to use the instrument and posts the scoring instructions on its website. The European Organization for Research and Treatment of Cancer (EORTC) has developed a QOL instrument targeted more specifically for cancer patients and containing modifications for different types of cancer. The instrument consists of 30 items divided among five functional subscales (physical, role, cognitive, emotional, and social), three symptom subscales (fatigue, pain, and nausea), as well as global health and quality

scales.^[3,4] Other instruments exist, but these two are the ones primarily used. The question of whether to use generic instruments, such as the SF-36, or specific instruments, such as the EORTC questionnaire, remains open.^[4–7] The advantage of using more generic instruments is that tests have been conducted to demonstrate their validity.^[8–10]

DEVELOPING QUALITY-OF-LIFE INSTRUMENTS

A specific QOL instrument can be developed for a specific clinical trial, one that is closely tailored to the aims and outcomes of the trial. However, care must be taken to ensure the validity of the instrument. The first step in the developmental process is to decide what important characteristics need to be measured. It is also important to determine any constraints. The major requirements of a QOL instrument are reliability, validity, and responsiveness. Generally, reliability is defined as freedom from measurement error. Validity is defined as the extent to which the survey measures what it is supposed to measure. Responsiveness is the extent to which the survey measures change over time.^[3] There are also some practical issues, including ease of use. For example, compliance with completing a QOL form decreases with the length of the form. Moreover, any instrument should focus on three different domains: physical, psychological (mental, emotional), and social. A brief outline is given in Suhonen et al. (Table 1).^[11]

Content validity suggests that the items in the instrument are meaningful to the quality of life to be studied. The language in the instrument must be clear, precise, and should result in clear subject responses. If there is any garbled language, it must be changed before the instrument is used. The instrument has construct validity if the questions represent specific topics that should be included. For example, suppose that there are 25 questions in the instrument. The questions should relate to several subtopics such as

Table 1 Development of Quality-of-Life Instrument

Purpose	Method
Examine concepts	Literature review
Develop instrument questions	Construction of instrument
Assess content validity	Expert review
Assess clarity of items	Expert review
Test instrument	Pilot sample
Test reliability and validity	Statistical methods
Test stability of instrument	Test–retest
Assess construct validity	Principal component analysis

physical functioning, mental functioning, and social functioning. Before any statistical analysis, the investigator should have an idea as to which questions relate to which topic. A principal component analysis (or factor analysis) is a statistical method for grouping such responses to specific subtopics. The statistical grouping should match the investigator's expectations. Although construct validity is a complex issue, it can be somewhat routinely investigated.^[12–14] Reliability indicates that similar questions receive similar responses. An instrument can be reliable without being valid, and care must be taken to examine both issues.

For a recent cancer trial, the author of this article developed a QOL instrument. The principal investigator (PI) of the study wanted to restrict the form to one page because the coinvestigators did not want to use a form that was longer.

The form was primarily restricted to physical and emotional symptoms, asking the patients to note the presence or absence of such symptoms. The purpose of the instrument was to have patient input in the definition of utilities that are used in adjustments to survival analysis. The instrument was reviewed by all participating clinicians before its use in the study. The questions were based upon potential adverse effects from using the drug, as identified in previous studies. In particular, previous studies demonstrated the acute adverse effects of chills, malaise, and arthralgias, and chronic adverse effects of fatigue, anorexia, weight loss, and depression.^[15] The documented depression reached the clinical criteria for major depression in some patients.

Statistical analysis to demonstrate the validity of the instrument was performed on 1200 patients who had up to 3 years of follow-up, approximately halfway to completion of the study (Table 2). There are a number of values that are statistically significant between treatment and control groups. The percentages listed in Table 2 indicate the number in each category identifying a symptom. Note that with 1200 patients, almost all the factors are statistically significant. The data were split at the 12-month interval because treatment was given for 1 full year. Interestingly, a number of the symptoms persisted beyond the treatment period.

Patients were also asked to identify, on a scale of 1 to 10, their general quality of life and general physical condition. For the first 12 months, there was a slight difference between treatment and control (Figs. 1 and 2). There was also an improvement in the treatment group before and

Table 2 Examination of Quality-of-Life Factors

Symptom	0–12 months			Over 12 months		
	No drug	Drug	P-value	No drug	Drug	P-value
Weak	22 (17%)	55 (50%)	<0.0001	4 (6%)	8 (15%)	0.1417
Irritable	11 (9%)	27 (25%)	0.0009	6 (9%)	9 (16%)	0.2658
Tired	31 (24%)	86 (78%)	<0.0001	12 (19%)	21 (38%)	0.0209
Depressed	13 (10%)	33 (30%)	0.0001	3 (5%)	9 (16%)	0.0375
Forgetful	8 (6%)	29 (26%)	<0.0001	3 (5%)	10 (18%)	0.0202
Feverish	3 (2%)	30 (27%)	<0.0001	1 (2%)	3 (5%)	0.2469
Anxious	13 (10%)	26 (24%)	0.0055	0 (0%)	24 (22%)	<0.0001
Confused	0 (0%)	24 (22%)	<0.0001	0 (0%)	5 (9%)	0.0145
Hungry	3 (2%)	16 (14%)	0.0006	4 (6%)	8 (14%)	0.1417
Sad	3 (2%)	25 (23%)	<0.0001	1 (2%)	6 (11%)	0.0325
Nervous	10 (8%)	27 (24%)	0.0004	1 (2%)	8 (14%)	0.0082
Nausea	5 (4%)	27 (24%)	<0.0001	2 (3%)	1 (2%)	0.6405
Constipation	0 (0%)	13 (12%)	<0.0001	2 (3%)	2 (4%)	0.89
Diarrhea	6 (5%)	26 (24%)	<0.0001	2 (3%)	4 (7%)	0.3121
Headaches	11 (9%)	34 (31%)	<0.0001	2 (3%)	7 (13%)	0.0512
Muscle pain	25 (20%)	56 (51%)	<0.0001	5 (8%)	15 (27%)	0.0052
Shortness of breath	9 (7%)	19 (17%)	0.0154	0 (0%)	7 (13%)	0.0035
Swelling	22 (17%)	16 (14%)	0.5611	1 (2%)	2 (4%)	0.4806
Loss of appetite	1 (1%)	31 (28%)	<0.0001	0 (0%)	1 (2%)	0.2825

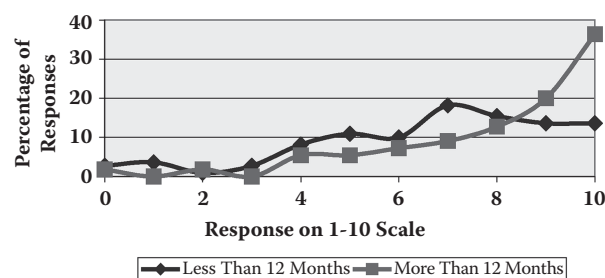


Fig. 1 Quality of life changes before and after the 12-month mark for patients in treatment arm

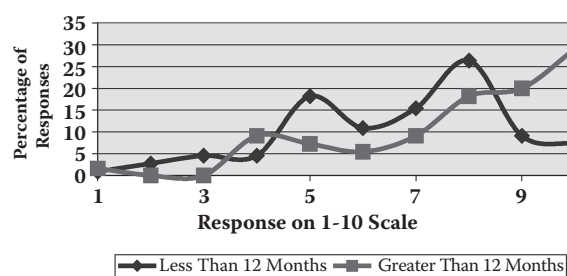


Fig. 2 Physical condition before and after the 12-month mark for control arm

after the termination of treatment. A much higher proportion of satisfaction with QOL (29% vs. 7%) in the control group indicates the highest level of physical condition after treatment. A similar increase (36% vs. 13%) in answers occurred in the highest level of overall quality of life. The proportions remained relatively stationary for the control group. However, there was a difference in responses in the treatment group in the 12 months of treatment compared with the time period after treatment.

Similarly, drug treatment was shown to interfere with a patient's employment. Patients taking treatment reported greater absenteeism from work (Fig. 3). The patients found that the drug generally interfered with their daily routine (Fig. 4).

Factor analysis was performed to determine how the variables interacted with each other (Table 3). Note that the use of the drug is strongly related to physical variables such as weakness and fever but not as strongly related to mental variables. Similarly, time of follow-up is strongly

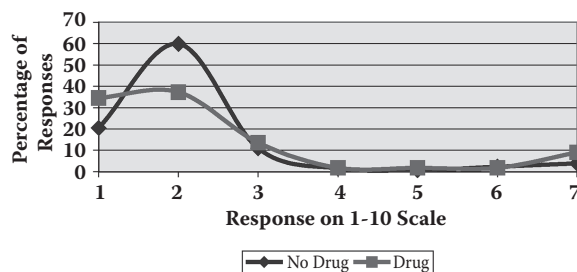


Fig. 3 Time absent from work at less than 12 months

Table 3 Factor Loadings of Quality-of-Life Variables

Factor variable	Loading
1	
Irritable	0.55
Depressed	0.6
Forgetful	0.5
Anxious	0.71
Confused	0.61
Hungry	0.44
Sad	0.72
Nervous	0.6
2	
Weak	0.45
Feverish	0.58
Constipation	0.6
Diarrhea	0.49
Loss of appetite	0.57
Drug group	0.57
3	
Percent of routine	0.41
Tired	0.52
Muscle pain	0.55
Shortness of breath	0.57
Swelling	0.58
4	
Weight change	0.56
Physical condition	0.84
Quality of life	0.84
5	
Follow up month	0.62
Absent from work	0.58
Unable to do routine	0.5

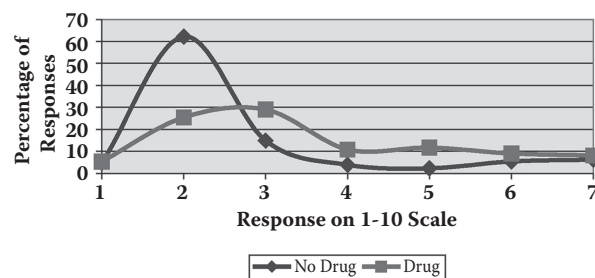


Fig. 4 Interfering with routine at less than 12 months

related to the percentage of time a patient is absent from work and unable to perform a normal routine.

Social status was examined, in part, through questions concerning life changes such as marriage, divorce, or retirement (Fig. 5).

There was actually a higher proportion of patients from the control group who applied for social security and Medicare than from the treatment group. However, the proportions were so small that none were statistically significant.

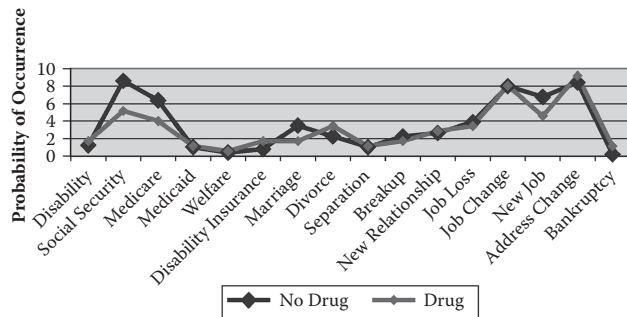


Fig. 5 Comparison of life factors

Q-TWIST ANALYSIS

The general Q-twist method is a quality-adjusted survival analysis that integrates quality of life considerations into treatment comparisons.^[16] The survival time is adjusted according to the quality of life experienced. The equation is defined by^[17]

$$\text{Q-twist} = (u_{\text{TOX} \times \text{TOX}}) + \text{TWiST} + (u_{\text{REL} \times \text{REL}})$$

where TOX is the time with treatment toxic side effects; TWiST is the time without symptoms of disease and toxic side effects; REL is the time following recurrence; u_{TOX} is utility, or value of TOX relative to TWiST; u_{REL} = utility, or value of REL relative to TWiST.

The average duration of TOX, TWiST, and REL must be estimated from clinical data. The value assigned depends on the level of toxicity. Often, the toxicity definitions used are defined only by physicians.^[15]

No input from patients is used to define patient utilities.^[18–20] It is preferable to design a measurement of quality of life prior to enrolling patients, rather than to determine the level of toxicity once the patient data have been collected.

The Q-twist analysis is a form of quality-adjusted life years (QALYs). In this type of analysis, decisions are made based upon preferences. For example, how many years of life (x) is a patient willing to give up to have some years in good health (y). The primary problem with QALYs analysis is that these preferences are usually examined as hypothetical questions, and it is questionable just how relevant the responses are to actual situations. Similarly, for the Q-Twist, the major problem is in defining the utilities u_{TOX} and u_{REL} . It is clear that even minor changes in TOX value can result in substantially different outcomes and decisions.^[3] Therefore care must be taken to validate the utility weights used in the equation. The weights are defined using the factor loadings in Table 3.

The current standard practice is to vary the value of the utility weight to determine where changes in outcomes can occur. For example, consider the survival difference between control and treatment (Fig. 6).

Consider the adjustments to the survival. Assume that the toxic effect is chronic and continues through the duration of the time on medication. Assume that the utility is equal to 0.5 when compared to the time off the drug. The adjusted survival curves are given in Fig. 7, along with additional measures of toxicity. Note that, although the end result is closer, the adjusted survival still indicates the benefit of treatment.

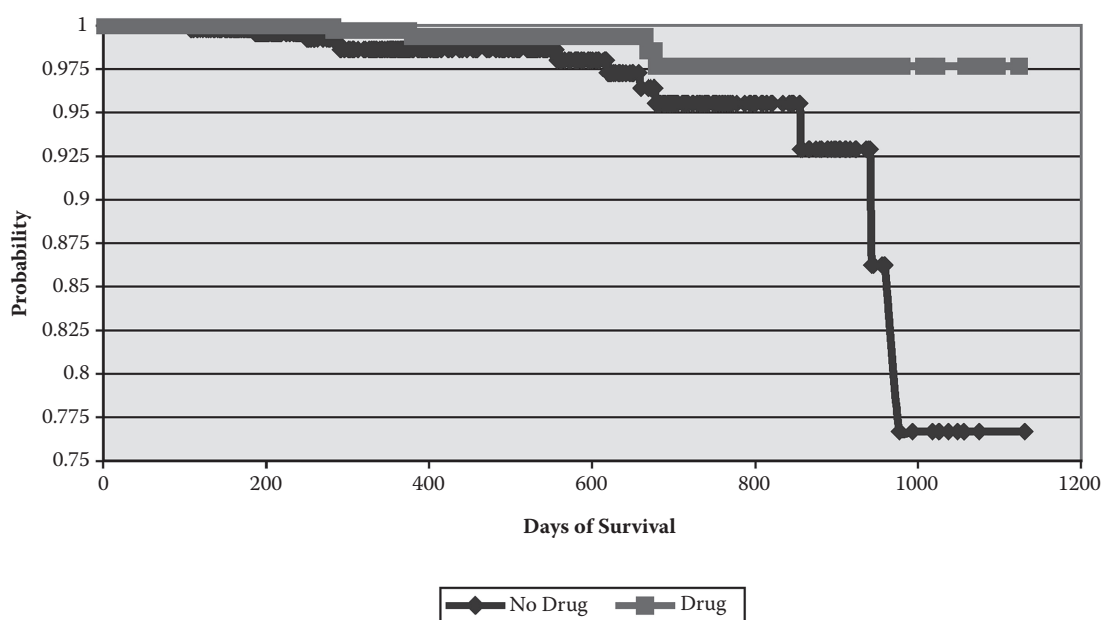


Fig. 6 Survival differences between control and treatment

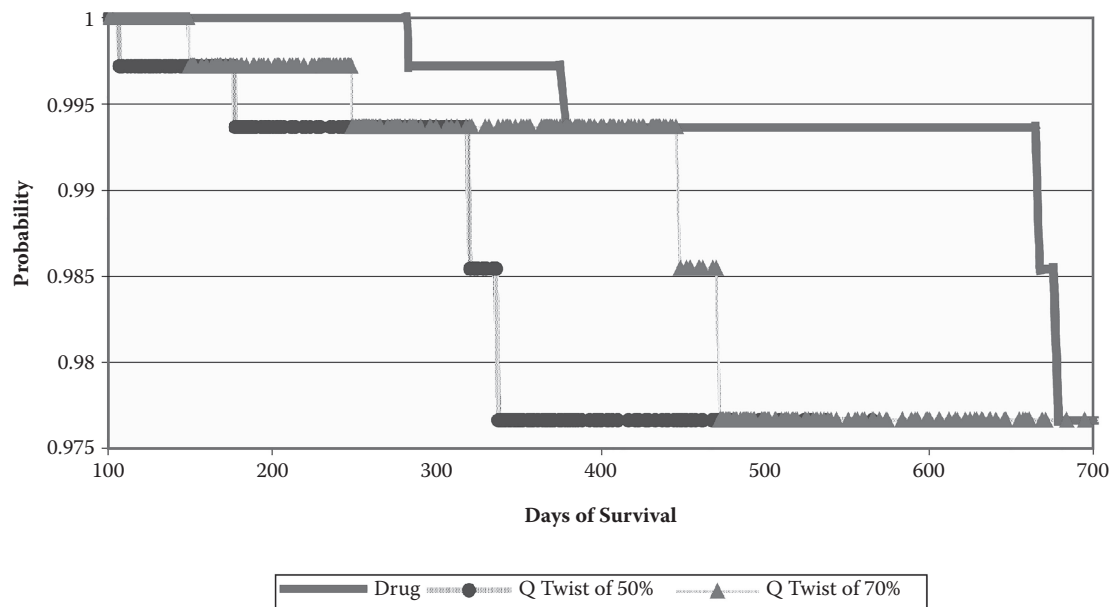


Fig. 7 Survival differences between control and treatment using sensitivity analysis of Q-twist adjustments

CONCLUSION

Because of the toxicity of treatments for various forms of cancer, it is helpful to determine whether the treatment adds or detracts from the overall quality of life. It is possible to use preexisting questionnaires that measure QOL issues. However, it is also possible to develop a valid, reasonable tool that is specific to a clinical trial. Care must be taken to ensure that the tool satisfies construct validity. If possible, both a new and an established tool can be used simultaneously to demonstrate convergent validity.

REFERENCES

1. Corporation, R. *SF-36*; 2001 Rand Corporation. Available at: <http://www.rand.org/health/surveys/sf36item/>.
2. Myles, P.S.; Hunt, J.O.; Fletcher, H.; Solly, R.; Woodward, D.; Kelly, S. Relation between quality of recovery in hospital and quality of life at 3 months after cardiac surgery. *Anesthesiology* **2001**, *95* (4), 862–867.
3. Langenhoff, B.; Krabbe, P.; Wobbes, T.; Ruers, T. Quality of life as an outcome measure in surgical oncology. *British J. Surg.* **2001**, *88* (5), 643–652.
4. Ringash, J.; Bezjak, A. A structured review of quality of life instruments for head and neck cancer patients. *Head Neck*. **2001**, *23* (3), 201–213.
5. Batista-Miranda, J.E.; Diez, M.D.; Bertran, P.A.; Vilavicencio, H. Quality-of-life assessment in patients with benign prostatic hyperplasia: Effects of various interventions. *Pharmacoeconomics* **2001**, *19* (11), 1079–1090.
6. Kremer, B.; Klimek, L.; Bullinger, M.; Mosges, R. Generic or disease-specific quality of life scales to characterize health status in allergic rhinitis? *Allergy* **2001**, *56* (10), 957–963.
7. Bernheim, J.L. How to get serious answers to the serious question: “How have you been?”: Subjective quality of life (QOL) as an individual experiential emergent construct. *Bioethics* **1999**, *13* (3–4), 272–287.
8. Findler, M.; Cantor, J.; Haddad, L.; Gordon, W.; Ashman, T. The reliability and validity of the SF-36 health survey questionnaire for use with individuals with traumatic brain injury. *Brain Inj.* **2001**, *15* (8), 715–723.
9. Kosinski, M.; Keller, S.D.; Ware, J.E., Jr.; Hatoum, H.T.; Kong, S.X. The SF-36 Health Survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: Relative validity of scales in relation to clinical measures of arthritis severity. *Med. Care* **1999**, *37* (5 Suppl.), MS23–MS39.
10. Andresen, E.M.; Gravitt, G.W.; Aydelotte, M.E.; Podgorski, C.A. Limitations of the SF-36 in a sample of nursing home residents. *Age Ageing* **1999**, *28* (6), 562–566.
11. Suhonen, R.; Valimaki, M.; Katajisto, J. Developing and testing an instrument for the measurement of individual care. *J. Adv. Nurs.* **2000**, *32* (5), 1253–1263.
12. Marinus, J.; Ramaker, C.; van Hilten, J.J.; Stiggelbout, A.M. Health related quality of life in Parkinson’s disease: A systematic review of disease specific instruments. *J. Neurol., Neurosurg. Psychiatr.* **2002**, *72* (2), 241–248.
13. Carlsson, M.; Hamrin, E.; Lindqvist, R. Psychometric assessment of the Life Satisfaction Questionnaire (LSQ) and a comparison of a randomised sample of Swedish women and those suffering from breast cancer. *Qual. Life Res.* **1999**, *8* (3), 245–253.
14. van Wegberg, B.; Bacchi, M.; Heusser, P. The cognitive-spiritual dimension—An important addition to the assessment of quality of life: Validation of a questionnaire

(SELT-M) in patients with advanced cancer. *Ann. Oncol.* **1998**, 9 (10), 1091–1096.

15. Borden, E.C.; Parkinson, D. Interferons: Effectiveness, toxicities, and costs. *Ann. Intern. Med.* **1996**, 125 (7), 614–615.
16. Tsiatis, A.A.P. Estimating the distribution of quality-adjusted life with censored data. *Am. Heart J.* April **2000**, 139 (4), S177–S181.
17. Schwartz, C.E.; Cole, B.F.; Gelber, R.D. Measuring patient-centered outcomes in neurologic disease. Extending the Q-TWiST method. *Arch. Neurol.* **1995**, 52 (8), 754–762.
18. Levy, V.; Porcher, R.; Delabarre, F., et al. Evaluating treatment strategies in chronic lymphocytic leukemia: Use of quality-adjusted survival analysis. *J. Clin. Epidemiol.* **2001**, 54 (7), 747–754.
19. Wirt, D.P.; Giles, F.J.; Oken, M.M.; Solal-Celigny, P.; Beck, J.R. Cost-effectiveness of interferon alfa-2b added to chemotherapy for high-tumor-burden follicular non-Hodgkin's lymphoma. *Leuk. Lymphoma* **2001**, 40 (5, 6), 565–579.
20. Lafuma, A.; Dreno, B.; Delaunay, M., et al. Economic analysis of adjuvant therapy with interferon alpha-2a in stage II malignant melanoma. *Eur. J. Cancer* **2001**, 37 (3), 369–375.

Clinical Trials: Meta-Analysis using R

Ding-Geng (Din) Chen

School of Social Work, University of North Carolina, Chapel Hill, North Carolina, U.S.A.;

Department of Biostatistics, Gillings School of Global Public Health, University of North

Carolina, Chapel Hill, North Carolina, U.S.A.; Department of Statistics, University of

Pretoria, Pretoria, South Africa

Abstract

Given the globalization of drug development, clinical trials for biopharmaceuticals are most often conducted as multicenter, multiregional trials. Meta-analysis presents an efficient, effective method for combining the statistical data from large, global clinical trials and for increasing the reliability of statistical findings and the conclusions based on those data. This entry presents an overview of meta-analysis and emphasizes the use of the classical fixed-effects and random-effects models to synthesize summary statistics from multiple studies. In addition, meta-regression is used to explain between-study heterogeneity. The meta-analysis models are illustrated using real meta-data collected in 13 clinical trials to assess the efficacy of Bacillus Calmette-Guerin vaccine to prevent tuberculosis; the meta-analyses are implemented in open source R software.

Keywords: DerSimonian-Laird estimator; Fixed-effects model; Heterogeneity; Meta-analysis; Meta-regression; Random-effects model; Weighted-mean estimator.

INTRODUCTION

The development and testing of new biopharmaceuticals has become globalized because many regulatory bodies in different countries require safety and efficacy data obtained from a local patient sample before approving a new drug. To meet these requirements, clinical trials for biopharmaceuticals are most often conducted as multicenter, multiregional trials that offer some distinct advantages over single-site trials. For example, multiregional clinical trials (MRCT) tend to have a larger total sample sizes and are typically conducted in multiple, targeted regions to meet regulatory requirements, and thus, speed-up approval and reduce the time lag in getting a product to market. Global MRCTs can cover multiple regions in one country or can include several countries (e.g., trials in Japan and the United States) or broad global regions (e.g., Europe, Asian-Pacific countries, South America, and Africa). However, ever-changing and differing data requirements across countries can mean biopharmaceutical developers have to institute additional animal or clinical studies that increase development costs and delays patients' access to new drugs. A call for harmonization of these regulatory standards has prompted the U.S. Food and Drug Administration (FDA), the European Medicines Agency (EMA), and the Japanese Pharmaceuticals and Medical Devices Agency (PMDA) to release guidelines regarding the synthesis and meta-analysis of data from MRCTs as a step

toward international harmonization of scientific standards and regulatory requirements. Despite some advances in the harmonization of MRCTs, many challenges remain related to the potential for regional and between-study heterogeneity. Researchers must give careful consideration to this heterogeneity, which stems from the differing regulatory requirements in different countries regarding study design and practices for conducting an MRCT. Specifically, heterogeneity is generated from: (1) differences in the endpoints required to meet regional regulatory requirements because regional authorities have differing standards for primary, co-primary, or key secondary endpoints; (2) different time points required for hypothesis testing because of across-regions inconsistencies in the expected treatment duration or the primary time point for measuring treatment response; (3) differences in experimental designs that result from modifications made to match the differing characteristics of the various regions in an MRCT; (4) differing margins of non-inferiority as determined by different regulatory agencies; and (5) different analytic patient populations as defined in the analysis plans designed for different regions. Any one of these differences could lead to inflated type-I and type-II errors from the perspective of statistical analysis. Thus, it is imperative to follow the guidelines promoting global harmonization that have been provided by health and regulatory authorities such as the International Conference on Harmonization (<http://www.ich.org>).

Among the various methods available for synthesizing data from MRCTs, meta-analysis is a statistical method that is increasingly popular and widely used for examining the validity of the effect-sizes from each component trial in an MRCT, for quantifying the heterogeneity between the effect-sizes, and for providing an estimate of the overall pooled effect-size with optimal precision. With the rising costs and time investment associated with developing MRCTs, many trials are carried out with relatively small samples of patients, resulting in underpowered studies that lack the statistical power to be useful in detecting clinical trial effect-size. As a result, this small trial phenomenon increases the likelihood that separate clinical trials in an MRCT will produce conflicting results. However, by using meta-analytic methods to pool the estimated effect-size from each constituent trial in an MRCT, the data from a larger number of patients can be used and the statistical power can be increased.

In the literature, the term meta-analysis is traditionally defined as a statistical procedure for systematically combining data from multiple studies to reach a single conclusion with greater statistical power than any of the component studies. Most often in the literature, the term is used to refer to the meta-analysis of summary statistics from published studies, or summary statistics meta-analysis (SS-MA). The purpose of a meta-analysis of MRCT is to combine the estimates of study effect-sizes using the well-established approaches of the fixed-effects and random-effects models, which differ fundamentally on whether between-study heterogeneity is incorporated. These traditional meta-analysis models have played an increasingly important role in biopharmaceutical, health, and medical sciences, and meta-analyses of clinical trial data have led to numerous scientific discoveries. For example, since 2000, *The New England Journal of Medicine* has published more than 500 articles that reported results from a meta-analysis. In addition, since 2000, an increasing number of books have been published on SS-MA methods and applications.^[1–5]

This entry is organized to present an overview of methods and then an illustration of the methods. Section “Overview of meta-analysis” gives an extensive overview of meta-analysis using summary statistics. Specifically, the section presents a review of the classical SS-MA using; (1) fixed-effects and random-effects MA models; (2) an overview of quantification of heterogeneity with Q-statistic, τ^2 index, H index, and I^2 index; and (3) a review of the use of meta-regression to explore the heterogeneity with study-level moderators. Section “Meta-analysis using R” then illustrates the meta-analysis using a real dataset obtained from 13 clinical trials conducted to assess the impact of a *Bacillus Calmette–Guerin* (BCG) vaccine in the prevention of tuberculosis (TB). Detailed analysis using open source R software is illustrated in this section so readers can follow the meta-analysis for their own data analysis. The entry ends with concluding remarks.

OVERVIEW OF META-ANALYSIS

Summary Statistics and the Sources of Variations

In a typical meta-analysis with summary statistics, K independent studies are obtained to estimate a parameter of interest, such as the effect-size of efficacy between a new treatment and a control, β_k ($k = 1, 2, \dots, K$). This analytic approach can be applied to a broad range of study designs, including single-arm or multiple-arm studies, randomized controlled trials, and observational studies. For ease of illustration, this discussion focuses on MA with two-arm studies, where β_k is some form of the effect-size between the two groups. The most popular choices for β_k for a continuous outcome are the mean difference or the standardized mean difference; for a dichotomous outcome, the most popular choices are odds ratio, risk ratio, and risk difference. In most cases, an estimated $\hat{\beta}_k$ of the true β_k and its associated standard error (SE) could be directly extracted from each study. The ultimate goal of meta-analysis is to produce an optimal estimate of the overall population effect-size by pooling the estimates $\hat{\beta}_k$ ($k = 1, 2, \dots, K$) from individual studies using appropriate statistical models.

Two sources of errors or variations exist in these summary statistics of $\hat{\beta}_k$ from different studies, with one source being the within-study variation and the other source being the between-study variation. Within-study variation is caused by sampling error, which is random or nonsystematic within each study. In contrast, between-study variation results from systematic differences among studies. If the between-study variation can be verified as being zero, the effect estimates $\hat{\beta}_k$ are considered homogeneous—otherwise, the effect estimates are heterogeneous. In MA, the assumption of homogeneity states that β_k ($k = 1, 2, \dots, K$) is the same in all studies, i.e.,

$$\beta_1 = \beta_2 = \dots = \beta_K = \beta. \quad (1)$$

If these studies are homogeneous, then two commonly used meta-analysis models can be used—the fixed-effects MA model and random-effects MA model.

Fixed-Effects Meta-Analysis

As shown in Eq. 1, a fixed-effects meta-analysis assumes homogeneity when the underlying population effect-sizes β_k are constant across all studies. A typical fixed-effects model is described as:

$$\hat{\beta}_k = \beta + \epsilon_k; k = 1, 2, \dots, K, \quad (2)$$

where $\hat{\beta}_k$ represents the effect-size for study K and β is the overall global population effect-size. The ϵ_k are the sampling error with mean 0 and KNOWN variance $\hat{\sigma}_{\beta_k}^2$ which can be extracted or calculated from the individual studies.

In general, the ε_k is assumed to follow a normal distribution $N(0, \hat{\sigma}_{\beta_k}^2)$. A pooled estimate of β in fixed-effects MA models is given by the weighted least square estimation,

$$\hat{\beta}_F = \frac{\sum_{k=1}^K w_k \hat{\beta}_k}{\sum_{k=1}^K w_k}, \quad (3)$$

and the variance of $\hat{\beta}_F$ can be expressed as:

$$\text{Var}(\hat{\beta}_F) = 1 / \sum_{k=1}^K w_k \quad (4)$$

where a popular choice of weight is $w_k = 1 / \hat{\sigma}_{\beta_k}^2$ and variance $\hat{\sigma}_{\beta_k}^2$ is estimated from study K. Hence, the 95% confidence interval of β_F is given by,

$$\hat{\beta}_F - z_{0.025} \sqrt{\text{Var}(\hat{\beta}_F)} \leq \beta \leq \hat{\beta}_F + z_{0.025} \sqrt{\text{Var}(\hat{\beta}_F)}, \quad (5)$$

where $z_{0.025}$ denotes the 2.5%-percentile of the standard normal distribution. Similarly, a statistical z-test can be formulated as:

$$z = \frac{\hat{\beta}_F - \beta}{\sqrt{\text{Var}(\hat{\beta}_F)}} \quad (6)$$

to be used to test $H_0: \beta = 0$.

Random-Effects Meta-Analysis

When meta-analyzing effect-sizes from a group of studies, it may be impractical to follow the assumption of the fixed-effects model that the K true effect-sizes are the same for all studies. When attempting to synthesize a group of studies using meta-analysis, it is reasonable to expect that the studies have enough in common to allow the data from those studies to be combined for statistical inference; however, it would be impractical to require that all studies in the group have identical true effect-sizes. It is impossible for independent studies to be identical in every aspect. Therefore, heterogeneity is likely to exist in many meta-analyses. The model that takes heterogeneity into account is the following random-effects meta-analysis models:

$$\hat{\beta}_k = \beta + b_k + \varepsilon_k, k = 1, 2, \dots, K, \quad (7)$$

where for study K, $\hat{\beta}_k$ represents the observed effect-size and β the global population effect-size. b_k is now the random-effects with mean 0 and variance τ^2 representing the between-study heterogeneity, and ε_k is the sampling error with mean 0 and variance $\hat{\sigma}_{\beta_k}^2$. It is assumed that b_k and ε_k are independent and follow normal distributions $N(0, \tau^2)$ and $N(0, \hat{\sigma}_{\beta_k}^2)$, respectively. Let $\beta_k = \beta + b_k$, $k = 1, 2, \dots, K$.

Then the random-effects model given in Eq. 7 can be simplified as:

$$\hat{\beta}_k = \beta_k + \varepsilon_k, \quad (8)$$

where β_k represents the true effect-size for study K. All β_k ($k = 1, 2, \dots, K$) are random samples from the same normal population,

$$\beta_k \sim N(\beta, \tau^2) \quad (9)$$

rather than being a constant in the fixed-effects MA in Eq. 2. Further, the marginal variance of $\hat{\beta}_k$ is given by,

$$\text{Var}(\hat{\beta}_k) = \tau^2 + \hat{\sigma}_{\beta_k}^2, \quad (10)$$

which is composed of two sources of variation, i.e. the between-study variance τ^2 and within-study variance $\hat{\sigma}_{\beta_k}^2$. If the between-study variance $\tau^2 = 0$, the random-effects MA models given in Eq. 7 would reduce to the fixed-effects MA models given in Eq. 2.

Similar to the fixed-effects MA models, the within-study variance $\hat{\sigma}_{\beta_k}^2$ can be obtained or calculated from each study K ($k = 1, 2, \dots, K$). However, the information for determining between-study variance τ^2 is often not available, and the methods commonly used for assessing between-study heterogeneity include the DerSimonian-Laird's method of moments (MM) described in the study DerSimonian and Laird,^[6] the maximum likelihood estimation (MLE) method described in the study by Hardy and Thompson,^[7] and the restricted maximum likelihood (REML) method described in the study by Raudenbush and Bryk.^[8] As the most commonly used estimator, MM is a distribution-free and non-iterative approach whereas both MLE and REML are parametric methods and need iteration for estimating τ^2 and the related meta-regression parameter β described in the section "Meta-regression."

The MM uses the Q-statistic for testing the assumption of homogeneity:

$$Q = \sum_{k=1}^K w_k (\hat{\beta}_k - \hat{\beta}_F)^2 \quad (11)$$

where $w_k = 1 / \hat{\sigma}_{\beta_k}^2$ is the weight from the Kth study, $\hat{\beta}_k$ is the Kth study effect-size, and $\hat{\beta}_F$ is the global overall effect estimated from Eq. 3. It can be seen that Q is calculated: (1) to compute the deviations of each effect-size from the meta-estimate and square them [i.e., $(\hat{\beta}_k - \hat{\beta}_F)$]; (2) to weight these values by the inverse-variance for each study; and (3) then to sum these values across all K studies to produce a weighted sum of squares to obtain the heterogeneity measure Q.

From Eq. 11, it can be shown that the expected value of Q is:

$$E(Q) = (K - 1) + U \times \tau^2 \quad (12)$$

$$\text{where } U = \sum_{k=1}^K w_k - \frac{\sum_{k=1}^K w_k^2}{\sum_{k=1}^K w_k}.$$

Under the assumption of no heterogeneity (all studies have the same effect-size), then τ^2 would be zero and $E(Q) = df = K - 1$. Based on this heterogeneity measure Q , the test of heterogeneity is conducted to address the null hypothesis that the effect-sizes β_k from all studies share a common effect-size β (i.e., the assumption of homogeneity) and then test this hypothesis where the test statistic is constructed using Q as a central χ^2 distribution with degrees of freedom of $df = K - 1$. However, a cautionary note is warranted—this χ^2 -test using the Q -statistic has poor statistical power to detect true heterogeneity for a meta-analysis with a small number of studies K , but excessive power to detect negligible variability with a large number of studies as discussed in the study by Harwell^[9] and Hardy and Thompson.^[7] Thus, a nonsignificant test using Q -statistic from a small number of studies can lead to an erroneous selection of a fixed-effects model when possible true heterogeneity exists among the studies, and vice versa. The inability to conclude statistically significant heterogeneity in a meta-analysis of a small number of studies at the 0.05 level of significance is similar to failing to detect statistically significant treatment-by-center interaction in an MRCT. In these settings, many analysts might choose to conduct the test of homogeneity at the 0.10 level as a means of increasing the power of the test.

From Eq. 12, the MM estimate of τ^2 can be shown as follows:

$$\hat{\tau}^2 = \max\left(0, \frac{Q - (K - 1)}{U}\right) \quad (13)$$

The truncation at zero in Eq. 13 ensures the variance estimate is nonnegative.

Therefore, the estimate of the global population effect-size in random-effects MA is given by:

$$\hat{\beta}_R = \frac{\sum_{k=1}^K w_k^* \hat{\beta}_k}{\sum_{k=1}^K w_k^*} \quad (14)$$

where $w_k^* = 1/(\hat{\sigma}_{\beta_k}^2 + \hat{\tau}^2)$. The variance of $\hat{\beta}_R$ is simply,

$$\text{Var}(\hat{\beta}_R) = 1/\sum_{k=1}^K w_k^*$$

and the 95% confidence interval can be constructed by

$$\hat{\beta}_R - z_{0.025} \sqrt{\text{Var}(\hat{\beta}_R)} \leq \beta \leq \hat{\beta}_R + z_{0.025} \sqrt{\text{Var}(\hat{\beta}_R)}.$$

The statistical test can be similarly formulated by:

$$z = \frac{\hat{\beta}_R - \beta}{\sqrt{\text{Var}(\hat{\beta}_R)}} \quad (15)$$

to be used to test $H_0 : \beta = 0$ in the random-effects MA framework.

Quantify Heterogeneity in Meta-Analysis

Although the Q -statistic can be used to test the existence of heterogeneity, it does not report the extent of such heterogeneity. Given this limitation of the Q -statistic, several indices are proposed to quantify heterogeneity, including the τ^2 , H , and I^2 indices.

The τ^2 Index

The τ^2 index is defined as the variance of the true effect-size as seen in the random-effects MA model given in Eq. 13. Since it is impossible to observe the true effect-size, we cannot calculate this variance directly, but we can estimate the variance from the observed data using Eq. 11 as follows:

$$\hat{\tau}^2 = \frac{Q - (K - 1)}{U} \quad (16)$$

which is the well-known DerSimonian-Laird method of moments for τ^2 in Eq. 13. Even though the true variance τ^2 can never be less than zero, the estimated variance $\hat{\tau}^2$ can sometimes be less than zero because of the sampling error leading to $Q < K - 1$. When this occurs, the estimated $\hat{\tau}^2$ is set to zero.

As used in the random-effects MA model, the τ^2 index is also an estimate for the between-studies variance in the meta-analysis.

The H Index

Another index or measure of heterogeneity is the H index, which was proposed in Higgins and Thompson^[10] and is defined as follows:

$$H = \sqrt{\frac{Q}{K - 1}} \quad (17)$$

The H index is based on the fact that when no heterogeneity exists, then $E[Q] = K - 1$. In this case, H should be 1.

The confidence interval for the H index is derived in Higgins and Thompson^[10] based on the assumption that the natural logarithm of $\ln(H)$ follows a standard normal distribution. Accordingly,

$$LL_H = \exp\{\ln(H) - |z_{\alpha/2}| \times SE[\ln(H)]\} \quad (18)$$

$$UL_H = \exp\{\ln(H) + |z_{\alpha/2}| \times SE[\ln(H)]\} \quad (19)$$

where LL and UL denote the lower- and upper-limits of the CI, $z_{\alpha/2}$ is the $\alpha/2$ -quantile of the standard normal

distribution, and $SE[\ln(H)]$ is the standard error of $\ln(H)$ and is estimated by:

$$SE[\ln(H)] = \begin{cases} \frac{1}{2} \frac{\ln(Q) - \ln(K-1)}{\sqrt{2Q - \sqrt{2K-3}}} & \text{if } Q > K \\ \sqrt{\frac{1}{2(K-2)} \left(1 - \frac{1}{3(K-2)^2}\right)} & \text{if } Q \leq K \end{cases} \quad (20)$$

Since $E(Q) \approx K - 1$ as seen in Eq. 12, the H index should be greater than 1 to measure the relative magnitude of heterogeneity among all studies included in the meta-analysis. If the lower limit of this interval is greater than 1, then H is statistically significant.

The I^2 Index

To measure the proportion of observed heterogeneity from the real heterogeneity, Higgins and Thompson^[10] and Higgins et al.^[11] proposed the I^2 index as follows:

$$I^2 = \left(\frac{Q - (K-1)}{Q} \right) \times 100\%, \quad (21)$$

which represents the ratio of excess dispersion to total dispersion and is similar to the well-known R^2 in classical regression that represents the proportion of the total variance that can be explained by the regression variables.

As suggested from Higgins et al.^[11] values of the I^2 index around 25%, 50%, and 75% could be considered as *low*-, *moderate*-, and *high*-heterogeneity, respectively. As noted in their article, about half of the meta-analyses of clinical trials in the *Cochrane Database of Systematic Reviews* have reported an I^2 index of zero, and the other half of reviewers have reported evenly distributed I^2 indices between 0% and 100%.

Mathematically, the I^2 index can be represented using the H index as follows:

$$I^2 = \frac{H^2 - 1}{H^2} \times 100\% \quad (22)$$

This expression allows us to use the results from the H index to give a confidence interval for the I^2 index using the expressions in Eqs. 18 and 19 as follows:

$$LL_{I^2} = \left[\frac{(LL_H)^2 - 1}{(LL_H)^2} \right] \times 100\%$$

$$UL_{I^2} = \left[\frac{(UL_H)^2 - 1}{(UL_H)^2} \right] \times 100\%$$

Since I^2 represents the percentage, any of these limits computed as negative is set to zero. In cases when the lower limit of I^2 is greater than zero, then the I^2 is considered to be statistically significant.

Meta-Regression

Although the random-effects meta-analysis model can account for between-study heterogeneity, this model cannot explain or explore how these heterogeneities originated relative to other study-level variables. When study-level variables—commonly called moderators—are available, meta-regression is used to investigate the association between these moderators and the reported effect-sizes. Meta-regression is in fact the typical weighted regression when study-level moderators are associated with the reported effect-sizes as the dependent variable and their variance is used as weights. From this viewpoint, meta-regression is merely typical multiple regression applied to study-level data, and therefore, the theory of regression can be directly applied to meta-regression.

The following introduces two types of meta-regression, which are built on the fixed-effects and random-effects MA models, respectively. Suppose that there are p moderators $X_1, X_2, \dots, X_p$ and K independent studies. The fixed-effects meta-regression model is given by:

$$\hat{\beta}_k = \beta_0 + \beta_1 x_{k1} + \dots + \beta_p x_{kp} + \epsilon_k \quad (23)$$

where $x_{k1}, \dots, x_{kp}$ ($k = 1, 2, \dots, K$) denote the observed values of the p moderators $X_1, X_2, \dots, X_p$ for study K and $\beta_0, \beta_1, \dots, \beta_p$ are regression coefficients. The effect-size $\hat{\beta}_k$ and sampling error ϵ_k have the same definitions as in fixed-effects MA models given in Eq. 2. The model assumes that the variation in effect-sizes can be fully explained by these moderators.

The random-effects meta-regression can be obtained by adding random-effects b_k to the fixed-effects model in Eq. 23:

$$\hat{\beta}_k = \beta_0 + \beta_1 x_{k1} + \dots + \beta_p x_{kp} + b_k + \epsilon_k, k = 1, 2, \dots, K$$

where b_k is assumed to be independent with a mean 0 and variance τ_2 . Unlike the fixed-effects meta-regression, where all variability in effect-sizes can be explained by the moderators $X_1, X_2, \dots, X_p$, the random-effects meta-regression assumes that the model can explain only part of the variation and the remaining variance can be accounted for by random-effects b_k .

Is is notable that the meta-regression technique is most appropriate when a meta-analysis includes a large number of studies. Further, because the moderators and the outcome in meta-regression are all study-level summary statistics (e.g., patient mean age, proportion of female patients), the relation between these moderators and the outcome might not directly reflect the relation between subject scores and subject outcomes, causing aggregation bias. Therefore, those conducting meta-regression should give careful consideration to the interpretation of the results.

META-ANALYSIS USING R

BCG Vaccine Clinical Trials

This section uses a dataset compiled from 13 clinical trials conducted to assess the efficacy of a BCG vaccine in the prevention of TB. The dataset has been used elsewhere to illustrate meta-analysis and meta-regression. For example, these data have been used by Everitt and Hothorn^[12] (see Table 12.2), Hartung et al.^[2] (see Table 18.8), Borenstein et al.^[3] (see Table 20.1) and Chen and Peace^[5] (see Table 7.1). In addition, the dataset has been used in the papers by van Houwelingen et al.^[13] to illustrate a tutorial in statistical analyses, and by Viechtbauer^[14] to provide an overview for those interested in learning how to conduct meta-analysis in R with *metafor* package. The source dataset was first reported in the study by Colditz et al.^[15]

Notably, the sources noted above vary in the numbers reported even though all used the same dataset from the same publications. The data tables reported from Everitt and Hothorn,^[12] Borenstein et al.,^[3] and Colditz et al.^[15] are the total cases in both the BCG test group and the control group. However, the dataset reported in van Houwelingen et al.^[13] and in Viechtbauer^[14] for the R *metafor* library are the numbers of TB “negative cases.” This entry uses the TB “negative case” data structure which is reproduced here for complete presentation in Table 1.

In this table, *author* denotes the authorship from the 13 studies, *year* is the publication year of these 13 studies, *tpos* is the number of TB-positive cases in the BCG vaccinated group, *tneg* is the number of TB-negative cases in the BCG vaccinated group, *cpos* is the number of TB-positive cases in the control group, *cneg* is the number of TB-negative cases in the control group, *ablat* denotes the absolute latitude of the study location (in degrees), and

alloc denotes the method of treatment allocation with three levels of random, alternate, or systematic assignment.

The purpose of the original meta-analysis was to quantify the efficacy of the BCG vaccine against TB, which was facilitated by a random-effects meta-analysis. The original study concluded that the BCG vaccine significantly reduced the risk of TB-negative in the presence of significant heterogeneity. The heterogeneity was partially explained by the study sites’ geographic latitudes. The following section uses this dataset to illustrate the meta-analysis procedures described in Section 2 with a step-by-step implementation in R.

Fixed-Effects and Random-Effects Meta-Analysis

Effect-Size Calculation

We illustrate the meta-analysis using the R package *metafor*. To use this package, the dataset must first be loaded into R using the following R code chunk:

```
# load the R library
> library("metafor")
# load the BCG data
> data("dat.bcg", package = "metafor")
```

To conduct a meta-analysis, we need to convert the 2×2 contingency table in Table 1 into study-specific effect-size (ES). As a typical ES, we use the log odds-ratio which can be specified as *measure* = “OR” in R function *escalc* which stands for “effect-size calculation” as shown in following R code chunk:

```
> dat = escalc(measure="OR", ai=tpos,
  bi=tneg, ci=cpos, di=cneg, data=dat.
  bcg, append = TRUE)
```

Table 1 Data from studies on efficacy of BCG vaccine for preventing TB

Trial	Author	Year	tpos	tneg	cpos	cneg	ablat	Alloc
1	Aronson	1948	4	119	11	128	44	Random
2	Ferguson & Simes	1949	6	300	29	274	55	Random
3	Rosenthal et al.	1960	3	228	11	209	42	Random
4	Hart & Sutherland	1977	62	13536	248	12619	52	Random
5	Frimodt-Moller et al.	1973	33	5036	47	5761	13	Alternate
6	Stein & Aronson	1953	180	1361	372	1079	44	Alternate
7	Vandiviere et al.	1973	8	2537	10	619	19	Random
8	TPT Madras	1980	505	87886	499	87892	13	Random
9	Coetzee & Berjak	1968	29	7470	45	7232	27	Random
10	Rosenthal et al.	1961	17	1699	65	1600	42	Systematic
11	Comstock et al.	1974	186	50448	141	27197	18	Systematic
12	Comstock & Webster	1969	5	2493	3	2338	33	Systematic
13	Comstock et al.	1976	27	16886	29	17825	33	Systematic

The option `append=TRUE` above is to append the calculated ES (i.e. log odds-ratio in Eq. 3.5 and named as “yi”) and its variance in Eq. 3.6 (i.e., named “vi”) to the original data which can be seen as follows:

trial	author	year	tpos	tneg	cpos	cneg	ablat	yi	vi
1	Aronson	1948	4	119	11	128	44	-0.9387	0.3571
2	Ferguson & Simes	1949	6	300	29	274	55	-1.6662	0.2081
3	Rosenthal et al	1960	3	228	11	209	42	-1.3863	0.4334
4	Hart & Sutherland	1977	62	13536	248	12619	52	-1.4564	0.0203
5	Frimodt-Moller et al	1973	33	5036	47	5761	13	-0.2191	0.0520
6	Stein & Aronson	1953	180	1361	372	1079	44	-0.9581	0.0099
7	Vandiviere et al	1973	8	2537	10	619	19	-1.6338	0.2270
8	TPT Madras	1980	505	87886	499	87892	13	0.0120	0.0040
9	Coetzee & Berjak	1968	29	7470	45	7232	27	-0.4717	0.0570
10	Rosenthal et al	1961	17	1699	65	1600	42	-1.4012	0.0754
11	Comstock et al	1974	186	50448	141	27197	18	-0.3408	0.0125
12	Comstock & Webster	1969	5	2493	3	2338	33	0.4466	0.5342
13	Comstock et al	1976	27	16886	29	17825	33	-0.0173	0.0716

It can be found that 11 of 13 trials (except trials 8 and 12) have negative log odds-ratios, which indicate that the BCG vaccine reduced the risk of TB for these 11 studies, but not for the trials of 8 and 12. The statistical significance can be examined by constructing the t-statistic using the ratio of “yi” to the squared-root of “vi” or the 95% confidence intervals as shown in Fig. 1. Examining this figure, we can see that only studies 2, 3, 4, 6, 7, 10, and 11 (i.e., 7 out of 13) are statistically significant.

Fixed-Effects Meta-Analysis

The above ES calculation leads to meta-analysis of the 13 studies together. Assuming all 13 studies are homogenous,

the fixed-effects meta-analysis described in Section Fixed-Effects Meta-Analysis can be performed using R function “rma” as follows:

```
# Fixed-effects MA
> meta.FE = rma(yi, vi, data = dat,
  method="FE")
# Print the summary
> meta.FE
Fixed-Effects Model (k = 13)
Test for Heterogeneity: Q(df = 12)
= 163.1649, p-val < .0001
```

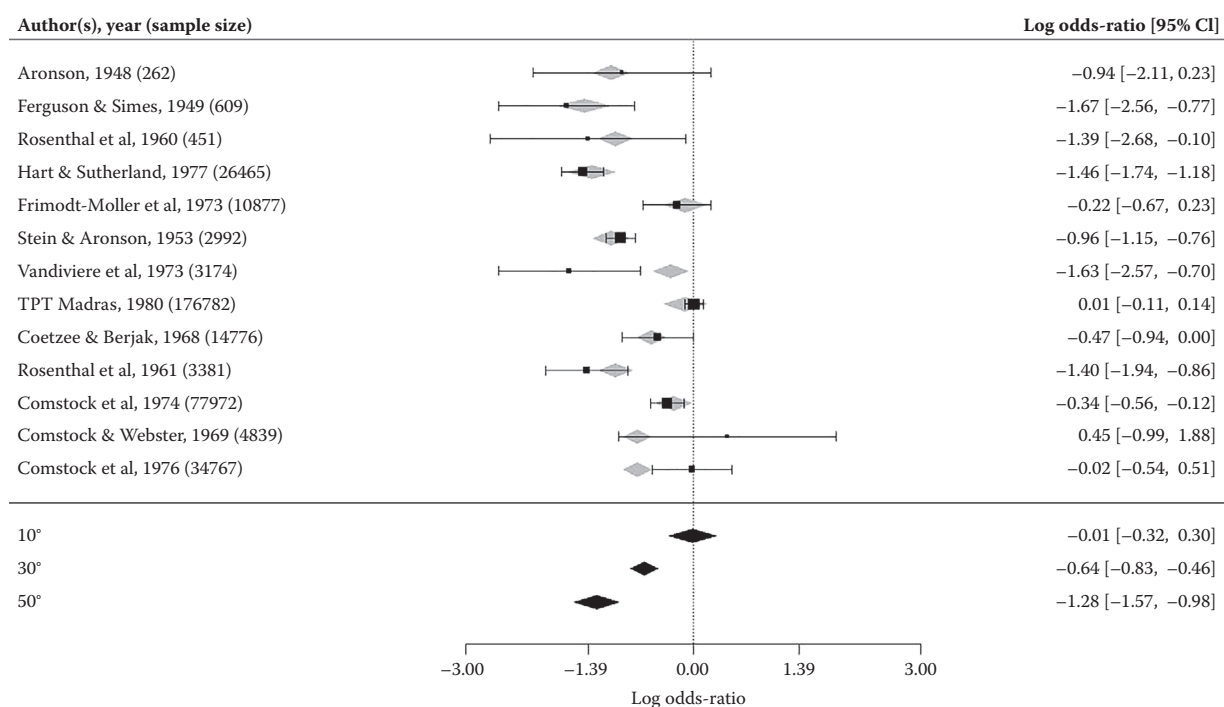


Fig. 1 Forest plot before/after meta-regression

Model Results:

```
estimate se      zval    pval   ci.lb   ci.ub
-0.4361 0.0423 -10.3190 <.0001 -0.5190 -0.3533
```

This shows that the estimated global log odds-ratio is -0.436 with a SE of 0.0423 and a p -value < 0.0001 . These values mean that when the 13 studies are analyzed together (i.e., meta-analysis), the global overall effect of the BCG vaccine is found to be statistically significant in preventing TB. However, the test of heterogeneity (i.e., $Q = 163.1649$ with $df = 12$, p -value < 0.0001 from χ^2 -test) is highly statistically significant, which indicates that the assumption of homogeneous studies in fixed-effects meta-analysis is not appropriate, and therefore, the random-effects meta-analysis should be explored.

Random-Effects Meta-Analysis

The DerSimonian-Laird random-effects meta-analysis models can be implemented easily as follows:

```
# Call 'rma' with method="DL" to fit the
random-effects MA
> meta.DL = rma(yi, vi, data = dat,
method="DL")
# Print the summary
> meta.DL
Random-Effects Model (k = 13; tau^2
estimator: DL)
tau^2 (estimated amount of total
heterogeneity): 0.3663 (SE = 0.2659)
tau (square root of estimated tau^2
value): 0.6053
I^2 (total heterogeneity / total
variability): 92.65%
H^2 (total variability / sampling
variability): 13.60
```

Model Results:

```
estimate se      zval    pval   ci.lb   ci.ub
-0.4361 0.1923 -10.3190 0.0001 -1.1242 -0.3706
```

The global overall estimate from the random-effects meta-analysis model is $\hat{\beta} = -0.747$ with $SE = 0.192$ and p -value $= 0.001$, which indicates the results of the 13 studies reach statistical significance, if meta-analyzed together. In this random-effects meta-analysis, the estimated between-study heterogeneity $\hat{\tau}^2 = 0.3663$ with $I^2 = 92.65\%$ and $H^2 = 13.60$ demonstrates highly and statistically significant heterogeneity. However, to explain this heterogeneity, we will need to use meta-regression.

Meta-Regression

To explain the heterogeneity, we can make use of the geographic latitude data that indicate the location where the clinical trials were conducted. This random-effects meta-regression can be implemented using the following R code chunk:

```
# random-effects meta-regression
> metaReg = rma(yi, vi, mods = ~ablat,
method="DL", data = dat)
# Print the meta-regression results
> metaReg
Mixed-Effects Model (k = 13; tau^2
estimator: DL)
tau^2 (estimated amount of residual
heterogeneity): 0.0480 (SE = 0.0451)
tau (square root of estimated tau^2
value): 0.2191
I^2 (residual heterogeneity /
unaccounted variability): 56.17%
H^2 (unaccounted variability / sampling
variability): 2.28
R^2 (amount of heterogeneity accounted
for): 86.90%
Test for Residual Heterogeneity: QE(df =
11) = 25.0954, p-val = 0.0088
Test of Moderators (coefficient(s) 2):
QM(df = 1) = 26.1628, p-val < .0001
```

Model Results:

```
estimate se      zval    pval   ci.lb   ci.ub
intrcpt  0.3030 0.2109  1.4370 0.1507 -0.1103  0.7163
ablat    -0.0316 0.0062 -5.1150 <.0001 -0.0437 -0.0195
```

As can be seen from the output, with only *ablat*, the estimated residual heterogeneity $\hat{\tau}^2$ dropped to 0.048 ($SE = 0.0451$) from the previous random-effect meta-analysis without *ablat* where the heterogeneity was estimated at 0.3663 . This change in estimates indicates that the moderator *ablat* itself accounts for $(0.3663 - 0.0480) / 0.3663 = 87\%$ of the total amount of heterogeneity, and the absolute latitude is significantly related to the effectiveness of the BCG vaccine in preventing TB, which can be quantified in the estimated meta-regression equation as follows:

$$\log(OR) = 0.3030 - 0.0316 \times ablat \quad (24)$$

The estimated equation and the entire meta-regression summary is graphically displayed in Fig. 2. In Fig. 2, the meta-regression in Eq. 24 is plotted using a solid line and its 95% CI bounds are plotted in dashed dots. The horizontal line from $OR = 0$ indicates no BCG vaccine effects. The

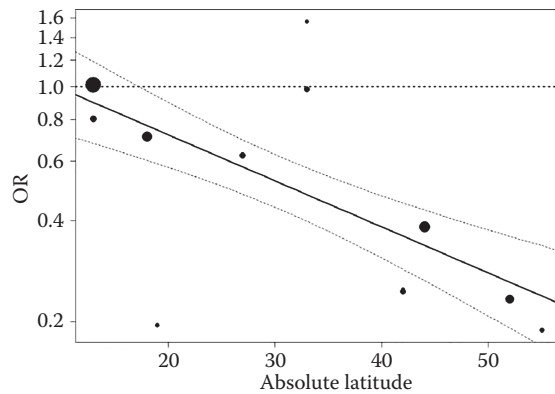


Fig. 2 Random-effects meta-regression for BCG data on OR on absolute latitude

13 studies are plotted in 13 solid dots where the sizes of the dots are proportional to their sample sizes.

Both this equation and Fig. 2 show that the BCG vaccine was more effective as the absolute latitude increased. When the *ablat* is less than 20 degrees and close to zero (i.e., a study performed closer to equator), the effect-size would be close to zero (as evidenced by the nonsignificant intercept parameter and Fig. 2, which means the vaccine had no real effect on prevention of TB). As *ablat* increases, say to latitude of 60 degrees, the log OR as calculated from Eq. 24 is -1.593 which corresponds to an OR of 0.20.

This property is further illustrated in Fig. 1, where the right side of the figure is the point estimate of log odds-ratio and the 95% CIs for each of the 13 studies. The middle part is a graphical representation of these 95% CIs in solid line segments. The shaded diamonds are the 95% CIs after the meta-regression adjustment. The three dark diamonds near the bottom of the figure are the 95% CIs for the random-effects meta-regression at three latitudes of 10, 30, and 60 degrees. We can see that at 10 degrees, the estimated log odds-ratio is -0.01 and the 95% CI is $(-0.32, 0.30)$, indicating that the tested BCG vaccine produced no statistically significant benefit.

CONCLUSION

This entry gave a detailed overview for the classical statistical meta-analysis including the fixed-effects and random-effects meta-analysis models. Linked with quantification of the heterogeneity in meta-analysis, the indices of τ^2 , H , and I^2 were introduced along with the meta-regression to take into account the between-study heterogeneity and variations from all studies in meta-analysis. To facilitate the understanding of meta-analysis techniques and for interested readers to use these models for their own meta-analysis, a real-world dataset is used as an example to illustrate how to do a meta-analysis by giving

a step-by-step implementation in R so that readers can follow the meta-analysis models and reuse the R programs for their own meta-analysis.

For ease of illustration, this entry focused on meta-analysis of two-arm clinical trials. Another direction for meta-analysis is to consider multiple-treatment comparisons or mixed-treatment comparisons, which are called network meta-analysis. Network meta-analysis combines multiple-treatment comparisons as a network to achieve a more statistically powerful analysis; therefore, network meta-analysis is a key tool for evidence-based biopharmaceutical research and development. Readers interested in network meta-analysis are referred to R package “netmeta” for more information and R implementations.

REFERENCES

1. Whitehead, A. *Meta-Analysis of Controlled Clinical Trials*; John Wiley & Sons, Inc.: New York, 2003.
2. Hartung, J.; Knapp, G.; Sinha, B.K. *Statistical Meta-Analysis with Applications*; John Wiley & Sons, Inc.: Hoboken, NJ, 2008.
3. Borenstein, M.; Hedges, L.V.; Higgins, J.P.T.; Rothstein, H.R. *Introduction to Meta-Analysis*; Wiley: Chichester, 2009.
4. Pigott, T.D. *Advances in Meta-Analysis*; Springer: New York, 2012.
5. Chen, D.G.; Peace, K.E. *Applied Meta-Analysis Using R*; Chapman & Hall/CRC: Boca Raton, FL, 2013.
6. DerSimonian, R.; Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* **1986**, *7*, 177–188.
7. Hardy, R.J.; Thompson, S.G. Detecting and describing heterogeneity in metaanalysis. *Stat. Med.* **1998**, *17*, 841–856.
8. Raudenbush, S.W.; Bryk, A.S. Empirical bayes meta-analysis. *J. Educat. Stat.* **1985**, *10* (2), 7598.
9. Harwell, M. An empirical study of hedges homogeneity test. *Psychol. Methods* **1997**, *2*, 219–231.
10. Higgins, J.P.T.; Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **2002**, *21*, 1539–1558.
11. Higgins, J.P.T.; Thompson, S.G.; Deeks, J.J.; Altman, D.G. Measuring inconsistency in meta-analyses. *Br. Med. J.* **2003**, *327*, 557–560.
12. Everitt, B.; Hothorn, T. *A Handbook of Statistical Analyses Using R*; Chapman & Hall/CRC: Boca Raton, FL, 2006.
13. van Houwelingen, H.C.; Arends, L.R.; Stijnen, T. Tutorial in biostatistics: Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Stat. Med.* **2002**, *21*, 589–624.
14. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *J. Stat. Software* **2010**, *36* (1), 1–48.
15. Colditz, G.A.; Brewer, T.F.; Berkey, C.S.; Wilson, M.E.; Burdick, E.; Fineberg, H.V.; Mosteller, F. Efficacy of BCG vaccine in the prevention of tuberculosis: Metaanalysis of the published literature. *J. Am. Med. Assoc.* **1994**, *271*, 698–702.

Clinical Trials: Response-Adaptive Repeated Measurement Designs

Keumhee Carriere Chough

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada

Yuan Yuan Liang

Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, San Antonio, Texas, U.S.A

Abstract

This entry reviews general allocation strategies for building repeated measurement designs, provides simulations to illustrate the properties of adaptive designs, and details applications for this method.

Keywords: Evaluation function; Mean squared error (MSE); Multiple-objective design; Optimality criteria; Repeated measurement design; Response-adaptive design (RAD); Self and mixed carryover effects model.

INTRODUCTION

Drug development is complex and costly. It requires researchers to test the potential of numerous chemical compounds to treat the disease. Use of prior experiences and/or accumulated information can advance the development of clinical trial designs. Clinical designs can greatly assist in this pursuit, by utilizing past experiences to update the protocols of the trials continuously based on evidence.

When using response-adaptive designs (RADs), the trial is modified on the basis of available responses in order to achieve a specific goal. Response-adaptive allocation literature has a long history. Hu and Rosenberger^[1] raised important questions regarding response-adaptive randomization in a rigorous mathematical framework and systematically addressed various issues. RADs have traditionally been constructed to satisfy a single objective, for example, obtaining a precise estimate of a treatment effect,^[2] minimizing the number of subjects required for a study,^[3] or maximizing the allocation of subjects to beneficial treatments.^[4–6] More recently, Liang and Carriere^[7] studied an adaptive allocation strategy to construct multiple-objective repeated measurement designs, to increase estimation precision and treatment benefit by recognizing efficacy of treatments or treatment groups (sequences). In this chapter, we provide unified strategy to construct designs for both continuous and dichotomous outcomes.

This chapter is organized as follows. The section “Design Allocation Strategy” reviews the general allocation strategies for building repeated measurement designs. “Design Model” describes the application details of the general allocation rule for clinical trials with continuous and

dichotomous outcomes. “Simulations” presents the properties of the adaptive designs via simulations in practical two-treatment two- or three-period repeated measurement designs with continuous and dichotomous outcomes. Finally, we present our conclusions and suggestions for further work.

DESIGN ALLOCATION STRATEGY

To construct multiple-objective response-adaptive designs, we pre-specify N (total number of subjects) and λ , the percentage weight given to the primary objective, and then $1 - \lambda$ is the percentage weight given to a secondary objective. Here, the first objective may be the typical one, maximizing the information matrix, and the second objective may be to increase favorable treatment experiences in the trial. Depending on the nature of trials, these objectives can be revised.

Basically, the usual optimal design construction methods are advocated to adaptively determine the allocation rule for solving multiple objectives,^[2,8–10] as follows.

- Step 1: Choose a desired evaluation function g . Then, the first m ($m < N$) patients are randomly assigned to all possible treatments or treatment sequences.
- Step 2: Assuming that the l^{th} ($m + 1 \leq l \leq N$) patient will be treated by treatment sequence k , calculate the information matrix $\hat{A}_l^k$ based on first $l - 1$ patients' responses, and the evaluation function $g_{l-1,k}$ for each treatment sequence k , where the domain of k is all possible treatment sequences, to allocate the l^{th} patient.

Liang and Carriere^[7] have given some examples of defining evaluation functions for different types of outcomes. Without loss of generality, we assume that a higher value of $g_{l-1,k}$ indicates a better treatment sequence from this point forward.

Step 3: Choose the treatment sequence k^* from the s possible treatment sequences for the l^{th} patient, in such a way that

$$\Lambda(l, k^*) = \max_{k=1 \dots s} \Lambda(l, k)$$

where

$$\Lambda(l, k) = \lambda \frac{\Theta(\hat{A}_l^k)}{\Theta(\hat{A}_l^{(o)})} + (1 - \lambda) \frac{g_{l-1,k}}{g_{l-1,k_{l-1}^{(B)}}}. \quad (1)$$

Eq. 1 is called the selection criterion, which is designed to achieve two objectives by creating a balance between them. We choose the criterion in such a way that the first part aims to favor a treatment sequence that maximizes the information matrix measured by its determinant (D-optimality), trace (A-optimality), or maximum eigenvalue (E-optimality). The second part aims to favor a treatment sequence that gives the best performance on the evaluation function, among all of s treatment sequences. At each stage, one treatment sequence $k_l^{(o)}$ maximizes the optimality criterion, $\Theta(\cdot)$, and possibly another treatment sequence $k_{l-1}^{(B)}$ has the best value on the evaluation function, $g_{l-1,k}$. Prior to the experiment, investigators can choose a parameter $\lambda \in [0, 1]$ to weight and balance the two objectives. When $\lambda = 1$, we have the traditional criterion of a response-adaptive design problem. When $\lambda = 0$, the efficacy of the treatment sequence is the only concern. The choice is often driven by what the investigators want to emphasize. In situations where more than one treatment sequence achieves the maximum criterion score, we can randomly assign the l^{th} patient to one of them.

Step 4: Repeat steps 2 to 3 until all N patients have been allocated.

Note that the above adaptive approach is applicable to both discrete and continuous responses, under suitable model assumptions.

DESIGN MODEL

Continuous Outcomes

Traditionally, models for repeated measurement designs have used simple first-order carryover effects.^[11,12] Recently,

Afsarinejad and Hedayat^[13] proposed a model that allows for two different types of carryover effects from each treatment—the self and mixed carryover effects. They also studied optimal two-period repeated measurement designs with two or more treatments under a model with fixed subject effects. The model we consider is

$$y_{ijk} = \mu + \pi_i + \tau_{d_k[i,j]} + (1 - \delta_{ijk})\gamma_{d_k[i-1,j]} + \delta_{ijk}\phi_{d_k[i-1,j]} + \xi_{jk} + \varepsilon_{ijk} \quad (2)$$

where y_{ijk} denotes the response variable for the j^{th} subject given treatment sequence k in period i , $i = 1, 2, \dots, p$, $j = 1, 2, \dots, N_k$, $k = 1, 2, \dots, s$, p is the number of periods, N_k is the number of subjects given treatment sequence k , s is the number of treatment sequences, μ , π_i , ξ_{jk} , and ε_{ijk} are the overall mean, period, subject, and measurement error effects, respectively^[7]. Denote $d_k[i, j]$ as the treatment used for subject j given treatment sequence k in period i . Assume that ξ_{jk} and ε_{ijk} are mutually independent random effects with mean 0 and variance σ_ξ^2 and σ_ε^2 , respectively. Then $\tau_{d_k[i,j]}$ is the treatment effect of a treatment $d_k[i, j]$, and both $\gamma_{d_k[i-1,j]}$ and $\phi_{d_k[i-1,j]}$ represent carryover effects of a treatment $d_k[i-1, j]$ given in period $(i-1)$ into the next period for subject j given treatment sequence k with an indicator variable δ_{ijk} taking 1 if $d_k[i, j] = d_k[i-1, j]$ and 0 otherwise.

Thus, this model considers two different kinds of carryover effects: $\gamma_{d_k[i-1,j]}$ is the carryover effect of one treatment followed by a different treatment, called the *mixed carryover effect*, while $\phi_{d_k[i-1,j]}$ is the carryover effect of a treatment onto itself, called *self carryover effect*, with $\gamma_{d_k[0,j]} = \phi_{d_k[0,j]} = 0$. In following the steps in the section “Design Allocation Strategy”, calculating the information matrix and Criterion (1) is rather straightforward with continuous responses.^[7]

Dichotomous Outcomes

When the outcome is dichotomous, we use a multivariate Bernoulli model, and we assume that the responses are mutually independent. Let π_r be the success probability of treatment A in the r^{th} period, $r = 1, 2, \dots, p$; when $r = p+1, p+2, \dots, 2p$, let π_r be the success probability of treatment B in the $(r-p)^{\text{th}}$ period. To illustrate the adaptive strategy, we consider two treatment situations.

To allocate the l^{th} patient (or sometimes called the l^{th} stage), denote N_{kl} the number of subjects receiving treatment sequence k , where $k = 1, 2, \dots, 2^p$. $S_l = (S_{1A|}, \dots, S_{pA|}, S_{1B|}, \dots, S_{pB|})^T$, where S_{itl} denotes the number of successes of treatment t in the i^{th} period, where $i = 1, 2, \dots, p$ and $t = A$ or B . For example, $S_{1A|}$ represents the number of successes of A in the first period at the l^{th} stage. Let $S_l[r]$ be the r^{th} element of S_l , where $r = 1, 2, \dots, 2p$.

The likelihood function based on the first l patients is then

$$L_l = \prod_{r=1}^6 \pi_r^{S_l[r]} (1 - \pi_r)^{(N_{l_l}[r] - S_l[r])}$$

where if $1 \leq r \leq p$, $N_{l_l}[r]$ denotes the total number of patients receiving treatment A in the l^{th} period; if $p+1 \leq r \leq 2p$, $N_{l_l}[r]$ denotes the total number of patients receiving treatment B in the $(r-p)^{th}$ period. For example of a two-treatment three-period case with treatment sequences AAA, AAB, ABA, ABB, BBB, BBA, BAB, and BAA, they are

$$\begin{aligned} N_{l_l}[1] &= N_{1l} + N_{2l} + N_{3l} + N_{4l}, \\ N_{l_l}[2] &= N_{1l} + N_{2l} + N_{7l} + N_{8l}, \\ N_{l_l}[3] &= N_{1l} + N_{3l} + N_{6l} + N_{8l}, \\ N_{l_l}[4] &= N_{5l} + N_{6l} + N_{7l} + N_{8l}, \\ N_{l_l}[5] &= N_{3l} + N_{4l} + N_{5l} + N_{6l} \quad \text{and} \\ N_{l_l}[6] &= N_{2l} + N_{4l} + N_{5l} + N_{7l}. \end{aligned}$$

Let $N_{l_l} = (N_{l_l}[1], N_{l_l}[2], \dots, N_{l_l}[6])^T$.

The log-likelihood function ℓ_l becomes

$$\ell_l \propto \sum_{r=1}^6 (S_l[r] \log \pi_r + (N_{l_l}[r] - S_l[r]) \log(1 - \pi_r)).$$

The expected Fisher information matrix up to the l^{th} patient, A_r , which is a 6×6 diagonal matrix, becomes

$$A_l = \text{Diag} \left(E \left(\frac{S_l[r]}{\pi_r^2} + \frac{N_{l_l}[r] - S_l[r]}{(1 - \pi_r)^2} \right) \right)$$

where $r = 1, 2, \dots, 6$.

It is easy to show that the maximum likelihood estimation of unknown parameter π_r at stage l is obtained as $\hat{\pi}_r = \frac{S_l[r]}{N_{l_l}[r]}$, where $r = 1, 2, \dots, 6$.

Result 1: In two-treatment three-period repeated measurement designs with eight possible treatment sequences: {AAA, AAB, ABA, ABB, BBB, BBA, BAB, BAA} the expected Fisher information matrix on the $(l+1)^{th}$ stage is as follows. Given the history by l patients and under the assumption that the $(l+1)^{th}$ patient receiving treatment sequence k , $k = 1, 2, \dots, 8$,

$$A_{l+1}^k = \text{Diag} \left(\frac{S_{l+1}[r]}{\pi_r^2} + \frac{N_{l+1}[r] - S_{l+1}[r]}{(1 - \pi_r)^2} \right)$$

where $S_{l+1} = S_l + \text{Diag}(\pi \times \mathbf{u}_k)$, $\pi = (\pi_1, \pi_2, \dots, \pi_6)^T$, $N_{l+1} = N_{l_l} + \mathbf{u}_k^T$, $S_{l+1}[r]$ and $N_{l+1}[r]$ are the r^{th} element of S_{l+1} and N_{l+1} , respectively, and $\mathbf{u}_k$ is the k^{th} row vector of the matrix $\mathbf{U}$ defined below

$$\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_8)^T = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 \end{pmatrix}$$

$r = 1, 2, \dots, 6.$

This result can be generalized to apply to a multi-period situation, with $\mathbf{u}_k = (d(1, k), \dots, d(2p, k))$. When $1 \leq r \leq p$, $d(r, k) = 1$ if the treatment in the r^{th} period of the treatment sequence k is the treatment A; $d(r, k) = 0$, otherwise. When $p+1 \leq r \leq 2p$, $d(r, k) = 1$ if the treatment in the $(r-p)^{th}$ period of the treatment sequence k is the treatment B; $d(r, k) = 0$, otherwise. And $k = 1, 2, \dots, 2^p$, $r = 1, 2, \dots, 2p$.

In the spirit of the play-the-winner principle, an evaluation function is defined as the average number of successes for each treatment sequence, that is,

$$g_{kl} = \frac{\mathbf{u}_k \times S_l}{N_{kl}}$$

where $k = 1, 2, \dots, 8$, and $\mathbf{u}_k$ as defined in *Result 1* above.

For a given value of λ , under the optimality criteria of the estimated information matrix, $\Theta(\cdot)$, we choose a treatment sequence k^* to the $(l+1)^{th}$ patient by maximizing Λ ,

$$\Lambda = \lambda \frac{\Theta(\hat{A}_{l+1}^k)}{\Theta(\hat{A}_{l+1}^{(O)})} + (1 - \lambda) \frac{g_{l,k}}{g_{l,k_l^{(B)}}}$$

where $\Theta(\cdot)$, $k_l^{(O)}$ and $k_l^{(B)}$ are defined as in the section “Design Allocation Strategy”. Liang and Carriere^[14] give more details on the response adaptive construction for repeated measurement designs with dichotomous outcomes.

SIMULATIONS

We construct response-adaptive two-treatment two- and three-period repeated measurement designs. To assess the efficiency of an adaptive design, a matrix of mean squared error (MSE) for θ is computed

$$MSE = E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T],$$

where θ is a vector of parameters of interest, and $\hat{\theta}$ is an estimate of θ .

In simulation studies, MSE is estimated by

$$MSE = \frac{\sum_{b=1}^B (\hat{\theta}^{(b)} - \theta)(\hat{\theta}^{(b)} - \theta)^T}{B},$$

where $\hat{\theta}^{(b)}$ is a maximum likelihood estimator of θ obtained in the b^{th} simulation run for the total B number of simulations.

Denote MSE_1 as the matrix of MSE for a proposed adaptive design and MSE_0 as the matrix for a reference design. Based on A-, D-, or E- optimality criteria,^[14] the relative efficiency (RE) of the adaptive design compared with the reference design. When $RE = a > 1$, the adaptive design is $(a - 1) \times 100\%$ more efficient than the reference design. When $RE = a < 1$, the adaptive design is only $a \times 100\%$ as efficient as the reference design.

Next we consider the situation for continuous and dichotomous outcomes, respectively. We consider $\lambda = 1, 0.9, 0.7, 0.3$, and 0 , and $N = 10$ (or 12), $40, 80$, and 100 . To smooth out the randomness, we report the average allocation results to treatment sequences from 5,000 repetitions. In addition, we compare the adaptive designs with fixed optimal designs using the relative efficiency. The computations were done in R version 2.6.2 (The R Foundation for Statistical Computing, Vienna, Austria, 2008).

Continuous Outcomes

We assume that $\sigma_\xi^2 = 2, \sigma_\epsilon^2 = 1$, and $\mu = 100$. For two-period three-period designs, there are four possible treatment sequences: AA, AB, BA , and BB . We consider two sets of parameters: (i) $\pi = \tau = \gamma = \phi = 0$ (no treatment difference); and (ii) $\pi = \tau = \phi = 2.5, \gamma = -2.5$ (for treatment performance in decreasing order from best to worst: AA, BA, AB , and BB). π, τ, γ and ϕ are the period effect contrast, direct treatment effect contrast, mixed carryover effect contrast, and self carryover effect contrast, respectively. For three-period designs, there are eight possible treatment sequences: $AAA, AAB, ABA, ABB, BBB, BBA, BAB$, and BAA . Similar to the two-period case, we consider two sets of parameters as well: (i) $\pi_2 = \pi_3 = \tau = \gamma = \phi = 0$ (no treatment difference); and (ii) $\pi_2 = \pi_3 = \tau = \phi = 2.5, \gamma = -2.5$ (with treatment difference), where π_2 and π_3 are dummy variables for the second- and third-period effects, respectively.

Two-Treatment Two-Period Designs

Four different treatment sequences are available for assignment. Suppose the first four patients are assigned equally to the four sequences, and the rest of the patients are to be allocated adaptively.

Table 1 shows the average number of patients assigned to each of the four treatment sequences. When there is no treatment difference, the columns I in Table 1 show that for all combinations of N and λ values, the adaptive allocation rule assigns an approximately equal number of subjects to each of the four treatment sequences, as expected, leading to the usual optimal four-sequence two-period design. When there is a treatment difference, the rule again assigns

an equal number of subjects to each of the four treatment sequences if $\lambda = 1$. This is similar to the traditional fixed optimal design. This is because estimation precision is the only consideration when $\lambda = 1$. When $\lambda < 1$, more patients are assigned to treatment sequence AA , which is the best combination setting. The rest of the patients, in decreasing order, receive treatments BA, AB , or BB , consistent with the setting.

Estimation of each parameter of interest with its standard error (not shown) indicates that in all cases the estimated values are very close to the true values of the parameters, and behave consistently as expected for increasing λ and N , allowing a tradeoff between treatment benefit and statistical optimality.

Two-Treatment Three-Period Designs

In two-treatment three-period designs, eight different treatment sequences are available for assignment. At the initial stage, eight patients are assumed to enter in the study, one for each treatment sequence. For $\pi_2 = \pi_3 = \tau = \gamma = \phi = 0$, since there are no treatment advantage, the columns I in Table 2 show that adaptive designs assign an approximately equal number of subjects to each of the eight treatment sequences when $\lambda = 0$. When $\lambda = 1$, adaptive designs assign an approximately equal number of subjects to a dual block (a treatment sequence and its dual with treatments in a reverse order), and more subjects are given to ABB and its dual, which the fixed optimal design uses for comparing two treatments under the traditional model. However, adaptive designs use six of the eight sequences rather uniformly.

When treatment differences are present, the mean vector of each treatment sequence, in decreasing order, is AAA, BAA, AAB (equivalently, ABA), BAB (equivalently, BBA), ABB , and BBB . Table 2 shows that, when $\lambda = 1$, adaptive designs assign an approximately equal number of subjects to a dual block, and more subjects are given to ABB and its dual. However, as $\lambda < 1$ and decreasing, adaptive designs assign more subjects to the best treatment sequence AAA and fewer subjects to the worst treatment sequence BBB . This is consistent with the setting. Similar to the two-period case, the allocation is rapidly skewed as λ decreases. When treatment differences are present, the impact of using low λ is evident. If the emphasis is placed on increasing the treatment benefit ($\lambda = 0$), Design AAA is optimal, followed by BAA .

Figure 1 shows the estimated relative efficiency (RE) of adaptive designs relative to the reference design using two sequences ABB and its dual, the popular fixed optimal design (under the traditional model) for estimating $\theta = (\tau, \gamma, \phi)^T$ under A-, D-, and E-optimality, and for estimating the treatment contrast τ , respectively. Adaptive design is shown more efficient than the reference design if $RE > 1$. Fig. 1 illustrates that response-adaptive designs are more efficient than the traditional design for large λ under

Table 1 Estimated Numbers of Patients Receiving Each Treatment Sequence Using Adaptive Two-Period Designs: Continuous Outcome

N	λ	N_{AA}		N_{AB}		N_{BA}		N_{BB}	
		I	II	I	II	I	II	I	II
10	1	2.49	2.50	2.51	2.50	2.44	2.45	2.56	2.55
	0.9	2.49	3.00	2.51	2.00	2.50	3.00	2.50	2.00
	0.7	2.49	3.00	2.51	2.00	2.50	3.00	2.50	2.00
	0.3	2.49	3.25	2.51	1.97	2.50	2.99	2.50	1.79
	0	2.50	6.49	2.53	1.01	2.51	1.50	2.47	1.00
40	1	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
	0.9	10.00	10.72	10.00	9.51	10.00	10.52	10.00	9.25
	0.7	10.01	12.66	10.00	8.36	10.00	11.33	9.99	7.65
	0.3	10.07	21.61	9.96	4.06	9.98	11.65	10.00	2.68
	0	10.27	35.63	9.86	1.01	10.09	2.36	9.78	1.00
80	1	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00
	0.9	20.00	22.80	20.00	18.22	20.00	21.60	20.00	17.38
	0.7	20.00	30.12	20.01	14.04	20.02	24.39	19.97	11.45
	0.3	20.00	54.30	20.03	4.75	20.06	18.09	19.91	2.86
	0	19.76	74.68	20.46	1.02	20.55	3.30	19.23	1.00
100	1	25.00	25.00	25.00	25.00	25.00	25.00	25.00	25.00
	0.9	25.01	29.30	24.99	22.47	25.00	27.26	25.00	20.97
	0.7	25.01	40.36	24.99	16.10	25.02	31.04	24.99	12.50
	0.3	25.22	72.20	24.89	4.93	25.04	20.00	24.86	2.87
	0	24.69	93.97	25.57	1.02	24.81	4.01	24.93	1.00

Note: Entries in the columns I are based on 5,000 computer replications under two-treatment two-period repeated measurement designs with $\pi = \tau = \gamma = \phi = 0$, and those to the right are for $\pi = \tau = \phi = 2.5$, $\gamma = -2.5$, with $\sigma_\xi^2 = 2$, $\sigma_\varepsilon^2 = 1$, and $\mu = 100$.

Table 2 Estimated Numbers of Patients Receiving Each Treatment Sequence Using Adaptive Three-Period Designs: Continuous Outcome

N	λ	N_{AAA}		N_{AAB}		N_{ABA}		N_{ABB}		N_{BBB}		N_{BBA}		N_{BAB}		N_{BAA}	
		I	II	I	II	I	II	I	II	I	II	I	II	I	II	I	II
10	1	1.00	1.00	1.01	1.00	1.48	1.49	1.51	1.51	1.00	1.00	1.01	1.01	1.48	1.48	1.51	1.51
	0.9	1.00	1.00	1.01	1.00	1.50	1.80	1.49	1.20	1.00	1.00	1.01	1.00	1.49	1.19	1.50	1.81
	0.7	1.00	1.00	1.02	1.00	1.53	1.92	1.45	1.08	1.00	1.00	1.02	1.00	1.52	1.09	1.46	1.91
	0.3	1.01	1.17	1.07	1.07	1.47	1.73	1.45	1.03	1.01	1.00	1.07	1.01	1.46	1.07	1.46	1.92
	0	1.26	2.51	1.25	1.05	1.26	1.06	1.25	1.00	1.24	1.00	1.25	1.01	1.23	1.00	1.26	1.37
40	1	1.01	1.01	5.99	5.99	5.97	5.97	7.03	7.03	1.01	1.01	5.99	5.99	5.97	5.97	7.03	7.03
	0.9	1.01	1.01	6.01	6.25	5.96	6.85	7.02	6.04	1.01	1.01	6.01	5.21	5.96	5.28	7.03	8.35
	0.7	1.23	1.36	5.71	7.62	6.00	8.49	7.06	3.50	1.25	1.11	5.70	3.96	5.98	3.17	7.08	10.79
	0.3	2.51	11.27	4.90	3.82	6.01	6.94	6.56	1.29	2.55	1.01	4.83	1.44	5.96	1.64	6.68	12.59
	0	4.98	29.54	5.01	1.15	5.04	1.13	5.03	1.00	5.06	1.00	5.03	1.01	4.80	1.01	5.05	4.16
80	1	1.01	1.01	13.00	13.00	11.84	11.84	14.16	14.15	1.01	1.01	13.00	13.00	11.84	11.84	14.16	14.15
	0.9	1.03	1.03	13.04	15.34	11.73	14.81	14.19	9.88	1.03	1.02	13.04	11.05	11.75	8.36	14.18	18.51
	0.7	2.05	3.46	11.99	16.95	11.84	18.75	14.11	3.84	2.06	1.25	11.99	5.95	11.90	3.58	14.06	26.22
	0.3	4.91	38.18	9.57	4.61	11.66	9.22	13.78	1.33	4.87	1.02	9.98	1.60	12.10	1.75	13.13	22.29
	0	9.96	65.93	9.80	1.19	10.21	1.17	10.42	1.00	9.61	1.00	10.16	1.01	9.97	1.01	9.87	7.69
100	1	1.01	1.01	16.45	16.46	14.61	14.60	17.93	17.93	1.01	1.01	16.45	16.46	14.61	14.60	17.93	17.93
	0.9	1.08	1.08	16.53	20.23	14.64	19.38	17.75	11.02	1.07	1.05	16.54	13.49	14.66	9.31	17.73	24.44
	0.7	2.58	6.03	14.96	20.76	14.74	23.55	17.68	3.88	2.54	1.27	15.09	6.31	14.91	3.71	17.50	34.49
	0.3	6.14	54.47	12.21	4.68	14.79	9.62	16.81	1.32	5.95	1.02	12.41	1.64	15.13	1.81	16.56	25.44
	0	12.29	84.89	12.82	1.19	12.60	1.17	13.00	1.00	11.68	1.00	12.93	1.01	12.16	1.01	12.52	8.73

Note: Entries are based on 5,000 computer replications under two-treatment three-period repeated measurement designs with $\pi_2 = \pi_3 = \tau = \gamma = \phi = 0$ (columns I) and $\pi_2 = \pi_3 = \tau = \phi = 2.5$, $\gamma = -2.5$ (columns II) respectively, with $\sigma_\xi^2 = 2$, $\sigma_\varepsilon^2 = 1$, and $\mu = 100$.

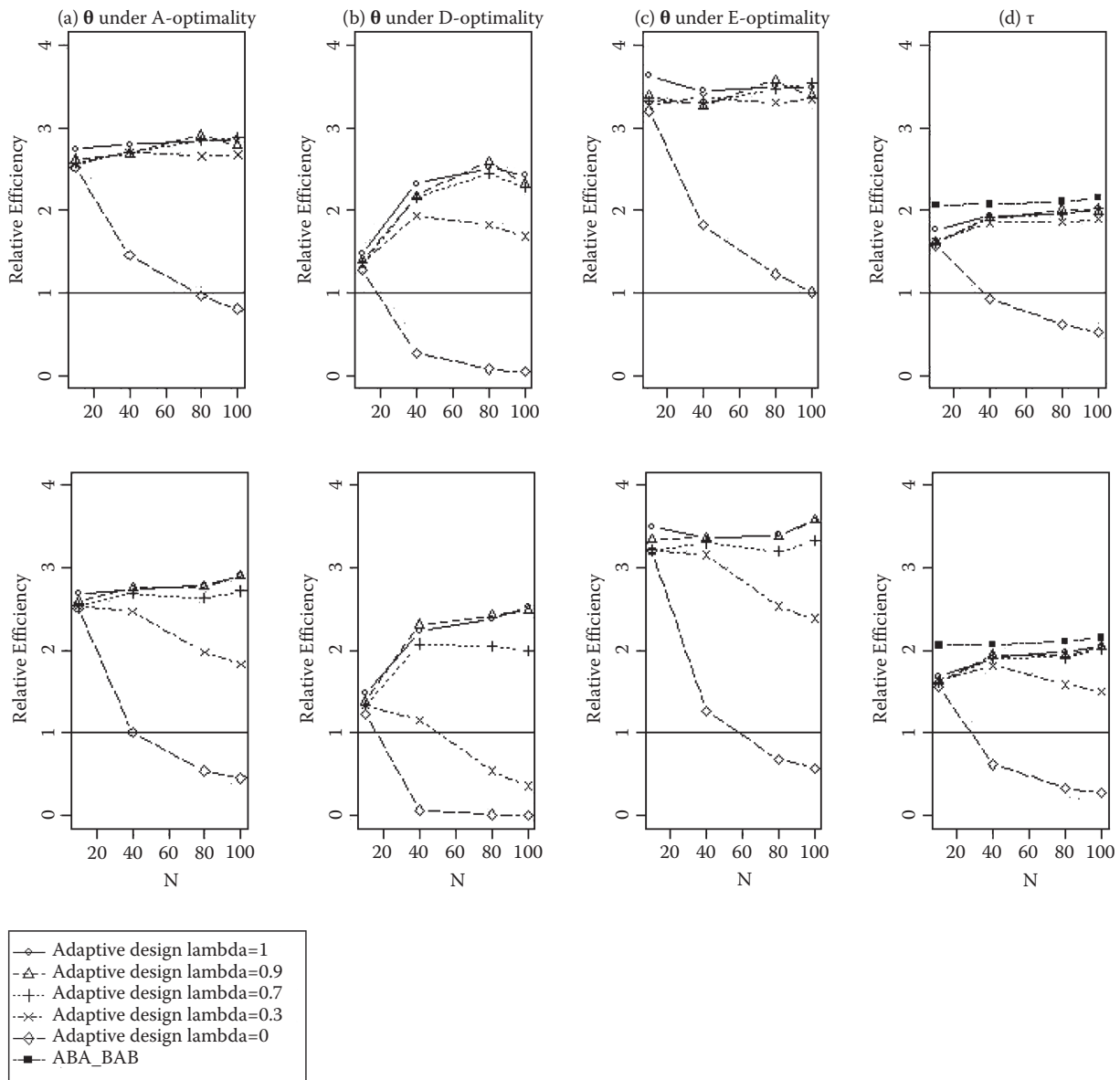


Fig. 1 Relative efficiencies of adaptive two-treatment three-period designs: continuous outcomes

Note: The fixed design *ABB/BBA* is the reference design. Relative efficiencies in the upper row are calculated by using 5,000

computer replications with $\pi_2 = \pi_3 = \tau = \gamma - \phi = 0$; the bottom row with $\pi_2 = \pi_3 = \tau = \phi = 2.5$, $\gamma = -2.5$. And in all cases, we assume

$\sigma_\epsilon^2 = 2$, $\sigma_\tau^2 = 1$, and $\mu = 100$. $\theta = (\tau, \gamma, \phi)^T$. Since Design *ABA/BAB* cannot estimate the self carryover effect contrast $\left(\varphi = \frac{\phi_A - \phi_B}{2}\right)$, the efficiency is compared only for the direct treatment effect contrast $\left(\tau = \frac{\tau_A - \tau_B}{2}\right)$

all three optimality criteria. However, when $\lambda = 0$, adaptive designs are generally less efficient in terms of MSE. When a treatment effect is absent (upper row of Fig. 1), adaptive designs with $\lambda = 0$ utilize all possible treatment sequences equally (Table 2), which apparently reduces design efficiency; when a treatment effect is present (bottom row of Fig. 1), adaptive designs with $\lambda = 0$ use treatment sequence *AAA* heavily, followed by *BAA* and the unbalance clearly reduces the design efficiency, measured by MSE, although it is the best for estimating θ .

For estimating the direct treatment contrast, the design using *ABA* and *BAB* is the best because it is the optimal design under model (2)^[13]. However, this design cannot estimate self-carryover effects. Nonetheless, adaptive designs with a large value of λ are nearly as efficient as the fixed design (Fig. 1, panel d), while also taking the treatment performance into consideration. In general, using a large λ has no real consequences for the traditional notion of design efficiency, especially when the sample size is large. Use of $\lambda = 0$ has low efficiency relative to the reference

Table 3 Estimated Numbers of Patients for Each Treatment Sequence under Two-Treatment Two-Period Design: Binary Outcome

N	λ	N_{AA}		N_{AB}		N_{BA}		N_{BB}	
		I	II	I	II	I	II	I	II
12	1	3.09	3.12	2.91	2.84	2.91	2.94	3.09	3.10
	0.9	3.05	3.11	2.94	2.88	2.95	3.02	3.06	2.99
	0.7	3.00	3.26	2.97	2.90	2.97	2.98	3.06	2.86
	0.3	3.04	3.52	2.95	2.90	3.00	2.97	3.01	2.61
	0	2.93	3.55	2.88	2.83	2.99	2.98	3.20	2.64
40	1	11.23	11.18	8.77	8.82	8.77	8.82	11.23	11.18
	0.9	10.29	11.51	9.65	9.62	9.72	9.74	10.34	9.13
	0.7	10.09	13.23	9.57	9.54	9.99	9.67	10.35	7.56
	0.3	9.99	14.15	10.08	9.45	10.21	9.64	9.72	6.76
	0	9.80	13.87	9.65	9.16	10.17	10.08	10.38	6.89
80	1	23.19	23.25	16.81	16.75	16.81	16.75	23.19	23.25
	0.9	20.30	25.23	19.76	19.15	19.77	19.23	20.17	16.39
	0.7	20.65	30.39	19.37	18.24	19.07	19.09	20.91	12.28
	0.3	19.62	31.19	20.30	18.38	20.38	18.86	19.70	11.57
	0	19.46	29.55	19.28	18.68	20.54	19.81	20.72	11.96
100	1	29.00	29.11	21.00	20.89	21.00	20.89	29.00	29.11
	0.9	25.16	33.07	25.03	23.60	25.19	23.87	24.62	19.46
	0.7	24.84	38.87	24.93	23.73	24.99	24.13	25.24	13.27
	0.3	24.76	40.15	25.08	22.78	24.44	23.10	25.72	13.97
	0	24.89	38.08	23.73	22.70	25.32	25.08	26.06	14.14

Note: Entries are based on 5,000 simulations with $\pi_r = 0.5, r = 1, 2, 3, 4$ (columns I) and $\pi_1 = 0.5, \pi_2 = 0.6, \pi_3 = 0.4$, and $\pi_4 = 0.5$ (columns II).

design, as N increases. These extreme results are due to implementing an objective while the efficiency is measured by a MSE, different from a measure of treatment efficacy.

Dichotomous Outcomes

Based on the multivariate Bernoulli model as described in section “Dichotomous Outcomes”, we construct response-adaptive two-treatment two-and three-period repeated measurement designs. We note here that the literature is sparse on optimal design methodologies for dichotomous responses. Clinical trials with an interest in binary or nominal outcome measurements have adopted optimal designs constructed for continuous responses. Various related research have dealt with how to efficiently analyze binary data upon adopting the traditional optimal design constructed for continuous responses under assumptions, inappropriate for actual measurements of interest. This chapter examines the merits of such approaches and construct adaptive designs, which are the most efficient under a given setting.

For two-period designs, the two sets of parameters are: (i) $\pi_r = 0.5$, for $r = 1, 2, 3, 4$ (no treatment difference); and (i) $\pi_1 = 0.5, \pi_2 = 0.6, \pi_3 = 0.4$, and $\pi_4 = 0.5$ (with treatment difference). For three-period designs, the two sets of parameters are: (i) $\pi_r = 0.5, r = 1, 2, \dots, 6$ (no treatment difference);

and (ii) $\pi_1 = 0.5, \pi_2 = 0.6, \pi_3 = 0.7, \pi_4 = 0.4, \pi_5 = 0.5$, and $\pi_6 = 0.6$ (with treatment difference). We assume that at the initial stage four patients were assigned for two-period designs and eight patients were assigned for three-period designs, one for each of the treatment sequences. We then consider allocating the rest of the patients adaptively.

Two-Treatment Two-Period Designs

Table 3 summarizes the estimated average number of patients receiving each treatment sequence based on the 5,000 simulation results. We can see that in the case of no treatment effect (columns I), the strategy assigns an equal number of subjects to a treatment sequence and its dual when $\lambda = 1$, with a slight preference to treatment sequences AA and BB , indicating no need for complex administration in the absence of treatment effects. When $\lambda < 1$, an approximately equal number of subjects is assigned to each of the four treatment sequences.

When there are treatment effects with $\lambda = 1$, the patterns are similar to the equal success probability case, with equal numbers of subjects to a dual block, allocating more subjects to treatment sequences AA and BB . However, when $\lambda < 1$, the adaptive design successfully assigns more patients to the best treatment sequence, which is a treatment

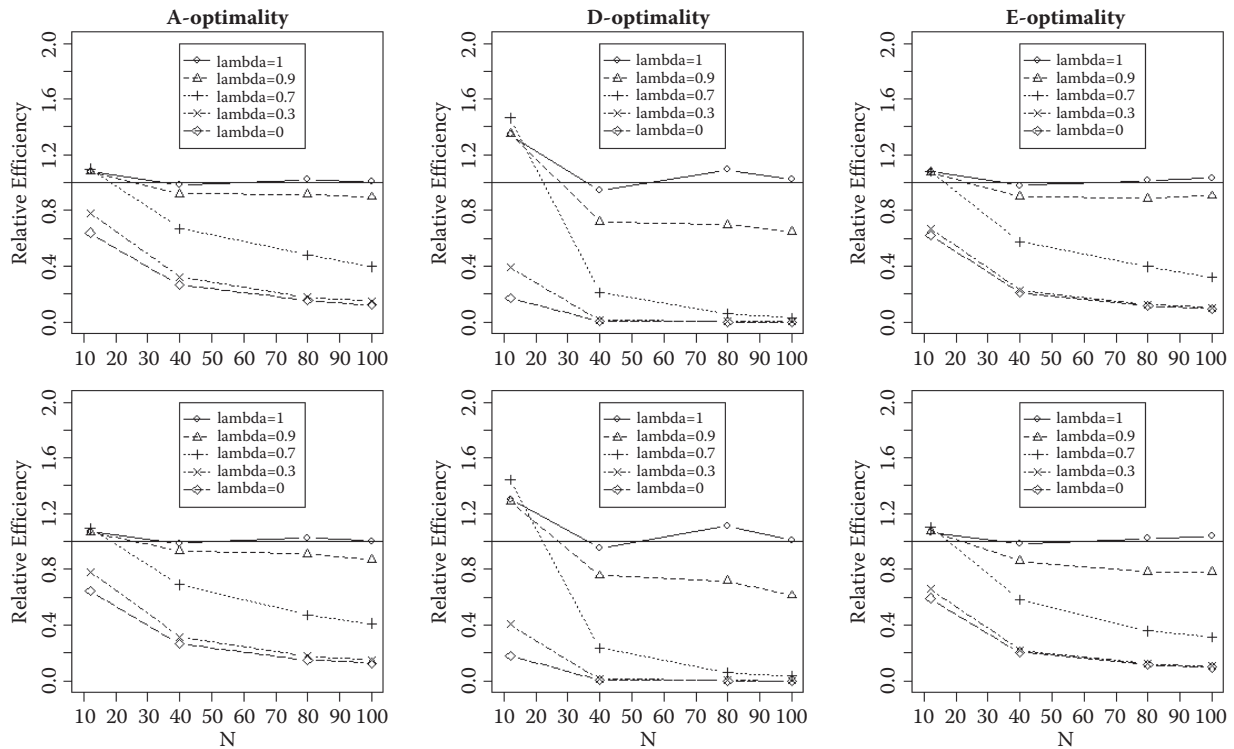


Fig. 2 Relative efficiency of θ for two-period two-treatment designs: binary outcome

Note: $\theta = (\pi_1, \pi_2, \pi_3, \pi_4)^T$. The fixed design $AB/BA/AA/BB$ is the reference design. Relative efficiencies are calculated on the basis of 5,000 simulations with $\pi_r = 0.5, r = 1, 2, 3, 4$ (upper row), and $\pi_1 = 0.5, \pi_2 = 0.6, \pi_3 = 0.4$, and $\pi_4 = 0.5$ (bottom row)

sequence AA followed by BA, AB and BB , even when sample size is small (e.g., $N = 12$). In addition, as N increases, the proportion of patients receiving the best treatment increases; whereas the proportion of patients receiving the worst treatment (BB) decreases. But the allocation results are more or less uniform unlike the continuous data case for the treatment effect setting considered.

Figure 2 illustrates the relative efficiency of the adaptive design to the reference design (the fixed optimal design for continuous data). Again we see that the design with a large λ serves both purposes quite well. Designs with a small value of λ are not efficient according to the conventional measure, but are no doubt the best strategy under another measure. Adopting the fixed optimal design for continuous responses will clearly be a misleading strategy.

The success probabilities of treatments in different periods and the estimation precision are also calculated for both equal and unequal treatment cases. Although not shown, they both show that the spread of the estimates of π_r decreases when the total number of patients increases, consistent with the setting.

Two-Treatment Three-Period Designs

Again, we observe that the similar assignment of patients is observed as in the continuous case, but in a lesser degree (Table 4). When all treatments are equally effective with

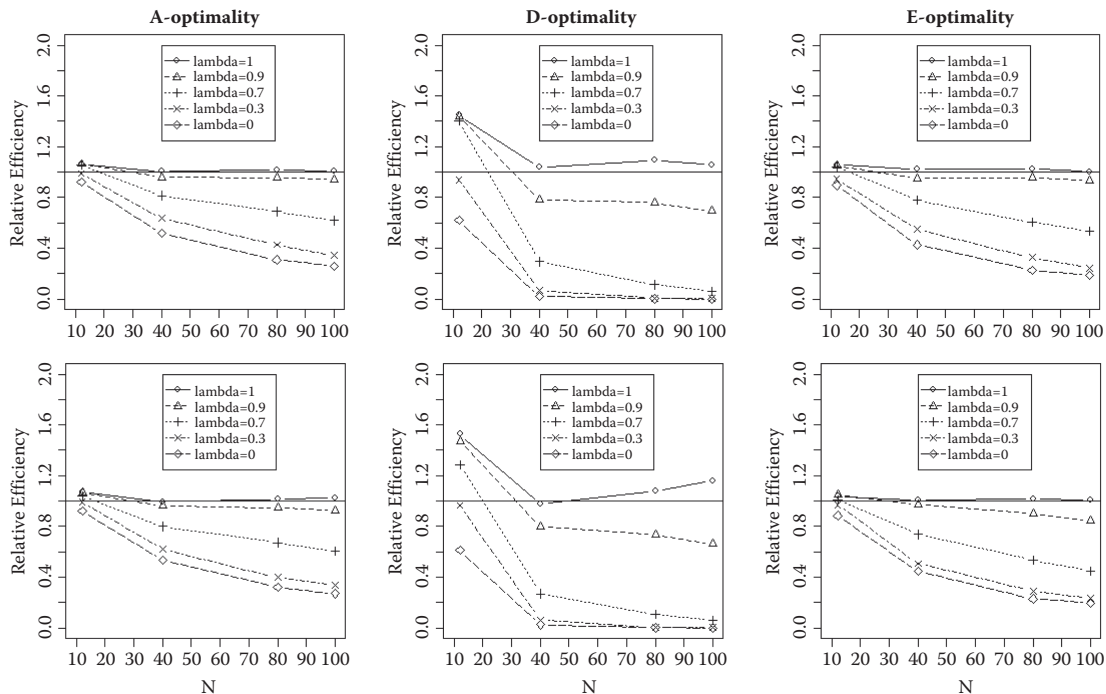
$\pi_r = 0.5, r = 1, 2, \dots, 6$ and when $\lambda < 1$, an approximately equal number of patients are assigned to each treatment sequence. When $\lambda = 1$, an equal number of patients are assigned to dual blocks with treatment sequences ABB and its dual having slightly more patients than other blocks, similar to the continuous case, but in a much less degree. When the success probabilities are not equal with $\pi_1 = 0.5, \pi_2 = 0.6, \pi_3 = 0.7, \pi_4 = 0.4, \pi_5 = 0.5$, and $\pi_6 = 0.6$, adaptive designs still use all eight sequences rather significantly. However, the impact of using low λ is not as pronounced as in the continuous case. Nonetheless, if the emphasis is placed on increasing the treatment benefit ($\lambda = 0$), it is best to allocate more to AAA , followed by BAA , and BBB is the worst, as it is the least effective treatment sequence.

These findings are evident in the relative efficiencies against the fixed optimal design with sequences ABB and BAA with an equal number of subjects per treatment sequence. Figure 3 illustrates the relative efficiency of the adaptive and reference designs under the A-, D-, and E- optimality, respectively. Similar to the two-period designs, the adaptive design with $\lambda < 1$ has low efficiency compared to the reference design. However, since the reference design is not necessarily the optimal choice in this situation, our conclusion is such that clinical trials must be carefully planned with a clear idea on not only the covariance structure^[12] but also the type and scale of outcome measurements desired and the trial objectives.

Table 4 Estimated Numbers of Patients for Each Treatment Sequence under Two-Treatment Three-Period Design: Binary Outcome

N	λ	N_{AAA}		N_{AAB}		N_{ABA}		N_{ABB}		N_{BBB}		N_{BBA}		N_{BAB}		N_{BAA}	
		I	II	I	II	I	II	I	II	I	II	I	II	I	II	I	II
12	1	1.33	1.37	1.39	1.37	1.50	1.52	1.78	1.71	1.32	1.31	1.40	1.42	1.50	1.49	1.78	1.81
	0.9	1.47	1.56	1.50	1.47	1.50	1.51	1.51	1.46	1.49	1.44	1.50	1.52	1.51	1.48	1.52	1.56
	0.7	1.48	1.60	1.51	1.50	1.50	1.53	1.51	1.44	1.47	1.38	1.50	1.51	1.52	1.46	1.51	1.58
	0.3	1.48	1.63	1.48	1.50	1.48	1.55	1.51	1.45	1.53	1.35	1.49	1.46	1.52	1.45	1.51	1.61
	0	1.43	1.55	1.43	1.49	1.42	1.46	1.48	1.42	1.47	1.36	1.54	1.47	1.59	1.52	1.64	1.73
40	1	3.84	4.10	4.30	4.44	5.07	5.07	6.79	6.39	3.84	4.10	4.30	4.44	5.08	5.07	6.78	6.39
	0.9	5.13	5.70	5.04	4.97	4.94	5.23	4.88	4.82	5.14	4.58	5.08	4.64	4.91	4.88	4.88	5.18
	0.7	4.95	6.50	4.88	5.47	4.91	5.41	4.98	4.45	5.11	3.69	4.89	4.49	5.14	4.55	5.14	5.44
	0.3	4.97	6.97	5.17	5.63	4.98	5.67	5.08	4.27	5.10	3.36	4.90	4.28	4.85	4.17	4.95	5.65
	0	4.92	7.20	4.93	5.38	4.83	5.39	5.04	4.20	4.99	3.31	4.97	4.22	5.06	4.29	5.26	6.01
80	1	7.67	8.17	8.72	8.92	10.15	10.10	13.46	12.81	7.67	8.18	8.72	8.91	10.14	10.10	13.47	12.81
	0.9	10.03	12.18	9.87	10.41	10.07	10.90	10.05	9.13	9.89	8.27	9.99	9.06	10.09	9.30	10.01	10.75
	0.7	10.10	15.99	9.96	10.96	10.02	11.46	9.80	7.82	9.95	6.09	10.30	8.17	9.77	8.21	10.10	11.30
	0.3	9.96	17.15	10.38	11.73	9.67	11.72	10.13	7.71	9.80	5.03	9.79	7.52	9.97	7.36	10.30	11.78
	0	9.50	17.45	9.99	11.39	9.77	11.57	9.45	7.33	9.98	4.98	10.14	7.61	10.60	7.87	10.57	11.80
100	1	9.70	10.28	10.92	11.11	12.70	12.66	16.68	15.94	9.70	10.28	10.92	11.12	12.70	12.67	16.68	15.94
	0.9	12.74	16.32	12.07	13.71	12.70	12.84	12.58	11.13	12.48	9.94	12.29	11.76	12.96	11.11	12.18	13.19
	0.7	12.26	21.26	12.31	13.93	12.74	14.26	12.34	9.68	12.60	7.16	12.57	9.91	12.82	9.64	12.36	14.16
	0.3	12.80	23.59	12.08	14.72	12.57	14.82	12.38	8.82	12.47	5.68	12.82	8.84	12.38	8.56	12.50	14.97
	0	12.19	22.91	12.73	13.88	11.92	14.52	12.44	9.23	12.19	5.52	12.32	9.82	12.86	9.06	13.35	15.06

Note: Entries are based on 5,000 simulations with $\pi_r = 0.5, r = 1, 2, \dots, 6$ (columns I) and $\pi_1 = 0.5, \pi_2 = 0.6, \pi_3 = 0.7, \pi_4 = 0.4, \pi_5 = 0.5$ and $\pi_6 = 0.6$ (columns II).

**Fig. 3** Relative efficiency of θ for three-period two-treatment designs: binary outcome

Note: $\theta = (\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6)^T$. The fixed design ABB/BAA is the reference design. Relative efficiencies are calculated on the basis of 5,000 simulations with $\pi_r = 0.5, r = 1, 2, \dots, 6$ (upper row) and $\pi_1 = 0.5, \pi_2 = 0.6, \pi_3 = 0.7, \pi_4 = 0.4, \pi_5 = 0.5$ and $\pi_6 = 0.6$ (bottom row)

CONCLUSIONS

In this paper, we discussed the design allocation strategy to construct adaptive repeated measurement designs with both continuous and dichotomous responses/outcomes. We provide a detailed allocation rule for constructing practically useful repeated measurement designs adaptively for two-treatment and two-and three-period trials. We demonstrated that, as expected, the adaptive designs may not be as efficient as the fixed designs in terms of the mean squared error, but they successfully allocate more patients to better treatment sequences, no doubt the best strategy depending upon the study goals (Tables 1 to 4). In dichotomous outcome case, adaptive designs use all treatment sequences and the clinical trials simply adopting the conventional fixed optimal designs constructed for continuous responses may fail to achieve the study goal.

The investigator can predetermine the value of λ to balance the two objectives of increasing estimation precision and decreasing the proportion of patients receiving inferior treatments. A large value of λ places relatively heavy emphasis on the estimation precision. When $\lambda = 1$, the allocation rule becomes the usual response adaptive design as considered by other researchers.^[2] A small value of λ emphasizes the performance/benefit of the treatment. When $\lambda = 0$, the allocation rule becomes a typical play-the-winner rule.^[4] In addition, our simulated experiments show that the design with a high value of $\lambda < 1$ can result in allocating patients to effective treatment sequences without much loss of estimation precision.

The allocation rule can be generalized to construct adaptive t -treatment p -period repeated measurement designs. The main challenge is to narrow down the number of treatment sequences out of t^p possibilities, which may increase substantially as the number of treatments and periods increase. We may wish to keep the trials manageable by reducing the number of sequences, although this choice has the danger of possibly omitting some relevant sequences.

The value of λ can be chosen by investigators and is used to balance the two objectives of increasing estimation precision and decreasing the proportion of patients

receiving inferior treatment sequences. We find that design efficiency is a skewed function of λ that decreases sharply as λ decreases.

REFERENCES

1. Hu, F.; Rosenberger, W.F. *The Theory of Response-Adaptive Randomization in Clinical Trials*; Wiley and Sons: New York, 2006.
2. Kushner, H.B. Allocation rules for adaptive repeated measurements designs. *J. Stat. Plan. Inference* **2003**, *113*, 293–313.
3. Armitage, P. *Sequential Medical Trials*; Blackwell: Oxford, 1975.
4. Zelen, M. Play the winner rule and the controlled clinical trial. *J. Am. Stat. Assoc.* **1969**, *64*, 131–146.
5. Wei, L.J.; Durham, S. The randomized play-the-winner rule in medical trials. *J. Am. Stat. Assoc.* **1978**, *73*, 840–843.
6. Rosenberger, W.F.; Stallard, N.; Ivanova, A.; Harper, C.; Ricks, M. Optimal adaptive designs for binary response trials. *Biometrics* **2001**, *57*, 909–913.
7. Liang, Y.; Carriere, K.C. Multiple-objective response-adaptive repeated measurement designs for clinical trials. *J. Stat. Plan. Inference* **2009**, *139*, 1134–1145.
8. Laska, E.M.; Meisner, M.; Kushner, H.B. Optimal crossover designs in the presence of carryover effects. *Biometrics* **1983**, *39*, 1087–1091.
9. Kershner, R.P. Optimal 3-period 2-treatment crossover designs with and without baseline measurements. *Proc. Biopharm. Sec. Am. Stat. Assoc.* **1986**, *81*, 152–156.
10. Atkinson, A.C.; Donev, A.N.; Tobias, R.D. *Optimum Experimental Designs, with SAS*. Oxford University Press: Oxford, 2007.
11. Hedayat, A.; Afsarinejad, K. Repeated measurement designs, II. *Ann. Stat.* **1978**, *6* (3), 619–628.
12. Carrière, K.C. Cross-over designs for clinical trials. *Stat. Med.* **1994**, *13*, 1063–1069.
13. Afsarinejad, K.; Hedayat, A.S. Repeated measurement designs for a model with self and mixed carryover effects. *J. Stat. Plan. Inference* **2002**, *106*, 449–459.
14. Liang, Y.; Carriere, K.C. Multiple-Objective response-adaptive repeated measurement designs for clinical trials with dichotomous outcomes. *Proceedings of the 6th Workshop on Simulation, St. Petersburg, Russia*, 1084–1090.

Clinical Trials: Selection of Control

Irving K. Hwang

Irving Consulting Group, Pluckemin, New Jersey, U.S.A., and University of Medicine and Dentistry of New Jersey (UMDNJ), New Brunswick, New Jersey, U.S.A.

Abstract

This entry discusses the importance of a control group or placebo in a clinical trial by presenting historical information and statistical indicators supporting their importance, and also offers useful alternatives and design modifications.

INTRODUCTION

Why is control necessary in clinical trials? The rationale is being well documented in ICH E-10 Guidance—Choice of Control Group and Related Design Issues in Clinical Trials.^[1] First, the control serves as a comparator that allows for discrimination of patient outcomes (changes in symptoms, signs, morbidity, or mortality) caused by the test treatment from outcomes caused by other factors, such as the natural progression of the disease, patient and/or investigator expectations, or other treatment. The experience gained on control provides the critical information, if the patients had not been treated with the test treatment or if they had been treated with a different treatment (e.g., a standard effective drug). Second, the inclusion of control can minimize bias through randomization and blinding. “Bias,” as defined in ICH E9 Guidance—Statistical Principles for Clinical Trials,^[2] means the systematic tendency of any aspects of the design, conduct, analysis, and interpretation of the results of clinical trials to make the estimate of a treatment effect deviate from its true value.

Control can be “historical” or “concurrent.” A historical control is a control selected from a defined patient population in a similar group of patients studied in historical trials. The patient population may no longer be the same due to medical care advancement and life style changes via health awareness. A concurrent control is one selected from the same patient population as the test treatment group and treated in the same trial concurrently in adherence to the study protocol. To ensure that the test and control groups are comparable at the trial start, randomization is usually used to minimize the chance of patient selection bias. To ensure that the test and control groups are similarly treated and evaluated during the study, blinding is usually used to minimize the treatment and evaluation bias. Whether a trial design includes these features is a critical determinant of its quality and integrity, especially for a Phase IIb–III confirmatory trial.

TYPES OF CONTROLS

As depicted in ICH E-10, controls in clinical trials can be classified on the basis of the treatment received and whether the test and chosen control groups are treated concurrently vs. externally/historically. There are generally five types. The first four (i.e., no treatment control, placebo control, active control, and dose–response control) are concurrent controls where the test and control groups are selected from the same patient population and treated concurrently usually via randomization of treatments. A variation of the four concurrent controls on the treatment received is multiple control (e.g., include both placebo and active controls in a single trial). The fifth type is external/historical controls. Because of lack of randomization and patient comparability of the test and control groups and its inability to minimize bias, this type of control is used only in exceptional cases. The five types of controls are no-treatment concurrent control, placebo concurrent control, active concurrent control, dose–response concurrent control, and external (historical) control.

No-Treatment Concurrent Control

In a no-treatment controlled trial, patients are usually randomly assigned to test treatment or no study treatment, e.g., test treatment vs. surgery or watchful waiting. Patients and investigators are not blinded to treatment assignment. Because the trial cannot be fully blinded, this can affect all aspects of the trial, including patient retention, patient management, and all aspects of endpoint assessment. However, it is often possible to have a blinded observer/rater conduct the endpoint assessment. This blind-observer approach should always be considered in trials that cannot be blinded, but it does not solve the other problems associated with knowing the treatment assignment such as investigator and/or patient expectation.

Placebo Concurrent Control

The placebo-controlled, double-blind trial has been the “gold” standard in drug development for many decades. The utilities of placebo-controlled trials are well established. In this design, patients are randomly assigned to a test drug or to an identical-appearing inactive drug (i.e., placebo). A placebo is a “dummy” drug that appears as identical as possible to the test drug with respect to physical characteristics such as color, weight, taste, and smell, but it does not contain the test drug or any other active ingredient of drugs. Such trials are almost always double-blind, with both patients and investigator unaware of treatment assignment. The treatments can be titrated or given at one or multiple, fixed doses. The placebo concurrent control design, by including a group that receives a dummy treatment, via randomization and blinding, controls for the so-called “placebo” effect. Simply, the placebo effect indicates improvement in a patient resulting from being treated other than from the true pharmacologic effect of an active drug. Therefore, the use of placebo concurrent control will be able to separate the true pharmacologic effect of the test drug from the other confounded effects (e.g., natural progression of the disease, regression to the mean, patient and/or investigator expectation, the effect of being treated in a trial, and subjectiveness of diagnosis/assessment). Placebo-controlled trials also have the ability to distinguish adverse effects caused by a drug from those resulting from underlying disease or intercurrent illness. The placebo-controlled trial, using randomization and blinding, generally minimizes patient and investigator bias. However, use of a placebo control may raise problems of ethics, acceptability, and feasibility, when an effective treatment is available for the condition (e.g., mortality and/or irreversible morbidity) under study. It is often possible to address the ethical or practical limitations of placebo-controlled trials by using modified study designs such as inclusion of multiple doses of the test drug or a known active-control drug, including the use of a pseudo-placebo (an ineffective low-dose, active-control drug).

Active Concurrent Control

An active-controlled (positive-controlled) trial is one in which a test treatment is compared with a known effective active treatment. Active-controlled trials have two distinct objectives with respect to demonstrating efficacy: (1) to demonstrate efficacy by showing superiority of the test treatment to the active-control treatment or (2) to show efficacy of the test treatment by showing it is non-inferior or equivalent to a known effective treatment. In an active (positive)-controlled trial, patients are randomly assigned to the test treatment or to an active-control treatment. Such trials are usually double-blind via the use of double dummies, but this is not always feasible. Many oncology trials, e.g., are difficult or impossible to blind

because of different regimens, routes of administration, and toxicities. Active-controlled trials are generally considered to pose fewer ethical and practical problems than placebo-controlled trials because all patients receive active treatment. Active-controlled trials also can, if properly designed, provide information about relative efficacy. The most crucial design question for an active-controlled trial is whether the trial is intended to show superiority or to show noninferiority/equivalence. When this design is used to show superiority of test treatment to an active-control drug, the approach is straightforward like the placebo-controlled trials. When superiority (significant difference between the test and active-control treatments) is shown, “assay sensitivity,” as defined in ICH E-10, of the trial is established, and efficacy of the test treatment is demonstrated. However, when it is used to show noninferiority/equivalence of the test treatment to an active-control treatment, assay sensitivity will need special consideration; it cannot be demonstrated directly, but rather deduced via “historical evidence of sensitivity-to-drug-effects” and “appropriate trial conduct,” as they are well addressed in ICH E-10. The discussion of these important issues will be deferred to the next section.

Dose–Response Concurrent Control

In a randomized, fixed-dose, dose–response-controlled trial, patients are randomized to one of several fixed-dose test treatment groups via titration or fixed dosing. Dose–response trials are usually double-blind. Dose–response trials are carried out to establish the relation between dose and efficacy and adverse effects and/or to demonstrate efficacy. Evidence of efficacy could be based on significant differences in pairwise comparisons between dosing groups or between dosing groups and placebo, or on evidence of a significant positive trend with increasing dose, even if no two groups are significantly different. In addition to fixed-dose test treatment groups, it may include a placebo (zero dose) and/or fixed-dose active-control treatment groups. A potential problem is that a positive dose–response trend (i.e., a significant regression of the response on dose), without significant pairwise differences, can establish test treatment efficacy, but may leave uncertainty as to which doses (other than the maximum dose) are actually effective.

External (Historical) Control

An external- or historical-controlled trial compares a group of patients receiving the test treatment with a group of patients external to the study, rather than to an internal control group consisting of patients from the same patient population randomly assigned to a different treatment. The external (historical) control can be a group of patients treated at an earlier time (historical control) or a group treated during the same time period but in a

different setting (external control). Lack of randomization and inability to minimize bias, the utilities of external/historical control are generally limited.

ASSAY SENSITIVITY, HISTORICAL EVIDENCE OF SENSITIVITY-TO-DRUG-EFFECTS, AND NONINFERIORITY MARGIN

As defined in ICH E-10, assay sensitivity is a property of a clinical trial defined as the ability to distinguish an effective treatment from a less effective or ineffective treatment. Assay sensitivity is important in any trial but has different implications for trials intended to demonstrate superiority vs. noninferiority/equivalence. (Note that most active-controlled equivalence trials are practically noninferiority trials intended to establish the efficacy of a new treatment. The new treatment can be shown noninferior as well as superior to the active-control drug, not limited to equivalence per se. Hence, the term of equivalence will be dropped hereafter in this document.) In the superiority trial setting whether the control is placebo or an active drug, assay sensitivity refers to the ability of a specific trial to detect a difference between treatments, if one exists. For active-controlled noninferiority trials, assay sensitivity requires the presence of a control drug effect of a minimum size so that a specific trial, properly designed and conducted, has the ability not to falsely conclude an ineffective treatment noninferior. Because the actual effect size of the active control in the trial is not measured relative to placebo, the presence of assay sensitivity must be deduced. If a trial intended to demonstrate efficacy by showing superiority of a test treatment to control lacks assay sensitivity, it will fail to show that the test treatment is superior and will fail to lead to a conclusion of efficacy. In contrast, if a trial intended to demonstrate efficacy by showing a test treatment to be noninferior to an active control lacks assay sensitivity, the trial may find an ineffective treatment to be noninferior and could lead to an erroneous conclusion of efficacy. Therefore, when a finding of noninferiority is used as evidence of efficacy, the determination of assay sensitivity is critical. In this case, the evidence of assay sensitivity in the trial may be deduced from:

1. Historical evidence of sensitivity-to-drug-effects.

It means that similarly designed trials in the past regularly distinguished effective treatments from less effective or ineffective treatments. Historical evidence of sensitivity-to-drug-effects should be evaluated at the beginning of designing a noninferiority trial. Specifically, it should be determined that, in the specific therapeutic area under study, appropriately designed and conducted trials that used a well-defined dose of a specific active treatment, or other treatments with similar effects, reliably showed an effect of at least a minimum size.

2. Appropriate trial conduct.

It means that the conduct of the currently noninferiority trial did not undermine its ability to distinguish effective treatments from less effective or ineffective treatments. Appropriateness of trial conduct can only be fully evaluated after the trial is completed. The design should be similar to the historical trials in all aspects as possible (i.e., satisfy the “constancy assumption”), except that the current trial is intended to demonstrate noninferiority of the test treatment to the active control, whereas the historical trials were mostly placebo-controlled trials involving the selected active-control treatment. Most importantly, the trial should be conducted with high quality.

Demonstration of efficacy by showing noninferiority of the test treatment to an active control of proven effectiveness, therefore, rests on a critical assumption that the selected active-control treatment does have a treatment effect in that trial, i.e., with proven historical evidence of sensitivity-to-drug-effects and constancy in effect size. The effect of the active-control treatment has been shown with a minimum size based on historical placebo-controlled trials. That is, if the test treatment had a much smaller effect than that observed in the historical data (or no effect), then the trial would not have concluded noninferior, and in turn, deduced as effective. Even when historical experience indicates that the trials in a particular therapeutic area do support sensitivity-to-drug-effects, it still could be undermined by whether the specific trial actually had assay sensitivity. Because there are many trial-quality-related factors that could reduce the trial assay sensitivity, e.g., poor patient selection, poor drug compliance and excessive dropouts, use of disease-interfering concomitant medications, use of inappropriate or insensitive endpoints, biased assessment of endpoints, inadequate sample sizes and power, and/or vast variability.

Together with historical evidence of sensitivity-to-drug-effects, appropriate trial conduct provides evidence of assay sensitivity in the new trial. A successful active-controlled trial for noninferiority thus involves four critical steps:

1. Determining that historical evidence of sensitivity-to-drug-effects exists. Without this determination, demonstration of efficacy from a showing of noninferiority is not possible and should not even be attempted.
2. Designing a trial to satisfy the constancy assumption. The trial design (e.g., study population, concomitant therapy, endpoints, and run-in periods) needs to adhere closely to the design of the trials for which historical evidence of sensitivity-to-drug-effects has been determined.
3. Setting an appropriate margin. An acceptable noninferiority margin should be chosen, taking into

account the historical data and relevant clinical and statistical considerations.

4. Conducting the trial with high quality. The trial conduct should also adhere closely to that of the historical trials and should be of high quality.

Prior to the trial start, a noninferiority margin, δ , is selected. This margin is the degree of clinical inferiority of the test treatment to the active-control treatment that the trial will attempt to exclude statistically. If the confidence interval for the difference between the test and control treatments excludes the margin, the test treatment can be declared noninferior and thus effective; otherwise, the test treatment cannot be declared noninferior and cannot be considered effective. In this case, “noninferior” means that at least some degree of the efficacy of the active control has been preserved in the test treatment.

Choosing and setting an appropriate margin for an active-controlled noninferiority trials can be difficult. The margin chosen for a noninferiority trial cannot be greater than the smallest effect size that the active-control treatment would be reliably expected based on historical evidence of sensitivity-to-drug-effects. The margin generally is identified based on past experience in placebo-controlled trials of adequate design under conditions similar to those planned for the new trial, but could also be supported by historical positive dose–response or active-controlled superiority trials. The determination of the margin in a noninferiority trial is based on both statistical reasoning and clinical judgment, should reflect uncertainties in the evidence on which the choice is based, and should be suitably conservative.

A major difficulty in the design and conduct of a noninferiority trial is to choose the noninferiority margin appropriately. In many cases, an appropriate margin cannot be chosen because historical evidence of sensitivity-to-drug-effects cannot be supported. Or, the specific trial may not have assay sensitivity (e.g., improper dose/regimen, poor trial conduct, noninferiority expectation, and inadequate statistical power). In either case, a noninferiority trial cannot be properly designed and conducted.

For detailed discussion on active-controlled noninferiority (or equivalence) trials including margin determination, refer to ICH E-10 Guidance,^[1] Hwang and Morikawa,^[3] Hauck and Anderson,^[4] Temple and Ellenberg,^[5] Ellenberg and Temple,^[6] Ng,^[7] Hwang,^[8] and other relevant documents in this encyclopedia.

USEFUL ALTERNATIVES AND DESIGN MODIFICATIONS

Among the five types of controls discussed thus far the most useful ones are the concurrent controls, namely, placebo, active, and dose–response concurrent

controls. Each has ethical or practical limitations, as well as the issues of assay sensitivity and historical evidence of sensitivity-to-drug-effects (active-control noninferiority trials specific) to be satisfied. The following alternatives as suggested^[1,3] are practically useful:

1. Include placebo in addition to the active-control drug (e.g., a three-arm trial) to establish internal assay sensitivity.

It is often possible and advantageous to use more than one kind of control in a single study, e.g., use of both a placebo and an active control. Inclusion of placebo in an active-control trial, for example, a three-arm trial, will add internal evidence of trial assay sensitivity. The three-arm trial with both a known effective active drug and placebo as controls. Such a trial measures the drug effect size (test–placebo difference), and it allows for the comparison of the test and active-control treatments in a setting where assay sensitivity is established by the direct active control–placebo comparison. It is particularly informative even when the test drug and placebo are nondistinguishable in the trial. If the active control is superior to placebo, the trial does have assay sensitivity and will provide evidence that the test drug is not effective. On the other hand, if neither drug, the test, nor the known effective active-control drug can be distinguished from placebo, the trial is proven as not having evidence of assay sensitivity. That is, this three-arm design can readily evaluate whether a failure to distinguish the test drug from placebo is due to ineffectiveness of the test drug, or simply whether a trial does not have assay sensitivity to detect a difference when it exists. Similarly, trials can use several doses of test drug and several doses of an active control, with or without placebo, as previously defined in dose–response concurrent control. This design may be useful for active drug comparisons and establishing relative potency.

2. Show superiority of the test treatment to the low dose of an active-control treatment.

One may design a trial to show superiority of the test treatment to a low dose of a properly selected active-control drug (i.e., “a pseudo” placebo). Using a pseudo-placebo as control, the advantages of placebo control remain, while the ethical concern may be alleviated.

3. Demonstrate a significant dose–response relationship of the test treatment and show difference between doses.

Design a trial to establish a significant dose–response relationship for the test drug and demonstrate pairwise differences between doses (e.g., the highest vs. the lowest) of the test treatment. The dose–response trial is not very useful for dose determination if the trial failed to show any significant

differences between doses. The failure may arise due to a flat dose–response, a narrow therapeutic window, or inadequate statistical power. Again, inclusion of a placebo control (zero dose) will be beneficial.

4. Use the design modifications for placebo control such as the add-on, replacement, early escape, or randomized withdrawal trial, as appropriate.

In many cases, an add-on, replacement, early escape, or randomized withdrawal trial has been shown to be very useful. An add-on trial is a placebo-controlled trial in which the test drug or placebo is added on to a baseline standard treatment via randomization (e.g., in cancer, antiepileptic, and CHF trials). This design is useful only when standard therapy is subeffective, and it has the advantage of providing evidence of improved clinical outcome/response. A variation of this design is the replacement trial, in which the test or placebo is added by random assignment to conventional therapy and then the conventional therapy is withdrawn, usually by tapering. This design has often been used in trials to study steroid-sparing/replacement and antiepileptic drugs. Another design is an early-escape trial, in which therapy of trial patients whose clinical status worsens or fails to improve to a defined level is promptly withdrawn. The need to change therapy will become the primary endpoint in this case. Further modification is a randomized withdrawal trial, in which patients receiving a test drug for a fixed time period are then randomized to continued treatment with the test drug or placebo. The observed difference between the test drug and placebo would demonstrate a treatment effect of the test drug. The postwithdrawal study period can be of fixed duration, use early escape, or time to event.

SELECTION OF CONCURRENT CONTROL

The selection of control can be affected by the availability of treatments and medical practices in specific regions. Assuming that all trials are appropriately designed and carried out, the potential usefulness of the principal concurrent controls (i.e., placebo, active, and dose–response) in specific situations for demonstrating efficacy can be summarized as follows:

- Placebo concurrent control: Show test treatment effects and measure effect size.
- Active concurrent control for superiority: Show test treatment effects and provide comparative efficacy.
- Active concurrent control for noninferiority: May show test treatment effects and provide comparative efficacy pending proven historical evidence of sensitivity-to-drug-effects.

- Dose–response concurrent control: Show test treatment effects and dose–response relationship.
- Multiple concurrent controls (placebo and active): Show test treatment effects, measure effect size, and provide comparative efficacy.
- Multiple concurrent controls (placebo and dose–response): Show test treatment effects, measure effect size, and dose–response relationship.
- Multiple concurrent controls (active and dose–response): Show test treatment effects, dose–response relationship, and may provide comparative efficacy pending proven historical evidence of sensitivity-to-drug-effects.

Based on the above-listed potential usefulness for each type of control, the following basic questions may be asked sequentially to select the proper control(s) for demonstrating test treatment efficacy:

1. Is proven effective (standard) treatment available?
If the answer is “yes,” proceed to the next question. If the answer is “negative,” one must select one of the following (concurrent) controls to show superiority of the test treatment to control with an appropriate design: (1) placebo-controlled with or without design modifications, as appropriate, (2) dose–response-controlled, (3) active-controlled (superiority), or (4) no-treatment controlled with design modifications, as appropriate.
2. Is the proven effective treatment lifesaving and/or known to prevent irreversible morbidity?
If the answer is “negative,” proceed to the next question. If the answer is “yes,” one can select the proven effective treatment as active control to show either superiority or noninferiority (if there is proven historical evidence of sensitivity-to-drug-effects). One may also select the placebo control with appropriate design modifications (e.g., test treatment vs. placebo in an add-on trial).
3. Is there proven historical evidence of sensitivity-to-drug-effects for an appropriately designed and conducted trial?

If the answer is “yes,” one can select any type of control(s) including active control for noninferiority, as appropriate. Otherwise, same as in Question 1 with the “negative” answer, one must select one of the options (i.e., a–d) listed above—it will be infeasible to design and conduct a noninferiority trial.

CONCLUSION

The selection of control is always critical in designing a clinical trial. It affects trial feasibility including ethic consideration, patient recruitment, dose/regimen determination, and endpoint assessment, etc. It also affects many other features of trial design, conduct, and interpretation.

Ultimately, it may affect trial inferences, credibility, and regulatory acceptability.

The utilities of the external/historical controls are rather limited as compared to the concurrent controls (i.e., placebo, active, and dose–response) due to lack of randomization and ability to minimize bias. Each concurrent control with or without modifications has its potential usefulness and limitations. There are also specific issues such as assay sensitivity and historical evidence of sensitivity-to-drug-effects (in particular for active-controlled noninferiority trials), as defined in ICH E-10 Guidance,^[1] need to be well understood.

An appreciation of these critical issues ensures that a proper control is selected for a appropriately designed and conducted clinical trial for demonstrating treatment effectiveness; insurance is given to achieve the trial's intended objectives with scientific and regulatory credibility.

REFERENCES

1. ICH E-10 Guidance. *Choice of Control Group and Related Issues in Clinical Trials. Guidance for Industry*; CDER/CBER, Food and Drug Administration, May 2001, 1–48.
2. ICH E-9 Guidance. *Statistical Principles for Clinical Trials. Guidance for Industry*; CDER/CBER, Food and Drug Administration, September. 1998, 1–46. http://www.fda.gov/cder/guidance/ICH_E9-fnl.PDF (accessed January 2002).
3. Hwang, I.K.; Morikawa, T. Design issues in noninferiority/equivalence trials. *Drug Inf. J.* **1999**, *33*, 1205–1218.
4. Hauck, W.W.; Anderson, S. Some issues in the design and analysis of equivalence trials. *Drug Inf. J.* **1999**, *33*, 109–118.
5. Temple, R.; Ellenberg, S. Placebo-controlled trials and active-controlled trials in the evaluation of new treatments. Part 1: Ethical and scientific issues. *Ann. Intern. Med.* **2000**, *133* (6), 455–463.
6. Ellenberg, S.; Temple, R. Placebo-controlled trials and active-controlled trials in the evaluation of new treatments. Part 2: Practical issues and special cases. *Ann. Intern. Med.* **2000**, *133* (6), 464–470.
7. Ng, T.-H. Choice of delta in equivalence testing. *Drug Inf. J.* **2001**, *35*, 1517–1527.
8. Hwang, I.K. Choosing the Noninferiority Margin: How Creative Can We Be? Presented at the FDA-Industry Statistical Workshop, Bethesda, MD, January 17–18, 2002. <http://www.fda.gov/cder/guidance/4155fnl.htm> (accessed January 2002).

Clinical Trials: Vaccine Development

Ivan S. F. Chan, William W. B. Wang, and Joseph Heyse

Merck Research Laboratories, West Point, Pennsylvania, U.S.A.

Abstract

This entry reviews some of the special considerations in vaccine development, including an increased interest in surrogate markers of activity, a special need for assessing immune responses, and a need for more specificity in endpoint definitions of clinical efficacy based on exposure.

INTRODUCTION

Clinical trials designed to demonstrate the safety and efficacy of new vaccines have a rich history. In 1954, the largest medical experiment in history tested a vaccine to prevent poliomyelitis, one of the most feared childhood diseases. This was the Salk vaccine, named after Dr. Jonas Salk. Close to 2 million children across the United States and Canada participated in this field trial. The vaccine became one of a group of required vaccinations for all children in the United States. Meier^[1,2] discussed the key statistical considerations of that landmark trial. In many ways, the development of clinical trial methods for vaccines has paralleled rather than coincided with the development of trial methods for new drugs and other interventions. Even in the general area of primary prevention research, vaccine trials occupy a distinct place.

Vaccines are biological products that work primarily by introducing antigen or attenuated live virus into the body. Vaccination triggers an immune response in the form of antibodies and T-cell immunity that are specific for the target infectious agent. The presence of antibodies and T-cell immunity plays an important role in limiting the spread of most pathogens and preventing disease infection. Immunologic memory of immune responses enables the body to rapidly produce a large amount of antigen-specific antibodies and T-cells when the infectious agent is detected. The objective of vaccination is to educate the immune system by inducing the immune memory in the absence of the unpleasant clinical features associated with the target infection. Booster doses of vaccines are sometimes administered after several years to ensure that the immune memory is maintained.

Vaccines are used for the prevention of disease and play a critical role in public health policy. Widespread vaccination in a population not only protects vaccinees who are exposed to the infectious agent, but can change the exposure to infection for people who are not vaccinated.

Because vaccines are developed for administration to millions of healthy subjects, often children, there is a premium on safety. A comprehensive evaluation is made to assure that the benefits of vaccination outweigh the risks. Vaccine review at the United States Food and Drug Administration (FDA) is undertaken by the Center for Biologics Evaluation and Research (CBER). The Centers for Disease Control and Prevention (CDC) are also integrally involved in establishing vaccination policy.

OVERVIEW

Clinical trials in vaccine development generally have four phases.^[3,4] Phase I trials are early, small-scale human studies that explore the safety and immunogenicity of multiple dosage levels of a new vaccine. Phase II trials assess the safety, immunogenicity, and, sometimes, efficacy of selected doses of the vaccine, and generate hypotheses for later testing. Phase III trials, usually large in scale, seek to confirm the efficacy of the vaccine in the target population or prove the consistency of manufacturing processes. This is the last stage of clinical studies before licensure of the vaccine is requested. Phase IV trials are often conducted after licensure to collect additional information on the safety, immunogenicity, or efficacy of the vaccine to meet regulatory commitments or postmarketing objectives. Some sponsors further categorize phase IV trials and refer to postmarketing studies as phase V.

This article reviews some of the special considerations in vaccine development, including an increased interest in surrogate markers of activity (often termed as correlates of protection), a special need for assessing immune responses, a need for more specificity in endpoint definitions of clinical efficacy based on exposure, a greater need for confidence in safety, the unique regulatory process, and special health economic considerations of the public health impact of vaccination programs.

REGULATORY PROCESS FOR VACCINE LICENSURE

Vaccines have been successfully developed by using weakened live viruses, inactivated viruses, purified bacterial proteins and glycoproteins, and recombinant, pathogen-derived proteins. In the United States, vaccine development and licensing applications are regulated by the Center for Biologics Evaluation and Research within the Food and Drug Administration. The Center for Biologics Evaluation and Research's mission is to "protect and enhance the public health through regulation of biological and related products including blood, vaccines, and biological therapeutics..."^[5] The manufacture of biological products such as vaccines presents unique regulatory concerns in relation to quality control, consistency, stability, microbial contamination, product administration, and source of the material.^[6] The Center for Biologics Evaluation and Research is responsible for ensuring the safety, purity, and efficacy of biological products intended for use in the treatment, prevention, or cure of diseases or conditions in humans.^[5]

Vaccines are the most common class of biologic products regulated by CBER. The regulatory process starts with a filing of an Investigational New Drug application (IND), when preliminary testing of a new vaccine shows promising results in animals. All clinical trials under an IND are reviewed by CBER. The licensing application for a new vaccine is called the Biologic License Application (BLA), which is similar to the New Drug Application for a drug. When phase I/II studies are near completion and the planning of phase III (pivotal) trials is underway, the sponsor will request a pre-phase III meeting with CBER during the period of late phase II/pre-phase III to discuss a variety of issues, which may include designs of phase III trials, statistical analysis strategies, proposed indications of the vaccine, product profiles, manufacturing process, and facilities. Once phase III trials have been conducted and the sponsor is preparing for the submission of a licensing application, a pre-BLA meeting is usually scheduled with CBER to discuss a variety of topics such as the format and timing of submission, structure of clinical data, and any potential issues that may lead to refusal to file. The last step of applying for vaccine licensure is to file a BLA with CBER. Under the Prescription Drug User Fee Act of 1992, a standard review cycle (12 mo) is assigned for all submissions requiring the review of clinical data that cannot be classified as priority.^[5] The Center for Biologics Evaluation and Research's review of submissions that do not contain clinical data (such as filing of manufacturing changes) are to be completed within 6 mo. During the review of a BLA, CBER may request appropriate FDA advisory committee review of the new vaccine in terms of its efficacy, safety, and public health implications.

The manufacturing process is a fundamental characteristic of the vaccine product. As a result, the description

of chemistry, manufacturing, and controls of the vaccine is particularly important in the licensing application. Because of the inherent variability in the manufacturing process, the concept of a generic drug has not been applied to vaccines. In fact, vaccines made under different manufacturing processes could be considered distinctly different products, although they are indicated for the same disease protection.^[6]

ASSESSING VACCINE EFFICACY

One of the most critical steps of evaluating a new vaccine is to assess the protective efficacy of the vaccine against the target disease. If early phases (phase I/II) of clinical trials have demonstrated that a new vaccine is safe and able to induce immune responses, an efficacy trial (usually phase III, large scale) will often be conducted to evaluate whether the vaccine, as compared with a control, can completely prevent the disease of interest or at least reduce the incidence and/or the severity of the disease in the target population. When the control is a placebo injection or a vaccine that has no effect on the disease, the trial can estimate absolute vaccine efficacy. When the control is an already licensed vaccine for the same disease indication, the trial can estimate relative vaccine efficacy.^[6]

Absolute vaccine efficacy is best assessed in a prospective, randomized, double-blind, placebo-controlled trial in which participants are randomly assigned to receive either a new vaccine (T) or a placebo control (C). Let P_T and P_C represent the true disease incidence rates among N_T vaccinees and N_C controls randomly assigned in the trial, respectively. The vaccine efficacy, denoted by π , measures the relative reduction of the disease incidence in the vaccine group compared with the placebo group. A widely used measure of vaccine efficacy (π) is given by

$$\pi = 1 - R, \quad \text{where } R = P_T/P_C \quad (1)$$

Here R represents the relative risk of contracting the disease between the vaccine and placebo groups. In general, it is assumed that a new vaccine candidate will not be worse than placebo ($P_T \leq P_C$). A vaccine is 100% efficacious ($\pi = 1$) if it prevents the disease completely ($P_T = 0$), and it has no efficacy ($\pi = 0$) if $P_T = P_C$. Technically, however, the efficacy π can be negative if the vaccine group has a higher disease incidence rate than the placebo group.

In designing a vaccine efficacy trial, one needs to ensure that the study has sufficiently high power to test the hypothesis

$$H_0 : \pi \leq \pi_0 \text{ versus } H_1 : \pi > \pi_0 \quad (2)$$

where π_0 denotes the minimal level of efficacy considered to be acceptable for the new vaccine. This is equivalent to requiring a high level of confidence that the vaccine

efficacy is greater than the prespecified lower bound (π_0). While π_0 could be 0, as is typically the case in establishing therapeutic efficacy of a new drug, considering a lower bound of 0 in vaccine trials involving healthy participants is usually not sufficient. In fact, it is a regulatory requirement to demonstrate that the protective efficacy of a new vaccine is significantly greater than some nonzero lower bound ($\pi_0 > 0$); this requirement was established so that researchers define more precisely the benefit of the vaccine and justify the risk of vaccinating potentially millions of healthy individuals.^[7] This nonzero lower bound is often called a “super efficacy” requirement. Orenstein, Bernier, and Hinman^[8] gave an excellent review of the design considerations for vaccine efficacy evaluation in terms of case definition, case finding, vaccination status ascertainment, and assuring comparability of vaccinated and unvaccinated groups. Lachenbruch^[9] discussed how the specificity and sensitivity of the case definition affects the vaccine efficacy estimate and concluded that a less-specific case definition will lead to a severe underestimation of the vaccine efficacy. Ellenberg and Dixon^[10] and Rida and Lawrence^[11] discussed many important statistical issues related to HIV vaccine trials.

Vaccine Efficacy Based on the Ratio of Two Binomial Proportions

In a randomized vaccine efficacy trial, the number of disease cases in the vaccine and control groups can be assumed to follow independent binomial distributions with parameters (N_T, P_T) and (N_C, P_C), respectively. This assumption is practical if the trial has a relatively short duration or all participants are completely followed through the end of the trial. If so, then a natural estimate of vaccine efficacy is

$$\hat{\pi} = 1 - \hat{R}, \quad \text{where } \hat{R} = \hat{P}_T / \hat{P}_C \quad (3)$$

Here $\hat{P}_T$ and $\hat{P}_C$ are the observed proportions of participants in the vaccine and control groups, respectively, who develop disease during the trial.

In this case, asymptotic methods for analyzing vaccine efficacy (hypothesis test and confidence interval) are direct applications of methods for estimating the relative risk of two binomial proportions; these methods have been extensively discussed in the literature.^[12–18] Of common use is the Z-type method proposed by Miettinen and Nurminen.^[7,13,16]

$$Z_E = \frac{\hat{P}_T - (1 - \pi_0)\hat{P}_C}{\hat{\sigma}_0 / \sqrt{N_T}} \quad (4)$$

where

$$\hat{\sigma}_0 = \left[\tilde{P}_T(1 - \tilde{P}_T) + \frac{1}{u}(1 - \pi_0)^2 \tilde{P}_C(1 - \tilde{P}_C) \right]^{1/2}, \quad u = N_C / N_T \quad (5)$$

and ($\tilde{P}_T, \tilde{P}_C$) are the constrained maximum likelihood estimates of (P_T, P_C) under the null hypothesis given in Eq. (2) based on the observed responses ($\hat{P}_T, \hat{P}_C$). Detailed expressions of (T, C) are given in Refs. [7,13]. The vaccine efficacy is demonstrated (rejecting the H_0) at the one-sided α (usually 2.5%) significance level if $Z_E \leq -Z_\alpha$, where Z_α is the 100(1 - α) percentile of the standard normal distribution. This Z_E test is equivalent to the likelihood score test^[16] and performs very well in many practical applications.^[7,13,16] In addition to hypothesis testing, this method provides a test-based confidence interval that ensures consistent inference with the corresponding p -value. Asymptotic formulas for sample size and power calculations for vaccine efficacy (or relative risk) studies can be found in reports by O'Neill,^[19] Farrington and Manning,^[15] Blackwelder,^[16] and Nam.^[18] Exact inference of vaccine efficacy has been proposed by Chan^[20] and discussed by Chan and Bohidar^[7] in terms of power and sample size calculation. An example of sample size and power calculation based on the Z_E test is given in “Sample Size and Power Calculation for Vaccine Efficacy Trials.”

Vaccine Efficacy Based on the Ratio of Two Rates

When the disease incidence rate is very low, a large-scale study is usually required to demonstrate vaccine efficacy. For sufficiently large sample sizes and small incidence of disease, the numbers of cases in the vaccine and placebo groups may be approximated by independent Poisson distributions with rate parameters λ_T ($\approx N_T P_T$) and λ_C ($\approx N_C P_C$), respectively. In this case, the number of cases in the vaccine group, given the total number of cases (S), is distributed as binomial (S, θ) where

$$\theta = \frac{\lambda_T}{\lambda_T + \lambda_C} = \frac{N_T P_T}{N_T P_T + N_C P_C} = \frac{R}{R + u} = \frac{1 - \pi}{1 - \pi + u} \quad (6)$$

where $u = N_C / N_T$

Because θ is decreasing in π , the efficacy hypothesis in Eq. (2) is equivalent to

$$H_0 : \theta \geq \theta_0 \text{ versus } H_1 : \theta < \theta_0 \quad (7)$$

where $\theta_0 = (1 - \pi_0)/(1 - \pi_0 + u)$. Blackwelder^[16] discussed many asymptotic methods for constructing hypothesis tests and interval estimation regarding vaccine efficacy. For testing the classical null hypothesis with $\theta_0 = 0$, an exact conditional inference based on the Clopper and Pearson method^[21] has been proposed.^[22,23] Chan and Bohidar^[7] discussed a generalization of this exact conditional method for testing efficacy hypotheses with nonzero lower bounds. They found that the exact conditional test works extremely well for low incidence rates. This exact conditional method has been used to analyze the efficacy of vaccines for the prevention of hepatitis A,^[24] herpes zoster, and cervical cancer.

This exact conditional method^[7] can also be used to design a study whose goal is to accrue a fixed number of events instead of running for a fixed duration. Once the total number of events is fixed, the power of the study depends on incidence rates only through the true efficacy, not the disease incidence rate. Thus one can avoid a situation sometimes encountered in a fixed-duration trial in which the anticipated power was not achieved at the end of the trial because the unexpectedly low incidence rates resulted in too few events.^[7] In addition, the Poisson assumption also allows the exact conditional method to handle person-time data easily. This flexibility is important in long-term studies, in which follow-up often differs between the treatment groups. Sample size and power calculation based on this exact conditional method are discussed in Chan and Bohidar,^[7] and an example is given in “Sample Size and Power Calculation for Vaccine Efficacy Trials.”

In longitudinal vaccine efficacy trials, time-to-event type analyses such as the Cox proportional hazards model^[25] may also be used to compare the disease outcomes between the treatment groups. In this case, the vaccine efficacy may be estimated as the ratio of the hazard rates between vaccine and placebo groups.

Trials to Estimate Relative Vaccine Efficacy

When the control is an already licensed vaccine, a randomized efficacy trial can estimate the relative efficacy (π) of a new vaccine via the relative risk ($R = P_T/P_C$) based on the relationship $\pi = 1 - R$. If the absolute efficacy of the control (π_C) has been established, one can estimate the absolute efficacy of the new vaccine (π_T) through the indirect argument^[26,27]

$$\pi_T = 1 - R(1 - \pi_C) \quad (8)$$

A trial to compare the relative efficacy focuses on the relative risk (R). It is usually designed as a noninferiority trial and tests the hypothesis

$$H_0 : R \geq R_0 \text{ versus } H_1 : R < R_0 \quad (9)$$

where R_0 (> 1) is a prespecified noninferiority margin or relative risk threshold. The trial is powered to show that the relative risk R is less than R_0 such that the lower bound of the confidence interval for the estimated absolute vaccine efficacy (π_T) is greater than $(1 - R_0(1 - \pi_C))$ with a high likelihood.^[26] The choice of the noninferiority margin should be based on clinical, regulatory, and statistical judgments so that ruling out a relative risk of R_0 or larger between the new and control vaccines will imply that the new vaccine preserves a large proportion of the efficacy of the control vaccine. Although a noninferiority margin that preserves 50% of the treatment effect has been proposed for the evaluation of drug treatments,^[28–30] there is a general perception that a narrower margin should be used in

preventive vaccine trials because the vaccine will be given to healthy individuals for prophylaxis.

Other Considerations for Vaccine Efficacy Estimates

The vaccine efficacy estimates discussed in previous sections measure the direct effect of the vaccine on the prevention of the disease infection (reduction of disease incidence). Sometimes a vaccine may reduce both the incidence and the severity of the target disease. For example, the chickenpox vaccine has been shown to be more than 90% efficacious in preventing chickenpox in children 1 to 12 yr of age.^[31–33] Furthermore, in vaccinated children who developed chickenpox, the severity (number of lesions and fever) was generally much milder than in unvaccinated children who contracted chickenpox.^[33] Chang, Guess, and Heyse^[34] proposed a combined measure of efficacy, which considers both incidence and severity in evaluating the total direct effect of a vaccine. Recent clinical trials have used this composite efficacy measure, called burden-of-illness, to evaluate the efficacy of new vaccines for the prevention of rotavirus diarrhea in children and herpes zoster in elderly persons. The burden-of-illness approach has also been adapted to compare injection-site symptoms of two hepatitis A vaccines.^[35]

Besides the direct benefits, vaccination also often produces indirect benefits to unvaccinated persons by reducing person-to-person transmission of the disease in the population. Examples of vaccines conferring indirect benefits include the oral polio vaccine, the *Haemophilus influenzae* type B vaccine, and the measles–mumps–rubella vaccine. This indirect effect of the vaccine is called herd immunity.^[36] Extensive work has focused on design and analysis in vaccine trials to estimate the indirect effect and the vaccine efficacy for the reduction of disease transmission rate or secondary attack rate.^[4,37–40] This vaccine efficacy on disease transmission rate is an important measure for evaluating new vaccines intended to prevent diseases with cluster exposures, such as HIV infection.

Investigation of waning (or long-term) vaccine efficacy is also an important issue in vaccine trials. If long-term follow-up data on persons who received vaccines and those who received placebo are available, one can divide the study duration into different time intervals, estimate vaccine efficacy for each interval, and then examine the estimates for trends.^[4] Durham et al.^[41] used a nonparametric survival method to estimate the long-term efficacy of a cholera vaccine in the presence of waning protection. In many instances, long-term follow-up data on placebo controls are not available because these participants would have been offered vaccination once the vaccine has been shown to be efficacious. In such cases, long-term disease breakthrough rates among vaccinees can be compared with age-specific rates from the unvaccinated susceptible population to evaluate the long-term vaccine efficacy.^[33]

Time-to-event analyses can also be performed to examine whether breakthrough rates among vaccinees change over time. However, assessing waning vaccine efficacy without concurrent controls may be subject to bias because the disease epidemiology may change over time, particularly with the increased vaccination coverage after licensure of the vaccine.

Sample Size and Power Calculation for Vaccine Efficacy Trials

As discussed in “Vaccine Efficacy Based on the Ratio of Two Binomial Proportions” and “Vaccine Efficacy Based on the Ratio of Two Rates,” there are many methods of calculating sample size and power for vaccine efficacy trials. Here we briefly describe the methods based on the commonly used asymptotic Z_E test and the exact conditional test in a placebo control trial setting. For testing the null hypothesis $H_0: \pi \leq \pi_0$ against a specific alternative $H_1: \pi = \pi_1$ ($\pi_1 > \pi_0$), the asymptotic power of an α level Z_E test is given by

$$1 - \beta = \Phi \left\{ \frac{1}{\sigma_1} \left[-Z_\alpha \bar{\sigma}_0 + \sqrt{N_T} P_C (\pi_1 - \pi_0) \right] \right\} \quad (10)$$

where σ_1 is the value of the expression in Eq. (5) when $\tilde{P}_T$ and $\tilde{P}_C$ are replaced with P_T and P_C , respectively, $\bar{\sigma}_0$ is the limiting value of $\tilde{\sigma}_0$ in Eq. (5) obtained by calculating $(\tilde{P}_T, \tilde{P}_C)$ at the point $(\tilde{P}_T, \tilde{P}_C) = (P_T, P_C)$, and Φ is the standard normal distribution function.^[7] Similarly, the asymptotic sample size for an α level Z_E test with $(1 - \beta)$ power is given by

$$N_T = (Z_\alpha \bar{\sigma}_0 + Z_\beta \sigma_1)^2 / [P_C (\pi_1 - \pi_0)]^2 \quad (11)$$

and $N_C = u N_T$

For example, suppose we design a study to test the efficacy hypothesis with a lower bound of π_0 of 0.2 at the one-sided 2.5% level, and we would like to achieve 95% power if the true vaccine efficacy is 0.8. Assuming equal sample sizes and a disease incidence rate of 0.6% in the placebo control group, the total number of subjects needed for the study will be approximately 10,838 based on Eq. (11).

To calculate the power using the exact conditional method, we first compute the critical value Y_C corresponding to the one-sided α level test. Given a total number of cases (T) desired for the study and $\theta_0 = (1 - \pi_0)/(1 - \pi_0 + u)$ under the null hypothesis, Y_C can be determined as the maximum number of cases in the vaccine group such that the null hypothesis is rejected: $\Pr[Y \leq Y_C | Y \sim \text{Binomial}(T, \theta_0), H_0] \leq \alpha$. Then the power is given by Chan and Bohidar^[7] as

$$1 - \beta = \Pr[Y \leq Y_C | Y \sim \text{Binomial}(T, \theta_1), H_1] \\ = \sum_{k=0}^{Y_C} \theta_1^k (1 - \theta_1)^{T-k} \quad (12)$$

where $\theta_1 = (1 - \pi_1)/(1 - \pi_1 + u)$. One can also use Eq. (12) to iteratively determine the total number of cases (T) required for the study to achieve the desired power. Because the unconditional expected value of T is $(N_T P_T + N_C P_C)$, we can estimate the number of participants needed for the study as

$$N_T = T / [(1 - \pi_1 + u) P_C] \text{ and } N_C = u N_T \quad (13)$$

For the above example, a total number of cases of 37 will ensure the study to have at least 95% power to show the vaccine efficacy. Based on the disease incidence rate of 0.6% in the placebo control group, the total number of subjects required for enrollment is estimated to be 10,278 using Eq. (13), which compares favorably with the sample size of 10,838 obtained based on the asymptotic formula.

ASSESSING VACCINE IMMUNOGENICITY

Vaccine immunogenicity clinical trials study the immune response to vaccination, which are usually measured by serum antibody titers or T-cell responses. Many immunogenicity trials are conducted in the early phases of vaccine development to assess whether the vaccine can induce immunity before large-scale efficacy trials are performed. In addition, immunogenicity of a vaccine is typically assessed in an efficacy trial to determine whether an immune marker to the vaccine can be used as a surrogate or correlate of disease protection (see the section “Immune Surrogate and Correlate of Protection for Vaccines” for more details).

Once an immune surrogate or correlate of protection has been established, immunogenicity trials can be used as efficient alternatives (for time and economical considerations) to efficacy trials in evaluating the effectiveness of future vaccines. After the successful demonstration of vaccine efficacy, for example, immunogenicity trials are typically used to demonstrate the consistency of vaccine manufacturing process; to prove the vaccine effectiveness when manufacturing processes, storage conditions, or vaccination schedules are modified; to justify concomitant use with other vaccines; and to develop a new combination vaccine with multiple components. The primary objective of such immunogenicity studies is to demonstrate that a new (or modified) vaccine is noninferior (or equivalent) to the current vaccine by ruling out a prespecified clinically relevant difference in the immune response.

General Approaches to Immunogenicity Analysis

Immunogenicity Endpoints

Two types of immunologic endpoints are commonly used to assess vaccine immunogenicity: (1) an immune

response rate, defined as the percentage of participants who achieve a certain level of immune response after vaccination (yes/no); and (2) a geometric mean titer (GMT) or geometric mean concentration of immune response after vaccination. When a particular level of immune response has been shown to be correlated with disease protection, the percentage of participants achieving the “protective level” after vaccination is usually considered the primary endpoint for immunogenicity analyses. In studying participants with preexisting immunity, additional endpoints may include the percentage of participants who develop a certain (e.g., fourfold) fold-rise and the geometric mean fold-rise in immune response from before to after vaccination.

Approaches for Analyzing Immune Response Rates

For the analyses of immune response rates, we consider the general setting comparing a new vaccine (T) to a control (C) in a randomized study. If the control is a placebo, the study is designed as a superiority trial. If the control is a licensed vaccine, the study is often designed as a noninferiority trial. Using the notation of “Assessing Vaccine Efficacy,” let P_C and P_T represent the true immune response rates of the control (with N_C subjects) and the new vaccine group (with N_T subjects), respectively. Then the number of successes (i.e., participants who achieve a certain level of immune response) in the control and new vaccine groups is independently distributed as binomial (N_C, P_C) and binomial (N_T, P_T), respectively. In general, the study can be designed to test the statistical hypothesis (superiority or noninferiority)

$$H_0: P_T - P_C \leq -\delta \text{ versus } H_1: P_T - P_C > -\delta \quad (14)$$

where δ is a prespecified small quantity defining the noninferiority margin. The nominal significance level for the one-sided test is generally set at half of the conventional significance level for a two-sided test for the difference in proportions. This approach has been adopted in regulatory environments, as suggested in the International Conference on Harmonization E9 (ICH E9) guidelines.^[42]

When $\delta = 0$, the hypothesis (Eq. [14]) reduces to the classical one-sided hypothesis aimed to show that the new vaccine is superior to the control. Standard asymptotic methods for testing a null hypothesis of no difference in two proportions can be used.^[43] In addition, exact methods are available for testing this hypothesis^[44,45] and for constructing confidence intervals.^[46,47] Finally, some discussions and debates of one-sided versus two-sided hypothesis for superiority testing can be found in Peace^[48] and Fisher.^[49]

When δ is a small positive quantity, the hypothesis above (Eq. [14]) aims to show that the new vaccine is noninferior (or equivalent) to the control vaccine. Considerations for the choice of noninferiority margin δ are discussed below at the end of “General Approaches to Immunogenicity Analysis.” Asymptotic statistical tests of the noninferiority

hypothesis (Eq. [14]) for two independent treatment groups with a dichotomous endpoint have been extensively discussed in the literature (e.g., Refs. [13,15,50–52]). Many authors have proposed Z-type test statistic with different standard error estimates.^[13,51,52] Of common use and better performance is the Z-type method proposed by Miettinen and Nurminen:^[13,15]

$$Z_D = \frac{\hat{P}_T - \hat{P}_C + \delta}{\tilde{\sigma}_{02} / \sqrt{N_T}} \quad (15)$$

where

$$\tilde{\sigma}_{02} = \left[\tilde{P}_T(1 - \tilde{P}_T) + \frac{1}{u} \tilde{P}_C(1 - \tilde{P}_C) \right]^{1/2}, \quad (16)$$

$$u = N_C / N_T$$

and $(\tilde{P}_T, \tilde{P}_C)$ are the constrained maximum likelihood estimates of (P_T, P_C) under the null hypothesis given in Eq. (14) based on the observed responses $(\hat{P}_T, \hat{P}_C)$. Detailed expressions of $(\tilde{P}_T, \tilde{P}_C)$ are given in Refs.^[13,15] The noninferiority is established (H_0 rejected) at the one-sided α significance level if $Z_D \geq Z_\alpha$, where Z_α is the 100(1 - α) percentile of the standard normal distribution. For immunogenicity trials with small sample sizes, as are often conducted in the early phases of vaccine development, exact tests and confidence intervals^[20,46,53] can be used to compare treatments and test the hypothesis in Eq. (14).

Some immunogenicity trials also involve a hypothesis regarding whether the new vaccine induces an adequate immune response compared with a historical control.^[54] The statistical hypothesis of interest is that $H_0: P_T \leq p_0$ versus $H_1: P_T > p_0$, where p_0 is the prespecified lower limit determined from the historical control data. The statistical criterion for this hypothesis is equivalent to requiring that the lower bound of the confidence interval on the single group immune response is larger than the prespecified lower limit. A confidence interval for immune response rate can be computed by using asymptotic or exact methods for a single binomial proportion.

Approaches for Analyzing Geometric Mean Titers

For the comparison of GMTs, an immunogenicity study is usually designed to test the following general statistical hypothesis:

$$H_0: \text{GMT}_T / \text{GMT}_C \leq K \text{ versus } H_1: \text{GMT}_T / \text{GMT}_C > K \quad (17)$$

where GMT_C and GMT_T represent the GMTs of the control and new vaccine groups, respectively, and K is a prespecified positive quantity relevant to the comparison of fold-difference between the two GMTs. Titers are usually log-transformed in the statistical analysis. The confidence interval for individual GMTs is often calculated on the basis of the asymptotic t -distribution.

When $K = 1$, the hypothesis (Eq. [17]) reduces to the classic one-sided hypothesis of superiority of the new vaccine compared with the control. When $0 < K < 1$, the rejection of the null hypothesis H_0 in 17 at a prespecified α level (usually one-sided $\alpha = 0.025$) leads to the conclusion that the new vaccine is noninferior to the control with respect to the GMTs, with an interpretation that a difference in GMTs (control/new) is $< 1/K$ -fold. This statistical criterion corresponds to the lower bound of the corresponding two-sided $(1 - 2\alpha)$ 100% confidence interval on the ratio of GMTs being greater than K . Some considerations for the choice of noninferiority margin K are given next. The statistical testing and confidence interval estimation regarding GMTs can be performed by using an analysis of variance (ANOVA), an analysis of covariance, or a linear mixed-effects model that includes natural log of the antibody titer as the dependent variable and uses treatment group, baseline, and stratification variables as explanatory variables.

In addition to these parametric analyses on GMTs, graphical displays of reverse cumulative distribution curves of $\Pr(X \geq x)$,^[55] which give percentages of participants with titers greater than or equal to varying levels of titers (x), are commonly used for exploratory evaluation of immune response to vaccination. Nonparametric methods have also been proposed to estimate the overlap or the proportion of similar response in distributions, which can be used to measure the similarity between two distributions of immune response.^[56]

Considerations for Noninferiority/Equivalence Margins

In general, the choice of δ and γ , called noninferiority or equivalence margins, depends on the level of correlation between immune responses and the vaccine efficacy, the variability of immunogenicity measures, and the importance of these endpoints in the trial. The choice should be based on statistical reasoning as well as on clinical and regulatory judgments so that ruling out a difference of δ in immune response rates (or a fold-difference of $1/K$ in GMTs) between the new vaccine and the control will demonstrate that the effectiveness of the new vaccine is comparable to that of the control. For example, in studies of the hepatitis A vaccine VAQTA®, a Δ of 10 percentage points is typically used as the noninferiority margin for comparing immune response rates, which are usually greater than 90%.^[57] A K of 0.67, 0.5, or 0.25 (corresponding to 1.5-, 2-, or 4-fold difference) are often used in comparing GMTs. These noninferiority margins should be discussed proactively between the sponsor and regulatory agencies. Regulatory agencies usually apply the same requirements of equivalence margins for the same endpoints when reviewing similar products from different companies. More general discussions on the choice of noninferiority or equivalence margins can be found in Temple,^[28] Jones et al.,^[29]

Ebbutt and Frith,^[30] the ICH E9 guidelines,^[42] the ICH E10 guidelines,^[58] Siegel,^[59] and Wiens.^[60]

Different Types of Vaccine Immunogenicity Studies

Superiority Immunogenicity Study

Immunogenicity studies are often conducted in the early phases of vaccine development to assess vaccine immunogenicity compared with a placebo. In addition, immunogenicity trials can also be performed to claim superiority of one vaccine over another vaccine from a different manufacturer with respect to immune responses. Such immunogenicity trials should be designed as a superiority (difference-detection) trial rather than an equivalence or noninferiority trial. In this case, the conventional null hypothesis of no difference should be tested against a one-sided alternative or two-sided alternative. The analysis strategy follows the methods for $\Delta = 0$ or $K = 1$ described in the section “General Approaches to Immunogenicity Analysis.”

Dose Response Immunogenicity Study

During vaccine development, a dose–response study may be conducted to assess the immunologic response across different dosage levels of the vaccine and thus determine the minimum effective and safe dose. For many live-virus vaccines, the potency of the vaccine may decay over time. A dose–response study will help investigators understand the kinetic of potency decay, determine the release and end-expiry dosage level, and specify the shelf life of the vaccine. Immune response rates and GMTs are often co-primary endpoints in this type of study. The existence of a dose–response trend in the immune response rate can be tested by using the Cochran–Armitage (C–A) trend test.^[61,62] A dose–response trend with respect to GMTs can be evaluated by using an ANOVA model. Multiple testing procedures^[63,64] can be used to control the overall type I error rate if multiple trend tests or comparisons are made.

Consistency Lots Study

Vaccines are biological products, and their manufacturing process generally varies much more widely than does that of chemical drug products. Before a vaccine can be licensed, the sponsor must demonstrate consistency of the vaccine manufacturing process through analytical and clinical testing. A clinical consistency lot study typically uses three lots of vaccines made from the same manufacturing process (called consistency lots) and a control vaccine (see an example in Ref. [57]), and is intended to (1) rule out a clinically significant difference in either direction between any two pairs of the three consistency lots; and (2) rule out a clinically significant difference

between the combined immune response of three consistency lots and the response of the control. A consistency lot study is typically designed as a two-sided equivalence study with respect to both an immune response rate and a GMT.

The hypothesis of consistency among three lots of vaccine with respect to immune response rate can be written as

$$\begin{aligned} H_0: |P_i - P_j| \geq \Lambda \text{ for at least one pair } (i, j) \\ \text{versus } H_1: |P_i - P_j| < \Lambda \text{ for every } (i, j) \text{ pair} \end{aligned} \quad (18)$$

where P_i and P_j are the immune response rates for lots i and j , $1 \leq i < j \leq 3$. Testing this hypothesis is equivalent to testing three pairs of noninferiority hypothesis tests,^[65] all of which must be rejected for consistency to be declared among the three lots. The nominal level of significance for the consistency lots hypothesis is usually set at the $\alpha = 0.05$ level, at which all six hypotheses are tested. The statistical criterion is equivalent to requiring that all two-sided 90% confidence intervals for the three pairwise differences in immune response rates fall within the interval of $(-\delta, \delta)$. The analysis approach for each noninferiority hypothesis can follow the methods described in “General Approaches to Immunogenicity Analysis.” Wiens and Iglewicz^[66] proposed a slightly improved statistic to test the hypothesis (Eq. [18]). Once the consistency among the three vaccine lots has been demonstrated, the immune response for the three lots can be pooled and compared with the response of the control in a noninferiority hypothesis setting. The hypothesis testing strategy for clinical lot consistency with respect to GMTs can be structured in a fashion similar to that for the immune response rate. The sample size and power for consistency lot studies can be calculated by using simulation.^[65] More discussions on sample size and power calculations can be found in the section “Sample Size and Power Considerations for Vaccine Immunogenicity Studies.”

Bridging Study

After vaccine licensure, manufacturing processes, storage conditions, routes of administration, or dosing schedules may be changed to improve production yields, potency stability, or convenience of vaccination schedule. It is a regulatory requirement to conduct a bridging study (comparing a modified vaccine with the current vaccine) to demonstrate that such changes do not have adverse effects on vaccine effectiveness. If an immune marker correlates with disease protection (see “Immune Surrogate and Correlate of Protection for Vaccines” for more details), immunogenicity bridging trials can be conducted instead of efficacy trials to show clinical equivalence or noninferiority of the modified vaccine to the current vaccine.

An immunogenicity bridging trial is commonly designed as a noninferiority trial aimed at excluding a clinically significant difference in the immune response between the modified vaccine and the current vaccine. The analysis strategy usually follows the methods described in “General Approaches to Immunogenicity Analysis.” The noninferiority hypothesis is usually tested at $\alpha = 0.025$. The key for determining the study hypothesis is the choice of the noninferiority margin, namely, a clinically tolerable difference in proportions (Λ) or fold-difference in GMTs (K) between the modified vaccine and the current vaccine. As discussed in “General Approaches to Immunogenicity Analysis,” the choice of Λ and K depends on the level of correlation between immune responses and the vaccine efficacy, the variability of immunogenicity measures, and the importance of these endpoints.

Combination or Multivalent Vaccine Study

A combination or multivalent vaccine consists of two or more live organisms, inactivated organisms or purified antigens combined by the manufacturer or mixed immediately before administration. The vaccine is intended to prevent multiple diseases, prevent one disease caused by different strains or serotypes of the same organism, or reduce the number of injections and health care visits.^[67] The immunogenicity of all the vaccine components in the combination or multivalent vaccine should be studied to rule out the clinically significant differences in immune response rates or GMTs between the combined vaccine and the separate but simultaneously administered antigens.^[57,67,68] In addition, acceptable levels of immunogenicity should be demonstrated for each serotype or component.^[67,69] In most cases, the endpoints for each component are considered co-primary, and the success requires the demonstration of similarity for all components (see case 2 in Ref. [70]). In some cases, the success requires the demonstration of similarity for any M of the K components (see case 3 in Ref. [70]). For this kind of problems, Capizzi and Zhang^[70] and Ruger^[71] discussed multiplicity adjustments using the ordered test statistics for individual endpoints or using “hybrid” decision rules.

Immunologic Persistence Study

The clinical development phase of a vaccine often takes several years before reaching its licensure. However, the duration of vaccine-induced immunity is expected to be considerably longer than the duration of the clinical studies. Thus there are both conceptual and methodological challenges in assessing whether immunity is waning and in estimating long-term immunologic persistence (defined as the presence of vaccine-induced antibody or cell-mediated immunity over time). To this end, immunologic studies (often open-labeled) are conducted, even after vaccine licensure, to annually measure the immune responses over

multiple years. Life-table or time-to-event analyses^[72] can be used to analyze the cumulative immunologic persistence rate.^[33] In addition, several modeling strategies have been proposed to predict the duration of vaccine-induced immunity based on the extrapolation of observed antibody data from the clinical studies.^[73–75] Although prediction from these extrapolation models relies on many assumptions, they may be useful for developing vaccine booster recommendation and epidemiologic surveillance.^[75]

Sample Size and Power Considerations for Vaccine Immunogenicity Studies

Sample size and power calculation for vaccine immunogenicity trials usually follow the same principles and formula as in the drug trials. Appropriate sample size and power estimations for immunogenicity studies depend on (1) reasonable assessments of source and magnitude of immunologic variability; (2) meaningful determinations of clinically significant difference/threshold in immune response rates or geometric mean titers; (3) clearly stated study hypotheses and endpoints; and (4) well-planned statistical analysis strategies.

Bridging Study

For bridging studies that are designed to test the noninferiority hypothesis (Eq. [14]) with respect to immune response rates, there are many methods of calculating the sample size and power. Of common use is the sample size formula proposed by Farrington and Manning,^[15] which is based on the Z -type test statistic (Eq. [15]). To have a $1 - \beta$ power to claim noninferiority at one-sided α level, the approximate sample size for testing the hypothesis (Eq. [14]) should be

$$N_T = \{Z_\alpha \bar{\sigma}_{02} + Z_\beta \sigma_{12}\}^2 / (P_T - P_C + \delta)^2 \quad (19)$$

and $N_C = uN_T$

where u is the prespecified sample size ratio between the control and the test vaccine groups, P_C and P_T are the expected immune response rates for control and test vaccine groups in the planned study, respectively, Z_φ is the $100(1 - \varphi)$ percentile of the standard normal distribution, σ_{12} is the value of the expression in 16 when the constrained maximum likelihood estimates $\hat{P}_T$ and $\hat{P}_C$ are replaced with P_T and P_C , respectively, and $\bar{\sigma}_{02}$ is the limiting value of $\hat{\sigma}_{02}$ in Eq. (16) obtained when $(\hat{P}_T, \hat{P}_C)$ are calculated taking $(\hat{P}_T, \hat{P}_C) = (P_T, P_C)$. Similarly, the asymptotic power of a one-sided α level Z_D test is given by

$$1 - \beta = \Phi \left\{ \frac{1}{\sigma_{12}} \left[-Z_\alpha \bar{\sigma}_{02} + \sqrt{N_T} (P_T - P_C + \delta) \right] \right\} \quad (20)$$

where Φ is the standard normal distribution function.

To illustrate the sample size calculation under this setting, let us assume that one needs to design a bridging study to rule out a 10 percentage-point difference in immune response rate between a new test vaccine and a control vaccine. The noninferiority margin in the study hypothesis (Eq. [14]) is 10% (i.e., $\Lambda = 0.1$). Assuming the hypothesis is tested at one-sided $\alpha = 0.025$ level and the expected immune response rates for both groups are 90% in the planned study (i.e., $P_C = P_T = 0.9$), then in order to have 95% power to rule out a 10 percentage-point difference between the equally sized test and control groups, ~250 subjects per group are required to be evaluable for the primary analysis. Assuming a 10% nonevaluable rate because of protocol violation, loss to follow-up, invalid assay results, etc., ~280 subjects should be enrolled in each group.

For bridging studies that are designed to test the noninferiority hypothesis in Eq. (17) with respect to the GMTs, the sample size calculation is similar to that for testing the conventional hypothesis of no difference under the lognormal assumption. Let l be the standard deviation of the log-transformed antibody titer. In order to achieve $1 - \beta$ power in testing the hypothesis (Eq. [17]) at the one-sided α level, the sample size for the test vaccine group is:

$$N_T = (1 + 1/u)(Z_\alpha + Z_\beta)^2 \sigma^2 / [\log(R_{\text{GMT}}) - \log(\gamma)]^2 \quad (21)$$

and $N_C = uN_T$

where u is the prespecified sample size ratio between the control and the test vaccine groups and R_{GMT} ($\text{GMT}_T / \text{GMT}_C$) represents the ratio of GMTs between the new and control vaccine groups under the alternative hypothesis. Similarly, the asymptotic power of a one-sided α level test based on the normal approximation is given by

$$1 - \beta = \Phi \left\{ Z_\alpha + \frac{1}{\sigma} \sqrt{\frac{N_T}{(1 + 1/u)}} [\log(R_{\text{GMT}}) - \log(\gamma)] \right\} \quad (22)$$

Consistency Lot Study

When planning a consistency lot study to compare three clinical lots, no closed-form formula exists for sample size calculation, and simulation study can be used to adequately plan the sample size.^[65] For power estimation, Wiens, Heyse, and Matthews recommended performing simulations with the three population parameters under the alternative hypothesis of being not all equal. For example, for hypothesis testings with respect to the immune response rate, they recommended performing simulations under the setup that the true immune response rates for the three groups are P_1 , P_1 , and $P_1 + 0.5\Lambda$, instead of being all equal. This gives a conservative estimate of power when the alternative hypothesis is true.^[65]

Combination or Multivalent Vaccine Study

In a combination or multivalent vaccine study, the hypotheses regarding each component or each serotype are often considered as co-primary. In planning such a study, sample size for each primary hypothesis can be calculated using formulas 19, 21, or other related sample size formulas for that hypothesis. To address the concerns of multiplicity, one approach is to plan the power for each primary hypothesis large enough such that the overall power is acceptable under the assumption of independence among the multiple hypotheses. This approach is generally conservative (estimating lower bound of power) because the correlation between components is usually positive. For example, for a multivalent vaccine with k components, one can plan the power for hypothesis testing on each component to be at least $1 - \beta/k$ so that the overall power is guaranteed to be no less than $1 - \beta$.

ASSESSING VACCINE SAFETY

A complete evaluation of safety is an important component in all vaccine development programs. The safety considerations of vaccines differ from those of the pharmaceutical products because of the fundamental role of vaccines in public health initiatives. Vaccines are administered to millions of healthy individuals around the world with the objective of preventing infectious diseases. Many vaccines are mandated for entry into public schools, day care and preschool programs, and even universities. Safety evaluations begin with the earliest preclinical experiments and continue throughout the clinical development program and marketing life span of the vaccine. Phase IV postlicensure regulatory commitments typically include a formal safety evaluation in a large managed health care setting. In addition to postmarketing surveillance systems monitored by individual corporate sponsors, the FDA and the CDC have developed a national adverse event reporting system for U.S. licensed vaccines.

The ICH E9 guidelines^[42] include specific recommendations on safety evaluation in clinical trials that pertain to both drugs and vaccines. Safety is always an important consideration in vaccine trials, although these trials are mostly designed with a sample size related to a primary hypothesis of efficacy or immunogenicity. It is important to monitor all clinical trials for adverse events that become apparent. Independent data monitoring committees are often established in large multicenter trials to assess the safety and efficacy data at predefined study intervals. The committee may recommend to the sponsor to modify or terminate a trial on the basis of an evaluation of the risks and benefits to the study participants.

The methods and measurements chosen to establish the safety of a vaccine depend on many factors, including the type of vaccine (e.g., live attenuated viruses or bacterial

vaccines) and its specific mechanism for eliciting immune responses. Vaccination can cause allergic or anaphylactic reactions because it induces the immune system. The reactions are typically local to the site of the infection, such as swelling, tenderness, or redness. Systemic reactions, such as fever or muscle ache, can also occur. The vaccine may also produce an illness similar to that produced by the wild-type infectious agent. Coplan et al.^[76] developed and validated a standardized vaccine report card for use in vaccine trials. The pediatric and adult instruments were developed to ensure that the wording was understandable, to promote ease of use, and to enhance accuracy and completeness of reporting. The validation phase showed that the use of a standardized vaccine report card provides a practical and complete method of eliciting complaints after vaccination.

With the increasingly aggressive vaccination schedule in children, an interest in the development of combination vaccines has been expressed. Ellenberg^[77] discussed the statistical considerations in evaluating the safety of combination vaccines. Although combination vaccines generally include individual components that have been previously studied and used, new or more severe adverse reactions might occur in the combination product. Clinical trials usually compare the combination vaccine to the individual components. Although the safety database may not need to be as large as that for the novel vaccines with new immunogens, a reliable evaluation of safety should still be performed. The CBER^[67] developed guidelines for the evaluation of combination vaccines.

All safety variables encountered in a clinical trial require attention, and the protocol should specify the analytical approach. All adverse events should be reported, regardless of whether they are considered related to vaccination. When studies include hypotheses about specific adverse events, the usual statistical framework is appropriate. However, the main difficulty with the statistical analysis of adverse events is the possibility of new and unanticipated reactions. Treating these safety variables as confirmatory information is not appropriate. For this type of safety evaluation, it is best to apply descriptive statistics supplemented by confidence intervals. Measures of risk differences or relative risk are useful for this purpose. Reporting individual p -values can sometimes be used to help evaluate a specific difference of interest or as a “flagging” device applied to many safety variables to highlight possible vaccine effects that are worth further attention.

The ICH E9 guidelines recognize the importance of considering multiplicity in the interpretation of safety data.^[42] They noted the need to quantify the type I error by using statistical adjustments for multiplicity while recognizing the concerns of type II errors in safety assessment. Mehrotra and Heyse^[78] proposed a two-step use of adjusted p -values based on the false discovery rate method of Benjamini and Hochberg^[79] as a flagging method. They used data from three vaccine clinical trials to illustrate

the proposed double false discovery rate approach and to reinforce the potential effect of failure to account for multiplicity.

Because vaccines will typically be administered to millions of otherwise healthy individuals, evaluating vaccines for rare but serious adverse events is important. Common reactions to vaccines are readily identified in the investigational phase of clinical development programs. However, rare but serious events may be missed in the clinical trials. To exemplify the problem, we refer to the “rule of three,”^[80] which states that $3/n$ is an upper 95% confidence bound for the binomial probability P when no events occur among n independent trials. Applying this rule to a vaccine program of $n = 5000$ participants gives an upper bound estimate of $P = 0.0006$ for adverse events not observed in the studies. Administration of the vaccine to 1 million persons could yield 600 severe reactions.

Several activities have been put in place to protect against this situation. Clinical development programs include prelicensed safety studies to build up the safety database. In addition, sponsors undertake Phase IV studies to address important safety concerns. For example, Black et al.^[81] reported the results of a prospective postmarketing evaluation of the safety and effectiveness of varicella vaccination in 89,753 adults and children in the preventive care program of the Northern California Kaiser Permanente Medical Care Program. These types of studies are possible in large managed care organizations because of the large patient populations and the sophisticated use of automated clinical databases of hospitalizations, emergency department visits, and clinic visits. Other types of epidemiologic investigations of vaccine safety are also undertaken. Ascherio et al.^[82] reported the results of a nested case-control study of a possible association between hepatitis B vaccination and the risk of multiple sclerosis. The study was prompted by spontaneous reports of multiple sclerosis development after vaccination. Ascherio et al.^[82] reported no such association in the context of this well-conducted evaluation.

The FDA and the CDC have created the Vaccine Adverse Event Reporting System (VAERS) for postmarketing safety surveillance.^[83] This system accepts reports of adverse events that may be associated with U.S. licensed vaccines from health care providers, manufacturers, and the public. The reports are continually monitored for any unexpected patterns or changes in rates of adverse events. Niu, Eriwn, and Braun^[84] used the empirical Bayes data-mining techniques of DuMouchel^[85] in VAERS to investigate the detection of intussusception after rotavirus vaccination.

Postmarketing evaluations are complex and have their drawbacks.^[86,87] In addition, many persons may be at risk during the early phases of the vaccine integration to the marketplace. Ellenberg^[77] argued that the most reliable way to assess causality in safety is a controlled study and that the size of clinical trials of new vaccines must

be increased to be able to detect serious rare effects. She proposed sample sizes on the order of 60,000 to 80,000 participants using the methods appropriate under simple binomial sampling. This information would increase the confidence in the safety of vaccines and would be a valuable resource for assessing spontaneous reports of adverse events after licensure. Overall, a comprehensive and continual assessment of the safety of a vaccine reduces the risk that a vaccine may cause severe reaction to a small proportion of vaccinees.

The association between rotavirus vaccination and intussusception provides an excellent example of the issues relating to evaluating vaccine safety. Intussusception is a rare but serious bowel condition that occurs in young children 3 mo to 2 yr of age. It occurs when a portion of the bowel folds in on itself and causes a constriction that interrupts blood flow. The incidence of intussusception is approximately 1 case per 2000 person-yr. In prelicensure studies of a rhesus-human reassortant rotavirus tetravalent vaccine (RRV-TV), 5 cases of intussusception were identified among 10,054 vaccinated infants and 1 case was identified among 4633 placebo recipients. This result was not statistically significant, and Rennels et al.^[88] concluded that the findings did not support an apparent association between intussusception and RRV-TV. The RRV-TV was licensed on August 31, 1998, and was recommended by the Advisory Committee on Immunization Practices as a three-dose series given at 2, 4, and 6 mo of age.

Murphy et al.^[89] reported on an investigation that was initiated after 9 cases of intussusception were reported to VAERS. They undertook a thorough data collection exercise and analyzed data for 429 infants with intussusception and 1763 matched controls in a case-control analysis, as well as for 432 infants with intussusception in a case series analysis. A case series analysis^[90,91] estimates the relative incidence of clinical events in defined time intervals after vaccination compared with control periods using only cases. Murphy et al.^[89] concluded a strong causal association between intussusception and RRV-TV vaccination, especially within 1 week after the first dose. Niu, Eriwn, and Braun^[84] used empirical Bayes data-mining methods on the VAERS intussusception data and showed that even earlier detection of rare serious adverse events was possible. The Advisory Committee on Immunization Practices withdrew its recommendation of the vaccine, and the vaccine manufacturer voluntarily withdrew its product from the market.^[92]

Sadoff et al.^[93] described the study design considerations necessary to detect an increased risk of intussusception in a randomized clinical safety study. By using extensive monitoring for intussusception cases through multiple stopping boundaries, they noted that efficiencies in the design are possible. Still, studies to detect rare safety outcomes require very large sample sizes. The complex design described by Sadoff et al.^[93] used Monte Carlo simulation methods for a study that involves at least 60,000 infants. This sample size

is consistent with the sample sizes generally proposed by Ellenberg^[77] for vaccine safety studies.

As a postscript, it is interesting to note that 4 of the 5 cases of intussusception among RRV-TV recipients occurred within 2 weeks of vaccination. The single case in placebo recipients occurred several weeks after vaccination. This observation supports O'Neill's^[94] recommendation to consider the timing of events in the analysis of safety data.

ADDITIONAL STATISTICAL ISSUES FOR ANALYZING VACCINE TRIALS

Intention-to-Treat or Per-Protocol Analysis

A typical intention-to-treat population for vaccine trials includes all randomly assigned participants, regardless of whether they met study entry criteria, the vaccine they actually received, and subsequent withdrawal or deviation from the protocol. In comparison, a typical per-protocol population includes a subset of randomly assigned participants who met the study entry criteria, received study vaccines as prescribed by the protocol, completed the study without major protocol violations, and had endpoint measurements at specific time points. Some vaccine efficacy studies also use a modified intention-to-treat population, which includes only participants who complete all doses of the assigned vaccine and disease cases occurring after some specific time following the full series of vaccination to capture the full effect of the vaccine. More general discussions of these analysis populations in clinical trials can be found in Gillings and Koch,^[95] the ICH E9 guidelines,^[42] Lachin,^[96] and Lange.^[97]

For the analysis of vaccine efficacy, an intention-to-treat analysis generally provides unbiased assessment of vaccination strategy and controls false-positivity rates in placebo-controlled efficacy trials. However, this analysis may artificially inflate the false-positivity rate in trials aimed to show equivalent/noninferior efficacy or immunogenicity of two vaccines. A per-protocol or modified intention-to-treat analysis of vaccine efficacy generally aims to assess the biological effect of the vaccine. Historically, the per-protocol or modified intention-to-treat analysis of vaccine efficacy has often been considered the primary approach, and the intention-to-treat analysis has been considered the secondary approach.^[98] This is probably because bias is less an issue in vaccine trials involving healthy participants; high adherence is usually achievable and missing data/dropouts as a result of adverse experience are usually very low.^[99] In addition, the intention-to-treat and modified intention-to-treat/per-protocol analyses usually yield the same conclusion about the vaccine efficacy. However, if bias is a major concern of the trial (such as in open-label trials), or the goal is to compare the effectiveness of vaccination strategy (such as comparing

a three-dose regimen at 1, 2, and 6 mo versus 2, 6, and 12 mo), then intention-to-treat analysis should be used.^[99]

A per-protocol analysis is usually the primary approach for immunogenicity, which is designed to assess the biological or immunologic effects of vaccination. In addition, an intention-to-treat or modified intention-to-treat analysis is typically planned to support or confirm the per-protocol analysis. Safety analysis of vaccine trials is typically based on a modified intention-to-treat population that includes all participants vaccinated and for whom safety follow-up data are available.

Missing/Coarse Data and Loss to Follow-Up

Missing data and loss to follow-up may occur in vaccine studies. These issues present fewer problems in estimating the efficacy, immunogenicity, or safety of vaccines than in drug trials. Because vaccination is usually administered once or as a few doses over time, treatment adherence is typically very high. In addition, as mentioned earlier, few dropouts are a result of vaccine-related adverse events. In practice, missing data and loss to follow-up from vaccine trials are generally a result of the causes unrelated to the study vaccine. Therefore missing data arising from vaccine trials can reasonably be assumed to be missing at random or even missing completely at random. In an efficacy trial, one can estimate the loss of information (disease cases) because of loss to follow-up and perform sensitivity analyses. Multiple imputation techniques^[100] are useful in these analyses.

Serial-dilution assays are commonly used to measure immune responses. Without a standard curve, the reported immune responses from a serial-dilution assay are often interval-censored, which can be considered as a special form of coarse data.^[101] The lower limit of the interval is routinely reported and treated as the exact titer value in immunogenicity analyses. However, ignoring the coarseness of the immune responses may result in severe bias in parameter estimation. Different approaches have been discussed for addressing this issue.^[102]

Analyses with Stratification Variables

Many vaccine trials are conducted in multiple study centers or are stratified by variables such as age, sex, or other prognostic factors. In these cases, stratified analysis can be used to increase clinical trial (or assay) sensitivity and to reduce estimation bias.^[103,104] An overall vaccine effect can be estimated as a weighted average across strata, and the weighting scheme may depend on whether the stratification factor is prognostic or nonprognostic, or whether treatment-by-stratification interaction exists.^[105–108]

Interim Analyses

In large-scale vaccine efficacy trials, formal interim analyses for potential early stopping of the trial are often

planned. Interim monitoring of efficacy and safety is typically performed by an independent data monitoring committee. The FDA has recently developed a draft guidance on the establishment and the operation of these committees.^[109] Determination of early stopping because of overwhelming efficacy is usually guided by statistical stopping criteria such as those of O'Brien and Fleming^[110] and Lan and DeMets.^[111] Trial termination because of safety concerns occurs far less frequently in vaccine trials than in drug trials because participants receive only one or a limited number of vaccinations over a short period. Of course, futility analysis (for early stopping of the trial because of a low likelihood of success at study end) can be performed to minimize resource utilization.^[112–114]

IMMUNE SURROGATE AND CORRELATE OF PROTECTION FOR VACCINES

A critical task in vaccine development is identifying immune response markers (antibody or T-cell response) that can be used as surrogates for clinical vaccine efficacy so that more efficient evaluation of new vaccines or manufacturing process changes can be performed using the immune surrogates. Immune surrogates are best evaluated in a prospectively planned efficacy trial. The classic methods for validating surrogate endpoints^[115–119] may be used to assess the validity of an immune surrogate. In particular, these methods are applicable for vaccines aimed at increasing the immune responses of persons who may have preexisting immunity; in these situations, the validation criteria are similar to the validation of CD4 counts and viral RNA levels as surrogates for AIDS. However, these classic methods may break down in evaluating pediatric vaccines, which are targeted toward persons who are immune-naïve (without preexisting immunity before vaccination). If a vaccine is highly immunogenic, the immune responses of vaccinated persons will clearly be separated from that of unvaccinated persons (who have no immunity). As a result, the immune response status will be completely confounded with the vaccination status, and it will be very difficult, if not impossible, to apply Prentice's criteria^[115] to show that given an immune response, the probability of developing the disease is the same regardless of whether a person is vaccinated or not.

Because of the above-mentioned reasons, vaccine researchers tend to assess the relationship between immune responses and protective efficacy via the concept of "correlate of protection." By examining vaccine failures, researchers aim to establish a "protective level" of an immune response that can be used to determine whether a person is "completely protected" or "not protected and still susceptible to the disease."^[120–122] On a population basis, this "protected level" means that 100% of the persons who have immune responses at this level or higher will be completely protected from the disease.

In some instances, it is difficult to identify a clear-cut value of the immune response as the "protective level" because protective efficacy tends to increase with higher immune responses. For example, clinical trials with a live attenuated varicella (Oka/Merck) vaccine have shown a high protective efficacy against chickenpox^[31–33] and an inverse relationship between the frequency of chickenpox breakthroughs among vaccinees and the post-vaccination antibody titer measured by the glycoprotein enzyme-linked immunosorbent assay.^[123–125] However, no particular level of antibody can be considered a clear-cut "protective level." Chan et al.^[124] have proposed the use of statistical models (such as accelerated failure time models) to link the whole distribution of antibody titers and long-term disease protection conferred by vaccination, and Li et al.^[125] have proposed the use of the concept of "approximate protective level" at which a high proportion (e.g., > 95%) of individuals are expected to be protected from the disease. If feasible, the use of placebo group and different vaccine dosages in the same study will also enhance the assessment of the immune correlate. Once an immune marker is validated as a "correlate of protection," vaccine efficacy can be subsequently evaluated through this immune marker. This will facilitate efficient evaluation of vaccine manufacturing process changes, vaccine consistency lots, concomitant vaccination, or new combination vaccines. Chan et al.^[124] have also proposed the use of statistical models to predict vaccine efficacy based on the immune responses collected in immunogenicity trials.

ECONOMIC EVALUATION OF VACCINE

There is increasing interest in conducting economic evaluations of medical interventions. Several textbooks review the study types and methods used to evaluate the cost effectiveness of health care programs.^[126–128] Cost-effectiveness analysis identifies all relevant costs involved and evaluates the expected health gains derived in particular programs. The goal of economic evaluation is to maximize the health benefits per dollar spent. The usual measure is the cost-effectiveness ratio, defined as the ratio of the incremental costs of undertaking the health care program to the incremental health effects. Cost-effectiveness analysis is often used as an aid to national public health decision making.

Edmunds, Medley, and Nokes^[129] argued that while most cost-effectiveness analyses are based on the benefits to individual persons for vaccine programs, the population is the appropriate target for the mass immunization program. Therefore the indirect effects (herd immunity) are important and should be included in the cost-effectiveness analysis. This approach measures the health effect of the vaccination program in terms of the proportion of the total population that will be protected after mass immunization. The effects are a result of both the efficacy of the vaccine

and a reduction in the number of infectious individuals in the population. Edmunds, Medley, and Nokes^[129] called this a dynamic model of an infectious disease and showed that the cost-effectiveness ratio seen with the use of this approach tends to decline over time.

Using the dynamic approach involves developing a population-based mathematical model of the spread of the infection. Halloran et al.^[130] developed such a model for routine varicella immunization of preschool children in the United States. Using newly available data, Brisson et al.^[131] updated Halloran et al.'s estimates. The Halloran model has been used as the basis for conducting cost-effectiveness analysis of varicella vaccinations in several countries.^[132]

REFERENCES

- Meier, P. Polio trial: An early efficient clinical trial. *Stat. Med.* **1990**, *9*, 13–16.
- Meier, P. The Biggest Health Experiment Ever: The 1954 Field Trial of the Salk Poliomyelitis Vaccine. In *Statistics: A Guide to the Study of the Biological and Health Sciences*; Holden-Day, 1977.
- Lynch, C.J.; Lachenbruch, P.A. Statistical issues in biologics submissions to the FDA. *Drug Inf. J.* **1996**, *30*, 921–932.
- Halloran, M.E. Vaccine Studies. In *Biostatistics in Clinical Trials*; Armitage, P., Colton, T., Eds.; John Wiley & Sons: New York, 2001, 479–486.
- Sensabaugh, S.M. A primer on CBER's regulatory review structure and process. *Drug Inf. J.* **1998**, *32*, 1011–1030.
- Lachenbruch, P.A.; Horne, A.D.; Lynch, C.J.; Tiwari, J.; Ellenberg, S.S. Biologics. In *Encyclopedia of Biopharmaceutical Statistics*; Chow, S.C., Ed.; Marcel Dekker, Inc.: New York, 2000, 47–54.
- Chan, I.S.F.; Bohidar, N.R. Exact power and sample size for vaccine efficacy studies. *Communications in Statistics. Commun. Stat., Part A-Theory Methods* **1998**, *27*, 1305–1322.
- Orenstein, W.A.; Bernier, R.H.; Hinman, A.R. Assessing vaccine efficacy in the field. *Epidemiol. Rev.* **1988**, *10*, 212–241.
- Lachenbruch, P.A. Sensitivity, specificity, and vaccine efficacy. *Control. Clin. Trials* **1998**, *19*, 569–574.
- Ellenberg, S.S.; Dixon, D.O. Statistical issues in designing clinical trials of AIDS treatments and vaccines. *J. Stat. Plan. Inference* **1994**, *42*, 123–135.
- Rida, W.N.; Lawrence, D.N. Some statistical issues in HIV vaccine trials. *Stat. Med.* **1994**, *13*, 2155–2177.
- Gart, J.J. Approximate tests and interval estimation of the common relative risk in the combination of 2×2 tables. *Biometrika* **1985**, *72*, 673–677.
- Miettinen, O.; Nurminen, M. Comparative analysis of two rates. *Stat. Med.* **1985**, *4*, 213–226.
- Gart, J.J.; Nam, J. Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness. *Biometrics* **1988**, *44*, 323–338.
- Farrington, C.P.; Manning, G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Stat. Med.* **1990**, *9*, 1447–1454.
- Blackwelder, W.C. Sample size and power for prospective analysis of relative risk. *Stat. Med.* **1993**, *12*, 681–691.
- Ewell, M. Comparing methods for calculating confidence intervals for vaccine efficacy. *Stat. Med.* **1996**, *15*, 2379–2392.
- Nam, J. Power and sample size for stratified prospective studies using the score method for testing relative risk. *Biometrics* **1998**, *54*, 331–336.
- O'Neill, R.T. On sample sizes to estimate the protective efficacy of a vaccine. *Stat. Med.* **1988**, *7*, 1279–1288.
- Chan, I.S.F. Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. *Stat. Med.* **1998**, *17*, 1403–1413.
- Clopper, C.J.; Pearson, E.S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **1934**, *26*, 404–413.
- Gail, M. Power computations for designing comparative Poisson trials. *Biometrics* **1974**, *30*, 231–237.
- Guess, H.A.; Thomas, J.E. A rapidly converging algorithm for exact binomial confidence intervals about the relative risk in follow-up studies with stratified incidence-density data. *Epidemiology* **1990**, *1*, 75–77.
- Werzberger, A.; Mensch, B.; Kuter, B.; Brown, L.; Lewis, J.; Sitrin, R.; Miller, W.; Shouval, D.; Wiens, B.; Calandra, G. A controlled trial of a formalin-inactivated hepatitis A vaccine in healthy children. *N. Engl. J. Med.* **1992**, *327* (7), 453–457.
- Cox, D.R. Regression models and life-tables (with discussion). *J. R. Stat. Soc., B.* **1972**, *34*, 187–220.
- Chan, I.S.F. Proving non-inferiority or equivalence of two treatments with dichotomous endpoints using exact methods. *Stat. Methods Med. Res.* **2002**, *in press*.
- Wang, S.J.; Hung, H.M.J.; Tsong, Y. Utility and pitfalls of some statistical methods in active controlled clinical trials. *Control. Clin. Trials*, **2002**, *23* (1), 15–28.
- Temple, R. Problems in interpreting active control equivalence trials. *Account. Res.* **1996**, *4*, 267–275.
- Jones, B.; Jarvis, P.; Lewis, J.A.; Ebbutt, A.F. Trials to assess equivalence: The importance of rigorous methods. *Br. Med. J.* **1996**, *313*, 36–39.
- Ebbutt, A.F.; Frith, L. Practical issues equivalence trials. *Stat. Med.* **1998**, *17*, 1691–1701.
- Weibel, R.E.; Neff, B.J.; Kuter, B.J.; Guess, H.A.; Rothenberger, C.A.; Fitzgerald, A.J.; Connor, K.A.; McLean, A.A.; Hilleman, M.R.; Buynak, E.B.; Scolnick, E.M. Live attenuated varicella virus vaccine. Efficacy trial in healthy children. *N. Engl. J. Med.* **1984**, *310* (22), 1409–1415.
- Kuter, B.J.; Weibel, R.E.; Guess, H.A.; Matthews, H.; Morton, D.H.; Neff, B.J.; Provost, P.J.; Watson, B.A.; Starr, S.E.; Plotkin, S.A. Oka/Merck varicella vaccine in healthy children: Final report of a 2-year efficacy study and 7-year follow-up studies. *Vaccine* **1991**, *9* (9), 643–647.
- Vessey, S.J.R.; Chan, C.Y.; Kuter, B.J.; Kaplan, K.M.; Waters, M.; Kutzler, D.P.; Carfagno, P.A.; Sadoff, J.C.; Heyse, J.F.; Matthews, H.; Li, S.; Chan, I.S.F. Childhood vaccination against varicella: Persistence of antibody, duration of protection, and vaccine efficacy. *J. Pediatr.* **2001**, *139* (2), 297–304.

34. Chang, M.N.; Guess, H.A.; Heyse, J.F. Reduction in burden of illness: (A) new efficacy measure for prevention trials. *Stat. Med.* **1994**, *13*, 1807–1814.
35. Su, L.; Tucker, R.; Frey, S.E.; Gress, J.O.; Chan, I.S.; Kuter, B.J.; Guess, H.A. Measuring injection-site pain associated with vaccine administration in adults: A randomized, double-blind, placebo-controlled clinical trial. *J. Epidemiol. Biostat.* **2000**, *5* (6), 359–365.
36. Fine, P.E.M. Herd immunity: History, theory, practice. *Epidemiol. Rev.* **1993**, *15*, 265–302.
37. Halloran, M.E.; Haber, M.; Longini, I.M., Jr.; Struchiner, C.J. Direct and indirect effects in vaccine efficacy and effectiveness. *Am. J. Epidemiol.* **1991**, *133* (4), 323–331.
38. Haber, M.; Halloran, M.E.; Longini, I.M., Jr.; Watelet, L. Estimation of vaccine efficacy in non-randomly mixing populations. *Biom. J.* **1995**, *37*, 25–38.
39. Rida, W.N. Assessing the effect of HIV vaccination on secondary transmission. *Stat. Med.* **1996**, *15*, 2393–2404.
40. Longini, I.M., Jr.; Sagatelian, K.; Rida, W.N.; Halloran, M.E. Optimal vaccine trial design when estimating vaccine efficacy for susceptibility and infectiousness from multiple populations. *Stat. Med.* **1998**, *17*, 1121–1136.
41. Durham, L.K.; Longini, I.M.; Halloran, M.E.; Clemens, J.D.; Nizam, A.; Rao, M. Estimation of vaccine efficacy in the presence of waning: Application to cholera vaccine. *Am. J. Epidemiol.* **1998**, *147* (10), 948–959.
42. ICH E9 Expert Working Group. Statistical principles for clinical trials: ICH harmonized tripartite guidelines. *Stat. Med.* **1999**, *18*, 1905–1942.
43. Fleiss, J.L. *Statistical Methods for Rates and Proportions*; John Wiley & Sons: New York, 1981.
44. Barnard, G.A. Significance test for 2×2 tables. *Biometrika* **1947**, *34*, 123–138.
45. Suissa, S.; Shustern, J.J. Exact unconditional sample sizes for the 2×2 binomial trial. *J. R. Stat. Soc., A* **1985**, *148*, 317–327.
46. Chan, I.S.F.; Zhang, Z.X. Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics* **1999**, *55*, 1202–1209.
47. Agresti, A.; Min, Y. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* **2001**, *57*, 963–971.
48. Peace, K.E. One-sided or two-sided p-values: Which most appropriately address the efficacy issues. *J. Biopharm. Stat.* **1991**, *1*, 133–138.
49. Fisher, L.D. The use of one-sided tests in drug trials: An FDA advisory committee member's perspective. *J. Biopharm. Stat.* **1991**, *1*, 151–156.
50. Nam, J. Sample size determination in stratified trials to establish the equivalence of two treatments. *Stat. Med.* **1995**, *14*, 2037–2049.
51. Dunnett, C.W.; Gent, M. Significance testing to establish equivalence between treatments with special reference to data in the form of 2×2 tables. *Biometrics* **1977**, *33*, 593–602.
52. Blackwelder, W.C. Proving the null hypothesis in clinical trials. *Control. Clin. Trials* **1982**, *3*, 345–353.
53. Röhm, J.; Mansmann, U. Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biom. J.* **1999**, *41*, 149–170.
54. Marsano, L.S.; West, D.J.; Chan, I.S.F.; Hesley, T.M.; Cox, J.; Hackworth, V.; Greenberg, R.N. A two-dose hepatitis B vaccine regimen: Proof of priming and memory responses in young adults. *Vaccine* **1998**, *16* (6), 624–629.
55. Reed, G.F.; Meade, B.D.; Steinhoff, M.C. The reverse cumulative distribution plot: A graphic method for exploratory analysis of antibody data. *Pediatrics* **1995**, *96*, 600–603.
56. Stine, R.A.; Heyse, J.F. Nonparametric measures of overlap. *Stat. Med.* **2001**, *20*, 215–236.
57. Frey, S.; Dagan, R.; Ashur, Y.; Chen, X.; Ibarra, J.; Kollaritsch, H.; Mazur, M.H.; Poland, G.A.; Reisinger, K.; Walter, E.; Van Damme, P.; Braconier, J.H.; Uhnou, I.; Wahl, M.; Blatter, M.M.; Clements, D.; Greenberg, D.; Jacobson, R.M.; Norrby, S.R.; Rowe, M.; Shouval, D.; Simmons, S.S.; van Hattum, J.; Wennerholm, S.; O'Brien Gress, J.; Chan, I.S.F.; Kuter, B. Interference of antibody production to hepatitis B surface antigen in a combination hepatitis A/hepatitis B vaccine. *J. Infect. Dis.* **1999**, *180* (6), 2018–2022.
58. ICH E10 Guideline. Choice of Control Group and Related Issues in Clinical Trials. In *International Conference on Harmonization (ICH)*; July 2000.
59. Siegel, J.P. Equivalence and noninferiority trials. *Am. Heart J.* **2000**, *139* (4), 166–170.
60. Wiens, B.L. Choosing an equivalence limit for noninferiority or equivalence studies. *Control. Clin. Trials* **2002**, *23* (1), 2–14.
61. Cochran, W.G. Some methods for strengthening the $_2$ -test. *Biometrics* **1954**, *10*, 417–451.
62. Armitage, P. Tests in linear trends in proportions and frequencies. *Biometrics* **1955**, *11*, 375–386.
63. Tukey, J.W.; Ciminera, J.L.; Heyse, J.F. Testing the statistical certainty of a response to increasing doses of a drug. *Biometrics* **1985**, *41*, 295–301.
64. Rom, D.M.; Costello, R.J.; Connell, L.T. On closed test procedures for dose-response analysis. *Stat. Med.* **1994**, *13*, 1583–1596.
65. Wiens, B.L.; Heyse, J.F.; Matthews, H. Similarity of Three Treatments, with Application to Vaccine Development. In *ASA Proceedings of the Biopharmaceutical Section*; **1996**, 203–206.
66. Wiens, B.L.; Iglewicz, B. Design and analysis of three treatment equivalence trials. *Control. Clin. Trials* **2000**, *21*, 127–137.
67. FDA. *Guidance for Industry for the Evaluation of Combination Vaccines for Preventable Diseases: Production, Testing and Clinical Studies*, Center for Biologics Evaluation and Research, Food and Drug Administration, 1997. Internet: <http://www.fda.gov/cber/guidelines.htm>.
68. Blackwelder, W.C. Similarity/Equivalence Trials for Combination Vaccine. In *Combined Vaccines and Simultaneous Administration*; Williams, J.C., Goldenthal, K.L., Burns, D.L., Lewis, B.P., Eds.; Ann. N.Y. Acad. Sci., **1995**, 754, 321–328.
69. Hausdroff, W.P. The need for and design of combination vaccines. *Drug Inf. J.* **1998**, *32*, 13–17.
70. Capizzi, T.P.; Zhang, J. Testing the Hypothesis that Matters for Multiple Primary Endpoints. In *ASA Proceedings of the Biopharmaceutical Section*; 1994, 460–465.

71. Ruger, B. The maximal significance level of the test which consists of rejecting H_0 when K of a given test leads to rejection (German). *Metrika* **1978**, *25*, 171–178.
72. Lawless, J.F. *Survival Models and Methods for Lifetime Data*; John Wiley & Sons: New York, 1982.
73. Van, D.P.; Thoelen, S.; Cramm, M.; DeGroote, K.; Safary, A.; Meheus, A. Inactivated hepatitis A vaccine: Reactogenicity, immunogenicity, and long-term antibody persistence. *J. Med. Virol.* **1994**, *44*, 446–451.
74. Wiens, B.L.; Bohidar, N.; Pigeon, J.; Egan, J.; Hurni, W.; Brown, L.; Kuter, B.; Nalin, D. Duration of protection from clinical hepatitis A disease after vaccination with VAQTA®. *J. Med. Virol.* **1996**, *49*, 235–241.
75. Pigeon, J.G.; Bohidar, N.R.; Zhang, Z.X.; Wiens, B.L. Statistical models for predicting the duration of vaccine-induced protection. *Drug Inf. J.* **1999**, *33* (3), 811–819.
76. Coplan, P.; Chiacchierini, L.; Nikas, A.; Sha, J.; Baumritter, A.; Beutman, K.; Cassidy, W.; Sawyer, M.; Watson, B.; Heyse, J.; Guess, H. Development and evaluation of a standardized questionnaire for identifying adverse events in vaccine clinical trials. *Pharmacoepidemiol. Drug Saf.* **2000**, *9*, 457–471.
77. Ellenberg, S.S. Safety considerations for new vaccine development. *Pharmacoepidemiol. Drug Saf.* **2001**, *10*, 1–5.
78. Mehrotra, D.V.; Heyse, J.F. Multiplicity Considerations in Clinical Safety Analyses. In *Proceedings of the Biopharmaceutical Section*; American Statistical Association, 2001, *in press*.
79. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc., B.* **1995**, *57*, 289–300.
80. Jovanovic, B.D.; Levy, P.S. A look at the rule of three. *Am. Stat.* **1997**, *51*, 137–139.
81. Black, S.; Shinefield, H.; Ray, P.; Lewis, E.; Hansen, J.; Schwalbe, J.; Coplan, P.; Sharrar, R.; Guess, H. Postmarketing evaluation of the safety and effectiveness of varicella vaccine. *Pediatr. Infect. Dis. J.* **1999**, *12*, 1041–1046.
82. Ascherio, A.; Zhang, S.M.; Hernan, M.A.; Olek, M.J.; Coplan, P.M.; Brodovicz, K.; Walker, A.M. Hepatitis B vaccination and the risk of multiple sclerosis. *N. Engl. J. Med.* **2001**, *344* (5), 327–332.
83. Chen, R.T.; Rastogi, S.C.; Mullen, J.R.; Hayes, S.W.; Cochi, S.L.; Donlon, J.A.; Wassilak, S.G. The vaccine adverse reporting system (VAERS). *Vaccine* **1994**, *12*, 42–50.
84. Niu, M.T.; Eriwn, D.E.; Braun, M.M. Data mining in the US vaccine adverse event reporting system (VAERS): Early detection of intussusception and other events after rotavirus vaccination. *Vaccine* **2001**, *19*, 4627–4634.
85. DuMouchel, W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am. Stat.* **1999**, *53*, 177–190.
86. Ellenberg, S.S.; Chen, R.T. The complicated task of monitoring vaccine safety. *Public Health Rep.* **1997**, *112* (1), 10–20.
87. Brewer, T.; Colditz, G.A. Postmarketing surveillance and adverse drug reactions: Current perspectives and future needs. *JAMA* **1999**, *281* (9), 824–829.
88. Rennels, M.B.; Parashar, U.D.; Holman, R.C.; Le, C.T.; Chang, H.-G.; Glass, R.I. Lack of an apparent association between intussusception and wild or vaccine rotavirus infection. *Pediatr. Infect. Dis. J.* **1998**, *10*, 924–925.
89. Murphy, T.V.; Gargiullo, P.M.; Massoudi, M.S.; Nelson, D.B.; Jumaan, A.O.; Okord, C.A.; Zanardi, L.R.; Setia, S.; Fair, E.; LeBaron, C.W.; Wharton, M.; Livongood, M.R. For the rotavirus intussusception investigation team: Intussusception among infants given an oral rotavirus vaccine. *N. Engl. J. Med.* **2001**, *344*, 564–572.
90. Farrington, C.P. Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics* **1995**, *51*, 228–235.
91. Farrington, C.P.; Nash, J.; Miller, E. Case series analysis of adverse reactions to vaccines: A comparative evaluation. *Am. J. Epidemiol.* **1996**, *143*, 1165–1173.
92. Cohen, J. Rethinking a vaccine's risk. *Science* **2001**, *293*, 1576–1577.
93. Sadoff, J.; Heyse, J.F.; Heaton, P.; Dallas, M. Designing a Study to Evaluate Whether the Merck Rotavirus Vaccine is Associated with a Rare Adverse Reaction (Intussusception). In *U.S. Food and Drug Administration Workshop: Evaluation of New Vaccines: How Much Safety Data*, Bethesda, MD; November 14–15, 2000.
94. O'Neill, R.T. Statistical concepts in the planning and evaluation of drug safety from clinical trials in drug development: Issues in international harmonization. *Stat. Med.* **1995**, *14*, 1117.
95. Gillings, D.; Koch, G. The application of the principle of intention-to-treat to the analysis of clinical trials. *Drug Inf. J.* **2000**, *34*, 569–578.
96. Lachin, J.M. Statistical consideration in the intent-to-treat principle. *Control. Clin. Trials* **2000**, *21*, 167–189.
97. Lange, S. The all randomized/full analysis set (ICH E9)—May patients be excluded from the analysis? *Drug Inf. J.* **2001**, *35*, 881–891.
98. Peduzzi, P.; Donta, S.; Iwane, M. A review of the design of vaccine efficacy trials and a proposal for the design of the VA cooperative study of active immunotherapy of HIV infection. *Control. Clin. Trials* **1997**, *18* (5), 397–419.
99. Horne, A.D.; Blackwelder, W.C. Statistical Evaluation of Preventive Vaccine Efficacy: Current Approaches. In *American Statistical Association Joint Meetings*, Dallas, TX; August 8–12, 1998.
100. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; John Wiley & Sons: New York, 1987.
101. Heitjan, D. Ignorability and coarse data: Some biomedical examples. *Biometrics* **1993**, *49*, 1099–1109.
102. Moulton, L.H.; Halsey, N.A. A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics* **1995**, *51*, 1570–1578.
103. Fleiss, J.L. Analysis of data from multiclinic trials. *Control. Clin. Trials* **1986**, *7*, 267–275.
104. Senn, S. Some controversies in planning and analysing multi-centre trials. *Stat. Med.* **1998**, *17*.
105. Gail, M.; Simon, R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* **1985**, *41*, 361–372.
106. Mehrotra, D.V.; Chan, I.S.F. Testing for Treatment by Stratum Interaction in Stratified Comparative Binomial Trial. In *American Statistical Association Joint Meetings*; August 13, 2000.

107. Mehotra, D.V. Stratification issues with binary endpoints. *Drug Inf. J.* **2001**, 35, 1343–1350.
108. Wiens, B.L.; Heyse, J.F. Testing for Interaction in Studies of Noninferiority. In *Proceedings of the Biopharmaceutical Section*; American Statistical Association, 2001, *in press*.
109. FDA. *Guidance for Clinical Trial Sponsors, on the Establishment and Operation of Clinical Trial Data Monitoring Committees. Draft*, Center for Biologics Evaluation and Research, Food and Drug Administration, November, 2001. Internet: <http://www.fda.gov/cber/guidelines.htm>.
110. O'Brien, P.C.; Fleming, T.R. A multiple testing procedure for clinical trials. *Biometrics* **1979**, 35(3), 549–556.
111. Lan, K.K.G.; DeMets, D.L. Discrete sequential boundaries for clinical trials. *Biometrika* **1983**, 70, 659–663.
112. Lan, K.K.G.; Simon, R.; Halperin, M. Stochastically curtailed tests in long-term clinical trials. *Commun. Stat., C.* **1982**, 1, 207–219.
113. Ware, J.H.; Muller, J.E.; Braunwald, E. The futility index: An approach to the cost-effective termination of randomized clinical trials. *Am. J. Med.* **1985**, 78, 635–643.
114. Pallay, A. On the timing of a futility analysis in clinical trials. *J. Biopharm. Stat.* **2000**, 10 (3), 399–405.
115. Prentice, R.L. Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat. Med.* **1989**, 8, 431–440.
116. Freedman, L.; Graubard, B.I.; Schatzkin, A. Statistical validation of intermediate endpoints for chronic disease. *Stat. Med.* **1992**, 11, 167–178.
117. Lin, D.Y.; Fleming, T.R.; DeGruttola, V. Estimating the proportion of treatment effect explained by a surrogate marker. *Stat. Med.* **1997**, 16, 1515–1527.
118. Buyse, M.; Molenberghs, G. The validation of surrogate endpoints in randomized experiments. *Biometrics* **1998**, 54, 1014–1029.
119. DeGruttola, V.G.; Clax, P.; DeMets, D.L.; Downing, G.J.; Ellenberg, S.S.; Frideman, L.; Gail, M.H.; Prentice, R.; Wittes, J.; Zeger, S.L. Consideration in the evaluation of surrogate endpoints in clinical trials: Summary of a national institutes of health workshop. *Control. Clin. Trials* **2001**, 22, 485–502.
120. Anderson, P. The protective level of serum antibodies to the capsular polysaccharide of *H. influenzae* type b. *J. Infect. Dis.* **1984**, 149, 1034.
121. Chen, R.T.; Markowitz, L.E.; Albrecht, P.; Stewart, J.A.; Mofenson, L.M.; Preblud, S.R.; Orenstein, W.A. Measles antibody: Reevaluation of protective titers. *J. Infect. Dis.* **1990**, 162, 1036–1042.
122. Wilson, J.N.; Nokes, D.J.; Dimmock, N.J. Analysis of the relationship between immunogenicity and immunity for viral subunit vaccines. *J. Med. Virol.* **2001**, 64 (4), 560–568.
123. White, C.J.; Kuter, B.J.; Ngai, A.; Hildebrand, C.S.; Isganitis, K.L.; Patterson, C.M.; Capra, A.; Miller, W.J.; Krah, D.L.; Provost, P.J. Modified cases of chickenpox after varicella vaccination: Correlation of protection with antibody response. *Pediatr. Infect. Dis. J.* **1992**, 11, 19–23.
124. Chan, I.S.F.; Li, S.; Matthews, H.; Chan, C.; Vessey, R.; Sadoff, J.; Heyse, J. Use of statistical models for evaluating antibody response as a correlate of protection against varicella. *Stat. Med.*, *in press*.
125. Li, S.; Chan, I.S.F.; Matthews, H.; Heyse, J.F.; Chan, C.Y.; Kuter, B.J.; Kaplan, K.M.; Vessey, S.J.R.; Sadoff, J.C. Childhood vaccination against varicella: Inverse relationship between 6-week postvaccination varicella antibody response and likelihood of long-term breakthrough infection. *Pediatr. Infect. Dis. J.* *in press*.
126. Drummond, M.F.; McGuire, A. *Economic Evaluation in Health Care, Merging Theory with Practice*; Oxford University Press, 2001.
127. Drummond, M.F.; O'Brien, B.; Stoddart, G.L.; Torrance, G.W. *Methods for the Economic Evaluation of Health Care Programmes*; Oxford University Press, 1997.
128. Gold, M.R.; Siegel, J.E.; Russell, L.B.; Weinstein, M.C. *Cost-Effectiveness in Health and Medicine*; Oxford University Press, 1996.
129. Edmunds, W.J.; Medley, G.F.; Nokes, D.J. Evaluating the cost-effectiveness of vaccination programmes: A dynamic perspective. *Stat. Med.* **1999**, 18, 3263–3282.
130. Halloran, M.E.; Cochi, S.L.; Lieu, A.L.; Wharton, M.; Fehrs, L. Theoretical epidemiologic and morbidity effects of routine varicella immunization of preschool children in United States. *Am. J. Epidemiol.* **1994**, 140, 81–104.
131. Brisson, M.; Edmunds, W.J.; Gay, N.J.; Law, B.; DeSerres, G. Analysis of varicella vaccine breakthrough rates: Implications for the effectiveness of immunisation programmes. *Vaccine* **2000**, 18, 2775–2778.
132. Schuette, M.C.; Hethcote, H.W. Modeling the effects of varicella vaccination programs on the incidence of chickenpox and shingles. *Bull. Math. Biol.* **1999**, 61, 1031–1064.

Cluster Randomized Trials: Sample Size

Sin-Ho Jung

Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, U.S.A.

Abstract

In a cluster randomization trial, clusters consisting of multiple subjects are randomized to two or more treatment arms, and observations are collected from the subjects of each cluster while receiving the same intervention. In such a trial, inferences are usually made at subject level while randomization is done at cluster level. Since the subjects in each cluster share some common frailties, their responses are usually positively correlated, both the sample size calculation and the subsequent statistical analysis must take into account the correlation. In this chapter, we investigate some estimators of binary response rate and sample size method for comparing the response rates between two arms using these estimators when the number of subjects is random among clusters.

Keywords: Binary data; Intraclass correlation coefficient; Varying cluster size; Weighted estimator.

1 INTRODUCTION

Over the past decades there has been increasing popularity of cluster randomization trials in various clinical fields including dentistry and ophthalmology. In such trials, clusters consisting of multiple subjects, called subunits, are randomized to two or more treatment arms, and observations are collected from the subjects of each cluster while receiving the same intervention. So, inferences are usually made at subunit level while randomization is done at cluster level. Since the responses of subjects within a cluster are usually positively correlated due to the shared factors among them, both the sample size calculation and the subsequent statistical analysis must take into account the intraclass dependency among subunits.

In estimating sample size for cluster randomization trials one usually assumes that the number of subunits within each cluster, called cluster size, is constant across clusters for simplicity. However, cluster sizes are not constant in many cluster randomization trials. The sample size formula with for binary outcomes with equal cluster size m is obtained by the standard formula developed for studies randomizing subjects by a variance inflation factor, $[1 + (m - 1)\rho]$, where ρ is the intraclass correlation coefficient (ICC). When the cluster sizes vary, one may replace m with an estimate of the average cluster size as in Manatunga et al.^[1] This approximation is likely to underestimate the actual required sample size (Donner and Klar^[2]).

When cluster size varies among clusters, Kang, Ahn and Jung^[3] evaluated investigated naive sample size calculation methods assuming that cluster sizes are fixed at its mean value or its maximum value. Ahn, Jung and Kang^[4] proposed

weighted chi-squared test statistics for clustered binary outcomes using equal weights to subunits, weights to clusters, and optimal weights. In this chapter, we present a sample size formula for each of these weighted chi-squared tests.

We may choose a time to event variable as the primary endpoint of a cluster randomization trial. Jung and Jeong^[5] proposed weighted log-rank tests and Jung^[6] proposed a sample size calculation method for cluster randomization trials with a time to event endpoint.

2 WEIGHTED CHI-SQUARED TEST STATISTICS

In this chapter, we assume that clusters are randomized between two treatment arms, but the extension to more than two arms is straightforward. Suppose that n_k clusters are randomized to arm k ($= 1, 2$), and cluster i ($= 1, \dots, n_k$) in arm k has m_{ki} subunits. For subunit j ($= 1, \dots, m_{ki}$) of cluster i ($= 1, \dots, n_k$) in arm k ($= 1, 2$), let y_{kij} be a binary random variable taking 1 for response and 0 for no response. By clustered data, we assume that $(y_{ki1}, \dots, y_{kim_k})$ have a common marginal response probability $P(y_{kij} = 1) = p_k$ and an ICC, $\rho_k = \text{corr}(y_{kij}, y_{kij'})$ for $j \neq j'$. We want to test $H_0: p_1 = p_2$ vs. $H_1: p_1 \neq p_2$. Let $y_{ki} = \sum_j y_{kij}$, $y_k = \sum_i y_{ki} = \sum_i \sum_j y_{kij}$, $M_k = \sum_i m_{ki}$, $M = M_1 + M_2$ and $n = n_1 + n_2$.

Let $\hat{p}_{ki} = m_{ki}^{-1} \sum_{j=1}^{m_{ki}} y_{kij}$ denote an estimator of p_k from cluster i in arm k . Jung and Ahn^[7] propose a family of weighted estimator of p_k

$$\hat{p}_k = \frac{\sum_{i=1}^{n_k} w_{ki} \hat{p}_{ki}}{\sum_{i=1}^{n_k} w_{ki}},$$

where $w_{ki} \geq 0$ are weights assigned to cluster i in arm k . Jung, Kang and Ahn^[8] propose a sample size method for one-sample testing using the weighted estimator. Using these results, Ahn, Jung and Kang^[4] develop weighted chi-squared test statistics Z^2 to compare response rates between two arms, where

$$Z = \frac{\sqrt{n}(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}$$

Where $\hat{\sigma}_k^2 = n \sum_{i=1}^{n_k} w_{ki}^2 \hat{\sigma}_{ki}^2 / \left(\sum_{i=1}^{n_k} w_{ki} \right)^2$ is a variance estimator of $\sqrt{n} \hat{p}_k$,

$$\hat{\sigma}_{ki}^2 = \frac{\hat{p}\hat{q}\{1 + (m_{ki} - 1)\hat{p}\}}{m_{ki}}$$

is a variance estimator of $\hat{p}_{ki}$, $\hat{p} = \sum_{k=1}^2 \sum_{i=1}^{n_k} w_{ki} \hat{p}_{ki} / \sum_{k=1}^2 \sum_{i=1}^{n_k} w_{ki}$ is an estimator of response probabilities under H_0 : $p_1 = p_2$, and $\hat{q} = 1 - \hat{p}$. Under H_0 , for large n_1, n_2 , Z approximately $N(0, 1)$. Hence, we can reject H_0 if $|Z| > z_{1-\alpha/2}$, the $100(1 - \alpha/2)$ -th percentile of $N(0, 1)$.

Weights $w_{ki} = m_{ki}$ assign equal weights to subunits, $w_{ki} = 1$ assign equal weights to clusters, and

$$w_{ki} = \frac{\hat{\sigma}_{ki}^{-2}}{\sum_{i'=1}^{n_k} \hat{\sigma}_{ki'}^{-2}}$$

yield an optimal test statistic, where

$$\hat{\sigma}_{ki}^2 = \frac{\hat{p}\hat{q}\{1 + (m_{ki} - 1)\hat{p}\}}{m_{ki}},$$

$\hat{p}$ is an estimator of ICC. One of most popular ICC estimators is the analysis of variance (ANOVA) estimator, given by

$$\hat{p} = \frac{\text{msc} - \text{mse}}{\text{msc} + (\tilde{m} - 1)\text{mse}},$$

where

$$\text{msc} = \sum_k \sum_i m_{ki} (\hat{p}_{ki} - \hat{p}_k)^2 / (n - 2),$$

$$\text{mse} = \sum_k \sum_i y_{ki} (1 - \hat{p}_{ki}) / (M - n),$$

$$\tilde{m} = \left(M - \sum_k M_k^{-1} \sum_i m_{ki}^2 \right) / (n - 2), \hat{p}_{ki} = y_{ki} / m_{ki} \text{ and } \hat{p}_k = y_k / M_k.$$

3 SAMPLE SIZE CALCULATION

We derive a sample size formula of the weighted chi-squared tests assuming equal ICC with ρ for the two arms and a nearby alternative hypothesis $\sqrt{n}(p_1 - p_2) = o_p(1)$. Let $a_k = n_k/n$ denote the allocation proportion of arm k . Then, $\hat{\sigma}_k^2$ converges to

$$\sigma_k^2 = a_k^{-1} E(w_{ki}^2 \hat{\sigma}_{ki}^2) / \{E(w_{ki})\}^2$$

It is easy to show that

$$\sigma_k^2 = \frac{\bar{p}\bar{q}\{(1-\rho)\mu_{(1)} + \rho\mu_{(2)}\}}{a_k \mu_{(1)}^2}$$

for $w_{ki} = m_{ki}$,

$$\sigma_k^2 = \frac{\bar{p}\bar{q}\{(1-\rho)\mu_{(-1)} + \rho\}}{a_k}$$

for $w_{ki} = 1$, and

$$\sigma_k^2 = \frac{\bar{p}\bar{q}}{a_k \gamma}$$

for the optimal weights, where $\bar{p} = a_1 p_1 + a_2 p_2$, $\mu_{(l)}$ denotes the l -th moment of distribution of cluster sizes m_{ki} and

$$\gamma = E \left\{ \frac{m_{ki}}{(1-p) + m_{ki}p} \right\}.$$

Given n , the power of a test with a 2-sided significance level α is

$$1 - \beta = \bar{\Phi} \left(z_{1-\alpha/2} - \frac{\sqrt{n}\Delta}{\sqrt{\sigma_1^2 + \sigma_2^2}} \right),$$

where $p_1 - p_2 = \Delta$ under H_1 , $\bar{\Phi}(\cdot)$ is the survivor function of the standard normal distribution, and $\sigma^2 = a_k \sigma_k^2$. Hence, the sample size is given as

$$n = \frac{\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{a_1 a_2 \Delta^2} \quad (1)$$

or $n_k = a_k n$.

In summary, the sample size calculation for a cluster randomization trial may be proceeded as follows.

- Select a test statistic to be used for the final analysis
- Specify design parameters
 1. $\Delta = p_1 - p_2$ under H_1
 2. $(\alpha, 1 - \beta)$

Table 1 Distribution of cluster sizes, $f(m) = P(m_{ki} = m)$

m	5	6	7	8	9	10	11	12	13
$f(m)$	2/32	1/32	3/32	2/32	7/32	9/32	2/32	3/32	3/32

3. ICC ρ
4. Distribution of m_{ki}
5. Allocation proportions (a_1, a_2)
- Calculate the sample size using formula (1)

4 EXAMPLE

Suppose that we want to develop a teratologic animal study to randomize pregnant rats between a control diet and a chemically treated diet during pregnancy and lactation, and compare the 21-day survival rates of the pups between the two arms. We take the data from Weil^[9] to specify the design parameters for our study. Based on the Weil^[9] data, we assume $\rho = 0.25$ and a distribution of cluster sizes, $f(m) = P(m_{ki} = m)$, as given in Table 1. Suppose that we want to show that the 21-day survival rate is about $p_1 = 0.9$ for the control diet and $p_2 = 0.75$ for the chemically treated diet. Under this design setting, we have $\bar{p} = 0.825$, $\mu_{(-1)} = 0.112$, $\mu_{(1)} = 9.47$, $\mu_{(2)} = 94.03$, and $\sigma^2 = 4.929 \times 10^{-2}$ for the equal weights to subunits, $= 4.823 \times 10^{-2}$ for the equal weights to clusters, and $= 4.802 \times 10^{-2}$ for the optimal weights. Hence, for $(\alpha, 1 - \beta) = (0.05, 0.9)$ and a balanced allocation $a_1 = a_2 = 1/2$, we need $n = 92$ ($n_1 = n_2 = 46$), $n = 90$ ($n_1 = n_2 = 45$) and $n = 88$ ($n_1 = n_2 = 44$) using equal weights for subunits, equal weights for clusters and optimal weights, respectively.

REFERENCES

1. Manatunga, A.K.; Hudgens, M.G. Chen, S. Sample size estimation in cluster randomized studies with varying cluster size. *Biometrical Journal* **2001**, *43* (1), 75–86.
2. Donner, A.; Klar, N. Design and analysis of cluster randomization trials in health research. *Int. J. Epidemiol.* **2001**, *30* (2), 407–408.
3. Kang, S.H.; Ahn, C.; Jung, S.H. Sample size calculation for dichotomous outcomes in cluster randomization trials with varying cluster size. *Drug Information Journal* **2003**, *37* (1), 109–114.
4. Ahn, C.; Jung, S.H.; Kang, S.H. An evaluation of weighted chi-square statistics for site-specific binary data. *Drug Information Journal* **2003**, *37* (1), 91–99.
5. Jung, S.H.; Jeong, J.H. Rank tests for clustered survival data. *Lifetime Data Analysis* **2003**, *9* (1), 21–33.
6. Jung, S.H. Sample size calculation for weighted rank tests comparing survival distributions under cluster randomization: a simulation method. *Biopharmaceutical Statistics* **2007**, *17* (5), 839–849.
7. Jung, S.; Ahn, C. Estimation of response probability in correlated binary data: A new approach. *Drug Information Journal* **2000**, *34* (2), 599–604.
8. Jung, S.; Kang, S.; Ahn, C. Sample size calculations for clustered binary data. *Statistics in Medicine* **2001**, *20*, 1971–82.
9. Weil, C.S. Selection of the valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. *Food and Cosmetics Toxicology* **1970**, *8* (2), 177–182.

Cluster Trials

John S. F. Preisser

Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, North Carolina, U.S.A.

Bing Lu

Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, U.S.A.

Abstract

This entry provides several examples of cluster trials and considers, in detail, the design and analysis of a cluster trial in which an intervention is targeted at changing physician behavior to increase cancer screening in primary medical care practice settings.

INTRODUCTION

Cluster trials are studies that evaluate interventions delivered to intact social groups or clusters, such as communities, churches, schools, workplaces, and medical practices, while outcomes are measured on members of those groups. Today, cluster trials are conducted in social science, medicine, public health, and, to a growing degree, biopharmaceutical science. Cluster trials have been known by many names including “cluster-unit trials,” “cluster-randomized trial,” “group-randomized trials,” and “community trials.” The latter term recognizes the long history of cluster trials in assessing the effect of social interventions in communities. While “cluster trial” may be the most generic term used, it reflects the diversity of settings that use cluster assignment to test interventions. Examples of cluster trial settings include randomized trials with a small number of clusters, non-randomized trials with many clusters, and many settings (e.g., medical practices) where the ratio of the number of clusters in a trial to the average number of subjects per cluster is somewhere along the spectrum between the relative extremes of the typical longitudinal study (large ratio) and the classical community trial (small ratio). This article provides several examples of cluster trials and considers, in detail, the design and analysis of a cluster trial in which an intervention is targeted at changing physician behavior to increase cancer screening in primary medical care practice settings.

The distinctive feature of cluster trials is the presence of intraclass correlation (ICC) among members within groups that arises from restricting randomization to groups instead of to individuals. The ICC will increase the variance of a statistic for assessing intervention beyond what would be expected with random assignment of members to intervention conditions. For clusters of equal sizes m in a posttest-only design, the variance inflation factor (VIF), also referred to as the design effect (DE),

is $\varphi = 1 + (m - 1) \rho$, where ρ is denoted as ICC. Ignoring the correlation associated with clustering may lead to inflation of type I error or the increased chance to declare a statistically significant difference among treatment conditions when no such difference exists. Assignment of interventions by cluster, instead of by individual, decreases statistical power. Therefore to achieve a pre-specified level of power, the total number of individuals enrolled in a cluster trial must exceed the number enrolled in a comparable clinical trial that assigns treatments to individuals. Both the design and analysis of cluster trials require special care.

Another common feature of cluster trials is the small number of clusters enrolled as compared with a relatively large number of subjects within clusters. Many community trials enroll 10 communities or fewer because of the prohibitive financial costs and logistical challenges associated with implementing an often multifaceted intervention to inhabitants of sizable geographic areas, such as villages, towns, cities, or regions. These studies may enroll several hundreds of subjects per community. This feature contrasts with the data structure associated with clustered data from a longitudinal study where individuals are the clusters and each has maybe three, four, or five observations. Because of the large cluster sizes in cluster trials, the presence of a typically small intraclass correlation translates into a substantial design effect that must be considered in the design and statistical analysis. In effect, design and analysis considerations are driven mainly by the number of clusters as opposed to the total number of individuals. Moreover, with a small number of clusters, the degrees of freedom available to estimate cluster-level statistics for assessing treatment differences are limited.

Much as randomization is a cornerstone of the individual-randomized clinical trial, so randomization is applied, when possible, to the cluster trial. Randomization of clusters to intervention conditions lends significant scientific credibility to the study and its subsequent evaluation.

This is especially so when the number of clusters is small. However, in this context, there is no guarantee that simple randomization will result in important baseline covariates being balanced across treatment conditions. Therefore consideration has been given to alternative randomization schemes, including randomization by stratification and within matched pairs. Statistical analysis should consider adjusting for baseline imbalance between treatment groups likely to be present.

WHY ARE CLUSTER TRIALS CONDUCTED?

Given the reduction in statistical power associated with cluster trials, there must be clear motivation for conducting a cluster trial vs. an individual randomized trial. Three reasons a cluster trial may be deemed necessary are as follows: (i) the inherent nature of the intervention; (ii) administrative, political, or ethical considerations; and (iii) concern for the scientific integrity of the study such as issues of blinding. Some cluster trials are motivated by multiple reasons.

Interventions in many trials are, by nature, group-oriented. For example, the COMMIT Trial was a randomized community trial designed to bring diverse organizations, institutions, and individuals together to conduct activities to encourage smoking cessation among the residents of 22 communities.^[1] The philosophy behind the multifaceted intervention was that the community-wide messages against smoking would reach the vast majority of residents and would alert smokers to opportunities for cessation. Similar multi-level community-wide strategies have been employed to reduce underage drinking.^[2,3]

Many cluster trials in the medical practice setting also employ interventions that are, by nature, group-oriented. For example, an intervention may attempt to change physician behavior by targeting medical office practice patterns, while outcomes are collected on patients through surveys or their medical charts. Recognizing that physicians may be at different stages in their readiness to change practice behavior, the Partners for Prevention Project (PARTNERS) employed a multifaceted intervention designed to increase their cancer screening behavior.^[4] The outcomes, measured on a sample of patients in a practice, were whether or not they had been screened for different types of cancer. The cholesterol education and research trial (CEART) is a cluster randomized trial testing the effectiveness of translating the ATP III (Adult Treatment Panel III) cholesterol guidelines into clinical practice, with primary care physician practices as the unit of randomization and patients as the unit of data collection. The prevention arm received an intervention of academic detailing, a computerized patient activation tool placed in the primary care providers office waiting room, and an interactive ATP III Cholesterol Guideline decision support tool made for a hand help device (PDA); the usual care arm received minimal

academic detailing, no computerized patient activation tool, and a PDA without any guidelines.^[5] A multifaceted intervention was also used in a cluster randomized trial to increase physician compliance to guidelines for prescribing antihypertensive and cholesterol-lowering drugs for the primary prevention of cardiovascular disease.^[6]

Cluster trials are used in public health studies to combat infectious diseases because interaction among members within communities means that community indices of health are of primary concern. A cluster trial in 10 communities in Uganda tested whether community-level control of sexually transmitted diseases via a mass antibiotic treatment program would reduce the incidence of HIV infection.^[7] Although drugs were administered within families, the intervention conditions were assigned at the community level because infection rates in communities were the outcomes of interest. In a study to combat sexually transmitted diseases through the promotion of female condom use in 12 agricultural villages in Kenya, political considerations led to a randomized cluster trial.^[8] Because village leaders, religious leaders, and teachers have pronounced influence on inhabitants, it was felt important to include, as part of the intervention, interviews seeking their opinions. Interventions were assigned at the village level in part because of the political consideration to be all inclusive in these communities where social networks were very strong. An education program based on group meetings was employed. A third example is a matched-pair cluster trial in which villages in northwest Iran were randomized to intervention or control with the goal to evaluate the effects of insecticide-impregnated dog collars of reducing the incidence rate of leishmaniasis, an infectious disease, in children.^[9] Transmission is by the bite of a sand fly, and domestic dogs were the principal host. The study showed that application of the collars to all domestic dogs in the intervention villages reduced infection in children of those villages relative to control villages. Villages were separated by distances greater than sand fly flight ranges supporting the use of villages as clusters.

Some interventions are applied at the level of individuals, but because of concerns for the scientific integrity of the study, the intervention conditions are assigned to groups of individuals. A common concern is bias that may result from cross-contamination of the treatment conditions. If subjects who receive the control treatment are in close contact to those that receive the intervention, then the former may become exposed to the intervention and a biased estimate of the treatment effect may result. Although an individual randomized trial would have been possible in the Kenyan female condom use study, contamination within sites was a concern. On the other hand, little contamination was expected in the community trial because “sites were distant enough to preclude social or sexual contacts between intervention and control participants, none of whom own private vehicles.” In an intervention study comparing two preoperative drug regimens

for an eye disease, groups of patients undergoing surgery in the same day were randomized to the same regimen to address concern about the possibility of incorrect regimen assignments because of the heavy volume of daily operations.^[10]

COMMON CLUSTER TRIAL DESIGNS

Study designs for cluster trials can be classified as single-point or multiple time-point designs and as to whether any form of stratification or matching is used. The two most common designs for multiple time-point cluster trials are the nested cross-sectional and nested cohort designs. In both designs, the same clusters are followed over time. In a nested cross-sectional design, different random samples of subjects are selected within each cluster over time, whereas in a nested cohort design, the same subjects are followed over time. There are many factors to consider in choosing between these two designs. These include the rate of attrition of study subjects, length of the trial, population stability, effect of measurement on future responses, statistical efficiency, cost, and objective of the study.^[11] Cross-sectional designs are used when the focus is on the change in a population over time. Cohort designs address questions of within-individual change, and they are more powerful barring dropout. However, a long intervention period might result in substantial dropout in a cohort design.^[12] Because each has its relative strengths and weaknesses, the design of the COMMIT trial incorporated both cross-sectional and cohort components. In the Kenyan female condom use study, creation of cohort of female employees took advantage of the stable populations of permanent employees at the 12 sites, along with the relatively short intervention period.^[8] Many cluster randomized trials in medical practices such as the CEART trial use nested cohort designs with a limited number of clusters.^[5] However, the PARTNERS trial used repeated cross-sectional samples because the focus of inference was on physician screening rates and not on individual patients.^[4]

Studies having a single time point have been called posttest-only designs.^[13] For example, there is no baseline response in the cohort component of the COMMIT trial because all persons were smokers at baseline. Other studies record baseline outcome variables but do not use these data in the primary analysis. If a study randomizes a large number of groups in each condition, then randomization will likely have made the communities in the conditions comparable at baseline. Because the inclusion of the outcome at baseline as a response variable in the model typically reduces power to detect an intervention effect, the analyst may not want to include it unless inclusion is deemed necessary. On the other hand, use of the baseline value of the primary outcome as a covariate in the model for the individual level data, possible in a cohort but not a cross-sectional design, may increase statistical power.

Matching is a common design strategy in cluster trials with a few to moderate number of clusters. In matched-pair cluster randomized trials, pairs of clusters are matched with respect to one or more cluster characteristics that may be linked to the outcome. Within pairs, one cluster is randomized to intervention and the other receives control. Well-made matches not only diminish the chance of baseline imbalance in covariates, but also increases statistical power to detect an intervention effect. Assessment of smoking quit rates in the COMMIT trial of the intervention was based on 11 matched community pairs with matches based on population size and baseline smoking prevalence. In the insecticide-impregnated dog collar trial, villages were matched according to baseline infection rates. Matching in the design does not necessarily need to be followed by a matched-pair analysis. If matches were ineffective, then a matched-pair analysis may be less statistically efficient than an unmatched analysis.^[1] In the case of trials with small numbers of clusters, matching may be overused. For example, loss to follow-up of a single member of a pair may lead to a serious power loss.

STATISTICAL ANALYSIS OF CLUSTER TRIALS

There are several statistical approaches to the analysis of cluster-unit trials. Direct hypothesis testing approaches include tests applied to cluster summary statistics and variants of common tests applied to individual-level data that are adjusted for clustering. Alternatively, modeling approaches for individual level data are broadly applicable because they can easily handle covariates, unequal cluster sizes, and multiple time points or treatment conditions. Models that are extensions of generalized linear models provide the context for hypothesis tests of intervention effects using population-averaged or conditional models. For simplicity, we restrict attention to cluster trials with only two treatment conditions: intervention and control. Binary outcomes are treated here because analogous methods for continuous responses tend to be better established.

Cluster-Level Analysis

One approach to the analysis of cluster trials is to treat the cluster as the unit of analysis by creating a summary statistic for each cluster such as the proportion of respondents in the cluster exhibiting the outcome following the treatment, in the case of a binary response. For unmatched studies, the standard two-sample *t*-test may be used to compare intervention and control conditions. One criticism of this approach is that the assumptions that the cluster-specific event rates are normally distributed with equal variances are not strictly satisfied. These assumptions will clearly be violated if there is variation in cluster sizes. However, when the number of clusters assigned to each intervention group is equal, the *t*-test is known to be fairly robust to

violations of the underlying assumptions.^[12] The t -test also applies to unweighted cluster-level mean responses for continuous outcomes.^[12]

The parametric assumptions of the t -test may be avoided by using non-parametric procedures such as the Wilcoxon rank sum test. Randomization tests, which use the actual values of observations rather than their ranks, tend to be more powerful.^[12] In the COMMIT trial, randomization tests were applied to the within matched-pair difference in community level smoking quit rates.^[1] Exact p -values are determined from the permutation distribution of the test statistic under the null hypothesis of no difference between conditions. These summary statistic methods may be adapted to adjust for individual-level covariates.^[1]

Adjusted Chi-Square Test for Individual-Level Data

Another approach to the analysis of binary outcomes in cluster trials is to adjust the standard Pearson chi-square test for the clustering effect. Let $\hat{P}$ be the overall event rate observed in the study, $\hat{P}_k$, $k = 1, 2$, be the event rates for the k th treatment condition, and M_k be the total number of individuals in condition k . Then an adjusted one degree of freedom chi-square statistic is

$$\chi_A^2 = \sum_{k=1}^2 \frac{M_k (\hat{P}_k - \hat{P})^2}{\phi_k \hat{P}(1 - \hat{P})} \quad (1)$$

where the correction factor ϕ_k depends on the intraclass correlation and the cluster sizes.^[12] In the case of equal cluster sizes m , $\phi_1 = \phi_2 \equiv \phi$, and $\phi = 1 + (m - 1)\rho$, where ρ is the average pairwise correlation within clusters assumed to be constant across clusters. The parameter ϕ is known as the variance inflation factor because it represents the multiplicative factor by which the variance of cluster summary statistics are inflated because of the correlation arising from the cluster design relative to a study design that randomizes treatments to individuals who are statistically independent of one another. If the data are uncorrelated ($\rho = 0$), then Eq. 1 is the standard chi-square test. An estimate of ρ can be obtained by applying standard analysis of variance to the binary outcomes:

$$\hat{\rho} = \frac{MSB - MSW}{MSB + (m - 1)MSW} \quad (2)$$

where

$$MSB = \sum_{k=1}^2 \sum_{i=1}^{n_k} m (\hat{P}_{ki} - \hat{P}_k)^2 / (K - 2) \quad (3)$$

and

$$MSW = \sum_{k=1}^2 \sum_{i=1}^{n_k} m \hat{P}_{ki} (1 - \hat{P}_{ki}) / (M - K) \quad (4)$$

where $\hat{P}_{ki}$ is the event rate observed for the i th cluster in the k th condition, n_1 and n_2 are the number of clusters assigned to the intervention and control conditions, respectively, $M = M_1 + M_2$, and $K = n_1 + n_2$ is the total number of clusters. For continuous responses, the two-sample t -test can be adjusted with a measure of the pooled variance that depends on an estimate of ρ .^[12] The ICC can also be estimated when cluster sizes are variable.^[12]

Individual-Level Analysis with Population-Averaged Models

The population-averaged or marginal model approach extends generalized linear models to account for clustered data. Population-averaged models provide a natural approach for the analysis of cluster-unit trials because differences in population means or proportions between intervention and control conditions are typical parameters of interest. A population-averaged model parameter for the intervention effect describes how the response average changes across various subsets of the study population where subsets are defined by covariate values. These models are commonly estimated with the generalized estimating equations procedure.^[14] Let $\mathbf{Y}_i$ be the response vector for cluster or community $i = 1, \dots, K$, where $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ij}, \dots, Y_{im_i})$ is the vector of responses from the m_i individuals in the i th cluster. Let μ_i be the vector of population marginal means, $E[\mathbf{Y}_i] = \mu_i$. A generalized linear model is $g(\mu_{ij}) = X_{ij}'\beta$ where $g(\cdot)$ is the link function and X_{ij} is the $p \times 1$ vector of covariates for the j th individual in the i th cluster. Estimation of β is performed by solving

$$\sum_{i=1}^K \mathbf{D}_i'(\mathbf{X}_i, \beta) \mathbf{V}_i^{-1} [\mathbf{Y}_i - \mu_i(\beta)] = 0 \quad (5)$$

where $\mathbf{D}_i(\mathbf{X}_i, \beta) = \partial \mu_i / \partial \beta$, $\mathbf{V}_i = \mathbf{A}_i \mathbf{R}_i \mathbf{A}_i'$, where $\mathbf{A}_i = \text{diag}\{v_{ij}^{1/2}\}$ is a diagonal matrix with $v_{ij} = \text{var}(Y_{ij})$ and $\mathbf{R}_i(\alpha)$ is a working correlation matrix that depends on an unknown nuisance parameter vector α . For binary data, $v_{ij} = \mu_{ij}(1 - \mu_{ij})$. The GEE estimate $\hat{\beta}$, solved by iteratively reweighted least squares, has an asymptotically Gaussian distribution with mean β and covariance matrix estimated by $(\sum_{i=1}^K \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i)^{-1}$, the model-based variance estimator which provides valid inference when $\mathbf{R}_i$ is correctly specified [i.e., $\text{corr}(\mathbf{Y}_i) = \mathbf{R}_i$]. Otherwise $\text{var}(\hat{\beta})$ may be estimated by the “robust” or empirical covariance estimator^[15,16] that consistently estimates $\text{var}(\hat{\beta})$ even if $\mathbf{R}_i(\alpha)$ does not equal the true correlation matrix. This strategy permits the use of simple working models for the correlation structure which is considered a nuisance factor in the analysis that tests for intervention effects.

The most common specification of α for a cluster trial is a single parameter, $\rho = \text{corr}(Y_{ij}, Y_{ij'})$, denoting a common or exchangeable correlation, for all pairs of elements from $\mathbf{Y}_i$. A common approach is to use simple method-of-moment

estimators for the elements of α ,^[15] however, this procedure does not provide standard error estimates. Standard errors are desirable to provide information regarding a plausible range of correlation values when performing sample size calculations in planning future cluster trials. In the case of binary data, define the $[m_i(m_i - 1)/2 \times 1]$ vectors $\mathbf{Z}_i = \{Z_{i12}, Z_{i13}, \dots, Z_{i(mi-1)mi}\}'$ and $\rho_i = \{\rho_{i12}, \rho_{i13}, \dots, \rho_{i(mi-1)mi}\}'$ where $Z_{ijj}' = r_{ij}r_{ij}'$, $r_{ij} = (Y_{ij} - \mu_{ij})/[\text{var}(Y_{ij})]^{1/2}$, and $\rho_{ijj}'(\alpha) = E[Z_{ijj}']$. A formal specification of the GEE is completed by defining estimating equations for the correlation parameters α :^[17]

$$\sum_{i=1}^K \mathbf{E}_i' \mathbf{W}_i^{-1} [\mathbf{Z}_i - \rho_i(\beta, \alpha)] = 0 \quad (6)$$

where $\mathbf{E}_i = \partial \rho_i / \partial \alpha$ and $\mathbf{W}_i$ is a working covariance matrix for $\mathbf{Z}_i$. The form of $\mathbf{E}_i$ in Eq. 6 is determined from the model for the pairwise correlation. Here the pairwise correlations are modeled as a linear function of pair-specific (typically, cluster-specific) covariates. The combination of Eqs. 5 and 6 gives consistent estimation of β even if the model for α is misspecified. A common working covariance matrix $\mathbf{W}_i$ is given by a diagonal matrix with elements

$$\text{var}(Z_{ijj}') = 1 + (1 - 2\mu_{ij})(1 - 2\mu_{ij}')\{\mu_{ij}(1 - \mu_{ij}')\mu_{ij}' \times (1 - \mu_{ij}')\}^{-1/2} \rho_{ijj}' - \rho_{ijj}'^2 \quad (7)$$

Use of weights $\mathbf{W}_i = \mathbf{I}_{mi(mi-1)/2}$ and a further refinement for the correction factor p gives the all-available-pairs correlation estimator.^[18] Empirical standard errors for α are used because $\mathbf{W}_i$ is a diagonal working correlation matrix.^[17]

One approach with GEE is to avoid estimation of α by using a working independence matrix and the empirical variance estimator for $\hat{\beta}$. However, modeling the correlation structure may increase statistical efficiency of the GEE regression parameter estimates.^[15] Another advantage with the second approach is that the correlation estimates may be used in statistical power calculations to plan future studies.^[3] Finally, estimation of α from a carefully specified model enables one to use the model-based variance estimator for $\hat{\beta}$. It may have performance superior to the robust variance estimator because the latter is an inefficient estimator of the true variance when cluster sizes are large.^[16]

Individual-Level Analysis with Mixed Effects Models

The subject-specific or mixed effects model approach extends generalized linear models to account for clustered data through specification of random cluster effects. A generalized linear mixed model (GLMM) is $g(\nu_{ij}) = \mathbf{X}_{ij}'\beta^* + \mathbf{Z}_{ij}'\mathbf{U}_i$ where ν_{ij} denotes $E(Y_{ij}|U_i)$. The GLMM for cluster trials^[19,20] assumes the following: (i) the conditional distribution of Y_{ij} given U_i , given by $f(y|U)$, follows an

exponential family generalized linear model; (ii) given U_i , the responses from a cluster Y_{ij} , $j = 1, \dots, m_i$, are independent; and (iii) the U_i are independent and identically normally distributed with density $f(U_i; \gamma)$ having covariance matrix G . The likelihood function is a function of the unknown parameters β^* and the parameter elements of γ of G , and the data $\mathbf{Y} = (\mathbf{Y}_1', \dots, \mathbf{Y}_K')'$:

$$L[\beta^*, \gamma, \mathbf{Y}] = \prod_{i=1}^K \prod_{j=1}^{m_i} f(Y_{ij} | U_i; \beta^*) f(U_i; \gamma) dU_i \quad (8)$$

Because the random effects, U_i , are not observed, they are integrated out to obtain the likelihood for the observable $\mathbf{Y}_i$, $i = 1, \dots, K$. This integration is difficult when $\mathbf{Y}_i$ is not normally distributed; in these cases, numerical approximation is used because the integral has no closed form. Linear mixed model analysis is well established for cluster unit trials with continuous responses.^[11,13]

The choice between population-averaged and subject-specific models should be based on the scientific question of interest. Use of a non-linear mixed effects model leads to regression parameters that have simple interpretations for covariates that vary within a cluster. However, for covariates that do not vary within clusters, the interpretation of such cluster-specific coefficients can be difficult or misleading because they measure a contrast in covariates when the random effects are controlled for, that is, are held equal. Intervention status of a community is a cluster-specific variable in cluster-unit trials because each community is studied under exactly one condition. The mixed logistic model gives an intervention effect whose meaning is conditional on the random effect and so the interpretation is restricted to the particular community, or, more broadly, to the conceptual population of communities with the same values of the unobserved random effect. The contrast of interest relies on the difference in the response under intervention and control conditions for a given cluster, clearly a non-observable event. For this reason, a population-averaged model has been recommended for the analysis of cluster-specific covariates.^[21,22]

There are other important differences in the analysis of cluster trials between the two general approaches. The regression parameters in the population-averaged model and the conditional model are different parameters with $\beta \neq \beta^*$, generally. In particular, the logistic-normal mixed model for binary data tends to give intervention parameter estimates that are larger than their population-averaged model counterparts. Specifically, for the random intercept model with $U_i \sim N(0, \sigma^2)$, $\beta \approx \beta^* / \sqrt{1 + 0.35\sigma^2}$.^[22] Finally, aside from the linear model and a few simple cases of the loglinear model for Poisson data, there are no simple relationships between variance components from generalized linear mixed models and marginal correlations defined for the distinct types of pairs of observations in a cluster.^[23]

For cluster trials with continuous response variables, mixed linear models are commonly used.^[13] If the design is balanced, exact tests are available with analysis of variance techniques. Generally, however, regression parameters in mixed linear models are estimated using maximum likelihood techniques. Regression parameters in these models, which are also called hierarchical linear models, have both population-averaged and conditional model interpretations, and unlike the binary data case, $\beta = \beta^*$. While GEE may be used for linear models, mixed model analysis is usually chosen because of its superior finite sample properties when the likelihood is correctly specified. Sandwich variance estimation may be used in mixed model analysis if robustness to the assumed covariance structure is desired.

DETERMINING SAMPLE SIZE

Taking the intracluster correlation into account in the planning phase of a study is critical because failure to do so may result in an underpowered study. The number of clusters in each condition (assuming $n_1 = n_2 \equiv n$) needed for testing the intervention with equal numbers of clusters per group using a two-sided test with α significance level and power $1 - \beta$ is

$$n = \frac{\phi(\sigma_1^2 + \sigma_2^2)(z_{1-\alpha/2} + z_{1-\beta})^2}{m\delta^2} \quad (9)$$

where m is the number of subjects per cluster, $\phi = 1 + (m - 1)\rho$, $\sigma_k^2 = \text{var}(Y_{ij} | \text{condition } k)$, δ is the difference to be detected, and z_c is the $(100 \times c)$ th percentile of the standard normal distribution. Furthermore, let $\Phi(\cdot)$ define its cumulative distribution function. Power $(1 - \beta)$ is given by

$$1 - \beta = \Phi\left(\frac{\delta}{\sqrt{\phi(\sigma_1^2 + \sigma_2^2) / nm}} - z_{1-\alpha/2}\right) \quad (10)$$

In the case of continuous outcomes, equal variance ($\sigma^2 \equiv \sigma_1^2 = \sigma_2^2$) is usually assumed and δ is the hypothesized difference in intervention and control population means.^[12] When n is small, use of the t -distribution instead of the normal distribution will improve these calculations. For binary data, Eq. 9 becomes

$$n = \frac{\phi\{P_1(1 - P_1) + P_2(1 - P_2)\}(z_{1-\alpha/2} + z_{1-\beta})^2}{m(P_1 - P_2)^2} \quad (11)$$

using the event rate notation of “Adjusted Chi-Square Test for Individual-Level Data.”

Often in primary care, the number of clusters (practices) is fixed. Based on Eq. 11, the cluster size m is derived:

$$m = \frac{(1 - \rho)\{P_1(1 - P_1) + P_2(1 - P_2)\}(z_{1-\alpha/2} + z_{1-\beta})^2}{n(P_1 - P_2)^2 - \rho\{P_1(1 - P_1) + P_2(1 - P_2)\}(z_{1-\alpha/2} + z_{1-\beta})^2}$$

Note that the denominator could be negative. That means n is too small, such that for the given α , one could not achieve the desired power $1 - \beta$ for any value of m .

Extensions of this formula for pretest-posttest cluster trial designs are available.^[13]

ANALYSIS OF CANCER SCREENING DATA

A representative sample of over 100 non-academic family physicians and general internists throughout Colorado was randomly assigned into four treatment conditions directed at physician and office practice behavior to increase screening for breast, cervical, prostate, and skin cancer. One physician was selected per primary care medical practice. The data considered here involve 1,209 patients from the 55 physicians that received either the control or full intervention condition; data on two partial intervention conditions are not included. Control refers to usual care, while intervention involved regional orientation and continuing contact by telephone and office visits at six-month intervals over a two-year period from 1992 to 1994. From each practice, medical charts from approximately 22 male and 22 female patients between the ages of 42 and 74 were selected at each of three time points, baseline (preintervention), and one and two years post intervention.

The analysis considers one outcome from the female patients, a binary indicator for clinical breast exam within the year previous to the year 2 follow-up study visit. Observed screening rates were $\hat{P}_1 = 266 / 594 = 44.8\%$ in the intervention group, $\hat{P}_2 = 252 / 615 = 41.0\%$ in the control group, and $\hat{P}_2 = 518 / 1209 = 42.8\%$ overall. A naive analysis of these data ignores possible intrapractice correlation of responses; the Pearson chi-square test statistic is 1.787 ($p = 0.181$). The conclusion from this incorrect analysis is that the intervention did not significantly increase screening for breast cancer with clinical breast exam. The incorrect analysis is liberal because it ignores the positive intrapractice correlation. Applying a standard two-sample t -test to the physician level screening rates in Table 1 comparing the unweighted mean screening rates of 44.78% for intervention practices and 40.99% for control practices gives $t = 0.70$ ($p = 0.488$). Alternatively, the ANOVA estimate of the intracluster correlation is $\rho = 0.125$, and the adjusted chi-square statistic is $\chi_A^2 = 0.494$ ($p = 0.482$). Thus relative to the results of these two tests, the naive analysis gives an inappropriately small p -value.

The GEE approach is to test the intervention with hypothesis test $H_0: \beta_1 = 0$ in the context of the logistic model:

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{1i} \quad (12)$$

where $\mu_{ij} = E(Y_{ij}) = P(Y_{ij} = 1)$ is the probability of clinical breast exam in the preceding year for the j th patient in the

Table 1 Number of Women Current for Clinical Breast Exam after Practice-Oriented Intervention among 22 Patients Enrolled by Each of 55 Colorado Primary Care Medical Practices Partners for Prevention Project in 1994

Control group (<i>n</i> = 28)	2	2	3	4	4	4	4	6	6	7	8	8	8	9	10	10	10	10 ^a	11	11	11	13	14	14	14	15	16
Intervention group (<i>n</i> = 27)	0	1	5	5	6	6	6	8	8	8	9	9	10	11	11	11	11	11	12	13	13	13	14	14	16	17	18

^aThis medical practice enrolled only 21 patients.**Table 2** Results from GEE and GLMM^a Analyses for the Proportion of Women Current for Clinical Breast Exam after Practice-Oriented Intervention among 22 Patients Enrolled by Each of 55 Colorado Primary Care Medical Practices in the Partners for Prevention Project, 1994

	GEE		GLMM	
	Estimate	Std. err.	Estimate	Std. Err.
Intercept	-0.3646	0.1539 ^b	-0.4203	0.1770
Intervention	0.1551	0.2184 ^b	0.1675	0.2520
ρ	0.1204	0.0272	—	—
σ^2	—	—	0.6548	0.1825

^aFit with SAS PROC NLMIXED.^bModel-based standard errors.

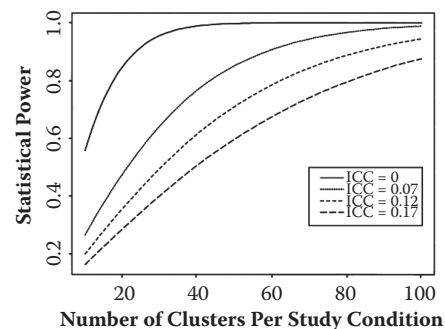
i th practice in the intervention ($x_{i1} = 1$) or control group ($x_{i1} = 0$) at posttest. The analogous GLMM is

$$\text{logit}(v_{ij}) = \beta_0^* + u_i + \beta_1^* x_{i1} \quad (13)$$

where $v_{ij} = E(Y_{ij}|u_i) = P(Y_{ij} = 1|u_i)$ is the conditional probability of clinical breast exam in the preceding year for the j th patient in the i th practice. Table 2 reports the results of each of these models. Although intervention practices had higher rates of clinical breast exam, results of both models show that the difference between intervention and control practices in screening rates was not statistically significant. From the GEE analysis, a randomly selected woman from an intervention practice had an estimated $\exp(0.155) = 1.168$ times higher odds to be current for clinical breast exam than a randomly selected woman from a control practice (95% confidence interval: 0.761–1.792). The interpretation from the GLMM is somewhat different: a randomly selected woman under the intervention condition had an estimated $\exp(0.168) = 1.182$ times higher odds to be current for clinical breast exam than a randomly selected woman from the same practice under the control condition (95% confidence interval: 0.722–1.938). This odds ratio represents a contrast that is not observed in the data because each practice is observed under exactly one condition. The presence of the practice random effect u_i in the model means that β_1^* is the log odds of cancer screening for the i th practice. The population average intercept as estimated from the GLMM is $-0.4203 / \sqrt{1 + 0.35\sigma^2} = -0.420 / 1.133 = -0.371$ which is close to the estimate from GEE ($\hat{\beta}_0 = -0.365$). The population average intervention effect as estimated from the GLMM is $-0.168 / 1.133 = 0.148$ which is reasonably close to the GEE estimate ($\hat{\beta}_1 = 0.1551$). Differences may

be a result of the bias inherent in the respective estimation methods, including the approximation involved in evaluating Eq. 8 in fitting the GLMM.

Suppose now that a future intervention is planned to increase clinical breast exam rates among female patients and sample size calculations are required. Assume the control group screening rate is $P_1 = 0.41$ and the follow-up visit intervention group screening rate is $P_2 = 0.51$. Power calculations are based on the GEE estimate $\hat{\rho} = 0.12$ and additional values that are the estimated upper and lower 95% confidence limits, 0.07 and 0.17. For example, using Eq. 10 with $\sigma_k^2 = P_k(1 - P_k)$, $\delta = 0.10$, and $\rho = 0.12$, $n = 62$ medical practices per study condition would provide 80% power for a two-sided 0.05 significance level test for a two-condition cluster trial with $m = 22$ patients per practice. Fig. 1 presents power as a function of n and ρ . A curve for $\rho = 0$ illustrates how failing to account for intraclass correlation can result in gross overestimation of study power. Lastly, an estimate of ICC from the

**Fig. 1** Power for binary outcome, $m = 22$ individuals per cluster

GLMM results, for a latent continuous outcome, is given by $\hat{\rho} = \hat{\sigma}^2 / \{[\hat{P}(1 - \hat{P})]^{-1} + \hat{\sigma}^2\} = 0.138$.^[13]

OTHER ISSUES IN CLUSTER TRIALS

Multiple Time Points

Both the GLMM and GEE approaches allow specification of general models for the mean structure. For cluster trials with multiple time points, a model may include terms for the intercept, time effects, main effects for conditions, and time by condition interactions. In the GLMM approach, specification of multiple random effects accounts for different sources of variation arising from multiple time-point designs.^[13] This provides a correlation structure that is richer than the simple exchangeable correlation. Implementation of GEE with commercially available software has been more restricted with respect to providing realistic correlation structures for multiple time-point cluster trials. GEE methods and sample size formulae specifying correlation structures based on certain study designs have been proposed.^[3] Use of GEE based on Eqs. 5 and 6 provides estimation in general correlation models.

Adjusting for Covariates

Analysis with mixed models or GEE can accommodate cluster-level or individual-level covariates. The multiple time-point nested cohort design is a special case that allows one to model the change score from baseline as a response variable. Modeling baseline response as a covariate may increase precision and power. More generally, including covariates in the model that are correlated with the response may increase statistical power as well as adjust for baseline imbalances. In non-randomized trials, it is important to adjust for covariates to account for imbalance in the treatment assignment.

Non-Randomized Trials

Randomization, a common feature in cluster trials, is not always possible. In large government- and foundation-sponsored cluster trials, the intervention communities are sometimes chosen by administrators of such funds.^[3] Study investigators may then only choose the comparison communities for the evaluation. A consequence of cluster non-randomized trials is that clusters selected to receive intervention may differ from those not selected with respect to baseline covariates. Those differences may lead to biased estimates of intervention effect. Traditional methods of adjustment, matching, stratification, and covariance analysis, may be employed. One large non-randomized study to reduce underage drinking used propensity scores to choose the control communities for evaluation.^[3]

Finite Sample Performance

Particularly in cases of unbalanced or categorical data, inference for cluster trials relies on test statistics with approximate distributions. Valid inference based on GEE or GLMM methods requires that K is sufficiently large. Likelihood ratio tests in GLMM analyses may be unreliable in small samples.^[12] Obtaining valid inference with GEE and using empirical sandwich covariance estimation require approximately 40 clusters,^[20,24] otherwise, the sandwich estimator tends to underestimate the true variance. This, in turn, tends to increase the type I error of the Wald test that uses the sandwich estimator. Mancl and DeRouen and Kauermann and Carroll have proposed alternative bias-corrected covariance estimators for a small sample of clusters.^[16,25] Simulation study results for cluster trials show that these bias-corrected covariance estimators perform better than uncorrected sandwich estimators in terms of bias and coverage probabilities.^[26] Additionally, introduction of bias-correction into the estimating equations for the intraclass correlations may considerably improve their point and interval estimation.^[27]

Cluster Trials in Pharmaceutical Studies

Due to inherent features of pharmaceutical product development processes, cluster trials are seldom used in pharmaceutical studies. Recently, some community based cluster trials were designed to evaluate interventions to improve the effectiveness and safety of drugs. Mullany and colleagues have reported a community-based cluster trial to assess the effects of 4.0% chlorhexidine on omphalitis and mortality risk among newborns in southern Nepal.^[28] Clarke and colleagues designed a double-blind cluster-randomized trial in 30 primary schools in western Kenya to evaluate the effect of intermittent preventive treatment in reducing anemia and improving classroom attention and educational achievement in semi-immune schoolchildren. Schools were randomly assigned to treatment (sulfadoxine-pyrimethamine in combination with amodiaquine or dual placebo), and the prevalence of was assessed through cross-sectional surveys 12 months post-intervention.^[29] Another cluster trial has been used to evaluate the effectiveness of educational outreach visits to improve adverse drug reaction (ADR) reporting by pharmacists in Portugal. Four spatial-clusters of pharmacists were assigned to the intervention group and eleven to the control group. The intervention took the form of hour-long educational outreach visits tailored to training needs detected in a previous study, with a 13- to 16-month follow-up period.^[30]

CONCLUSION

Cluster trials are used in many areas of health research, and they have a particularly strong tradition in behavioral

and disease prevention research. By contrast, there are fewer cluster pharmaceutical trials; this article has briefly described some that have been conducted in recent years. Although cluster pharmaceutical trials have been mostly used in community settings in developing countries, their use is likely to broaden.

The primary feature of the cluster trial is the intracluster correlation that arises from assignment of treatment conditions to intact groups of individuals. This means that special consideration is needed in the planning and conduct of such trials that is not required in a clinical trial that randomizes treatments to individuals. Because of such challenges, research in how to best plan and analyze data from cluster trials will continue. This is especially true for categorical outcomes and in those situations characterized by a small number of clusters, each with a large number of subjects per cluster.

ACKNOWLEDGMENTS

The author thanks Stuart Cohen for providing the PARTNERS data. The PARTNERS project was funded by research grant HS 06992 from the Agency for Healthcare Research and Quality, formerly the Agency for Health Care Policy and Research.

REFERENCES

- Gail, M.; Mark, S.D.; Carroll, R.J.; Green, S.B.; Pee, D. On design considerations and randomization-based inference for community intervention trials. *Stat. Med.* **1996**, *15*, 1069–1092.
- Wagenaar, A.C.; Murray, D.M.; Gehan, J.P.; Wolfson, M.; Forster, J.L.; Toomey, T.L.; Perry, C.L.; Jones-Webb, R. Communities mobilizing for change on alcohol: Outcomes from a randomized community trial. *J. Stud. Alcohol.* **2000**, *61* (1), 85–94.
- Preisser, J.S.; Young, M.L.; Zaccaro, D.J.; Wolfson, M. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Stat. Med.* **2003**, *22*, 1235–1254.
- Cohen, S.J.; Halvorson, H.W.; Gosselink, C.A. Changing physician behavior to improve disease prevention. *Prev. Med.* **1994**, *23*, 191–284.
- Parker, D.R.; Evangelou, T.E.; Eaton, C.B. Intraclass correlation coefficients for cluster randomized trials in primary care: The cholesterol education and research trial (CEART). *Contemp. Clin. Trials* **2005**, *26*, 260–267.
- Fretheim, A.; Oxman, A.D.; Treweek, S.; Bjorndal, A. Rational prescribing in primary care (RaPP-Trial). A randomised trial of a tailored intervention to improve prescribing of antihypertensive and cholesterol-lowering drugs in general practice [ISRCTN48751230]. *BMC Health Serv. Res.* **2003**, *3* (1), 5. <http://www.biomedcentral.com/1472-6963/3/5> (accessed October 2003).
- Wawer, M.J.; Sewankambo, N.K.; Serwadda, D.; Quinn, T.C.; Paxton, L.A.; Kiwanuka, N.; Wabwire-Mangen, F.; Li, C.; Lutalo, T.; Nalugoda, F.; Gaydos, C.A.; Moulton, L.H.; Meehan, M.O.; Ahmed, S.; the Rakai Project Study Group and Gray, R. H. Control of sexually transmitted diseases for AIDS prevention in Uganda: A randomised community trial. *Lancet* **1999**, *353*, 525–535.
- Feldblum, P.J.; Bwayo, J.J.; Kuyoh, M.; Welsh, M.; Ryan, K.; Chen-Mok, M. The female condom and STDs: Design of a community intervention trial. *Ann. Epidemiol.* **2000**, *10*, 339–346.
- Gavagni, A.S. M.; Hodjati, M.H.; Mohite, H.; Davies, C.R. Effect of insecticide-impregnated dog collars on incidence of zoonotic visceral leishmaniasis in Iranian children: A matched-cluster randomised trial. *Lancet* **2002**, *360*, 374–379.
- Lam, P.T.; Poon, B.T.; Wu, W.K.; Chi, S.C.; Lam, D.S. Randomized clinical trial of the efficacy and safety of tropicamide and phenylephrine in preoperative mydriasis for phacoemulsification. *Clin. Exp. Ophthalmol.* **2003**, *31* (1), 52–56.
- Feldman, H.A.; McKinlay, S.M. Cohort versus cross-sectional design in large field trials. *Stat. Med.* **1994**, *13*, 61–78.
- Donner, A.; Klar, N. *Design and Analysis of Cluster Randomization Trials in Health Research*; Arnold: London, 2000.
- Murray, D.M. *Design and Analysis of Group-Randomized Trials*; Oxford University Press: New York, 1998.
- Diggle, P.J.; Heagerty, P.; Liang, K.Y.; Zeger, S.L. *The Analysis of Longitudinal Data*, 2nd Ed.; Oxford University Press: Oxford, 2002.
- Liang, K.-Y.; Zeger, S.L. Longitudinal data analysis using generalized linear models. *Biometrika* **1986**, *73*, 13–22.
- Kauermann, G.; Carroll, R.J. A note on the efficiency of sandwich covariance matrix estimation. *J. Am. Stat. Assoc.* **2001**, *96*, 1387–1396.
- Prentice, R.L. Correlated binary regression with covariates specific to each binary observation. *Biometrics* **1988**, *44*, 1033–1048.
- Sharples, K.; Breslow, N. Regression analysis of correlated binary data: Some small sample results for the estimating equations approach. *J. Stat. Comput. Simul.* **1992**, *42*, 1–20.
- Sashegyi, A.I.; Brown, S.; Farrell, P.J. Application of a generalized random effects regression model for cluster-correlated longitudinal data to a school-based smoking prevention trial. *Am. J. Epidemiol.* **2000**, *152*, 1192–1200.
- Bellamy, S.L.; Gibberd, R.; Hancock, L.; Howley, P.; Kennedy, B.; Klar, N.; Lipsitz, S.; Ryan, L. Analysis of dichotomous outcome data for community intervention studies. *Stat. Methods Med. Res.* **2000**, *9*, 135–159.
- Neuhaus, J.M.; Kalbfleisch, J.D.; Hauck, W.W. A comparison of cluster-specific and population averaged approaches for analyzing correlated binary data. *Int. Stat. Rev.* **1991**, *59*, 25–36.
- Zeger, S.L.; Liang, K.-Y.; Albert, P.S. Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **1988**, *44*, 1049–1060.
- Young, M.L.; Preisser, J. S.; Qaqish, B. F.; Wolfson, M. A comparison of subject-specific and population averaged models for count data from cluster-unit intervention trials. *Stat. Methods Med. Res.* **2007**, *16*, 167–184.
- Feng, Z.; Diehr, P.; Peterson, A.; McLerran, D. Selected statistical issues in group randomized trials. *Annu. Rev. Public Health* **2001**, *22*, 167–187.

25. Mancl, L.A.; DeRouen, T. A. A covariance estimator for GEE with improved small-sample properties. *Biometrics* **2001**, *57*, 126–134.
26. Lu, B.; Preisser, J.S.; Qaqish, B.; Suchindran, C.; Bangdiwala, S.; Wolfson, M. A Comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics* **2007**, *63*, 935–941.
27. Preisser, J.S.; Lu, B.; Qaqish, B.F. Finite sample adjustments in estimating equations and covariance estimators for intraclass correlations. *Stat. Med.* **2008**, *27*, 5764–5785.
28. Mullany, L.C.; Darmstadt, G.L.; Khatry, S.K.; LeClerq, S.C.; Katz, J.; Tielsch, J.M. Impact of umbilical cord cleansing with 4.0% chlorhexidine on time to cord separation among newborns in southern Nepal: a cluster-randomized, community-based trial. *Pediatrics* **2006**, *118*, 1864–1871.
29. Clarke, S.E.; Jukes, M. C.H.; Njagi, J.K.; Khasakhala, L.; Cundill, B.; Otido, J.; Crudder, C.; Estambale, B.; Brooker, S. Effect of intermittent preventive treatment of malaria on health and education in schoolchildren: a cluster-randomized, double-blind, placebo-controlled trial. *Lancet* **2008**, *372*, 127–138.
30. Herdeiro, M.T.; Polonia, J.; Gestal-Otero, J. J.; Figueiras, A. Improving the reporting of adverse drug reactions: a cluster-randomized trial among pharmacists in Portugal. *Drug. Safety* **2008**, *31*, 335–344.

Clustered Study Designs: Power Analysis

J. Zhang

Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, U.S.A.

C. Feng, W. Tang, and X. M. Tu

Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York, U.S.A.

J. Kowalski

Division of Oncology Biostatistics, Johns Hopkins University, Baltimore, Maryland, U.S.A.

Abstract

Presented in this entry is a review of existing methods for power analysis with a focus on regression and correlation analysis. Also discussed is the latest development in addressing the random missing data pattern.

INTRODUCTION

Power analysis for sample size calculation is an important component in the design and planning of modern clinical trials. It provides information for assessing the feasibility of a study to detect some prespecified effect size and for estimating the amount of resources necessary for its execution in both efficacy and effectiveness research. Over the past two decades, studies in biomedical, pharmacologic, and psychosocial research have evolved from simple cross-sectional to more elaborate longitudinal study designs. As longitudinal designs use subjects as their own controls, such study designs afford a unique opportunity to examine changes of outcomes of interest over time, providing critically important information for research in these areas such as modeling disease progression, treatment response patterns, treatment tolerability, and toxic side effects. However, such study designs have also brought about several methodologic issues, the most prominent of which are data clustering and missing data.

Although the past two decades have witnessed major advances in the development of statistical methods for addressing data clustering, particularly for longitudinal studies, most research efforts have been centered around data analysis, with little attention to power and sample size estimation. As a result, the latter development has been progressing at a much slower pace. Often, methods developed based on cross-sectional study designs are used to provide power and sample size estimates for longitudinal trials. Only recently has attention been paid to the effects of within-subject correlation (or intracluster variability) and missing data in such study designs. However, despite the heightened attention and nascent efforts, some critical issues remain unresolved, particularly the random missing data pattern.

In this article, we review existing methods for power analysis with a focus on regression and correlation analysis. In addition, we discuss our latest development in addressing the random missing data pattern. For convenience of exposition, we have organized the presentation into two parts, with Part I devoting to regression and Part II to correlation analysis. We illustrate the methods discussed with several examples that are motivated by real trials in biomedical and psychosocial research.

PART I. POWER ANALYSIS FOR REGRESSION MODELS

Consider a study of n clusters, indexed by i , each of size m ($1 \leq i \leq n$). Let Y_{it} denote the response and X_{it} the vector of predictors (or covariates) of the t th member from the i th cluster.

In statistics, the notion of cluster is a generic term used to describe groups of subjects and/or data that are more similar within each group than across different groups. The existence of such clusters invalidates the usual independent sampling assumption, and as a result, statistical methods developed based on independence of observations such as in most cross-sectional studies are no longer applicable to such data. Longitudinal studies are the most popular examples involving data clustering, as repeated assessments of an individual over time give rise to clustered responses.^[1] Clustered data also often arise from cross-sectional study designs, as in epidemiologic and psychosocial research, where subjects sampled from a common habitat such as families, schools, and communities are more similar than those from different habitats.^[2,3] For convenience, we assume no additional data clustering for longitudinal

studies other than repeated assessments and single-level data clusters (i.e., no nesting) for cross-sectional studies. Such an assumption applies to most of the study designs in practice. As single-level data clusters in cross-sectional studies can be treated as a special case of data clusters of longitudinal study designs,^[4] we will focus on the latter study designs only. We first review methods for power analysis in the absence of missing data.

Complete Data Case

Consider a longitudinal study with n subjects and m prespecified assessment times $t = 1, \dots, m$.

Let Y_{it} denote the response and $X_{it} = (X_{it1}, \dots, X_{itp})^\top$, the vector of covariates or explanatory variables from the i th subject at t th assessment. In many studies, the X_{it} may consist of baseline covariates and deterministic functions of such variables and time. Two most popular approaches for longitudinal study data are the generalized estimating equations (GEE) and mixed-effects models (MM).^[1,5–21] For data analysis, the latter class of models has the advantage of delineating the between- from the within-cluster variability, while the former has the much desired distribution-free property for providing robust inference about the between-cluster variability. In many applications, power analysis is usually conducted for detecting between-cluster variability, or differences in the expected value of the response. Thus, the advantage of MM in obtaining intraclass variance estimates is no longer an important consideration. On the other hand, the distribution-free property of GEE becomes especially desirable for power analysis; without real data, it would be impossible to validate a parametric model, and such a robust property will ensure more reliable estimates. Of course, the final decision on which approach to take ultimately rests upon which approach will be used for data analysis; it makes little sense to compute power based on an MM, when the study data will eventually be analyzed by GEE.

Regardless of which approach to use, we consider regression models that link X_{it} to Y_{it} through a linear predictor $X_{it}^\top \beta$ or a function of such a predictor, where β is the vector of parameters of interest. Under GEE, such a regression for a continuous, binary or count (frequency) response is defined by the marginal mean and variance of the response in terms of a generalized linear model:^[11,21]

$$\begin{aligned} E(Y_{it} | X_{it}) &= h(X_{it}^\top \beta) = h_{it}, \\ \text{Var}(Y_{it} | X_{it}) &= \lambda^2 v(h_{it}), \quad 1 \leq i \leq n, \quad 1 \leq t \leq m \end{aligned} \quad (1)$$

where $h^{-1}(\cdot)$ is a known link function, λ^2 , a scale parameter, and $v(\cdot)$ is a known function.^[22] Under MM, the within-subject correlations of the joint distribution of the repeated responses are explicitly accounted for by introducing some latent variables (random effects). For a continuous response, the linear mixed-effects model (LMM) is defined by

$$\begin{aligned} Y_{it} &= X_{it}^\top \beta + Z_{it}^\top \mathbf{b}_i + \epsilon_{it} \text{ or } Y_i = X_i \beta + Z_i \mathbf{b}_i + \epsilon_i, \\ \mathbf{b}_i &\sim N(0, D), \quad \epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{im})^\top \sim N(0, \sigma^2 I_m), \\ 1 &\leq t \leq m \end{aligned} \quad (2)$$

where D denotes the variance of the random effect $\mathbf{b}_i$, σ^2 , the variance of the model error ϵ_i , I_m , the $m \times m$ identity matrix, and $X_i = (X_{i1}, \dots, X_{im})^\top$ and $Z_i = (Z_{i1}, \dots, Z_{im})^\top$ are the design matrices for the fixed and random effect, respectively. Note that in Eq. (2), we have assumed that ϵ_{it} has a constant variance σ^2 . If this variance changes over time, the discussion to follow still applies, with $\sigma^2 I_m$ replaced by a diagonal matrix with σ_t^2 of on the t th diagonal.

As in Ref. [4], we limit our discussion of mixed-effects models to LMM in this article. A primary reason for focusing on LMM rather than more general mixed-effects models such as the generalized linear mixed-effects models for discrete responses is the large number of different approaches proposed for inference and the complexity of the asymptotic distributions of model estimates.^[5–9,12,14,18–20] Also, for GEE, we limit our consideration to Eq. (1) for continuous, binary or count responses. The proposed approach is readily generalized to GEE models for categorical responses.^[13]

For power analysis, we consider the class of general linear hypothesis of the form

$$H_0 : K\beta = 0, \quad H_a : K\beta = a \neq 0 \quad (3)$$

where K is a known full rank $s \times p$ matrix and a some $s \times 1$ vector of known constants. The linear contrasts (3) apply to virtually all types of hypotheses concerning β in practical studies.^[4]

Generalized Estimating Equations

Under GEE, estimates of β are obtained as solutions to the following estimating equations

$$U(\beta) = \frac{1}{n} \sum_{i=1}^n U_i = \frac{1}{n} \sum_{i=1}^n D_i^\top V_i^{-1} S_i = 0 \quad (4)$$

where

$$\begin{aligned} A_i &= \text{diag}[v(h_{it})], \quad Y_i = (Y_{i1}, \dots, Y_{im})^\top, \\ h_i &= (h_{i1}, \dots, h_{im})^\top, \quad V_i = A_i^{-1} W(\alpha) A_i^{-1}, \\ D_i &= \frac{\partial}{\partial \beta^\top} h_i, \quad S_i = Y_i - h_i, \quad U_i = D_i^\top V_i^{-1} S_i \end{aligned} \quad (5)$$

In Eq. (5), $\text{diag}[v(h_{it})]$ denotes a diagonal matrix with $v(h_{it})$ on the t th diagonal and $W(\alpha)$ a working correlation matrix modeled by the parameter vector α . The working correlation matrix, $W(\alpha)$, is generally not equal to the true within-subject correlation of Y and its role is to increase efficiency.^[11,23] Note that for power analysis, λ^2 and α are all specified so that Eq. (4) is a function of β only.

A GEE estimate, $\hat{\beta}$ obtained as a solution to Eq. (4), is both consistent and asymptotically normal, i.e.,

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta) &\rightarrow_d N(0, \Sigma_\beta = B^{-1} \Sigma_U B^{-1}) \\ B &= E(D_1^\top V_1^{-1} D_1), \Sigma_U = E(D_1^\top V_1^{-1} D_1) \\ \Sigma_U &= E(D_1^\top V_1^{-1} S_1 S_1^\top V_1^{-1} D_1) \end{aligned} \quad (6)$$

where $\rightarrow_d$ denotes convergence in distribution.^[24,25] For data analysis, α , B , and Σ_U are estimated by the observed data. When their estimates are substituted into Eq. (6), we obtain a robust asymptotic variance estimate of Σ_β . For power analysis, there is typically no data to estimate these parameters. Thus, to evaluate Σ_U , we must model the true conditional variance $\text{var}(Y_i | X_i)$ in addition to $W(\alpha)$. Two extreme cases are worth noting: (1) $W(\alpha)$ is the true correlation matrix and (2) $W(\alpha)$ is set equal to the identity matrix (working independence assumption). Although achieving maximum power, it is unrealistic to expect that a selected $W(\alpha)$ is the true correlation matrix. Thus, the identity matrix choice for $W(\alpha)$ may be preferable in real study designs to provide more realistic and useful power estimates. The asymptotic variance for each of the two case scenarios is given by

$$\Sigma_\beta = \begin{cases} E^{-1} [D_1^\top \text{Var}(Y_i | X_i) D_1] & \text{if } W(\alpha) \text{ is true correlation} \\ B^{-1} E^{-1} [D_1^\top A_1^{-1} \text{Var}(Y_i | X_i) A_1^{-1} D_1] B^{-1} & \text{if } W(\alpha) \text{ is set to identity matrix} \end{cases}$$

where E^{-1} denotes the inverse of the expectation and $B = E(D_1^\top A_1^{-1} D_1)$.

Under H_0 and H_a in Eq. (3), the quadratic

$$Q_n^2 = n \hat{\beta}^\top K^\top (K \Sigma_\beta K^\top)^{-1} K \hat{\beta} \quad (7)$$

has an asymptotic central and non-central χ_s^2 distribution, respectively,

$$H_0 = Q_n^2 \sim \chi_s^2(0), \quad H_a = Q_n^2 \sim \chi_s^2(c) \quad (8)$$

where $c = n a^\top (K \Sigma_\beta K^\top)^{-1} a$, a is the non-centrality parameter. Let $F_{\chi_s^2(c)}^2$ denote the cdf of $\chi_s^2(c)$. The power function, φ , for the linear contrast (3) for an α level of type I error is given by

$$\begin{aligned} \varphi(n, c, \alpha) &= 1 - F_{\chi_s^2(c)}^2(p1 - \alpha), \\ \alpha &= 1 - F_{\chi_s^2(0)}^2(p1 - \alpha) \end{aligned} \quad (9)$$

where p_γ denotes the γ th percentile of $\chi_s^2(0)$.

For a given sample size n , φ above provides the power for detecting H_a in Eq. (3). Alternatively, Eq. (9) can also be used to estimate the required sample size to achieve a prespecified power.^[4] In this case, φ is given, and the

minimum sample size required to reach φ is the solution to the first equation in (9). This non-linear equation is numerically solved by root-finding methods such as the Newton's algorithm.^[26] Most commercially available software packages implement one or more such root-finding methods. For example, in SAS, the function NLPFDD computes the root of a general non-linear function using Newton's method, where the required first-order derivatives are either analytically calculated or numerically approximated, depending on the complexity of the non-linear function. Sample size estimates are obtained by rounding off the solutions using the largest integer function. Later in this article, we will give some similar power functions in various situations, and the required sample sizes to achieve prespecified powers can be similarly computed.

Linear Mixed-Effects Model

Let $V_i = Z_i D Z_i^\top + \sigma^2 I_m$. For power analysis, D and σ^2 are known, and the maximum likelihood estimate (MLE) of β is given by Laird and Ware^[10] and Kowalski et al.^[27]

$$\hat{\beta} = \left(\sum_{i=1}^n X_i^\top V_i^{-1} X_i \right)^{-1} \left(\sum_{i=1}^n X_i^\top V_i^{-1} Y_i \right) \quad (10)$$

In addition, $\hat{\beta}$ has the limiting distribution

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, \Sigma_\beta = E^{-1} [X_1^\top (Z_1 D Z_1^\top + \sigma^2 I_m)^{-1} X_1]) \quad (11)$$

The Wald statistic Q_n^2 defined in Eq. (7) again has an asymptotic χ_s^2 distribution, except for a redefined Σ_β given by Eq. (11). The power function is again given by Eq. (9). Evaluation of Σ_β has been discussed in Ref. [4].

Missing Data Case

In almost all longitudinal clinical trials, missing data are inevitable, no matter how well a trial is planned and executed. One common approach to account for the effect of missing data is simply inflate sample size calculated based on complete data so that the adjusted sample size has a built-in protection against the loss of power owing to missing data. For example, if we anticipate that 10% of the subjects will drop out, we can use a sample size so that the trimmed 90% of the subjects will be sufficient to achieve a prespecified power. This approach, however, assumes that only subjects with complete data will be used for analysis, which is too conservative since GEE and LMM can use all available data. In addition, it does not account for the effect of missing data patterns. Consider for example the missing data patterns shown in Fig. 1 from a hypothetical study with three assessments on each of six subjects. Although the same proportion of total measurements (2/3) is missing

at each assessment point across the patterns, the power to detect between-assessment differences could differ greatly between the patterns. For example, the unbalanced case in (b) generally yields more power than (a), because the observations in (b) are obtained from distinct rather than the same clusters (individuals) as in (a). Although patterns (b) and (c) involve the same number of clusters, they will also yield different power functions if we are interested in testing differences between the means at different assessment points. For the three patterns in Fig. 1, the trimmed approach to inflate sample size is only appropriate for pattern (a).

Unlike data analysis where inference is typically conditioned on observed data, missing data patterns are completely unknown when planning a real study. The three scenarios in Fig. 1 are among the many possible patterns for the hypothetical study. For power analysis, we must consider these plus many more as potential missing data patterns. Currently, there is no general approach for addressing the random missing data patterns. Existing methods provide power estimates either conditional on a particular missing data pattern or based on certain marginal distributions of missing data patterns.^[28–30] As different missing data patterns generally give rise to different power functions, none of these approaches sufficiently addresses the dynamic nature of missing data patterns. For example, one such strategy is to condition on the available sample size at each assessment point. Clearly, such an approach does not distinguish the first two patterns in Fig. 1. Although methods developed for survey research and epidemiologic studies do account for differences in the number of clusters as well as cluster size,^[3,31] they do not address the time-ordered structure in longitudinal data. For example, patterns (b) and (c) are the same in terms of number of clusters and cluster size in each cluster. But, as noted earlier, the two patterns generally give rise to different power functions.

Recently, we developed a new approach to address the random missing data pattern. This approach explicitly models the random missing data pattern using random

indicators and integrates the missing data model into the power function. We outline this approach below. More details can be found in Ref. [32].

Define a vector of binary variables for indicating missing (or rather observed) data for each subject as follows.

$$r_{it} = \begin{cases} 1 & \text{if } Y_{it} \text{ is observed} \\ 0 & \text{otherwise, } \mathbf{r}_i = (r_{i1}, \dots, r_{im})^\top, \\ & R_i = \text{diag}(r_{it}), \quad 1 \leq i \leq n \end{cases} \quad (12)$$

For convenience, denote the baseline by $t = 1$ and assume no missing data at this time point, i.e., $r_{i1} \equiv 1$ ($1 \leq i \leq n$). Each configuration of $\mathbf{r}_i$ determines a missing data pattern and vice versa. In accordance with the literature, we assume that missing data only occurs in the response, not in the covariates. For a real study, $\mathbf{r}_i$ are observed once the study is completed. Data analysis proceeds by conditioning on the observed $\mathbf{r}_i$. For power analysis, $\mathbf{r}_i$ are unknown, and we must consider all possible values of this random vector in constructing the power function. Throughout the rest of the discussion, we assume missing completely at random (MCAR), i.e., given X_i , $\mathbf{r}_i$, and Y_i are independent.^[11,33–35] We discuss extensions to more general missing data mechanisms in section “Discussion.”

Generalized Estimating Equations

Let $\tilde{S}_i$ denote a subvector of S_i containing those S_{it} for which $r_{it} = 1$ ($1 \leq t \leq m$). Similarly, let $\tilde{D}_i$ and $\tilde{V}_i$ be the submatrices of D_i and V_i corresponding to the t s with $r_{it} = 1$ ($1 \leq t \leq m$). In data analysis, Y_{it} with $r_{it} = 1$ represent observed data, and inference about β under GEE is based on the observed-data estimating equations:

$$U(\beta) = \frac{1}{n} \sum_{i=1}^n \tilde{D}_i^\top \tilde{V}_i^{-1} \tilde{S}_i = 0 \quad (13)$$

Under MCAR, Eq. (13) defines a consistent and asymptotically normal estimate of β .^[11,35,36]

For power analysis, $\mathbf{r}_i$ are not observed, which makes Eq. (13) difficult to work with in terms of expressing power as a function of $\mathbf{r}_i$. To illustrate the alternative approach that will be used for the development of power analysis, consider the working independence assumption, i.e., $V_i = A_i$. In this special case, it is readily checked that Eq. (13) is equivalent to

$$U(\beta) = \frac{1}{n} \sum_{i=1}^n D_i^\top R_i V_i^{-1} R_i S_i = 0 \quad (14)$$

The asymptotic distribution of the estimate, $\hat{\beta}$, of the above estimating equations can be expressed as

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, \Sigma_\beta = B^{-1} \Sigma_U B^{-1}) \quad (15)$$

Subject	Pattern a Assessments			Pattern b Assessments			Pattern c Assessments		
	1	2	3	1	2	3	1	2	3
1	y_{11}	y_{12}	y_{13}	y_{11}	.	.	y_{11}	.	.
2	y_{21}	y_{22}	y_{23}	y_{21}	.	.	y_{21}	.	.
3	.	.	.	.	y_{32}	.	y_{31}	.	.
4	.	.	.	.	y_{42}	.	y_{41}	.	.
5	.	.	.	.	.	y_{53}	y_{51}	.	.
6	.	.	.	.	.	y_{63}	y_{61}	.	.

Fig. 1 Three missing-data patterns for a longitudinal study with six subjects and three assessment points (y_{it} denotes the response of the i th subject at time t with dots denoting missing data)

where B and Σ_U are given by

$$B = E(D_1^T R_1 V_1^{-1} D_1) \\ \Sigma_U = E(D_1^T V_1^{-1} R_1 \text{Var}(Y_1 | X_1) R_1 V_1^{-1} D_1) \quad (16)$$

Unlike Eq. (13), the estimating equations (14) explicitly involve r_{it} so that the asymptotic variance of $\hat{\beta}$ can be expressed as a function of $\mathbf{r}_i$. Given a model for the true variance $\text{Var}(S_i | X_i)$, B and Σ_U in Eq. (16) are readily evaluated. Note that B and Σ_U depend on R_i only through the first two moments of $\mathbf{r}_i$.

For non-diagonal choices of V_i , Eq. (14) is generally different from Eq. (13). Although estimates of β obtained from Eq. (14) are still consistent, they are different from the ones produced by Eq. (13), which are usually used in data analysis. Thus, we focus on the choice of V_i based on the working independence assumption in the discussion to follow. As noted earlier in Section “Complete Data Case,” power derived based on this special case is generally conservative.

The construction of the power function for testing the hypothesis (3) proceeds in the same way as in the complete data case. In particular, the quadratic Q_n^2 is again given by Eq. (7), but with Σ_β redefined by Eqs. (15) and (16). In addition, the form of the power function given in Eq. (9) remains unchanged.

As an example to illustrate the evaluation of B and Σ_U defined in Eq. (16), consider Eq. (1) with the identity link, i.e.,

$$E(Y_{it} | X_{it}) = X_{it}^T \beta, \quad \text{Var}(Y_{it}) = \sigma_i^2, \quad 1 \leq t \leq m, \\ 1 \leq i \leq n \quad (17)$$

The above linear model is quite popular in growth curve analysis. It is readily checked that B and Σ_U in Eq. (16) are given by

$$B = E(X_1^T R_1 A_1^{-1} X_1), \\ \Sigma_U = E[X_1^T A_1^{-1} R_1 \text{Var}(Y_1) R_1 A_1^{-1} X_1] \quad (18)$$

In the complete data case, $R_i = I_m$, and B and Σ_U can be expressed in closed form as a function of the first two moments of X_i .^[4] With missing data, Eq. (18) involves the random matrix R_i . In addition, R_i may depend on X_i , which further complicates the evaluation of these quantities. However, when R_i is independent of X_i , Eq. (18) reduces to

$$B = E(X_1^T E(R_1) A_1^{-1} X_1) \\ \Sigma_U = E\{X_1^T A_1^{-1} E[R_1 \text{Var}(Y_1) R_1] A_1^{-1} X_1\} \quad (19)$$

Both matrices above have a common form $E(X_1^T P X_1)$, where

$$P = E(R_1) A_1^{-1} \quad \text{or} \quad P = A_1^{-1} E[R_1 \text{Var}(Y_1) R_1] A_1^{-1}$$

depending on whether to evaluate B or Σ_U . In either case, P is a constant matrix and is readily expressed in closed form as a function of $\text{Var}(Y_1)$, $E(\mathbf{r}_1)$ and $E(\mathbf{r}_1 \mathbf{r}_1^T)$. Given P , $E(X_1^T P X_1)$ is readily evaluated

$$E(X_1^T P X_1) = \left[g_{lr} = \sum_{s=1}^m \sum_{t=1}^m p_{st} E(X_{1s} X_{1t}) \right] \quad (20)$$

where g_{lr} denotes the lr th element of $E(X_1^T P X_1)$ and p_{st} , the st th element of P . Thus, B and Σ_U are functions of the first two moments of the missing data indicator $\mathbf{r}_1$ as well as the first two moments of X_1 .

The more complex situation with R_i depending on X_i can be similarly discussed.^[32] However, B and Σ_U may not be in closed form, and their evaluations may be facilitated by Monte Carlo approximation.^[4,32]

Linear Mixed-Effects Model

Given D and σ^2 , we can express the MLE of β for the linear mixed-effects model defined in Eq. (2) in terms of the observed data, i.e., Y_{it} with $r_{it} = 1$.^[10,37] However, as in the case of GEE, it is difficult to work with such an expression and its associated asymptotic distributions for power analysis because it does not explicitly involve $\mathbf{r}_i$. Our strategy is to apply the approach developed for GEE to the marginalized linear mixed-effects model, which can be treated as a GEE model with a more structured variance.

To this end, let A_i , V_i , D_i , and S_i be defined in Eq. (5) with $h_i = X_i \beta$, $v(h_i) = Z_i^T D Z_i + \sigma^2$, and $W(\alpha) = I_m$. Consider the estimating equations (14) with these redefined components. The asymptotic distribution of the estimate from Eq. (14) is given by Eq. (15) with

$$B = E(X_1^T R_1 A_1^{-1} X_1) \\ \Sigma_U = E[X_1^T A_1^{-1} R_1 (Z_i D Z_i^T + \sigma^2) R_1 A_1^{-1} X_1] \quad (21)$$

Thus, evaluation of B and Σ_U is similar to Eq. (18) for the linear regression model under GEE (17) except that Eq. (21) has a more structured working variance A_i and true variance $Z_i D Z_i^T + \sigma^2$.

Note that unlike Eq. (17), the true variance now depends on Z_i . In most growth-curve analyses, Z_i is a function of time only and is also functionally independent of X_i . In this case, Eq. (21) can be evaluated in closed form as for the linear GEE model just discussed. When Z_i involves baseline covariates, the distribution of X_i must be known to evaluate Eq. (21). Evaluation of Eq. (21) in this more general case may be facilitated by Monte Carlo approximation.^[4,32]

For the linear contrast (3), the quadratic Q_n^2 defined in Eq. (7) again has an asymptotic χ_s^2 distribution, except for the redefined B and Σ_U given in Eq. (21). In addition, the power function remains the same general form as well.

Modeling Missing Data Indicators with Markov Processes

As seen in the preceding discussion, the asymptotic variance, Σ_{β} , is a function of moments of the missing data indicator $\mathbf{r}_i$. To evaluate these moments, it is convenient to model $\mathbf{r}_i$ using a two-state Markov process because r_{it} is a binary variable.^[38,39]

For $1 \leq s \leq t \leq m$, let

$$\mathbf{P}_i(s, t) = \begin{pmatrix} p_{i11}(s, t) & \bar{p}_{i11}(s, t) \\ p_{i01}(s, t) & \bar{p}_{i01}(s, t) \end{pmatrix},$$

$$p_{ik1}(s, t) = \Pr(r_{it} = 1 | r_{is} = k),$$

$$s < t, k = 0, 1 \quad (22)$$

where $p_{ik1}(s, t)$ is the transition probability from state k at time s to state 1 at time t of r_{it} and $\bar{p}_{ik1} = 1 - p_{ik1}$. We may model the transition matrix $\mathbf{P}_i(s, t)$ differently for different subgroups or even as a function of individual covariates. For convenience, we consider the homogeneous case in which $\mathbf{P}_i(s, t) = \mathbf{P}(s, t)$. In addition, we assume no missing data at baseline $t = 1$ so that $\Pr(r_{i1} = 1) = 1$. Using Eq. (22), we can easily express the moments of $\mathbf{r}_i$ in terms of the transition probabilities

$$E(r_{i1} r_{i2} \dots r_{iK}) = \prod_{k=1}^K p_{11}(t_{(k-1)}, t_k), \quad (23)$$

$$1 \leq t_1 < \dots < t_K \leq m, \quad 1 \leq K \leq m$$

where $t_0 \equiv 1$ and $p_{11}(t_{(-1)}, 1) \equiv \Pr(r_{i1} = 1)$ so that the above also yields a valid expression for $k = 1$.

The Markov model in Eq. (22) involves many parameters and may be difficult to use in the absence of real data as in most applications. One approach to this problem is the one-step Markov model defined by

$$\mathbf{P}(s, t) = \begin{pmatrix} p_{11}(s, t) & \bar{p}_{11}(s, t) \\ p_{01}(s, t) & \bar{p}_{01}(s, t) \end{pmatrix} = \prod_{u=s}^{t-1} \Phi(u),$$

$$\Phi(u) = \begin{pmatrix} \phi_{11}(u) & \bar{\phi}_{11}(u) \\ \phi_{01}(u) & \bar{\phi}_{01}(u) \end{pmatrix}, \quad (24)$$

$$\phi_{k1}(u) = p_{k1}(u, u+1), \quad \phi_{01}(1) = p_{01}(1, 2) = 0$$

$$1 \leq u < t \leq m, k = 0, 1$$

where $\bar{\phi}_{k1} = 1 - \phi_{k1}$. Under Eq. (24), $\mathbf{P}(s, t)$ is determined by the one-step transition probability matrix $\Phi(t)$, with $\phi_{k1}(t)$ describing the one-step transition probabilities from state k at time t to state 1 at $t + 1$, as depicted in Fig. 2. The constraint $\phi_{01}(1) = 0$ reflects the assumption of no missing data at baseline $t = 1$. These $\Phi(t)$ are easy to interpret and readily specified to account for different anticipated missing data processes. For example, by setting $\phi_{01}(t) = \phi_{11}(t)$ ($2 \leq t \leq m$), we model a process in which the occurrence

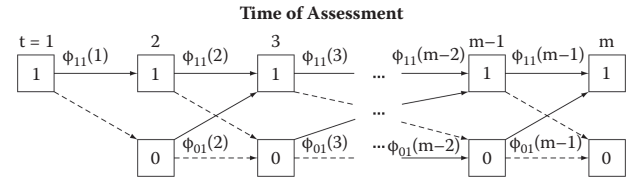


Fig. 2 One-step Markov model for missing data process with solid line-arrows indicating one-step transition probabilities $\phi_{k1}(t)$ from state k at time $t - 1$ to state 1 at time t of the missing data indicator r_{it} ($k = 0$ for a missing and $k = 1$ for a non-missing response) and dashed line-arrows indicating transitions (with probability $1 - \phi_{k1}(t)$) from state k at time $t - 1$ to state 0 at time t of the missing data indicator

of missing data at time t is independent of the missing data status at the earlier time $t - 1$. By allowing $\phi_{01}(t)$ to differ from $\phi_{11}(t)$, we can accommodate processes where missing data is either more or less likely to occur if it occurs at an earlier time. For example, by specifying $\phi_{01}(t) = 0.5\phi_{11}(t)$, we model a process in which missing data is twice more likely to occur if it occurs at the previous time.

In addition to directly specifying $\Phi(t)$, we may also determine these one-step transition probability matrices so that they yield the prespecified marginal means of $\mathbf{r}_i$. This latter approach has a more transparent interpretation. For example, if we anticipate that an expected q_t percent of data will be missing at time t , i.e., $E(r_{it}) = 1 - q_t$ ($2 \leq t \leq m$), we can solve for $\Phi(t)$ as a function of q_t using the following recursive expression

$$\mathbf{P}(1, t) = \prod_{u=1}^{t-1} \Phi(u) = \mathbf{P}(1, t-1) \Phi(t-1), \quad (25)$$

$$\mathbf{P}(1, 1) = I_2, \quad \phi_{01}(1) = 0, \quad 2 \leq t \leq m$$

where I_2 denotes the 2×2 identity matrix. The above also simultaneously yields $\mathbf{P}(1, t)$ ($2 \leq t \leq m$), which can be used to express the moments of $\mathbf{r}_i$ according to Eq. (23). Note that solving $\Phi(t)$ also requires specifying a relationship between $\phi_{01}(t)$ and $\phi_{11}(t)$ such as $\phi_{01}(t) = \phi_{11}(t)$, as discussed earlier (see Example 1.^[32]).

Note that if the one-step transition matrix is a constant, $\Phi(t) = \Phi$, we obtain a time-stationary Markov model.^[38,39] If in addition, $\phi_{11} = \phi_{01}$, then $\mathbf{P}(s, t) = \Phi$ and $E(r_{it}) = \phi_{11}$. In this special case, missing data occurs at a uniform rate $q = 1 - \phi_{11}$ over time. In addition, the moments in Eq. (23) admit a simpler expression

$$E(r_{i1} r_{i2} \dots r_{iK}) = \phi_{11}^K,$$

$$2 \leq t_1 < \dots < t_K \leq m, \quad 1 \leq K \leq m \quad (26)$$

As before, the case with $\phi_{11} > \phi_{01}$ ($\phi_{11} < \phi_{01}$) would imply that missing data is more (less) likely to occur if it occurs at an earlier time.

PART II. POWER ANALYSIS FOR CORRELATION MODELS

Complete Data Case

Consider a clustered study with n clusters, each of size m . As before, we assume independence across clusters and a common cluster size. Let $\{(X_{it}, Y_{it}); 1 \leq i \leq n; 1 \leq t \leq m\}$ denote a set of paired variables from the i th cluster. We assume that clustered data arise from longitudinal study designs or multivariate responses in cross-sectional studies, with clusters formed by individual subjects. We are interested in examining the changes in correlations $\rho_t = \rho_{xyt} = \text{Corr}(X_{it}, Y_{it})$ between X_{it} and Y_{it} over t .

This general setup applies to both longitudinal and cross-sectional study designs. For example, if n denotes the number of subjects and m the number of assessments in a longitudinal study, then ρ_t represent time-dependent correlations between X_{it} and Y_{it} . As another example for a clustered cross-sectional study design, consider the problem of assessing correlations among three response variables, U , V , and W . Suppose that we are interested in testing the null hypothesis: $H_0: \rho_1 = \rho_2$, where $\rho_1 = \text{Corr}(U, W)$ and $\rho_2 = \text{Corr}(V, W)$. To couch this inference problem within the general setup, simply let

$$m = 2, \quad X_{i1} = U_i, \quad X_{i2} = V_i, \quad Y_{i1} = Y_{i2} = W_i$$

For convenience, we refer to t as time and m as total number of assessments throughout the rest of the discussion.

Let $\rho = (\rho_1, \rho_2, \dots, \rho_m)^\top$. Consider again linear contrasts of the form

$$H_0: K_p \rho = 0 \quad \text{vs.} \quad H_a: K_p \rho = \mathbf{a} \neq 0 \quad (27)$$

where K is some $g \times m$ full rank ($g \leq m$) matrix of known constants and $\mathbf{a}$, a $m \times 1$ vector of known constants. The Pearson correlation is an estimate of ρ , and its asymptotic distribution has been used extensively for inference about ρ , particularly in psychiatric and psychosocial research.^[40–47] Let $\hat{\rho} = (\hat{\rho}_1, \hat{\rho}_2, \dots, \hat{\rho}_m)^\top$, with $\hat{\rho}_t$ denoting the Pearson estimate of ρ_t ($1 \leq t \leq m$). Then, under a multivariate normal assumption for $Z = (Z_1^\top, \dots, Z_m^\top)^\top$, with $Z_t = (X_t, Y_t)^\top$, the asymptotic normal distribution of $\mathbf{r}$ is given by

$$\sqrt{n}(\hat{\rho} - \rho) \rightarrow_d N(0, \Sigma_\rho = [\sigma_{st}])$$

$$\sigma_{st} = \begin{cases} (1 - \rho_s^2)^2 & \text{if } s = t; \\ \frac{1}{2} \rho_s \rho_t (\rho_{xst}^2 + \rho_{yst}^2 + \rho_{xyts}^2 + \rho_{xyts}^2) & \text{if } s \neq t \\ + (\rho_{xst} \rho_{yst} + \rho_{xyts} \rho_{xyts}) - & \\ - \rho_s \rho_{xst} \rho_{xyts} + \rho_{yst} \rho_{xyts} - \rho_t (\rho_{xst} \rho_{xyts} + \rho_{yst} \rho_{xyts}) & \end{cases} \quad (28)$$

where $\rightarrow_d$ denotes convergence in distribution and

$$\begin{aligned} \rho_{xyst} &= \text{Corr}(X_{is}, Y_{it}), \quad \rho_{xst} = \text{Corr}(X_{is}, X_{it}), \\ \rho_{yst} &= \text{Corr}(Y_{is}, Y_{it}), \quad 1 \leq s, t \leq m \end{aligned} \quad (29)$$

To use this asymptotic distribution, we must estimate the parameters in the asymptotic variance Σ_ρ . Let $P_g = [\rho_{gst}]$ ($g = X, Y$) denote the $m \times m$ matrix defined by the within-variable correlations (WVC), ρ_{gst} , and $P_{xy} = [\rho_{xyts}]$, the $m \times m$ matrix defined by the cross-variable correlations (CVC), ρ_{xyts} . Then, Σ_ρ is a function of the elements of these matrices. In addition, ρ corresponds to the diagonal of P_{xy} . With real data as in data analysis, consistent estimates of P_g and P_{xy} are readily computed. When no data is available, such estimates cannot be calculated, and other alternative estimates such as guesses from other similar studies must be considered. Thus, for power analysis, it is desirable to reduce the nuisance parameters as much as possible.

Note that Fisher's z transformation is often applied to improve the accuracy of the asymptotic distribution.^[48–51]

Within our context, this transformation, $T_F \hat{\rho} = \frac{1}{2} \ln \frac{1 + \hat{\rho}}{1 - \hat{\rho}}$,

is applied to each component of $\hat{\rho}$, i.e., $T_F(\hat{\rho})$ is defined as a vector

$$T_F(\hat{\rho}) = \left(\frac{1}{2} \ln \frac{1 + \hat{\rho}_1}{1 - \hat{\rho}_1}, \dots, \frac{1}{2} \ln \frac{1 + \hat{\rho}_m}{1 - \hat{\rho}_m} \right)^\top \quad (30)$$

By applying the Delta method,^[52] we obtain the asymptotic distribution of $T_F(\mathbf{r})$

$$\begin{aligned} \sqrt{n} [T_F(\hat{\rho}) - T_F(\rho)] &\rightarrow_d \\ N(0, \Sigma_{T_F} = D_\rho \Sigma_\rho D_\rho^\top), \quad D_\rho &= \frac{\partial}{\partial \xi} |_{\xi = \rho} T_F(\xi) \\ &= \text{diag}((1 - \rho_t^2)^{-1}) \end{aligned} \quad (31)$$

where $\text{diag}(a_t)$ denotes the diagonal matrix with a_t on the t th diagonal. For notational brevity, the asymptotic distribution (28) will be used in the following discussion.

In most applications, elements of WVC are typically modeled as a function of $|t - s|$, and a wide class of analytic functions is available for modeling $\rho_{gst} = \rho_g(|t - s|)$.^[53] For example, under the uniform compound symmetry assumption,^[1] $\rho_{gst} = \rho_g$ for all $s \neq t$ with $-1 < \rho_g < 1$. Such an assumption may oversimplify the correlation structure, but other alternatives are available to reach a compromise between practical utility and analytic tractability, especially when partial information is available about the variables. Modeling such correlation structure has been extensively discussed in the literature and is not repeated here.

For elements of the CVC P_{xy} , the problem unfortunately is much more complicated, and in particular, is not readily addressed by existing methods such as the approach

used for modeling the WVC. For example, consider a simple case with $m = 2$. If P_{xy} is modeled using a uniform compound symmetry assumption, then the null hypothesis $H_0: \rho_1 = \rho_0, \rho_2 = 1/2\rho_0$ would be contradictory to the assumed model, while $H_0: \rho_1 = \rho_2 = \rho_0$ is implied by such a model ($-1 < \rho_0 < 1$). On the other hand, the unstructured model with no constraint on the elements of P_{xy} would involve too many parameters to be practically useful in most applications. Thus, to use the asymptotic distribution for power analysis, it is desirable to develop an approach that not only reduces the number of parameters in P_{xy} , but also ensures this internal consistency with respect to the hypotheses for ρ .

Surrogacy Assumption

Consider the assumption

$$[Y_t|X_t, X_s] = [Y_t|X_t], \quad 1 \leq s \neq t \leq m \quad (32)$$

where $[Y|X]$ denotes the conditional distribution of Y given X . Although similar in appearance, the above is not a Markov assumption as often seen in stochastic process modeling,^[38] because the conditional distribution involves different rather than the same variable at different times. This condition is known as a surrogacy assumption in measurement error, or errors in variable, models literature.^[28,36,54] Within the current context, Eq. (32) stipulates that X_s contains no additional information that could be predictive of Y_t once X_t is known, i.e., a surrogate of X_t in predicting Y_t .

Under Eq. (32), we can express the elements of CVC P_{xy} in terms of ρ_t and the elements of the WVC P_{xy} by assuming a linear regression model. Let

$$Y_{it} = \beta_{0t} + \beta_{1t}X_{it} + \epsilon_{it}, \quad \epsilon_{it} \sim i.i.d.(0, \sigma_{\epsilon t}^2), \quad 1 \leq i \leq n, 1 \leq t \leq m \quad (33)$$

where (μ, σ^2) denotes a distribution with mean μ and variance σ^2 . Let σ_{gt} denote the standard deviation of g_t ($g = X, Y$). Then, since $\rho_t = \beta_{1t} \sigma_{xt} \sigma_{yt}^{-1}$, it follows from Eq. (32) and the law of iterated expectations that^[52]

$$\begin{aligned} \rho_{xyst} &= \sigma_{xs}^{-1} \sigma_{yt}^{-1} E[(X_s - \mu_{xs})(Y_t - \mu_{yt})] \\ &= \beta_{1t} \sigma_{xt} \sigma_{yt}^{-1} \rho_{xst} = \rho_t \rho_{xst} \end{aligned} \quad (34)$$

The above computations are diagrammatically illustrated in Fig. 3; ρ_{xyst} (dotted line) is calculated as a function of ρ_t (vertical line) and ρ_{xst} (horizontal line). Thus, given P_x , P_y , and ρ , P_{xy} is completely determined by Eq. (34) and is consistent with the hypotheses for ρ . The asymptotic variance of the limiting normal (28) becomes a function of only P_x , P_y , and ρ . Note that in Eq. (33), X_{it} is treated as an independent and Y_{it} , a dependent variable. Similar results are obtained by reversing the roles of the two variables.

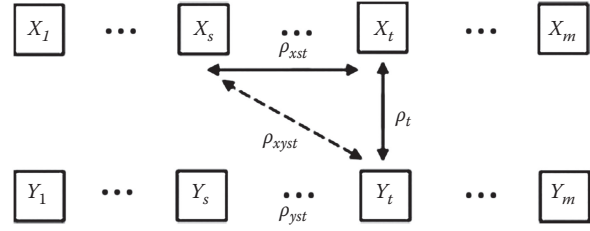


Fig. 3 Diagram illustrating the calculation of CVC correlation ρ_{xyst} using WVC correlations ρ_{xst} and ρ_t

In practice, we only need to assume one of the models. For convenience, we assume Eq. (33) in the following discussion. Note also that Eq. (34) follows from Eqs. (32) and (33). Although the reverse in general is not true, i.e., not all linear regression models satisfy Eq. (32), a large class of linear models does.^[55]

Power Functions Based on Asymptotic and Empirical Distributions

For the class of linear contrasts (27), it follows from the discussion in Part I that under H_0 and H_a , the quadratic quantity $Q_n^2 = n(K\hat{\rho} - \mathbf{a})^T(K\Sigma_r K^T)^{-1}(K\hat{\rho} - \mathbf{a})$ has approximately a central and non-central χ^2 distribution

$$H_0: Q_n^2 \sim X_g^2(0), \quad H_a: Q_n^2 \sim X_g^2(c) \quad (35)$$

where $c = n\mathbf{a}^T(K\Sigma_r K^T)^{-1}\mathbf{a}$, and $\chi_g^2(c)$ denotes a χ^2 distribution with g degrees of freedom and non-centrality parameter c . Thus, power function for the linear contrast (27) is similarly derived.^[55] As an alternative to the asymptotic results, power function may also be constructed based on the exact sampling distribution of $\hat{\rho}$, i.e., $T_F(\hat{\rho})$. For relatively small sample sizes, such an exact method provides more reliable power estimates, which are useful not only for planning small trials, but also for assessing the accuracy of the asymptotic approach. For example, we have used the exact distribution to evaluate the effect of the surrogacy as well as the normal distribution assumption in our previous investigation and found that the surrogacy assumption has little, while normality has a significant impact on power estimates.^[55] Details on exact distributions and their implementations using Monte Carlo simulations can be found in Ref. [55].

Missing Data Case

As discussed in section “Missing Data Case,” missing data arises frequently in real studies, particularly in longitudinal trials. Within the context of correlation analysis, either X_{it} or Y_{it} or both may be missing. By employing a similar strategy as in regression analysis, we define a vector of binary variables for indicating missing data as follows

$$r_{it} = \begin{cases} 1 & \text{if } (X_{it}, Y_{it}) \\ & \text{is observed} \\ 0 & \text{if } X_{it} \text{ or } Y_{it} \text{ or} \\ & \text{both is missing,} \end{cases} \quad \mathbf{r}_i = (r_{i1}, \dots, r_{im})^\top \quad (36)$$

As before, we again assume no missing data at baseline $t = 1$, i.e., $r_{i1} \equiv 1 (1 \leq i \leq n)$. In addition, we assume MCAR so that r_{it} is stochastically independent of (X_{it}, Y_{it}) .

In the presence of missing data, the Pearson estimates for the observed data can be expressed by using the missing data indicator as follows

$$\hat{\rho}_t = \frac{\sum_{i=1}^n r_{it} (X_{it} - \bar{X}_{\cdot t}) (Y_{it} - \bar{Y}_{\cdot t})}{\sqrt{\sum_{i=1}^n r_{it} (X_{it} - \bar{X}_{\cdot t})^2 \sum_{i=1}^n r_{it} (Y_{it} - \bar{Y}_{\cdot t})^2}} \quad (37)$$

where $\bar{U}_{\cdot t} = (\sum_{i=1}^n r_{it})^{-1} (\sum_{i=1}^n r_{it} U_{it})$. It can be shown that $\hat{\rho} = (\hat{\rho}_1, \dots, \hat{\rho}_m)^\top$ is a consistent estimate of ρ and asymptotically normal.^[56] The asymptotic distribution of $\hat{\rho}$ can be used to construct the power function for testing linear contrasts (27). The asymptotic variance in this case is a function of P_x , P_y , ρ , and the first two moments of $\mathbf{r}_i$. By modeling the moments of $\mathbf{r}_i$ similarly as in section “Modeling Missing Data Indicators with Maskov Processes” for regression models, we can again use the asymptotic distribution to construct the power function for linear contrasts (27).^[56]

ILLUSTRATION

In this section, we illustrate the approaches discussed with two worked examples. Although both examples are motivated by real study designs, we have altered the actual study designs to make them less complex and more suitable for exposition and illustration. Throughout the examples, we set type I error at $\alpha = 0.05$.

Example 1

In a longitudinal trial to compare treatment efficacy of a manualized information-motivation-behavioral skills intervention and a structurally equivalent health promotion control for HIV risk reduction with adolescent girls,^[57] we are interested in estimating power and sample size to detect some hypotheses concerning between-treatment differences at 3 (for short-term efficacy), 6, and 12 months (for long-term efficacy) postbaseline. For illustration purposes, we add a wait-list control, which, unlike the structurally equivalent health promotion, does not offer study participants any health promotion material other than have them assessed at baseline and the three prescheduled follow-up times. With the inclusion of the wait-list condition, we can also assess the placebo effect by the health promotion control.

The appropriate GEE model for this study design is the repeated measures analysis of variance (ANOVA)

$$E(Y_{kit}) = \mu_{ki}, \text{Var}(Y_{kit}) = \sigma_i^2, \quad 1 < i < nk, 1 \leq k \leq g, 1 \leq t \leq m \quad (38)$$

where g indicates the number of groups and n_k the group size. For the current study, $g = 3$ and $m = 4$, with $t = 1$ designating baseline. In addition, assume a common group size n , a common variance σ^2 , and a uniform compound symmetry structure $C_4(\rho)$ for the true correlation of the repeated responses.^[1–53] Under these assumptions, the true variance is given by $V = \sigma^2 C_4(\rho)$.

Now, consider testing for a constant between-group difference over the three postbaseline assessments

$$H_0 = \mu_{ki} - \mu_{(k-1)i} = 0 \quad \text{vs.} \quad H_a: \mu_{ki} - \mu_{(k-1)i} = a, \quad k = 2, 3, 2 \leq t \leq 4 \quad (39)$$

Note that for convenience and illustration purposes, we assumed a constant between-group difference in Eq. (39), though differential between-group changes may be more likely in the real study. In addition, we assumed no between-group difference at baseline so that the hypothesis only involves the means at postbaseline assessments. To eliminate the dependence of the power function on σ^2 , we divided the mean differences in Eq. (39) by σ so that a has the interpretation of the effect size.^[58] To express Eq. (39) in the form of Eq. (3), let

$$K = (\mathbf{c}(1, 2), \mathbf{c}(3, 4), \mathbf{c}(5, 6), -\mathbf{c}(1), -\mathbf{c}(3), -\mathbf{c}(5), -\mathbf{c}(2), -\mathbf{c}(4), -\mathbf{c}(6)), \mathbf{a} = a\mathbf{1}_6) \quad (40)$$

where $\mathbf{c}(i, \dots, j)$ denotes a 6×1 vector with 1 in the rows $i, \dots, j$ and 0 elsewhere, $\mathbf{1}_6$ a 6×1 vector of 1s, and $\beta = \mu$ with $\mu_k = (\mu_{k2}, \mu_{k3}, \mu_{k4})^\top$ and $\mu = (\mu_1^\top, \mu_2^\top, \mu_3^\top)^\top$.

Shown in Fig. 4 are power curves as a function of sample size and within-subject correlation for the hypothesis (39) with $a = 0.15$ under a time-stationary Markov model with a uniform missing data rate $q = 1 - \phi_{11} = 20\%$.

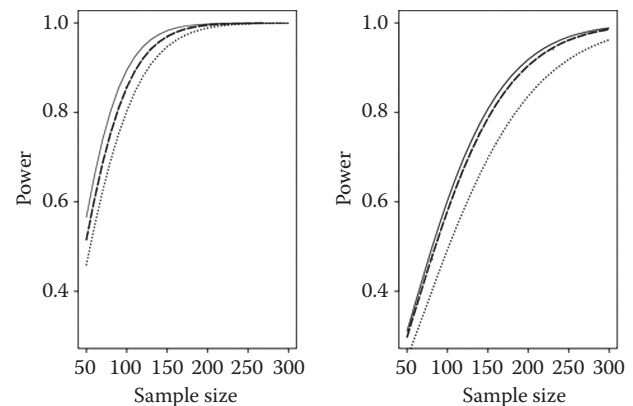


Fig. 4 Power curves for Example 1 with within-subject correlation $\rho = 0.2$ (left plot) and $\rho = 0.8$ (right plot) under (A) no missing data (—), (B) time-stationary Markov model with 20% missing data rate (---), and (C) a trimmed sample with 20% subjects removed (...)

For comparison purposes, we also plotted power based on the full sample and a trimmed sample by removing 20% of the subjects. As noted in section “Missing Data Case,” the trimmed approach is often used to guard against loss of power due to missing data and is justified based on a cross-sectional study design or a longitudinal study with a particular missing data pattern such as the one depicted in Fig. 1.

The plots in Fig. 4 show that the trimmed approach underestimates power with the amount of underestimation depending on the within-subject correlation ρ ; larger ρ leads to bigger differences between the trimmed and Markov model. As ρ measures the strength of correlation between repeated observations, smaller (larger) values of ρ give rise to less (more) correlated responses. Although both approaches have the same expected 20% missing data at each assessment point, they differ on how subjects contribute to the observed data over time. For the Markov model, missing data occurs randomly within each individual over time so that the individuals with observed responses are different across times. In contrast, the trimmed approach is based on a particular, fixed missing data pattern in which the observed responses are from the same subjects over time. Thus, there are more individuals who contribute to observed responses in the Markov model. As a result, this model yields more power than the trimmed approach. In addition, as ρ gets smaller, observations from the same individual become more independent and it matters less whether the observations are from the same individual or different individuals. On the other hand, when ρ gets larger, observations from the same subject become highly correlated, which further decreases the amount of independent information in the trimmed approach. Thus, the difference between the two approaches also varies as an increasing function of the within-subject correlation ρ .

In practice, the amount of missing data typically increases as time elapses, and the uniform missing data rate model is insufficient to capture such a dynamic trend. A one-step Markov model with time-varying transition probability matrices in Eq. (24) provides a way to accommodate a process that generates an increasing amount of missing data over time. Suppose that missing data increases over time at a rate given by

$$E(r_{it}) = p_{11}(1, t) = 0.80 + 0.10(3 - t), \quad 2 \leq t \leq 4 \quad (41)$$

Starting with 10% data missing at time 2, the missing data rate increases by 10% at each next assessment, though the averaged rate over the three assessments is still maintained at 20%. For convenience, we assume that the one-step transition probability to observe a response is independent of the response status at earlier times, i.e., $\phi_{01}(t) = \phi_{11}(t)$ ($2 \leq t \leq 3$). Given Eq. (41), we can use the recursive formula (25) to determine $\Phi(u)$ as well as $\mathbf{P}(1, t)$.

As there is no missing data at $t = 1$, $p_{01}(1, 2) = 0$. It follows that

$$\mathbf{P}(1, 2) = \begin{pmatrix} p_{11}(1, 2), & \bar{p}_{11}(1, 2) \\ p_{01}(1, 2), & \bar{p}_{01}(1, 2) \end{pmatrix} = \begin{pmatrix} 0.9, & 0.1 \\ 0, & 1 \end{pmatrix} = \Phi(1)$$

from which we obtain $\Phi(1)$. Similarly, by setting $t = 3, 4$ in Eq. (25) and solving for $\phi_{11}(u)$, we obtain

$$\Phi(2) = \begin{pmatrix} 0.8, & 0.2 \\ 0.8, & 0.2 \end{pmatrix}, \quad \Phi(3) = \begin{pmatrix} 0.7, & 0.3 \\ 0.7, & 0.3 \end{pmatrix}$$

Shown in Fig. 5 are power curves as a function of sample size for the hypothesis in Eq. (39) with $\mathbf{a} = (0.05, 0.05, 0.1, 0.1, 0.2, 0.2)^\top$ under the one-step time-varying Markov model. As before, we also plotted power based on the full and 20% trimmed sample. Note that unlike the uniform missing data rate model considered earlier, the trimmed sample does not have the same expected missing data rate as the Markov model at each assessment time. However, since the Markov model has an average rate of 20% missing data over the three assessment times, it is still sensible to compare this model with the trimmed approach.

Like Fig. 4, the plots in Fig. 5 also show that the difference between the two approaches varies as a function of the within-subject correlation. However, the trimmed approach no longer consistently underestimates power as in Fig. 4; in fact, it substantially overestimates power when $\rho = 0.8$. This overestimation by the trimmed approach is expected because with the redefined between-group difference vector $\mathbf{a}$, the ad hoc approach has more observations than the Markov model to detect a larger between-group difference at $t = 4$. Thus, with a general missing data process, power results from the trimmed approach become difficult to interpret as the bias can go either direction.

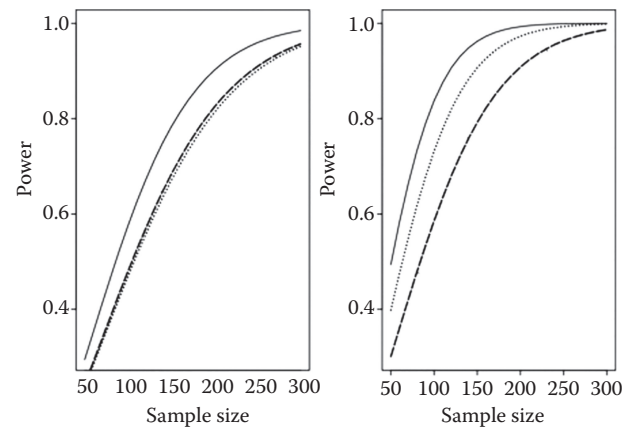


Fig. 5 Power curves for Example 1 with within-subject correlation $\rho = 0.2$ (left plot) and $\rho = 0.8$ (right plot) under (a) no missing data (—), (b) Markov model with time-vary one-step transition probability matrix (---), and (c) 20% trimmed sample (...)

Example 2

In an intervention study on posttraumatic stress disorder, one primary hypothesis is to determine whether the correlation between a biological and psychological measure changes over time. Suppose that there are five assessments; $t = 1, 2, 3, 4, 5$ with $t = 1$ designating baseline. Consider two sets of hypotheses

$$\begin{aligned} H_0 &= \rho_{xy} = (0.5, 0.5, 0.5, 0.5, 0.5)^\top \quad \text{vs.} \\ H_a &= \rho_{xy} = (0.5, 0.75, 0.75, 0.75, 0.75)^\top \\ H_0 &= \rho_{xy} = (0.5, 0.5, 0.5, 0.5, 0.5)^\top, \quad \text{vs.} \\ H_a &= \rho_{xy} = (0.5, 0.55, 0.60, 0.65, 0.750)^\top \end{aligned} \quad (42)$$

where H_a for the first hypothesis implies a constant change over the four assessments postbaseline, while H_a for the second tests for an upward trend from baseline to the last visit.

As in our earlier investigation with no missing data,^[55] we modeled P_x and P_y using a uniform compound symmetry correlation structure, $P_x = C5(\rho_x)$ and $P_y = C5(\rho_y)$, with $\rho_x(\rho_y)$ denoting the constant within-variable X_t (Y_t) correlation. For missing data, we assumed a time-stationary Markov process with a uniform missing data rate ϕ_{11} , as in Example 1.

Shown in Fig. 6 are the power curves (dotted line) for the two hypotheses with $\rho_x = \rho_y = 0.2$. For comparison purposes, we also plotted power based on the full sample (solid line) and a trimmed sample by taking out 80% of the individuals from the original sample (dashed line). The effect of missing data on power is clear from the plots in Fig. 6. In addition, it is interesting to see that the trimmed sample underestimated power as compared to the Markov modeling approach, though the amount of underestimation

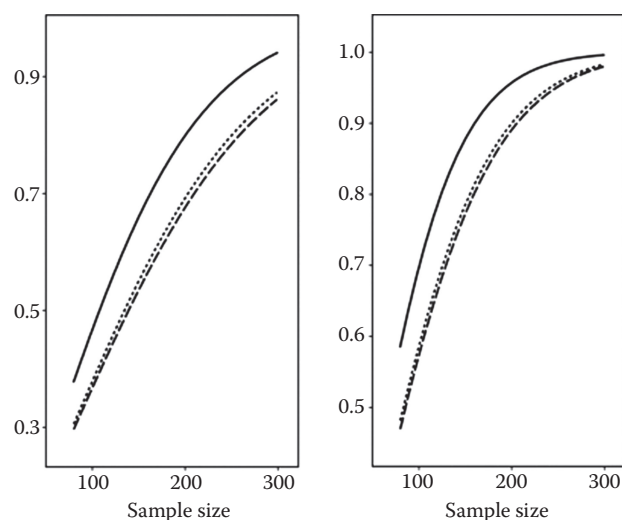


Fig. 6 Power curves for the first (left plot) and second set (right plot) of hypothesis in Example 2 with $\rho_x = \rho_y = 0.2$ under (a) no missing data (—), (b) 20% missing data per subject (...), and (c) a trimmed sample with 80% of the original subjects (-----)

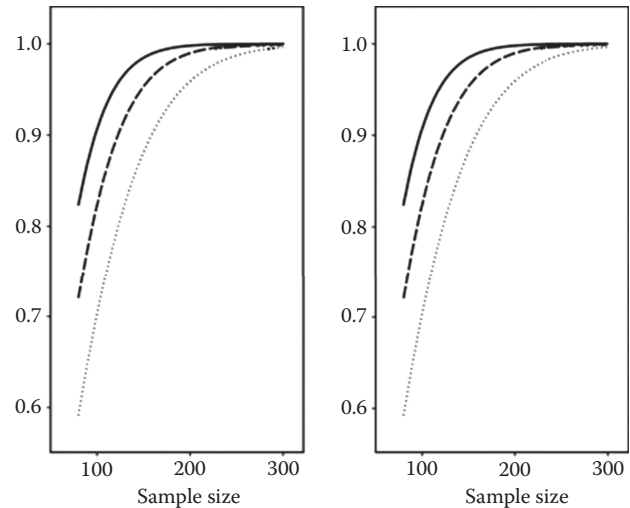


Fig. 7 Power curves for the first (left plot) and second set (right plot) of hypothesis in Example 2 with $\rho_x = \rho_y = 0.8$ under (a) no missing data (—), (b) 20% missing data per subject (...), and (c) a trimmed sample with 80% of the original subjects (-----)

is small for both hypotheses. To further explore the impact of the input parameters on power, we changed the values of the within-variable correlation ρ_x (ρ_y), from 0.2 to 0.8, while keeping the other input parameters the same as before. Shown in Fig. 7 are the resulting power curves for the two hypotheses. The plot again shows the impact of missing data on power. However, in comparison to Fig. 6, the power curves now exhibit a different pattern; the power curves of the two missing data samples are no longer close to each other and in addition, the trimmed sample overestimated power for both hypotheses. Thus, as in the regression model case, bias in the power estimates by the trimmed approach can go in either direction.

CONCLUSIONS

In this entry, we reviewed existing work on power analysis for longitudinal and other types of clustered study designs and focused on GEE and LMM for regression analysis and product-moment correlation for correlation analysis. In addition, we also presented our latest development in addressing the issue of missing data pattern. Although various methods have studied the effect of within-cluster correlation within different modeling contexts,^[1,4,29,30,59–61] the effect of missing data pattern has never been systematically examined within the general context of clustered study designs for regression as well as correlation analyses. Thus, this new development represents a major advance in research on power analysis for sample size calculation (68).

Our investigation of power as a function of the within-subject correlation in the presence of missing data

indicates that power is highly dependent on this parameter whether conducting regression or correlation analyses. This interplay between within-subject correlation and missing data illustrates the importance to address the issue of missing data pattern to provide reliable power estimates for real study designs. In studying the effect of random missing data patterns on power, we contrasted the proposed approaches with a widely used ad hoc method that accounts for missing data by inflating the sample size calculated based on a particular fixed missing data pattern. As illustrated by the examples in section “Illustration,” this trimmed approach in general provides unreliable estimates as bias in the power estimates by this approach can go either direction. Although other existing ad hoc methods may provide less biased estimates, they are more model specific and inapplicable to the general models considered in this entry.

One limitation of the work proposed to address missing data is the MCAR assumption. Accommodating more general missing data assumptions such as the missing at random (MAR) mechanism not only requires more complex models for the missing data indicator, but also different types of parameter estimates. For example, under MAR, the estimating equations considered in section “Part I. Power Analysis for Regression Models” no longer yield consistent estimates, and different type of estimating equations must be considered.^[35,37,62–67] We will address these more general missing data mechanisms in future research.

ACKNOWLEDGMENT

This research is supported in part by an NIH/NIDA R01-DA012249 grant (Feng, Tang, and Tu) and by an NIH/NIAD grant K22AI 51186 (Kowalski). We thank Prof. Chow at Duke University for inviting us to write this paper and for Associate Editor, Alison Cohen, for the Encyclopedia of Biopharmaceutical Statistics (EBS) and Cheryl Bliss-Clark, in the Department of Biostatistics and Computational Biology at University of Rochester, for assistance with manuscript preparation.

Minor updates have been made by the Editor-in-Chief in order to reflect developments since publication of the *Encyclopedia of Biopharmaceutical Statistics, Third Edition*.

REFERENCES

1. Diggle, P.J.; Heagerty, P.J.; Liang, K.Y.; Zeger, S.L. Analysis of Longitudinal Data, 2nd; Oxford University Press: New York, 2002.
2. Donner, A.; Klar, N. Cluster randomization trials in epidemiology: theory and application. *J. Stat. Planning Inference* **1994**, *42*, 37–56.
3. Manatunga, A.K.; Hudgens, M.G.; Chen, S. Sample size estimation in cluster randomization studies with varying cluster size. *Biomet. J* **2001**, *43*, 75–86.
4. Tu, X.M.; Kowalski, J.; Zhang, J.; Lynch, K.; Crits-Christoph, P. Power analyses for longitudinal trials and other clustered designs with missing data. *Stat. Med* **2004**, *23*, 2799–2815.
5. Agresti, A.; Booth, J.G.; Hobert, J.P.; Caffo, B. Random Effects Modeling of Categorical Response Data; Technical Report; Department of Statistics; University of Florida: Gainesville, FL, 2001.
6. Breslow, N.E.; Clayton, D.G. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc* **1993**, *88*, 9–25.
7. Davidian, M.; Gallant, A.R. The nonlinear mixed effects model with a smooth random effects density. *Biometrika* **1993**, *80*, 475–488.
8. Gilmour, A.R.; Anderson, R.D.; Rae, A.L. The analysis of binomial data by a generalized linear mixed model. *Biometrika* **1985**, *72*, 593–599.
9. Goldstein, H.; Rasbash, J. Improved approximations for multilevel models with binary responses. *J. Royal Stat. Soc. Ser. A* **1996**, *159*, 505–513.
10. Laird, N.; Ware, J. Random-effects models for longitudinal data. *Biometrics* **1982**, *38*, 963–974.
11. Liang, K.Y.; Zeger, S.L. Longitudinal data analysis using generalized linear models. *Biometrika* **1986**, *73*, 13–22.
12. Lin, X.; Breslow, N.E. Bias correction in generalized linear mixed models with multiple components of dispersion. *J. Am. Stat. Assoc* **1996**, *91*, 1007–1016.
13. Lipsitz, S.R.; Kyungmann, K.; Zhao, L. Analysis of repeated categorical data using generalized estimating equations. *Stat. Med* **1994**, *13*, 1149–1163.
14. Pinheiro, J.C.; Bates, D.M. Approximations to the log-likelihood function in the non-linear mixed-effects model. *J. Comput. Graph. Stat* **1995**, *4*, 12–35.
15. Prentice, R.L. Correlated binary regression with covariates specific to each binary observation. *Biometrics* **1988**, *44*, 1033–1048.
16. Prentice, R.L.; Zhao, L.P. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **1991**, *47*, 825–839.
17. Schluchter, M.D. Methods for the analysis of informatively censored longitudinal data. *Stat. Med* **1992**, *11*, 1861–1870.
18. Stiratelli, R.; Laird, N.; Ware, J.H. Random-effects models for serial observations with binary response. *Biometrics* **1984**, *40*, 961–971.
19. Wang, N.; Lin, X.; Gutierrez, G.; Carroll, R.J. Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *J. Am. Stat. Assoc* **1998**, *93*, 249–261.
20. Wolfinger, R. Laplace’s approximation for nonlinear mixed models. *Biometrika* **1993**, *80*, 791–795.
21. Zeger, S.L.; Liang, K.Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **1986**, *42*, 121–130.
22. McCullagh, P.; Nelder, J.A. Generalized Linear Models, 2nd Ed.; Chapman and Hall: London, 1989.
23. Pepe, M.S.; Anderson, G.L. A cautionary note on inferences for marginal regression models with longitudinal

- data and general correlated response. *Commun. Stat. Part A—Theory Methods* **1994**, 23, 939–951.
24. Chung, K.L. *A Course in Probability Theory*, 3rd Ed.; Academic Press: Berkeley, CA, 2001.
25. Serfling, R.J. *Approximation Theorems of Mathematical Statistics*; Wiley: New York, 1980.
26. Seber, G.A.F.; Wild, C.J. *Nonlinear Regression*; Wiley: New York, 1989.
27. Kowalski, J.; Mendoza-Blanco, J.R.; Tu, X.M.; Gleser, L.R. On the difference in inference and prediction between the joint and independent t-error models for seemingly unrelated regressions. *Commun. Stat. Theory Methods* **1999**, 28, 2119–2140.
28. Liang, K.Y.; Liu, X.H. Godambe, V.P. Estimating equations in generalized linear models with measurement error. In *Estimating Functions*; Eds.; Oxford University Press: Oxford, 1991, 47–63.
29. Wu, M. Sample size for comparison of changes in the presence of right censoring caused by death, withdrawal, and staggered entry. *Controll. Clin. Trials* **1988**, 9, 32–46.
30. Hedeker, D.; Gibbons, R.D.; Waternaux, C. Sample size estimation for longitudinal designs with attrition: comparing time-related contrasts between two groups. *J. Educ. Behav. Stat* **1999**, 24, 70–93.
31. Donner, A. Sample size requirements for stratified cluster randomization designs. *Stat. Med* **1992**, 11, 743–750.
32. Tu, X.M.; Zhang, J.; Kowalski, J.; Shults, J.; Feng, C.; Sun, W.; Tang, W. Power analyses for longitudinal study designs with missing data. *Stat. Med* **2007**, 26 (15), 2958–2981.
33. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, 63, 581–592.
34. Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*; Wiley: New York, 1987.
35. Robins, J.; Rotnitzky, A.; Zhao, L.P. Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data. *J. Am. Stat. Assoc* **1995**, 90, 106–121.
36. Kowalski, J.; Tu, X.M. A GEE approach to modeling longitudinal data with incompatible data formats and measurement error: application to HIV immune markers. *J. Royal Stat. Soc. Ser. C* **2002**, 51, 91–114.
37. DeGruttola, V.; Tu, X.M. Modeling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics* **1994**, 50, 1003–1015.
38. Hoel, P.G.; Port, S.C.; Stone, C.J. *Introduction to Stochastic Processes*; Waveland Press: Prospect Heights, IL, 1987.
39. Ross, S.M. *Introduction to Probability Models*; Academic Press: San Diego, CA, 1993.
40. Boik, R.J. A local parameterization of orthogonal and semi-orthogonal matrices with applications. *J. Multivar. Anal* **1998**, 67, 244–276.
41. Browne, M.W.; Shapiro, A. The asymptotic covariance matrix of sample correlation coefficients under general conditions. *Linear Algebra Appl* **1986**, 82, 169–176.
42. Elston, R.C. On the correlation between correlations. *Biometrika* **1975**, 62, 133–140.
43. Hsu, P.L. The limiting distribution of functions of sample means and applications to testing hypotheses. In *Proc. the First Berkeley Symposium on Mathematical Statistics and Probability*, 1949, 359–402.
44. Kollo, T.; Neudecker, H. Asymptotics of eigenvalues and unit-length eigenvectors of sample variance and correlation matrices. *J. Multivar. Anal* **1994**, 47, 283–300.
45. Neudecker, H.; Wesselman, A.M. The asymptotic distribution of the sample correlation matrix. *Linear Algebra Appl* **1990**, 127, 589–599.
46. Olkin, I.; Siotani, M. Ikeda, S. Asymptotic distribution of functions of a correlation matrix. In *Essays in Probability and Statistics*; Ed.; Shinko Tsusho: Tokyo, 1976, Chapter 16; 235–251.
47. Steiger, J.H.; Hakstian, A.R. The asymptotic distribution of elements of a correlation matrix: theory and applications. *Br. J. Math. Stat. Psychol* **1982**, 35, 208–215.
48. Dunn, O.J.; Clark, V.A. Correlation coefficients measured on the same individuals. *J. Am. Stat. Assoc* **1969**, 64, 366–377.
49. Dunn, O.J.; Clark, V.A. Comparison of tests of the equality of dependent correlation coefficients. *J. Am. Stat. Assoc* **1971**, 66, 904–908.
50. Machin, D.; Campbell, M.J.; Fayers, P.M.; Pinol, A.P.Y. *Sample Size Tables for Clinical Studies*; Blackwell Science: Oxford, 1997.
51. Mudholkar, G.S. Samuel, K.; Johnson, N.L. Fisher's z-transformation. In *Encyclopedia of Statistical Sciences*; Eds.; Wiley: New York, 1983, Vol. 3, 130–135.
52. Billingsley, P. *Probability and Measure*, 3rd Ed.; Wiley: New York, 1995.
53. Jennrich, R.I.; Schluchter, M.D. Unbalanced repeated-measures models with structured co-variance matrices. *Biometrics* **1986**, 42, 805–820.
54. Fuller, W.A. *Measurement Error Models*; Wiley: New York, 1987.
55. Tu, X.M.; Kowalski, J.; Crits-Christoph, P.; Gallop, R.J. Power analyses for correlations from clustered study designs. *Stat. Med* **2007**, 26 (15), 2958–2981.
56. Tu, X.M.; Feng, C.; Kowalski, J.; Tang, W.; Wang, H. Correlation analysis for longitudinal data: applications to HIV and psychosocial research, *Statistics in Medicine*, **2007**, 26 (22), 4116–4138.
57. Morrison-Beedy, D.; Carey, M.P.; Kowalski, J.; Tu, X.M. Group-based HIV risk reduction intervention for adolescent girls: evidence of feasibility and efficacy. *Res. Nurs. Health* **2005**, 28, 3–15.
58. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd Ed.; Lawrence Erlbaum Associates: New Jersey, 1988.
59. Lee, J.W.; DeMets, D.L. Sequential comparison of changes with repeated measurements data. *J. Am. Stat. Assoc* **1991**, 86, 757–762.
60. Muller, K.E.; LaVange, L.M.; Ramey, S.L.; Ramey, C.T. Power calculations for general linear multivariate models including repeated measures applications. *J. Am. Stat. Assoc* **1992**, 87, 1209–1226.
61. Rochon, J. Application of GEE procedures for sample size calculations in repeated measures experiments. *Stat. Med* **1991**, 17, 1643–1658.
62. Diggle, P.J.; Kenward, M.G. Informative drop-out in longitudinal data analysis. *Appl. Stat* **1994**, 43, 49–93.
63. Follman, D.; Wu, M.C. An approximate generalized linear model with random effects for informative missing data. *Biometrics* **1995**, 51, 151–168.

64. Hogan, J.W.; Laird, N.M. Model-based approaches to analysing incomplete longitudinal and failure time data. *Stat. Med* **1997**, *16*, 259–272.
65. Rotnitzky, A.; Robins, J.; Scharfstein, D.O. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Am. Stat. Assoc* **1998**, *93*, 1321–1339.
66. Wu, M.; Carroll, R. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **1988**, *44*, 175–188.
67. Wulfsohn, M.S.; Tsiatis, A.A. A joint model for survival and longitudinal data measured with error. *Biometrics* **1997**, *53*, 330–339.
68. Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. *Sample Size Calculation in Clinical Research*. 3rd Ed.; Taylor & Francis: New York, 2017.

Companion Diagnostics

Gene Pennello

Center for Devices and Radiological Health (CDRH), Food and Drug Administration (FDA),
Silver Spring, Maryland, U.S.A.

Jingjing Ye

Center for Drug Evaluation and Research (CDER), Food and Drug Administration (FDA),
Silver Spring, Maryland, U.S.A.

Abstract

Companion diagnostics have become important medical devices for tailoring treatment choices to individual patient characteristics. A companion diagnostic test is one that is essential for the safe and effective use of a therapeutic product. Companion biomarker tests measure or detect one or more biomarkers. We review statistical methods for evaluating companion diagnostics at different phases of development, including biomarker development, analytical validation, and clinical validation.

Keywords: Biomarker discovery; Biomarker strategy; Clinical utility; Enrichment; Omics-based classifier; Predictive biomarker.

I. INTRODUCTION

Companion diagnostic tests have become increasingly important to the advancement of precision medicine, i.e., personalizing or tailoring medical treatment or management decisions based on individual patient characteristics. The US Food and Drug Administration (FDA) defines an *in vitro* companion diagnostic device (or test) as one that is essential for the safe and effective use of a therapeutic product^[1]. The therapeutic product may be drug, biologic, novel product, or existing product with a new indication.

In their initial guidance on companion diagnostics^[1], FDA indicates that an *in vitro* diagnostic test could be essential for the safe and effective use of a therapeutic product if it is used to

- identify patients who are most likely to benefit from the therapeutic product,
- identify patients likely to be at increased risk for serious adverse reactions as a result of treatment with the therapeutic product,
- monitor response to treatment with the therapeutic product for the purpose of adjusting treatment (e.g., schedule, dose, discontinuation) to achieve improved safety or effectiveness,
- identify patients in the population for whom the therapeutic product has been adequately studied, and found safe and effective, i.e., there is insufficient information about the safety and effectiveness of the therapeutic product in any other population.

In a second, draft guidance^[2], FDA elaborates on principles for co-development of an *in vitro* companion diagnostic test with a therapeutic product, including regulatory processes and scientific principles to support contemporaneous marketing authorization of the companion diagnostic with the corresponding therapeutic. Topics covered by the guidance include analytical validation of the test (demonstration that the detection or measurement of analyte(s) by the test is accurate and reliable), clinical validation of the test with a prospective or retrospective study, and types of marketing claims supported by validation studies. Statistical principles are also covered, including clinical study designs and robustness of study conclusions to missing test results.

A companion diagnostic assay is usually based on detection or measurement of one or more biomarkers. The biomarkers may be directly related to the pharmacodynamic or pharmacokinetic properties of the corresponding therapy^[3]. In other instances, efforts are undertaken to discover one or more biomarkers empirically to develop a biomarker-based classifier or signature that can explain heterogeneity in a treatment effect.

Much but not all of the literature on companion diagnostics concerns oncology because of opportunities to develop highly effective cancer treatments for molecular targets that can be evaluated with biomarker measurements. In oncology, predictive biomarkers are defined as informing on likely outcomes with specific treatments (e.g., relative sensitivity or resistance), while prognostic biomarkers are defined as informing on likely outcomes independent of specific treatment (i.e. ability of tumor to proliferate, invade, and/or spread)^[4]. In general predictive

markers are candidates for the development of companion diagnostic tests, while prognostic markers may not be, although they could be clinically actionable. Well-known examples of prognostic signatures are Oncotype DX and MammaPrint for patients with stage I, estrogen receptor positive breast cancer^[5,6]

A large literature exists on principles for discovery and validation of biomarkers for treatment decision making. Guidelines are available regarding design, conduct, analysis, and reporting of clinical trials of a therapeutic in combination with a companion diagnostic test for treatment selection. Guidelines developed by researchers and professional medical societies include Deverka, Messner, et al.^[7], McShane, Cavenagh, et al.^[8], Fridlyand, Simon, et al.^[9], Matcher^[10], Micheel, Nass, et al.^[11], Moons, Kengne et al.^[12], Moore^[13], Khleif, Doroshow, et al.^[14], Ransohoff and Gourlay^[15], Simon, Paik, et al.^[16], Pepe, Feng, et al.^[17], Simon^[18], McShane, Altman, et al.^[19], and Bossuyt, Johannes, et al.^[20]. In addition, FDA and the National Institutes of Health (NIH) have jointly published a glossary on biomarkers, endpoints, and other tools (BEST)^[21].

In this entry, we review statistical methods for companion diagnostic tests. In lieu of the regulatory term, these tests have been called predictive marker tests and treatment selection tests in the literature. Our intention is to review developments that we consider important, not to convey FDA's current thinking on accepted statistical practices for companion diagnostics.

In Section II we briefly review statistical methods and principles in biomarker discovery and signature development. In Section III, we review different types of evaluations that may be required for validating companion diagnostic tests. In Section IV we review clinical study designs for validating a companion diagnostic test with its corresponding therapeutic product. In Section V we discuss selected statistical methods for evaluating the clinical utility of a companion diagnostic. In Section VI we consider missing test results as well as other types of non-missing test results that require special care in statistical analysis. In Section VII, we discuss follow-on companion diagnostics tests. Finally, in Section VIII we provide some concluding remarks.

II. BIOMARKER DISCOVERY AND SIGNATURE DEVELOPMENT

A strength of some emerging technologies is that a very large number of biomarkers can be interrogated simultaneously. For example, next generation sequencing of the whole genome can potentially evaluate 3.2 billion base-pairs of DNA in one sample. However, despite intense biomarker discovery efforts made possible with emerging technologies, relatively few biomarkers have been commercially developed into marketable tests for treatment decision making^[22]. Indeed, no complex classification signature

combining multiple biomarkers has been FDA-approved as a companion diagnostic^[23], despite that methodologies for classifier development abound, e.g., Fisher linear discriminant analysis, maximum likelihood discriminant rules, nearest-neighbor classifiers, classification trees including classification and regression trees (CART), aggregation classifiers including bagging and boosting, support vector machines (SVMs), random forests, Naïve Bayes classifiers and many more^[23–25].

While thousands of biomarkers may be simultaneously evaluated as candidate predictors of treatment efficacy or safety, typically most are uninformative for a clinical outcome, complicating efforts to identify signal from noise. Accompanying the simultaneous evaluation of a large number of candidate biomarkers is large random variability, which can easily generate spurious associations of biomarker values with clinical conditions of interest. To mitigate the large multiplicity problem, statistical methods have been used to control the false discovery rate.

In omics-based discovery research, a small sample size n is frequently used to evaluate a much larger number of candidate biomarkers p . In this small n large p setting, if the p variables are not co-linear they are typically capable of being used to develop a rule that classifies disease state correctly in all n patients with no prediction error. However, the development and evaluation of a classifier on the same dataset amounts to “double-dipping”, which can lead to large internal bias in the performance of the classifier, known as resubstitution bias^[27–29].

To increase confidence that a classification model can predict a clinical condition in out-of-sample testing, its performance may be cross-validated within a developmental dataset. In split-sample cross-validation, one part of a dataset is used to build the model and the remaining part is used to test it. If model building is automated, the data may be split repeatedly to train and test the model. Commonly, the classifier is developed on $k - 1$ parts and tested on the remaining k th part, with the process repeated k times for each possible selection of $k - 1$ parts for training. Performance is summarized (e.g., averaged) across the k test sets. As k increases, average performance across the test sets becomes less biased but more variable. In leave-one-out cross validation, $k = n$, and average performance is least biased but most variable. Alternatively, in one form of bootstrap cross-validation, a bootstrap dataset of the same size as the original dataset serves as the training dataset and units left out of the bootstrap dataset serve as the test dataset.

A critical step in developing a classifier is feature selection, that is, selection of the variables (e.g., biomarkers) used in the classifier as predictors for the clinical condition of interest. Feature selection is typically the largest source of bias in resubstitution performance and in cross-validation performance if it is not considered to be a model-building step^[27–28]. A common mistake is to select features based on an entire dataset and cross-validate only the steps used to build a model with the selected features.

As with all phases of medical product development, biomarker discovery is more likely to be successful to the extent that it is planned. Major discoveries that advance the field are often surprises (almost by definition), but are usually the result of carefully planned, reproducible research. Principles of experimental design are often crucial to employ in early phase studies^[30] to avoid generating biomarker associations with clinical conditions that are confounded by, for example, convenience sampling, sample test order, co-morbidities, analyte instability during collection/storage, variation in sample handling/processing, batch effects, geography, ethnicity, etc.

III. LEVELS OF DIAGNOSTIC TEST EVALUATION

Fryback and Thornbury^[31] organized the evaluation of diagnostics devices (tests) into six levels: technical efficacy, diagnostic accuracy efficacy, diagnostic thinking efficacy, therapeutic efficacy, patient outcome efficacy and societal efficacy. While they developed their conceptual model with a focus on diagnostic imaging systems, it is broadly applicable to the evaluation of any diagnostic test, in particular companion diagnostics. Commonly, diagnostic tests are evaluated for technical efficacy (measurement quality) and diagnostic accuracy (association of test result with a clinical condition of interest).

For in vitro diagnostics, technical efficacy (level 1) corresponds to analytical performance, the characterization of the test to detect or measure analytes accurately and reliably. Generally, in vitro diagnostics (IVDs) are analytically validated prior to clinical validation^[5]. To demonstrate analytical validity, studies are designed to evaluate various aspects of test measurement quality, which may include bias, precision, limit of detection, reagent stability, analytical specificity, and, when processed samples are used in place of clinical samples, commutability between sample types. Bias is evaluated by comparing a test measurement with the true (or accepted reference) value of the measurand. Measurement imprecision (variability) can be evaluated with replicate measurements taken under the same set of conditions, e.g., within a run (repeatability) or under different conditions, e.g., replicates tested in different labs (reproducibility). Reagent stability may refer to shelf-life, in-use life, or shipment. Analytical specificity refers to robustness of the measurement to interfering substances, cross-reactivity, or cross-contamination.

The Clinical Laboratory Standards Institute (CLSI) develops guidelines on clinical and laboratory practices, including protocols for evaluating various aspects of the analytical performance of laboratory measurement procedures. Evaluation Protocol guidelines include an overview of evaluations of analytical performance that may be necessary to perform for a measurement procedure^[32] and a harmonization of symbology and equations used to describe statistical equations and models in the guidelines^[33].

For many diagnostic tests, the focus of clinical evaluation is diagnostic accuracy – level 2 in the Fryback-Thornbury hierarchy – using such measures as sensitivity, specificity, diagnostic likelihood ratio, predictive value, and receiver operating characteristic (ROC) plot. In contrast, because companion diagnostics are associated directly with treatment recommendations, clinical evaluation is typically based on therapeutic and patient outcome efficacy (levels 3–4). Therapeutic efficacy refers to the frequency with which the diagnostic result helped the physician plan the management or treatment of the patient. Patient outcome efficacy – also called clinical or medical utility in literature – refers to the effect of the test on patient outcomes when used as intended.

IV. CLINICAL TRIAL DESIGNS

Many companion diagnostic tests provide a binary output (test negative or positive) to identify a population most likely to benefit from a therapeutic product. Other binary tests identify a population at risk for a serious adverse event if treated with the product. The tests may be based on measurements of one or more biomarkers that are thought to be predictive of these endpoints. For binary-valued companion biomarker tests that determine eligibility of treatment, three prospective Phase III trial designs are the biomarker-stratified (or interaction) design, the biomarker-strategy design, and the enrichment (or targeted) design (Fig. 1)^[34–35]. The designs could be generalized to evaluate companion diagnostic tests that discriminate among multiple treatment options.

A. Biomarker-Stratified Design

In a biomarker-stratified design, all patients are enrolled regardless of biomarker test value. Enrolled patients are randomized to the treatment or a control. The treatment effect (e.g., hazard ratio between treatment and control) is stratified by biomarker test value. If test value explains variation in treatment effect that is statistically and clinically significant, then the test could be considered essential for the safe and effective use of the treatment and therefore a companion diagnostic (Section V). For example, for a test to be considered a companion diagnostic, the treatment effect may need to be positive (beneficial) and large (clinically significant) in patients who are biomarker test positive, but close to zero or negative (harmful) in patients who are biomarker test negative.

For small to moderately sized trials, imbalances may occur by chance in the numbers of patients randomly assigned to the treatment and control groups within subsets defined by biomarker test value. The imbalance can lead to less precision in the treatment effect estimate relative to 1:1 perfect balance in treatment assignment. Stratification of the randomization by biomarker test value would ensure 1:1 balance and thus may increase precision of estimation, but is not necessary for unbiased evaluation of the treatment

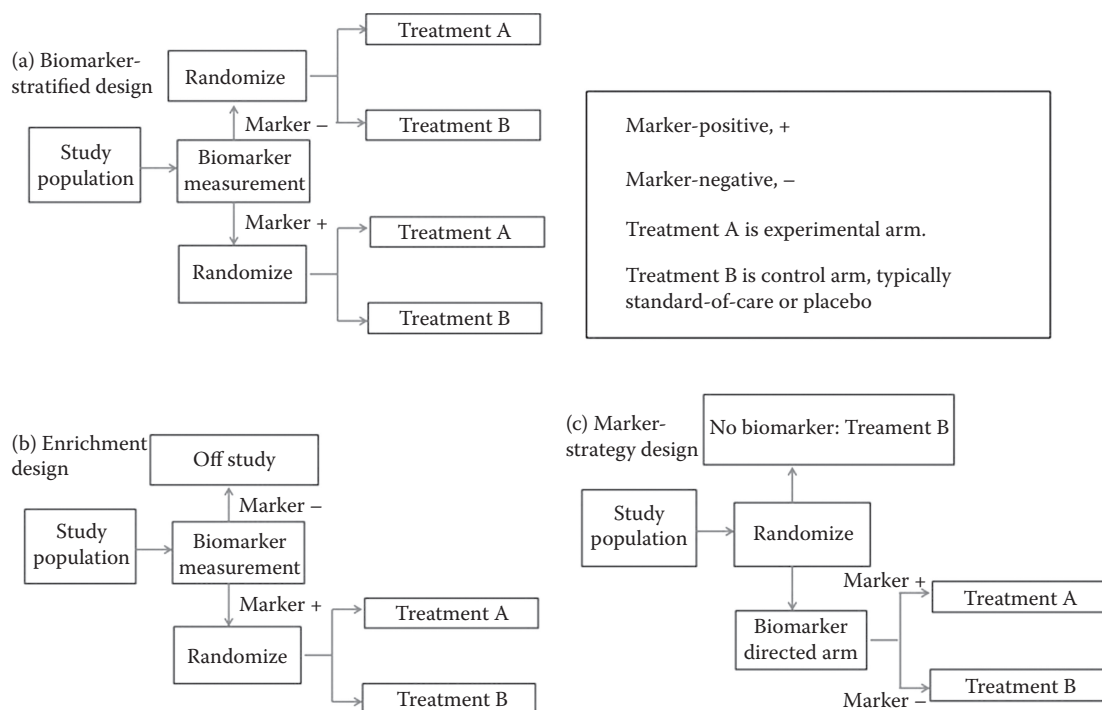


Fig. 1 Clinical Trial Designs for Treatments with Companion Diagnostics

effect within biomarker subsets because within the subsets treatment assignment is randomly generated^[36]. In general, the treatment effect within a biomarker subset is usually better estimated if adjusted for covariates observed to have effects on outcome and distributed differently between the treatment and control groups.

B. Enrichment Design

In an enrichment design, only patients with particular biomarker test values are enrolled into a trial of the investigational treatment. For example, in a predictive enrichment strategy^[37], the only patients enrolled are those with a biomarker test value indicating they are more likely to respond to treatment. The enrichment design requires that the test be available during enrollment to screen patients for inclusion into the trial depending on their biomarker test value.

An enrichment design may be necessary for a clinical trial to have equipoise in randomizing patients to a treatment or a control. While the biomarker-stratified design is efficient for evaluating differential treatment effects in biomarker subpopulations, its assignment of the treatment to some patients with biomarker test values that are supposed to predict lack of benefit or increased risk of an adverse event could be unethical. In general, if the predictive ability of the biomarker for an efficacy or safety outcome during treatment is based on a strong biological rationale or compelling Phase II evidence, and if the biomarker is measured well by the test (i.e., is accurate and reproducible), then clinical equipoise for a randomized trial of the treatment may only exist among the subset of

patients with biomarker test values that predict a favorable benefit-risk trade-off for the treatment.

An enrichment trial may show that the treatment is safe and effective in a population defined by biomarker test value. The enrichment trial is not designed to evaluate if the treatment is safe and effective in other populations. For example, an enrichment trial may demonstrate that the treatment effect is clinically and statistically significant in patients who are biomarker test positive. The test would then be indicated as a companion diagnostic to select biomarker test positive patients as the intended use population for the treatment. However, the treatment may also happen to be effective in biomarker test negative patients. In practice, these patients may be dissuaded from receiving the treatment and obtaining its benefits.

C. Biomarker Strategy Design

In a biomarker strategy design, patients are randomized to either have their biomarker test value used or not used to determine if they should receive the experimental treatment or the control^[40]. If the biomarker value is not used, the patient receives the control. If the biomarker test value is used, then a patient receives the treatment if they test positive for the biomarker and the control if they test negative. The design is an attempt to evaluate if test-guided treatment selection can improve average clinical outcome. However, as noted by many authors^[35,39–41], the design has at least two shortcomings

First, the effectiveness of the test-based strategy for selecting the treatment cannot be separated from a

homogenous treatment effect. In this case, even a random test will select a subset of patients for whom the treatment is effective relative to control. Similarly, if the treatment is effective in subgroups unrelated to the test strategy, then even a random test strategy will by chance select some patients in these subgroups. In both cases, the study will tend to show that outcomes are better in the test strategy arm than the control arm. These results may be over-interpreted to mean that the test is appropriately selecting patients for treatment when in fact it is selecting patients at random. To avoid over-interpretation of study results, a difference between the test strategy and control arm in an outcome would be compared with its null distribution if the test strategy was random. Unfortunately general methods for determining this null distribution are unclear.

Second, the biomarker strategy design is inefficient. Patients who are biomarker test negative will receive the control whether they are randomized to the test strategy arm or the control arm. These patients will dilute the average difference between the test strategy arm and the control arm in an outcome. More precisely, if a linear treatment effect on outcome equals Δ in biomarker test positive patients and the probability of testing positive is π , then the expected difference between the test strategy arm and the control arm is $\pi \Delta$. Consider that the sample size to power a study to detect a treatment effect is roughly inversely proportional to the square of the effect size. Then, the sample size would need to be $1/\pi^2$ times greater for the biomarker-strategy design than for an enrichment design in which only biomarker test positive patients are enrolled^[40–41]. For example if $\pi = 0.5$, the sample size would need to be 4 times larger for the marker strategy design than the enrichment design to detect the linear effect.

D. Prospective-Retrospective Design

In a so-called prospective-retrospective study, a prospective plan is developed for evaluating an existing randomized trial for a treatment and its companion diagnostic test. The existing trial would ideally enroll the same patient population and have the same objectives as the prospective-retrospective study. Uniform policies should be in place for sample collection, handling, and processing. The test would need to be analytically validated for stored samples^[16]. An adequate amount of sample material would need to be available for testing from enough patients in the trial for test results to be representative.

The biomarker-stratified design is often accomplished with retrospective study of a trial that was conducted prior to knowing that treatment efficacy or safety could have been predicted by a biomarker test. The retrospective study permits evaluation of the predictive ability of the biomarker test by examining the study for biomarker by treatment interaction on relevant outcomes.

A prospective-retrospective study has the potential to expedite validation of a companion diagnostic test for a

therapeutic. However, qualitatively, the level of evidence for a prospective-retrospective study is considered lower than for a prospective randomized clinical trial for several reasons, including the retrospective analysis is based on a clinical trial usually not explicitly designed for the retrospective analysis objective, is based on test results on available stored samples rather than on fresh samples, and could be subject to confounding due to uncontrollable factors^[16]. Therefore, a recommendation is that validation of a companion diagnostic test be based on more than one prospective-retrospective study yielding similar conclusions^[16].

E. Other Designs

Other fixed sample size designs for randomized trials of treatments in combination with biomarker tests have been described and evaluated^[34,38–42]. A variant on the enrichment design is to enroll only the subset of patients for whom biomarker test result changes the treatment decision from the pre-test planned therapy (i.e., the subset with level 4 therapeutic efficacy in Fryback and Thornbury's hierarchy). For example, to evaluate the clinical utility of procalcitonin to guide antibiotic therapy decision making for patients suspected to have lower respiratory tract infection, patients with a low procalcitonin level who would otherwise have been started on a course of antibiotic therapy could be randomized to receive the therapy or not, the general premise of the upcoming TRAP trial (<http://arlg.org/studies-in-progress>).

Some trials of treatments with companion biomarker tests are designed to be adaptive in various aspects, including sample size, enrollment (enrichment), randomization ratio, biomarker signature, and biomarker threshold (e.g.,^[43–48]). An adaptive threshold design is a special type of a biomarker analysis design^[47] for adaptive determination of a biomarker-based subset in whom a treatment is recommended. Adaptive designs that develop and validate a companion diagnostic in the same study blur the usual recommendation to perform these phases on separate datasets. However, a biomarker analysis design could be conceivable for the pivotal validation of a companion diagnostic with exceptional and highly robust analytical accuracy and reproducibility.

Basket and umbrella study designs are being used to increase clinical trial efficiency for evaluating therapies in oncology with molecular targets.^[44,49] In a basket trial, cancer patients are screened with a biomarker test for the molecular target of an investigational therapy, permitting its evaluation in multiple histological subtypes or tumor types (the basket). In an umbrella trial, a central infrastructure is used to conduct multiple enrichment subtrials of molecularly targeted drugs with companion diagnostics, with a study design that usually permits flexibility to add, drop or modify subtrials as evidence from the trial accumulates or emerges elsewhere.

V. CLINICAL EVALUATION OF COMPANION DIAGNOSTICS

A. Treatment by Biomarker Interaction

In the biomarker stratified design, an observed statistically significant interaction between treatment and biomarker on an outcome can be supportive of the biomarker test being clinically useful as a companion diagnostic, but by itself may be insufficient. Examples can be constructed where the same level of interaction can arise for a biomarker that is clearly useful for treatment decision making and for another biomarker that is clearly not^[50–52].

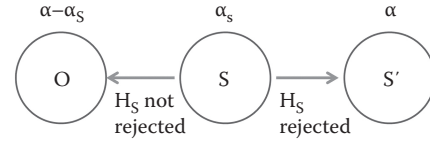
B. Time-to-Event Analysis

When the outcome is time to an event such as progression or death, Kaplan-Meier plots may be used to compare biomarker test negative and test positive subsets on the difference between treatment and control groups in risk of the event at specified follow-up times. For a biomarker test providing a quantitative measurement or continuous score value, the overall utility of the test as a companion to a treatment may be assessed by comparing treatment and control groups on the predictiveness curve, a plot of risk as a function of quantile in biomarker test value^[37].

C. Overall and Biomarker Subset Analyses

In biomarker-stratified trials, biomarker test positive patients are typically anticipated to be more likely to receive benefit from the treatment. However, the treatment effect may be of interest in the overall study population as well as in subsets defined by the biomarker test. A statistically significant treatment effect overall may be homogeneous across the biomarker-defined subsets or may be driven by a large treatment effect in the biomarker test positive subset S . In the latter situation, common sense would indicate that the treatment should not necessarily be given to everyone despite the overall treatment effect being statistically significant^[53].

To determine if a treatment should be given overall or just to a biomarker defined subset, Freidlin et al.^[54] propose a marker sequential test (MaST) procedure. The procedure consists of testing null hypotheses H_O for treatment effect in the overall study population O , H_S for treatment effect in biomarker test positive subset S , and $H_{S'}$ for treatment effect in biomarker negative subset $S' = O - S$, controlling the overall type 1 error rate at significance level α among these three hypotheses (Fig. 2). First, H_S is tested at level $\alpha_S < \alpha$. If H_S is rejected, then α_S is reused to test $H_{S'}$ at full level α . If H_S and $H_{S'}$ are both rejected, then the MaST procedure says the treatment can be given to everyone. If $H_{S'}$ is not rejected, then the MaST procedure says the treatment is indicated only for biomarker test positive subset S with the biomarker test indicated as a companion diagnostic for



O: overall population; S: test positive subset; S': test negative subset

If H_S rejected at α_S ,
then claim for subset S is met;

If $H_{S'}$ also rejected at α , overall claim met.
else if H_O rejected at $\alpha - \alpha_S$, overall claim met.

Fig. 2 Marker sequential test (MaST) procedure^[54]

selecting this subset for treatment. If H_S is not rejected, then H_O is tested at level $\alpha - \alpha_S$ as a Bonferroni correction for multiple testing. If H_O is rejected, then the MaST procedure says the treatment can be given to everyone.

D. Impact of Biomarker Analytical Accuracy on Clinical Performance

The analytical accuracy of a companion diagnostic can be used to inform on its capacity to predict treatment response. For example, the performance of the absence or presence of a biomarker $S = 0, 1$ to predict non-response or response $R = 0, 1$ to a treatment can be defined as

$$\delta = p_1 - p_0,$$

the difference in probability of response $p_s = \Pr(R = 1 | S = s)$ between patients with the biomarker ($S = 1$) and those without it ($S = 0$). Suppose an imperfect test is used to obtain the biomarker status as $S^* = 0, 1$ and its predictive values for the biomarker are $\pi_t = \Pr(S = 1 | S^* = t)$, $t = 0, 1$. Assume that any measurement error by the test is *non-differential* to response status, that is, $S^* | S, R = S^* | S$. Equivalently, $R | S, S^* = R | S$, i.e., given the true biomarker status, the test is non-informative for response. Then the performance of the test to predict response is

$$\delta^* = \delta a,$$

the performance δ of the biomarker attenuated by the factor

$$a = \pi_1 - \pi_0 = PPV + NPV - 1$$

where $PPV = \pi_1$ and $NPV = 1 - \pi_0$ are the positive and negative predictive values of the test^[55].

E. Causal Effect Estimands

The causal effect of a treatment within a biomarker-defined subset is considered fundamental to evaluating a biomarker test as a companion diagnostic to the treatment^[56–60].

Causal effects require consideration of the counterfactual outcome that would have occurred if a patient on treatment had instead been on control (or vice versa). Fortunately, in randomized trials, potential outcomes are independent of treatment assignment, permitting causal treatment effects to be estimated from the observed data alone^[61–62].

To evaluate a biomarker test, Simon^[60] proposes causal estimands for sensitivity, specificity, and predictive value. The positive predictive value (PPV) is the probability that time to event is longer for a test positive patient on treatment than if she were on control. The negative predictive value (NPV) is the probability that time to event is longer for a test negative patient on control than on treatment. Under assumptions permitting causal effect estimation, Simon derives estimands of these quantities as simple functions of proportional hazard ratios for biomarker test negative and positive patients and also for other non-proportional hazards models. Applying Bayes Theorem to the predictive values, true and false positive fractions (sensitivity, $1 - \text{specificity}$) can be derived as the probability of testing positive for the biomarker given time to event is, respectively, longer or shorter on treatment than on control.

F. Bayesian Methods

Because of their ability to draw inferences from complicated study designs and on complicated estimands, Bayesian methods can be useful for evaluating biomarker-defined subsets as companion diagnostics^[46,63–64]. Lunceford^[63] uses Bayesian methods to correct for disproportionate enrichment of the biomarker subset by viewing it as a missing data problem. Schnell et al.^[64] use Bayesian methods to determine a pair of credible subgroups in whom the probability of the individualized treatment effect being clinically significant is large or small, respectively.

G. General Considerations

Practical considerations have been proposed for deciding when a treatment should be given to everyone or to just a biomarker-defined subset based on a biomarker-stratified, all-comers trial^[51–52,65]. Broadly applicable statistical procedures for developing, calibrating, evaluating and comparing marker-guided treatment decision rules have been developed^[66]. Many other contributions, not cited here, have been made to the statistical evaluation of predictive biomarker tests.

VI. MISSING TEST RESULTS

For trials of treatments with companion diagnostic tests, test results may be missing in some patients. A sample may be unavailable or unevaluable. A patient may not have given consent to test their sample. The missing data mechanism is simplified if a sample is taken and consent is given to

test the sample prior to treatment assignment^[36]. With this design, the probability of a missing test value is independent of treatment assignment, increasing confidence that the observed treatment effect in the trial is internally valid among samples with non-missing test results.

Test results may be uninterpretable or equivocal. An uninterpretable test value may be considered not as missing but as a separate category of test result unless in practice the test will be repeated until a valid test result is obtained^[67]. An equivocal test result is a non-missing test value that is neither negative nor positive and therefore by design is a separate category of test result. Equivocal test results are sometimes called indeterminate.

When meaningful for analysis, imputation of missing test results can simplify likelihood-based inferences and increase generalizability compared with inferences based on patients with complete data. If test results are assumed to be missing at random, they may be imputed based on variables that could affect the test result. Variables that are most predictive of the missing test result are not necessarily patient and disease characteristics or even outcome, but characteristics of the sample, its handling, and its processing. Robustness of inferences to missing test data can be considered in a sensitivity or tipping point analysis^[68].

In retrospective studies, samples are archived and not tested until later with the companion diagnostic test. Archiving may affect the test result. A test result that is valid on a fresh sample may become invalid, uninterpretable or experience drift after the sample is stored for a period of time. If a test is to be applied to stored samples, analytical validity of the test on stored samples is important to characterize.

For many reasons the distribution of available test results on archived samples may be distorted relative to the distribution in fresh samples (e.g., tumors with larger volume may be overrepresented). This distortion may limit the generalizability of treatment effects observed in retrospective studies of archived samples.

In many co-development programs of a therapy and its companion diagnostic (CDx), an enrichment trial for the therapy is conducted before the CDx is ready. Because the CDx is unavailable, a clinical trial assay (CTA) is used to screen patients for the trial. The CTA and CDx are usually based on the same technology, but nonetheless can be discordant for some patients. To “bridge” the results of the trial from the CTA to the CDx, samples are saved for re-testing by the CDx when it becomes available. For samples that are unavailable or unevaluable in re-testing, the CDx test result is missing. For patients who are CDx positive but not enrolled into the trial because they were CTA negative, outcome data are missing. Performance goals and estimation strategies for bridging studies have been proposed^[69–70]. In one approach, if variation between the CTA and CDx is similar to variation within the CTA in repeated testing of the same subject, then the two tests could be considered interchangeable and expected to perform similarly.

VII. FOLLOW-ON COMPANION DIAGNOSTIC TESTS

After the first companion diagnostic for a treatment has been validated and granted marketing approval (by the US FDA or another regulatory agency), other companion diagnostics may be developed for the same indication. Unfortunately, a randomized clinical trial for a follow-on companion diagnostic may not be possible to conduct. Clinical equipoise is generally not present to randomize patients to the treatment or a control therapy because the approved companion diagnostic already indicates whether the patient should or should not receive the treatment.

In an attempt to infer whether a follow-on companion diagnostic will have clinical performance similar to an approved companion diagnostic for the same indication, the follow-on may be evaluated for concordance with the approved companion diagnostic^[71]. If the follow-on and the approved test have results that are concordant in a very high proportion of patient samples, then clinical performance for the follow-on could be expected to be similar to the approved test. The appropriateness of such a conclusion hinges upon the assumption that concordance study samples and patients are transportable to the clinical trial in which the approved test was evaluated.

Ideally, a concordance study of the follow-on with the approved companion diagnostic is conducted on samples from the randomized trial in which the approved companion diagnostic was evaluated. The clinical samples could be re-tested with the follow-on provided archiving of the samples is not detrimental to the quality of the follow-on test result. For biomarker-stratified all-comers trials, re-testing of the clinical samples would permit evaluation of clinical efficacy of the follow-on. For enrichment trials, re-testing with the follow-on may be conducted on all screened samples. However, only patients enrolled into the trial, i.e., those with a sample that passed the appropriate enrichment criterion (e.g., biomarker positive according to the approved test) would have clinical outcome information on which to evaluate clinical efficacy of the follow-on re-test result.

VIII. CONCLUSION

For many diagnostic tests, clinical validation consists of evaluating the association of the diagnostic test result with a current or future health condition in an observational study of diagnostic accuracy. In contrast, a companion diagnostic is typically evaluated for clinical utility in a randomized clinical trial of the corresponding therapeutic. Companion diagnostics have direct clinical consequences because they are intended for the safe and effective use of a corresponding therapeutic product. In the foreseeable future, companion diagnostics will continue to be crucial

for individualized safe and effective treatment of cancer and other diseases in efforts to improve health care through precision medicine.

REFERENCES

1. US FDA. In Vitro Companion Diagnostic Devices, US FDA: Silver Spring, MD, 2014, <http://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm262327.pdf> (accessed March 2017).
2. US FDA. Principles for Codevelopment of an In Vitro Companion Diagnostic Device with a Therapeutic Product. US FDA: Silver Spring MD, 2016. <https://www.fda.gov/ucm/groups/fdagov-public/@fdagov-meddev-gen/documents/document/ucm510824.pdf> (accessed March 2017).
3. Wang, L.; McLeod, H.L.; Weinshilboum, R.M. Genomics and drug response. *New Engl J Med* **2011**, *364* (12), 1144–1153.
4. Yamauchi, H.; Stearns, V.; Hayes, D.F. When Is a Tumor Marker Ready for Prime Time? A Case Study of c-erbB-2 as a Predictive Factor in Breast Cancer. *J Clin Oncol* **2001**, *19* (8), 2334–2356.
5. Paik, S.; Shak, S.; Tang, G.; Kim, C.; Baker, J.; Cronin, M.; Baehner, F.L.; Walker, M.G.; Watson, D.; Park, T.; Hiller, W.; Fisher, E.R.; Wickerham, D.L.; Bryant, J.; Wolmark, N. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New Engl J Med* **2004**, *351* (27), 2817–2826.
6. van-de-Vijver, M.J.; He, Y.D.; van't Veer, L.J.; Dai, H.; Hart, A.A.M.; Voskuil, D.W.; Schreiber, G.J.; Peterse, J.L.; Roberts, C.; Marton, M.J.; Parrish, M.; Atsma, D.; Witteveen, A.; Glas, A.; Delahaye, L.; van der Velde, T.; Bartelink, H.; Rodenhuis, S.; Rutgers, E.T.; Friend, S.H.; Bernards, R. A gene expression signature as a predictor of survival in breast cancer. *New Engl J Med* **2002**, *347* (25), 1999–2009.
7. Deverka, P.; Messner, D.; Dutta, T. *Effectiveness Guidance Document: Evaluation of Clinical Validity and Clinical Utility of Actionable Molecular Diagnostic Tests in Adult Oncology*, Center for Medical Technology Policy: Baltimore, MD, 2013.
8. McShane, L.M.; Cavenagh, M.M.; Lively, T.G.; Eberhard, D.A.; Bigbee, W.L.; Williams, P.M.; Mesirov, J.P.; Polley, M.C.; Kim, K.Y.; Tricoli, J.V.; Taylor, J.M.G.; Shuman, D.J.; Simon, R.M.; Doroshow, J.H.; Conley, B.A. Criteria for the use of omics-based predictors in clinical trials. *Nature* **2013**, *502* (7471), 317–320.
9. Fridlyand, J.; Simon, R.M.; Walrath, J.C.; Roach, N.; Buller, R.; Schenkein, D.P.; Flaherty, K.T.; Allen, J.D.; Sigal, E.V.; Scher, H.I. Considerations for the successful co-development of targeted cancer therapies and companion diagnostics. *Nat Rev Drug Discov* **2013**, *12* (10), 743–755.
10. Matchar, D.B. Introduction to the Methods Guide for Medical Test Reviews. AHRQ Publication No. 12-EHC073-EF. Chapter 1 of Methods Guide for Medical Test Reviews (AHRQ Publication No. 12-EHC017). Agency for Healthcare Research and Quality: Rockville,

- MD; June 2012. <https://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=1088> (accessed March 2017).
11. Micheel, C.M.; Nass, S.J.; Omenn, G.S. (eds.) For the Institute of Medicine. *Evolution of Translational Omics: Lessons Learned and the Path Forward*. National Academies Press: Washington, DC, 2012.
 12. Moons, K.G.M.; Kengne, A.P.; Woodward, M.; Royston, P.; Vergouwe, Y.; Altman, D.G.; Grobbee, D.E. Risk prediction models, I: Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* **2012**, *98* (9), 683–690.
 13. Moore, H.M. Biospecimen reporting for improved study quality (BRISQ). *Cancer Cytopathol* **2011**, *119* (2), 92–101.
 14. Khleif, S.N.; Doroshow, J.H.; Hait, W.N. Development Report: Advancing the Use of Biomarkers in Cancer Drug AACR-FDA-NCI Cancer Biomarkers Collaborative Consensus. *Clin Cancer Res* **2010**, *16* (13), 3299–3318.
 15. Ransohoff, D.F.; Gourlay, M.L. Sources of bias in specimens for research about molecular markers for cancer. *J Clin Oncol* **2010**, *28* (4), 698–704.
 16. Simon, R.M.; Paik, S.; Hayes, D.F. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J Natl Cancer I* **2009**, *101* (21), 1446–1452.
 17. Pepe, M.S.; Feng, Z.; Janes, H.; Bossuyt, P.M.; Potter, J.D. Pivotal Evaluation of the Accuracy of a Biomarker Used for Classification or Prediction: Standards for Study Design. *J Natl Cancer I* **2008**, *100* (20), 1432–1438.
 18. Simon, R. A Checklist for Evaluating Reports of Expression Profiling for Treatment Selection. *Clin Adv Hematol Oncol* **2006**, *4* (3), 219–224.
 19. McShane, L.M.; Altman, D.G.; Sauerbrei, W.; Taube, S.E.; Gion, M.; Clark, G.M. REporting recommendations for tumor MARKer prognostic studies (REMARK). *Nat Rev Clin Oncol* **2005**, *23* (36), 416–422.
 20. Bossuyt, P.M.; Reitsma, J.B.; Bruns, D.E.; Gatsonis, C.A.; Glasziou, P.P.; Irwig, L.M.; Lijmer, J.G.; Moher, D.; Rennie, D.; de Vet, H.C.W.; For the STARD Group. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Clin Chem* **2003**, *49* (1), 1–6.
 21. FDA-NIH Biomarker Working Group. BEST (Biomarkers, Endpoints, and other Tools) Resource [Internet]. Silver Spring (MD): Food and Drug Administration (US); 2016-. Glossary. 2016 Jan 28 [Updated 2016 Dec 22]. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK338448/>. Co-published by National Institutes of Health (US), Bethesda, MD.
 22. Rifai, N.; Gillette, M.A.; Carr, S.A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol* **2006**, *24* (8), 971–983.
 23. Beaver, J.A.; Tzou, A.; Blumenthal, G.M.; McKee, A.E.; Kim, G.; Pazdur, R.; Philip, R. An FDA Perspective on the Regulatory Implications of Complex Signatures to Predict Response to Targeted Therapies. *Clin Cancer Res* **2017**, *23* (6), 1368–1372.
 24. Dudoit, S.; Fridlyand, J.; Speed, T.P. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *J Amer Statist Assoc* **2002**, *97* (457), 77–87.
 25. Dudoit, S.; Fridlyand, J. Classification in microarray experiments. In *Statistical Analysis of Gene Expression Microarray Data* (Ed. Speed T). Chapman & Hall / CRC: Boca Raton, FL, 2003, 93–158.
 26. McDermott, J.E.; Wang, J.; Mitchell, H.; Webb-Robertson, B.; Hafen, R.; Ramey, J.; Rodland, K.D. Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data. *Expert Opin Med Diagn* **2013**, *7* (1), 37–51.
 27. Ambrose, C.; McLachlan, G.J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences* **2007**, *99* (10), 6562–6566.
 28. Simon, R.; Radmacher, M.D.; Dobbin, K.; McShane, L.M. Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. *J Natl Cancer I* **2003**, *95* (1), 14–18.
 29. Molinaro, A.M.; Simon, R.; Pfeiffer, R.M. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **2005**, *21* (15), 3301–3307.
 30. Oberg, A.; Vitek, O. Statistical design of quantitative mass spectrometry-based proteomic experiments. *J Proteome Res* **2009**, *8* (5), 2144–2156.
 31. Fryback, D.G.; Thornbury, J.R. The Efficacy of Diagnostics Imaging. *Med Decis Making* **1991**, *11* (2), 88–94.
 32. CLSI. *A Framework for Utilizing CLSI Guidelines to Evaluate Clinical Laboratory Measurement Procedures*; 2nd ed; CLSI report EP19. Clinical and Laboratory Standards Institute; Wayne, PA, 2015.
 33. CLSI. *Harmonization of Symbolology and Equations*. CLSI report EP36. Clinical and Laboratory Standards Institute; Wayne, PA, 2015.
 34. Buyse, M.; Michiels, S.; Sargent, D.J.; Grothey, A.; Matheison, A.; de Gramont, A. Integrating biomarkers in clinical trials. *Expert Rev Mol Diagn* **2011**, *11* (2), 171–182.
 35. Freidlin, B.; McShane, L.M.; Korn, E.L. Randomized clinical trials with biomarkers: Design issues. *J Natl Cancer I* **2010**, *102* (3), 152–160.
 36. Simon, R. Stratification and partial ascertainment of biomarker value in biomarker driven clinical trials. *J Biopharm Stat* **2014**, *24* (5), 1011–1021.
 37. US FDA. *Guidance on Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products*. US FDA: Silver Spring, MD, 2012.
 38. Mandrekar, S.J.; Sargent, D.J. Clinical Trial Designs for Predictive Biomarker Validation: Theoretical Considerations and Practical Challenges. *J Clin Oncol* **2009**, *27* (24), 4027–4034.
 39. Bossuyt, P.M.; Lijmer, J.G.; Mol, B.W. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet* **2000**, *356* (9244), 1844–1847.
 40. Simon, R. Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Personalized Medicine* **2010**, *7* (1), 33–47.
 41. Simon, R.; Maitournam, A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin Cancer Res* **2004**, *10* (20), 6759–6763.
 42. Buyse, M.; Michiels, S. Omics-based clinical trial designs. *Curr Opin Oncol* **2013**, *25* (3), 289–295.
 43. Jiang, W.; Freidlin, B.; Simon, R. Biomarker-adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect. *J Natl Cancer I* **2007**, *99* (13), 1036–1043.

44. Barker, A.D.; Sigman, C.C.; Kelloff, G.J.; Hylton, N.M.; Berry, D.A.; Esserman, L.J. I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy. *Clin Pharmacol Ther* **2009**, *86* (1), 97–100.
45. Freidlin, B.; Jiang, W.; Simon, R. The cross-validated adaptive signature design. *Clin Cancer Res* **2010**, *16* (2), 691–698.
46. Karuri, S.W.; Simon, R. A two-stage Bayesian design for codevelopment of new drugs and companion diagnostics. *Statist Med* **2012**, *31* (10), 901–914.
47. Baker, S.G.; Kramer, B.S.; Sargent, D.J.; Bonetti, M. Biomarkers, subgroup evaluation, and clinical trial design. *Discov Med* **2012**, *13* (70), 187–192.
48. Rosenblum, M.; Luber, B.; Thompson, R.E.; Hanley, Group sequential designs with prospectively planned rules for subpopulation enrichment. *Statist Med* **2016**, *35* (21), 3776–3791.
49. Mandrekari, S.J.; Dahlberg, S.E.; Simon, R. Improving clinical trial efficiency: thinking outside the box. *Am Soc Clin Oncol Educ Book* **2015**, *35*, e141–e147.
50. Gunter, L.; Zhu, J.; Murphy, S. Variable selection for qualitative interactions. *Stat Methodol* **2011**, *8* (1), 42–55.
51. Janes, H.; Pepe, M.S.; Bossuyt, P.M.; Barlow, W.E., Measuring the performance of markers for guiding treatment decisions. *Ann Intern Med* **2011**, *154* (4), 253–259.
52. McShane, L.M.; Polley, M.Y. Development of omics-based clinical tests for prognosis and therapy selection: the challenge of achieving statistical robustness and clinical utility. *Clin Trials* **2013**, *10* (5), 653–665.
53. Rothmann, M.D.; Zhang, J.J.; Lu, L.; Fleming, T.R. Testing in a prespecified subgroup and the intent-to-treat population. *Drug Information J* **2012**, *46* (2), 175–179.
54. Freidlin, B.; Korn, E.L.; Gray, R. Marker sequential test (MaST) design. *Clin Trials* **2014**, *11* (1), 19–27.
55. Pennello, G.A. Analytical and clinical evaluation of biomarkers assays: when are biomarkers ready for prime time? *Clin Trials* **2013**, *10* (5), 666–676.
56. Dunn, G.; Emsley, R.; Liu, H.; Landau, S. Integrating biomarker information within trials to evaluate treatment mechanisms and efficacy for personalised medicine. *Clin Trials* **2013**, *10* (5), 709–719.
57. Janes, H.; Pepe, M.S.; McShane, L.M.; Sargent, D.J.; Heagerty, P.J. The fundamental difficulty with evaluating the accuracy of biomarkers for guiding treatment. *J Natl Cancer I* **2015**, *107* (8), 1–4.
58. Huang, Y.; Gilbert, P.B.; Janes, H. Assessing treatment-selection markers using a potential outcomes framework. *Biometrics* **2012**, *68* (3), 687–696.
59. Sitlani, C.M.; Heagerty, P.J. Analyzing longitudinal data to characterize the accuracy of markers used to select treatment. *Statist Med* **2014**, *33* (17), 2881–2896.
60. Simon, R. Sensitivity, specificity, PPV, and NPV for predictive biomarkers. *J Natl Cancer I* **2015**, *107* (8), 153–156.
61. Rubin, D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Edu Psychol* **1974**, *66* (5), 688–701.
62. Rosenbaum, P.R. From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *J Amer Statist Assoc* **1984**, *79* (385), 41–48.
63. Lunceford, J.K. Clinical utility estimation for assay cutoffs in early phase oncology enrichment trials. *Pharm Stat* **2015**, *14* (3), 233–241.
64. Schnell, P.M.; Tang, Q.; Offen, W.W.; Carlin, B.P. A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Statist Med* **2016**, *72* (4), 1026–1036.
65. Millen, B.A.; Dmitrienko, A.; Ruberg, S.; Shen, L. A statistical framework for decision making in confirmatory multipopulation tailoring clinical trials. *Drug Information J* **2012**, *46* (6), 647–656.
66. Matsouaka, R.A.; Li, J.; Cai, T. Evaluating marker-guided treatment selection strategies. *Biometrics* **2014**, *70* (3), 489–499.
67. Begg, C.B.; Greenes, R.A.; Iglewicz, B. The influence of uninterpretability on the assessment of diagnostic tests. *J Chronic Dis* **1986**, *39* (8), 575–584.
68. Denne, J.S.; Pennello, G.; Zhao, L.; Chang, S.C.; Althouse, S. Identifying a subpopulation for a tailored therapy: Bridging clinical efficacy from a laboratory-developed assay to a validated in vitro diagnostic test kit. *Statist Biopharm Res* **2014**, *6* (1), 78–88.
69. Li, M. Statistical consideration and challenges in bridging study of personalized medicine. *J Biopharm Stat* **2015**, *25* (3), 397–407.
70. Bu, Y.; Zhou, X.H. Statistical evaluation of drug efficacy for bridging study in companion diagnostic test trials. *J Biopharm Stat* **2016**, *26* (6), 1118–1124.
71. Sharma, A.; Zhang, G.; Aslam, S.; Yu, K.; Chee, M.; Palma, J.F. Novel approach for clinical validation of the cobas KRAS mutation test in advanced colorectal cancer. *Mol Diagn Ther* **2016**, *20* (3), 231–240.

Comparability Studies: Statistical Considerations

Jason J.Z. Liao

Merck & Co., Inc., North Wales, Pennsylvania, U.S.A.

Li He and Robert Capen

Merck & Co., Inc., West Point, Pennsylvania, U.S.A.

Abstract

The purpose of this entry is to describe some statistical approaches that might be considered when designing and analyzing comparability studies for biological products or processes. In its most basic form, a comparability study seeks to compare a “test product” to a “reference product.” This is accomplished through evaluation of the extensive data obtained on both the test and reference products from analytical and biological assays as well as from nonclinical and clinical testing. To conclude two products or processes are comparable it is first necessary to define a standard for demonstrating comparability. Without a precise, quantitative definition that is accepted by regulatory authorities and the scientific community, it is impossible to prove conclusively that two products (or processes) are indeed comparable. But defining such a standard is not straightforward and often relies on expert opinion that, ideally, is driven by quality considerations, regulatory guidance and clinical safety or efficacy concerns. The usual bioequivalence approach that is appropriate for generics is not acceptable for biological products or processes due to their complexity. Regulatory guidance recommends following a step-wise approach utilizing a totality-of-evidence strategy, which requires collaboration among the pertinent experts, including statisticians. The demonstration of comparability does not necessarily imply that the test and reference products are identical but does ensure that any difference or change between them is not considered scientifically or clinically relevant.

Keywords: Clinical endpoint; Critical quality attribute; Equivalence margin; Equivalence testing; PK/PD; Plausibility interval; Prediction interval; Quality range; Tolerance interval; Totality-of-evidence.

INTRODUCTION

A comparability study is simply a study that seeks to compare two or more things. For example, a process modification is made and the post-change process is compared to the pre-change process, a replacement analytical method is validated and compared to the existing method, a new drug product image is developed and compared to a previously filed image, a biosimilar product is produced and compared to the reference product, etc. Each of these involves a comparison of a “test product” (post-manufacturing change, replacement assay, new image, biosimilar) to the “reference product” (pre-manufacturing change, existing assay, filed image, innovator). In this entry, “comparability” is defined in a broad sense to mean “(bio)similarity”, “bridging”, “equivalence”, etc. The demonstration of comparability does not necessarily mean that the quality attributes and other key parameters of interest of the reference and test product are identical, but that they are highly similar and that the existing knowledge is sufficiently predictive to ensure that any differences in quality attributes and the key parameters of interest have no adverse impact upon safety or efficacy

of the drug product. Determination of comparability can be based solely on quality considerations if the manufacturer can provide sufficient assurance of comparability through analytical studies related to quality attributes. Additional evidence from nonclinical or clinical studies is considered appropriate and required when quality data are insufficient to establish comparability.

Improvements are commonly applied to biological products in production methods, process and control test methods for product characterization, or changes in equipment or facilities both during development and after approval. There are three major reasons that a biological product manufacturer may seek to make changes to an existing process or product:

- **Competitiveness.** Improving product quality, yield, and stability as well as increasing manufacturing scale and enhancing customer experience helps the manufacturer maintain a competitive advantage.
- **Innovation.** Enhancements in manufacturing processes and test methods bring important and improved products to the market more efficiently and rapidly.

- Compliance. Regulatory agencies acknowledge that product and process changes are necessary for the biotechnology industry to advance. A manufacturer must stay in compliance as regulatory requirements evolve.

Similar biological products (“biosimilars”) may help reduce health care costs and increase accessibility to patients. A biosimilar is not an identical or “generic” copy of the reference product. It is, however, highly similar to the reference product, notwithstanding minor differences in clinically inactive components. To develop a biosimilar, a manufacturer may need to create a totally new process. It can even take highly experienced companies many years to devise the technologies necessary to cultivate and genetically engineer the living cells that produce the proteins and antibodies that are filtered and refined into successful biological drugs. Regardless of whether an existing process is updated or a new process is created for the purpose of developing a biosimilar, it is a regulatory requirement for the sponsor to demonstrate comparability between the test and reference products.

The purpose of this entry is to describe some statistical approaches that might be considered when designing and analyzing both analytical and clinical studies to compare a test product to a reference product. Ideally a comparability study would require data to be collected proactively through a designed experiment. When that is not feasible, historical data can be collected and evaluated instead. For example, only historical data are likely to be available on batches tested prior to making a change in the manufacturing process. The nature of the comparison depends on the relevance of the attribute being studied as well the type of data generated. On one extreme, a clinical study requires pre-specification of a number of important characteristics, including, sample size, power, confidence level, acceptance criteria, randomization plan, etc. in conjunction with a formal statistical test of hypothesis or equivalence. On the other extreme, for some analytical attributes a graphical comparison of chromatograms generated on some of the reference and test batches might be all that is required. Regulatory guidance recommends following a step-wise approach for demonstrating comparability using data obtained from analytical and biological assays as well as from nonclinical and clinical testing based on a totality of evidence strategy. But how exactly is comparability demonstrated? Without a clear, objective and scientifically relevant definition of what it means to be “comparable,” there is no way to prove conclusively that two products or processes are indeed comparable. In 1992 the FDA adopted the “80–125%” standard for clinical testing that is still applied today, namely, the 90% confidence interval for the ratio of bioavailability measures must fall entirely within 80%–125%. No such standard exists for non-clinical testing. Simply translating the clinical standard into a non-clinical setting (e.g., bioassay testing) is not a practical solution. This may seem obvious but it is nevertheless a point worth stating.

This entry is divided into two general categories: Analytical and Clinical. While the discussion focuses primarily on statistical aspects, Statistics is just one component of the overall evaluation strategy. Input is also required from other key areas including regulatory, quality, CMC, etc. Data from animal studies and pre-clinical potency testing may also be necessary to assess comparability completely but will not be discussed. An overall summary that also includes some notions on how a biologics manufacturer might objectively implement a “totality of evidence” strategy is provided in the Conclusions section.

EVALUATING THE ANALYTICAL COMPARABILITY OF A BIOLOGICAL PRODUCT

General Considerations

A biological product is defined as “a virus, therapeutic serum, toxin, antitoxin, vaccine, blood, blood component or derivative, allergenic product, or analogous product, or arsphenamine, or derivative of arsphenamine or any other trivalent organic arsenic compound, applicable to the prevention, treatment or cure of a disease or condition of human beings”.^[1] Biological products, like other drugs, are used for the treatment, prevention or cure of disease in humans. Biologics were 7 of the top 10 bestselling drugs in recent years including 6 mAbs.^[2–4] They are usually produced by living cells and are generally large complex molecules, which are more difficult to characterize by nature, and can have significant heterogeneity. In addition to structural complexity, the source materials and manufacturing processes for protein and antibodies may lead to a variety of product variants and impurities that can impact product efficacy and safety. Manufacturing biologics is a complicated process. A biologic manufacturing process depends on many steps of production and purification processes which can influence the biological and clinical properties. These sophisticated steps are monitored by sensitive analytical methods, which need to be adapted to specific biological products. Therefore, it is the process that defines a biological product.

Regulatory

The regulatory science on which comparability is based has subsequently been applied to the review and approval of biosimilars.^[5] Similar Biologics Medicinal Products (“biosimilars”) have been on the market in the European Union since early 2006. In the United States, the FDA prefers to use the term “comparability” for changes made by a single manufacturer and “similarity” for applying this same process to biosimilars. By contrast, in Europe, the European Medicines Agency uses the term “comparability” when evaluating both inter- as well as intra-manufacturing changes. More recently, the European Medicines Agency has used the expression “biosimilar comparability” to clarify the context

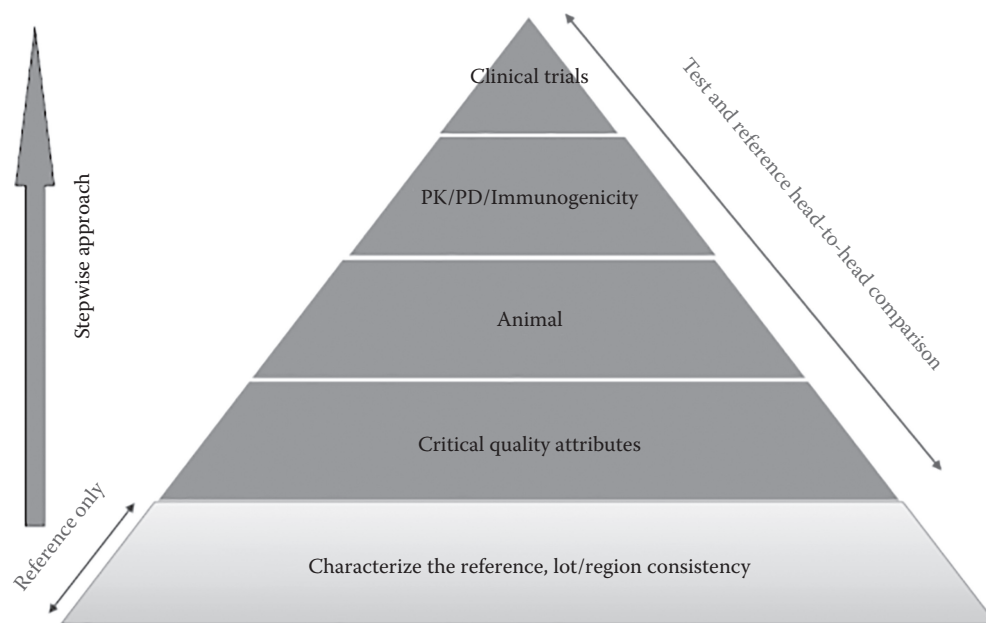


Figure 1 Stepwise diagram for comparability studies

but not to change the concept as a scientific matter.^[6] However, these terms are used interchangeably here. A comparability study is formally defined as a quantitative assessment of the extent of similarity of drug substances and drug products made before and after a bioprocess change or between two products.^[7] The purpose of comparability studies is to ensure that the test product does not have any significant or clinically meaningful difference or changes that impact the material/product/process quality, safety, purity, and potency or efficacy compared to the reference product.^[7–12] The information submitted in the application for approval may include structural and functional characterization, nonclinical evaluation, human PK and PD data, clinical immunogenicity data, and clinical safety and efficacy data. This stepwise approach can be illustrated graphically as shown in [Figure 1](#), where (1) extensive characterization of the reference product using fingerprint-like techniques and (2) assessment of the variability among different reference batches, obtained from different regions and/or different manufacturing sites and manufactured over a broad window of time are completed to define the reference target product profile before the comparability study is performed.

Quality

Quality attributes (QA's) characterize a biological product in terms of its structural, physico-chemical and functional properties. Assessing analytical comparability involves the evaluation of data obtained from batch release testing, characterization testing, comparative stability studies and forced degradation studies. It is a key component in the overall comparison between a test product and a corresponding reference product. In view of the fact that manufacturers often test a large number of QA's, Tsong, Dong, et al.,^[13] recommend that the QA's be categorized into three tiers based on

their potential impact on product quality and clinical outcome. Whether or not the data generated on each QA are amenable to statistical analysis should also be considered when categorizing each assay. As shown in [Figure 2](#), the critical quality attributes (CQA's) are those assigned to Tier 1 and comparability should be statistically evaluated based on a comparison of means following an equivalence testing strategy. A “quality range” approach is recommended for the QA's assigned to Tier 2, and for QA's in Tier 3, a graphical comparison is sufficient. Evaluation of Tier 3 assay data will not be discussed in this entry.

Any strategy developed to characterize and compare the data must account for batch-to-batch variability since that is an important component of the overall variability in the data. This is especially true if batches are obtained

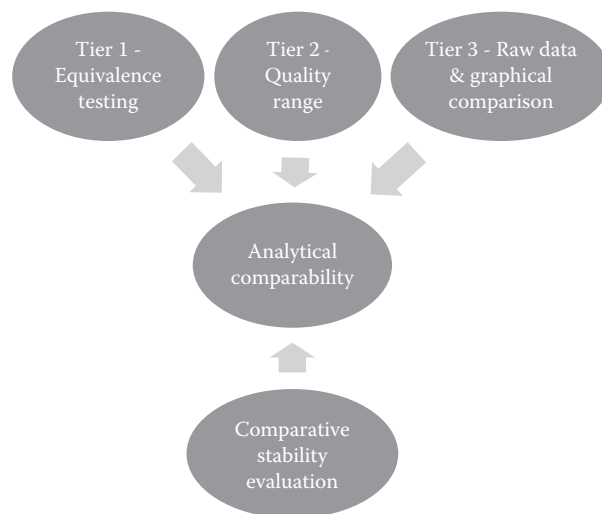


Figure 2 Investigational Strategies for Analytical Comparability Evaluation

from multiple regions (e.g., US and EU) or multiple manufacturing sites, are present in multiple images, or are of different ages. Selecting as many representative batches as possible and performing more assay runs for assays that have large variability (to define the reportable value) have direct impact on the relevance of the conclusions reached from the data analysis. This means that experience and judgment are necessary components to help guide the decision making process. However, because of manufacturing constraints, the number of available test product batches will usually be limited. When comparing a potential biosimilar to an innovator product, a manufacturer can purchase and test as many reference batches as necessary to generate an adequate sample size but it is not practical for the manufacturer to produce an equivalent number of test batches to achieve a balanced design. This is also true when a manufacturer is comparing batches produced after a manufacturing change to batches produced before the change. There may be a large number of existing pre-change batches that have data on the relevant attributes but likely only a few post-change batches will be available for testing. For comparative stability evaluation, the number of test or reference batches placed on stability will both be small, even under accelerated or stressed conditions.

Data Presentation

Some pharmacological data, such as potency or concentration, are generally right-skewed and have the characteristic that the standard deviation of the data increases proportionally to the mean. Standard statistical assumptions of normality and variance homogeneity are better satisfied if such data are first log transformed. Statistical calculations are performed on the transformed data and then converted back to the original scale. Conversion back to the original scale simply involves taking the anti-log of the metric calculated on the transformed data. Arithmetic means, standard deviations (SDs) and standard errors (SEs) become geometric means (GMs), geometric standard deviations (GSDs) and geometric standard errors (GSEs); “Mean $\pm$ SE” in the log scale becomes “GM $\times$ / GSE” in the original scale. It is recommended that the data are not rounded prior to statistical evaluation. The final results obtained from the statistical analysis should be rounded prior to presentation.

Before any formal comparison is performed, the collected data should be summarized using standard statistical approaches as described in Table 1. Unlike clinical endpoints, analytical data are heavily dependent on the assay, thereby limiting the interpretability of a 95% confidence interval (CI) as it relates to linking statistical difference or lack of difference to biological relevance. That is, a statistically significant difference does not necessarily imply the difference is biologically relevant. Nor does a conclusion of “not statistically significantly different” imply the compounds are biologically equivalent. Statistical findings must always be placed in the appropriate biological context.

Table 1 Recommended Data Summaries Prior to Formal Comparison

Data Source	Data Summary ^a
Reference	N ^b , Mean, SD ^c , 95% CI ^d , Box Plot
Test	N, Mean, SD, 95% CI, Box Plot
Reference + Test	95% CI for True Mean Difference

^aFor log-normal data perform analysis in log scale first and then convert back to the original scale; ^bSample size (number of batches); ^cStandard Deviation; ^dConfidence Interval

Tier 1 Evaluation: Statistical Equivalence Testing of Means

The measurements made on the reference product (X_R) are assumed to follow a normal distribution with mean μ_R and variance σ_R^2 , respectively. Likewise, the measurements made on the test product (X_T) are assumed to follow a normal distribution with mean μ_T and variance σ_T^2 , respectively. For a given equivalence margin, $\delta (> 0)$, the equivalence hypotheses can be stated as follows:

$$H_0 : |\mu_R - \mu_T| \geq \delta \text{ vs } H_1 : |\mu_R - \mu_T| < \delta \quad (1)$$

The most commonly used approach to testing H_0 is the two one-sided test (TOST) procedure.^[14–15] Recently equivalence testing has been expanded into much broader applications.^[16] Essentially TOST is a union-intersection test which decomposes the null hypothesis H_0 in Eq. 1 into two separate sub-null hypotheses $H_{01} : \mu_R - \mu_T \geq \delta$ and $H_{02} : \mu_R - \mu_T \leq -\delta$. Let ϕ_1 and ϕ_2 be the rejection regions, each at level α , for testing H_{01} and H_{02} , respectively. Then, by the union intersection principle, the rejection region for testing the hypothesis in Eq. 1 at level α is $\phi_1 \cup \phi_2$. More specifically, let $\bar{X}_R$, s_R^2 and n_R be the sample mean, sample variance and sample size for the reference product and let $\bar{X}_T$, s_T^2 and n_T be the sample mean, sample variance and sample size for the test product. Define

$$s_p^2 = \frac{(n_R - 1)s_R^2 + (n_T - 1)s_T^2}{n_R + n_T - 2} \text{ as the pooled sample variance.}$$

When it is feasible to assume that $\sigma_T^2 = \sigma_R^2$, the TOST procedure rejects the null hypothesis H_0 if

$$\frac{\bar{X}_R - \bar{X}_T + \delta}{s_p \sqrt{\frac{1}{n_R} + \frac{1}{n_T}}} \leq -t_\alpha \text{ or } \frac{\bar{X}_R - \bar{X}_T - \delta}{s_p \sqrt{\frac{1}{n_R} + \frac{1}{n_T}}} \geq t_\alpha \quad (2)$$

where t_α is the 100(1- α)th percentile of a T distribution with $n_R + n_T - 2$ degrees of freedom. When it is not feasible to assume $\sigma_T^2 = \sigma_R^2$, the test rejects H_0 in Eq. 1 if

$$\frac{\bar{X}_R - \bar{X}_T + \delta}{\sqrt{\frac{s_R^2}{n_R} + \frac{s_T^2}{n_T}}} \leq -t_\alpha^* \text{ or } \frac{\bar{X}_R - \bar{X}_T - \delta}{\sqrt{\frac{s_R^2}{n_R} + \frac{s_T^2}{n_T}}} \geq t_\alpha^* \quad (3)$$

where t_{α}^* is the $100(1-\alpha)\text{th}$ percentile of a T distribution with degrees of freedom approximated as

$$\left(\frac{s_R^2}{n_R} + \frac{s_T^2}{n_T} \right)^2 / \left(\frac{1}{n_R-1} \frac{s_R^2}{n_R} + \frac{1}{n_T-1} \frac{s_T^2}{n_T} \right) \quad (4)$$

Because of the close connection between hypothesis testing and confidence intervals, one expects that the TOST procedure is operationally identical to a procedure that compares some confidence interval for $\mu_R - \mu_T$ with the equivalence interval $(-\delta, \delta)$ and claims equivalence if this confidence interval is contained entirely within the equivalence interval. Westlake^[17] and Schuirmann^[18] noted one choice for such a confidence interval can be

$$(\bar{X}_R - \bar{X}_T) \pm t_{\alpha} \cdot s_p \sqrt{\frac{1}{n_R} + \frac{1}{n_T}} \text{ which has confidence level}$$

$(1-2\alpha)100\%$. Later Brown, Casella, et al.,^[19] called the relationship between the α level TOST and the above $(1-2\alpha)100\%$ confidence interval approach a coincidence. Berger and Hsu^[20] discussed examples where a $(1-2\alpha)100\%$ confidence interval approach is not a size α test, being either too liberal or too conservative, and also discussed cases where a $(1-\alpha)100\%$ confidence interval approach corresponds to an exact size α level TOST. These authors suggested that the practice of making a connection between a size α test and a $(1-2\alpha)100\%$ confidence interval be abandoned, and that $(1-\alpha)100\%$ confidence interval be reported instead when testing hypotheses of equivalence at a significance level of α . Although the above discussion has focused on symmetric equivalence intervals, equivalence testing can readily be extended to the asymmetric case whereby the equivalence interval is defined by $(-\delta_1, +\delta_2)$ for $\delta_1 > 0$, $\delta_2 > 0$ and $\delta_1 \neq \delta_2$. Only the symmetric case will be pursued in this entry.

Tier 1 Evaluation: Determination of the Equivalence Margin

The most crucial step in equivalence testing is the determination of the equivalence margin, δ , since the choice for it not only impacts the conclusion of the test but also the credibility of the study. Ideally δ would be prespecified and clinically meaningful or at least scientifically defensible. And while choosing a large value for δ increases the power to establish equivalence, it can also invite criticism from regulatory agencies unless such a choice is fully justified. In practice it is often quite challenging to provide a scientific basis for δ or link the choice to clinical efficacy endpoints or even Pharmacokinetic or pharmacodynamic (PK/PD) endpoints and consequently alternative approaches to determining δ need to be considered. As mentioned in the Introduction, it is common nowadays to choose $\log(1.25)$ for δ in bioequivalence studies in humans. In the context of a Tier 1 analytical evaluation, a one-size-fits-all approach is not practical since the impact

each CQA has on clinical efficacy or PK/PD endpoints, while unknown, is likely different. Also, the analytical methods used to evaluate the multiple CQA's possess different performance characteristics, such as different levels of precision. Before considering other approaches for determining δ , it should be emphasized that any statistically derived δ should be reviewed and agreed upon by appropriate subject matter experts before it is used in a statistical equivalence test.

The USP <1010>^[21] Approach

For comparing two analytical methods using statistical equivalence testing, Appendix E of USP <1010> describes an approach to the derivation of δ .^[21] This approach compares the specification limits for the quality attribute being measured with a [95/95] tolerance interval calculated based on the data generated by the reference method. In the context of comparing two procedures, the data from the reference or current procedure are used to generate the tolerance interval. In the context of assessing analytical comparability, the data from the reference product are used to derive the tolerance interval. Under the assumption of normality, a $[100(1-\alpha)/100p]$ tolerance interval (LTL , UTL) encloses $100p\%$ of the population of measurements with $100(1-\alpha)\%$ confidence.

Denote the lower and upper specification limits as LSL and USL respectively. The equivalence margin δ can be determined as the minimum of the two differences $LTL - LSL$ and $USL - UTL$, both of which need to be positive (see Figure 3). A mean shift of no more than δ implies a high probability that test product batches will be able to satisfy the same specification limits used to evaluate reference product batches. This method requires that specification limits be available and that the specification interval be wider than the 95/95 tolerance interval determined based on the reference data. When either of these conditions is not satisfied, the USP <1010>^[21] approach is not applicable. Because the width of a tolerance interval generally increases as the sample size decreases, this approach should only be used when there are a reasonable number of reference batches available for testing (say at least 20). This approach can easily be adapted to handle data from CQA's that only have a one-sided specification limit. In that case, a corresponding one-sided [95/95] tolerance bound would be constructed.



Figure 3 Illustration of the Derivation of $\delta = \min(A, B)$ using USP <1010> (Calculate $A = LTL - LSL$ for $LTL > LSL$ and calculate $B = USL - UTL$ for $USL > UTL$)

The Limentani, Ringo et al.,^[22] Approach

For a particular CQA, Limentani, Ringo et al.,^[22] proposed that the equivalence margin be chosen as $\delta = \delta_0 + s_R^* \left[t_{1-\alpha, 2n-2} + t_{1-\beta, 2n-2} \right] \frac{2}{\sqrt{n}}$, where δ_0 is an acceptable shift as determined by appropriate subject matter experts (if δ_0 is unknown it can be conservatively set to zero); n is the common sample size (number of batches) for both test and reference products that is fixed before the determination of δ ; $s_R^* = s_R \sqrt{\frac{n-1}{\chi_{\gamma, n-1}^2}}$ is the upper $100(1-\gamma)\%$

confidence bound for the true standard deviation σ_R associated with the reference product, $t_{1-\alpha, 2n-2}$ ($t_{1-\beta, 2n-2}$) is the upper $100(1-\alpha)\text{th}$ ($100(1-\beta)\text{th}$) percentile of the Student's T distribution with $2n-2$ degrees of freedom and $\chi_{\gamma, n-1}^2$ is the upper $100(1-\gamma)\text{th}$ percentile of the Chi-Squared distribution with $n-1$ degrees of freedom. Here α and β are the allowed Type I and Type II errors of the TOST procedure and therefore this recommended δ is closely related to the power of the test. If a point estimate, s_R , of σ_R is used in the definition of the δ instead of the upper confidence bound, s_R^* , then the corresponding TOST will have power equal to $1-\beta$, when the sample sizes for reference and test products are indeed equal. As δ is a function of the sample size n , it can be large when n is small.

The Tsong, Dong, et al.,^[13] Approach

The statistical determination of δ should follow a unified approach for all Tier 1 CQA's and allow for possibly different variabilities among the corresponding analytical methods. Tsong, Dong et al.,^[13] propose to use a sample size and variance adjusted margin that rewards large sample sizes but controls the power of the TOST procedure for small sample sizes. Similar to Limentani, Ringo et al.,^[22] they consider that δ be represented as $\delta_0 + k\sigma_R$, where $\delta_0 \geq 0$ is an acceptable shift as determined by subject matter experts ($\delta_0 = 0$ by default); $k > 0$ is some positive number and σ_R is the true standard deviation associated with the reference product. Let n be the common sample size (number of batches) for both reference and test products and assume $\sigma_T^2 = \sigma_R^2 = \sigma^2$, where σ_T represents the true standard deviation associated with the test product. If the true mean difference (i.e., true effect) to be detected is also defined as a multiple η of σ_R ($\eta > 0$) then for $\delta_0 = 0$ the power of the TOST procedure defined in Eq. 2 can be expressed as a function of n , α , k and η :

$$\begin{aligned} \text{Power} &= \beta(n, \alpha, \eta, k) \\ &= E_X \left[\Phi \left(\frac{k-\eta}{\sqrt{2/n}} - t_{\alpha, 2(n-1)} \sqrt{\frac{X}{2(n-1)}} \right) \right. \\ &\quad \left. - \Phi \left(\frac{-k-\eta}{\sqrt{2/n}} + t_{\alpha, 2(n-1)} \sqrt{\frac{X}{2(n-1)}} \right) \right] \end{aligned} \quad (5)$$

where the expectation is taken with respect to X which follows a Chi-squared distribution with $2(n-1)$ degrees of freedom. Having chosen a fixed $\eta = 1/8$, as well as fixed values for n and α , they consider determining k , or equivalently δ , by assigning, for each value of n , a value for power based on the function $f(n) = 1 - \exp(-0.53948 - 0.14694n + 0.00205n^2)$. The function $f(n)$ is determined based on the rationale that larger power should correspond to larger sample sizes but still be reasonable for small sample sizes. Specifically, based on discussions between FDA statisticians with FDA biologists, initial power values were assigned to five different sample sizes, ranging from 65% power for $n = 4$ to 95% power for $n = 25$.^[13] To generate a generally applicable power function, a monotonic exponential function was fit through the five values to obtain the function $f(n)$ described above. To illustrate the methodology, given $\eta = 1/8$ and $\alpha = 0.05$, for specific n , say $n = 10$, a value for k can be found so that the above expectation in Eq. 5 is equal to $f(10) = 0.84$. The process of finding k in this example can be illustrated using Figure 4 below.

Alternatively, one can fix k and report the corresponding value for α . This is their final proposed approach. Specifically, given $\eta = 1/8$ and $k = 1.5$, for a specific n , a value for α can be found so that the above expectation in Eq. 5 is equal to $f(n)$. This process is illustrated in Figure 5 below using the same inputs as above.

In situations where one or more of the three conditions, $\delta_0 = 0$, $\sigma_T^2 = \sigma_R^2 = \sigma^2$ or $n_T = n_R = n$, cannot be assumed, the power function $\beta(n, \alpha, \eta, k)$ in Eq. 5 can be adjusted

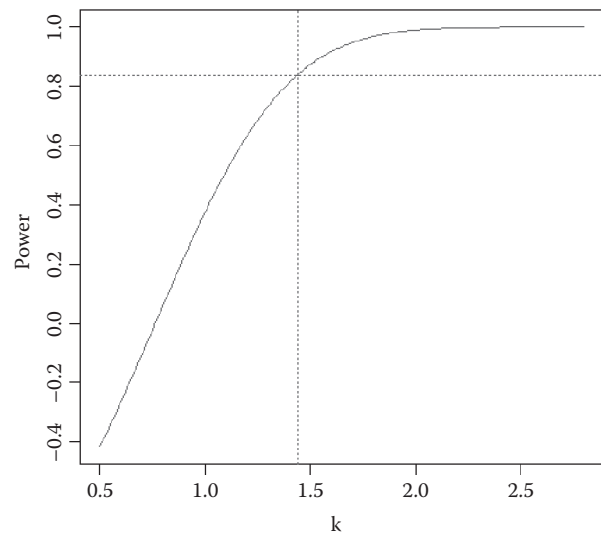


Figure 4 Illustration of the determination of k for fixed n , α , and η based on Tsong, Dong, et al.,^[13]. The blue solid curve is the graph of $\beta(n, \alpha, \eta, k)$ as a function of k , with $n = 10$, $\alpha = 0.05$ and $\eta = 1/8$. The horizontal red dotted line represents the power assigned to $n = 10$, namely $f(10)$, which equals 0.84 in this example. The vertical red dotted line represents the value of k where $\beta(n, \alpha, \eta, k) = f(10)$. In this example, $k = 1.4$

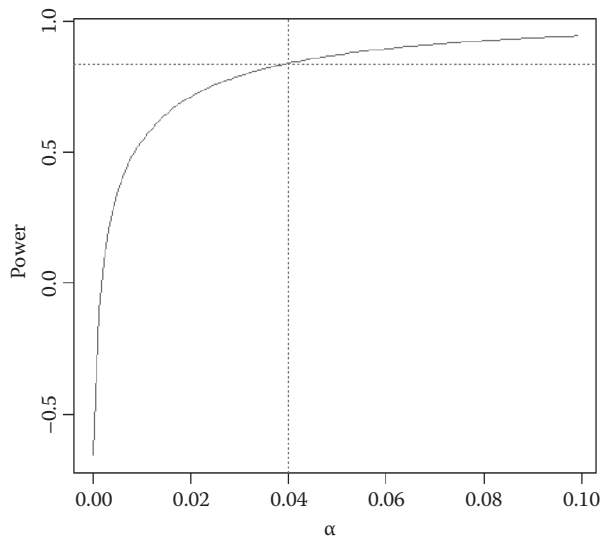


Figure 5 Illustration of the determination of α for fixed n, k , and η based on Tsong, Dong, et al.,^[13]. The blue solid curve is the graph of $\beta(n, \alpha, \eta, k)$ as a function of α , with $n = 10, k = 1.5$, and $\eta = 1/8$. The horizontal red dotted line represents the power assigned to $n = 10$, namely $f(10)$, which equals 0.84 in this case. The vertical red dotted line represents the value of α where $\beta(n, \alpha, \eta, k) = f(10)$. In this example, α equal 0.04

accordingly and the calculations described above would be based on the adjusted power function instead.

The Liao and Darken^[23] Approach

Chow^[24] pointed out some concerns regarding the approach recommended by Tsong, Dong et al.,^[13]. One main concern is that the model structure assumed that one observation was obtained per batch for both reference and test products. In fact, this assumption was also made in Limentani, Ringo, et al.,^[22] and by the USP <1010>^[21] approach for the determination of δ . Under this assumption, batch-to-batch variability is not estimable separately from analytical variability. In reality there is often heterogeneity among the batches and it is therefore important to separate between-batch variability from within-batch (or analytical) variability when estimating σ_R . Otherwise, σ_R will likely be underestimated resulting in overly narrow equivalence interval. Another concern is with regard to the recommended choice of $k = 1.5$ in their proposed approach, since no scientific justification for this particular selection was provided.

Liao and Darken^[23] and Liao^[25] considered the data structure where multiple observations are taken from multiple batches for both reference and test products. Specifically, let X_{ij} be the j th observation taken from the i th reference batch, $i = 1, \dots, m_X; j = 1, \dots, n_X$, with m_X being the number of reference batches and n_X being the number of observations obtained on each reference batch. Similarly, let Y_{ks} be the s th observation taken from the k th test product batch, $k = 1, \dots, m_Y; s = 1, \dots, n_Y$, with m_Y being the number

of test product batches and n_Y being the number of observations obtained on each test product batch. They considered the following random effect model for X_{ij} and Y_{ks} :

$$X_{ij} = \mu_X + \tau_X + \varepsilon_{ij} \text{ and } Y_{ks} = \mu_Y + \tau_Y + \delta_{ks} \quad (6)$$

where μ_X (μ_Y) is the overall mean for reference (test) product; $\tau_X \sim N(0, \sigma_{\tau_X}^2)$ and $\tau_Y \sim N(0, \sigma_{\tau_Y}^2)$ are random batch effects for reference and test product, respectively, with $\sigma_{\tau_X}^2$ and $\sigma_{\tau_Y}^2$ being the corresponding process or batch-to-batch variability; ε_{ij} is the random error term for the reference product, assumed to be identically and independently distributed as $N(0, \sigma_\varepsilon^2)$ and δ_{ks} is the random error term for the test product assumed to be identically and independently distributed as $N(0, \sigma_\delta^2)$. Here σ_ε^2 and σ_δ^2 represent the corresponding within-batch or assay variabilities.

It is assumed that (1) the ε_{ij} 's and δ_{ks} 's are independently distributed and (2) τ_X and τ_Y are independent of each other. Note that in this set-up the variability of any reference (or test) measurement contains two components: one representing batch-to-batch variability and the other within-batch or assay variability. The proposed acceptance criterion for assessing equivalence is derived based on the reasoning that any difference observed when comparing reference product to itself is a result of only random variation and not some underlying non-random cause. Such random variation can be quantified by a *plausibility interval* for the differences obtained from reference-to-reference comparison. Specifically, for $k > 0$, the plausibility interval is defined as

$$\left(-k \sqrt{2(\hat{\sigma}_{\tau_X}^2 + \hat{\sigma}_\varepsilon^2)}, +k \sqrt{2(\hat{\sigma}_{\tau_X}^2 + \hat{\sigma}_\varepsilon^2)} \right) \quad (7)$$

where $\hat{\sigma}_{\tau_X}^2$ is an estimate of the reference batch-to-batch variability and $\hat{\sigma}_\varepsilon^2$ is an estimate of the reference assay variability. Equivalency is claimed if (1) the calculated $[100(1 - \alpha)/100p]$ tolerance interval for the differences between the reference and test products fall within the pre-determined plausibility interval and (2) the estimated mean ratio of reference to test must be within some pre-specified range, for example (0.8, 1.25), as might be used in an average bioequivalence study. Based on a simulation study, the authors suggest that k in Eq. 7 can be chosen to be 3. Their simulation studies revealed that the requirement on the mean ratio is necessary since it adds protection against making a Type I error and that it is important to account for batch-to-batch variability as it plays a significant role in the power of the test. The assessment procedure developed here is suitable for both paired and non-paired data structures. However, the mathematical formula to calculate the plausibility interval is different depending on whether the data are paired or non-paired. The ability to design a comparability study to exploit a paired-data structure is advantageous since it reduces the total variability in the data by removing batch to batch variability resulting in a greater

probability of claiming equivalence when there is truly no difference between the test and reference products.

Tier 2 Evaluation: Quality Range

Analytical results for each test designated for Tier 2 analysis are evaluated using criteria based on constructing intervals defined by the assay results from the batches of reference product. Since methods for constructing such intervals often rely on the assumption of normality, the data should be plotted to confirm that this assumption is not seriously violated. For some attributes, the data may need to be transformed to better approximate a normal distribution. When severe departures from normality occur and none of the common transformations help alleviate the problem, alternative statistical methods (e.g., distribution-free methods) can be considered. A statistician should be consulted for advice.

Assume the measurements made on the reference product (X_R) follow a normal distribution with mean μ_R and variance σ_R^2 , respectively. Here σ_R represents the total standard deviation (batch-to-batch and within-batch) of the population. Following Tsong, Dong et al.,^[13] a probability interval or *quality range* for the reference is an interval of the form $\mu_R \pm f\sigma_R$ for some value f . If μ_R and σ_R are known, then about 95% of the population of measurements made on the reference batches will fall between $\mu_R \pm 2\sigma_R$ (95% coverage) and about 99.7% of the population of measurements will fall between $\mu_R \pm 3\sigma_R$ (99.7% coverage). There are a number of approaches used to estimate a quality range and the most common approaches are shown in Table 2. The choice opted for should be discussed with regulatory agencies prior to implementing it.

The Tolerance Interval Approach

As described previously, under the assumption of normality, a $[100(1-\alpha)/100p]$ tolerance interval (*LTL*, *UTL*) encloses $100p\%$ of the population of measurements with $100(1-\alpha)\%$ confidence. It is calculated as $\hat{\mu}_R \pm g_{(1-\alpha;p,n)}\hat{\sigma}_R$ where $g_{(1-\alpha;p,n)}$ is the tolerance factor and depends on the

coverage probability, confidence level and sample size, n , the number of reference batches. In most practical applications, $1-\alpha$ and p are typically chosen from the set of values {0.90, 0.95, 0.99}. Tables exist that provide exact values for $g_{(1-\alpha;p,n)}$ for various choices of p , α and n when the data arise from a simple random sample.^[26] For more complicated data structures, for example, nested data, a statistician should be consulted. In practice, assuming comparability, it is expected that the entire available test product batch data (reportable values) would fall within the tolerance interval constructed from the reference product batch data for each attribute. It is possible, though, that one batch might fall outside of the tolerance interval for a particular attribute. The implication of such a sporadic result would need to be discussed among the appropriate subject matter experts and any conclusion of comparability (or lack thereof) would be based on the totality of the evidence, not just this one result for one particular attribute. If two or more of the test product batches fall outside of the tolerance interval for a particular attribute, then this should be thoroughly investigated before an overall conclusion of comparability is made as this outcome would not be expected under the assumption of comparability. If the calculated tolerance interval is wider than the current release specification interval, then the specification interval should be used to gauge analytical comparability. A tolerance interval can be very wide when n is relatively small. Generally this approach is suitable when n is at least 20 or 25 reference batches.

The Prediction Interval Approach

Suppose a random sample of n reference product batches are independently obtained and assume the resulting measurements follow a normal distribution with mean μ_R and variance σ_R^2 , respectively. Let $\hat{\mu}_R$ and $\hat{\sigma}_R$ be the corresponding estimates obtained from the sample measurements. Following Hahn and Meeker,^[26] a two-sided $100(1-\alpha)\%$ prediction interval to contain the values obtained on all of m randomly selected test product batches is $\hat{\mu}_R \pm r_{(1-\alpha;m,n)}\hat{\sigma}_R$ assuming the observations obtained on the test product

Table 2 Common Approaches for Estimating a Quality Range

Approach	Description	Comment
Tolerance Interval	Cover a large proportion of the population with a stated confidence	Very wide with small samples sizes (number of reference batches); suitable when number of reference batches at least 20 or 25 batches.
Prediction Interval	High confidence that all of a pre-specified number of future batches included in the interval	Useful when the number of test batches is known <i>a priori</i> ; can be wide when the number of reference batches is small.
Process Capability Interval	Sample equivalent of the quality range ($\hat{\mu}_R \pm f\hat{\sigma}_R$)	Often used in setting acceptance criteria and is easy to understand; recommended by Tsong, Dong et al., ^[13]
Data Range	Minimum of data to Maximum of data	Coverage can be poor for small sample sizes; simple to use but not recommended

batches follow the same normal distribution as the measurements obtained on the reference product batches. Exact values for $r_{(1-\alpha;m,n)}$ are available.^[26] If one or more of the test product batches generate data falling outside of the prediction interval then it is likely that the data obtained on the test batches do not arise from the same population that generated the data obtained on the reference batches. This approach is useful when m is known in advance. Similar to a tolerance interval, a prediction interval can be wide when n is small, especially for large m .

The Process Capability Interval Approach

This approach simply replaces the population mean and variance in $\mu_R \pm f\sigma_R$ with their sample estimates, $\hat{\mu}_R$ and $\hat{\sigma}_R$. When the number of reference batches is large (> 50) this approach is viable. As n decreases, the increasingly poor coverage offered by the process capability interval is not compensated for by the simplicity of the calculations. For example, based on the observations obtained on $n = 14$ reference batches, suppose a naïve user constructs the interval $\hat{\mu}_R \pm 3\hat{\sigma}_R$. The user may believe that about 99.7% of the population of measurements will be enclosed by this interval. But, in fact, he or she would be mistaken since for $n = 14$ the tolerance factor associated with a [95/95] tolerance interval is 3.024 and therefore only approximately 95% of the population will be enclosed by the interval with 95% confidence.

The Data Range Approach

Reporting the minimum and maximum of the data is sensible but unless n is extremely large (in the order of hundreds), defining the quality range as the interval spanning the minimum observation to the maximum observation should be avoided since it will have poor coverage.

Tier 2 Evaluation: Variability Comparison

Besides generating a quality range that is expected to enclose the measurements obtained on the test product batches, it might also be important to construct a prediction bound to exceed the standard deviation of the test batch data based on the data obtained on the reference batches. Suppose independent observations are obtained on n randomly sampled reference batches. Further, suppose the data follow a normal distribution with mean μ_R and variance σ_R^2 , respectively. Let $\hat{\sigma}_R$ be the (reference) sample standard deviation. Following Hahn and Meeker,^[26] a one-sided $100(1 - \alpha)\%$ upper prediction bound to exceed the standard deviation ($\hat{\sigma}_T$) of the observations obtained on m test product batches is $\hat{\sigma}_R \sqrt{F_{(1-\alpha;m-1,n-1)}}$, assuming the observations obtained on the test product batches follow the same normal distribution as the measurements obtained on the reference product batches. $F_{(1-\alpha;m-1,n-1)}$ represents the

upper $100(1 - \alpha)\%$ th percentile point from an F distribution with numerator degrees of freedom equal to $m - 1$ and denominator degrees of freedom equal to $n - 1$. Exact values for $F_{(1-\alpha;m-1,n-1)}$ for various choices of α , m and n are available.^[26] If the standard deviation of the test batch data exceeds the prediction bound derived from the reference batch data then it is likely that the test data do not arise from the same population that generated the reference data. Since the same analytical method is used to assay both reference and test batches, the likelihood of such an occurrence is low unless there is a significant differences in the among-batch variability for the test product batches compared to the reference product batches. Completing the variability assessment is therefore important since such an outcome could impact how the quality range is derived.

Comparative Stability Evaluation: Statistical Equivalence Testing of Slopes

Stability is an important quality attribute for many products. When comparing a reference product to a test product, comparative stability testing is required to confirm that changes over time in each of the pertinent QAs are equivalent. In this section methods developed for evaluating trends (i.e., slopes) are discussed as part of the overall assessment of similarity or comparability. Assessing equivalence of the intercepts is not considered since that is essentially the same as testing the equivalence of means as discussed previously as part of a Tier 1 analysis.

The four approaches described in Table 3 are the most common strategies employed to compare slopes. Schofield^[27] evaluates each of them in terms of vaccine potency but his conclusions apply to any stability indicating attribute. He recommended that an equivalence testing approach be utilized for evaluating the slopes obtained from comparative stability testing. For attributes that are not considered crucial, a quality range type of approach is a reasonable alternative. With any equivalence testing approach, the choice of equivalence margin is critical.

Comparative Stability Evaluation: Choice of Equivalence Margin for Comparing Slopes

The Schofield^[27] Approach

For determining the equivalence margin, δ , assume that the reference product changes according to some known rate β and let T_E be the exposure time, which is the time from release to expiry. Denote the measurement of reference product at release as X and the measurement of reference product at expiry as Y . Assume that the distribution of the release data X and distribution of the degradation rate β can both be determined based on historical data and experience. Then simulate data x^* and β^* based on the release distribution and degradation rate distribution respectively

Table 3 Four Approaches for Comparing Slopes

Approach	Description	Comment
Comparison to Specification	Test product data are compared against established stability specification limits; comparability established if all data fall within the limits	Highly risky even when reference and test product are comparable since the probability that all measurements obtained on the test product batches are contained within the specification limits can be extremely small, especially when assay variability is large. Conversely, when the reference and test products are not comparable, the criteria could nevertheless be satisfied.
Quality Range	Requires that slopes derived from analysis of data generated on the test product batches fall within a range of slopes established from the analysis of the data obtained on the reference product batches	The limits of the quality range may not align with slope differences considered to be scientifically important. For example, even though all the data generated on the test product batches fall inside the quality range determined based on the data generated on reference product batches, the two sets of data can still have very different stability trends. On the other hand, the fact that some data corresponding to the test product fall outside of the quality range based on the reference product does not necessarily mean a difference between the stability trends in the two sets of the data.
Hypothesis Testing	Requires a decision to “not reject” the null hypothesis of parallelism: H_0 : Slope Difference = 0 vs H_1 : Slope Difference $\neq$ 0	The conclusion to reject or not reject H_0 depends highly on the variability in the data. Very precise data exhibiting little difference in slopes can nevertheless lead to the rejection of H_0 . Conversely, highly variable data can lead to the conclusion to not reject H_0 even if there visually appears to be a large difference in the slopes. Lastly, the null hypothesis of equal slope (parallelism) may not be of interest in practice
Equivalence Testing	Requires a decision to “reject the null hypothesis of non-equivalence” to demonstrate the slope difference is within an acceptance range. H_0 : Slope Difference > δ vs H_1 : Slope Difference $\leq$ δ	The choice of δ is critical and should relate to slope differences that have scientific or clinical relevance

to determine a measurement at expiry as $y^* = x^* - \beta^* T_E^*$ where T_E^* should be determined based on some criterion, for example, the expiry specification of the product. This process can be repeated many times to obtain sufficient measurements at expiry based on which the distribution of Y can be estimated. This distribution can then be used to establish a shift in reference measurements at expiry, ΔY , and the equivalence margin δ can then be determined as $\delta = \Delta Y / T_{SL}$, where T_{SL} is the shelf-life of the reference product.

This approach is not appropriate for attributes that do not change much or at all over time under normal storage conditions ($\beta \approx 0$). For such attributes, accelerated comparative stability studies are required since then β is likely not near 0. Also, in practice, it may be difficult to decide upon a proper value for β , especially at accelerated conditions since usually there is little experience with the stability of the reference product under these conditions. Preliminary studies can be performed to provide this information.

The Burdick and Sidor^[28] Approach

Assume the stability profiles are consistent from batch-to-batch and denote the measurements taken from i^{th} lot and j^{th} time point as X_{ij}^R and X_{ij}^T respectively for the reference

product and test product. Then following Burdick and Sidor,^[28] the stability data can be modeled as:

$$\begin{aligned} X_{ij}^R &= \mu^R + \tau_i^R + \beta^R t_j + \varepsilon_{ij}^R, \quad i = 1, \dots, n_R; j = 1, \dots, T_R \\ X_{ij}^T &= \mu^T + \tau_i^T + \beta^T t_j + \varepsilon_{ij}^T, \quad i = 1, \dots, n_T; j = 1, \dots, T_T \end{aligned} \quad (8)$$

where

1. μ^R and μ^T are the means at the initial time point for the reference and test products;
2. τ_i^R and τ_i^T are the random intercept terms for the reference and test products for the i^{th} batch, both of which assumed to be distributed as $N(0, \sigma_\alpha^2)$;
3. β^R and β^T are the true slopes for the reference and test products;
4. t_i is the i^{th} time point;
5. ε_{ij}^R and ε_{ij}^T are the residual error terms for the reference and test products, both of which assumed to be distributed as $N(0, \sigma_\varepsilon^2)$; Here ε_{ij}^R , ε_{ij}^T , τ_i^R and τ_i^T are assumed to all be independent.
6. n_R and n_T are the total number of batches included for the reference and test products;
7. T_R and T_T are the total number of time points for the reference and test products and generally taken to be the same value, T .

While the authors were interested in deriving the equivalence margin, δ , in an accelerated stability study under non-recommended conditions in the context of evaluating a process change, their approach can be utilized to compare slopes between a reference product and test product in the context of an analytical comparability evaluation. Their approach can also be used for product stored under recommended conditions although in this case the resulting δ must be gauged against existing product specifications. The risk that a test product batch fails its specification might be deemed acceptably small even if the difference in slopes is outside of the equivalence interval for one or more attributes. The strategy discussed in Schofield^[27] can also be used for deriving δ .

Under the model given in Eq. 8, the null and alternative hypotheses to be tested for demonstrating slope equivalence are $H_0: |\beta^R - \beta^T| \geq \delta$ vs. $H_1: |\beta^R - \beta^T| < \delta$. The TOST procedure previously discussed can then be modified to perform this test. That is, H_0 is rejected if the two-sided confidence interval for $\beta^R - \beta^T$ is contained within the equivalence interval $(-\delta, \delta)$. Since it is difficult to ascertain what difference in slopes might be considered acceptable, especially under accelerated conditions, Burdick and Sidor^[28] propose that the equivalence margin be determined based on *effect size*, which expresses the absolute difference in the true slopes in terms of the variability associated with the least squares estimator of β^R . That is, effect size (ES) is defined as $ES = |\beta^R - \beta^T|/\sigma_R$, where in this case σ_R is the standard deviation corresponding to the least squares estimate of the slope obtained from the data generated on the reference product batches. Upon rearranging the equation for ES, the authors define the equivalence margin as $\delta = ES \times \sigma_R$. Under the assumptions stated above for the

model given in Eq. 8, $\sigma_R = \sqrt{\frac{\sigma_e^2}{\sum_i (t_i - \bar{t})^2}}$ with $\bar{t} = \frac{1}{T} \sum_i t_i$.

Through a simulation study the authors show that an ES of 2 is a reasonable choice. However, the choice should be justified for each specific situation; subject matter expert input is required. The residual error variance term, σ_e^2 , can be determined from an analytical method qualification or validation study, or from a prior stability study performed on the reference product. In some cases the use of an upper confidence bound (instead of the point estimate) can be considered to account for the uncertainty in the estimate.

Alternative model formulations may be necessary to better describe the data. For example, including both random intercept and random slope terms in the model is feasible when variation beyond the residual variability could impact the slope estimates. Also, using a fixed intercept model makes sense when the data at the initial time point exhibit very little variation. For example, the level of impurity at time 0 may be nearly 0% (below the limit of quantitation) for all batches in the study. Sometimes a simpler model must be utilized to achieve convergence in the numerical algorithm used to estimate

the model parameters. Whatever the reason is for using an alternative model, the derivation of the ES and the corresponding equivalence margin must be modified accordingly.

EVALUATING THE CLINICAL COMPARABILITY OF A BIOLOGICAL PRODUCT

Pharmacokinetic (PK) comparability

Typically, an equivalence approach is used to demonstrate the comparability of pharmacokinetic (PK) measurements of the test product and the reference product. To demonstrate equivalence, the 90% confidence interval (CI) for the true geometric mean ratio (GMR) of the test product relative to the reference product for both area under the curve (AUC) and maximum serum concentration (Cmax) must fall between the predefined limits of 0.80 and 1.25.^[29,30] However, due to the usually high variability for biological products resulting from the complexity and the possible pleiotropic activities of biologics, the CI of the estimated GMR could be too wide with commonly acceptable sample size. Consequently, there is a high probability of the constructed CI falling out of the predetermined limits (0.8, 1.25), which can lead to an unacceptably high producer risk. Many different approaches have been proposed for high variability drug products to manage the producer risk under a reasonable sample size requirement. The type of acceptance criteria using the reference-scaled method proposed by the FDA working group rather than the traditional equivalence criteria is recommended for the demonstration of PK comparability in this situation. The acceptance criteria are scaled based on the variability of the reference product resulting in a reduction of the producer risk without increasing the consumer risk.

Let μ_T and μ_R be the true means of a logarithmic pharmacokinetic response (e.g., AUC or Cmax) corresponding to the test and the reference products, respectively. The classical paradigm for the determination of equivalence of a test product to the reference product is to require the 90% confidence interval for the difference of means between the test and the reference fall within the predetermined equivalence interval ($\log(0.8)$, $\log(1.25)$). This is equivalent to evaluating Eq. 1 with $\delta = \log(1.25) \approx 0.233$ (using log base e). This approach does not depend on the underlying variability, or the ratio of geometric means of the PK parameters. To reduce the producer risk when product variability is high, without resorting to unrealistically large sample sizes, Haidar, Davit et al.,^[31] and Haider, Makhoul, et al.,^[32] compared several approaches and proposed the following reference-scaled criterion

$$(\mu_T - \mu_R)^2 / \sigma_{WR}^2 - \theta_F \leq 0 \quad (9)$$

where $\theta_F = \frac{(\log(1.25))^2}{\sigma_{w0}^2} = \frac{(\log(1.25))^2}{0.25^2} \approx 0.8$ using a “switching” variability σ_{w0} as a tool to control the sponsor risk, and σ_{wR}^2 is the within-subject variance for the reference product for a cross-over design but represents the total variance for the reference product for a parallel design. Comparability is claimed if the upper 95% CI for the parameter on the left side of Eq. 9 is negative or zero, and the point estimate of the GMR (test/reference estimated geometric mean ratio) falls within (0.8, 1.25). This approach is currently recommended by the FDA. Comparing to the traditional equivalence approach, the reference-scaled approach could effectively decrease the sample size needed for the demonstration of comparability.^[32] The additional point-estimate constraint eliminates the potential that a test product with a large mean difference would enter the market due to large variability in the reference product.

In terms of risks in evaluating PK comparability, it is clear that a larger switching variability is needed to control the consumer risk, while a smaller switching variability is needed to control the producer risk. However, the FDA recommended approach uses a single fixed switching variability. When there is no mean difference between the reference and the test products, the ideal test will have a high probability of accepting comparability. If the reference product is compared to the reference itself (i.e., True GMR = 1), then it should always be comparable regardless of the variability. However, for this case, the FDA ACPS method only gives 74% chance of accepting comparability, and leads to a large producer risk. In other words, the switching variability is too large in this case and a smaller switching variability should be used. In the other direction, a switching variability that is too small will lead to a large consumer's risk. For a more reasonable balance and control of the two risks, Liao and Heyse^[33] proposed using two switching variabilities: one for controlling the producer risk when the true GMR = 1, where a high probability of concluding comparability is desired, and a second one for controlling the consumer risk when the true GMR = 1.25, where a low probability of concluding comparability is desired. For this purpose, the authors proposed a new reference-scaled approach using the criterion

$$(\mu_T - \mu_R)^2 / \sigma_{wR}^2 - 2\theta_L(GMR) \leq 0 \quad (10)$$

Where $\theta_L(GMR) = [\log(1.25) / (\sqrt{2} \times (3.763 \times GMR - 2.763) \times 0.152)]^2$, and σ_{wR}^2 is the within-subject variance for the reference product for a cross-over design but represents the total variance for the reference product for a parallel design. Here, the switching variability is a function of the GMR. The GMR should always be greater than or equal to 1. If it is believed that the true PK parameter corresponding to the test product is less than that of the reference product replace GMR with 1/GMR in Eq. 10. Comparability is claimed if the upper 95% CI for the parameter on the left side of Eq. 10 is negative or zero.

For the criteria described Eq. 9 or Eq. 10, an approximate linearized regulatory model of the form $\eta = X^2 - cY^2 \leq 0$ is suggested^[34] where $\eta = (\mu_T - \mu_R)^2 - \theta_F \sigma_{wR}^2 \leq 0$ for Eq. 9 and $\eta = (\mu_T - \mu_R)^2 - 2\theta_L(GMR) \sigma_{wR}^2 \leq 0$ for Eq. 10, which would require a more complicated calculation to construct the confidence interval. The sampling distribution of the approximate linearized regulatory model is complicated but can be approximated by a non-central t-distribution or even a normal distribution when the number of subjects is large.^[35] Thus, a non-central t-distribution or a normal approximation could be used to calculate the confidence limits rather than the more complicated linearized regulatory model as all three have similar operating characteristics.

The performance of the two reference scaled approaches was compared using a simulation study and by way of an example. As detailed in Liao and Heyse,^[33] when there is no subject-by-product interaction and no mean shift (true GMR = 1), the criterion given by Eq. 10 leads to an 85% probability of concluding comparability compared to 74% for the FDA approach (Eq. 9) thereby lowering the producer risk. In this case, the switching variability, σ_{w0} , is 0.152. When there is a mean shift, i.e., when the true GMR = 1.25, the switching variability $\sigma_{w0} = 0.292$ is used for controlling the consumer risk. Third, the approach in Eq. 10 gives a narrower boundary for a larger GMR, which is very reasonable since the probability of claiming comparability should be getting smaller when GMR is deviating from 1. In fact, optimal switching variabilities can be derived based on pre-specified consumer and producer risks, chosen to be acceptable to both the manufacturer and the regulatory agency. For two given switching variabilities, σ_{w1} and σ_{w2} , the values for $\theta_L(GMR)$ in Eq. 10 would be $\theta_L(GMR) = [\log(1.25) / (\sqrt{2} \times ((a+1) \times GMR - a) \times \sigma_{w1})]^2$, where $a = 4 \times \sigma_{w2} / \sigma_{w1} - 5$, σ_{w1} is the smaller switching variability controlling the producer risk and σ_{w2} is the larger switching variability controlling the consumer risk.

Pharmacodynamic (PD) profile comparability

It is challenging to assess pharmacodynamic (PD) profile comparability. Varki, Pequignot et al.,^[36] and Waller, Bronchud et al.,^[37] used the classical equivalence rule for PD comparison in terms of the area under the effect curve (AUEC) of the PD profile. Recently, the FDA issued a guidance document for conducting a PD comparability study using this rule.^[30] However, this may not be a meaningful approach since the interpretation for AUEC used in PD studies is not the same as for PK studies since PK describes what the body does to the drug while PD describes what the drug does to the body. Since the AUEC is summarized before the PD profile is compared directly, the same AUEC value does not necessarily imply the same PD profile. The objective of evaluating PD comparability is to directly compare the entire PD profile. A single AUEC assessment of the PD profile may not be sensitive enough.

To directly compare two PD profiles, Liao and Darken^[38] proposed a PD comparability index, f_{PD} , using the entire profile as shown in Eq. 11.

$$f_{PD} = \frac{\min(R_{\max}^R - R_{\min}^R, R_{\max}^T - R_{\min}^T)}{\max(R_{\max}^R - R_{\min}^R, R_{\max}^T - R_{\min}^T) + \sqrt{\frac{1}{n} \sum_{j=1}^n w_j (y_j^R - y_j^T)^2}} \quad (11)$$

where n is the total number of time points in the PD profile; y_j^R and y_j^T are the predicted PD responses at time t_j for the reference and the test product, respectively; w_j is the optional weight for controlling the contribution at time t_j to the overall evaluation, and $R_{\max}^R$ ($R_{\min}^R$) and $R_{\max}^T$ ($R_{\min}^T$) are the predicted maximum (minimum) profile response for the reference and the test product, respectively. The PD comparability index falls within $[0, 1]$ with larger values of f_{PD} implying closer agreement between the two PD profiles. An f_{PD} of 1 indicates 100% comparability, i.e., the two profiles are identical. PD comparability is claimed if (1) The approximate lower 95% confidence limit of the PD comparability index f_{PD} for the test against the reference is greater than a value δ_0 and (2) The point estimate of

δ_0 as the maximum of the two values. The second requirement in the acceptance criteria ensures that no major difference in the PD profiles passes the comparability assessment due to large variability in the reference product, which is not unusual for a biological product.^[31–32] Selecting a value of $\delta_1 = 0.9$, which is about 10% difference in the means, could be a reasonable value for use.

To estimate the value of f_{PD} in Eq. 11, the corresponding parameters can be obtained by fitting a model such as the Emax model or the extended Emax model^[39] or by using a non-parametric method to determine a smooth curve (spline) of the PD profile. To construct the lower 95% confidence limit for f_{PD} when comparing the reference profile to itself, a bootstrap method is recommended. The residuals are bootstrapped a large number of times (N) and the PD model is re-fitted N times using the bootstrapped data. The lower 5th percentile of the bootstrapped determinations of the PD comparability index is then used as the approximate 95% lower confidence limit. For comparing the test product profile to the reference product profile using non-parametric spline fits, the lower confidence band can be constructed using the method in Hastie and Tibshirani,^[40] as provided in Eq. 12.

$$\frac{\min(R_{\max}^R - R_{\min}^R, R_{\max}^T - R_{\min}^T)}{\max(R_{\max}^R - R_{\min}^R, R_{\max}^T - R_{\min}^T) + \sqrt{\max \left\{ \frac{1}{n} \sum_{j=1}^n w_j (y_j^R - y_j^T \pm z_{0.975} \sqrt{S_R^2 + S_T^2})^2 \right\}}} \quad (12)$$

f_{PD} is greater than a value δ_1 . The two boundaries δ_0 and δ_1 should be pre-specified and agreed to by regulatory authorities.

The boundary δ_0 in the acceptance criteria takes into account variability in the reference product. One way to choose δ_0 is to use the reference comparison as the basis for making a comparison between the test and the reference product. When the reference and the test PD profiles are very close, the two confidence interval limits, one from the reference against itself and the other from the test against the reference, are also close to each other, as expected. Considering this outcome, first calculate the approximate lower 95% confidence limit of the PD comparability index based on evaluating the reference against itself. Then, choose δ_0 equal to a fraction, c ($0 < c < 1$), of this approximate lower limit as the threshold for comparing the test profile against the reference profile. One suggestion is to choose $c = 0.9$. Another approach for determining δ_0 sets it to a fixed value regardless of the variability in the reference product. For example, to account for a 30% difference or change in the PD profiles when the maximum (minimum) response is the same for both the test and reference products, set δ_0 at 0.77. A conservative approach is to do both calculations and set

where S_k^2 , $k = R, T$ is the residual variance for the reference and the test product, respectively. If possible, the same subjects should be utilized to obtain information on the PK-PD relationship.

Clinical Efficacy Comparability

To demonstrate that the test and reference products exhibit comparable clinical efficacy, the available data from the historical trials comparing the reference product with placebo (P) and from new clinical comparability studies that directly compare the test and reference products are used. Without loss of generality, we assume the efficacy is measured on some arbitrary scale such that a larger value represents better efficacy. Further, we assume that the summary statistic measuring efficacy can be reasonably assumed to have a normal distribution, where the treatment effect can be measured as a, for example, mean difference between groups (log scale), log-odds ratio, log-relative risk, or log-hazard ratio.

Let γ_{RP} be the true effect of the reference product relative to the placebo and γ_{TR} the true effect of the test product relative to the reference product. Therefore,

$\gamma_{TR} + \gamma_{RP}$ represents the true effect of the test product relative to the placebo. Let B_{RP} and V_{RP} be the estimate of γ_{RP} and the corresponding estimated variance of B_{RP} , respectively, based on a meta-analysis of the data from the appropriate historical clinical trials. Similarly, let B_{TR} and V_{TR} be the estimate of γ_{TR} and the corresponding estimated variance of B_{TR} based on the data from a new comparability clinical trial. Three approaches to assess the comparability between the test and reference products are described below.

Non-inferiority (N) Approach

In this approach, given a margin δ , the goal is to demonstrate the proposed test product is not worse than the reference product by more than δ . The choice of the non-inferiority margin is widely debated in the literature. However, no matter how the margin is selected, for this approach to be valid it must not be larger than the true effect of the reference product relative to the placebo (i.e., δ must be less than or equal to γ_{RP}). One common approach for selecting δ is to pick a value that preserves at least some non-zero fraction, f , of γ_{RP} . This could be based on the lower limit of the 95 percent confidence interval for the reference product relative to the placebo, i.e., $\delta = (1 - f)(B_{RP} - 1.96\sqrt{V_{RP}})$, a common *fixed margin* method and part of the 95-95 approach where the lower limit of a 95% confidence interval for γ_{TR} is compared to this margin. A commonly used value that is referenced as a reasonable starting point for discussion in FDA guidance is $f = 0.5$, i.e., 50% effect preservation.^[41] Thus, the hypotheses for non-inferiority are $H_{01} : \gamma_{TR} \leq -\delta$ vs $H_{a1} : \gamma_{TR} > -\delta$. If a fixed-margin is used, non-inferiority is established if $B_{TR} - 1.96\sqrt{V_{TR}} > -\delta$, i.e.,

$$\frac{B_{TR} + (1 - f)B_{RP}}{\sqrt{V_{TR}} + \sqrt{(1 - f)^2 V_{RP}}} > 1.96 \quad (13)$$

If a *synthesis margin* method is used which also uses the variability information from the current comparability study, then non-inferiority is established if^[42]

$$\frac{B_{TR} + (1 - f)B_{RP}}{\sqrt{V_{TR}} + (1 - f)\sqrt{V_{RP}}} > 1.96 \quad (14)$$

Note that FDA guidance prefers the fixed margin method for the NI approach.^[41] However, it is of interest to note that the left side of these equations will always be larger for the synthesis method (the terms in the denominator are > 0 and $1/\sqrt{A+B} \geq 1/(\sqrt{A} + \sqrt{B})$) and hence the synthesis method is uniformly more powerful than the 95-95 fixed margin approach, or more generally, than any double confidence interval based fixed margin approach provided that the assay sensitivity and constancy are accounted for in order to control the type I error.^[41]

Equivalence Approach

The hypotheses for equivalence are $H_{02} : \gamma_{TR} \leq -\delta$ or $\gamma_{TR} \geq \delta$ vs $H_{a2} : -\delta < \gamma_{TR} < \delta$. If the fixed margin is used, then equivalence is established if $B_{TR} - 1.96\sqrt{V_{TR}} > -\delta$ and $B_{TR} + 1.96\sqrt{V_{TR}} < \delta$, i.e.,

$$\begin{aligned} \frac{B_{TR} + (1 - f)B_{RP}}{\sqrt{V_{TR}} + \sqrt{(1 - f)^2 V_{RP}}} &> 1.96 \quad \text{and} \\ \frac{B_{TR} - (1 - f)B_{RP}}{\sqrt{V_{TR}} + \sqrt{(1 - f)^2 V_{RP}}} &< -1.96 \end{aligned} \quad (15)$$

Using a similar argument as in the non-inferiority case, if a synthesis margin is used, then equivalence is established if

$$\begin{aligned} \frac{B_{TR} + (1 - f)B_{RP}}{\sqrt{V_{TR}} + (1 - f)\sqrt{V_{RP}}} &> 1.96 \quad \text{and} \\ \frac{B_{TR} - (1 - f)B_{RP}}{\sqrt{V_{TR}} + (1 - f)\sqrt{V_{RP}}} &< -1.96 \end{aligned} \quad (16)$$

Constrained Non-inferiority (cNI) Approach

It is a well-known fact that an equivalence approach requires much larger sample size to achieve the same power as the non-inferiority approach. However, the non-inferiority approach only guarantees that the test product is not inferior to the reference product. For the equivalence approach, it is necessary to show that the proposed test product is comparable to the reference product; it has neither decreased nor increased activity compared to the reference product.^[8-10] Hence these two approaches have different merits and concerns. Liao^[43] proposed a constrained non-inferiority (cNI) approach using a similar sample size as the NI approach but also addressing comparability. The cNI approach represents a middle ground where greater evidence is required to demonstrate that the test product is no worse than the reference product (compared to the NI approach) but it is less stringent than the equivalence approach. It properly evaluates that the test product is comparable to the reference product by 1) demonstrating that the test product is not inferior to the reference product and 2) demonstrating that the test shows no increased activity compared to the reference product for the clinical endpoint. If these two objectives are achieved, then the test and reference products are claimed to be comparable.

The proposed cNI approach for claiming comparability depends on two facts. The first is to pass the classical NI requirement to ensure the proposed test product is not inferior to the reference product. The second is to satisfy two constraints: 1) the point estimate of γ_{TR} is within the pre-defined mean boundary and 2) the 95% CI of γ_{TR} is within the predefined plausibility interval to ensure the proposed test product does not exhibit increased activity.

The plausibility interval can be constructed using the idea that any observed difference between the reference against itself should not be considered clinically meaningful, and is defined as $(-k\sqrt{2\sigma_R^2}, +k\sqrt{2\sigma_R^2})$, where σ_R^2 is the total variability for the reference. Liao^[43] used simulation and real datasets to demonstrate that the cNI approach had good performance in terms of power and type I error control. The cNI approach generally requires smaller sample sizes than required for the equivalence approach but still meets all necessary requirements for addressing clinical efficacy comparability.

Incorporating Multiple Batches

Recently, Liao and Yu^[44] presented a very interesting idea by including multiple batches in equivalence and non-inferiority studies to improve the study power. Batch variability exists in almost all products to varying degrees. Generally speaking, more batch variability is expected in biological products than pharmaceutical products. In a clinical study, the drug supply personnel usually buys the product from a single batch, which can be very challenging especially when many companies are vying for the same reference product for many other studies in the same time frame. When using one batch product in a large study with a long study period, there may be a high probability that the product will expire before the end of the study. If there is a large amount of inter-batch variability, a study that uses just one reference batch increases the risk of either not detecting an important effect (low power) or over-stating a statistically significant effect (truth inflation) due to the choice of reference batch used in the study. To overcome this potential risk and reduce the burden of the drug supply department, one straightforward solution is to buy the reference products from different batches manufactured at different times or at different sites/regions. More importantly, at the design stage, increasing the number of batches in the study can increase the study power while controlling the type I error and at the same time reduce the estimation bias of the treatment effect and its variance. The improvement is more significant when the between batch variability is large. Another benefit of using multiple batches is the robustness and reproducibility of the study results. By using more than one batch, the treatment effect is “averaged” and therefore more reflective of true difference. Using more than one batch will also reduce the risk of experiencing a possible clinical supply issue because the reference batch expired.

CONCLUSIONS

Evaluating the comparability of a test product to the reference product is a very complicated task requiring support by many different functional areas (nonclinical, analytical, regulatory, clinical, etc.). Demonstrating comparability

requires a stepwise approach to evaluate the “totality of evidence,” which may include information from structural and functional characterizations, nonclinical evaluation, analytical assessments, human PK/PD studies, clinical immunogenicity studies, and clinical safety and efficacy trials. No one piece of evidence is sufficient to establish comparability or to disprove it. The demonstration of comparability does not necessarily mean that the quality attributes and other key parameters of interest of the reference and test product are identical, but that they are highly similar and that the existing knowledge is sufficiently predictive to ensure that any differences in quality attributes and the key parameters of interest have no adverse impact upon safety or efficacy of the drug product.

In this entry, statistical considerations for evaluating comparability were discussed in the context of analytical testing and human clinical studies (PK/PD, efficacy). The challenges at each phase of the stepwise approach can be different. The criteria for a comparability study should be established at the planning stage with input from relevant subject matter experts. Communication between the sponsor and regulatory agency should begin very early as well. The appropriate design, quality attributes, reference selection, acceptance criteria and statistical methods should all be included in the discussion with the agencies and consensus should be reached prior to beginning the studies. Different agencies may have different opinions on what needs to be completed as part of the overall assessment of comparability. The sponsor should be prepared to design flexibility into their studies to accommodate these potentially differing opinions. To some extent, batch-to-batch variability exists for all products. Because acceptance criteria are often derived based on variability among the reference batches, using too few batches can result in unrealistic criteria that are difficult to meet. It is therefore important and practical to use more than one batch in designing a comparability study for improving the study power. Selecting batches that reflect the likely sources of variability is sensible. How many batches should be used depends on the magnitude of the between batch variability as well as on manufacturing constraints. Due to the complexity of comparability studies, it is recommended that the actual study design and data analysis should be carried out with the close collaboration of a statistician.

All of the discussion has focused on assessing comparability on an attribute-by-attribute basis, including clinical endpoints, if needed. This is somewhat of an abuse of terminology since comparability is a holistic concept, based on the totality of evidence. However, once all of the data have been collected and analyzed, a final overall decision has to be reached. Rarely will the data look so great (or so awful) that the conclusion is obvious. How is the idea of appraising “totality of evidence” implemented in practice? Consider the analytical evaluation of a potential biosimilar. The various assays might be grouped into different categories relating to immunogenicity, quality, process, safety,

etc. At some pre-determined time, such as prior to a key management review, convene a meeting of the relevant subject matter experts to review the data using some sort of scoring system. Calculate a summary measure for each assay category and compare to pre-specified cutoffs. Values above the cutoff increase the likelihood of reaching a comparability conclusion; values below the cutoff decrease the likelihood of reaching such a conclusion. This information is then reviewed with management and a decision is made about the program. As the program moves forward in development, this process can be repeated as new data or information becomes available. This provides a scientifically objective mechanism for making informed program decisions using input from all relevant experts.

REFERENCES

- Public Health Services Act 42 U.S.C. §262(i).
- Genetic Engineering & Biotechnology News (GEN, 2013) Top 20 best-selling drugs of 2012. March 5, 2013. Website: www.genengnews.com/insight-and-intelligence/top-20-best-selling-drugs-of-2012/77899775/?page=2 623.
- Genetic Engineering & Biotechnology News (GEN, 2014) The top 25 best-selling drugs of 2013. September 16, 2014. Website: <http://www.genengnews.com/insight-and-intelligence/the-top-25-best-selling-drugs-of-2013/77900053/>.
- Genetic Engineering & Biotechnology News (GEN, 2015) The Top 25 Best-Selling Drugs of 2014. February 23, 2015. Website: <http://www.genengnews.com/the-lists/the-top-25-best-selling-drugs-of-2014/77900383?page=2&q=top> selling drugs in 2014.
- Weise, M.; Bielsky, M.C.; De Smet, K.; Ehmann, F.; Ekman, N.; Giezen, T.J.; Gravanis, I.; Heim, H.K.; Heinonen, E.; Ho, K.; Moreau, A.; Narayanan, G.; Kruse, N.A.; Reichmann, G.; Thorpe, R.; van Aerts, L.; Vlemineckx, C.; Wadhwa, M.; Schneider, C.K. Biosimilars: what clinicians should know. *Blood* **2012**, *120* (26), 5111–5117.
- McCamish, M.; Woollett, G. The Continuum of Comparability Extends to Biosimilarity: How Much Is Enough and What Clinical Data Are Necessary? *Clinical pharmacology & Therapeutics* **2013**, *93* (4), 315–317.
- U.S. Food and Drug Administration (FDA). Guidance for industry: Q5E Comparability of Biotechnological/Biological Products Subject to Changes in Their Manufacturing Process. Washington, DC, 2001.
- U.S. Food and Drug Administration (FDA). Guidance for industry: Scientific considerations in demonstrating biosimilarity to a reference product, Washington, DC, 2012a.
- U.S. Food and Drug Administration (FDA). Guidance for industry: Quality considerations in demonstrating biosimilarity to a reference product, Washington, DC, 2012b.
- U.S. Food and Drug Administration (FDA). Guidance for industry: Biosimilars: Questions and answers regarding implementation of the biologics price competition and innovation act of 2009, Washington, DC, 2012c.
- Babbitt B.; Nick C. Considerations in establishing a US approval pathway for biosimilar and interchangeable biological products. *DIA Global Forum* **2011**, *3* (1), 44–49.
- Saurwein-Teissl, M.; Lang-Salchner, M.; Kirchlechner, T. Global biosimilar development, *DIA Global Forum* **2011**, *3* (1), 39–43.
- Tsong, Y.; Dong X.; Shen, M. Development of Statistical Methods for Analytical Similarity Assessment, *Journal of Biopharmaceutical Statistics* **2015** DOI:10.1080/10543406.2015.1092038
- Schuurmann, D. J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics* **1987**, *15* (6), 657–680.
- Chatfield, M.J.; Borman, P.J.; Damjanov, I. Evaluating change during pharmaceutical product development and manufacture-comparability and equivalence. *Qual. Reliability Eng. Int.* **2011**, *27* (5), 629–640
- Tuke, J.; Glonek, G. F. V.; Solomon, P. J. P-values, q-values and posterior probabilities for equivalence in genomics studies, **2012**, arXiv preprint arXiv:1202.0048.
- Westlake, W. J. Bioequivalence testing - a need to rethink. *Biometrics* **1981**, *37* (3), 589–594.
- Schuurmann, D. L. On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics* **1981**, *37* (3), 617.
- Brown, L.; Casella, G.; Hwang, J. Optimal confidence sets, bioequivalence, and the limaçon of Pascal. *Journal of the American Statistical Association* **1995**, *90* (431), 880–889.
- Berger, R. L.; Hsu, J. C. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* **1996**, *11* (4), 283–319.
- United States Pharmacopeia and National Formulary. General Chapter <1010> Analytical Data-Interpretation and Treatment, USP39-NF34. United States Pharmacopeia Convention: Rockville, MD, 2016, 767–782.
- Limentani, G. B.; Ringo, M. C.; Ye, F.; Bergquist, M. L.; MCSorley, E. O. Beyond the t-test: statistical equivalence testing. *Analytical Chemistry* **2005**, *77* (11), 221–A.
- Liao, J.J.Z.; Darken, P. F. Comparability of critical quality attributes for establishing biosimilarity. *Statistics in medicine* **2013**, *32* (3), 462–469.
- Chow, S.C. Challenging issues in assessing analytical similarity in biosimilar studies. *Biosimilars* **2015**, *5*, 33–39.
- Liao, J.J.Z. Statistical Methods for Comparability Studies. In *Nonclinical Statistics for Pharmaceutical and Biotechnology Industries* (L. Zhang and A. Stan, Ed.), Springer: New York, 2016, 675–694.
- Hahn, G. J.; Meeker, W. Q. *Statistical Intervals: A Guide for Practitioners*, John Wiley & Sons, Inc.: New York, 1991.
- Schofield, T. L. Maintenance of vaccine stability through annual stability and comparability studies. *Biologicals* **2009**, *37* (6), 397–402.
- Burdick, R. K.; Sidor, L. Establishment of an equivalence acceptance criterion for accelerated stability studies. *Journal of biopharmaceutical statistics* **2013**, *23* (4), 730–743.
- U.S. Food and Drug Administration (FDA). Guidance for industry: statistical approaches to establishing bioequivalence, Washington, DC, 2001.
- U.S. Food and Drug Administration (FDA). Guidance for industry: clinical pharmacology data to support a demonstration of biosimilarity to a reference product, Washington, DC, 2014.

31. Haidar, S.H.; Davit, B.M.; Chen, M.L.; Xu, L.X. Bioequivalence approaches for highly viable drugs and drug products, *Pharm. Res.* **2007**, *25*, 237–241.
32. Haidar, S. H.; Makhoulouf, F.; Schuirmann, D. J.; Hyslop, T.; Davit, B.; Conner, D.; Xu, L. X. Evaluation of scaling approach for the bioequivalence of highly variable drugs. *AAPS Journal* **2008**, *10*, 450–454.
33. Liao, J.J.Z.; Heyse, J.F. Biosimilarity for follow-on biologics. *Stat. Biopharm. Res.* **2011**, *3* (3), 445–455.
34. Hyslop, T.; Hsuan, F.; Holder, D.J. A small sample confidence interval approach to assess individual bioequivalence. *Stat. Med.* **2000**, *19*, 2885–2897.
35. Tothfalusi, L.; Endrenyi, L. Limits for the scaled average bioequivalence of highly variable drugs and drug products. *Pharm. Res.* **2003**, *20*, 382–389.
36. Varki, R.; Pequignot, E.; Leavitt, M.C.; Ferber, A.; Kraft, W.K. A glycosylated recombinant human granulocyte colony stimulating factor produced in a novel protein production system (AVI-014) in healthy subjects: a first-in human, single dose, controlled study. *BMC Clinical Pharmacology* **2009**, *9* (2), 1–8.
37. Waller, C.F.; Bronchud, M.; Mair, S.; Challand, R. Comparison of the pharmacodynamics profiles of a biosimilar filgrastim and Amgen filgrastim: results from a randomized, phase I trial., *Ann. Hematol.* **2010**, *89*, 971–978.
38. Liao, J.J.Z.; Darken, P. Biocomparability Studies: The Key to Success for Biosimilars. DIA meeting, Chicago, IL, 2011.
39. Sarkar, C.A.; Lauffenburger, D.A. Cell-level PK model of granulocyte colony-stimulating factor: implications for ligand lifetime and potency in vivo. *Molecular Pharmacology* **2003**, *63* (1), 147–158.
40. Hastie, T.J.; Tibshirani, R.J. Generalized Additive Models. Chapman and Hall: London, 1990.
41. U.S. Food and Drug Administration (FDA). Guidance for industry: non-inferiority clinical trials, Washington, DC, 2010.
42. Snapinn, S.; Jiang, Q. Controlling the type 1 error rate in non-inferiority trials. *Stat. Med.* **2008**, *27*, 371–381.
43. Liao, J.J.Z. A Constrained Non-inferiority Approach for Assessing Clinical Efficacy to Establish Biosimilarity, *Int J Clin Biostat Biom* **2015**, *1* (008), 1–7.
44. Liao, J.J.Z.; Yu, Z. Batch Effects on Design and Analysis of Equivalence and Non-inferiority Studies, ICSA symposium, Atlanta, GA, 2106.

BIBLIOGRAPHY

1. EMA Committee for Medicinal Products for Human Use (CHMP). Guideline on similar biological medicinal products. London, 2005.
2. Krishnamoorthy, K; Mathew, T. Statistical Tolerance Regions: Theory, Applications, and Computation. Wiley: Hoboken, NJ, 2009.
3. Chow, S.-C. Biosimilars: Design and Analysis of Follow-on Biologics. Chapman and Hall/CRC: Boca Raton, FL, 2014.
4. Chow, S.-C.; Liu, J.-P. Design and Analysis of Bioavailability and Bioequivalence Studies, 3rd Edition. Chapman and Hall/CRC: Boca Raton, FL, 2009.
5. World Health Organization (WHO). Guidelines on evaluation of similar biopharmaceutical products (SBPs). World Health Organization: Geneva, 2009.

Confidence Interval and Hypothesis Testing

Ivan S. F. Chan and Devan V. Mehrotra

Merck Research Laboratories, West Point, Pennsylvania, U.S.A.

Abstract

This entry reviews the general definition and properties of confidence intervals, describes the various methods of constructing confidence intervals including asymptotic and exact procedures, and discusses the duality between confidence interval estimation and hypothesis test.

INTRODUCTION

Parameter estimation is a primary objective in almost every kind of study, whether it is a clinical trial to evaluate the efficacy of a drug or vaccine, a cohort study of excess risk of disease due to toxic exposure, an epidemiological study of disease incidence, or a survey of public opinion. Because of the inherent variability in measurements and study populations, confidence intervals (CIs) are typically used to provide a measure of precision of parameter estimates. The concept of confidence intervals was first proposed by Neyman^[1,2] in the 1930s, and the use of confidence interval has since become an integral part of statistical analyses. In this article, we review the general definition and properties of confidence intervals, describe the various methods of constructing confidence intervals including asymptotic and exact procedures, and discuss the duality between confidence interval estimation and hypothesis test. We emphasize the importance of using the same methodology for confidence interval construction and hypothesis test in order to maintain consistency of statistical inferences.

DEFINITIONS OF CONFIDENCE INTERVALS

In this section we briefly review the general definition of confidence intervals and their properties. More comprehensive discussions can be found in Bickel and Doksum,^[3] Rohatgi,^[4] or Lehman.^[5]

Suppose $(X_1, X_2, \dots, X_n)$ is a random sample from a distribution F characterized by a parameter of interest θ . Let $\theta_L = L(X_1, \dots, X_n)$ and $\theta_U = U(X_1, \dots, X_n)$ be two statistics such that for all possible values of θ

$$\Pr(\theta_L \leq \theta \leq \theta_U) \geq 1 - \alpha \quad (1)$$

Then the random interval $[\theta_L, \theta_U]$ is called a level $(1 - \alpha)$ or a $100(1 - \alpha)\%$ (two-sided) confidence interval for θ . If the parameter θ is r -dimensional, then a confidence

set (or region) can be defined as a r -dimensional space $S = \{\theta: \theta_{Lj} \leq \theta_j \leq \theta_{Uj}, j = 1, \dots, r\}$ such that $\Pr(\theta \in S) \geq 1 - \alpha$ for all θ .

The quantity on the left of Eq. 1 is called the probability of coverage of the interval, and $(1 - \alpha)$ is called the (minimum) confidence level of the interval. If the distribution F is continuous, the equality in Eq. 1 can be achieved; whereas if F is discrete, not all levels $(1 - \alpha)$ are attainable and the confidence interval is generally conservative in the sense that the probability of coverage is at least $(1 - \alpha)$. An interpretation of a confidence interval is that if we perform many experiments (or studies) to evaluate θ and construct a $100(1 - \alpha)\%$ confidence interval based on the outcome of each experiment, then (approximately) at least $100(1 - \alpha)\%$ of these confidence intervals will bracket the true parameter θ .

In some situations, it may be of interest to obtain a lower or upper confidence bound for θ . A statistic $\theta_L = L(X_1, \dots, X_n)$ is called a $100(1 - \alpha)\%$ lower confidence bound for θ if

$$\Pr(\theta_L \leq \theta) \geq 1 - \alpha \quad \text{for all } \theta \quad (2)$$

Similarly, a statistic $\theta_U = U(X_1, \dots, X_n)$ is called a $100(1 - \alpha)\%$ upper confidence bound for θ if

$$\Pr(\theta_U \geq \theta) \geq 1 - \alpha \quad \text{for all } \theta \quad (3)$$

The random intervals $[\theta_L, \infty)$ and $(-\infty, \theta_U]$ are called one-sided $100(1 - \alpha)\%$ confidence intervals. This lower (or upper) confidence bound corresponds to the lower (or upper) bound of an equal-tailed two-sided $100(1 - 2\alpha)\%$ confidence interval.

CONSTRUCTION OF CONFIDENCE INTERVALS

For any given situation, there can be more than one confidence interval at level $1 - \alpha$. In addition, the construction of confidence intervals depends on the choice of statistic used. In general, the length $(W = \theta_U - \theta_L)$ or expected length $(E(W))$ of a confidence interval is used as a measure

of precision. Among all possible $100(1 - \alpha)\%$ confidence intervals based on some statistic T , we desire the one with the shortest length or expected length. However, a uniformly minimum length confidence interval may not exist.^[3] If the distribution of the statistic is symmetric about θ , the expected length of the (two-sided) confidence interval based on the statistic is minimized^[4] by choosing an equal-tailed confidence interval such that

$$\Pr(\theta < \theta_L) \leq \alpha_1 \quad \text{and} \quad \Pr(\theta > \theta_U) \leq \alpha_2, \quad (4)$$

where $\alpha_1 = \alpha_2 = \alpha/2$

The probabilities $\Pr(\theta < \theta_L)$ and $\Pr(\theta > \theta_U)$ are called the left- and right-tail error coverage probabilities, respectively. If the distribution of the statistic is not symmetric about θ , one can find a combination of α_1 and α_2 with $(\alpha_1 + \alpha_2 = \alpha)$ to minimize the expected length. Nevertheless, an equal-tailed confidence interval is often provided in practice for convenience or other considerations.

In many situations the construction of a confidence interval can be based on finding a pivot, which is a function of $(X_1, X_2, \dots, X_n)$ and θ whose distribution is independent of θ . For example, suppose $(X_1, X_2, \dots, X_n)$ is a random sample from a normal distribution with unknown mean (θ) and known standard deviation (σ). A natural (maximum likelihood, minimum variance, unbiased) estimate of θ is the sample mean ($\bar{X}$). Because $\bar{X}$ is normally distributed with mean θ and standard deviation $(\sigma/\sqrt{n})$ the function $\sqrt{n}(\bar{X} - \theta)/\sigma$ has a standard normal distribution (with mean 0 and variance 1) and thus can be considered as a pivot. Therefore, to obtain a 95% confidence interval for θ , we can find c_1 and c_2 such that

$$\Pr(c_1 \leq \sqrt{n}(\bar{X} - \theta)/\sigma \leq c_2) = 0.95 \quad (5)$$

Because the distribution of $\bar{X}$ is symmetric about θ , the shortest confidence interval is the equal-tailed interval with $c_1 = -z_{0.025}$ and $c_2 = z_{0.025}$ where $z_{0.025} (= 1.96)$ is the upper 2.5th percentile of the standard normal distribution. As a result, the 95% confidence interval for θ is $[\theta_L, \theta_U] = [\bar{X} - z_{0.025}\sigma/\sqrt{n}, \bar{X} + z_{0.025}\sigma/\sqrt{n}]$.

When a pivot is not readily available, one may consider using “approximate” pivots based on large sample approximations, and the resulting confidence intervals are usually called asymptotic or approximate confidence intervals. Some examples include the Wilson score confidence interval^[6] for a single binomial probability (θ) based on the (standard) normal approximation of $\sqrt{n}(\bar{X} - \theta)/\sqrt{\theta(1 - \theta)}$, and the approximate confidence interval for the Poisson parameter (θ) based on the (standard) normal approximation of the variance stabilizing transformation $\sqrt{n}(\sqrt{\bar{X}} - \sqrt{\theta})$.^[3] For parameter estimates involving complicated functions of the observed data, simulation techniques such as the bootstrap^[7,8] may be used to construct confidence intervals.

Another common approach to confidence interval construction is to invert a hypothesis test using the duality between hypothesis testing and interval estimation.^[3–5] Such intervals are called test-based confidence intervals. Suppose T is a test statistic for $H_0: \theta = \theta_0$ with size at most α . Let $A(\theta_0)$ be the acceptance region based on T . Then a $100(1 - \alpha)\%$ confidence set for θ based on the observation $(x_1, \dots, x_n)$ is

$$C(x_1, \dots, x_n) = \{\theta_0 : T(x_1, \dots, x_n) \in A(\theta_0)\} \quad (6)$$

A widely used method of constructing test-based confidence intervals is the tail method, which inverts two separate one-sided tests each having size at most $\alpha/2$: one for testing $H_0: \theta = \theta_0$ vs. $H_{11}: \theta < \theta_0$ and the other for testing $H_0: \theta = \theta_0$ vs. $H_{12}: \theta > \theta_0$. Suppose the distribution of T is monotone in θ , and the tail regions in favor of H_{11} and H_{12} are $\{T \leq T_{\text{obs}}\}$ and $\{T \geq T_{\text{obs}}\}$, respectively, where T_{obs} is the observed value of T . Then a $100(1 - \alpha)\%$ test-based confidence interval $[\theta_L, \theta_U]$ can be obtained by solving the two equations:

$$\Pr(T \leq T_{\text{obs}}; \theta_U) = \alpha/2 \quad \text{and} \quad \Pr(T \geq T_{\text{obs}}; \theta_L) = \alpha/2 \quad (7)$$

This tail method is frequently used in constructing “exact” confidence intervals (see, e.g., Clopper and Pearson,^[9] Cornfield,^[10] Santner and Snell,^[11] Zelterman,^[12] Chan and Zhang,^[13] Agresti and Min^[14]). Here the word “exact” refers to fact that the tail probabilities in Eq. 7 are evaluated under the exact sampling distribution of the test statistic. Some approximate confidence intervals such as Wilson’s score interval^[6] for a single binomial proportion and the confidence interval for the difference or ratio of two proportions proposed by Miettinen and Nurminen^[15] can also be considered test-based intervals obtained using the tail method. By construction, the tail method (Eq. 7) yields equal-tailed confidence intervals. If the distribution of T is continuous, then the tail method yields confidence intervals with coverage probability $1 - \alpha$ at all θ . When T is discrete, however, $1 - \alpha$ is the lower bound and thus the confidence interval is generally conservative.^[13,14]

For binomial parameters, Chen^[16] and Agresti and Min^[14] have recently proposed constructing test-based two-sided confidence intervals by inverting one two-sided test for $H_0: \theta = \theta_0$ vs. $H_1: \theta \neq \theta_0$, and showed the resulting confidence intervals are generally narrower than those obtained by inverting two one-sided tests, especially if the underlying distribution is not symmetric. Their confidence intervals preserve the overall coverage probability at $(1 - \alpha)$. However, they are usually not equal-tailed, and thus may not maintain consistency of inference if used in conjunction with one-sided hypothesis test as discussed below.

Table 1 Duality Between Hypothesis Test and Confidence Interval (CI) for Achieving Consistent Statistical Inferences in Comparing Two Treatment Groups Under Different Study Designs

Type of study	Statistical hypothesis on the difference ($\theta = \theta_T - \theta_C$)	Significance level of test	Corresponding criterion for rejecting H_0 based on a test-based CI for the difference θ
Difference	$H_0: \theta = 0$ vs. $H_1: \theta \neq 0$	Two-sided α	Entire two-sided $(1 - \alpha)100\%$ CI excludes 0
Superiority	$H_0: \theta \leq 0$ vs. $H_1: \theta > 0$	One-sided $\alpha/2$	One-sided $(1 - \alpha/2)100\%$ CI lower bound > 0 ; or two-sided $(1 - \alpha)100\%$ CI lower bound > 0
Super superiority	$H_0: \theta \leq \Delta$ vs. $H_1: \theta > \Delta$	One-sided $\alpha/2$	One-sided $(1 - \alpha/2)100\%$ CI lower bound $> \Delta$; or two-sided $(1 - \alpha)100\%$ CI lower bound $> \Delta$
Noninferiority	$H_0: \theta \leq -\Delta$ vs. $H_1: \theta > -\Delta$	One-sided $\alpha/2$	One-sided $(1 - \alpha/2)100\%$ CI lower bound $> -\Delta$; or two-sided $(1 - \alpha)100\%$ CI lower bound $> -\Delta$
Equivalence	$H_0: \theta \leq -\Delta$ or $\theta \geq \Delta$ vs. $H_1: -\Delta < \theta < \Delta$	Two one-sided each at α level	One-sided $(1 - \alpha)100\%$ CI lower bound $> -\Delta$ and one-sided $(1 - \alpha)100\%$ CI upper bound $< \Delta$; or entire two-sided $(1 - 2\alpha)100\%$ CI falls within (to Δ, Δ)

Note: Except for “difference” studies, all two-sided confidence intervals must be equal-tailed in order to maintain the consistency of inference with the corresponding hypothesis tests.

RELATIONSHIP WITH HYPOTHESIS TEST

There is a duality between confidence intervals and hypothesis tests. As described in the previous section, confidence intervals can be constructed based on a hypothesis testing procedure. These test-based confidence intervals are consistent with the hypothesis testing in the sense that both will lead to the same conclusion about the parameter of interest. Similarly, a hypothesis test procedure can be derived from a confidence interval.^[17] Suppose $C(x_1, \dots, x_n)$ is a $100(1 - \alpha)\%$ confidence interval for θ . Consider testing

$$H_0: \theta \in \Omega_0 \quad \text{vs.} \quad H_0: \theta \notin \Omega_0 \quad (8)$$

Then the test that rejects H_0 if and only if $C(x_1, \dots, x_n) \cap \Omega_0$ is a level α test for H_0 .

The duality between confidence intervals and hypothesis tests has a very important implication in making statistical inference. Many studies are designed to test a specific hypothesis, and confidence intervals are routinely provided to quantify the precision of the estimated parameter(s). Based on this duality, one can achieve the same conclusion whether the statistical inference is based on a hypothesis test or the corresponding confidence interval. For example, suppose we are interested in comparing two groups (e.g., test vs. control). Let θ_T and θ_C be the parameters of interest for the test and control groups, respectively, and $\theta = \theta_T - \theta_C$ be the difference between the two groups. Table 1 illustrates the duality between confidence intervals and hypothesis tests that will lead to consistent statistical inferences for comparing two treatment groups based on θ under a variety of study designs. The parameters θ_T and θ_C can be very general, such as the percentages of subjects responding to a treatment, means of some continuous response measures

(on raw or log-scale), log odds ratios, or log hazard rates. The Δ in noninferiority, equivalence, or super superiority studies represents a prespecified, small, positive quantity chosen as the clinically relevant difference or tolerance margin. For equivalence studies, the current recommended (ICH E9 Guideline^[18]) approach of two one-sided tests^[19] is used for the illustration.

CONCLUSION

It is important to note that consistent inferences are maintained only if the same method (such as test statistic or pivot) is used to construct both hypothesis test and confidence interval, or a test-based confidence interval is used in conjunction with the hypothesis test in the same study. For example, in a noninferiority study to compare two binomial proportions ($\theta = \theta_T - \theta_C$), the test statistic proposed by Miettinen and Nurminen^[15] is often used to test the specific hypothesis $H_0: \theta \leq -\Delta$ vs $H_1: \theta > -\Delta$ at the one-sided $\alpha/2$ level. As the Miettinen and Nurminen test uses a constrained maximum likelihood estimate of the variance, a test-based confidence interval using the same constrained maximum likelihood estimation method should be used to ensure the consistency of inference. In this case, a rejection of the null hypothesis at the one-sided $\alpha/2$ level will correspond to the lower bound of the one-sided $(1 - \alpha/2)100\%$ confidence interval, or the lower bound of the two-sided $(1 - \alpha)100\%$ confidence interval, being greater than $-\Delta$. In the case where a two-sided confidence interval is used with a one-sided hypothesis, the confidence interval should be equal-tailed in order to guarantee that the left-tail error coverage probability is $\leq \alpha/2$ and ensure the consistency of inference with the corresponding one-sided hypothesis test.

REFERENCES

1. Neyman, J. On the problem of confidence intervals. *Ann. Math. Stat.* **1935**, 6, 111–116.
2. Neyman, J. Outline of a theory of probability. *Philos. Trans. R. Soc. Lond. Ser. A* **1937**, 236, 333–380.
3. Bickel, P.J.; Doksum, K.A. *Mathematical Statistics*; Holden-Day: San Francisco, CA, 1977.
4. Rohatgi, V.K. *Statistical Inference*; Wiley: New York, 1984.
5. Lehmann, E.L. *Testing Statistical Hypotheses*, 2nd Ed.; Wiley: New York, 1986.
6. Wilson, E.B. Probable inference, the law of succession and statistical inference. *J. Am. Stat. Assoc.* **1927**, 22, 209–212.
7. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman and Hall: Boca Raton, FL, 1993.
8. Carpenter, J.; Bithell, J. Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Stat. Med.* **2000**, 19, 1141–1164.
9. Clopper, C.J.; Pearson, E.S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **1934**, 26, 404–413.
10. Cornfield, J. In *A Statistical Problem Arising from Retrospective Studies*, Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability IV, Neyman, J., E.D.; California Press: Berkeley, 1956; 135–148.
11. Santner, T.J.; Snell, M.K. Small-sample confidence intervals for $p_1 - p_2$ and p_1/p_2 in 2×2 contingency tables. *J. Am. Stat. Assoc.* **1980**, 75, 386–394.
12. Zelterman, D. *Models For Discrete Data*; Oxford University Press: Oxford, 1999.
13. Chan, I.S.F.; Zhang, Z.X. Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics* **1999**, 55, 1202–1209.
14. Agresti, A.; Min, Y. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* **2001**, 57, 963–971.
15. Miettinen, O.; Nurminen, M. Comparative analysis of two rates. *Stat. Med.* **1985**, 4, 213–226.
16. Chen, X. Confidence Intervals for the Difference of Two Independent Binomial Proportions in Small Sample Cases. *Technical Report*; BARDS, Merck Research Laboratories: Kenilworth, NJ, 2000.
17. Berger, R.L., Hsu, J.C. Bioequivalence trials, intersection-union tests and equivalence confidence sets (with discussion). *Stat. Sci.* **1996**, 11, 283–319.
18. ICH E9 Expert Working Group. Statistical principles for clinical trials: ICH harmonized tripartite guidelines. *Stat. Med.* **1999**, 18, 1905–1942.
19. Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* **1987**, 15, 657–680.

Confounding and Interaction

Catharine B. Stack

Glenmore, Pennsylvania, U.S.A.

Abstract

Confounding is a bias that must be controlled and something that can be addressed both in the design and in the analysis of a study. Interaction, on the other hand, is an inherent quality of the data, something to be explored once the data have been collected. The aim of this entry is to break down the meaning of these two phenomena and discuss how they are related.

INTRODUCTION

Confounding and interaction, or effect modification, are most commonly discussed in the epidemiological literature. For epidemiological studies, which are generally prospective or retrospective, nonrandomized studies, the principle of confounding is particularly critical. Theoretically speaking, confounding is not a concern for large, randomized studies in the pharmaceutical industry. But often randomizations are not effective in achieving balance across all important prognostic factors, particularly when trials are small or when the number of balancing factors is large. In any event, an appreciation for the issues surrounding confounding can aid in the development of informed study designs. An understanding of statistical interaction is relevant to all studies and is critical to performing meaningful analyses of the data once the trial has been conducted.

CONFOUNDING

Typically, in an epidemiological study, researchers are interested in exploring the relationship between a disease of interest and an exposure, or potential risk factor for the disease. In very basic terms, then, a confounding factor is something that is related both to the exposure and to the disease and that mixes with the exposure to distort the observed association between disease and exposure.^[1–3] A confounder may wholly or partially account for an association between the exposure and disease, or it may mask a true underlying association.

Although confounding results in bias, it differs from other biases, such as selection bias and information bias, in that it is not introduced into a study by the investigators or the subjects. Instead, confounding reflects actual underlying associations between the exposure of interest and the confounding factor and between the disease and the confounding factor. Hence, it is a reflection of the causal associations between variables in the study population.

A hypothetical example will help explain this concept and introduce a discussion of the criteria generally used to assess confounding. Suppose a clinical trial is conducted with 200 subjects and that 100 subjects are randomized to receive drug A and 100 are randomized to receive drug B. The primary outcome of the trial is either success or failure. Overall results are contained in [Table 1](#). These results would indicate that drug A is more effective than drug B because the success rate is 71% for drug A vs. 52% for drug B. Applying a chi-square test of no association to these data results in a p value of 0.006. Results of a more standard epidemiological analysis give an estimated odds ratio, or ratio of odds of success for drug A to that for drug B, of 2.26 with a 95% confidence interval of (1.26, 4.05).

However, when these data are tabulated by the level of a potential confounder or prognostic factor, age (< 70 vs. ≥ 70), a different conclusion emerges. Other examples of possible confounders would be smoking status, gender, previous therapy, and functional status at baseline. [Table 2](#) contains these results broken down by age category. For those subjects under 70, the success rate is 90% regardless of treatment, and for those subjects 70 or older, the success rate is 27% regardless of treatment. Accordingly, the estimated odds ratio is 1 in each strata, with 95% confidence intervals of (0.52, 1.94) for those under 70 and (0.37, 2.69) for those aged 70 and older. Hence, the apparent treatment difference in [Table 1](#) is spurious, due to age, which appears to be a confounder.

Evaluating Confounding

The general criteria used for assessing confounding^[1,3] specify that a confounder:

1. Must be associated with the exposure of interest.
2. Must be associated with the disease or outcome of interest.
3. Must not be a consequence of the exposure or of an intermediate step in the causal pathway of interest.

From criterion 1 we see that an extraneous factor can confound the relationship between treatment and outcome only if it is not evenly distributed in the treatment groups. In regard to the example depicted in [Tables 1](#) and [2](#), criterion 1 requires the proportion of subjects under 70 to differ in the two treatment groups. This criterion is in fact met in this example, because 70% on drug A are under 70 whereas 40% on drug B are in this younger age group.

Similarly, as required by criterion 2, the potential confounder must also be related to the outcome under study. Regardless of how unevenly the potential confounder is distributed in the treatment groups, unless it is also related to outcome there can be no confounding. If the potential confounder is not associated with outcome, there is nothing to mix or add to the association between the treatment and the outcome. In our example, age category, too, is related to outcome, because the success rate among those in the younger age group is 90% as opposed to 27% for those in the older group.

Criterion 3 is most relevant to retrospective epidemiological studies, where study populations are not followed prospectively. For studies in the pharmaceutical industry, this criterion is equivalent to stating that a confounder may not be a postbaseline factor or an intermediate result of the treatment under study. In our example, age certainly is not a consequence of treatment, and age category satisfies all criteria of confounding.

Once a particular risk factor satisfies the criteria for being a confounder, the next step is to evaluate the degree of confounding present. One way to do this is to compare the crude and adjusted estimates of association. For example, an estimated odds ratio calculated from the pooled data may be compared to individual odds ratios calculated within each stratum, or level, of the confounder. Alternatively, the crude estimate of association may be compared to a summary estimate of association that “adjusts” for the

confounder. Methods for finding such adjusted measures of association will be given under the section “Controlling Confounding Through Data Analysis.”

Referring back to our example, we can assess the degree of confounding introduced by age group by comparing the crude odds ratio, 2.26, to the stratum-specific odds ratios, which were both 1. For these hypothetical data, then, one would not want to base conclusions on an unadjusted analysis. The disparity between conclusions made based upon pooled data, as in [Table 1](#), and those made based upon stratified data, as in [Table 2](#), is an effect known as Simpson’s paradox.^[4] In practice, the amount of confounding requiring adjustment is a subjective judgment on the part of the investigator. There is no generally agreed-upon quantitative procedure for determining whether or not a given variable suspected of confounding should be controlled.

Significance tests should not be used to assess confounding.^[1,3] In particular, the *p* values associated with a standard table of baseline characteristics should not be used to determine whether or not a factor is a confounder. The results of these tests are heavily dependent on the size of a study rather than on the degree of confounding present. In addition, for a factor to satisfy the criteria of a confounder it must also be related to outcome. There are no tests able to assess the association of a potential confounder with both treatment group and outcome.

Controlling Confounding Through Design

Of course, the best way to control for confounding is through prevention. All potential confounders should be carefully considered at the time the study is designed. The most effective technique employed in the biopharmaceutical industry to control confounding is randomization. The goal of randomization is to ensure that each eligible subject is equally likely to be assigned to each treatment group. But randomizations are not always perfect. For small clinical trials, an imbalance in one or more prognostic factors is likely, and the resulting confounding can be problematic, given that sample sizes have been estimated assuming a large treatment effect. Even for a very large trial, where profound imbalance in prognostic factors is unlikely, perfect balance is also unlikely, and uncontrolled confounding can potentially affect trial results.^[8]

If a disease outcome under study has known, strong prognostic factors that could act as confounders, then researchers may take additional steps to ensure balance among these factors. One method is through restriction, whereby investigators include only patients with a certain prognostic factor in the trial. Because of the poor generalizability of the results from such a trial, this method may be used only for very small—say, Phase II—trials or in studies where the known relationship between the prognostic factor and disease is much stronger than the expected relationship between the experimental treatment

Table 1 Overall Success and Failure by Treatment

	Success	Failure	Total
Drug A	71	29	100
Drug B	52	48	100

Table 2 Results by Age Category

	Success	Failure	Total
Age < 70			
Drug A	63	7	70
Drug B	36	4	40
Age ≥ 70			
Drug A	8	22	30
Drug B	16	44	60

and disease. For example, in oncology trials, performance status at baseline is highly predictive of outcome. Hence, most cancer trials include some sort of performance level cutoff as part of its inclusion/exclusion criteria.

A stratified randomization might also be employed, whereby treatment assignments are made separately for each stratum, and balance is preserved within each level of the potential confounder. For example, in a study of obesity, where patterns of weight loss may differ by sex, one might chose to stratify by sex to ensure an equal distribution of males and females within each treatment group. Again, such a procedure is effective only for relatively small number of strata, particularly in multicenter trials. When such a stratified randomization is employed, the model used for analysis should still include the stratification factors.

One common method applied in the design of epidemiological studies to control confounding is matching. In a matched study, the investigator intentionally forces the comparison group to have the same or similar distribution as the index group with regard to a potential confounder or confounders. For example, in a retrospective study, controls, or subjects without disease, may be matched to cases or subjects with disease, based upon gender and 5-year age group (e.g., 40–45, 45–50, 50–55). A similar method, called minimization, has been proposed for use in clinical trials.^[5–8] Minimization assigns treatments to subjects sequentially, based upon each subject's individual stratification factors, so that balance may be achieved simultaneously across several prognostic factors. Treatment assignment is made by finding the assignment that “minimizes” imbalance over all prognostic factors considered. This method for assigning treatment is popular for clinical trials in oncology, where there are typically more factors to balance over, such as tumor histology, tumor stage, cell type, prior chemotherapy, and prior radiation.^[8] This method may also be useful for large, multicenter trials conducted internationally.

The primary disadvantage to such a sequential or dynamic assignment scheme is that randomization lists and prepackaged blinded drugs may not be prepared in advance or may not be allocated in numerical sequence. With the advent of telerandomizations and refined Internet communications, however, this method may become more of a practical reality.

Of course, the randomization employed is only as effective as the final comparison conducted. Only those groups assigned by randomization are truly comparable with regard to all potential confounders. Hence, comparing any other groups at the time of analysis defeats the purpose of the randomization. That is, a true intent-to-treat analysis should be conducted whenever possible. For example, bias may be introduced by comparing only those subjects who have completed a study or those with final observations. In such a situation, where final comparison groups differ from the initial groups randomized, methods adjusting for informed dropouts should be investigated.

Controlling Confounding Through Data Analysis

Once a strong confounder has been detected, the simplest way to control for this factor is to perform an analysis within each stratum of the confounder, or by performing a subgroup analysis. By performing such a subgroup analysis, one essentially finds estimates of the association between treatment and outcome, adjusted for the joint effect of the confounders. However, calculating group-specific estimates of association is not the most efficient method of analysis, and often a summary measure across subgroups is required.

Many methods exist for finding point estimates, significance tests, and confidence intervals on summary odds ratios.^[1] The most common method for conducting a test of association while controlling for other factors is to apply a Mantel–Haenszel chi-square test. The Mantel–Haenszel summary odds ratio is a weighted average of the stratum-specific odds ratios, and the Mantel–Haenszel chi-square statistic is used to test a null hypothesis of no association (summary odds ratio = 1). When the direction of association varies across strata, the Mantel–Haenszel test has very little power for detecting departures from no association, and opinions vary on the utility of a summary measure of association. One may choose to assess quantitatively the heterogeneity of stratum-specific odds ratios by applying a test. Again, many such tests exist, such as those proposed by Breslow and Day or Woolf,^[1] although all are unstable when sample sizes within strata are small.

As the number of potential confounders increases or the number of possible levels for a confounder increases, controlling confounding through stratification presents problems unless sample sizes are very large. In this setting, a more unified approach involving the fitting of a linear model with terms for each confounder is preferred. For dichotomous outcomes, such as success rates, a linear logistic model may be used; for continuous outcomes, such as a change from baseline in a laboratory measurement, an ANOVA/ANCOVA model may be employed.

For example, suppose that the outcome under study is a success rate and that x_1 represents treatment ($x_1 = 1$ for drug A, $x_1 = 0$ for drug B) and x_2 through x_p are potential confounders. Performing an analysis that controls for x_2 through x_p involves fitting the linear logistic model:

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (1)$$

where p is the probability of success and the logit (p) is defined as $\log [p/(1-p)]$. If the outcome is continuous, an analysis that controls for potential confounders involves fitting the ANOVA/ANCOVA model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (2)$$

where y is the continuous outcome measure.

Strong confounders will impact the estimate of β_1 , and by including them in such a linear model we are able to

find an estimate for β_1 that is adjusted for the confounding factors. For example, comparing the estimate of β_1 from the following models (3) and (4) provides an indication of the degree of confounding introduced by factor x_2 :

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 \quad (3)$$

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad (4)$$

Therefore, β_1 as estimated by model (4) may be interpreted as the treatment effect adjusted for factor x_2 . In fact, the logistic regression analysis in model (4) is equivalent to the Mantel–Haenszel analysis presented previously.

When controlling for confounding at the time of analysis, it is important to include all strong confounders, but including factors that have little impact on reducing bias will only increase the variability associated with the estimated treatment effect.^[9] Generally, factors should be included in a model only if they are known from prior studies to be causally related to outcome or if they have been used as a stratifying factor in the design of the trial.

INTERACTION

Interaction is a term referring to a change in the magnitude or direction of the association between treatment and outcome according to the level of a third variable. For example, if a drug is more effective in younger patients than in older patients, there would be an interaction between treatment and age. In this situation, the treatment effect is not independent of age. In the epidemiological literature, this phenomena is often referred to as “effect modification,” because this third variable “modifies the effect” of treatment on outcome. The subtle distinction, however, is that effect modification generally implies causality between the effect modifier and outcome, whereas “statistical interaction” does not.^[10]

Unlike confounding, interaction is not a bias or something that must be controlled in a study. Instead, interaction is something that should be explored so as to gain a greater understanding of the association between treatment and outcome. Interaction is a natural phenomenon occurring in the underlying population from which study participants are selected and will not vary based upon the design of the study.

Evaluating Interaction

An interaction effect is meaningful only within the context of a particular model. Suppose we are analyzing data from a trial with a continuous primary outcome of change from baseline in weight. The variable x_1 corresponds to treatment group (A or B), and x_2 corresponds to gender (male or female). An interaction between treatment and gender may be investigated by fitting the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma_1 (x_1 x_2) \quad (5)$$

where y represents the change from baseline in weight and γ_1 represents the interaction term between treatment and gender. If γ_1 is found to be significantly different from 0, then there is evidence of an interaction between gender and treatment, indicating that the effect of treatment and gender on the mean change from baseline in weight is non-additive. In other words, the joint effect of treatment and gender is different from the sum of their independent effects.

Often it is helpful to describe such interactions graphically. Plotting mean changes from baseline by treatment and gender can help in assessing the possible interactions present. For example, Fig. 1a depicts a situation in which the interaction is simply additive, Fig. 1b shows a deviation from an additive interaction in which response to drug B always exceeds that to drug A, and Fig. 1c depicts a situation in which response to drug B exceeds that to drug A in males, but response to drug A exceeds that to drug B in females.

Now suppose that our trial has a discrete outcome of success or failure. An interaction between treatment and gender may be investigated by fitting the model

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma_1 (x_1 x_2) \quad (6)$$

where the p is the probability of success and γ_1 again represents the interaction between treatment and gender. If γ_1 is found to be significantly different from zero in this model, we conclude that the relationship between success rate and treatment differs by gender or, using more standard epidemiological terminology, that the effect of treatment and gender on the log-odds for success is non-multiplicative. In other words, the joint effect of treatment and gender on the log-odds for success is different from the product of their independent effects.

For pharmaceutical studies, the interaction terms that are of most interest are those involving the treatment group or dose level. For example, in a dose–response analysis, one might want to examine whether or not the dose–response relationship varies by some third variable. This could be done by incorporating an interaction term in the model and evaluating its impact on model fit. Or, in a large, multicenter trial, one might want to test for a treatment-by-center effect.

Interaction terms including only prognostic factors are most useful in controlling interrelated confounding effects. In terms of analysis, it is recommended that one first consider and evaluate potential confounders, and then incorporate any interactions present. This order of analysis is recommended because the primary aim of most analyses is to assess the relationship between outcome and treatment, and including interaction terms with treatment while assessing confounding complicates this determination.

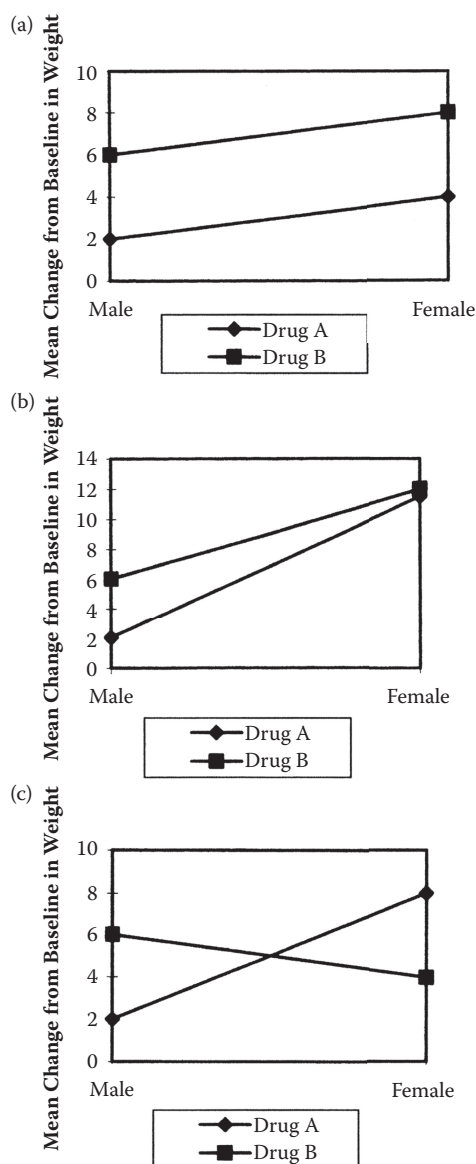


Fig. 1 (a) Additive interaction between treatment and gender. (b) Nonadditive interaction between treatment and gender in which response to drug B always exceeds that to drug A. (c) Nonadditive interaction between treatment and gender in which response “changes direction”

REGULATORY REQUIREMENTS

Regulatory guidelines do not specifically address the principle of confounding, but they do discuss the need for randomization in the design of clinical trials. Section 2.3.2 of the current version of the *Guideline on Statistical Principles for Clinical Trials*, proposed by the International Conference on Harmonization,^[11] discusses the role of randomization in clinical trials. These guidelines cite randomization as one of the critical techniques that should be adopted so as to provide “a sound statistical basis for the quantitative evaluation of the evidence relating to treatment effects.” In addition, the ICH recommends stratifying

on study center and other important prognostic baseline factors, particularly in small studies. Dynamic allocation procedures are discussed and advocated only after careful consideration of “the complexity of the logistics and potential impact on the analysis.”

Section 5.7 of ICH9 addresses the use of covariates, or possible confounders, and interaction terms in the data analysis. The guidelines stress the importance of “prestudy deliberations” concerning the potential covariates affecting outcome and advise that potential confounders be included in the analysis plan whenever possible. They also suggest, “when the potential value of an adjustment is in doubt, it is often advisable to nominate the unadjusted analysis as the one for primary attention, the adjusted analysis being supportive.”

In the discussion of interaction, the ICH guidelines recommend prespecification of any planned subgroup or interaction analyses in the analysis plan, adding that “in most cases, however, subgroup or interaction analyses are exploratory and should be clearly identified as such.” They go on to caution that “any conclusion of treatment efficacy (or lack thereof) or safety based solely on exploratory subgroup analyses are unlikely to be accepted.”

CONCLUSION

Confounding and interaction are different yet related phenomena. Confounding is a bias that must be controlled and something that can be addressed both in the design and in the analysis of a study. Interaction, on the other hand, is an inherent quality of the data, something to be explored once the data have been collected. A prognostic factor may be both a confounder and an interaction term, one of these, or neither. Confounding must be assessed qualitatively; interaction may be tested more quantitatively.

When designing a study, the most effective approach to block confounding is to employ randomization. To reduce further the possibility of confounding for known strong potential confounders, study subjects may be restricted by the level of the potential confounder, or a stratified or sequential treatment assignment may be followed.

When analyzing the results of a study, unexpected confounding can be controlled by fitting linear models with terms for such factors as well as any stratification factors used in the design of the study. Such a model may also be used to explore any interactions with treatment, or differences in treatment response based upon the level of another variable.

At some level, the goal of every study is to assess the true association between treatment and outcome. Control of confounding improves the estimate of this underlying association. Identification of treatment interaction terms may affect product labeling and prescribing, in the later phases of research, or may assist in targeting populations for further research, in the earlier stages.

REFERENCES

1. Schlesselman, J.J. *Case-Control Studies: Design, Conduct, Analysis*; Oxford University Press: New York, 1982.
2. Miettinen, O.S. Confounding and effect-modification. *Am. J. Epidemiol.* **1995**, *141*, 1113–1116.
3. Tongzhan, Z. *Principles of Epidemiology: Fundamentals and Study Design, Course notes*; Yale University School of Medicine: New Haven, CT, 1993.
4. Simpson, E.H. The interpretation of interaction in contingency tables. *J. R. Stat. Soc. B.* **1951**, *13*, 238–241.
5. Taves, D.R. Minimization: a new method of assigning patients to treatment and control groups. *Clin. Pharmacol. Ther.* **1974**, *15*, 443–453.
6. Pocock, S.J.; Simon, R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics.* **1975**, *31*, 103–115.
7. Freedman, L.S.; White, S.J. On the use of Pocock and Simon's method for balancing treatment numbers over prognostic factors in the controlled clinical trial. *Biometrics.* **1976**, *32*, 691–694.
8. Treasure, T.; MacRae, K.D. Minimisation: the platinum standard for trials? Randomization doesn't guarantee similarity of groups; minimisation does. *Br. Med. J.* **1998**, *317*, 362, 363.
9. Day, N.E.; Byar, D.P.; Green, S.B. Overadjustment in case-control studies. *Am. J. Epidemiol.* **1980**, *112*, 696–706.
10. Pearce, N. Analytical implications of epidemiological concepts of interaction. *Int. J. Epidemiol.* **1989**, *18*, 976–980.
11. *International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use—ICH Harmonised Tripartite Guideline E9: Statistical Principles for Clinical Trials*; Feb. 5, 1998.

Content Uniformity

John R. Murphy

Eli Lilly and Company, Indianapolis, Indiana, U.S.A.

Abstract

Content uniformity is a measure of the variation in the amount of active ingredient from one dosage to the next. This entry is concentrated on the process and importance of formal dose uniformity tests and discusses the different factors that can present problems in content uniformity of drug dosages.

INTRODUCTION

Content uniformity is a measure of the variation in the amount of active ingredient from one dosage to the next. In a perfect world, every dosage unit would contain exactly the same amount of drug. However, in the real world of manufacturing tablets, capsules, or other final dosage form, the active drug substance is diluted with excipients to mask taste, to enhance bioavailability, to add bulk, or to promote stability. In such processes, it is not always possible to obtain absolute homogeneous mixing of the drug with the excipients. Factors such as different densities, different particle sizes, and different particle shapes contribute to different settling tendencies and flow characteristics, which may cause slight variations in the amount of active ingredient present in the dose-sized partitions of the batch. Even if absolute homogeneity in the formulated drug product can be obtained, uniformity of active ingredient content in the final dosage forms would still not result because it is not possible to fill every capsule or to compress every tablet to contain exactly the same weight of drug product. In order to achieve adequately homogeneous mixtures of bulk drug product, manufacturers validate the mixing process with respect to time and other mixing parameters. However, even with a validated process, it is still necessary to explicitly demonstrate adequate content uniformity in the final dosage units on a batch-by-batch basis. This is usually carried out with a formal dose uniformity test.

OVERVIEW

All three major pharmacopeias, the United States Pharmacopeia and National Formulary (USP/NF),^[1] European Pharmacopeia (EP),^[2] and the Japanese Pharmacopeia (JP)^[3] have formal criteria for judging the uniformity of dosage units. While the tests are somewhat similar, there are important differences that can lead to cases where a product sample might meet the requirements of one test but not the criteria of the other. [Table 1](#) provides a

side-by-side comparison of the requirements of the three pharmacopeia. The reader will note that the wording in [Table 1](#) is different from the descriptions of the test given in the applicable pharmacopeia. This has been performed in order to facilitate comparison of the tests. Inspection of the three tests shows that the criteria in both the USP and the JP tests are tied to the mean content, in addition to controlling the allowable variation from the mean, while the EP deals only with the variation in the content of individual units from the mean content.

Although these tests should not be confused with a sampling and acceptance plan for lot release with respect to content uniformity, their performance can be evaluated using operating characteristic (OC) curves, which provide the probability that the test will be passed versus the hypothesized true mean and relative standard deviation (RSD) of the product. [Figs. 1–3](#) show the OC curves for the three compendial tests for a range of true means and RSD values. The curves can be generated using Monte Carlo simulation. As illustrated in [Fig. 1](#), when the true mean content is near 100% of label claim, the JP test and both USP tablet and capsule tests behave similarly, while the EP test is different. In [Fig. 2](#), where the true mean is less than 100%, the JP test is less likely to be passed than the USP tests because it places more emphasis on deviations from 100% of label than the USP. The OC curves of the EP are nearly the same as before because the operating characteristics are mostly unaffected by changes in the true mean content. [Fig. 3](#) illustrates the provision of the USP for cases when the rubric mean (average of the potency limits) is more than 100%. Here the JP is much tighter than the USP and EP because it employs similar treatment for deviations in both the high side and the low side of 100%. In these cases, the USP is seen to perform in a middle ground between the JP and the EP. Other cases of interest could be generated and examined in this way through the use of simulations and OC curves.

In the near future, pharmacopeial harmonization will lead to more common global compendial criteria. The tests for uniformity of dosage units is slated for harmonization,

Table 1 Requirements of the Three Major Pharmacopeia

	USP/NF	EP	JP
Sample sizes	10 + 20	10 + 20	10 + 20
Definition of M	M = greater of $\bar{X}$ and 100% and lesser of $\bar{X}$ and rubric mean	$M = \bar{X}$	$M = 100\%^a$
Capsules, etc.	$n1 = 1$ and $n2 = 3$	$n1 = 1$ and $n2 = 3$	–
Tablets, etc.	$n1 = 0$ and $n2 = 1$	$n1 = 0$ and $n2 = 1$	–
Stage 1	Pass: $n \leq n1$ and $RSD \leq 6.0\%$ and $m = 0$ Fail: $n > n2$ or $m > 0$	Pass: $n \leq n1$ and $m = 0$ Fail: $n > n2$ or $m > 0$	Pass: $\bar{X} - 2.2s \geq M - 15\%$ and $\bar{X} + 2.2s \leq M + 15\%$ and $m = 0$ Fail: $m > 0$
Stage 2	Stage 2: $n1 < n \leq n2$ Pass: $n \leq n2$ and $RSD \leq 7.8\%$ and $m = 0$ Fail: $n > n2$ or $m > 0$ or $RSD > 7.8\%$	Stage 2: $n1 < n \leq n2$ Pass: $n \leq n2$ and $m = 0$ Fail: $n > n2$ or $m > 0$	Stage 2: Neither pass nor fail Pass: $\bar{X} - 1.9s \geq M - 15\%$ and $\bar{X} - 1.9s \leq M + 15\%$ and $m = 0$ Fail: $m > 0$ or $\bar{X} - 1.9s < M - 15\%$ or $\bar{X} + 1.9s > M + 15\%$

Rubric mean = average of potency limits in the monograph; n = number of assays outside 85–115% of M; m = number of assays outside 75–125% of M; RSD = relative standard deviation.
^aUnless otherwise specified in the individual monograph.

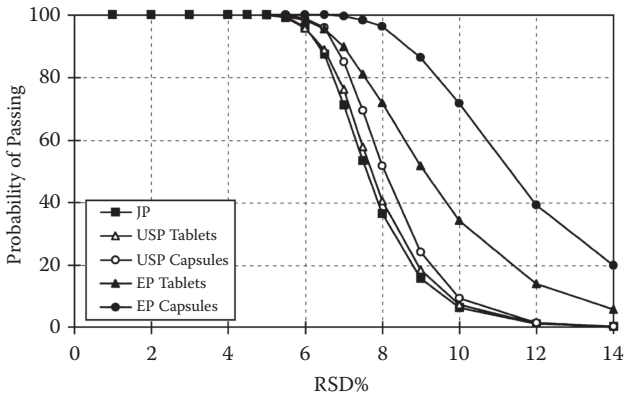


Fig. 1 Operating characteristic curves for current JP, USP, and EP when rubric mean = 100% and actual mean = 100%

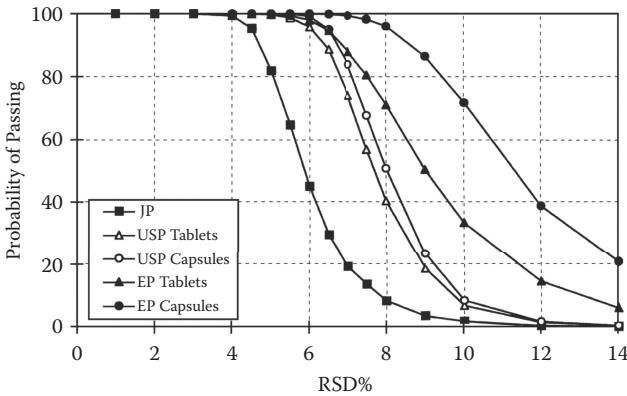


Fig. 3 Operating characteristic curves for current JP, USP, and EP when rubric mean = 104% and actual mean = 104%

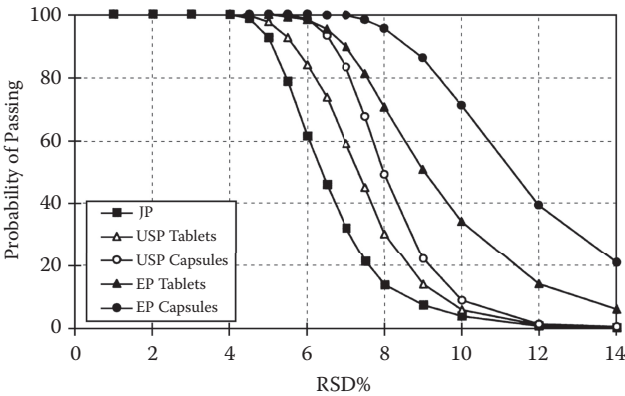


Fig. 2 Operating characteristic curves for current JP, USP, and EP when rubric mean = 100% and actual mean = 96%

and the next few years should see a single compendial test for uniformity of dosage units that is applicable to products in most major markets of the world. It is a natural extension of the concept of content uniformity to ask whether a simpler and more straightforward measure of variability, such as RSD, would be more appropriate for judging compliance. Even if the location of the mean content is kept in the equation, criteria similar to those in the JP have appeal because they are analogous to statistical tolerance limits. In such case, the criterion for passing the test is that there be a given level of confidence that a sufficiently high proportion of the dosage units fall within, e.g., 85–115% of label.

The expectation that content uniformity will be maintained throughout the shelf life of the product leads to the question of whether a manufacturer would have product

release criteria different from the pharmacopeial tests. By using the OC curves of the compendial tests, one could compare a candidate release criterion to see how it compares with the official test. Another way the OC curves of the compendial tests can be used is to determine the likelihood that the compendial requirements will be met for a product where the mean content and the variation of content is known. Bergum^[4] developed methods for determining acceptance limits for the sample mean and standard deviation that give a specified probability that the USP will be met. Sampson et al.^[5,6] examined alternate ways to assess content uniformity by exploiting the relationship between weight and content of the dosage unit, and they proposed sampling and acceptance strategies that could be used for product release.

CONCLUSION

In summary, content uniformity is an important product quality attribute that provides the patient assurance that no matter which tablet or capsule is taken, an acceptable amount of active drug will be delivered. Manufacturers perform validation to demonstrate that the mixing and finishing operations produce uniform dosage units with

respect to the content of active ingredient. Where applicable, batches are tested for content uniformity prior to release to provide assurance that the product will comply with compendial standards of uniformity. Currently, different compendial requirements are applied in different parts of the world, but international harmonization of the tests for uniformity of dosage units is under way.

REFERENCES

1. *United States Pharmacopeia, The United States Pharmacopeia XXIV and the National Formulary XIX*. the United States Pharmacopeial Convention Inc.: Rockville, MD, 2000.
2. *European Pharmacopeia*, 3rd Ed.; European Department for the Quality of Medicines, Council of Europe: Strasbourg, France, 1996.
3. *Japanese Pharmacopeia*, 3th Ed.; Society of Japanese Pharmacopeia: Tokyo, 1996.
4. Bergum, J.S. Constructing acceptance limits for multiple tests. *Drug Devel. Ind. Pharm.* **1990**, *16*, 2153–2166.
5. Sampson, C.B.; Breunig, H.L. *J. Qual. Technol.* **1971**, *3*, 170–178.
6. Sampson, C.B.; Comer, H.L.; Broadlick, D.E. *J. Pharm. Sci.* **1970**, *59*, 1653–1655.

Continual Reassessment Methods

Xue Lin

US Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

This entry introduces the Continual Reassessment Methods and its variations.

Keywords: Continual Reassessment Methods; Bayesian; Phase I clinical trials; dose escalation.

INTRODUCTION

In the drug development process, a drug usually needs to go through a series of clinical trials to show its efficacy and safety. And the first step in this process is usually to find a safe and efficacious dose. Phase I studies are usually for this purpose.

It is often assumed that as dose increases, the drug becomes more efficacious, however, at the same time, its toxicity increases too. The maximum tolerated dose (MTD) is the most efficacious dose with an acceptable safety profile. And safety is usually assessed by the frequency of dose limiting toxicity (DLT) occurrences.

Before the continual reassessment method was introduced in 1990 by O’Quigley, Pepe et al.,^[1] the so called “3+3” design was widely used to find the MTD. As the name indicates, in this design, cohorts of 3 subjects are treated at a dose level, and depending on the toxicity outcome, additional 3 subjects may be treated at the same dose level, or at a higher dose level. A detailed discussion of the “3+3” design can be found in^[2]. The “3+3” design was criticized for its inefficiency when the drug toxicity is over-estimated and its lack of transparency in the targeted toxicity rate^[1].

Besides “3+3” design, there are other methods for dose escalation. Simon et al.^[3] proposed to use the accelerated titration design in Phase I oncology trials. This study design used inpatient dose escalation and thus efficiently escalated the dose to the target level. Collins et al.^[4] proposed to use the real-time PK measurement to guide the dose escalation.

THE CONTINUAL REASSESSMENT METHOD

The general idea

The continual reassessment method is a Bayesian method. The general idea of this method is illustrated in Figure 1.

Figure 1 shows 5 doses. Each dose has a prior probability of DLT as shown in the left panel where the 95% credible interval is represented by a line segment at each dose. The horizontal dotted line represents the target DLT probability, which is 30% in this example. Once the trial begins and toxicity information accumulates, the prior belief on DLT probability gets updated, which results in the posterior distribution as shown in the right panel.

The prior DLT probability distributions at each dose level are different but are assumed to have the same parametric form and are functions of the corresponding dose level. Subjects enter the trial sequentially and the toxicity information on a subject is used to update the model parameter. The next subject will be treated at the dose level whose mean or median DLT probability is closest to the target probability.

The original CRM

The original continual reassessment method (CRM) was proposed by O’Quigley, Pepe et al.^[1] This section describes CRM, but first we will introduce some notations.

Suppose there are K dose levels denoted as $(d_1, d_2, \dots, d_K)$, Y_j is a binary variable defined as

$$Y_j = \begin{cases} 1, & \text{if the } j\text{th subject has DTL} \\ 0, & \text{otherwise} \end{cases} \quad j = 1, \dots, N$$

Let $p(d, \alpha)$ be the dose-toxicity function where d is the dose level and α is the unknown model parameter. And we require that $p(d_i, \alpha)$ equals the probability of DLT at dose level d_i , $i = 1, \dots, K$. The goal of the trial is to find the dose level whose mean probability of DLT is closest to the target γ .

Suppose the prior distribution of α is $p_0(\alpha)$. The likelihood function for the first j subjects are the following

$$L(Y_1, Y_2, \dots, Y_j | \alpha) = \prod_{i=1}^j p(d_i, \alpha)^{Y_i} [1 - p(d_i, \alpha)]^{1-Y_i}$$

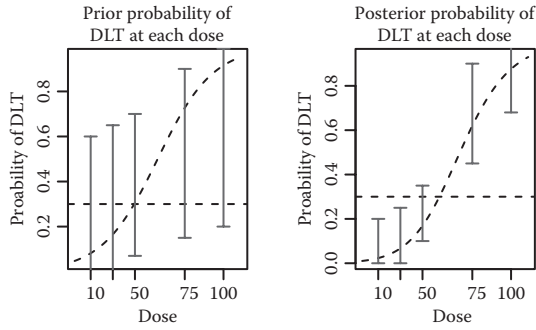


Fig. 1 Prior and posterior probability of dose limiting toxicity (DLT). The two figures show the general idea of Bayesian dose escalation method

Using Bayes' Theorem, we can get the posterior distribution of α as

$$\Pr(\alpha | Y_1, \dots, Y_j) = \frac{p_0(\alpha) * L(Y_1, Y_2, \dots, Y_j | \alpha)}{\int p_0(\alpha) * L(Y_1, Y_2, \dots, Y_j | \alpha) d\alpha}$$

With the updated (posterior) distribution of α , the posterior mean probabilities of DLT at each dose level are as the following

$$\theta_{i,j} = \int p(d_i, \alpha) \Pr(\alpha | Y_1, \dots, Y_j) d\alpha \quad (i = 1, \dots, K)$$

Then a dose level is chosen to minimize the distance to the target DLT probably γ as

$$\arg \min_{i=1, \dots, K} \Delta(\theta_{i,j}, \gamma)$$

where Δ is a distance function such as absolute difference. Then the $(j+1)$ th subject will be treated at this chosen dose level.

To ease the computational burden, the authors also offer alternatives to the direct calculation of $\Delta(\theta_{i,j}, \gamma)$. Instead one could calculate the posterior mean of α and plug it in the dose-toxicity function $p(d, \alpha)$, then find the d_i that minimize the distance between $p(d_i, \alpha)$ and γ , or back calculate d that produce the target DLT probability γ and then find a dose level d_i that is closed to the d .

Widely used dose-toxicity functions are hyperbolic function ($\{(\tanh(x) + 1)/2\}^a$), logistic model ($\frac{e^{2+ax}}{1+e^{2+ax}}$) and power model ($X^{\exp(a)}$).

MODIFIED CRM (mCRM)

Goodman^[5] proposed modifications to the original CRM. The major modifications are as follows:

- Instead of starting with the dose whose prior mean DLT probability is the target DLT probability, always starts with the lowest dose.

- Only escalate to a dose that is one level higher when the algorithm recommends a dose that is more than one level higher
- Treating at least 3 subjects at each dose level instead of 1 as CRM proposed

These modifications to the CRM are out of consideration of safety. Along the same line of thinking are the modifications proposed by Faries^[6] and Piantadosi and Fisher.^[7]

AN EXAMPLE OF mCRM

We illustrate how mCRM works via a simulated trial. We will use Goodman's version of the mCRM method.

Suppose the trial tests 5 dose levels: 10 mg, 50 mg, 75 mg, 100 mg and 125 mg, and their associated DLT probabilities are 5%, 15%, 30%, 45%, and 60% respectively. The target DLT probability is 0.30.

We choose the power model for the dose toxicity function:

$$p(d, \alpha) = d^{\exp(\alpha)}$$

For the prior distribution of α , we choose the Gamma distribution with both parameters equal to 1: $p_0(\alpha) = \text{Gamma}(1, 1)$

The prior distribution of DLT probabilities at each dose level is in Figure 2.

The first three patients enter the trial and they are treated at the lowest dose level which is 10 mg. The subjects' DLT outcomes are simulated from the true toxicity probability, which is 5%, and none of the first three subjects have DLT.

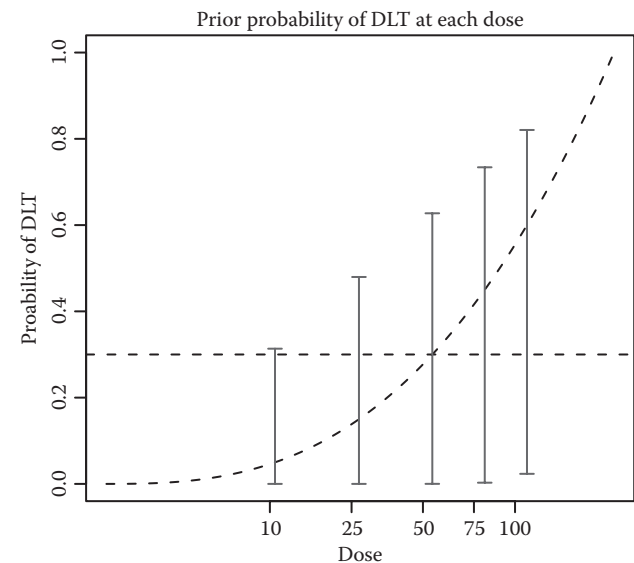


Fig. 2 Prior DLT distribution at each dose level for the simulated trial

Based on the toxicity outcome of these three subjects, the distribution of the dose-toxicity function parameter α is updated and we have a posterior distribution of K . The probability distribution of DLT at each dose level gets updated accordingly. Panel 1 of Figure 3 shows the posterior probability of DLT at each dose level after the toxicity outcome of the first cohort is available.

Although dose level 3 has a mean DLT probability closet to the target 30%, with mCRM, there is no dose skipping, so the next cohort of three subjects are treated at the dose level that is one level higher than the current dose, i.e. dose level 2.

And in the same way, the DLT probability is updated depending on the toxicity outcome for this second cohort and the algorithm recommends a dose level for the next cohort of subjects. Figure 3 shows how the posterior DLT probability distribution at each dose level continuously updated after each cohort of subjects are treated.

When the trial's stopping criteria is reached, which is at least 6 subjects have been treated at the recommended

dose level and a total of 18 subjects have been treated in the trial, the trial stops and the algorithm recommends dose level 3. And this dose level is then the MTD dose chosen and will be used for future Phase II trials.

EXTENSIONS OF CRM

There have been numerous extensions of the original CRM method. The original CRM assumes that drug toxicity monotonically increases as the dose increases. This monotonicity may be compromised in various clinical situations. Thall and Cook^[8] extended CRM to situations where both toxicity and efficacy are considered in dose finding. And Wages, Conaway and O'Quigley^[9] extended CRM to multi-agent trials in which only a partial order of toxicity can be assumed. In the original CRM methods, the dose-toxicity function is assumed and could be mis-specified. In addressing the uncertainty in the dose-toxicity function, Yin and Yuan^[10] proposes the Bayesian Model Averaging

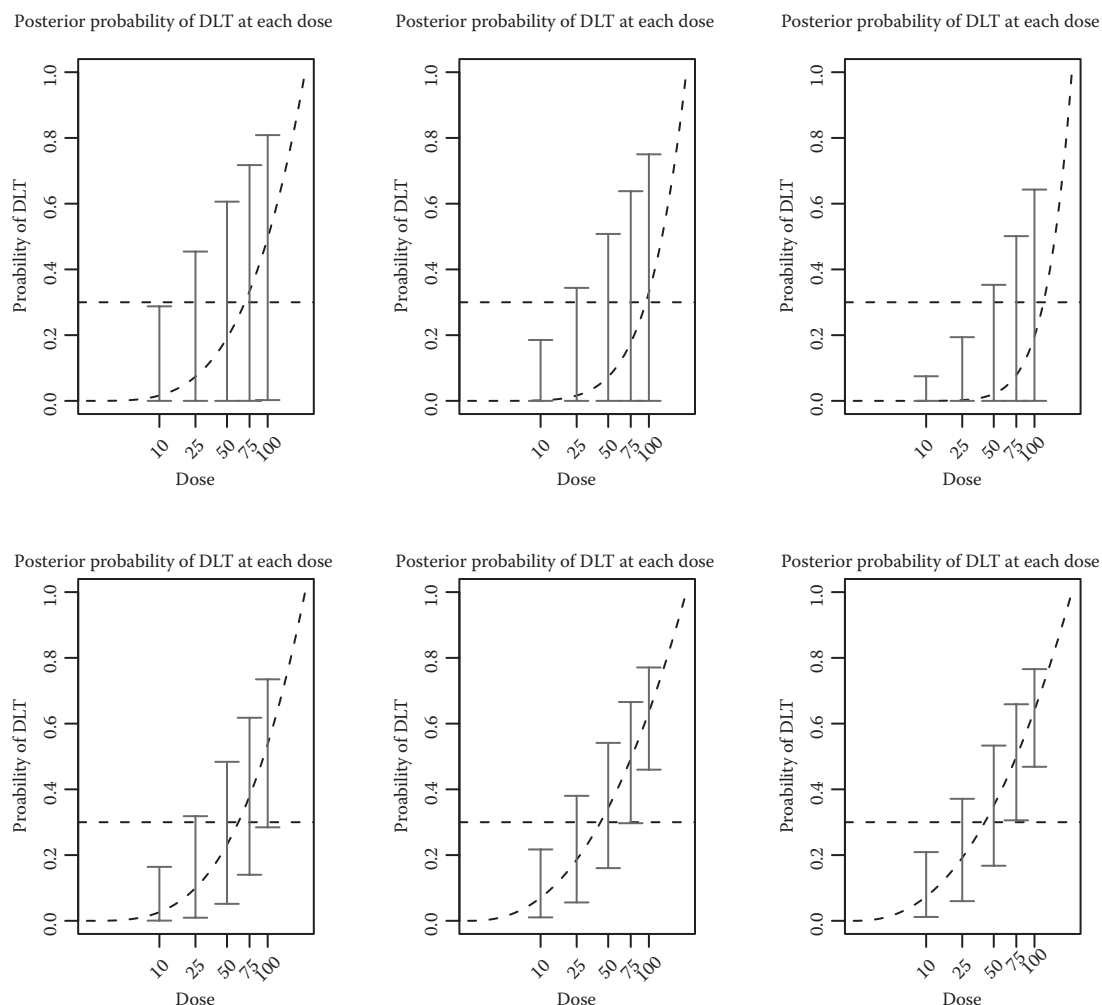


Fig. 3 Posterior DLT probability distribution at each dose level continuously updated after each cohort of subjects are treated

CRM (BMA-CRM). Their simulation study shows that BMA-CRM is robust to misspecification of the dose toxicity function and its improvement over CRM is prominent when the dose toxicity function is mis-specified by large amount.

DISCUSSION

Due to safety concerns, the modified CRM methods are often used in clinical trials instead of the original CRM method. Usually when a trial implements the CRM method or its variants, the operating characteristics of the method needs to be studied via simulation. The simulation needs to include various dose toxicity settings, the method's performance needs to be evaluated on aspects such as, the chances that the right dose is chosen, the proportion of subjects treated at each dose level, especially the percentage of subjects treated at the dose levels whose DLT probability are higher the target DLT, etc.

SOFTWARE

The MD Anderson cancer center has a software Kiosk (<https://biostatistics.mdanderson.org/softwaredownload/>) which provides computer software implementing the CRM and several of its variations. The CRM is also available in R as a package.

REFERENCES

1. O'Quigley, J.; Pepe, M.; Fisher, L. Continual Reassessment Method: A Practical Design for Phase I Clinical Trials in Cancer. *Biometrics* **1990**, *46* (1), 33–48.
2. Storer, B.E. Design and analysis of Phase I clinical trials. *Biometrics* **1989**, *45* (3), 925–937.
3. Simon, R.; Freidlin, B.; Rubinstein, L.; Arbuck, S.G.; Collins, J.; Christian, M.C. Accelerated titration designs for phase I clinical trials in oncology. *J Natl Cancer Inst.* **1997**; *89* (15):1138–1147.
4. Collins, J.M.; Zaharko, D.S.; Dedrick, R.L.; Chabner, B.A. Potential roles for preclinical pharmacology in phase I clinical trials. *Cancer Treat Rep* **1986**, *70* (1), 73–80.
5. Goodman, S.N.; Zahurak M.L.; Piantadosi, S. Some practical improvements in the continual reassessment method for phase I studies. *Stat. Med.* **1995**, *14* (11), 149–61.
6. Faries, D. Practical modificationis of the continual reassessment method for phase I cancer clinical trials. *J Biopharm Stat* **1994**; *4* (2), 147–64.
7. Piantadosi S.; Fisher J.D.; Grossman S. Practical implementation of a modified continuLal reassessment method for dose-findirng trials. *Cancer Chemother Pharmacol* **1998**; *41* (6) 429–436.
8. Thall, P.; Cook, J. Dose-Finding Based on Efficacy–Toxicity Trade-Offs. *Biometrics* **2004**; *60* (3) 684–693.
9. Wages, N. A.; Conaway, M.R.; O'Quigley, J. Continual Reassessment Method for Partial Ordering. *Biometrics* **2011**, *67* (4) 1555–1563.
10. Yin, G.; Yuan, Y. Bayesian Model Averaging Continual Reassessment Method in Phase I Clinical Trials. *JASA* **2009**, *104* (487), 954–968.

Contract Research Organization (CRO): Early 2000s

T. Y. Lee

Independent Consultant, Morristown, New Jersey, U.S.A.

Abstract

The contract research organization has become an important partner to pharmaceutical and biotechnology companies in the development of new compounds. This entry reviews the history and growth of the CRO and touches on its future in the pharmaceutical industry.

INTRODUCTION

Over the years, the pharmaceutical industry has developed a full-scale capability to carry out research from laboratory discovery to biological screening, animal toxicity studies, formulation, clinical pharmacology, stability, clinical trials, new drug applications (NDAs), and promotional marketing. During the late 1970s to early 1980s, the return of investment was so great that little attention was paid by the industry to the reaction of the consumer. Research and development (R&D) costs for the pharmaceutical industry have been doubling every 5 years since 1970; the average cost of bringing a drug to market was \$125 million in 1989 and climbed to \$231 million in 1993.^[1] “It was not unheard of for a potential shortfall in profitability to be remedied simply by a rise in product prices.”^[2] Although prescription drug prices have risen an average of 2.3 times faster than consumer prices since 1980,^[3] the probability of developing a successful drug is very small. According to the Center for the Study of Drug Development at Tuft's University, of every 5000 compounds that are synthesized, just five will enter clinical trials and just one will be approved by the Food and Drug Administration (FDA). As R&D costs increased steadily in the late 1970s, the use of contract research organization (CRO) has become very popular for handling tasks specific to pharmaceutical research and development, especially the conduct, data management, and statistical analysis of clinical studies.

INDUSTRY BACKGROUND

Pressures on the Pharmaceutical Industry from the Early 1990s

The funds from drug development has been decreasing, especially for the high-risk biotechnology products. For biotechnology companies, access to capital was drying up during the early 1990s and was continuing to do

so under the impact of health-care reform. In 1991, biotech companies raised \$6.5 billion; however, in 1992, just \$5 billion was raised. The value of the industry's stock fell 6% to \$45 billion, a loss of \$1.4 billion, while R&D expenses rose 14% to \$5.7 billion.^[4]

1. *Pricing Pressures.* While drug prices continued to outgrow the consumer price index, the public became discontent with the cost of drugs, hospitalization, doctors, and insurance. In early 1990, 37 million Americans lacked coverage for health insurance. In the meantime, the influx of generic drugs, the therapeutic substitution, and the reduction of tax credits for studies designed for marketing purposes have given the pharmaceutical industry a wake-up call. To counter the criticism of ever-rising health-care costs, health-care providers and insurance companies have formed their own respective super large organizations to consolidate their buying power for pharmaceuticals and health services. These powerful health maintenance organizations (HMOs) demand deeper price-cuts from the drug companies. In addition, the government demands rebates on sales to Medicare patients, who are the high-volume users of prescription drugs. Congressman Henry Waxman (D, California) stated, “If we are serious about controlling health-care costs, we must address the issue of prescription drug prices. Congress, as part of health reform, must find a way to balance profits and prices in a way fairer to the American consumer.”^[5]
2. *R&D Cost Increases Faster Than Price.* Despite the pricing pressure (or perhaps because of it), the pharmaceutical industry has continued to increase spending on R&D. The top 30 companies spent \$20 billions on R&D in 1993, representing 18.3% of sales, while 1994 sales grew by only \$6 billion.^[6] R&D costs have been nearly doubling every 5 years since 1970; and the average cost of bringing a drug to the U.S. market rose from \$125 million in 1989

to \$231 million in 1993 (just a little over 4 years). It cost an average of \$450 million to develop a new drug in 1997.

Alternate Approaches to Pharmaceutical Research and Development

In order to stay competitive and to maintain a healthy profit margin, pharmaceutical companies have adopted various approaches to respond to the pressures:

1. *Downsizing and early retirement.* As of July 1993, the pharmaceutical industry had already lost 18,000 jobs (according to the Pharmaceutical Manufacturers Association (PMA)). Many companies offered early retirement packages to reduce the workforce.
2. *Out-licensing nonmajor products.*
3. *Merger and consolidation.* Numerous pharmaceutical companies merged to expand their product lines. Once such mergers were consummated, the layoffs and early retirements followed. Companies that merged during this period included: SKF and Beecham, Dow and Merriam, Eastman Kodak and Sterling Drug, AHP and A.H. Robins, Merck and Johnson and Johnson Joint Venture, Hoffman LaRoche and Genentech, Merck and DuPont Joint Venture, Rh'one-Pulenc and Rorer, Bristol-Myers and Squibb, Roche and Syntax, AHP and American Cyanamid.
4. *Comarketing.* The sales forces of different companies are joining together to penetrate the market and to expand market share. For example, Glaxo Holding PLC worked with Roche on Glaxo's Zantac using Roche's 800-strong sales force. Smith Kline Beecham and DuPont comarketed SKB's Tagamet. Beecham and Upjohn joined forces to sell Beecham's heart-attack drug Eminase.
5. *Developing only the drugs with potential major market shares.*
6. *Developing one's own generic business.* Hoechst brought 51% of Copley Pharmaceuticals for \$546 million, which is 20 times Copley's 1992 sales of \$51.9 million. Merck established a separate generic house, West Point Pharma, to compete with manufacturers of generic drugs.
7. *Developing an over-the-counter (OTC) market that was not under the price limit.* Warner-Lambert and Wellcome PLC formed a joint venture to comarket OTC drugs.
8. *Price-cut to expand market shares.* Merck's Mevacor is an example of the vulnerability of products that enter crowded therapeutic classes. A 1 month's supply of Merck's Mevacor 20 mg was \$60 in April 1994. Sandoz launched Lescol 20 mg with a sale price of \$30/mo, which was less expensive than both Merck's Mevacor and BMS Pravachol (\$54/mo).
9. *Giving up research.*
10. *Farming out R&D to contract research organizations (CROs).* This accelerates R&D without tying up the internal resources of the pharmaceutical companies. To bring a product to market 1 month earlier would generate \$10 million in sales if the product had annual sales of \$120 million, and \$50 million in sales if the product had annual sales of \$500 million. It also brings in know-how in a new therapeutic area quickly. The CRO unit can be used as an independent quality control arm.

EMERGENCE OF THE CONTRACT RESEARCH ORGANIZATION INDUSTRY

Contract Research Organizations in the Public Market

As R&D costs increased steadily in the late 1970s, a few service organizations emerged to handle specific tasks for the pharmaceutical companies such as identifying investigator sites, monitoring, data processing, statistical analysis, report writing, and regulatory supports. By the mid-1980s, G.H. Besselaar, for example, had grown to several hundred employees to provide an almost full range of services to the pharmaceutical industry. By the late 1980s and early 1990s, pressure from the health-care reform forced pharmaceutical companies to staff at trough and to use CROs to supplement their resource need. The outsourcing of clinical research needs to CROs was established as a branch service industry when G.H. Besselaar was acquired by Corning Glass works in 1989.

Research and Development Spending and Duration of Marketing Exclusivity

Despite cost containment, pharmaceutical companies have continued to increase their spending on R&D, which exceeds \$35 billion/yr.^[7] The reasons for the continued elevation of R&D costs are numerous. The number of the NDAs increased from 3233 patients in 1985–1988 to 3567 in 1989–1992, and to 3936 patients in 1993–1996.^[8] The number of procedures per patient also increased from 256 in 1992 to 319 in 1993, to 331 in 1994, and to 372 in 1995.^[9] One of the reasons for the increased number of procedures is the globalization of drug development, which in turn increases the complexity of the regulations on clinical trials. The improved efficiency created even more pressures on competitors to market the new drugs faster. Therefore, the number of years of exclusivity enjoyed by the pharmaceutical companies have been reduced drastically.^[10] Inderal was approved in 1968 and enjoyed marketing exclusively for 10 years. Tagamet was approved in 1977 and had 7 years of exclusivity. The other

major drugs' approval dates (duration of exclusivity) were as follows:

Capoten	1980	(5 yr)
Prozac	1988	(4 yr)
Diflucan	1990	(2 yr)
Recombinant	1992	(1 yr)
Invirase	1995	(0.25 yr)

The pressure from generic-drug producers has never reduced. Brand-name drugs typically lose 70% of their market share at the end of the first year following the introduction of the generic version. Therefore, utilization of CROs becomes a strategic decision.

GROWTH OF THE CONTRACT RESEARCH ORGANIZATION INDUSTRY

Increase of Outsourcing by Pharmaceutical and Biotechnology Companies

As of the end of 1997, the annual spending on R&D by pharmaceutical and biotechnology companies had exceeded \$35 billion. Two-thirds of this spending (approximately \$22 billion) was devoted to clinical research. The CRO industry was awarded approximately \$2.5 billion and is expected to grow 10% a year.

Full-Service Contract Research Organizations Grow Rapidly

Full-service CROs are growing rapidly due to the following reasons:

1. The incentives for drug firms to outsource clinical testing.
2. The continued expansion of biotechnology research.
3. The greater clinical testing requirements being imposed on medical device manufacturers.
4. Market share gains resulting from industry consolidation.
5. The rewards conferred on the first product to market.

Pharmaceutical companies are beginning to recognize the advantages of developing broad as well as deep relationships with a limited number of vendors. CROs able to service the full continuum of R&D needs have the following capabilities:

1. A broad range of therapeutic expertise in designing clinical trials.
2. The ability to efficiently collect, analyze, and manipulate data from many patients with various clinical conditions across many geographically dispersed sites.

3. The ability to provide a full range of trial management services to customers who wish to narrow their relationships to a limited number of CROs for drug development activities.
4. The ability to conduct trials at diverse geographical sites, allowing simultaneous filings of registration packages in several major jurisdictions.^[11]

Upstream and Downstream Growth of Contract Research Organizations

By the end of 1994, there were more than 200 Drug Information Association (DIA)-registered CROs in the United States,^[12] and a similar number of CROs were established in Western Europe. To stay competitive, the wider range of services is an important incentive for the CROs to attract pharmaceutical clients. Starting in 1996, the CRO industry realized that it needs more diversification from the traditional clinical monitoring, data processing, statistical analysis, and regulatory support. The leading CROs ventured upstream into central laboratory, drug-packaging, and health-economic services. Starting in 1996, large CROs brought contract sales service organizations. "Although there have been moves to harmonize product registration and requirements, nevertheless, it is still a multidomestic market with different medical rules, regulations, laws, and medical practices in every country."

Some other CROs acquired organizations that provide prelaunch product strategy services, tactical sales, and reimbursement support services. Some CROs invested in companies that specialized in combinatorial chemistry and functional genomics. To accelerate the recruitment of patients for clinical trials, companies with calling centers became the target of acquisition in 1997 and 1998.

FUTURE OF THE CONTRACT RESEARCH ORGANIZATION INDUSTRY IN THE PACIFIC RIM

Market Size for Pharmaceuticals

The Japanese pharmaceutical market is about \$55 billion/yr, or 20% of the world market, with an annual growth rate of 10%. "Until recently, there was little recognition of the need for CROs, in Japan."^[14] The Ministry of Health, Labor and Welfare (MHW) has established a study group (Honma Group) to review the need for CROs in Japan. The Honma Group supports the use of CROs in Japan if they are properly controlled. CRO groups are being established by various multinational CROs.

The need for CRO support in Taiwan to conduct clinical trials is evidenced by the presence of multinational CROs in recent years. A similar need for CROs is observed in Australia, Southeast Asia, India/Pakistan, Eastern Europe, and Latin America.

Critical Issues to Ensure the Quality of Clinical Trials

Critical issues to ensure the quality of clinical trials include the following:

1. *The attitude of medical professionals.* In most Pacific Rim countries (except China), medical professionals enjoy a special social status. The CRO monitors have to deal with the attitude of "I am a doctor and I know what I am doing. Don't come here to tell me what I have to do." Investigators with this type of attitude tend to generate case reports with a higher percentage of errors.
2. *Informed consent.* It will require great effort to educate both investigators and patients that clinical research cannot be done without the proper protection of the rights of the patients. The language of informed consent should be balanced and should be understandable by the local participants.
3. *Transparency of budgeting and payment.* Each budget item should be clearly identified and accountable and should reflect the needs of local trials.
4. *Respect for intellectual property.* Respect for intellectual property will encourage multinational companies to conduct more clinical trials in this region.
5. *Time availability of the investigators and coordinators.* Investigators in some research institutes have very limited time to do research. Their understanding of the efforts required to conduct well-controlled clinical trials is not compatible with the actual time required.
6. *Simplifications of the governmental approval process.* The approval process should be independent of the relationship between the sponsor and governmental officials. The approval process should also be transparent to the public.
7. *Affordable central laboratories.* In the near future, the clinical data from Asia will routinely be part of the multinational submission. The cost of data generated from the Asia-Pacific Rim should be proportional to the local cost of living.

Extension of Traditional Contract Research Organization Services

Traditionally, pharmaceutical and biotechnology companies treat R&D and sales and marketing as their core value and as proprietary. They guard and manage them with in-house employees. In 1997, a major shift in the FDA policy toward direct marketing by pharmaceutical companies to the potential consumer was announced. This policy shift opens a new avenue for CROs to provide contract sales services (contract sales organizations (CSOs)) to pharmaceutical and biotechnology companies. The growth of contract

sales is expected to accelerate, and the mergers of CSOs will probably intensify in the coming years.

Although CROs have ventured upstream to acquire discovery and research arms, such as combinatorial chemistry, functional genomics, and special assays, the future of such services is not clear. As site management organizations (SMOs) have continued to grow in number, to consolidate, the experience of mixing CROs and SMOs has not been very impressive.

To accelerate pharmaceutical development, information technology (IT) is critical to the efficient processing of data. More CROs are acquiring firms that provide the technology to enhance integrated information processing, real-time status, information sharing, and global data harmonization. The IT services in this field will grow; however, the development and maintenance cost are high and the success rate is usually low.

CONCLUSION

The CRO has become an important partner to pharmaceutical and biotechnology companies in the development of new compounds. The CRO has also provided services in drug registration. The total spending on R&D by pharmaceutical and biotechnology companies is estimated to exceed \$35 billion. Two-thirds of this is spent on clinical trials, and the CROs have captured only 10% of this. Of the top 20 pharmaceutical companies, 19 have invested in emerging markets such as China, India, Russia, and Latin America. These markets are accessible to clinical research for fast patient enrolments because of the improvement in adopting the practice of the International Conference on Harmonization (ICH) and Good Clinical Practices (GCPs). The large CROs have been following the path of pharmaceutical companies into the Asia-Pacific market since 1996. The Asia market for CROs will expand rapidly in the next 10 years. Therefore, the potential for further growth for traditional CRO is enormous. Originating in 1997, the CRO has acquired marketing and sales capabilities to extend services downstream. The estimate for worldwide marketing and sales expenses (\$36 billion) is very close to worldwide R&D spending. If growth of the traditional CROs is any indication of the potential growth of contract sales and marketing (CSOs), then the growth and consolidation of CSOs will be accelerated. Larger players from other industries may join in to capture the growth. The traditional nature of CROs has changed, and they will evolve into a formidable industry.

REFERENCES

1. Thomas, S. *Insight* **1996**, 4, 11–14.
2. DiMasi, J.A.; Hansen, R.W.; Grabowski, H.G.; Lasagna, L. *J. Health Econ.* **1991**, 10, 107–142.
3. Sperry, P. *Investor's Business Daily*; August 24, 1993.

4. Lee, T.Y. *Appl. Clin. Trials* **1994**, 3, 36–38.
5. *The Star Ledger*; February, 3, 1994.
6. IMS, *Strategic Management Review*; 1995.
7. Kendle Internationals Inc., *Prospectus*; 1997.
8. Boston Consulting Group, 1994.
9. *DataEdge*; 1997.
10. PhRMA. *Facts and Figures*; 1997.
11. J.C. Bradford and Co., *Healthcare Service Basic Report*; 1998.
12. Drug Information Association. *Pharmaceutical Contract Support Organizations*, 1994.
13. *Center Watch Weekly*. March, 9, **1998**, 1 (30).
14. Bentley, S. Clinical research and CROs in Japan. *Appl. Clin. Trials* April **1997**, 1, 30–34.

Contract Research Organization (CRO)

Tamara Pinkett

Biostatistics, IQVIA, Durham, North Carolina, U.S.A.

Jill Dawson

Corporate Communications Consultant, Danville, California, U.S.A.

Abstract

A contract research organization (CRO) is a company or organization that contracts with pharmaceutical, biotechnology, and medical device companies to assist in their research and development processes. This article highlights the purpose, history, and growth of CROs. This article is an overview of the services provided by the CRO with a focus on the role of the biostatistician within a CRO.

INTRODUCTION

Pharmaceutical companies continue to face pressure due to factors such as health-care cost containment efforts, patent expiries, optimizing patient access to acutely needed drugs, and rising research and development (R&D) costs. Contract research organizations (CROs) have evolved to help address these pressures. A CRO is a company or organization that contracts with pharmaceutical, biotechnology, and medical device companies to provide specialized services to assist in their R&D processes.

The CRO marketplace originated in the 1940s and 1950s and started to emerge in its present form in the late 1970s and the early 1980s.^[1] Today's CROs range from small, niche specialty groups to large, international full-service organizations. In this diverse industry, the top 10 CROs represent more than half of the world market, with the rest divided between 700–1,000 small and mid-sized companies.^[2]

A decade ago, the primary role of CROs was to run clinical trials. On average, it takes at least 10 years to bring a drug to market and costs about USD 2.6 billion,^[3] with several thousand compounds being explored in drug discovery in order to develop one approved drug. Drugs being developed are moved through a series of phases. Typically, this is divided into three phases, with a fourth phase being postapproval investigations. In the United States, before any clinical trials involving human subjects can be initiated, an Investigational New Drug (IND) application must be submitted to the U.S. Food and Drug Administration (FDA). After the IND application is submitted, the FDA typically has 30 days to reject the application; if the FDA does not reject it, approval is automatically granted.

Phase 1 trials involve 20–100 healthy volunteers, last for several months, and are conducted to evaluate the candidate drug's safety and appropriate dosing. Approximately

70% of drugs move to the next phase. Phase 2 trials involve up to several hundred people with the disease or condition, last usually for several months to 2 years, and are executed to investigate the efficacy and side effects of the drug. About 33% of drugs move to Phase 3 trials, which involve 300–3,000 volunteers who have the disease or condition, generally last for 1–4 years, and are performed to evaluate efficacy and monitor adverse reactions.

Around 25%–30% of drugs move to the next phase. At this point, a New Drug Application (NDA) can be submitted, triggering the regulatory review of the nonclinical and clinical data. If the data indicate that the drug has sufficient benefit to outweigh its risks, the FDA will approve the product for commercial sales, but may require additional clinical trials to evaluate populations not investigated before approval or to monitor the safety of the compound in the wider population in which the drug will be prescribed. These studies would be executed in what is known as Phase 4 of clinical development. Phase 4 trials involve several thousand volunteers who have the disease or condition and are conducted to further evaluate the safety and efficacy of the drug. The types of services a CRO may provide through the product life cycle are shown in Table 1.

Today, as sponsor organizations become ever leaner, they are partnering with CROs to gain access to skill sets, including broad therapeutic expertise, that are no longer available in-house. Additionally, in the last decade, as biopharmaceutical drug developers has shifted the activities from North America and the European Union to regions in Asia such as India, China, and Korea, they have used CROs to expand their global reach. Full-service partnering and portfolio management, including joint strategic decision-making, and expanding from the clinical realm into the commercial one, is becoming more widespread, driving growth in the CRO market.

Table 1 Examples of services a CRO may offer

Clinical Solutions and Services	
Project management and clinical monitoring:	Late phase and advisory services:
• Clinical data management • Study design and operational planning • Investigator/site recruitment • Site and regulatory start-up • Clinical monitoring • Project management • Digital patient services	• Product development strategy consulting • Regulatory and compliance consulting • Process and IT implementation consulting
Clinical trial support services:	Commercial services:
• Clinical data management • Biostatistical services • Central laboratories • Bioanalytical laboratories • Cardiac safety and ECG laboratory services • Safety and pharmacovigilance operations • Phase I clinical pharmacology units	• Contract sales • Market entry/market exit • Integrated channel management • Patient engagement services • Market access and commercialization consulting • Brand and scientific communications • Medical education
Strategic planning and design:	Outcome/observational:
• Biomarkers, genomics, and personalized medicine • Model based drug development • Planning and design • Regulatory affairs services	• Observational studies • Product and disease registries • Comparative effectiveness studies • Payer/provider solutions

Zion Research predicted in June 2017 that the world-wide CRO market would reach USD 59.42 billion in 2020, increasing at a compound annual growth rate of 9.8% between 2015 and 2020.^[4] Industry Standard Research projected in March 2017 that the market would grow at an average annual rate of 6.8% from 2016 to 2021^[5]; and Grand View Research Inc. forecast in January 2016 that the CRO industry would reach USD 45.2 billion by 2022.^[6]

The regulatory space continues to evolve at a rapid pace, with CROs offering particular expertise and access to regulatory opinion due to their broad range of activities for many clients in many therapeutic areas. Regulators are increasingly accepting novel study design alternatives, such as adaptive trial designs, which use accumulating data to evaluate efficacy more efficiently and precisely in particular populations, drawing on cumulative experience, and master protocols, which allow multiple drugs to be evaluated in the same trial. Adaptive designs have forced CROs to be more flexible in their trial designs, since the demand for adaptive designs has been growing at approximately 35% per year over the last decade.^[7] These designs have required more continuous data monitoring, use of Data Monitoring Committees on a more regular basis, and greater statistical expertise. Another area of growth has been in real-world evidence, where data are analyzed from electronic health records, surveys, and other sources.

Among the opportunities for collaboration between biopharma companies and CROs are the potential to

- **Increase predictability** by creating study plans based on evidence, in order to mitigate operational risks, reducing the number of protocol amendments and increasing the likelihood of regulatory approval.
- **Shorten timelines** by using historical data to select appropriate sites, reducing the number of low- and non-enrolling sites, improving the predictability of enrollment rates, and reducing the time to last patient in.
- **Maximize asset value** by basing development decisions on real-world insights and analytics to generate relevant evidence earlier in the process. This can help optimize coverage and reimbursement, and reduce the time to peak sales.

THE ROLE OF THE BIOSTATISTICIAN WITHIN A CRO

Biostatisticians apply statistical methods to drug development in many areas including methodology, modeling pharmacokinetics, bioequivalence and bioavailability testing, monitoring product safety, and product quality assessment.^[8] Within CROs, biostatisticians are involved in every step of clinical research including trial design,

protocol development, data management and monitoring, data analysis, and clinical trial reporting.^[9] The importance of this role is demonstrated by the fact that the largest CROs may employ as many as 1,000 biostatisticians across 10 or more countries.

There are multiple differences between the business models of CROs and pharmaceutical companies (see Table 2). This is reflected in differing roles for statisticians. In CROs, statisticians typically work in a variety of therapeutic areas across all phases of trials, whereas in pharma companies, they tend to work in one therapeutic area, usually concentrating on one particular phase of development.

Biostatistical services span all study phases and therapeutic areas, including work on clinical development plans, protocols, case report forms, statistical analysis plans, and data analysis and interpretation. Biostatisticians also provide support for clinical study reports and for regulatory submissions; they may also attend meetings with regulators and offer expert testimony. Additional examples of roles are shown in Table 3.

At CROs, expert-level biostatisticians are used for such tasks as the following:

- Clinical development plans and other strategic program planning, such as NDA planning
- Protocol development

- Consulting on complex design or analysis issues
- Methodology development
- Internal review of all tables, study reports, and text
- Expert testimony (i.e., statistical theory, regulatory requirements/expectations)

Another interesting area of expert-level biostatistical input may include personalized medicine. For example, recent developments in mobile health include health-related mobile web applications for providers and potentially for patients. Applications include personalized dosing to target drug exposure via population pharmacokinetic models and randomized concentration-controlled trials. On June 16, 2017, FDA commissioner Dr. Scott Gottlieb announced an FDA initiative to accelerate the approval of mobile web applications.^[10]

At CROs, there are typically two key roles within biostatistics. First, are the statisticians, who typically have a master's degree or doctorate in statistics, biostatistics, or a related discipline. The biostatistician acts as the overall biostatistical team lead and project contact, and the role involves planning the study and data analyses, coordinating the production and review of data analyses and outputs, drafting or reviewing the clinical report, and interacting with the customer. Second, are the statistical programmers, who have extensive SAS programming skills. The primary

Table 2 A comparison between pharmaceutical companies and CROs

Pharmaceutical companies vs CROs	
Pharmaceutical companies	CROs
• Discover and own drugs • Responsible for all aspects of drug development from early phase through to marketing and packaging • Statisticians tend to work in one therapeutic area, usually concentrating on one particular phase of development	• Do not “own” drugs • Large CROs offer full service, getting involved in all aspects of drug development from early phase through to marketing and packaging • Statisticians work in a variety of therapeutic areas across all phases of trials • Offer cost containment (supplement existing work force), speed, international drug development, experience, specialized services

Table 3 Biostatistical services in CROs

Services		
• Clinical development plans • Protocol and data collection tool development • Statistical analysis plans including table shells • Analysis interpretation and reporting (analysis datasets, tables, listings, graphics) • Data Monitoring Committee support • Patient profiles	• Standardized dataset (CDISC) authorized supplier • Clinical study report support • Regulatory submission (e.g., IND, NDA, sNDA, MAA, BLA, and CTD) • Feasibility analysis • Nonclinical analysis • Regulatory representation • Expert testimony • Modeling and simulation	• Monitoring optimization • Prescription to over-the-counter switch trials • Postmarketing manuscript support • Fraud detection • Pharmacogenomics • Consultancy • Adaptive clinical trials • Biomarkers • Flexible resourcing models

CDISC = Clinical Data Interchange Standards Consortium; sNDA = supplemental New Drug Application; MAA = Marketing Authorization Application; BLA = Biologics License Application; CTD = Common Technical Document.

responsibilities of this role include developing programs to produce statistical analyses and outputs, verifying programs, and carrying out independent verification for all analytical outputs. Either role requires that person to be very detail-oriented. Depending on the CRO, the disciplines may be performed by different staff or be combined into one role.

ACKNOWLEDGMENTS

The authors thank Dr. Russell Reeve for his helpful discussion about the advances in pharmaceutical research and development, the demand for CROs, and the role of the (bio)statistician.

REFERENCES

1. Walsh, R. A History of: Contract Research Organisations (CROs). Pharmaphorum, posted 2/10/2012. https://pharmaphorum.com/views-and-analysis/a_history_of_contract_research_organisations_cros/ (accessed February 12, 2018).
2. Saias, P.; Kapadia, A. CROs, Convergence, and Commercial Opportunities. <http://www.kpmg-institutes.com/content/dam/kpmg/healthcarelifesciencesinstitute/pdf/2016/cros-convergence-and-commercial-opportunities.pdf> (accessed February 12, 2018).
3. Tufts Center for the Study of Drug Development Press Release Online. Cost to Develop and Win Marketing Approval for a New Drug Is \$2.6 Billion, posted Nov 18, 2014. http://csdd.tufts.edu/news/complete_story/pr_tufts_csdd_2014_cost_study (accessed February 12, 2018).
4. Zion Market Research. Global Contract Research Organization (CRO) Market Will Reach \$59.42 Billion by 2020. GlobeNewswire, posted June 05, 2017. <https://globenewswire.com/news-release/2017/06/05/1008190/0/en/Global-Contract-Research-Organization-CRO-Market-Will-Reach-59-42-Billion-by-2020.html> (accessed February 12, 2018).
5. Industry Standard Research. ISR Projects Strong Growth in the CRO Market. Industry Standard Research Report Abstract, posted Mar 10, 2017. <https://www.isrreports.com/isr-projects-strong-growth-cro-market/> (accessed February 12, 2018).
6. Grand Review Research Press Room Page Online. Healthcare CRO Market to Reach \$45.2 Billion by 2022, posted Jan 2016. <http://www.grandviewresearch.com/press-release/global-healthcare-cro-market> (accessed February 12, 2018).
7. Reeve, R. Critique of Trends in Phase II Trials. PowerPoint presentation, International Conference on Advances in Interdisciplinary Statistics and Combinatorics, Greensboro, NC, October 10–12, 2014.
8. U.S. Food and Drug Administration. Biostatisticians, update posted Sept 22, 2017. <https://www.fda.gov/Drugs/ScienceResearch/ucm294601.htm> (accessed February 12, 2018).
9. CROS NT. How well do you understand the role of Biostatistics in Clinical Research? posted Sept 9, 2014. <http://crosnt.com/role-biostatistics-clinical-research/> (accessed February 12, 2018).
10. Gottlieb, S. Fostering Medical Innovation: A Plan for Digital Health Devices. FDA Voice, posted June 15, 2017. <https://blogs.fda.gov/fdavoices/index.php/2017/06/fostering-medical-innovation-a-plan-for-digital-health-devices/> (accessed February 12, 2018).

Controlled Clinical Trials: *P*-Values, Evidence, and Multiplicity Considerations

Mohammad F. Huque, Mohamed Alosch, and Satya D. Dubey

Office of Biostatistics, CDER, FDA, Silver Spring, Maryland, U.S.A.

Rafia Bhore

Department of Clinical Sciences, University of Texas, Southwestern Medical Center, Dallas, Texas, U.S.A.

Abstract

This entry aims to clarify the role of *p*-values in hypothesis testing by examining methods for dealing with multiplicity, discussing various testing principles, and providing methods of analysis.

Keywords: Clinical trials; Evidence; Multiplicity; Null hypothesis; *P*-value.

INTRODUCTION

Multiplicity is widespread in clinical trials and is of increasing concern because many trials now aim at multiple objectives of different regulatory and scientific importance. Potential sources of multiplicity in trials include comparison of several dose groups, investigation of treatment effects for multiple endpoints observed at multiple time points, superiority and non-inferiority tests, subgroup analysis, variable and model selections, and analysis for demographic factors and for targeted regions in multi-regional trials. These and other sources of multiplicity in a trial can make the traditional α -level (e.g., at $\alpha = 0.05$ or 0.01) statistical tests and evaluations for various objectives of the trial not necessarily valid. That is, multiplicity in a trial can increase the likelihood of finding some treatment effects by chance alone, in the absence of no such treatment effects. If the probability of such chance findings, traditionally known as false-positive or type I error rate, is inflated in a trial due to multiple testing then such a trial is said to have a multiplicity problem. Solutions for this problem can generally be found when holding to some key statistical principles, methods, and clinical ideas, which are the focus of this article.

The remainder of this article is organized as follows: in the first few sections we clarify concepts of inferential *p*-values, “substantial evidence,” and “replication,” as these concepts relate to proper interpretation of results of a trial in the context of robustness of evidence and multiplicity of findings. Subsequent sections cover multiplicity concept, familywise error rate and its control, commonly used methods for addressing multiplicity, multiple hypotheses testing principles, and multiplicity considerations for some clinical trials. This article is intended for those interested

in the design, conduct, analysis, and interpretation of results of controlled clinical trials.

P-VALUE AND ITS MEANING

Clinical trial publications extensively report *p*-values as evidence of positive findings. If a *p*-value is small, it is taken as evidence against the so-called “null hypothesis” of no treatment effect, and subsequently for claiming a treatment effect. Such statistical evidence is often confused with the regulatory term “substantial evidence,” which is a much broader concept used in the regulatory decision-making. We will clarify these concepts, as understanding the concept of substantial evidence in relation to the statistical evidence may help in better planning of a single or multiple trials for a given indication.

Within the hypothesis-testing framework, the definition of *p*-value is straightforward but its interpretation is not. *p*-value is different from the type I error rate that is pre-specified, even though both use tail areas under the null hypothesis distribution of the test statistic. Its value is not known until the experiment is over. Suppose that a vector $\mathbf{y}$ represents observed measurements on an endpoint in an adequate well-controlled trial, and $T(\mathbf{y})$ is a value of the test statistic $T(\mathbf{Y})$, which is a statistical measure of the treatment effect, such as *t*-test or a log-rank statistic. Let δ be an unknown parameter of the treatment effect, such that values of $\delta \in \Omega_0$ are not desirable to the experimenter and he/she wants to rule it out. For example, if δ is the true treatment difference defined as the true cure rate by the new treatment minus the true cure rate by the control treatment, then $\delta \leq -\Delta$ means that the new treatment is inferior to the control by a tolerable margin of Δ . Similarly, if this

$\delta \leq 0$, this means that the new treatment is not more efficacious than the control. These δ values are then in Ω_0 , which are counter to what the experimenter desires to conclude from the experiment. A p -value associated with the null hypothesis H_0 : $\delta \in \Omega_0$ is then defined as the probability that $T(\mathbf{Y})$ is “more extreme” than the observed $T(\mathbf{y})$ given that H_0 is true. Here “more extreme” is in the direction against the null hypothesis.

This p -value is a routine statistical measure of uncertainty used in all kinds of scientific and medical experiments to date, and is taken as a measure of evidence against the null hypothesis if it is small, for example, less than 0.05. The approach taken in this regard is the method of contradiction.^[1–3] That is, assume that the null is true and show that the observed data conflict with this assumption. The smaller the p -value, the greater is the degree of this conflict, thereby giving greater support to the idea that the null hypothesis is not true.

The p -value as defined above is a random variable, uniformly distributed over the $[0, 1]$ interval under the null hypothesis. This distribution property under the null hypothesis is true regardless of whether a given study is small, large, highly powered, or underpowered. This property has led to the use of a cutoff point of p -value, such as 0.05, for claiming evidence against the null, meaning that the chance is 1 in 20 that an experiment with no treatment effect would result in a p -value of 0.05 or less. This error rate is known as the type I error rate in the context of the Neyman-Pearson theory of hypothesis testing.^[4] It is easy to show that this type I error rate is much higher than 0.05 if multiple null hypotheses were tested simultaneously without any statistical adjustments for multiplicity.^[5,6]

However, it is important to recognize that a small p -value could also arise in an experiment which has nothing to do with the evidence against the null hypothesis. An unusual problem or property of data at hand can give rise to a small p -value. Randomization and blinding techniques in clinical trials protect results from bias, but despite these measures, there are often problems. For example, in chronic disease areas and prevention trials, patients are usually treated and observed for longer periods, resulting in dropouts. A p -value claiming a treatment effect in this situation may be confounded by informative dropout effects depending on the extent of dropouts. Also, clinical trials that do not analyze all randomized patients may lose validity of inference that randomization provides, and even if all randomized patients were included, an imputation technique used for missing data may have failed to preserve the natural course of the disease and variability in the data. In addition, the sampling distribution of the statistic $T(\mathbf{Y})$, because of unusual behavior of the data, may be flat-tailed, having a larger p -value than that calculated by a generic standard formula. Furthermore, clinical trials often exhibit multiplicity issues missed at the design and analysis stages, and hence leading to questionable p -values.

Despite the above concerns, the p -value, when appropriately calculated, validated, adjusted for multiplicity, and properly interpreted, does provide evidence against the null hypothesis if it is small. The smaller the p -value, the greater is this evidence. In using a p -value, however, an important point to know is that it is not a probability of the null hypothesis, as misunderstood by some. It is rather a measure of the degree of the conflict of the data with the null hypothesis, producing a result only when this measure of conflict for a given data exceeds a pre-specified level.

Evidence based on the p -value approach is often called a “frequentist” approach for making inference. There are other approaches. Bayesian and the “likelihood” approaches are the main ones. The Bayesian approach requires specifying a model for the sampling data as the likelihood function $L(\mathbf{y}|\theta)$ conditional on parameters θ , as well as a prior distribution on θ . This setup considers the parameters as random variables and all inferences use the posterior distribution, which is the distribution of the parameters θ , conditional on the observed data. In contrast, the frequentist approach considers sampling observations as a random sample and parameters as fixed unknown quantities. The likelihood approach utilizes the likelihood function for making inference, justifying that once the sample data is observed, the likelihood function contains all necessary information about the unknown parameter θ . There have been arguments about these inference procedures claiming advantage of one over the other by citing special examples. However, we feel that a p -value that is derived from a sufficient test statistics and for which the test is constructed by the Neyman-Pearson likelihood ratio principle is quite appropriate for making inference for many controlled clinical trials.

EVIDENCE AND REPLICATION

The concept of evidence for making regulatory decisions has been around for sometime.^[7–9] It includes, in addition to statistical evidence, evidence from other areas, including epidemiology, basic sciences, ethics, and good clinical practices. Another concept embedded in the regulatory standard is the concept of replication. This implies that the collective evidence derived from multiple studies or in special cases from a single study should be reliable and reproducible with high probability. This replication probability idea is a valuable concept for assuring that a claimed result of treatment effect when clinically interpretable is a persuasive result.

The Kefauver-Harris amendment of the Food, Drug, and Cosmetic (FDC) Act of 1962 defined the concept of “substantial evidence” for the effectiveness of the drug in Section 505(d) of the Act as “evidence consisting of adequate and well-controlled investigations.” This has been interpreted in general to imply at least two adequate well-controlled confirmatory studies (as defined in 21 CFR

Table 1 Conditional and Posterior Predictive Probabilities of Finding a Statistically Significant Result at the 0.05 Significance Level in a Second Study for Various Observed p -Values in the First Study

Observed p -value (1st study)	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
Conditional power method	0.50	0.61	0.73	0.80	0.86	0.91	0.94
Bayesian predictive probability method	0.50	0.58	0.67	0.73	0.77	0.83	0.86

314.126), each convincing on its own as evidence of efficacy of a new drug. The justification for such a requirement is the need for independent substitution of experiments often referred to as the need for replication of the findings.^[8,9]

Thus, the requirement for substantial evidence has been interpreted to go beyond observing evidence from a single trial for the purpose of providing further assurance by requiring replication of study findings. The concept of replication is intended to provide additional assurance of the efficacy of the drug to be used for the patient population after weighing against possible safety risks. A large, well-designed, and adequately powered multi-center study with small p -value is expected to provide stronger evidence than a small study with p -value close to 0.05. Thus, in special situations, one might consider findings of such a large study with a small p -value (e.g., $p < 0.001$) as having established the substantial evidence sought, as long as this result is internally consistent and clinically meaningful. To account for such situations, for example in difficult disease areas where it might be unethical to do a second study, the U.S. Congress amended Section 505(d) of the Food, Drug, and Cosmetic Act by issuing Section 115(a) of the Food and Drug Administration Modernization Act (FDAMA), which states that the Agency in special situations may consider “data from one adequate and well-controlled clinical investigation and confirmatory evidence” to constitute substantial evidence. Useful discussions about the two-study versus single-study evidence can be found in Koch^[10] and Huque^[11] published as commentaries to a paper by Shun et al.^[12]

Following Goodman,^[13] the probability of replicating a statistically significant result in a second study depends on the true treatment effect size. On assuming that the true treatment effect size is the same as the observed treatment effect size in the first study, the probability of replicating efficacy for a second (would-be-conducted) study is simply the conditional power for this effect size. Mathematically, suppose that d_1 is the observed treatment difference for the first study and z_1 its corresponding treatment difference value on the normal Z -scale, then the probability of replicating a significant level α result in a second future trial is given by $\Pr\{|Z_2| > z_{\alpha/2} | \theta = z_1\}$, where Z_2 is the test statistic for this second study, θ being the true treatment effect size on the Z -scale, and $z_{\alpha/2}$ is the cutoff point such that the tail area under the standard normal density curve from $z_{\alpha/2}$ to infinity is $\alpha/2$. This probability of replication is then given by $\Pr\{|Z_2| > z_{\alpha/2} | \theta = z_1\} = 1 - \Phi(z_{\alpha/2} - z_1) + \Phi(-z_{\alpha/2} - z_1)$, for $z_1 \geq z_{\alpha/2}$.

Such a probability of finding a statistically significant result in a replicated experiment can also be calculated by the predictive probability approach of Geisser.^[14] Suppose that the posterior probability density of θ , using a flat prior, is $p_1(\theta | z_1) = N(\theta; z_1, 1)$ and the probability density of the second to-be-replicated trial is $f_2(z_2 | \theta) = N(z_2; \theta, 1)$. Then the posterior predictive probability density of Z_2 given z_1 is

$$f(z_2 | z_1) = \int_{-\infty}^{\infty} f_2(z_2 | \theta) p_1(\theta | z_1) d\theta = \frac{1}{2\sqrt{\pi}} \exp\left\{-\frac{1}{4}(z_2 - z_1)^2\right\} \\ \sim N(z_1, \sqrt{2})$$

Table 1 displays probabilities of finding a statistically significant result at the 0.05 significance level in a second identical study for various p -values observed in the first study by the above two approaches. These results show that a p -value of 0.05 found in a single experiment is not persuasive evidence, as the probability of replicating this result in a second study is only 50%. It is only when the p -value in the first study is less than 0.001 that one would expect a probability of 85% or greater for finding a result statistically significant in a second study. Note that the posterior predictive probabilities in Table 1 are consistently smaller than conditional probabilities when observed first-study p -values range from 0.0005–0.025.

MULTIPLICITY CONCEPT

A clinical trial that may not have any multiplicity issue has two groups, one with the treatment and the other a control group, and has a single clinical endpoint of importance, such as mortality, that can characterize a meaningful benefit of the test treatment. This is the case of testing a specified single null hypothesis of a single clinical endpoint without any interim analysis. However, even for this simplest case, as the data used is the sample data of this endpoint, its outcomes have associated uncertainties due to sampling variations. As a result, a positive treatment effect comparison result can appear with certain probability due to chance alone when in reality there is no such treatment effect. This error probability is called the *false-positive error rate* or the *type I error rate*. By convention, the type I error rate, also called alpha (α) usually is set at 0.05 (i.e., 5% or 1 in 20). This means that there is a 5% chance that results as extreme as those observed in the trial could have occurred by chance alone. It also means that if the investigator performed 20 similar trials, one of them might yield

a false-positive result. The type II error results from concluding that there is no treatment effect when in fact there is a treatment effect; in other words, it is a false-negative result. In the language of hypothesis testing this is accepting the null hypothesis when in fact the null hypothesis is false. The *type II error* rate, also called beta (β), is usually set at 0.10 or 0.20 (i.e., 10% or 20%). Power is defined as $100 \times (1 - \beta)\%$ and is correspondingly set at 90% or 80%.

The above simple case is based on the premise that there is one treatment comparison in order to answer one research question on a single endpoint. However, trials in general are rarely that simple. Many trials now aim at multiple objectives of differing regulatory and scientific importance, and usually test more than one research question or perform treatment comparisons on more than one endpoint at more than one time point. Consequently, the potential to draw a false positive conclusion, that is, the type I error rate, may increase. The trial may have the opportunity to win for treatment effect in multiple ways. This is analogous to taking a black box filled with 19 white marbles and one black marble, and asking a volunteer to choose one marble from the box at random blindly. The probability of choosing the black marble by chance is one in twenty, or 5%. On the other hand, if one increases the number of black marbles in the box, say, on keeping the same total 20 marbles in the box, the likelihood of choosing a black marble by chance increases. As one increases the number of black marbles in the box, the likelihood of pulling a black marble out by chance continues to increase. In other words, there are multiple ways to achieve a successful outcome in this situation. This example illustrates the issue of multiple testing or *multiplicity* that is, testing more than one treatment comparison or hypothesis or endpoint in a single trial.

Potential sources of multiplicity in trials include comparison of several treatments or dose groups, investigation of treatment effects for multiple endpoints measured at multiple time points, superiority and non-inferiority tests with interim analyses, subgroup analysis, variable and model selections, and analysis for demographic factors and for targeted regions in multi-regional trials. These and other sources of multiplicity in a trial can make the traditional α -level (e.g., at $\alpha = 0.05$ or 0.01) statistical tests and evaluations for various objectives of the trial no longer valid. That is, multiplicity in a trial can substantially increase the likelihood of finding some treatment effects by chance alone, in the absence of no such treatment effects. This phenomenon, which increases the frequency of chance findings, is traditionally known as the *inflation of the type I error rate*. A trial design is said to have multiplicity issues when it faces such an inflation of the type I error rate in conducting multiple statistical tests.

For example, if the probability of finding a significant result by chance is 0.05 in making a single comparison, then the unadjusted probability of finding a significant result by chance can be as high as 0.23, 0.40, 0.64, and 0.92 in

making 5, 10, 20, and 50 comparisons, respectively. These calculations follow from $1 - (1 - 0.05)^m$ with m being the number of independent comparisons made each at level $\alpha = 0.05$. These are substantial increases in the type I error rate compared to 0.05, i.e., 1 in 20 times. In this case, application of statistical approaches to adjust the results for multiplicity generally becomes necessary in order to control for the chance of erroneous conclusions. However, it is worth noting here that there are clinical trial situations that do not raise multiplicity issues even though the trial may perform multiple statistical tests. A few of these cases are covered in the section “Clinical Decision Rules.” An example is the testing for co-primary endpoints in a trial, which requires that each specified endpoint demonstrate treatment efficacy for characterizing a clinically meaningful benefit of the treatment.

FAMILYWISE ERROR RATE AND ITS CONTROL

In statistical testing of a family of multiple null hypotheses when the interest is to reject at least one of them for a treatment effect, a commonly used error rate is the familywise error rate usually denoted by the acronym FWER or FWR. This is a generalization of the usual type I error rate for a single hypothesis testing problem to multiple hypotheses testing problems. Consider testing m multiple (null) hypotheses $H_1, \dots, H_m$, where some hypotheses may be true and some may be false but their true status are unknown. This uncertainty gives rise to a total of $2^m - 1$ combination of true and false hypotheses as *null hypothesis configurations*. Suppose that for an arbitrary null hypothesis configuration, one calculates the probability of committing at least one type I error. Then this probability is called the familywise error rate (FWER) for that configuration. A test procedure is said to “control” FWER at level α for a specific null hypothesis configuration if FWER for that configuration is less than α .

A null hypothesis configuration of special interest is the one in which all m null hypotheses are assumed to be true null hypotheses. Such a null hypothesis configuration is often called a *global* or a *complete null hypothesis*. This is also referred to as an *intersection null hypothesis* of dimension m . If a test procedure controls FWER at a specified level α (e.g., $\alpha = 0.05$) only for the global null hypothesis then it is said to control FWER *weakly* at that level. However, if a multiple hypothesis testing procedure controls FWER at level α for all possible null hypotheses configurations, that is, for all possible combinations of true and false null hypotheses, then it is said to control FWER *strongly* at that level.^[5] Often FWER control is meant FWER control in the strong sense if there is no ambiguity.

In testing multiple (null) hypotheses $H_1, \dots, H_m$, often the interest is in making inference about the individual hypotheses H_j for each $j = 1, \dots, m$. This is achieved by controlling FWER at level α for all possible null hypotheses

configurations, including the global null hypothesis. As an illustration, consider a two-arm cardiovascular trial for assessing the effect of a new therapy against a control for three primary endpoints: all-cause mortality, stroke, and myocardial infarction (MI). The clinical interest for this trial is in making endpoint specific claims of efficacy of this therapy. The investigator then seeks to find the effect of the new therapy individually for each of these three endpoints. In this case, there are a total of seven null hypothesis configurations for making endpoint specific inferences, and one needs to control FWER for all these seven null hypotheses configurations. Therefore, this case requires strong control of FWER.

COMMONLY USED METHODS FOR ADDRESSING MULTIPLICITY

There are many statistical methods tailored for different situations that control FWER strongly. They can be broadly classified into two types: single-step and multi-step procedures. Single-step procedures that control FWER strongly and provide adjusted p -values, also easily offer adjusted confidence intervals for the magnitude of the treatment effects. On the other hand, multi-step procedures generally preserve the power of the tests more efficiently, but their adjusted p -values do not easily convert to adjusted confidence intervals. In the following, Bonferroni and Šidák procedures are single-step, whereas Holm's and Hochberg procedures are multi-step. Multi-step procedures have been of two types, *step-down* and *step-up* procedures. In step-down procedures, one starts with the smallest p -value (i.e., the most significance test) and then steps-down to the next smallest p -value (i.e., the next most significance test) and so on. In contrast, step-up procedures proceed in the reverse direction. They start with the largest p -value (i.e., the least significant test) and step-up to the second largest p -value and finally to the smallest p -value (i.e., the most significant test). In the following section, we briefly discuss the Holm's procedure as an example of a step-down procedure and the Hochberg procedure as an example of a step-up procedure. The subsequent section reviews some commonly used single and multi-step procedures for addressing multiplicity in clinical trials, starting with simple ones.

Bonferroni Test

The Bonferroni test is a single-step procedure and is a commonly used test for the primary objectives of clinical trials, as it works under all correlations and joint distributions of the test statistics. The procedure applies directly to p -value of the individual hypothesis tests when a given family of m hypotheses to test, and reliably and conservatively controls FWER in the strong sense. The Bonferroni test is simple to apply. One chooses positive weights $w_1,$

$w_2, \dots, w_m$ for the m individual hypotheses H_j (for $j = 1, \dots, m$) according to some clinically meaningful criteria (e.g., by the clinical importance of the m hypotheses) with the sum $(w_1 + w_2 + \dots + w_m) \leq 1$, and calculates weight adjusted values $q_j = p_j/w_j$, where p_j is the observed p -value on testing H_j in the trial. One then rejects H_j for those $j = 1, \dots, m$ for which $q_j < \alpha$. For simplicity of application it is customary to take equal weights $w_j = 1/m$. With these equal weights one rejects H_j for any j when $p_j < \alpha/m$.

Holm's Test

The Holm's^[15] test is a step-down procedure, and like the Bonferroni test, it is a commonly used method for the primary objectives of a trial. Similar to the Bonferroni test, it provides FWER control in the strong sense for all types of correlations among test statistics and their joint distributions. It can be implemented with weights as well as without weights. In the unweighted Holm's test, one orders the p -values of the individual m hypothesis tests in increasing order as $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$. The test begins by comparing the smallest p -value $p_{(1)}$ to α/m . If this $p_{(1)}$ is less than α/m , the hypothesis $H_{(1)}$ associated with this p -value is rejected. The testing then continues to the next step comparing the next smallest p -value $p_{(2)}$ to $\alpha/(m-1)$. The hypothesis $H_{(2)}$ associated with $p_{(2)}$ is then rejected if $p_{(2)} < \alpha/(m-1)$. This testing continues by comparing the subsequent ordered p -values to the corresponding alpha critical values of $\alpha/(m-2), \dots, \alpha$, as long as the previous hypotheses are rejected. Once a hypothesis is not rejected, the remaining hypotheses with larger ordered p -values are not rejected. The Holm's test is in general more powerful than the Bonferroni test. As an example, consider a clinical trial that produces two-sided p -values of 0.015, 0.024 and 0.045 for the three primary endpoints of the trial. The unweighted Bonferroni test will conclude statistical significance only for the endpoint with p -value of 0.015 because 0.015 is less than $0.05/3$. On the other hand, the Holm's test will conclude statistical significance for all three endpoints.

The weighted version of the Holm's test can be carried out as follows.^[16] Let $w_1, w_2, \dots, w_m$ be the positive weights as defined above and $q_j = p_j/w_j$ for $j = 1, \dots, m$. Then order q_j values in increasing order as $q_{(1)} \leq q_{(2)} \leq \dots \leq q_{(m)}$, and let $w_{(1)}, w_{(2)}, \dots, w_{(m)}$ be the corresponding weights. Further, let $W_{(j)} = \sum_{i=j}^m w_{(i)}$ for $j = 1, \dots, m$, be the total (for $j = 1$) and partial sums (for $j > 1$) of these weights. The test starts by comparing the smallest $q_{(1)}$ to $\alpha/W_{(1)}$. If $q_{(1)}$ is less than $\alpha/W_{(1)}$, the hypothesis associated with this $q_{(1)}$ is then rejected. The testing then continues to the next step comparing the next smallest $q_{(2)}$ to $\alpha/W_{(2)}$ to check if $q_{(2)} < \alpha/W_{(2)}$ for the rejection of the hypothesis associated with $q_{(2)}$. If this hypothesis is rejected, then testing continues to comparing the subsequent ordered q_j values to corresponding alpha critical values of $\alpha/W_{(j)}$ as long as all previous hypotheses were rejected. The weighting scheme increases the power for

those hypotheses with larger weights at the cost of reducing the power for those hypotheses with smaller weights. Note that if the weights are all equal, that is, $w_j = 1/m$ for $j = 1, \dots, m$, then the weighted Holm's test reduces to the unweighted Holm's test.

Hommel and Hochberg Tests

Hommel and Hochberg^[17,18] have proposed step-up tests that can be more powerful than the Holm's test. The Hochberg test has become popular because of its simplicity, and its use is well accepted for independent and for certain types of positively correlated test statistics. However, in general, it does not control FWER in the strong sense^[19] for all dependent hypotheses. For this reason, it is seldom used for the primary objectives of a trial. The Hochberg test with equal weights uses the same p -values and alpha critical values as those used in the Holm's test. However, it is a step-up test, and as such, it starts with the largest p -value, and rejects all hypotheses with smaller p -values when a p -value is less than its alpha critical value. The Hochberg test can also be used with weights.^[20] Consider a trial that produces two-sided p -values of 0.035, 0.040, and 0.045 for the three primary endpoint tests. The unweighted Bonferroni and Holm's tests will not conclude any of these three p -values as statistically significant at the 0.05 level, but the unweighted Hochberg test will conclude all three p -values as statistically significant at this level.

PAAS and Šidák's Test

Moyé^[21] has proposed the prospective alpha allocation scheme (PAAS) for testing multiple hypotheses of a trial. This has slight advantage in power over the Bonferroni test. Its use is well recognized for independent and positively correlated test statistics. However, for other cases it does not guarantee FWER control in the strong sense. Adjusted p -values $q_j = p_j/w_j$ and alpha critical values $\alpha_j = w_j\alpha$ are calculated on assuming independence between test statistics to satisfy the equation

$$(1 - w_1\alpha)(1 - w_2\alpha) \cdots (1 - w_m\alpha) = (1 - \alpha).$$

In the Šidák's procedure,^[22] each individual hypothesis test in a family of m hypotheses is carried out at the alpha critical value of $\alpha^* = 1 - (1 - \alpha)^{1/m}$. This method similar to PAAS is recommended when the comparison tests are either independent or positively correlated. It is conservative in declaring significance similar to Bonferroni and PAAS.

Extensions of Šidák's Test

Šidák's independence adjustment formula $\alpha^* = 1 - (1 - \alpha)^{1/m}$, although conservative for positively correlated test statistics, is very attractive and easy to use. It readily converts

to simultaneous confidence intervals. Therefore, there has been interest in ad hoc modifications of this independence formula that incorporate correlation information for making it less conservative when testing for treatment effects of moderate to highly correlated endpoints of early phase trials or for secondary and safety endpoints of a trial, and for other situations where strict FWER control in the strong sense is not needed. In this regard, one suggestion^[23] was to replace m by $m' = \sqrt{m}$. Surprisingly, this simple algorithm works well for highly correlated tests. Subsequently, Dubey,^[24] based on the ideas of Tukey et al.^[23] and Armitage,^[25] proposed a "multi-correlated endpoint metric," in which the multiplicity adjustment for each of the m response variables depends on how that response variable "on the average" is inter-correlated with the remaining $(m - 1)$ response variables. This multi-correlated endpoint metric approach has been further improved^[26,27] for better FWER control. These improvements easily provide adjusted confidence intervals which are often considered necessary to report for clinical trials.

Fixed and Flexible Fixed-Sequence (Fallback) Methods

In the fixed sequence (FS) method, hypotheses are tested in a pre-specified order at the full alpha level as long as all preceding hypotheses are rejected. Once a testing sequence breaks (that is, a hypothesis is not rejected) subsequent hypotheses cannot be tested. This method is useful for situations when a hypothesis has to be tested first otherwise there is no meaning in testing the second hypothesis. For example, some osteoporosis prevention trials in post-menopausal women test two primary endpoints, which are the incidence of endometrial hyperplasia (EH) and the percent change in lumbar spine bone mineral density (BMD) at one year. Unless the incidence of EH, which is a critical safety endpoint, is ruled at a level x (e.g., $x = 4\%$) by an α -level test with some other clinical safety constraints, such as the point estimate of EH not to exceed 1% in a one year trial, it may not be meaningful to test for efficacy for the BMD endpoint.

A disadvantage of the fixed-sequence testing method is that once a hypothesis in the sequence is not rejected, no efficacy can be claimed for the subsequent hypotheses regardless of their clinical relevance or how small the p -values were, if they were tested. This has been the cause of difficulties in interpreting some clinical trials. An example is the carvedilol trials.^[28,29] In this case, the experimental treatment failed to show a statistically significant result for the planned primary endpoint, namely change in the exercise tolerance which was tested first at the 5% α -level, but the p -value for mortality endpoint was quite small—smaller than the 1% level. From a statistical perspective, as all alpha was spent in testing the exercise endpoint, the mortality endpoint was not considered statistically significant. However, clinicians were not

convinced by this argument, and with some debate and clinical evaluation of the findings, they trumped the result in favor of treatment benefit for the mortality endpoint. The flexible fixed-sequence (FFS) methods, commonly known by the name “fallback” methods,^[30,31] then came about as improvements of the trial design for the purpose of avoiding these types of situations. This method allows testing of subsequent hypotheses in a sequence even if a test in the testing sequence is not rejected.

In the case of testing two null hypotheses H_1 and H_2 , the conventional fallback method tests H_1 at level α_1 , where $\alpha_1 < \alpha$; both α_1 and the total α are pre-specified. If H_1 is rejected, then H_2 is tested at the full significance level of α , otherwise it is tested at the reduced significance level of $\alpha_2 = \alpha - \alpha_1$. The case of testing more than two hypotheses in the fallback method has been straightforward. In addition, extension of this fallback method for parametric models, known as the parametric fallback method, is also available.^[32]

The Adaptive Alpha Allocation Approach (4A) and the Consistency Ensured Methods

The fallback method and its extensions are attractive in saving alpha for testing additional hypotheses. However, they do not go beyond PAAS for increasing the chance of a positive trial in rejecting the second hypothesis H_2 if the trial does not establish treatment efficacy on the first hypothesis H_1 . The adaptive alpha allocation approach (4A) by Li and Mehrotra^[33] is an interesting improvement in this regard. It takes into account the result of testing H_1 for determining the significance level for testing H_2 , consequently leading to some increase in power for testing H_2 . Like the fallback method, the 4A method tests the first hypothesis H_1 at level α_1 , which is at a level somewhat less than the full α . The test procedure tests the second hypothesis H_2 at the significance level α_2 which is determined adaptively by using a non-increasing function of the observed *p*-value associated with H_1 . Interestingly, this α_2 can either exceed or be less than α_1 , even when the first hypothesis H_1 is not rejected at level α_2 . It exceeds α_1 when the *p*-value associated with H_1 is small enough, and it is less than α_1 when the *p*-value for testing H_1 is large enough. Thus, in this method, a weaker result for the first hypothesis requires a much stronger result for the second hypothesis.

For clinical trial applications, the 4A method in hierarchical testing of two primary endpoints—a single primary endpoint and its alternative—takes advantage of the level of evidence of the treatment effect in the first endpoint to determine the significance level for testing for the second endpoint. However, like other methods, it allows for establishing efficacy for the second alternative primary endpoint even if the results for the first endpoint were very weak, or were in the opposite direction (i.e., in the harmful direction). This raises a potential

interpretation issue in study findings when each of the two primary endpoints is allowed to establish efficacy yet their findings are unexpectedly quite inconsistent with each other. Some methods are now available for addressing this case, which require establishing a certain degree of evidence of positive result for the first hypothesis at a minimum before proceeding to test for the second hypothesis.^[34–36] These methods aim to ensure a certain degree of consistency in efficacy findings and are particularly useful in testing for treatment effect in a targeted subgroup of patients when the total population *p*-value—if not statistically significant in favor of the treatment—has to show some consistency of result with the subgroup finding.

Dunnett's Test

This is to date a classical and a frequently used test for many randomized laboratory experiments and even for clinical trials where multiple-treatment means are compared to that of the control; the multiple treatments are often multiple doses of the same treatment.^[5] This is an easy to use method and allows both one-sided and two-sided significance tests and confidence interval calculations. The multi-step versions of the test are also available.^[37] However, these tests are primarily designed for continuous endpoints under certain general linear model assumptions.

Scheffe, Tukey, and Tukey-Kramer Procedures

These procedures give great flexibility for many randomized biological experiments in which variations between experimental units are not of major concern and the experimenter is not able to specify specific comparisons or contrasts in advance. Scheffe's procedure provides a way of looking into all possible linear contrasts of *m* treatment means without worrying about the type I error rate inflation. Similarly, Tukey and Tukey-Kramer procedures provide tests and confidence intervals for all possible pairwise comparisons of the treatment means.^[5,38] However, as the number *m* increases, these methods become quite conservative in declaring significance in comparison to approaches that pre-specify a small number of test contrasts. Note that these tests are designed for continuous endpoints when certain general linear model assumptions hold.

Exact Tests of Contrasts

Given a specific family of test contrasts, software is now available that provides exact tests and simultaneous confidence intervals for many situations using simulation techniques.^[39] This technology assures that a multiple testing procedure (MTP) can be found for a specific application that is neither conservative nor liberal as long as the family of test contrasts is specified in advance.

MULTIPLE HYPOTHESES TESTING PRINCIPLES

Often clinical trials designs lead to complex multiplicity issues for which standard methods may not readily apply. However, there are some useful multiple hypotheses testing principles that can help in developing a suitable statistical testing strategy for solving such a multiplicity problem. Below we introduce some of these principles. Interestingly, some of the well-known multiplicity adjustment methods follow from one or more of these principles. For example, the Hochberg method was derived by applying the closed testing principle to the Simes's global test procedure.^[18] The sequential and parallel gatekeeping methods and their variations follow from the hierarchical testing and closed testing principles.

Union-Intersection (UI) and Intersection-Union (IU) Testing Principles

This is a useful principle for testing multiple hypotheses and for obtaining simultaneous confidence intervals or regions.^[5,40] The potential of this method for multivariate data, correlated endpoints, and specific interest decision rules is still being researched. As before, let H_j (for $j = 1, \dots, m$) represent m component null hypotheses, and let C_j 's be their corresponding rejection regions, and C their union. Then the α -levels, that is, the sizes of C_j 's are so adjusted that the overall size of C is $\leq \alpha$. In this setup, as the overall rejection region is the union of the individual rejection regions of the component null hypotheses, the acceptance region is the intersection of the complement of these individual rejection regions, leading to a UI scenario. The overall null hypothesis having this region of rejection C is often called the intersection null hypothesis and the alternative as the alternative union hypothesis. In the UI scenario, the intersection hypothesis being false implies that at least one of the component null hypotheses is false, and the intersection hypothesis being true implies that all component hypotheses are true.

An alternate situation often arises, for example, when given m multiple primary clinical endpoints, a clinical decision rule may require significant results in all the m endpoints. This situation gives rise to an intersection-union (IU) testing scenario. The overall rejection region is the intersection of the individual rejection regions of the component nulls and the acceptance region is the union of the complement of these individual rejection regions. The null hypothesis for this case is then the union of the component null hypotheses. There is no multiplicity adjustment for IU testing if each test is performed at the same level α (e.g., $\alpha = 0.05$), but it impacts the type II error rate, and consequently, the size of the trial. The maximum type I error rate for IU testing is the same as α if each individual test is performed at level α . Often, clinical decision rule in a clinical trial requires that r out of the m multiple endpoints show effectiveness. This scenario requires combining the

IU and UI strategies in setting up the null and alternative hypotheses.

Global Testing

Global tests are designed to test the global null hypothesis for a given family of null hypotheses. In this hypothesis, one assumes that all null hypotheses are true null hypotheses. A null hypothesis of the type $H_0^{(m)}: \delta_1 = 0, \delta_2 = 0, \dots, \delta_m = 0$ is a global null hypothesis of dimension m , also it is called the intersection hypothesis of dimension m . In this hypothesis, $\delta_j = 0$, for example, could be the parameter indicating the treatment effect of the j th endpoint in testing for m endpoints. A global test follows the union-intersection principle and is generally used to find a combined treatment effect across several endpoints. It can be more powerful than tests for individual endpoints in the presence of such a combined effect. On the other hand, it can be less powerful than the individual endpoint tests if the treatment effect is located in a very few isolated endpoints.

However, once a global test result is found statistically significant, the usual interest is to find out which endpoints or which component hypotheses are specifically responsible for such a treatment effect. One way to achieve this specificity goal is to apply the so-called "closure principle" to the global test method used. This is addressed in the next section. Some examples of global tests are Hotelling's T^2 test, O'Brien's OLS/GLS tests, O'Brien's rank-sum global test, Simes' test, Läuter exact test, and the likelihood ratio-based global tests.^[26,41–43]

Closed Testing Principle

The closed testing principle, proposed by Marcus et al.,^[44] is an important general purpose principle for solving many multiple testing problems that arise in all sorts of pharmaceutical experiments including clinical trials. The principle is fairly simple and basic. Many multiple testing methods, such as Holm's,^[15] Hommel's,^[17] and Hochberg's^[18] tests follow from this principle. The first step in this principle, when given a family of m (null) hypotheses to test, is to construct a closed family of intersection hypotheses. That is, if any two intersection hypotheses belong to this family, then their intersection also belongs to this family. By definition, it also includes the m individual hypotheses and the global null hypothesis. For example, given a family of three endpoint hypotheses H_1, H_2 , and H_3 to test, the family $\{H_1 \cap H_2 \cap H_3, H_1 \cap H_2, H_1 \cap H_3, H_2 \cap H_3, H_1, H_2, H_3\}$ is such a closed family under intersection. The notation $H_1 \cap H_2 \cap H_3$ stands for the intersection hypothesis of dimension 3, where $\delta_1 = 0, \delta_2 = 0$, and $\delta_3 = 0$; $H_1 \cap H_2$ for the intersection hypotheses of dimension 2, where $\delta_1 = 0$ and $\delta_2 = 0$; H_1 for the hypothesis $\delta_1 = 0$; and so on.

The closure principle states that a testing procedure that controls FWER in the strong sense can be constructed by testing each intersection hypothesis (and each individual

hypothesis) in this family using a suitable α -level test that controls FWER at least in the weak sense. Such an α -level test can be an appropriate α -level global test. This procedure rejects a hypothesis of this family if all intersection hypotheses in that family containing that specific hypothesis are rejected. In the example of the three endpoint hypotheses, in order to reject H_1 at level α , each of the four hypotheses $H_1 \cap H_2 \cap H_3$, $H_1 \cap H_2$, $H_1 \cap H_3$, and H_1 must be rejected by a suitable α -level test. Lehman et al.,^[45] for example, applied the closure principle with the OLS and GLS global tests for inferences on individual endpoints.

The closed testing has some interesting features. In order to claim a result for a specific endpoint, it requires that this endpoint must show level α test results by itself and also in association with all other endpoints in the family in all possible combinations. This is a sensible scientific requirement given the possibilities of multiple scenarios of effects and no effects in the multiple endpoints in association with the given endpoint of interest. A closed testing procedure is useful for the case when specific endpoints have modest treatment effects. In this case, individual endpoints may fail to show results, but jointly they may show a result on collecting strengths from each other. This concept is similar to that in meta-analysis of multiple studies, when individual studies may fail to show results but collectively they may have a result. For example, for the three-endpoint case, H_1 and H_2 may not be rejected at level α because of modest treatment effects for each, but $H_1 \cap H_2 \cap H_3$, $H_1 \cap H_2$ may be rejected each at level α establishing a result for α joint effect of endpoints 1 and 2.

There has been some interest in investigating shortcut approaches to performing closed tests. In this regard, it is interesting to note that the close testing principle with the Bonferroni test or the weighted Bonferroni test of intersection hypotheses of different dimensions, leads to the corresponding unweighted and weighted Holm's tests as shortcut methods. In this exercise, the Bonferroni test would reject an intersection hypothesis of dimension k , if the minimum of p -values associated with the individual hypotheses in that intersection is less than α/k .

Partitioning Principle

The partitioning principle is another interesting idea for performing multiple hypotheses tests. However, at present only limited literature is available on this topic.^[46,47] The idea is to first partition the union of hypotheses into disjoint hypotheses. Then test each disjoint hypothesis at the same level α . This, similar to close testing, does not compromise the FWER control at level α . Note that the partitioning principle does not depend on the closed family of intersection hypotheses. An attractive feature of this principle is that it can also provide simultaneous confidence intervals.

As an example, consider a clinical trial in which, for a given clinical endpoint X , two doses D_1 and D_2 are compared to a control D_0 . Let θ_0 , θ_1 , and θ_2 be expected parameter values of X for these dosages. Consider the following parameter space partitions: $A = \{\theta_1 \leq \theta_0, \theta_2 \leq \theta_0\}$, $B = \{\theta_1 \leq \theta_0, \theta_2 > \theta_0\}$, $C = \{\theta_1 > \theta_0, \theta_2 \leq \theta_0\}$, and $D = \{\theta_1 > \theta_0, \theta_2 > \theta_0\}$. Note that A , B , C , and D are disjoint and add up to the total joint parameter space of θ_1 and θ_2 . One can then test the following hypotheses in sequential steps as shown.

- Step 1: Test the null H_{01} : $\omega \in A$ versus H_{a1} : $\omega \in A^c$ (complement of A). If H_{01} is rejected then conclude that at least $\theta_1 > \theta_0$ or $\theta_2 > \theta_0$, and go to Step 2 else stop.
- Step 2: H_{02} : $\omega \in B$ versus H_{a2} : $\omega \in (A \cup B)^c = \{\theta_1 > \theta_0\}$. If H_{02} is rejected then conclude that $\theta_1 > \theta_0$, else stop.
- Step 3: Test the null H_{03} : $\omega \in C$ versus H_{a3} : $\omega \in (A \cup B \cup C)^c = \{\theta_1 > \theta_0, \theta_2 > \theta_0\}$. If H_{03} is rejected then conclude that in addition to $\theta_1 > \theta_0$, also $\theta_2 > \theta_0$.

Hierarchical Testing Principle/Framework

Because various objectives of a trial differ in importance with respect to regulatory and scientific utility, a useful and efficient statistical framework for addressing multiplicity problems of trials is to rank-order the objectives of the trial into primary and secondary objectives. In addition, if there are several primary objectives, then they can be rank-ordered as well provided that this is clinically meaningful. In the statistical terminology, this rank-ordering of the objectives of a trial is known as hierarchical ordering of families of hypotheses of a trial. In this framework, one addresses multiplicity issues for the primary objectives first independent of the other objectives of the trial as if the other objectives did not exist. This framework has an advantage in that it allows recycling of used alpha, either in full or part (depending on the testing scheme), from a family of hypotheses to the next family for those hypotheses of the previous family which are statistically significant.^[48–53]

Note that the primary objective of the trial traditionally determines whether the trial is positively successful or not with respect to planned primary endpoints. If a trial is positive, it can characterize clinically meaningful benefits of the study treatment. Primary endpoints for this purpose are usually important clinical endpoints. These are few in number, usually 1–4, and trials are normally sufficiently powered for these endpoints for detecting meaningful treatment differences. After primary objectives of the trial are meaningfully met, and the trial is found to be a positive trial, then one can go ahead and address multiplicity for secondary objectives according to a prospectively planned test strategy. However, solving multiplicity problems for the primary and secondary objective requires formulating the problem and its solutions at the design stage.

Gatekeeping Testing Strategies

Gatekeeping testing strategies are useful in solving many clinical trial multiplicity problems that can be framed in terms of testing of hierarchical families of multiple hypotheses. The families of hypotheses are examined sequentially with each family serving as a gatekeeper for the subsequent families. Two types of gatekeeping testing strategies are popular: *parallel* and *serial*. In the serial strategy, one tests hypotheses within each family (gate) using a method that controls the familywise error rate for each gate and proceeds to the next gate when all of the hypotheses in the current gate are rejected.^[52] On the other hand, the parallel gatekeeping strategy requires only the rejection of at least one hypothesis in each gate.^[53]

As an illustration, consider a trial in which the primary objective is to show superiority of a test treatment to placebo for endpoints A and B, and then to test for superiority for the two endpoints C and D after this primary objective is met. This problem fits into the gatekeeping testing framework of the two families of hypotheses $F_1 = \{H_A, H_B\}$ and $F_2 = \{H_C, H_D\}$ as these two families are hierarchically ordered. There can be twotypes of clinical decision criteria for meeting the primary objective. In one, both A and B need to show statistically significant treatment benefits, because both endpoints A and B are needed to characterize a meaningful treatment benefit. In the other, a statistically significant treatment benefit in either A or B is sufficient to characterize treatment benefit. In the first type the F_1 gate in the gatekeeping method is of serial and in the second type it is of parallel nature.

Gatekeeping testing strategies are based on the closed testing principle that requires tracking $2^n - 1$ intersection hypotheses, n being the total number hypotheses across all families of hypotheses, as such, algorithms^[53] become messy for many applications. However, these methods guarantee FWER control at level α in the strong sense as long as the intersection hypotheses tests in the closed testing algorithms have the FWER control at level α . If the parallel gatekeeping procedure uses the weighted Bonferroni test method for testing intersection hypotheses in its closed testing algorithm, then it is called the Bonferroni parallel gatekeeping procedure. This procedure is of special interest as it allows a simple shortcut method for clinical trial applications.

This shortcut method is built around the concept of *rejection gain factor*. A formula for calculating this rejection gain factor, and some examples of applications of this shortcut method, can be found in Dmitrienko-Tamhane et al. articles.^[50,51] This Bonferroni based shortcut method basically allows recycling of alpha by this gain factor formula from one family to the next for those hypotheses that are rejected in the previous family. Consider, for example, $F_1 = \{H_A, H_B\}$, and say that H_A was tested at level $\alpha_1 = w_1\alpha = 0.035$ and H_B at level $\alpha_2 = w_2\alpha = 0.01$, with $w_1 + w_2 = 0.9 < 1$, and in this testing, H_A was rejected but H_B was

not rejected. Then the alpha of 0.035, associated with H_A , can be recycled (i.e., passed) from F_1 to F_2 giving a total alpha of $0.035 + 0.005 = 0.04$ for testing H_C and H_D in F_2 by the Holm's test.

MULTIPLICITY CONSIDERATIONS FOR SOME CLINICAL TRIALS

Clinical Decision Rules

A useful road map for identifying and addressing multiplicity for confirmatory control clinical trials is to outline at the design stage of the trial the “win” scenarios, which are sometimes referred to as clinical decision rules. This win criteria would specify how a positive clinical decision regarding the effectiveness of the study treatment is going to be reached based on the results of one or more primary endpoints or based on one or more dose-control treatment comparisons. These win scenarios are basically a set of prospectively defined decision rules in terms of statistical results of the primary endpoints and primary treatment comparisons that determines whether the trial is successful or not. If the trial is successful in this way, then one is able to characterize meaningful benefits of the study treatment. Such criteria represent a clinically acceptable way of assessing benefits of a treatment for the disease and patient population under consideration. Statistically, such win criteria for a clinical trial refer to alternative hypotheses and the corresponding complements to their null hypotheses situations. In the following discussion, we list some examples of such win criteria or clinical decision rules in terms of primary endpoint(s) for trials that compare a treatment to a control.

1. Given a single specified primary endpoint that can characterize treatment benefit, the treatment needs to be superior to the control for this endpoint.
2. Given several specified primary endpoints that can independently characterize treatment benefit, the treatment needs to be superior to the control for at least one of them.
3. Given m ($m \geq 2$) specified primary endpoints, the treatment needs to be superior to the control for all of them for a meaningful treatment benefit.
4. Given three specified primary endpoints E_1 , E_2 , and E_3 , the treatment needs to be superior to the control either for E_1 or for both E_2 and E_3 .
5. Given three specified primary endpoints E_1 , E_2 , and E_3 , the treatment needs to be superior to the control either for (E_1 and E_2) or for (E_1 and E_3).
6. Given two specified primary endpoints, the treatment needs to be superior to the control for one and at least marginally superior to the control for the other.
7. Given two specified primary endpoints, the treatment needs to be superior to the control for one and at least non-inferior to the control for the other.

8. Given several endpoints that are in a hierarchy that specifies the order in which they are to be evaluated, and significance of all preceding endpoints is required in order to proceed to the next one.
9. Given m specified primary endpoints, at least k of them needs to be statistically significant in favor of the treatment and the remaining $(m-k)$ endpoints show at least a trend in the right direction.

Note that decision rules for wining in examples 1, 3, 7, and 8 do not require any multiplicity adjustment. Each test can be carried out at the same level α without any adjustment. The null hypothesis structure for some of the above rules may not be straightforward like the simple intersection or union null hypothesis of m individual hypotheses; it may involve a mixture of these. Solutions to these problems can found on applying some of the principles and concepts introduced in the section “Multiple Hypotheses Testing Principles”.

Primary and Secondary Endpoint Concepts

Primary endpoints in clinical trials play an important role in characterizing treatment benefits and in achieving the main objectives of the trial. Key design features of the trial generally depend on the primary endpoint specifications, such as trial size, data monitoring, and analysis strategies. Secondary endpoints also play an important role in providing additional information, for example, for interpreting results of the primary endpoints, for expanding or providing additional characterization of treatment effects, and also for generating new hypotheses for future trials.^[54–58] Secondary endpoints by themselves are generally considered not sufficient for characterizing treatment benefits. However, if a secondary endpoint is an endpoint like mortality or irreversible morbidity, it could reach the status of a primary endpoint.

Composite Endpoint Issues

Often trials, particularly cardiovascular trials, combine events from component endpoints to construct a composite in order to increase the number of events for such an endpoint. An example is the first occurrence of death, myocardial infarction (MI), or recurrent angina for a cardiovascular trial. Such a construct gives more power to the statistical test for treatment difference if each component included in the composite has some sensitivity for detecting treatment difference. However, if a component endpoint included in the composite has no such sensitivity, it may instead hurt the cause by adding more noise to the composite.

A composite-endpoint-based statistical test provides an overall test for the component endpoints in the composite. Testing of the individual components in the composite requires adjustment for multiplicity when claiming treatment efficacy for specific components. However, difficulties arise in an analysis of a specific individual component

of the composite, as the events of the other components become censored observations. In addition, clinical interpretation of the composite is often difficult when its softer components drive the result or all its components do not trend positively.

There are some key considerations and issues requiring attention when using a composite endpoint as a primary endpoint.^[59–62] For example, the components of a composite endpoint must be clinically relevant, consistent or coherent with each other, and interpretable for the disease and the intervention under study. One should prospectively define the composite endpoint itself and all its components along with their ascertainment methods. Component endpoints are to add to the total treatment effect with induced treatment effects in a similar direction. Endpoint ascertainment methods are to capture accurately both the occurrence and non-occurrence of the component events. The results of the component endpoints are to be fully disclosed and displayed along with the results of their composite to allow a meaningful interpretation of the composite endpoint result.

A composite endpoint is usually not recommended as a primary endpoint when its components can be of widely differing importance to patients and can occur with differing frequency, and their magnitude of effects can vary markedly. Such a situation can complicate interpreting the result of a composite.

Approaches for Reducing the Burden of Multiplicity

Clinical trials generally test many hypotheses. This can easily overload a trial with excessive multiplicity burden. Therefore, one of the considerations in designing a trial is to try to gain efficiency by reducing the burden of multiplicity as much as possible. In this regard, several approaches may be taken. A popular approach is to order the families of hypotheses hierarchically as well as, where possible, the hypotheses within each family. Families of null hypotheses are said to be hierarchically ordered or ranked if earlier families serve as gatekeepers in the sense that one tests hypotheses in a given family if the preceding gatekeepers have been successfully passed. The two commonly used hierarchical families of endpoints in a clinical trial are the family of primary endpoints and the family of secondary endpoints. These two families are hierarchically ordered with the property that rejections or non-rejections of null hypotheses of primary endpoints do not depend on the results of the secondary endpoints. This allows the use of gatekeeping testing strategies that can considerably reduce the multiplicity burden in controlling the familywise error rate.

The use of dependence information between the test statistics, if available, can also help in reducing the extent of multiplicity. Another approach in this regard is to combine several endpoints to a single or to a few composite endpoints, if certain conditions are met and if such composite endpoints are clinically acceptable.

Dose-Response and Dose-Control Comparison Trials

These trials are generally carried out for early phase trials in contrast to confirmatory trials. Given the dose-response data on an endpoint at m doses of a test treatment, the early phase trials fit a set of suitable parametric dose-response models, such as E -max and logistic models and their variants. A working model, which is the best candidate for describing the dose response relationship, is then selected from these fitted models on performing statistical tests of model-contrasts. This working model is then used for proof of activity of the test treatment in the dose range of the trial and for deciding which few doses should be selected for the confirmatory trial. In this approach, the need for multiplicity adjustments arises for addressing uncertainty in model selection. Bretz et al.^[63] have provided a method for handling this multiplicity issue. Some early trials also use multiple comparisons based approach which treats dose as a qualitative factor with a very few assumptions, if any, about the underlying dose response model. This approach is often used for identifying minimum effective dose (MED) dose.^[64]

Confirmatory trials with multiple doses, however, are dose-control comparison trials with a basic design that includes m doses ($d_1, \dots, d_m$) of the test treatment, placebo (d_0), and at least one dose of a relevant active control. The value of m is generally small and the goal of such a trial is to prove overall efficacy of the test treatment and for its targeted doses, and in addition, if a dose is effective, to assess its standing in comparison to the active control of the trial. Bauer et al.^[65] have investigated this design in detail and have proposed various inferential strategies for dealing with questions of interest and multiplicity issues. There are variations of this basic design: some assume order restriction among doses and some do not; some exclude placebo for ethical reasons; some exclude the active control arm; and some perform superiority as well as non-inferiority tests. Some trials make dose-comparisons on multiple primary and secondary endpoints with logical restrictions, such as a dose-comparison for a secondary endpoint cannot be made unless such a comparison for the primary endpoint is statistically significant. These dose-control comparison designs raise multiplicity problems at advance levels. These advance multiplicity problems are often solved by the tree-structured gatekeeping procedures,^[66] which combine the serial and parallel gatekeeping features with logical restrictions.

Three-Arm “Gold Standard” Trials

Active control trials when conducted with three treatment arms, a placebo (P), an approved active control (S) and the test treatment (T) are often referred to as three-arm “gold standard” trials. These trials are common for non-serious diseases when treating patients in the trial with placebo is

not unethical. The objective of such a trial is normally two-fold; that is, the trial has assay sensitivity and the new treatment retains efficacy in comparison to the control by more than a prespecified percentage λ_0 . Showing assay sensitivity requires testing $H_1: \mu_S - \mu_P \leq 0$ versus $H_{a1}: \mu_S - \mu_P > 0$ and confirming that the size of the effect $\mu_S - \mu_P$ is in line with the previous historical experience in the patient population studied in this trial. The parameters μ_P , μ_S , and μ_T stand for the expected mean responses for the three treatment arms; a greater treatment mean implies a greater treatment benefit. Showing at least λ_0 percent retention requires testing $H_2: \lambda \leq \lambda_0$ versus $H_{a2}: \lambda > \lambda_0$, where $\lambda = (\mu_T - \mu_P)/(\mu_S - \mu_P)$.

A rigorous statistical solution to the problem that would address all aspects of this problem is a topic of current research. However, some statistical literature is available that address multiple comparison issues of such trials.^[67–69] This design would normally require allocating more patients to the active arms than to placebo, depending on the size of the treatment effects, specification of λ_0 , and variability of the measurements.

Drug Combination Trials

Some diseases require treatment with a combination drug. A two-active combination drug trial has four treatment arms: a combination arm, two component arms, and a placebo arm. The trial aims to show the contribution to efficacy of each component and that the combination is superior to each component. Sometimes these trials are three-arm trials without placebo, provided that each component is an approved product for efficacy and safety. Each comparison test is an α -level test and there is no type I error rate inflation issue. However, the type II error is an issue. It is often harder to show the superiority of the combination over the most effective component. Therefore, the trial may need design improvements to reduce variability.

The win criterion for efficacy for a drug combination trial can become complicated if two or more primary endpoints or two or more components are required in a trial to show the benefit of the combination over its approved components. For example, for the case of two endpoints, the win criterion for efficacy of the combination may be to show superiority for one endpoint and at least non-inferiority for the other endpoint when comparing the combination to each component.^[70] Sometimes multi-dose factorial designs are employed for the assessment of combination drugs for serving dual purpose, to provide confirmatory evidence that the combination is more effective than either component and to identify an effective and safe dose combination or a range of useful dose combinations.^[71]

Subgroup Analysis

Subgroup analyses are widespread in clinical trials, especially for large confirmatory trials. Sometimes the purpose of this analysis is exploratory to find out whether the

efficacy results are consistent across clinically relevant subgroups, such as race, gender, age, and region. Sometimes it is used to explore the possibility of large treatment differences in certain subgroups—for example, patients whose disease is caught early versus patients whose disease has progressed to a certain stage. However, subgroup analyses generally pose analytical challenges and difficult inferential issues including multiplicity and lack of power issues. Interpretation of subgroup analysis results can be problematic unless it is pre-specified at the design stage based on certain background factors that are clinically relevant. It is well understood in the clinical trial literature that subgroup analyses are an important part of the analysis of comparative trials, but their results are frequently exaggerated and often misinterpreted leading to misguided research and suboptimal patient care.^[72–74]

Among the factors which reduce the reliability of subgroup findings is that often subgroup analyses are data driven and unplanned. The biological plausibility of positive findings is argued after the data has been analyzed and results seen. These analyses may have been based on improper subgroups, characterized by a variable measured after randomization and potentially affected by treatment and study procedures. Such unplanned subgroup analyses usually produce false positive results because of bias and multiplicity issues, and also produce false negative results because of lack of power. Therefore, their *p*-values in general cannot be validly interpreted.

On the other hand, interpretation of subgroup findings is enhanced if the clinical trial had planned at the design stage to evaluate treatment effects in a targeted subgroup of patients selected by means of a classifier (such as a biomarker classifier) that is able to select potential responders to the treatment. A subgroup analysis, whether planned or unplanned, can experience a significant loss of power in detecting a treatment effect. However, this loss may not occur for such a targeted subgroup if its treatment effect size happens to be sufficiently greater than that for the total patient population of the trial. Several authors have addressed this type of design and their analyses covering both the multiplicity and power issues of treatment comparisons.^[35,75,76] These designs can potentially improve the chance of trial success.

CONCLUDING REMARKS

In this article, we have clarified the role of *p*-values in hypothesis testing and how it provides evidence in favor of a scientific claim in the area of controlled clinical trials. A small *p*-value is an indicator that the data at hand are in conflict with the null hypothesis, and in doing so, it provides evidence in favor of the alternative hypothesis; the smaller this *p*-value the greater this evidence. However, many undesirable factors that could occur in a clinical trial, including poor design and conduct of the trial, could create

bias and make a *p*-value small without having anything to do with the treatment effect. Therefore, clinical trial results must be carefully evaluated to rule out such bias before concluding a treatment benefit.

We have also tried to clarify the difference between statistical evidence and substantial evidence concepts used in regulatory decision making. Substantial evidence for the benefit of a new drug or treatment generally requires statistically conclusive results from at least two adequate and well-controlled trials. In special cases, depending on the seriousness of the disease and ethical and clinical considerations, a single trial result with a sufficiently small *p*-value may provide this evidence. Goodman^[13] shows that if a trial result has an associated *p*-value of 0.05, then the probability of replicating such a result in a second study at the 0.05 significance level is only 50%. It is only when the *p*-value is 0.001 that this replication probability increases to 90%. Other investigators of *p*-values also emphasize that *p*-values in general overestimate the evidence against the null hypothesis. Therefore, we support the idea of a *p*-value less than 0.001 for a result that could qualify substantial evidence status for a treatment benefit claim derived from a single trial, provided such a result after adjusting for multiplicity satisfies within-trial consistency and is clinically meaningful.

Multiplicity is common in clinical trials and requires special attention. Multiplicity inflates the type I error rate, which is the probability of concluding that the treatment is effective when in reality it is not. We have clarified some general concepts relevant to controlling the type I error rate in the presence of multiplicity. In addition, we have explained some basic principles behind some standard procedures, including those of union-intersection and intersection-union methods, closure, fixed-sequence, fallback, partitioning, and gatekeeping testing. We have also elucidated when to use a single-step and when to use a multi-step procedure. A single-step procedure that controls FWER strongly has more clinical utility—in addition to controlling the type I error probability for specific contrasts, it also easily provides their adequate interval estimates. Thus, such a method is useful when, in addition to testing of multiple hypotheses, estimation of treatment effect contrasts are also important. On the other hand, a multi-step procedure usually preserves the power of the test more efficiently.

In dealing with multiple-endpoint issues in clinical trials, it is important to define a family of null hypotheses of main interest considering claims to be made for the treatment effect. Such a family should be tightly chosen based on clinical and statistical considerations, such as desired sensitivity and relevancy of the endpoints with regard to the disease and treatment to be studied in the trial. A closed testing procedure is usually recommended for such a family, as it provides a better inferential picture across multiple endpoints at the global and marginal hypothesis levels. If two main families of null hypotheses are defined, one family for the primary endpoints and the other for

the secondary endpoints, there are ways to approach this problem, which we have addressed in this article.

Often a composite endpoint is constructed out of the multiple individual endpoints. A test for the composite endpoint is a test for an overall effect of the multiple endpoints in the composite. Testing of individual components in the composite requires adjustment for multiplicity, unless these analyses are intended only for the interpretation purposes and not for making new or additional claims of treatment benefits. In general, the use of a composite endpoint as a primary endpoint is not recommended when its components are likely to be of widely differing importance and can occur with differing frequency, and their magnitude of effects can markedly differ. Such a situation can complicate interpreting the result of a composite.

In conclusion, because the subject matter of multiplicity in clinical trials is truly vast and because of space considerations, this article has only briefly addressed relevant issues, and its scope has been limited to basic concepts and practices for addressing multiplicity issues in clinical trials. Some important multiplicity topics, such as interim analyses and adaptive designs, have been omitted.

DISCLAIMER

Views expressed in this article are those of the authors and not necessarily of the U.S. Food and Drug Administration (FDA). No official support or endorsement of the material presented in this article by the FDA is intended or should be inferred.

REFERENCES

- Schervish, M.J. *P*-values what they are and what they are not. *Am. Stat.* **1996**, 50 (3), 203–206.
- Berger, O.J.; Berry, D.A. Statistical analysis and the illusion of objectivity. *Am. Sci.* **1988**, 76, 159–165.
- Hung, H.M.J.; O'Neill, R.T.; Bauer, P.; Köhne, K. The behavior of the *P*-value when the alternative hypothesis is true. *Biometrics* **1997**, 53, 11–22.
- Lehman, E.L. *Testing Statistical Hypotheses*; John Wiley & Sons: New York, 1985.
- Hochberg, Y.; Tamhane, A.C. *Multiple Comparison Procedures*; John Wiley & Sons: New York, 1987.
- Huque, M.F.; Sankoh, A.J. A reviewer's perspective on multiple endpoint issues in clinical trials. *J. Biopharm. Stat.* **1997**, 7, 545–564.
- Dubey, S.D. Robust Statistical Evidence for Establishing Effectiveness of New Pharmaceuticals. *Proceedings of the Section on Biopharmaceuticals, American Statistical Association*; Alexandria, VA, 1990; 292–299.
- Dubey, S.D.; Chi, G.Y.H.; Kelly, R.E. *The FDA and the IND/NDA Statistical Review Process in Chapter 3 of Statistics in the Pharmaceutical Industry*, 3rd Ed.; Buncher, R.C.; Tsay, J.Y.; Eds.; Chapman & Hall/CRC Taylor and Francis Group: New York, 2006; 55–78.
- US FDA. The Quantity of Evidence to Support Effectiveness. *Guidance for Industry: Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products*; **1998**, <http://www.fda.gov/cder/guidance/>
- Koch, G.G. Commentary on “Statistical consideration of the strategy for demonstrating clinical evidence of effectiveness—one larger vs. two smaller pivotal studies”. *Stat. Med.* **2005**, 24, 1639–1646.
- Huque, M.F. Commentary on “Statistical consideration of the strategy for demonstrating clinical evidence of effectiveness—one larger vs. two smaller pivotal studies”. *Stat. Med.* **2005**, 24, 1646–1651.
- Shun, Z.; Chi, E.; Durrleman, S.; Lloyd Fisher, L. Statistical consideration of the strategy for demonstrating clinical evidence of effectiveness—one larger vs. two smaller pivotal studies. *Stat. Med.* **2005**, 24, 1619–1637.
- Goodman, S.N. A comment on replication, *p*-values and evidence. *Stat. Med.* **1992**, 11, 875–879.
- Geisser, S. *Predictive Inference*; Chapman and Hall: New York, 1993.
- Holms, S. A simply sequential rejective multiple test procedure. *Scand. J. Stat.* **1979**, 6, 65–70.
- Zhang, J.; Quan, H.; Ng, J.; Stepanavage, M.E. Some statistical methods for multiple endpoints in clinical trials. *Control Clin. Trials*. **1997**, 18, 204–221.
- Hommel, G.A. Stagewise rejective multiple test procedure based on modified Bonferroni test. *Biometrika* **1988**, 75 (2), 383–386.
- Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **1988**, 75, 800–802.
- Sarkar, S.K. Some probability inequalities for ordered MTP random variables; a proof of the Simes conjecture. *Ann. Stat.* **1998**, 26, 494–504.
- Tamhane, A.C.; Liu, W. On weighted Hochberg procedures. *Biometrika* **2008**, 95, 279–294.
- Moyé, L.A. Alpha calculus in clinical trials. Considerations and commentary for the new millennium. *Stat. Med.* **2000**, 19, 767–779.
- Šidák, Z. Rectangular confidence regions for the means of multivariate normal distributions. *J. Am. Stat. Assoc.* **1967**, 62, 626–633.
- Tukey, J.W.; Ciminera, J.L.; Heyse, J.F. Testing the statistical certainty of a response to increased doses of a drug. *Biometrics* **1985**, 41, 295–301.
- Dubey, S.D. Adjustment of *p*-values for multiplicities of intercorrelating symptoms. In *Statistics in the Pharmaceutical Industry*, 2nd Ed.; Buncher, R.C.; Tsay, J.Y.; Eds.; Marcel Dekker Inc.: New York, 1994, 513–527.
- Armitage, P. Two areas of controversy in the design and analysis of clinical trials. *Recent Topics Clin. Trials Eval.* **1985**, 13 (Suppl. III), 29–40.
- Sankoh, A.J.; Huque, M.F.; Dubey, S.D. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Stat. Med.* **1997**, 16, 2529–2542.
- Huque, M.F.; Quasem, M.; Rahman, A.; Dubey, S.D. An improved ad-hoc multiplicity adjustment method for correlated response variables. *Stat. Biopharm. Res.* **2009**, DOI: 10.1198/sbr.2009.0052.
- Fisher, L.D.; Moyé, L.A. Carvedilol and the Food and Drug Administration approval process. *Control Clin. Trials* **1999**, 20, 16–39.

29. Packer, M.; Bristow, M.R.; Cohn, J.N.; Colucci, W.S.; Fowler, M.B.; Gilbert, E.M.; Shusterman, N.H. The effects of carvedilol on morbidity and mortality in patients with chronic heart failure. *New Engl. J. Med.* **1996**, *334*, 1349–1355.
30. Wiens, B.L. A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharm. Stat.* **2003**, *2*, 211–215.
31. Wiens, B.L.; Dmitrienko, A. The fallback procedure for evaluating a single family of hypotheses. *J. Biopharm. Stat.* **2005**, *15*, 929–942.
32. Huque, M.F.; Alosch, M. A flexible fixed-sequence testing method for hierarchical ordered correlated multiple endpoints in clinical trials. *J. Stat. Plan. Inference* **2008**, *138*, 321–335.
33. Li, J.; Mehrotra, D. An efficient method for accommodating potentially underpowered primary endpoints. *Stat. Med.* **2008**, *27*, 5377–5391.
34. Song, Y.; Chi, G.Y. A method for testing a prespecified subgroup in clinical trials. *Stat. Med.* **2007**, *26*, 3535–3549.
35. Alosch, M.; Huque, M.F. A flexible strategy for testing subgroups and overall population. *Stat. Med.* **2009**, *28*, 3–23.
36. Alosch, M.; Huque, M.F. A flexible strategy for testing subgroups and overall population, *Proceedings of the Biopharmaceutical Section of the Joint Statistical Meetings* held at Salt Lake City, UT, 2007, 192–203.
37. Dunnett, C.W.; Tamhane, A.C. Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Stat. Med.* **1991**, *10*, 939–947.
38. Westfall, P.H.; Tobias, R.D.; Rom, D.; Wolfinger, R.D.; Hochberg, Y. *Multiple Comparisons and Multiple Tests Using the SAS System*; SAS Institute Inc.: Cary, NC, 1999.
39. *SAS System*; SAS Institute Inc.: Cary, NC, 2002.
40. Roy, S.N.; Bose, R.C. Simultaneous confidence interval estimation. *Ann. Math. Stat.* **1953**, *24*, 513–536.
41. Sankoh, A.J.; Huque, M.F.; Russel, H.K.; D'Agostino, R. Global two-group multiple endpoint adjustment methods applied to clinical trials. *Drug Inf. J.* **1999**, *33*, 119–140.
42. Läuter, J. Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics* **1996**, *52*, 964–970.
43. Tamhane, A.C.; Logan, B.R. Accurate critical constants for the one-sided approximate likelihood ratio test of a normal mean vector when the covariance matrix is estimated. *Biometrics* **2002**, *58*, 176–182.
44. Marcus, R.; Peritz, E.; Gabriel, K.R. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **1976**, *63*, 655–660.
45. Lehman, W.; Wassmer, G.; Reitmeier, P. Procedures for two sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* **1991**, *47*, 511–521.
46. Stefansson, G.; Kim, W.C.; Hsu, J.C. On confidence sets in multiple comparisons. *Statistical. In Decision Theory and Related Topics IV*; Gupta, S.S.; Berger, J.O.; Eds.; Academic Press: New York, 1988.
47. Finner, H.; Strassburger, K. The partitioning principle: A powerful tool in multiple decision theory. *Ann. Stat.* **2002**, *30*, 1194–1213.
48. Hommel, G.; Bretz, F.; Maurer, W. Powerful shortcuts for multiple testing procedures with special reference to gatekeeping strategies. *Stat. Med.* **2007**, *26*, 4063–4073.
49. Bretz, F.; Maurer, W.; Brannath, W.; Posch, M. A graphical approach to sequentially rejective multiple test procedures. *Stat. Med.* **2009**, *28*, 586–604.
50. Dmitrienko, A.; Tamhane, A.C.; Wang, X.; Chen, X. Stepwise gatekeeping procedures in clinical trial application. *Biom. J.* **2006**, *48* (6), 984–991.
51. Dmitrienko, A.; Tamhane, A.C. Gatekeeping procedures with clinical trial applications. *Pharm. Stat.* **2007**, *6*, 171–180.
52. Westfall, P.H.; Krishen, A. Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. *J. Stat. Plan. Inference* **2001**, *99*, 25–40.
53. Dmitrienko, A.; Offen, W.; Westfall, P. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Stat. Med.* **2003**, *22*, 2387–2400.
54. O'Neill, R.T. Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. *Control Clin. Trials* **1997**, *18*, 550–556.
55. Davis, C.E. Secondary endpoints can be validly analyzed, even if the secondary endpoint does not provide clear statistical significance. *Control Clin. Trials* **1997**, *18*, 557–560.
56. Prentice, R.L. Discussion: On the role and analysis of secondary outcomes in clinical trials. *Control Clin. Trials* **1997**, *18*, 561–567.
57. Moyé, L.A. P -value interpretation in clinical trials: The case of discipline. *Control Clin. Trials* **1999**, *20*, 40–49.
58. D'Agostino Sr., R.B. Editorial: Controlling alpha in a clinical trial: The case for secondary endpoints. *Stat. Med.* **2000**, *19*, 763–766.
59. Moyé, L.A. *Multiple Analyses in Clinical Trial*; Springer-Verlag: New York, 2003.
60. Lubsen, J.; Kirwan, B.A. Combined endpoints: can we use them? *Stat. Med.* **2002**, *21*, 2959–2970.
61. Montori, V.M.; Permyer-Miranda, G.; Ferreira-González, I.; Busse, J.W.; Pacheco-Huergo, V.; Bryant, D.; Alonso, J.; Akl, E.A.; Domingo-Salvany, A.; Mills, E.; Wu, P.; Schünemann, H.J.; Aeschke, R.; Guyatt, G.H. Validity of composite end points in clinical trials. *BMJ* **2005**, *330*, 594–596.
62. Chi, G.Y.H. Some issues with composite endpoints in clinical trials. *Fundam. and Clin. Pharmacol.* **2005**, *19*, 609–619.
63. Bretz, F.; Piheiro, J.C.; Branson, M. Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* **2005**, *61*, 738–748.
64. Tamhane, A.C.; Hochberg, Y.; Dunnett, C.W. Multiple test procedures for dose finding. *Biometrics* **1996**, *52*, 21–37.
65. Bauer, P.; Röhm, J.; Maurer, W.; Hothorn, L. Testing strategies in multiple-dose experiments including active control. *Stat. Med.* **1998**, *17*, 2133–2146.
66. Dmitrienko, A.; Wiens, B.L.; Tamhane, A.C.; Wang, X. Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Stat. Med.* **2006**, *26*, 2465–2478.
67. Pigeot, I.; Schäfer, J.; Röhm, J.; Hauschke, D. Assessing the therapeutic equivalence of two treatments in comparison with a placebo group. *Stat. Med.* **2003**, *22*, 883–899.
68. Koch, A.; Röhm, J. Hypothesis testing in the “gold standard” design for proving the efficacy of an experimental treatment relative to placebo and a reference. *J. Biopharm. Stat.* **2004**, *14* (2), 315–325.

69. Hauschke, D.; Pigeot, I. Rejoinder to establishing efficacy of a new experimental treatment in the gold standard design. *Biom. J.* **2005**, *47*, 782–786.
70. Logan, B.R.; Tamhane, A.C. Superiority inferences on individual endpoints following non-inferiority testing in clinical trials. *Biom. J.* **2008**, *50* (5), 693–703.
71. Hung, H.M.J. Evaluation of a combination drug with multiple doses in unbalanced factorial design clinical trials. *Stat. Med.* **2000**, *19*, 2079–2087.
72. Lagakos, S.W. The challenge of subgroup analyses—reporting without distorting. *N. Engl. J. Med.* **2006**, *354*, 1667–1669.
73. Yusuf, S.; Wittes, J.; Probstfield, J.; Tyroler, H.A. Analysis and interpretation of subgroups of patients in randomized clinical trials. *JAMA* **1991**, *266*, 93–98.
74. Wang, R.; Lagakos, S.W.; Ware, J.H.; Hunter, D.J.; Drazen, J.M. Statistics in medicine—reporting of subgroup analyses in clinical trials. *N. Engl. J. Med.* **2007**, *35* (21), 2189–2194.
75. Simon, R.; Maitournam, A. Evaluating the efficacy of targeted designs for randomized clinical trials. *Clin. Cancer Res.* **2004**, *10*, 6759–6763.
76. Maitournam, A.; Simon, R. On the efficiency of targeted clinical trials. *Stat. Med.* **2005**, *24*, 329–339.

Correlated Probit Model

Ralitza V. Gueorguieva

Department of Epidemiology and Public Health, Yale University, New Haven, Connecticut, U.S.A.

Abstract

The goals of this entry are to familiarize the reader with correlated probit models and to provide enough information and relevant references so that these models can be successfully applied using available statistical software.

INTRODUCTION

Probit models are commonly used to describe the relation between one or more independent variables and a quantal response (e.g., alive or dead; mild, moderate, or severe degree of illness). They are often motivated from the idea of a latent continuous variable underlying the quantal response and the existence of thresholds, such that if the unobserved continuous variable is below a certain threshold then the observed quantal variable is in a certain response category. For example, if toxicity is below the critical threshold the subject is alive, otherwise the subject is dead. However, the threshold concept and the existence of an underlying continuous variable are not necessary for model interpretability. The usual probit model is used when the sample consists of n independent cases with a single quantal outcome variable, while multivariate probit models are extending the usual probit models to cases of several quantal outcome variables observed on the same individual. In contrast to probit models for independent data, correlated probit models are useful in analyzing one or more quantal response variables when these variables are repeatedly measured within clusters or over time. For example, it may be of interest to model the probability of the presence of a particular illness or symptom in the members of the same family or on the same individual over the course of follow-up in a longitudinal study. As measurements on individuals within the same family or repeated measurements on the same individual are likely to be more highly correlated than measurements on unrelated individuals, special statistical methods are necessary to account for this correlation.

The goals of this entry are to familiarize the reader with correlated probit models and to provide enough information and relevant references so that these models can be successfully applied using available statistical software. To achieve this goal we first introduce the probit, multivariate probit, and correlated probit models, followed by

a discussion on model interpretation, identifiability, and model fitting. An illustrative example of a correlated probit model with both multiple response variables and clustering is provided, followed by a brief overview of available statistical software packages that can be used for model fitting.

HISTORICAL OVERVIEW

Probit Model for Uncorrelated Data

In many applications an observed binary variable y can be assumed to result from a dichotomization of an unobserved (latent) continuous variable y^* ranging from minus infinity to plus infinity. Larger values of y^* are observed as $y = 1$ while smaller values of y^* are observed as $y = 0$. Motivation for this representation often comes from studies of dose–response relationships where, for example, the response of interest is whether an individual receiving a certain dose of a toxic chemical dies.^[1,2] The latent variable y^* is assumed to be linearly related to the observed covariates through the model

$$y_i^* = x_i^T \beta + \varepsilon_i, \quad \varepsilon_i \sim \text{i.i.d. } N(0, \sigma^2) \quad (1)$$

where x_i is a vector of covariates and β is a parameter vector of fixed effects. The observed binary variable is linked to the latent continuous variable by

$$y_i = I\{y_i^* \geq \tau\} \quad (2)$$

where τ represents a threshold.

Then

$$\begin{aligned} P(y_i = 0) &= P(y_i^* < \tau) = P(x_i^T \beta + \varepsilon_i < \tau) \\ &= \Phi\left(\frac{\tau - x_i^T \beta}{\sigma}\right) \end{aligned}$$

$$P(y_i = 1) = P(y_i^* \geq \tau) = P(x_i^T \beta + \varepsilon_i \geq \tau) \\ = \Phi\left(\frac{x_i^T \beta - \tau}{\sigma}\right)$$

where Φ is the cumulative distribution function of a standard normal random variable.

Without loss of generality the threshold τ can be set equal to zero because it can be absorbed in the intercept of the linear predictor ($\beta_0^* = \beta_0 - \tau$). The error variance σ^2 is not estimable from the data because y_i^* is not observed, rather it is only known whether y_i^* is positive or negative. Usually σ^2 is taken to equal 1. Bliss^[3] is credited with the development of the method of probit analysis and the study by Finney^[1] is a useful reference of early work.

Like dichotomous data, ordinal data are also often assumed to arise from an underlying latent continuous variable. Let y_i^* be related to covariates as described in Eq. (1) and let

$$y_i = r \quad \text{if } \tau_{r-1} \leq y_i^* < \tau_r \\ \text{for } r = 1, 2, \dots, R, \tau_0 = -\infty, \\ \tau_R = +\infty, \tau_1 < \tau_2 < \dots < \tau_{R-1} \quad (3)$$

Then

$$P(y_i = r) = P(\tau_{r-1} \leq y_i^* < \tau_r) \\ = \Phi\left(\frac{\tau_r - x_i^T \beta}{\sigma}\right) - \Phi\left(\frac{\tau_{r-1} - x_i^T \beta}{\sigma}\right)$$

This ordered probit model was first suggested by Aitchison and Silvey^[4] who considered only a single independent variable. McKelvey and Zavoina^[5] later extended Aitchison and Silvey's work to the case of multiple independent variables. For identifiability reasons the variance σ^2 is set equal to 1, and either the intercept is set equal to zero, or the thresholds are reparametrized as follows: $\tau_1 = 0, \tau_2^* = \tau_2 - \tau_1, \dots, \tau_{R-1}^* = \tau_{R-1} - \tau_{R-2}$.

Probit models have no significant advantage over logit models for univariate data, but in multivariate data, the probit formulation allows to fully exploit the correlation structure of the data. By taking into account the information on covariates, multivariate probit modeling can be regarded as extension of tetrachoric and polychoric correlation.

Multivariate Probit Model for Multiple Variables or Correlated Data on the Same Variable

Extensions of the above methods to include multiple response variables or clustered data have been considered in the literature. Ashford and Sowden^[2] introduced a multivariate probit model based on an underlying multivariate normal distribution. This model is appropriate when either multiple response variables are measured on the same

individual or when there are repeated measures on the same variable within individual. In the bivariate case there are two correlated underlying latent variables such that

$$\begin{pmatrix} y_{i1}^* \\ y_{i2}^* \end{pmatrix} \sim N \left[\begin{pmatrix} x_{i1}^T \beta \\ x_{i2}^T \beta \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right], \quad (4) \\ y_{i1} = I(y_{i1}^* \geq 0) \text{ and } y_{i2} = I(y_{i2}^* \geq 0)$$

The joint probabilities for the two binary variables can be obtained based on the univariate normal cumulative distribution function Φ and on the bivariate normal cumulative distribution function $\Phi^{(2)}$

$$P(y_{i1} = 0, y_{i2} = 0) = P(y_{i1}^* < 0, y_{i2}^* < 0) \\ = \Phi^{(2)}(-x_{i1}^T \beta, -x_{i2}^T \beta)$$

$$P(y_{i1} = 0, y_{i2} = 1) = P(y_{i1}^* < 0, y_{i2}^* \geq 0) \\ = P(y_{i1}^* < 0) - P(y_{i1}^* < 0, y_{i2}^* < 0) \\ = \Phi(-x_{i1}^T \beta) - \Phi^{(2)}(-x_{i1}^T \beta, -x_{i2}^T \beta)$$

and the remaining joint probabilities are similarly determined. The marginal probabilities for each variable are

$$P(y_{i1} = 0) = \Phi(-x_{i1}^T \beta), \quad P(y_{i1} = 1) = \Phi(x_{i1}^T \beta), \\ P(y_{i2} = 0) = \Phi(-x_{i2}^T \beta), \quad P(y_{i2} = 1) = \Phi(x_{i2}^T \beta)$$

and hence do not depend on the error correlation of the variables.

Ashford and Sowden's generalization provides a full information maximum likelihood solution for the bivariate case but is computationally intractable for more than two response variables.^[6] Special algorithms have been developed for certain simple types of correlation structures. For example, Ochi and Prentice^[7] introduced a correlated generalization of the multivariate probit model of Ashford and Sowden to fit regression models to exchangeable binary data, i.e., binary data where the correlations between any two binary responses within a cluster are the same. Such situations may occur when repeated measures are taken on subjects within clusters (e.g., students within schools) and there is no particular ordering of subjects within clusters.

In contrast, or in addition to, assuming underlying multivariate normal distribution, many authors have used random effects or a smaller number of latent variables to impose correlation between quantal variables. Random effects models assume that the correlation among repeated measures arises because there is natural heterogeneity across individuals and that this heterogeneity can be represented as a probability distribution. Because of such heterogeneity, observations within subject are likely

to be more highly correlated than observations on different (unrelated) subjects. We now define a general random effects model for binary data, which is often referred to as a correlated probit model.

Correlated Probit Model for Multiple Variables or for Correlated Data on the Same Variable

Herein we assume that repeated observations occur within subject, although we could consider repeated measures within other clustering variables (e.g., school, family) rather than subject. Let y_{ij}^* be the latent response variable for the j th occasion ($j = 1, \dots, n_i$) of the i th subject ($i = 1, \dots, n$), let x_{ij} and z_{ij} be covariate vectors for the j th occasion of the i th subject, and let b_i be a random effects vector for the i th subject. Then the correlated probit model is defined as follows

$$\begin{aligned} y_{ij}^* &= x_{ij}^T \beta + z_{ij}^T b_i + \varepsilon_{ij}, \quad b_i \sim \text{i.i.d. } N(0, \Sigma_b), \\ \varepsilon_{ij} &\sim \text{i.i.d. } N(0, \sigma_\varepsilon^2) \end{aligned} \quad (5)$$

where the random effects and errors are independent. The observed binary response for each occasion within subject is $y_{ij} = I(y_{ij}^* \geq 0)$.

For this model, correlation between repeated observations on the same subject (or within the same cluster) is imposed via the random effects. Thus this model is a special case of the generalized linear mixed model with probit link for the binary response.^[8] However, it is also possible to assume that the errors within individual are correlated as in the multivariate probit model.^[4] When there is a single random intercept and the errors are not correlated, or when there are no random effects and the errors are equicorrelated, this model reduces to Ochi and Prentice's model with an equicorrelated variance–covariance structure. The correlated probit model is very flexible in modeling covariate effects and in accounting for the correlation structure of various types of data; however, there are models more appropriate for data with factor-analytic structure as described by Muthén^[9] and by Bock and Gibbons^[10] or for multilevel data as described by Hedeker and Gibbons,^[11] and by Gibbons and Hedeker.^[12] Also, the assumption of normality of the random effects can be relaxed,^[13] and the variance–covariance parameters can be assumed to depend on covariates.^[14]

Ordinal data is regarded as a generalization of binary data but often poses special challenges for analysis. Random effects regression models for ordinal regression have been considered a number of authors.^[11,15,16] The correlated probit model in Eq. (5) can be generalized to ordinal data by assuming that the observed data are related to the latent continuous variables by Eq. (3).

Correlated data can result from observations on different outcome variables within individual and/or from repeated observations on the same outcome variable

within individual. In the first case it may be more sensible to impose correlation on the error terms, while in the latter case it may make more sense to assume common subject-level random effects with or without correlated errors. Temporal or spatial data can be modeled by assuming a special structure on the errors, while clustering is usually more appropriately modeled by random effects. Some situations can be equivalently modeled by random effects or by random errors. Arguably the most complex scenarios arise when there are both multiple variables (possibly mixtures of binary, ordinal, and continuous) and repeated measures over time. These scenarios require imposing correlation at both the random effects level and at the error level and include all other situations as special cases. It is these latter scenarios that will be the focus of our discussion herein.

Because of complexity and computational difficulties with models for repeated measures on multiple binary, ordinal, and continuous outcomes, it is only in recent years that methods and software have been developed to deal with such data. Catalano and Ryan^[17] and Fitzmaurice and Laird^[18] considered models for clustered bivariate binary and continuous outcomes, and fitted a marginal model for one of the outcomes followed by a conditional model for the other. Gueorguieva and Agresti^[19] extended Chan and Kuk's^[20] probit-normal model for binary data to mixtures of binary and continuous responses and performed joint maximum likelihood estimation of the outcomes via a stochastic expectation maximization (EM) algorithm. Regan and Catalano^[14] considered a generalization of Ochi and Prentice's method for equicorrelated binary and continuous data, while Catalano^[21] and Regan and Catalano^[22] used the generalized estimating equations (GEE) approach to fit a model to a bivariate response consisting of an ordinal and a continuous variable. Geys et al.^[23] considered joint marginal mean models for developmental toxicity studies. Dunson, Chen, and Harry^[24] pioneered a Bayesian approach for jointly modeling cluster size and repeatedly measured binary and continuous outcomes, while Gueorguieva^[25] demonstrated how to perform maximum likelihood estimation on an equivalent model. Herein, we focus on the latter approach that allows for maximum likelihood estimation using existing software.

CORRELATED PROBIT MODEL FOR REPEATEDLY MEASURED COMBINATIONS OF BINARY, ORDINAL, AND CONTINUOUS RESPONSE VARIABLES

We now describe a correlated probit model that encompasses a large number of the models mentioned above as special cases. Let y_{ijv} be the v th response variable ($v = 1, \dots, V$) on occasion j ($j = 1, \dots, n_i$) for individual i ($i = 1, \dots, n$), where y_{ijv} is continuous, binary, or ordinal. Assume an underlying latent variable y_{ijv}^* for each ordinal or binary

response y_{ijv} and that the observed and the latent responses are related as in Eqs. (2) and (3) with, in general, different thresholds for each ordinal variable. The correlated probit model becomes

$$\begin{aligned} y_{ijv}^* &= x_{ijv}^T \beta + z_{ijv}^T b_i + \varepsilon_{ijv}, \\ b_i &\sim \text{i.i.d. } N(0, \Sigma_b), \\ \varepsilon_i &= (\varepsilon_{ij1}, \varepsilon_{ij2}, \dots, \varepsilon_{ijv})^T \sim \text{i.i.d. } N(0, \Sigma_\varepsilon) \end{aligned} \quad (6)$$

where there are, in general, different covariate vectors for each variable. Note that the parameters and the random effects can be shared or separate for different variables. Identifiability constraints can be imposed by setting all diagonal elements of Σ_ε corresponding to ordinal and binary responses equal to 1, setting all intercepts in the linear predictors for ordinal responses equal to 0, and setting the binary thresholds equal to 0.

RELATIONSHIP TO OTHER MODELS

Model (6) contains the models in Eqs. (1), (4), and (5) as special cases. It is also a generalization of the correlated probit model of Gueorguieva and Agresti^[19] to include ordinal responses, of the model of Catalano^[21] to include binary and continuous response and of the two-level formulation of Grilli and Rampichini's^[26] multilevel correlated probit model to include binary and continuous responses.

One can rewrite model (5) as a three-level correlated probit model by introducing occasion within subject-level random effects η_{ij} and assuming independence of the response variables conditional on subject- and occasion-level random effects. That is, the errors and the occasion-level random effects are related by the equation

$$\varepsilon_{ijv} = \gamma_v^T \eta_{ij} + e_{ijv}$$

and additional identifiability constraints on γ_v are imposed. The model formulation then becomes

$$y_{ijv}^* = x_{ijv}^T \beta + z_{ijv}^T b_i + \gamma_v^T \eta_{ij} + e_{ijv} \quad (7)$$

where $\eta_{ij} \sim \text{i.i.d. } N(0, I)$ and is independent of $e_{ij} = (e_{ij1}, e_{ij2}, \dots, e_{ijv})^T \sim \text{i.i.d. } N(0, D)$ (D —diagonal) and of b_i . This reparametrization makes the model an extension of the multivariate generalized linear mixed model^[27,28] that includes ordinal data. Model (7) belongs to the class of generalized linear latent and mixed models (GLLMM)^[29] and hence maximum likelihood estimation can be performed using the *gllamm* function in the software package Stata. However, model (7) has a larger number of random effects than the corresponding model (6) that, in general, requires more computational resources for model fitting.

INTERPRETATION

The regression parameters in the two-level correlated probit model have subject-specific interpretation. To illustrate consider the case of a single predictor variable for all outcomes, for example, say dose. Then the regression coefficients are interpreted as the expected changes in the continuous responses (or latent continuous responses) for a particular subject should dose for this subject be increased by one unit. For binary and ordinal outcomes, these are not the same as the expected changes in the responses for the population as a whole. However, because of the use of the probit link function, the marginal parameters can be obtained from the subject-specific parameters and the variance components parameters.^[30]

This latter point is one of the advantages of using probit models instead of logit models. Another advantage is that with a single random effect and only one observation per level of the random effect, the correlated probit model reduces to the ordinal probit model,^[31] while this is not true for the logit model. In addition, the probit model allows for correlated errors, which is usually an indispensable assumption when multiple variables are involved or when the data are temporally/spatially correlated, while the logit model makes such a generalization difficult.

IMPLIED CORRELATION STRUCTURE

The correlated probit models (6) and (7) are very flexible in accounting for the correlation structure of the data. The variance–covariance matrix between the continuous and latent continuous responses is implied by the underlying multivariate linear mixed model for normal data. Hence, it is very flexible in modeling both multiple variables and spatially/temporally correlated data. These models can be extended to multilevel data following the GLLMM approach and therefore can also accommodate hierarchical sampling designs. Correlations among binary variables and among ordinal variables are estimated by the maximum likelihood estimates of the correlations among the corresponding latent continuous variables and are called tetrachoric and polychoric correlations,^[32] respectively. Correlations between continuous and binary variables, and between continuous and ordinal variables are also estimated by the corresponding latent correlations and are called biserial and polyserial correlations, respectively.^[33]

ESTIMATION IN CORRELATED PROBIT MODELS

General Approaches for Maximum Likelihood Estimation

The focus in correlated probit models has mostly centered on maximum likelihood estimation. Advantages

of likelihood-based procedures include the availability of a well-established inferential framework and flexible treatment of missing data. Methods for maximum likelihood estimation of simple probit models are described by Long,^[34] among others.

Ashford and Sowden^[2] first proposed an iterative procedure for maximum likelihood estimation in bivariate probit models; however, their method was difficult to extend to higher dimensions because of the need to evaluate multivariate normal orthant probabilities.^[6] A simple but not efficient estimation for multivariate probit can be performed by first estimating the regression coefficients by univariate probit and then estimating all unknown elements of the error correlation matrix by considering pairs of responses involving only bivariate normal integrals.

Full information maximum likelihood estimation can be carried out in a direct or indirect fashion. As the marginal likelihood does not have a closed form expression, direct maximum likelihood estimation requires an approximation of the marginal likelihood (or score and information equations) and then an iterative procedure to obtain the maximum likelihood estimates of the parameters. The approximation can be carried out numerically using a method such as Gaussian quadrature^[11,25,35] or stochastically.^[36] Gaussian quadrature is preferred for lower dimensions while stochastic approximations can be used for even higher dimensions. Popular iterative procedures include Newton–Raphson^[37] and Fisher scoring.^[11] Gibbons and Hedeker^[12] extended the marginal maximum likelihood approach to a simple design with three-level data by using non-adaptive Gaussian quadrature for nested integration.

Indirect maximization is typically based on the EM algorithm because if the latent variables are observed then the correlated probit model (5) reduces to a general linear mixed model for normal data. At the E-step of the algorithm, the expectation of the likelihood of the complete data given the observed data and the current parameter estimates need to be approximated and then at the M-step, this conditional likelihood needs to be maximized to update the parameter estimates. The approximations at the E-step can be numerical (for example, Gaussian quadrature),^[16,38] stochastic (for example, Monte Carlo approximation),^[16,19,20] or analytical (for example, Taylor series expansion).^[15]

Special algorithms are available for specific types of the correlation structure for which closed form expressions of the probabilities are available.^[7,14]

Marginal Maximum Likelihood Estimation via Adaptive Gaussian Quadrature

We now provide some detail of marginal maximum likelihood estimation of the correlated probit model (6). Details about the estimation of special cases of that model are provided in Refs. [25,26,39].

By way of example, we consider here the case of two variables, where both response variables are measured at the same occasions. The marginal log-likelihood involves integration over the random effects distribution. Let Ψ denote the vector of all parameters. The marginal likelihood is $\ln L(y; \Psi) = \sum_{i=1}^n \ln L_i(y_i, \Psi)$, where

$$L_i(y_i, \Psi) = \int \prod_{j=1}^{n_i} f(y_{ij1}, y_{ij2} | b_i; \Psi) \times \frac{1}{|2\pi \Sigma_b|^{1/2}} \exp\left(-\frac{1}{2} b_i \Sigma_b^{-1} b_i\right) db_i$$

As the integral does not have a closed form expression when at least one binary or ordinal response variable is present, we need to use an approximation. When the dimension of the random effects vector is not large (up to four or five) we can use adaptive Gaussian quadrature as implemented in PROC NLMIXED in SAS.^[40] The *general* statement in PROC NLMIXED allows one to explicitly write the log-likelihood in the integrand as long as analytical expression is available.

When both response variables are binary, then

$$f(y_{ij1}, y_{ij2} | b_i) = \prod_{k=0}^1 \prod_{l=0}^1 P(y_{ij1} = k, y_{ij2} = l)^{I(y_{ij1}=k, y_{ij2}=l)}$$

where each of these normal quadrant probabilities is obtained as shown for model (4) and the parameter vector is omitted for simplicity of notation. For example

$$P(y_{ij1} = 0, y_{ij2} = 0) = P(y_{ij1}^* < 0, y_{ij2}^* < 0) = \Phi^{(2)}(-lp_{ij1}, -lp_{ij2})$$

where

$$lp_{ij1} = x_{ij1}^T \beta + z_{ij1}^T b_i \quad \text{and} \\ lp_{ij2} = x_{ij2}^T \beta + z_{ij2}^T b_i$$

When both response variables are ordinal, the conditional likelihood is similarly handled, and details are provided by Grilli and Rampichini.^[26]

When one response variable is continuous and one is ordinal (binary) we can rewrite the joint conditional density of the bivariate response given the random effects as the product of the univariate normal density of the continuous outcome $y_{ij1} | b_i$ and the conditional multinomial density of the ordinal outcome $y_{ij2} | y_{ij1}, b_i$. That is

$$\begin{aligned}
f(y_{ij1}, y_{ij2} | b_i) &= f_1(y_{ij1} | b_i) f_2(y_{ij2} | y_{ij1}, b_i) \\
f_1(y_{ij1} | b_i) &= \frac{1}{(2\pi\sigma_{\varepsilon 1}^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma_{\varepsilon 1}^2} (y_{ij1} - lp_{ij1})^2 \right\} \\
f_2(y_{ij2} | y_{ij1}, b_i) &= \Phi \left(\frac{\tau_1 - lp_{ij2}^*}{\sigma^*} \right) I(y_{ij2} = 1) \\
&\quad \left\{ \Phi \left(\frac{\tau_2 - lp_{ij2}^*}{\sigma^*} \right) - \Phi \left(\frac{\tau_1 - lp_{ij2}^*}{\sigma^*} \right) \right\}^{I(y_{ij2}=2)} \dots \\
&\quad \left\{ 1 - \Phi \left(\frac{\tau_{D-1} - lp_{ij2}^*}{\sigma^*} \right) \right\}^{I(y_{ij2}=D)}
\end{aligned}$$

where

$$\begin{aligned}
lp_{ij1} &= x_{ij1}^T \beta + z_{ij1}^T b_i, \\
lp_{ij2} &= x_{ij2}^T \beta + z_{ij2}^T b_i + \sigma_{\varepsilon 12} / \sigma_{\varepsilon 1}^2 (y_{ij1} - lp_{ij1}), \quad \text{and} \\
\sigma^* &= \sqrt{1 - \sigma_{\varepsilon 12}^2 / \sigma_{\varepsilon 1}^2}
\end{aligned}$$

In theory, this approach is extended to the case of multiple binary, ordinal, and continuous outcomes by specifying the multivariate conditional density of the observed continuous outcomes and then conditional on the continuous outcomes specifying a multinomial likelihood for all binary and ordinal outcomes, with category probabilities obtained using the multivariate normal cumulative density function. In practice, this approach works only with at most two ordinal responses, because of software limitations (SAS functions exist to calculate bivariate, but not higher order multivariate normal probabilities). The approach easily extends to modeling the variances and covariances of the random effects and/or errors as functions of the variables in the model. The three-level formulation of the model can be fitted using *gllamm* in Stata and is more flexible in handling randomly missing data.^[25,26]

Both the two-level and the three-level approaches are limited to about four-dimensional random effects and/or random error distributions because adaptive Gaussian quadrature becomes either computationally exhaustive or intractable. Alternative simulation methods based on the EM algorithm can be used for estimation.^[19]

Other Types of Estimation

Because of the complexity of full information maximum likelihood, a number of authors have considered alternative estimation methods. Muthen^[41] handled more than two response variables by using limited information generalized least squares and implemented the solution by Fisher scoring. Regan and Catalano^[22] and Geys et al.^[23] used GEE to fit correlated probit models for mixed continuous

and discrete outcomes. Spiess and Hamerle^[42] also considered GEE but supplemented with pseudoscore equations for association parameters. Qu et al.^[43,44] estimated both regression and association parameters using GEE. Spiess^[45] compared via a Monte Carlo experiment the latter three approaches and showed that the Spiess and Hamerle estimate is most efficient among non-ML estimates. Chib and Greenberg^[46] developed practical simulation-based Bayesian and non-Bayesian analysis of correlated binary data. Dunson, Chen, and Harry^[24] considered a fully Bayesian approach. Avery, Hansen, and Hotz^[47] developed quasi-maximum likelihood methods under temporal correlation and certain orthogonality conditions.

Software for fitting generalized linear mixed models can be used when all outcomes are of the same type and when the errors are assumed uncorrelated. Methods for fitting generalized linear mixed models are described by McCulloch and Searle,^[8] Diggle, Liang, and Zeger,^[48] Fahrmeir and Tutz,^[49] and Agresti.^[50]

INFERENCE

As mentioned above, one of the advantages of likelihood-based procedures include availability of well-established inferential framework. Thus, nested models can be compared using likelihood ratio tests. The significance of regression parameters and variance components not on the boundary can be tested using likelihood ratio, Wald or score tests. Testing of variance components on the boundary is more problematic because the asymptotic distribution is no longer chi-square. Comparison of models with different covariance structures can be based on information criteria, such as Akaike information criterion and Schwartz Bayesian criterion. Alternatively, if Bayesian analysis is undertaken, Bayes factors can be employed.^[46]

EXAMPLE

Developmental Toxicity Application^[51]

This was a study to examine the teratogenic effects of ethylene glycol conducted by the National Toxicology Program. During organogenesis pregnant mice were exposed to ethylene glycol at one of four different dose levels: 0, 0.75, 1.5, and 3 mg/kg. Fetal weight and a binary malformation indicator for each fetus j within litter i , and litter size s_i were recorded. It was of interest to estimate the dose effect on adverse outcomes (malformation and low fetal weight). Descriptive statistics for the developmental toxicity data are available in Table 1. These data were previously analyzed by many authors.^[14,17–19,24,25]

Although litter size is not considered an adverse outcome per se, we choose to model it herein together with fetal weight and malformation because of the potential estimation

Table 1 Descriptive Statistics for the Developmental Toxicity Study in Mice

Dose (mg/kg)	Dams	Mean litter size	Range litter size	Mean fetal weight	SD of fetal weight	Number malformed	Percent malformed
0	25	12.16	7–16	0.972	0.098	1	0.34
0.75	24	11.71	7–15	0.877	0.104	26	9.42
1.50	22	10.45	1–15	0.764	0.107	89	38.86
3.00	23	10.13	4–14	0.704	0.124	126	57.08

bias when it is ignored^[24] and because we would like to illustrate the use of the correlated probit model in a complicated design, involving a mixture of binary, ordinal, and continuous variables at different levels of nesting. We now consider a variation of the model of Gueorguieva^[25] with a correlated probit formulation for litter size. This is also the most general model, which was considered in conference paper.^[52]

$$\text{Weight}_{ij} = \alpha_1 + \beta_1 \text{dose}_i + b_{i1} + \varepsilon_{ij1}$$

$$\text{Latent Malf}_{ij} = \alpha_2 + \beta_2 \text{dose}_i + b_{i2} + \varepsilon_{ij2}$$

$$P(\text{Litter size}_i = k | b_i) = \Phi(\tau_k - \beta_3 \text{dose}_i - b_{i3}) - \Phi(\tau_{k-1} - \beta_3 \text{dose}_i - b_{i3})$$

$$\begin{pmatrix} b_{i1} \\ b_{i2} \\ b_{i3} \end{pmatrix} \sim \text{i.i.d. } N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{b1}^2 & \rho_{b12}\sigma_{b1}\sigma_{b2} & \rho_{b13}\sigma_{b1} \\ \rho_{b12}\sigma_{b1}\sigma_{b2} & \sigma_{b2}^2 & \rho_{b23}\sigma_{b2} \\ \rho_{b13}\sigma_{b1} & \rho_{b23}\sigma_{b2} & 1 \end{pmatrix} \right)$$

independent of

$$\begin{pmatrix} \varepsilon_{ij1} \\ \varepsilon_{ij2} \end{pmatrix} \sim \text{i.i.d. } N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\varepsilon1}^2 & \rho\sigma_{\varepsilon1} \\ \rho\sigma_{\varepsilon1} & 1 \end{pmatrix} \right)$$

We assume a random intercept for each variable and that these intercepts are correlated. This formulation implies an exchangeable correlation structure for each variable; that is, the correlation between fetal weight as measured on two fetuses within the same litter is constant and similarly, the correlation between latent malformation (i.e., malformation indicator) as measured on two fetuses within the same litter is also constant. We assume that there is a random intercept for litter size for convenience of modeling the correlation structure of the data.

Explicit in the model formulation are the necessary restrictions for theoretical and empirical identifiability of the model. Namely, the variance of the random effect for

litter size is assumed to be equal to 1, the error variance corresponding to the binary malformation variable is assumed to be equal to 1, and there is no intercept for the cluster level variable, litter size. Because of the correlated probit model formulation for litter size we also need to assume that the thresholds are ordered: $\tau_1 < \tau_2 < \dots < \tau_{D-1}$. Litter size ranges from 1 to 16; however, some litter sizes are not present and hence we estimate threshold parameters only for the litter sizes available in the data set.

The above model was fitted using PROC NL MIXED in SAS, and the results are presented in [Tables 2](#) and [3](#). Fetal weight and litter size decrease with dose while individual's probability of malformation increases with dose ([Table 2](#)). These findings are not surprising and, in a similar form, can be obtained by performing marginal analysis of each variable. However, marginal analyses will not give information about the correlation structure of the data and will not allow estimation of the probability of at least one adverse event (either low fetal weight or malformation), which is of considerable interest in risk assessment.

Based on the results in [Table 3](#), there is a significant positive correlation between fetal weight measured on two different fetuses within a litter and between malformation measured on two different fetuses within a litter. There was also negative correlation between fetal weight and malformation, whether measured on the same or on different fetuses within a litter, as well as between weight of a fetus within a litter and litter size. The latter result is intuitive because the more fetuses there are within litter, the more likely it is that the fetuses will be smaller. It is interesting however that the negative association between fetal weight and malformation observed in the raw data persists even after dose is controlled for in the model. Contrary to the finding concerning fetal weight there is no correlation between malformation and litter size.

Although the model we considered is quite general, there are still other potential extensions that can be considered. For example, the variances and covariances can be modeled as functions of the dose.

While the exchangeable correlation assumption does not appear restrictive in this data example with litter as the clustering variable, it is usually not appropriate for longitudinal data where random slopes are commonly assumed in addition to random intercepts and where special structures of the error variance–covariance matrix (e.g., autocorrelated) may be appropriate.

Table 2 Maximum Likelihood Estimates for the Parameters in the Developmental Toxicity Study in Mice

Response variable	Parameter	Maximum likelihood estimate (SE)
Fetal weight	Intercept (α_1)	0.944 (0.015)
	Dose (β_1)	−0.087 (0.009)
	SD ^a of random intercept (σ_{b1})	0.089 (0.007)
	SD ^a of random error (σ_{e1})	0.095 (0.002)
Malformation	Intercept (α_2)	−2.307 (0.198)
	Dose (β_2)	0.915 (0.102)
	SD ^a of random intercept (σ_{b2})	0.779 (0.098)
	SD ^a of random error (σ_{e2})	1.00 (0.00) ^b
Litter size	Dose (β_3)	−0.384 (0.136)
	SD ^a of random intercept (σ_{b3})	1.00 (0.00) ^b

^aSD denotes standard deviation.

^bThese parameters are fixed at the value of 1 for identifiability.

Table 3 Maximum Likelihood Estimates for Correlations between Response Variables in the Developmental Toxicity Study in Mice

Level	Variable 1	Variable 2	Maximum likelihood estimate (SE)
Within <i>i</i> th litter	Fetal weight on <i>j</i> th fetus	Fetal weight on <i>k</i> th fetus	0.47 (0.04)
	Malformation on <i>j</i> th fetus	Malformation on <i>k</i> th fetus	0.38 (0.06)
	Fetal weight on <i>j</i> th fetus	Malformation on <i>k</i> th fetus	−0.27 (0.05)
	Fetal weight on <i>j</i> th fetus	Malformation on <i>j</i> th fetus	−0.29 (0.06)
Between <i>i</i> th litter size and variables measured in <i>i</i> th litter	Size of <i>i</i> th litter	Fetal weight on <i>j</i> th fetus	−0.21 (0.07)
	Size of <i>i</i> th litter	Malformation on <i>j</i> th fetus	−0.04 (0.09)

SOFTWARE

In the early 1990s, Lessafre and Molenberghs^[32] observed that multivariate probit analysis was a neglected procedure in medical statistics because of the lack of software to perform such analysis. Therefore, they created an MS-DOS program for variable selection and model fitting of bivariate probit models. Owing to the recent computational advances, there are now several programs that can be used to fit more sophisticated probit models in higher dimensions. We already demonstrated how the popular SAS PROC NLMIXED can be used for a complicated example. The *gllamm* function in Stata^[29,53] can be used for the same model, and unlike PROC NLMIXED it can handle multilevel models. Preisler^[54] performs marginal maximum likelihood using Gaussian quadrature employing GLIM, while MIXOR was especially created for random effects analysis of repeatedly measured ordinal outcomes. Gibbons^[55] provides a nice description of the MIXOR implementation and lists additional software for binary, ordinal, and continuous data.

CONCLUSIONS

In this entry we presented a description of correlated probit models and demonstrated how such models can be fitted using standard software. Historically, probit models have not been as popular as logit models mainly because of the lack of odds-ratio interpretation and the computational difficulties in obtaining estimates for such models. However, probit models have appealing interpretation in the context of an underlying continuous response, and the computational problems have been minimized in recent years. Unlike logit models, probit models seamlessly allow for specification of correlated errors, which is an important and often indispensable assumption for spatial or temporal data. In addition, there is a direct correspondence between multivariate and univariate probit models, correlation structure modeling is flexible, and there is a convenient relationship between subject-specific and marginal model parameters. Hence, the correlated probit model has undergone a dramatic change from a “neglected procedure in medical statistics” into an accepted and useful model for correlated data.

ACKNOWLEDGMENT

I wish to thank Mr. Brian Pittman for his helpful comments.

REFERENCES

1. Finney, D.J. *Probit Analysis*; Cambridge University Press: Cambridge, 1964.
2. Ashford, J.R.; Sowden, R.R. Multi-variate probit analysis. *Biometrics* **1970**, *26*, 535–546.
3. Bliss, C.I. The method of probits. *Science* **1934**, *79*, 38–39.
4. Aitchison, J.; Silvey, S.D. The generalization of probit analysis to the case of multiple responses. *Biometrika* **1957**, *44*, 131–140.
5. McKelvey, R.D.; Zavoina, W. A statistical model for the analysis of ordinal level dependent variables. *J. Math. Sociol.* **1975**, *4*, 103–120.
6. Kiefer, N.M. Multivariate probit. *Encyclopedia of Statistical Sciences*; John Wiley and Sons: New York, 1985, Vol. 6, 108–110.
7. Ochi, Y.; Prentice, R.L. Likelihood inference in a correlated probit regression model. *Biometrika* **1984**, *71*, 531–543.
8. McCulloch, C.E.; Searle, S.R. *Generalized, Linear, and Mixed Model*; John Wiley & Sons: New York, 2001, 220–246.
9. Muthén, B.O. A structural probit model with latent variables. *J. Am. Stat. Assoc.* **1979**, *74*, 807–811.
10. Bock, R.D.; Gibbons, R.D. High dimensional multivariate probit analysis. *Biometrics* **1996**, *52*, 1183–1194.
11. Hedeker, D.; Gibbons, R.D. A random-effects ordinal regression model for multilevel analysis. *Biometrics* **1994**, *50*, 933–944.
12. Gibbons, R.D.; Hedeker, D. Random effects probit and logistic regression models for 3-level data. *Biometrics* **1997**, *53*, 1527–1537.
13. Gibbons, R.D.; Lavigne, L.V. Emergence of childhood psychiatric disorders: a multivariate probit analysis. *Stat. Med.* **1998**, *17*, 2487–2499.
14. Regan, M.M.; Catalano, P.J. Likelihood models for clustered binary and continuous outcomes: application to developmental toxicology. *Biometrics* **1999**, *55*, 760–768.
15. Harville, D.A.; Mee, R.W. A mixed model procedure for analyzing ordered categorical data. *Biometrics* **1984**, *40*, 393–408.
16. Tutz, G.; Hennevogel, W. Random effects in ordinal regression models. *Comput. Stat. Data Anal.* **1996**, *22*, 537–557.
17. Catalano, P.J.; Ryan, L.M. Bivariate latent variable models for clustered discrete and continuous outcomes. *J. Am. Stat. Assoc.* **1992**, *87*, 651–658.
18. Fitzmaurice, G.M.; Laird, N.M. Regression models for a bivariate discrete and continuous outcome with clustering. *J. Am. Stat. Assoc.* **1995**, *90*, 845–852.
19. Gueorguieva, R.V.; Agresti, A. A correlated probit model for joint modeling of clustered binary and continuous responses. *J. Am. Stat. Assoc.* **2001**, *96*, 1102–1112.
20. Chan, J.S.K.; Kuk, A.Y.C. Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics* **1997**, *53*, 86–97.
21. Catalano, P.J. Bivariate modeling of clustered continuous and ordered categorical outcomes. *Stat. Med.* **1994**, *16*, 883–900.
22. Regan, M.M.; Catalano, P.J. Bivariate dose-response modeling and risk estimation in developmental toxicity. *J. Agric. Biol. Environ. Stat.* **1999**, *43*, 217–237.
23. Geys, H.; Regan, M.M.; Catalano, P.J.; Molenberghs, G. Two latent variable risk assessment approaches for mixed continuous and discrete outcomes from developmental toxicity data. *J. Agric. Biol. Environ. Stat.* **2001**, *6*, 340–355.
24. Dunson, D.; Chen, B.; Harry, J. A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* **2003**, *59*, 521–530.
25. Gueorguieva, R.V. Comments about joint modeling of cluster size and binary and continuous subunit-specific outcomes. *Biometrics* **2005**, *61* (3), 862–867.
26. Grilli, L.; Rampichini, C. Alternative specifications of multivariate multilevel probit ordinal response models. *J. Educ. Behav. Stat.* **2003**, *28*, 31–44.
27. Gueorguieva, R.V. A multivariate generalized linear mixed model for joint modeling of clustered outcomes in the exponential family. *Stat. Model.: Int. J.* **2001**, *1*, 177–194.
28. Dunson, D.B. Bayesian latent variable models for clustered mixed outcomes. *J. Royal Stat. Soc.: Ser. B* **2000**, *62*, 355–366.
29. Rabe-Hesketh, S.; Pickles, A.; Skrondal, A. GLLAMM: a class of models and a stata program. *Multilevel Modell. Newsl.* **2001**, *13*, 17–23.
30. Zeger, S.L.; Liang, K.-Y.; Albert, P.S. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **1988**, *44*, 1049–1060.
31. McCulloch, C.E. Maximum likelihood variance components estimation for binary data. *J. Am. Stat. Assoc.* **1994**, *89*, 330–335.
32. Lesaffre, E.; Molenberghs, G. Multivariate probit analysis: a neglected procedure in medical statistics. *Stat. Med.* **1991**, *10*, 1391–1403.
33. Drasgow, F. Polychoric and polyserial correlations. In *Encyclopedia of Statistical Sciences*; Kotz, L.; Johnson, N.L., Eds.; Wiley: New York, 1988, Vol. 7, 69–74.
34. Long, S. *Regression Models for Categorical and Limited Dependent Variables*; Sage Publications: London, 1997.
35. Gibbons, R.D.; Bock, R.D. Trends in correlated proportions. *Psychometrika* **1987**, *52*, 113–124.
36. Vijverberg, W.P.M. Monte Carlo evaluation of multivariate normal probabilities. *J. Econ.* **1997**, *76*, 281–307.
37. Kim, K. A bivariate cumulative probit regression model for ordered categorical data. *Stat. Med.* **1995**, *14*, 1341–1352.
38. Jansen, J. On the statistical analysis of ordinal data when extravariation is present. *Appl. Stat.* **1990**, *39*, 75–84.
39. Gueorguieva, R.V.; Sanacora, G. A latent variable model for joint analysis of repeatedly measured ordinal and continuous outcomes. In *Proceedings of the 18th International Workshop on Statistical Modelling*, 2003; Verbeke, G.; Molenberghs, G.; Aerts, M.; Fiews, S., Eds.; Katholieke Universiteit Leuven: Leuven, 2003; 171–176.
40. Wolfinger, R. Fitting nonlinear mixed models with the new NLMIXED procedure. 24th SAS Users Group International Conference Proceedings 1999, Miami Beach, FL, Paper 287.

41. Muthen, B.O. *LISCOMP: Analysis of Linear Structural Equations with a Comprehensive Measurement Model*; Scientific Software International: Chicago, IL, 1988.
42. Spiess, M.; Hamerle, A. On the properties of GEE estimators in the presence of invariant covariates. *Biomet. J.* **1996**, *388*, 931–940.
43. Qu, Y.; Williams, G.W.; Beck, G.J.; Medendorp, S.V. Latent variable models for clustered dichotomous data with multiple subclusters. *Biometrics* **1992**, *48*, 1095–1102.
44. Qu, Y.; Piedmonte, M.R.; Williams, G.W. Small sample validity of latent variable models for correlated binary data. *Comm. Stat. Simul. Comput.* **1994**, *23* (1), 243–269.
45. Spiess, M. A mixed approach for the estimation of probit models with correlated responses: some finite sample results. *J. Stat. Comput. Simul.* **1998**, *61*, 39–59.
46. Chib, S.; Greenberg, E. Analysis of multivariate probit models. *Biometrika* **1998**, *85* (2), 347–361.
47. Avery, R.B.; Hansen, L.P.; Hotz, V.J. Multiperiod probit models and orthogonality condition estimation. *Int. Econ. Rev.* **1983**, *24*, 21–35.
48. Diggle, P.J.; Liang, K.Y.; Zeger, S.L. *Analysis of Longitudinal Data*; Oxford Science Publications; Clarendon Press: Oxford, 1994, 169–189.
49. Fahrmeir, L.; Tutz, G. *Multivariate Statistical Modelling Based on Generalized Linear Models*; Springer-Verlag: New York, 2001, 283–330.
50. Agresti, A. *Categorical Data Analysis*; Wiley & Sons: Hoboken, NJ, 2002, 491–537.
51. Price, C.J.; Kimmel, C.A.; Tyl, R.W.; Marr, M.C. The developmental toxicity of ethylene glycol in rats and mice. *Toxicol. Appl. Pharmacol.* **1985**, *81*, 825–839.
52. Gueorguieva, R.V. Joint modelling of cluster size and binary and continuous outcomes. In *Proceedings of the 19th International Workshop on Statistical Modelling*; Biggeri, A.; Dreassi, E.; Lagazio, C.; Marchi, M., Eds.; Firenze University Press: Florence, 2004, 199–203.
53. Rabe-Hesketh, S.; Skrondal, A.; Pickles, A. Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata J.* **2002**, *2*, 1–21.
54. Preisler, H.K. Maximum likelihood estimates for binary data with random effects. *Biomet. J.* **1998**, *30* (3), 339–350.
55. Gibbons, R.D. Mixed-effects models for mental health services research. *Health Serv. Outcomes Res. Methodol.* **2000**, *1* (2), 91–129.

Cost-Effectiveness Analysis

Heejung Bang

*Division of Biostatistics, Department of Public Health Sciences,
University of California, Davis, California, U.S.A.*

Abstract

Effectiveness and cost are the key components when comparing health-care options and making treatment decisions. Cost-effectiveness analysis (CEA) is the methodological approach that informs this evaluation. In this entry, we review the basic concepts and statistical methods, which are the foundations of CEA, and discuss quantitative and qualitative issues, which are quite unique to CEA.

Keywords: Cost-effectiveness; Clinical trial; Health outcome; QALY; Willingness to pay.

INTRODUCTION

It was an expensive way to live, so I decided to look for cheaper lodgings. But how do you find comfortable rooms at a reasonable price?

—Dr. Watson (*The Adventures of Sherlock Holmes*)

The rising economic burden of health care is a significant concern across the globe, particularly in countries with aging populations and inflated costs associated with the development, testing, marketing, and distribution of the new therapeutic and technological interventions. Most societies seek to provide the best or affordable health care to their citizens, often with very limited resources.

Cost had played an important role even in the presumably first controlled clinical trial of the modern era. Although the effect of oranges and lemons was apparent on the cure of scurvy, James Lind (1716–1794) hesitated to recommend them because they were too expensive. Nearly 50 years after, the British Navy eventually made lemon juice a part of the diet and was replaced by lime juice because it was cheaper.

Many interventions/regimens that are widely used in usual care or daily practice e.g., vitamin C, aspirin, chlorination of water, antibiotics, birth control pill, insulin, antiretroviral therapy, etc., produce large treatment effects. However, such discoveries are rare, so that research results often point to innovations with smaller effects or encourage the use of existing interventions for different target populations or goals, such as personalized medicine to identify an optimal therapy, dose or subgroup, or off-label use. The reality is that it is a long and expensive process to develop a new intervention (drug, treatment, technology, program, and policy) that is

significantly better than standard care, and this impacts greatly on cost relative to perceived health benefit. A notable example is a cancer drug that extends survival by 1.2 months at a cost of \$80,000.^[1] Efforts are made not only to develop interventions that could be comparable to standard care in terms of efficacy or cost but also to convey other advantages in terms of accessibility, tolerability, or convenience.

In the quest to determine which intervention produces the greatest value, cost is a commonly adopted outcome, secondary to the primary clinical outcome. Cost is a factor in virtually all comparative outcomes and decisions. Cost-effectiveness analysis (CEA) compares the relative expenditures and outcomes of competing strategies, with a goal to identify which provides the maximum aggregate health benefit for a given level of resources—or, equivalently, which provides a given level of health benefit at the lowest cost.

In this entry, we review the statistical methods used in CEA and discuss related concepts, quantitative and qualitative issues, and controversies that are somewhat unique to CEA. A previous version of this entry provides a general overview from economics perspectives;^[2] this newer version focuses on statistical methods and concepts that may be useful to practitioners and statisticians who may not specialize in health economics. Various journals, websites, and textbooks provide guidelines on the conduct and presentation of CEA. However, unlike effectiveness analysis (EA) in which consensus on methodology and implementation has been reasonably well established (enabling regulatory agencies, pharmaceutical companies, clinicians, and researchers to learn and use the same or well-agreed methods and data, while permitting reasonable flexibility), there are additional challenges and nuances associated with CEA.

STATISTICAL MEASURES

Let C_i and E_i denote cost and effectiveness for i th group, respectively, $i = 1$ (new), 2 (standard or control), where two interventions are compared head-to-head:

- If $E_1 > E_2$ and $C_1 < C_2$, new is more effective and less costly (dominate; accept)
- If $E_1 < E_2$ and $C_1 > C_2$, new is less effective and more costly (dominate; reject)
- If $E_1 > E_2$ and $C_1 > C_2$, new is more effective and more costly (trade-off)
- If $E_1 < E_2$ and $C_1 < C_2$, new is less effective and less costly (trade-off)
- If $E_1 \approx E_2$, compare C_1 and C_2 (cost minimization)
- Compare E_1 and E_2 (EA)

The scenarios above can be visualized in a cost-effectiveness (CE) plane, and CEA is particularly relevant when a new intervention is more effective and more costly. The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Task Force (2005) recommended that CEA be performed even when clinical effectiveness fails to be demonstrated. This is an ideal recommendation and indeed $E_1 \approx E_2$ is the goal of equivalence testing (e.g., generic drug); yet, some mathematical challenges could be entailed, as described below.

CE Plane

The CE plane is a simple, intuitive graphical tool that presents the difference in effectiveness (in X-axis) and the difference in cost (in Y-axis) between two interventions in the four quadrants (see Fig. 1a).^[4] The CE plane is essential for any CEA to prevent numerical or interpretational error (e.g., 2/2 vs. -2/-2), similar to descriptive statistics as a prerequisite for any data analysis. The CE

plane can be used with a point estimate of the incremental cost-effectiveness ratio (ICER), along with bootstrapped versions that can characterize its sampling variability and confidence interval (CI). Furthermore, the five decision regions in the CE plane may be used as a reference figure or yardstick to assist the final judgement. (see Fig. 1b). Of note is that the parameter and estimator in notation are not distinguished.

ICER

The most widely adopted measure in CEA is the ICER. Trade-off through “increment” (relative, marginal, or net) denoted by Δ is a key concept and parameter in economics. The ICER is defined as $\Delta C / \Delta E$. It is interpreted as additional cost needed to improve effectiveness (of a continuous variable or count) by 1 unit and conventionally estimated by:

$$\text{ICER}_{\text{mean}} = \frac{\text{mean}(C_1) - \text{mean}(C_2)}{\text{mean}(E_1) - \text{mean}(E_2)} \quad (1)$$

If effectiveness is binary (e.g., cure or no cure), its expected value or mean becomes the probability. The ICER also is resulted from a maximization problem with the Lagrange multiplier.

After computing the ICER, its value could be compared with the willingness to pay (WTP) for a 1-unit increment in E, denoted by λ (lambda), which serves as the ICER threshold. A sample decision is if $\text{ICER} < \lambda$, then the new intervention is cost-effective, not otherwise. More discussions on λ are provided below. The ICER can be interpreted as slope from (0,0) to (x,y) on the CE plane. Some analysts create a “league table” using the ICERs for ranking purposes.^[5]

Some measures utilize specialized names but can be understood within the ICER framework. For example,

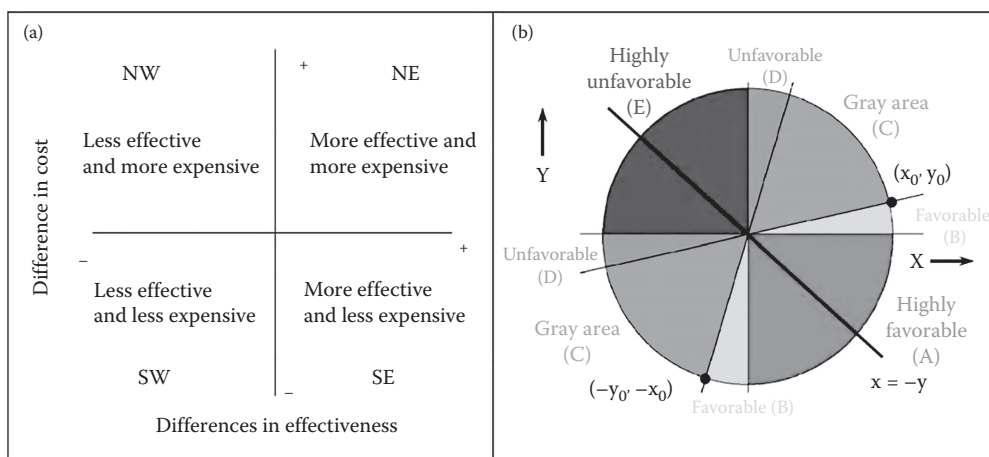


Fig. 1 Cost-effectiveness plane (a: left) with five decision regions (b: right)

Source: (b) Reproduced from Obenchain.^[3] ©2008 by permission of the author and copyright holder.

“cost per case prevented” is the difference in cost between the two interventions divided by the associated difference in the number of cases of a bad event.

Incremental Net Benefit (INB)

Despite wide acceptance and popularity, the ICER has some methodological weaknesses. First, ICER as a ratio statistic of the two differences can be $\pm$ infinity when the denominator (=difference in effectiveness) is 0, in which case CI construction could be troublesome. Also, the two totally opposite results (e.g., 2/2 vs. $-2/-2$) can have the same ICER, and negative ICERs are difficult to interpret and may not be properly ordered.^[6]

An alternative way to view the ICER is that overall outcome pairs are standardized by expressing both of their components in the same units. With the help of λ , the ICER threshold, the ICER can be expressed as $(x,y) = (\lambda\Delta E, \Delta C)$ or $(\Delta E, \Delta C/\lambda)$, naturally connecting the ICER can be expressed with the notion of (net) benefit (see sample calculations of different measures in Table 1). $INB = \lambda\Delta E - \Delta C$ overcomes the ICER’s limitations and puts E and C in the additive/linear scale, conveying improved statistical properties. As is clear in the formula, INB is a function of λ , a generally unknown constant, which is desired but not required when we use ICER. If $INB > 0$, we conclude that the net benefit of an intervention exceeds the cost or that the new treatment is cost-effective; virtually, the same rule is used for ICER and INB. INB (in Y-axis) as a function of λ (in X-axis) can be drawn in a figure—along with CI or band—over plausible values of λ .

CE Acceptability Curve

Another variant of ICER or INB is cost-effectiveness acceptability curve (CEAC). A plot can be drawn with X-axis = λ and Y-axis = $P(INB > 0)$ across a range of λ , and it can show the estimated probabilities that the new treatment is cost-effective for different values of WTP.^[7]

Average Cost-Effectiveness Ratio (ACER)

A simple and intuitive measure that has been used in cost analysis and CEA is the ACER:

$$ACER_i = \frac{\text{mean}(C_i)}{\text{mean}(E_i)} \quad (2)$$

Some interpret ACER as ICER when the control group offers zero effect and zero cost. Not using “relative” measure leads some to reject the use of the ACER or to limit its use to restrictive or simplistic CEA contexts (e.g., non-competing choices).^[2,8–10]

Some properties of ACER might be appealing from statistical perspectives as exploratory or supporting measures. For instance, ACER is well defined for one or more groups and accords with the traditional framework of estimation and testing for single or multiple groups, whereas ICER requires a comparator (and unknown threshold λ). The Neyman–Pearson hypothesis testing paradigm and its role in efficient allocation provide some theoretical support for ACER.^[11] Also, the ACER’s denominator is further from zero, so its point and CI estimates tend to be numerically stable.

Various measures based on the notion of “cost per case (or patient)” might be viewed within the ACER framework. For example, “cost per (usable) response” is used in survey research and “cost per patient in care” is used in HIV care.^[12,13]

DESIGN, METHODS, AND STATISTICS-RELATED ISSUES

Types of CEA; What is in the Denominator?

The basic steps in any CEA are to identify, measure, value, and compare costs. How to define cost(s) for a single person (raw or individual patient data) is the first step in any cost-related analysis. However, *do we really know what costs how much?* Inherently, the complete enumeration of each cost component and identification of a correct/accurate cost are complicated and quite subjective processes. Oftentimes, the cost for the i^{th} patient, C_i , must be estimated with high uncertainty, unlike E_i , which can be measured in a more objective or standardized fashion (e.g., survival time, response, blood pressure, and depression) in the EA. Inclusions/exclusions of various cost contributors

Table 1 Sample calculations of cost-effectiveness ratios and net benefits

Treatment alternative	Cost	Effect (anxiety-free weeks)	Cost-effectiveness ratio	Monetary value of the effect (or benefit)	Net benefit
Standard	\$1,000	40	ACER = \$25	$\lambda E = 20,000$	NB = 19,000
New	\$2,000	50	ACER = \$40	$\lambda E = 25,000$	NB = 23,000
Difference (Δ)	\$1,000	10	ICER = \$100	$\lambda\Delta E = \$5,000$	INB = 4,000

Note: ACER = average cost-effectiveness ratio; ICER = incremental cost-effectiveness ratio; INB = incremental net benefit; NB = net benefit; λE = the monetary value of effect is benefit ($B \equiv \lambda E$). (Remark: Outcome in natural units = effect, and monetary value = benefit = effect (in units) $\times$ value per unit.) Here, we have artificially assumed that societal willingness to pay for patient outcome (λ) equals \$500 per anxiety-free week. NB = monetary value of effect – cost; INB is calculated with an assumed value of $\lambda = \$500$; effect difference $\times$ \$500 – cost difference.

Source: Adapted from Hoch and Dewa.^[7] ©2008.

and time period substantially impact on the final estimate and analysis of costs. How to determine cost is not really a statistical question; sources and methods to be used (such as, reimbursement, cost-to-charge ratio, consumer price index, microcosting, top-down vs. bottom-up, etc.) are in the domains of economics and finance. Of note is that the cost estimates can be highly dependent on the method used.

Once a patient-level total cost—assumed to be independent and identically distributed over different individuals, while (longitudinal/intermediate) costs within individual are correlated—is derived, the next step is to estimate the population summary measure. A fundamental statistical parameter is $E(C)$, representing the expected value or mean of a variable C_i , where C denotes cost and i indexes individuals here. Oftentimes, $E(C)$ per se is of primary interest, for instance to estimate the cost of colorectal cancer. In the CEA, the arithmetic mean of costs in the untransformed/original scale is the primary parameter.

Within the same ICER formula, we use different names depending on what constitutes the denominator. To list, “cost minimization” is simply to compare the costs in two groups, ignoring the denominator (effectiveness), particularly when the denominator is close to 0. [Of note is, the seamless use of ICER with the CE plane in a unified framework, whether the denominator is 0 or not (which is not always apparent to determine), has been proposed.^[14]] When we use clinical outcomes such as survival time or life years (LY) for the effectiveness measure, it is called the CEA. Alternatively, if we use quality-adjusted life year (QALY) in place of LY, then a more specialized term will be the cost-utility analysis (CUA). As QALY is regarded as the most widely recommended measure of effectiveness, many use CEA and CUA interchangeably. QALY seeks to combine multidimensional consequences of health status into a one-dimensional metric with health utilities—a weight of 0 for death and 1 for perfect health. This approach is especially useful for interventions that extend life but at the expense of side effects or lower quality or that reduce morbidity instead of mortality. When the effectiveness is converted into monetary values, it is called the cost-benefit analysis (CBA). This analysis allows the numerator and denominator to take same scale/unit, so that health-related programs and policies may be more easily compared. CBA provides a clear decision rule—undertake an intervention if the monetary value of its benefit exceeds its cost. However, due to a general reluctance to characterize health benefits in monetary terms, CUA and CEA are more popular.

Study Design: Experiment Trumps Observation?

A (well-designed and conducted) randomized controlled trial (RCT) is scientifically the best study design for EA in practice, wherein a “head-to-head” on average comparison can be achieved. However, whether this is also the gold standard for CEA is questionable. The RCT design aims to

compare A versus B for a relatively short term on a homogeneous population, generally involving lengthy inclusion/exclusion criteria and protocol-driven tests for research purposes. Thus, generalizability of the findings to the real-world or usual practice in a broader population may be limited. While RCTs can minimize confounding and selection biases, it may not be superior in other aspects, such as valid and meaningful cost estimates. Also, new drug acquisition price is *unknown* at the clinical trial stage. Observational evidence (e.g., obtained from registry, medical records and expert opinions on actual medical practice) is inevitable in CEA, often with heavily assumption-driven modeling.

Skewness

As cost data tend to be highly skewed, so is the ICER. Thus, in addition to the mean, which is the norm, other summary measures and non-parametric approaches (e.g., median, quartiles, log-transformation, and bootstrap instead of Fieller for a CI) have been used. Consideration related to data distribution should be given. When cost data are log-transformed, it has been reported that an incorrect hypothesis is frequently used.^[15] Standard non-parametric methods, analyses of transformed costs and measures other than the arithmetic mean, are generally deemed inappropriate in CEA because the cost of treating all patients is needed as the basis for health-care policy decisions.^[16] Notwithstanding, the use of nonparametric methods for skewed or non-normal data may be practically unavoidable due to their dominance; e.g., the Wilcoxon test instead of (standard or bootstrap) t-test for cost comparison. Supporting roles of median/quantile-based methods for cost data also have been suggested; for instance, mean versus median-based analyses can yield meaningfully different results, in which case a conclusion may not be readily available, e.g., more evidence is needed.^[14,17]

In addition, median survival time has been regarded as a preferred outcome measure in the EA and statistics literature (due to mathematical issues), as long as it is estimable with time-to-event or survival data, say, when the Kaplan–Meier curve reaches 0.5.^[18] Thus, there are discrepancies in key parameters in EA versus CEA—median survival time for EA and mean survival time or QALY for CEA.^[19]

Discounting

Cost is meaningful only with an accompanying reference year (and location). When we deal with costs that accumulate across time, conversion is generally utilized. Discounting is justified because people tend to prefer dollars today versus dollars in the future (even with zero inflation). CEA usually incorporates this time preference that includes dollars over more than 1 year. Annual discount rates of 3% and 5% are the most popular choices for cost, while other rates are considered in sensitivity analyses. Similar to cost, people tend to want current health more

than future health. Discounting in health effect and the inflation in cost (which is distinct from discounting) are less often agreed upon.^[5,20,21]

Censoring

A common effectiveness measure in prospective health studies is time-to-event, e.g., survival time/LY or QALY. When survival time is censored, i.e., the patient did not die by the end of observation or follow-up period, the associated cost is also censored such that the observed cost is merely a lower limit of the true cost. Censoring complicates statistical analyses and should be adjusted. Indeed, censoring is routinely adjusted in EA in practice, which is why we use the Kaplan–Meier estimate or the Cox model. Moreover, censoring could influence a parameter’s identifiability or estimability and may determine which parameter should be primary. For example, the mean cost and mean survival time, the two key components in the CEA, are not always estimable with censored data. As a remedy, the *restricted* mean is often used and its theoretical issues have been studied in the statistical literature.^[22] The underlying mechanism calling for special attention is that, even when censoring is uninformative (or random) for survival time (which is assumed in most EA), censoring is *informative* for cost, due to the positive correlation between cost at event of interest and cost at censoring. Interestingly, this informative, induced censoring also happens to QALY. As such, cost and QALY may be mathematically more similar than are LY and QALY in some aspects. Thus, most standard survival analyses (e.g., Kaplan–Meier) are valid for censored survival data but not for censored cost or QALY

data. Various mean cost and ICER estimators with censoring have been proposed.^[22,23]

Modeling and Sensitivity Analyses

It was previously stated that RCT may not be the gold standard for CEA. Moreover, individual patient data for long term are often unavailable. In these circumstances, an alternative approach is decision analytic modeling, especially when we are interested in lifetime cost. Decision modeling compares the expected costs and consequences of decision options by synthesizing information from multiple data sources and applying mathematical techniques, usually with computer software.^[24] Modeling (e.g., Markov model, decision tree, and simulation) can be used where inputs are provided from best available sources, including concurrent or existing data and literature (see Fig. 2).^[20,24] Careful design and planning, along with a protocol documenting data sources, methods, assumptions, and sensitivity analysis (including substantially different perspectives, short vs. long term, and best vs. worst case scenarios), are critical.

Models and results should be sensible, transparent, reproducible, and well documented. Modeling results are only as good as: 1) the data inputs provided to the model; 2) the structure of the model; and 3) correct implementation. It has been demonstrated that different models, such as new vs. old (standard) model with one outdated component (i.e., assumption 9), can produce the estimates with a 10 times difference in a nutritional modeling and prediction.^[25] Wrong assumptions and investigator biases (e.g., protocol not fully prospective, unblinded) are big

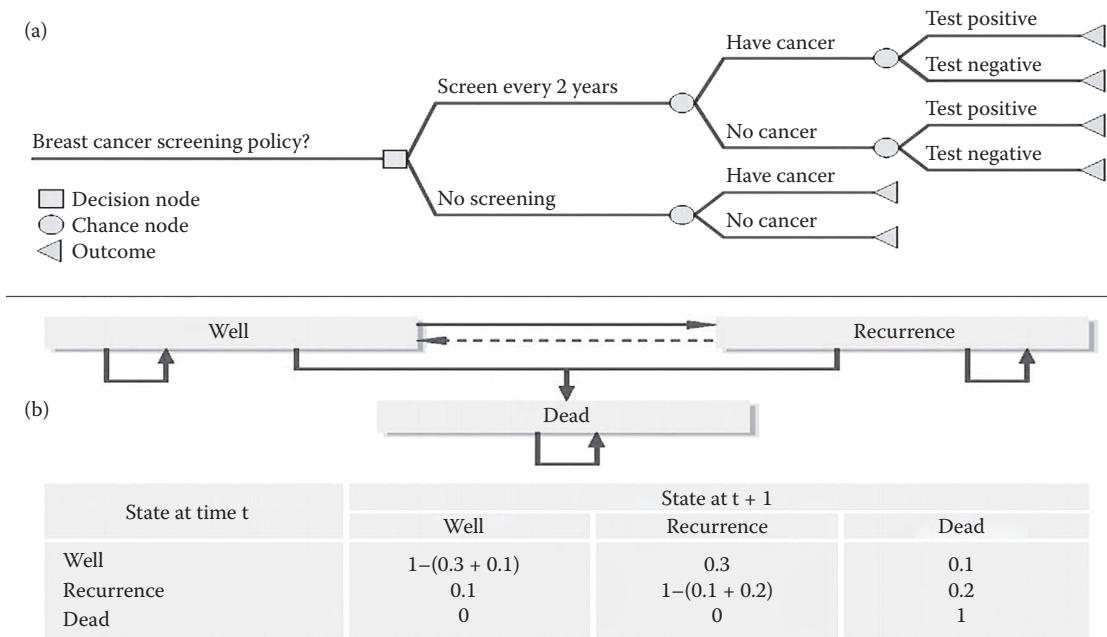


Fig. 2 Decision tree (a: upper) and Markov diagram (b: lower)
Source: Reproduced from Petrou & Gray.^[24] ©2011 BMJ Publishing Group, Ltd.

threats in modeling, although the choices and mistakes made can be honest.

An economic study from a cardiovascular RCT presented actual/estimated ICERs and projected ICERs over time—actual follow-up of the first 3.5 years and projecting it for 12 years under three hypothetical scenarios (see Fig. 3).^[26] This kind of plot may serve as a good example in CEA, demonstrating that trial-based (utilizing individual patient data) and modeling-based economic evaluations can be complementary.

Budget Impact Analysis as Companion

Budget impact analysis (BIA) may be required by reimbursement authorities, along with a CEA, as part of a listing or reimbursement submission or freestanding economic evaluations for budget or resource planning. A BIA addresses the expected changes in the expenditure of a health-care system before versus after the adoption of a new intervention. The ISPOR Task Force provided a reporting guide on BIA in 2007 and an update in 2014. Similar methodological issues covered in this entry are applied to BIA and CEA, including perspectives, types of costs, time horizon, discounting, input and data sources, uncertainty, and validation.

Sample Size

Every study needs sample size (N) and statistically supported sample size is always preferred whenever possible. Naturally, calculation of N for economic evaluation shares much in common with calculation of N for efficacy/

effectiveness evaluation, and a sizable number of publications and formulae are available to guide N adequate for a given CEA (see Glick for sample formulae and literature review).^[27] A goal of N/power calculation for CEA can be to identify the likelihood that an experiment will allow us to be confident that a therapy has a good or bad value when we adopt a particular WTP.

However, it is easily visible that the complexity associated with sampling uncertainty for economic assessments (e.g., standard deviation of cost) and methodological issues and choices (e.g., cost comparison only vs. ICER vs. INB, which threshold, hypothesis testing vs. estimation) can complicate these N calculations. It is reasonable to expect to assume N required for CEA is generally larger, more vague (thus, justifying it as “rule of thumb”), compared to the EA counterpart. Oftentimes, trialists do not have a sufficient power/N for a primary hypothesis testing in the EA, which can render CEA more underpowered. Moreover, it may not be ethical to accrue a larger number of patients for a well-powered CEA, which is generally a secondary aim in clinical trials. These issues further justify the pragmatic necessity of (very) large databases for economic studies, outside or beyond clinical trials, despite other limitations.

CONTROVERSIES

CEA or Not; QALY or Not?

The Patient-Centered Outcomes Research Institute (PCORI) was established by the 2010 Patient Protection and Affordable

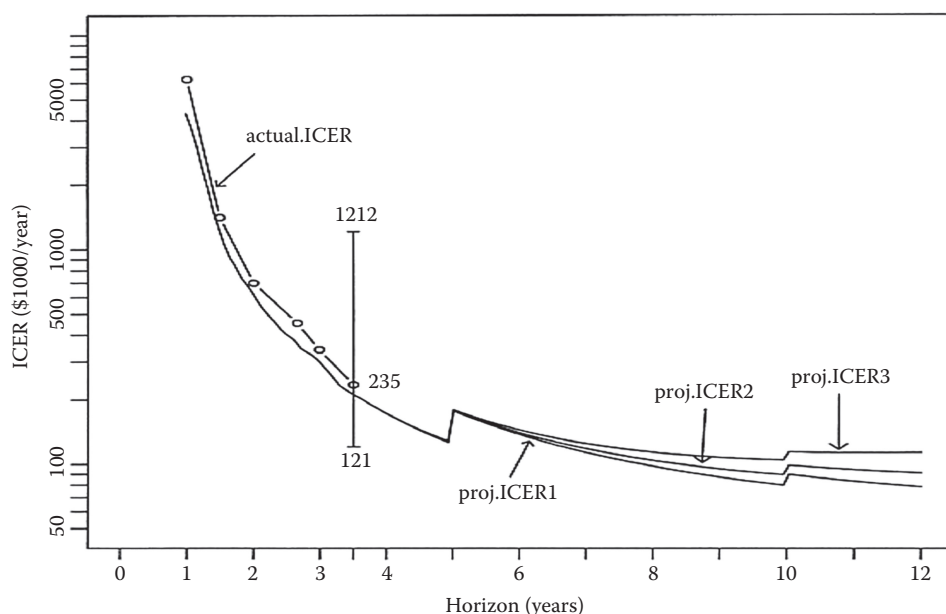


Fig. 3 ICER over years: Actual and projected connected for short vs. long terms

Source: Reproduced from Zwanziger et al.^[26] ©2006 Elsevier.

Care Act in the United States. It provides methodology guidelines such as:

The Patient-Centered Outcomes Research Institute ... shall not develop or employ a dollars per quality adjusted life year (or similar measure that discounts the value of a life because of an individual's disability) as a threshold to establish what type of health care is cost effective or recommended. The Secretary shall not utilize such an adjusted life year (or such a similar measure) as a threshold to determine coverage, reimbursement, or incentive programs under title XVIII. (The Patient Protection and Affordable Care Act)

In brief, the PCORI did not approve CEA and the use of QALY. Banning the use of a specific analysis or metric is quite unusual. In comparison, in 1996, the first U.S. Panel on Cost-Effectiveness in Health and Medicine (composed of physicians, health economists, ethicists, and other health policy experts) after 2 years of deliberation, recommended that CEA should use QALY as a standard metric to identify and assign value to health outcomes.^[28,29] Furthermore, the UK National Institute for Health and Care Excellence determines CE directly based on QALY valuations. Although it is understood that varying policies, processes, and decisions exist across different countries, agencies, and times, the PCORI's decision demonstrates how controversial CEA and its key measure can be.

In the CEA community, QALY is recommended and used in the "reference case" analysis. In contrast, in the EA community, QALY is rarely chosen as a primary outcome. Adding to this discrepancy in reporting, converting thousands of potential clinical outcomes (in natural health units) into QALY format and establishing consensus are formidable or nearly infeasible.

Which Perspective?

Total cost and CEA depend on the stakeholders' perspectives. Although the perspectives of society, institution, payer, and patient are all important, the societal perspective carries the most weight for policy decisions. From the comprehensive societal viewpoint that is relevant to all affected parties, all potential costs—direct, indirect, hidden, or opportunity—should be considered, regardless of who pays the costs or who receives the benefits. Patient- and institution-centered perspectives or ascertainment of actual costs spent/incurred, which is more explicit and objective, may provide additional information for ancillary or sensitivity purposes.

The Second U.S. Panel on Cost-Effectiveness in Health and Medicine (formed in 2012 and revised recommendations published in 2016) recommended that all CEA should report two reference case analyses—one based on a health-care sector perspective and another based on a societal perspective.^[29] Along the line, the Panel provided an "impact inventory," which is a structured table that contains consequences (both inside and outside the formal health-care sector), to clarify the scope and boundaries of the two reference case analyses.

Silence of λ , in the Eye of Policy Maker?

λ , the WTP, has both a positive property (it is a constant) and a negative property (it is arbitrary, unknown, and silent).^[3,30] The most frequently cited numerical value for λ is \$50,000 for one additional QALY gained. Other thresholds have been used, such as £3* \$100 K, CAN \$20 K, UK £20 K, among others. Ubel questioned the price of life and why it doesn't increase at the rate of inflation.^[31] Furthermore, it may be unlikely that \$50,000 per QALY is consistent with societal preferences in the modern medicine. Naturally, the criteria for judging CE are different at different times and in different places.

One versus Multiple Methods and Analyses?

In the CEA context, according to Hoch,^[10] "Published articles may be rejected by certain journals in part because of a reviewer's disagreement with the CEA methodology ... As studies that use CEA continue to proliferate, this conflict is sure to reach a flash point ... Journal editors may be forced to moderate methodological disagreements." On the other hand, Jiang, Wu, and Williams^[32] that stated "Different results can be more useful than blinded applications of only one method, and uninformative answer can be better than misleading one." The norm in CEA methodology is the combination of arithmetic mean, ICER, QALY, and societal perspective. Other methods may be considered in ancillary or sensitivity analyses.

Another less-discussed issue in CEA is related to multiplicity, selective reporting, confounding, and (external) validation which have been extensively discussed in other disciplines (e.g., EA using observational data). For instance, publication bias and selective reporting could be serious concerns; a majority of CEA publications showed ICER < \$50,000/QALY.^[33] Due to the heterogeneity of data sources, the variety of analytic methods/approaches, and the inherent subjectivity in outcomes, exploration rather than definitive conclusion may be a realistic goal.

CONCLUSION

In general, methodologies, implementation, and interpretation of EA, including meta-analysis, are reasonably well established and standardized; thus, the same or well-agreed methods, tests, computing software, and data sets can be used around the world (e.g., FDA, clinical trial course in biostatistics program, and clinical trial guidelines such as <http://www.consort-statement.org/>). In comparison, cost analysis and, to an even greater extent, CEA are destined to be more complicated and nuanced, where a set of recommendations and several guidelines are available (e.g., 1st and 2nd U.S. Panels on Cost-Effectiveness in Health and Medicine; WHO-CHOICE, <http://www.who.int/choice/cost-effectiveness/en/>; and ISPOR's CHEERS).

It is not uncommon to have one or two results for EA but 10 different results for CEA for a single aim/hypothesis. A valid comparative EA alone requires significant time and effort and is generally achieved via a well-done RCT, which is launched after sufficient cumulative evidence/justification. Its CEA counterpart would require even more time and work.

It is important to recognize that CEA can aid decision-making but does not make the decision for us. Other issues that must be considered are burden of disease, degree of uncertainty, CEA or unavailability of cost data, value judgement, non-economic or non-medical reasons for adoption or switching, monopoly prevention, and political price of saying “no” to some patients or treatments. Many health policy decisions are complex and controversial processes. Transparency in CEA would be essential, including data and model/algorithm sharing. The realistically best strategy may be to plan CEA alongside EA whenever feasible; beyond that, CEA follow-up that integrates more evidence and data can be conducted or awaited.

ACKNOWLEDGMENT

The author thanks Ms. Caron Modeas for editing service. The entry was partly supported by the National Institutes of Health through grant UL1 TR001860.

REFERENCES

1. Fojo, T.; Grady, C. How much is life worth: Cetuximab, non-small cell lung cancer, and the \$440 billion question. *J. Natl. Cancer Ins.* **2009**, *101* (15), 1044–1048.
2. Heshmat, S. Cost-effectiveness analysis. In *Encyclopedia of Biopharmaceutical Statistics*, 3rd Ed.; Chow, S.-C., Ed.; CRC Press: Boca Raton, FL, 2012, 363–368.
3. Obenchain, R.L. ICE preference maps: Nonlinear generalizations of net benefit and acceptability. *Health Serv. Outcomes Res. Method.* **2008**, *8*, 31–56.
4. Black, W.C. The CE plane: A graphic representation of cost-effectiveness. *Med. Decis. Making.* **1990**, *10* (3), 212–214.
5. Drummond, M.F.; Sculpher, M.J.; Torrance, G.W.; O'Brien, B.J.; Stoddart, G.L. *Methods for the Economic Evaluation of Health Care Programmes*, 3rd Ed.; Oxford University Press: Oxford, 2005.
6. Willan, A.; Lin, D. Incremental net benefit in randomized clinical trials. *Stat. Med.* **2001**, *20*, 1563–1574.
7. Fenwick, E.; Claxton, K.; Sculpher, M.J. Representing uncertainty: The role of cost-effectiveness acceptability curves. *Health Econ.* **2001**, *10* (8), 779–787.
8. Hoch, J.S.; Dewa, C.S. A clinician's guide to correct cost-effectiveness analysis: Think incremental not average. *Can. J. Psychiatry.* **2008**, *53* (4), 267–2674.
9. Laska, E.M.; Meisner, M.; Siegel, C. The usefulness of average cost-effectiveness ratios. *Health Econ.* **1997**, *6* (5), 497–504.
10. Hoch, J.S. An illustration of a current debate in cost-effectiveness analysis: Average cost-effectiveness ratios vs. incremental cost effectiveness ratios. *Assoc. Health Serv. Res. Meeting.* **1999**, *16*, 348–349.
11. Weinstein, M.; Zechkhauser, R. Critical ratios and efficient allocation. *J. Public Econ.* **1973**, *2* (2), 147–157.
12. Rosen, S.; Long, L.; Sanneb, I., Eds. *Cost and Cost-effectiveness of Antiretroviral Treatment Delivery*; Elizabeth Glaser Pediatric AIDS Foundation: Boston, MA, 2009.
13. Asch, D.A.; Christakis, N.A.; Ubel, P.A. Conducting physician mail surveys on a limited budget: A randomized trial comparing 2 bill versus 5 bill incentives. *Med. Care.* **1998**, *36* (1), 95–99.
14. Bang, H.; Zhao, H. Median-based incremental cost-effectiveness ratio (ICER). *J. Stat. Theory Pract.* **2012**, *6* (3), 428–442.
15. Zhou, X.H.; Melfi, C.A.; Hui, S.L. Methods for comparison of cost data. *Ann. Intern. Med.* **1997**, *127* (8 Pt 2), 752–756.
16. Barber, J.A.; Thompson, S.G. Analysis of cost data in randomized trials: An application of the non-parametric bootstrap. *Stat. Med.* **2000**, *19* (23), 3219–3236.
17. Bang, H.; Zhao, H. Cost-effectiveness analysis: A proposal of new reporting standards in statistical analysis. *J. Biopharm. Stat.* **2014**, *24* (2), 443–460.
18. Brookmeyer, R. Median survival time. *Enc. Biostat.* **2005**, *5*, 1–4. Accessed at: <http://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat06043/pdf>.
19. Guyot, P.; Welton, N.; Ouwens, M.; Ades, A. Survival time outcomes in randomized, controlled trials and meta-analyses: The parallel universes of efficacy and cost-effectiveness. *Value Health.* **2011**, *14* (5), 640–646.
20. Robinson, R. Cost-effectiveness analysis. *BMJ.* **1993**, *307*, (6907), 793–795.
21. Keeler, E.B.; Cretin, S. Discounting of life-saving and other nonmonetary effects. *Manage. Sci.* **1983**, *29* (3), 300–306.
22. Zhao, H.; Cheng, Y.; Bang, H. Some insight on censored cost estimators. *Stat. Med.* **2011**, *30* (19), 2381–2388.
23. Willan, A.R.; Lin, D.Y.; Cook, R.J.; Chen, E.B. Using inverse-weighting in cost-effectiveness analysis with censored data. *Stat. Methods Med. Res.* **2002**, *11* (6), 539–551.
24. Petrou, S.; Gray, A. Economic evaluation using decision analytical modelling: Design, conduct, analysis, and reporting. *BMJ.* **2011**, *342*, d1766.
25. Brown, A.W.; Hall, W.D.; Thomas, D.; Dhurandhar, N.V.; Heymsfield, S.B.; Allison, D.B. Order of magnitude misestimation of weight effects of children's meal policy proposals. *Child Obes.* **2014**, *10* (6), 542–544.
26. Zwanziger, J.; Hall, W.J.; Dick, A.W.; Zhao, H.; Mushlin, A.I.; Hahn, R.M.; Wang, H.; Andrews, M.; Mooney, C.; Wang, H.; Moss, A.J.; The cost effectiveness of implantable cardioverter-defibrillators: Results from the MADIT-II. *JACC.* **2006**, *47* (11), 2310–2318.
27. Glick, H.A. Sample size and power for cost-effectiveness analysis (part I). *Pharmacoeconomics.* **2011**, *29* (3), 189–198.
28. Neumann, P.J.; Weinstein, M.C.; Legislating against use cost-effectiveness information. *NEJM.* **2010**, *363* (16), 1495–1497.
29. Sanders, G.D.; Neumann, P.J.; Basu, A.; Brock, D.W.; Feeny, D.; Krahn, M.; Kuntz, K.M.; Meltzer, D.O.; Owens, D.K.; Prosser, L.A.; Salomon, J.A.; Sculpher, M.J.;

- Trikalinos, T.A.; Russell, L.B.; Siegel, J.E.; Ganiats, T.G.; Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: Second panel on cost-effectiveness in health and JAMA. **2016**, *316* (10), 1093–1103.
30. Gafni, A.; Birch, S. Incremental cost-effectiveness ratios (ICERs): The silence of the lambda. *Soc. Sci. Med.* **2006**, *62* (9), 2091–2100.
31. Ubel, P.A.; Hirth, R.A.; Chernew, M.E.; Fendrick, A.M.; What is the price of life and why doesn't it increase at the rate of inflation? *Arch. Intern. Med.* **2003**, *163* (14), 1637–1641.
32. Jiang, G.; Wu, J.; Williams, G.R.; Fieller's interval and the Bootstrap-Fieller interval for the incremental cost-effectiveness ratio. *Health Serv. Outcomes Res. Method.* **2000**, *1* (3), 291–303.
33. Cohen, J.T.; Neumann, P.J.; Weinstein, M.C. Does preventive care save money? Health economics and the presidential candidates. *NEJM.* **2008**, *358*, 661–663.

Covariate-Adjusted Adaptive Dose-Finding: Early Phase Clinical Trials

Peter F. Thall and Hoang Q. Nguyen

Department of Biostatistics, The University of Texas, M.D. Anderson Cancer Center, Houston, Texas, U.S.A.

Abstract

Most methods for sequentially adaptive dose-finding in phase I clinical trials characterize outcome as an indicator Y_T of toxicity and assume that patients are homogeneous. This entry reviews a new method which requires an informative prior on covariate effect parameters, obtained from historical data.

Keywords: Adaptive design; Bayesian design; Individualized treatment; Phase I clinical trial; Phase I/II clinical trial.

INTRODUCTION

Most methods for sequentially adaptive dose-finding in phase I clinical trials characterize outcome as an indicator Y_T of toxicity and assume that patients are homogeneous. Dose-finding is based on an algorithm involving $(x, Y_T)^{[1,2]}$ or model-based criteria defined in terms of $\pi_T(x, \theta) = \Pr(Y_T = 1 | x, \theta)^{[3,4]}$ where x denotes dose and θ is a model parameter vector. Doses are chosen adaptively for successive patient cohorts, with the main scientific goal to choose a dose having an acceptable probability of toxicity from a set $\mathcal{X} = \{x_1, \dots, x_K\}$. There are two useful extensions of this paradigm. The first accounts for effects of patient covariates, $\mathbf{Z} = (Z_1, \dots, Z_q)$, and so that $\pi_T(x, \theta)$ becomes $\pi_T(x, \mathbf{Z}, \theta)$. This provides a basis for using each patient's $\mathbf{Z}$ to improve dose selection, so-called “patient-specific” or “individualized” dose assignment.^[5-9] The second extension uses an indicator Y_E of efficacy along with Y_T to choose doses, giving a so-called “Phase I/II” design,^[10-18] with the aim to choose an x having acceptably large $\pi_E(x)$ and small $\pi_T(x)$.

We review a new method^[19] that combines these two extensions by basing dose-finding on $\mathbf{Y} = (Y_E, Y_T)$ while also accounting for $\mathbf{Z}$. For $k = E, T$ let $\pi_k(x, \mathbf{Z}, \theta) = \Pr(Y_k = 1 | x, \mathbf{Z}, \theta)$ for a patient with covariates $\mathbf{Z}$ treated with dose x , with $\pi(x, \mathbf{Z}, \theta) = (\pi_E(x, \mathbf{Z}, \theta), \pi_T(x, \mathbf{Z}, \theta))$, denoted respectively by π_E, π_T and π for brevity. The method requires an informative prior on covariate effect parameters, obtained from historical data. Elicited lower limits on π_E and upper limits on π_T for each of a representative set of covariate vectors are used to construct bounding functions that determine the acceptability of each element of $\mathcal{X}$ for each $\mathbf{Z}$. For a reference vector $\mathbf{Z}^*$, the physician must specify several equally desirable target values of π

that are used to construct criteria for comparing acceptable doses. Rather than choosing one best dose, the data from the trial are used to select $\mathbf{Z}$ -specific x for future patients.

PROBABILITY MODEL

Regression Models for Efficacy and Toxicity

Let $D_n = \{(\mathbf{Y}_i, \mathbf{Z}_i, x_{[i]}), i = 1, \dots, n\}$ denote the data from the trial's first n patients where $x_{[i]}$ is the dose of patient i , and denote the historical data by $\mathcal{H} = \{(\mathbf{Y}_i, \mathbf{Z}_i, \mathcal{T}_{[i]}), i = 1, \dots, n_H\}$, where $\{\mathcal{T}_1, \dots, \mathcal{T}_m\}$ are the historical treatments and $\mathcal{T}_{[i]}$ is the i th patient's treatment. Let unsubscripted T denote either a dose or historical treatment. For a patient with covariates $\mathbf{Z}$ treated with T let $\pi_{a,b}(T, \mathbf{Z}, \theta) = \Pr(Y_E = a, Y_T = b | T, \mathbf{Z}, \theta)$, for $a, b \in \{0, 1\}$ so that $\pi_E(T, \mathbf{Z}, \theta) = \Pr(Y_E = 1 | T, \mathbf{Z}, \theta) = \pi_{1,1}(T, \mathbf{Z}, \theta) + \pi_{1,0}(T, \mathbf{Z}, \theta)$ and $\pi_T(T, \mathbf{Z}, \theta) = \Pr(Y_T = 1 | T, \mathbf{Z}, \theta) = \pi_{1,1}(T, \mathbf{Z}, \theta) + \pi_{0,1}(T, \mathbf{Z}, \theta)$. For link function g , denote the linear terms $\eta_k = g(\pi_k)$. The bivariate model is determined by the marginals $\pi_E = g^{-1}(\eta_E)$ and $\pi_T = g^{-1}(\eta_T)$ and one association parameter, ψ . There are many possible models, including the Gumbel

$$\pi_{a,b} = \pi_E^a (1 - \pi_E)^{1-a} \pi_T^b (1 - \pi_T)^{1-b} + (-1)^{a+b} \psi \pi_E^a (1 - \pi_E)^{1-a} \pi_T^b (1 - \pi_T)^{1-b}, \quad (1)$$

with $-1 \leq \psi \leq 1$, and bivariate copulas,^[20] including the Gaussian copula

$$C_\psi(u, v) = \Phi_\psi(\Phi^{-1}(u), \Phi^{-1}(v)), \quad 0 \leq u, v \leq 1, \quad (2)$$

where Φ is the $N(0, 1)$ cdf and Φ_ψ is the bivariate $N(0, 1)$ cdf with correlation ψ . This gives $\pi_{0,0} = \Phi_\psi(\Phi^{-1}(1 - \pi_E), \Phi^{-1}(1 - \pi_T))$, $\pi_{1,0} = 1 - \pi_T - \pi_{0,0}$, $\pi_{0,1} = 1 - \pi_E - \pi_{0,0}$ and $\pi_{1,1} = \pi_E + \pi_T + \pi_{0,0} - 1$.

Given a bivariate model, the following forms of $\eta_E(T, \mathbf{Z}, \theta)$ and $\eta_T(T, \mathbf{Z}, \theta)$ provide a basis for using H to learn about covariate effects and accounting for joint effects of $(x, \mathbf{Z})$ on π based on the current trial data, D_n . The general form is

$$\eta_k(\mathcal{T}, \mathbf{Z}, \theta) = \beta_k \mathbf{Z} + \sum_{j=1}^m (\mu_{k,j} + \xi_{k,j} \mathbf{Z}) \mathbf{1}(\mathcal{T} = \mathcal{T}_j) + \{\phi_k(x, \alpha_k) + \gamma_k \mathbf{Z}\} \mathbf{1}(\mathcal{T} = x), \quad (3)$$

for $k = E, T$, where $\mathbf{1}(A)$ indicates the event A , and $\phi_E(x, \alpha_E)$ and $\phi_T(x, \alpha_T)$ characterize the main dose effects on π_E and π_T . For each k , the covariate main effects are $\beta_k = (\beta_{k,1}, \dots, \beta_{k,q})$, interactions between $\mathbf{Z}$ and historical treatment T_j are $\xi_{k,j} = (\xi_{k,j,1}, \dots, \xi_{k,j,q})$ and dose-covariate $(x - \mathbf{Z})$ interactions are $\gamma_k = (\gamma_{k,1}, \dots, \gamma_{k,q})$. For $\mathcal{H}$, the expression (3) is

$$\eta_k(T_j, \mathbf{Z}, \theta) = \mu_{k,j} + \beta_k \mathbf{Z} + \xi_{k,j} \mathbf{Z}, \quad (4)$$

for $j = 1, \dots, m$ and $k = E, T$,

where $\mu_k = (\mu_{k,1}, \dots, \mu_{k,m})$ are historical treatment main effects. For the trial, (3) is

$$\eta_k(x, \mathbf{Z}, \theta) = \phi_k(x, \alpha_k) + \beta_k \mathbf{Z} + x \gamma_k \mathbf{Z}, \quad \text{for } k = E, T. \quad (5)$$

The functions ϕ_E and ϕ_T should reflect the application. For example, for a cytotoxic agent, $\phi_T(x, \alpha_T) = \alpha_{T,0} + \alpha_{T,1} x$ with $\Pr(\alpha_{T,1} > 0) = 1$ gives $\pi_T(x, \mathbf{Z}, \theta)$ increasing in x , while a quadratic function $\phi_k(x, \alpha_k) = \alpha_{k,0} + \alpha_{k,1} x + \alpha_{k,2} x^2$ allows π_k to be non-monotone in x , which may be appropriate for biologic agents.

Denote $\alpha = (\alpha_E, \alpha_T)$, $\mu = (\mu_E, \mu_T)$, $\beta = (\beta_E, \beta_T)$, $\gamma = (\gamma_E, \gamma_T)$ and $\xi = (\xi_{E,1}, \xi_{T,1}, \dots, \xi_{E,k}, \xi_{T,k})$ so that $\theta = (\mu, \beta, \xi, \gamma, \alpha, \gamma)$. The likelihood for the current trial data is

$$L(D_n | \theta) = \prod_{i=1}^n \prod_{a=0}^1 \prod_{b=0}^1 \{\pi_{a,b}(x_{[i]}, \mathbf{Z}_i, \theta)\}^{\mathbf{1}\{\mathbf{Y}_i=(a,b)\}}.$$

Denoting $D_n^{\mathcal{H}} = \mathcal{H} \cup D_n$, the posterior is

$$p(\alpha, \gamma, \beta, \psi | D_n^{\mathcal{H}}) \propto L(D_n | \theta) p(\alpha, \gamma) p(\beta, \psi | \mathcal{H}). \quad (6)$$

Prior Specification

A fit of H yields the informative posterior $p = (\mu, \beta, \xi, \gamma | \mathcal{H})$. For covariate-adjusted dose-finding, the informative marginal posterior $p = (\beta, \psi | \mathcal{H})$ is essential, since it is used

as the prior on (β, ψ) at the start of the trial. Because α and γ account for effects of the experimental agent, their priors are non-informative. For prior means, set $E(\gamma) = 0$, and obtain prior $E(\alpha)$, by eliciting means of $\pi_E(x_j, \mathbf{Z}^*, \theta)$ and $\pi_T(x_j, \mathbf{Z}^*, \theta)$ at two or more values of x_j and solving for $E(\alpha)$, using the least squares method of Thall and Cook^[15] if the number of elicited means is larger than $\dim(\alpha)$.

To determine prior variances, assume all entries of α and γ have prior variance σ_0^2 , and calibrate σ_0^2 to control the effective sample size (ESS) of the prior $p\{\pi_k(x_j, \mathbf{Z}^*, \theta)\}$ for all $k = E, T$ and x_j . One may do this by matching the first two moments of each $\pi_k(x_j, \mathbf{Z}^*, \theta)$ with a beta $(a_{k,j}, b_{k,j})$ and approximating the ESS of $p\{\pi_k(x_j, \mathbf{Z}^*, \theta)\}$ by $a_{k,j} + b_{k,j} = E\{\pi_k(x_j, \mathbf{Z}^*, \theta)\} [1 - E\{\pi_k(x_j, \mathbf{Z}^*, \theta)\}] / \text{var}\{\pi_k(x_j, \mathbf{Z}^*, \theta)\} - 1$. A more general definition of ESS is given by Morita, Thall, and Mueller.^[21] To obtain a non-informative prior on dose effects and allow $p = (\beta, \psi | \mathcal{H})$ and the trial's data to dominate the dose-finding decisions, specify σ_0^2 to ensure that all $a_{k,j} + b_{k,j}$ are reasonably small. One also must control $\sigma^2(\alpha_{k,2} x^2) = \text{var}(\alpha_{k,2} x^2)$ relative to $\sigma^2(\alpha_{k,1} x) = \text{var}(\alpha_{k,1} x)$ for each k to avoid the quadratic effect $\alpha_{k,2} x^2$ being too large relative to the first order effect $\alpha_{k,1} x$ and thus misrepresenting ϕ_k . To do this, fix $\sigma(\alpha_{k,1}) = \sigma_0$, and examine effects of the ratio $\sigma(\alpha_{k,2} x^2) / \sigma(\alpha_{k,1} x) = x \sigma(\alpha_{k,2}) / \sigma(\alpha_{k,1}) = x \sigma(\alpha_{k,2}) / \sigma_0$ in the range 0.1 to 1.0 on the design's operating characteristics (OCs), and choose $\sigma(\alpha_{k,2})$ to ensure good OCs and reasonable prior ESS.

COVARIATE-ADJUSTED DOSE-FINDING

At any interim point in the trial, quantities computed from the posteriors $p(\alpha, \gamma, \beta, \psi | D_n^{\mathcal{H}})$ are used to choose doses adaptively. To account for both x and $\mathbf{Z}$, two different criteria are used, the first to determine whether x is *acceptable* for given $\mathbf{Z}$, and the second to quantify the *desirability* of each π in terms of a tradeoff between π_E and π_T .

Dose Acceptability

The following construction obtains a lower bound on $\pi_E(x, \mathbf{Z}, \theta)$ and an upper bound on $\pi_T(x, \mathbf{Z}, \theta)$ as $\mathbf{Z}$ is varied. To begin, a representative set of covariates, $\{\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(K)}\}$, must be specified. For each $\mathbf{Z}^{(j)}$, elicit the smallest probability of efficacy, $\underline{\pi}_E^{(j)}$, and the largest probability of toxicity, $\bar{\pi}_T^{(j)}$, that the physician will allow for a patient with covariates $\mathbf{Z}^{(j)}$. Denote $\zeta_k(\mathbf{Z}) = E(\beta_k \mathbf{Z} | H)$. To construct a bounding function for $\pi_E(x, \mathbf{Z}, \theta)$, treat the K pairs $\{(\zeta_E(\mathbf{Z}^{(j)}), \underline{\pi}_E^{(j)}), j = 1, \dots, K\}$, of estimated linear terms and elicited lower bounds on π_E like regression data and fit a smooth curve with $\zeta_E(\mathbf{Z}^{(j)})$ the predictor and $\underline{\pi}_E^{(j)}$ the outcome. A linear function $\underline{\pi}_E(\zeta_E) = a + b \zeta_E$ or quadratic $\underline{\pi}_E(\zeta_E) = a + b \zeta_E + c \zeta_E^2$ work well. Denoting the regression estimate by $\hat{\pi}_E(\zeta_E)$, the *efficacy lower bounding function*

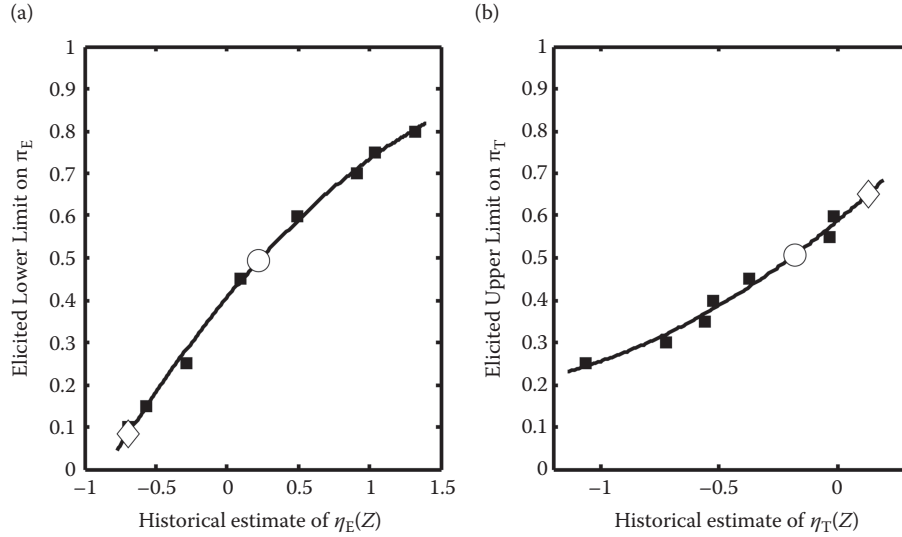


Fig. 1 Dose acceptability bounding functions. Each box represents a pair $(\zeta_E(\mathbf{Z}^{(j)}), \pi_E^{(j)})$ in (a) and a pair $(\zeta_T(\mathbf{Z}^{(j)}), \pi_T^{(j)})$ in (b). An open diamond denotes bounds for a 58-year-old patient with poor cytogenetics, and an open circle denotes bounds for a 48-year-old patient with intermediate cytogenetics

is $\pi_E(\mathbf{Z}) = \hat{\pi}_E \circ \zeta_E(\mathbf{Z})$. Similarly, fit a curve to the K pairs $\{(\zeta_T(\mathbf{Z}^{(j)}), \pi_T^{(j)}), j=1, \dots, K\}$, obtain an estimate $\hat{\pi}_T(\zeta_T)$, and define the *toxicity upper bounding function* to be $\bar{\pi}_T(\mathbf{Z}) = \hat{\pi}_T \circ \zeta_T(\mathbf{Z})$. These values should be plotted to allow any $\pi_E^{(j)}$ or $\pi_T^{(j)}$ to be adjusted if desired. Fig. 1 gives the scattergrams of elicited values and bounding functions for the illustrative trial. Given D_n , the set of acceptable doses for a patient with covariates $\mathbf{Z}$ is

$$A_n(\mathbf{Z}) = \{x \in \chi : \Pr[\pi_E(x, \mathbf{Z}, \boldsymbol{\theta}) < \bar{\pi}_E(\mathbf{Z}) \mid D_n^{\mathcal{H}}] < p_E \text{ and } \Pr[\pi_T(x, \mathbf{Z}, \boldsymbol{\theta}) > \bar{\pi}_T(\mathbf{Z}) \mid D_n^{\mathcal{H}}] < p_T\} \quad (7)$$

The cutoffs p_T and p_E should be calibrated to obtain good OCs, with values around 0.90 to 0.95 generally giving good designs. The set $A_n(\mathbf{Z})$ changes adaptively during the trial for each $\mathbf{Z}$. If $A_n(\mathbf{Z})$ is empty then no dose is acceptable for that patient. If $A_n(\mathbf{Z})$ consists of a single dose then that dose is used by default. If $|A_n(\mathbf{Z})| \geq 2$ then a criterion for choosing the patient's dose from $A_n(\mathbf{Z})$ is needed.

Dose Desirability

To select one dose from $A_n(\mathbf{Z})$, the following criteria for comparing doses may be used. These are defined in terms of a *target contour*, C , of π values that characterizes the tradeoff between π_E and π_T . A general approach for constructing C is to elicit target probability pairs $\pi_1^*, \dots, \pi_c^*$ considered to be equally desirable for a reference patient with covariates $\mathbf{Z}^*$ and define C to be a smooth, strictly increasing curve obtained by treating π_T as a function of π_E and fitting the elicited points. Fig. 2 gives the elicited

values and C in the illustration. An exact method for defining C based on three elicited targets is given by Thall and Cook.^[17]

Given C , the desirability $\delta_n(x, \mathbf{Z})$ of a dose x for a patient with covariates $\mathbf{Z}$ may be defined in several ways. For example, one may define $\delta_n(x, \mathbf{Z})$ to be the posterior probability given $D_n^{\mathcal{H}}$ of the set $\{\pi : \pi_E \geq \pi_E' \text{ and } \pi_T \leq \pi_T', \exists \pi' \in C\}$ of $\pi(x, \mathbf{Z})$ that are at

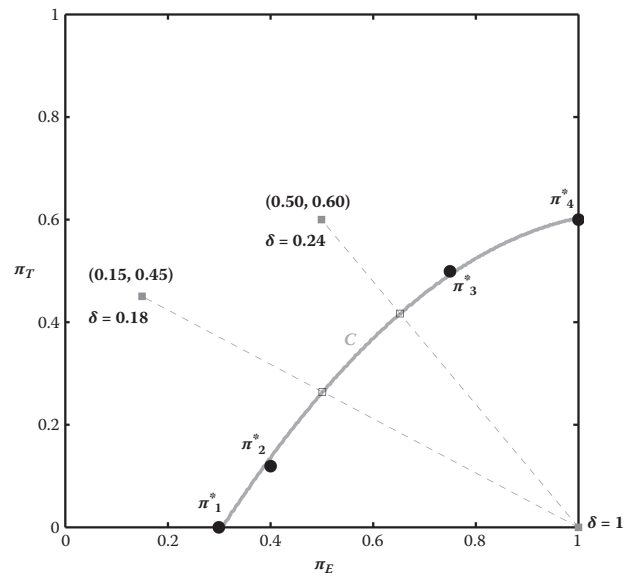


Fig. 2 Elicited targets and tradeoff contour. The four equally desirable elicited π pairs are denoted by shaded circles located along the fitted contour C . The two dashed line segments illustrate that the π pair (0.50, 0.60), which has desirability 0.24, is preferred over the pair (0.15, 0.45), which has desirability 0.18

least as desirable as a pair on C . Here, we use the following geometric definition based on the point π_c where the straight line segment L_π through π and the ideal point $(1,0)$ intersects C , as shown in Fig. 2. Denoting Euclidean distance by $\|\cdot\|$, the desirability of $\pi \in [0,1]^2$ is defined as $d_c(\pi) = \exp(-\|\pi - (1,0)\| / \|\pi_c - (1,0)\|)$. This function may be used to compare π pairs since $d_c(\pi)$ decreases from its maximum $d_c(1,0) = 1$ as π moves away from the ideal pair $(1,0)$ along any line L_π . Given d , the set of π having $d_c(\pi) = d$ form an *isocontour*, or *indifference* set in $[0,1]^2$. For a patient with covariates $\mathbf{Z}$, the desirability of x is $\delta_n(x, \mathbf{Z}) = d_c(E\{\pi(x, \mathbf{Z}, \theta) | D_n^H\})$, and the best acceptable dose is the x maximizing $\delta_n(x, \mathbf{Z})$.

TRIAL CONDUCT

Designing a trial using this methodology requires one to first analyze available historical data, which is used as a basis for choosing a model among several candidates^[19] and computing $p(\beta, \psi | \mathcal{H})$ under the selected model. One then must establish priors on α and δ doses χ , maximum sample size, N , a set of representative $\mathbf{Z}$, and acceptability and desirability criteria. During the trial, for each new patient $A_n(\mathbf{Z})$ is computed based on that patient's $\mathbf{Z}$ and the current $p(\theta | D_n^H)$, and the following decision rules are applied.

1. If $A_n(\mathbf{Z})$ is empty, do not treat the patient with any $x \in \chi$.
2. If $A_n(\mathbf{Z})$ is not empty, treat the patient using the dose x maximizing $\delta_n(x, \mathbf{Z})$.
3. If $A_n(\mathbf{Z}^{(j)})$ is empty for all $j = 1, \dots, K$, then stop the trial and declare the new agent unacceptable for any patient at any dose.
4. After the trial is completed, select doses for future patients using the acceptability and desirability criteria based on $p(\theta | D_n^H)$.

Rule (1) is a covariate-specific combined futility and safety stopping rule. This provides a formal basis for decisions based on a patient's particular characteristics that typically must be made subjectively by the physician when a design that ignores covariates is used.

ILLUSTRATION

Our illustration is a phase I/II trial of a chemo-protective agent (CPA), given at one of five dose levels 12, 15, 18, 21, or 24 mg/m² daily for three days, combined with fixed doses of the standard drugs idarubicin (IDA) and cytosine arabinoside in untreated acute myelogenous leukemia (AML) patients under age 60. The therapeutic goal of chemotherapy for AML is to achieve complete remission (CR), defined as recovery of circulating blood cells and immature cells in the bone marrow to normal levels. For the trial, E is the event that the patient is alive and in CR at day 42 of treatment, and T is defined as severe mucositis, diarrhea, pneumonia, or sepsis within 42 days. The design was motivated by the problem that higher doses of IDA increase both π_E and π_T , and the hypothesis that the CPA would decrease the risk of IDA-induced toxicity. Additional details are given in Thall, Nguyen, and Estey.^[19]

The covariates used in the design were Age, with $Z_1 = .01(\text{Age} - 45)$ for numerical stability, and cytogenetic abnormality (Cyto), classified in terms of prognosis as Good, Intermediate or Poor, with $Z_2 = 1$ (Cyto = Good) and $Z_3 = 1$ (Cyto = Poor). Quadratic dose effects $f_k(x, \alpha_k) = \alpha_{k,0} + \alpha_{k,1}x + \alpha_{k,2}x^2$ for both $k = E$ and T were assumed, with dose levels coded as $x = -2, -1, 0, 1, 2$. Analyses of historical data from 693 untreated AML patients age <60 treated with chemotherapy showed that, among a set of models defined in terms of bivariate distributional form and link function, the Gaussian copula with probit link gave the best fit to H . Table 1 gives posterior means and standard deviations of $\pi_E(\mathbf{Z}, \theta)$ and $\pi_T(\mathbf{Z}, \theta)$ under

Table 1 For Each Representative $\mathbf{Z}$, Posterior Means of $\pi(\mathbf{Z}, \theta)$ under the Gaussian Copula with Probit Link and Desirabilities Based on the Historical Data and Corresponding Elicited Bounds on $\pi_E(\mathbf{Z}, \theta)$ and $\pi_T(\mathbf{Z}, \theta)$

Age	Cyto	Posterior Mean _{sd}		d_c	Elicited Bounds	
		$\pi_E(\mathbf{Z}, \theta)_{sd}$	$\pi_T(\mathbf{Z}, \theta)_{sd}$		$\underline{\pi}_E(\mathbf{Z})$	$\bar{\pi}_T(\mathbf{Z})$
27	Good	0.90 _{.03}	0.15 _{.04}	0.73	0.80	0.25
27	Interm	0.69 _{.04}	0.30 _{.04}	0.45	0.60	0.40
27	Poor	0.39 _{.06}	0.36 _{.05}	0.28	0.25	0.45
48	Good	0.85 _{.04}	0.24 _{.05}	0.60	0.75	0.30
48	Interm	0.59 _{.02}	0.43 _{.02}	0.33	0.50	0.50
48	Poor	0.29 _{.04}	0.49 _{.04}	0.21	0.15	0.55
58	Good	0.81 _{.05}	0.29 _{.06}	0.53	0.70	0.35
58	Interm	0.54 _{.03}	0.49 _{.03}	0.29	0.45	0.60
58	Poor	0.25 _{.04}	0.55 _{.05}	0.18	0.10	0.65

Table 2 Simulation Scenarios. Dose-covariate Interactions $\gamma = (\gamma_{E, \text{Age}}, \gamma_{E, \text{CytoPoor}}, \gamma_{E, \text{CytoGood}}, \gamma_{T, \text{Age}}, \gamma_{T, \text{CytoPoor}}, \gamma_{T, \text{CytoGood}})$ are (0,0,0,0,0,0) under Scenarios 1 and 3, and (0,.6,-.8,0,-.6,.8) under Scenarios 2 and 4

	Dose level		1	2	3	4	5
Scenario 1	27, Good	$\pi_E, \pi_T d_c$	0.57, 0.04	0.70, 0.06	0.89, 0.07	0.96, 0.08	0.97, 0.09
		$\pi_E, \pi_T d_c$	0.52	0.61	0.79	0.84	0.84
	48, Interm	$\pi_E, \pi_T d_c$	0.18, 0.20	0.28, 0.24	0.55, 0.28	0.74, 0.31	0.79, 0.33
		$\pi_E, \pi_T d_c$	0.25	0.28	0.38	0.47	0.49
	58, Poor	$\pi_E, \pi_T d_c$	0.03, 0.30	0.07, 0.35	0.21, 0.40	0.39, 0.43	0.46, 0.45
		$\pi_E, \pi_T d_c$	0.18	0.18	0.21	0.26	0.27
Scenario 2	27, Good	$\pi_E, \pi_T d_c$	0.15, 0.30	0.46, 0.16	0.89, 0.07	0.99, 0.02	1.00, 0.01
		$\pi_E, \pi_T d_c$	0.22	0.39	0.79	0.95	0.99
	48, Interm	$\pi_E, \pi_T d_c$	0.18, 0.20	0.28, 0.24	0.55, 0.28	0.74, 0.31	0.79, 0.33
		$\pi_E, \pi_T d_c$	0.25	0.28	0.38	0.47	0.49
	58, Poor	$\pi_E, \pi_T d_c$	0.41, 0.02	0.24, 0.12	0.21, 0.40	0.14, 0.73	0.04, 0.93
		$\pi_E, \pi_T d_c$	0.42	0.30	0.21	0.13	0.09
Scenario 3	27, Good	$\pi_E, \pi_T d_c$	0.43, 0.03	0.76, 0.07	0.94, 0.12	0.90, 0.17	0.65, 0.31
		$\pi_E, \pi_T d_c$	0.43	0.66	0.79	0.70	0.42
	48, Interm	$\pi_E, \pi_T d_c$	0.10, 0.15	0.35, 0.28	0.68, 0.38	0.58, 0.47	0.24, 0.65
		$\pi_E, \pi_T d_c$	0.24	0.29	0.40	0.31	0.16
	58, Poor	$\pi_E, \pi_T d_c$	0.01, 0.23	0.10, 0.40	0.33, 0.50	0.24, 0.60	0.05, 0.76
		$\pi_E, \pi_T d_c$	0.19	0.18	0.22	0.17	0.11
Scenario 4	27, Good	$\pi_E, \pi_T d_c$	0.08, 0.24	0.54, 0.20	0.94, 0.12	0.97, 0.06	0.94, 0.04
		$\pi_E, \pi_T d_c$	0.21	0.42	0.79	0.89	0.88
	48, Interm	$\pi_E, \pi_T d_c$	0.10, 0.15	0.35, 0.28	0.68, 0.38	0.58, 0.47	0.24, 0.65
		$\pi_E, \pi_T d_c$	0.24	0.29	0.40	0.31	0.16
	58, Poor	$\pi_E, \pi_T d_c$	0.28, 0.01	0.31, 0.14	0.33, 0.50	0.07, 0.85	0.00, 0.99
		$\pi_E, \pi_T d_c$	0.35	0.32	0.22	0.10	0.07

this model for each $\mathbf{Z}$ in the representative set of nine covariate vectors. The posterior mean historical covariate effects were $\zeta_E(\mathbf{Z}) = 0.262 - 1.308Z_1 + 0.818Z_2 - 0.785Z_3$ and $\zeta_T(\mathbf{Z}) = -0.232 + 1.619Z_1 - 0.542Z_2 + 0.148Z_3$. The acceptability bounding functions, derived from the elicited covariate-specific limits given in Table 2, were the quadratic functions

$$\pi_E(\mathbf{Z}) = 0.4063 + 0.4078\zeta_E(\mathbf{Z}) - 0.0806\{\zeta_E(\mathbf{Z})\}^2$$

and

$$\pi_T(\mathbf{Z}) = 0.5890 + 0.4739\zeta_T(\mathbf{Z}) + 0.1403\{\zeta_T(\mathbf{Z})\}^2.$$

These bounding functions are plotted in Fig. 1. The acceptability criteria were applied with $p_E = p_T = 0.90$. The target contour C was obtained by fitting a quadratic to the four target pairs (0.75, 0.50), (0.30, 0), (1, 0.60), (0.40, 0.12), considered equally desirable improvements over the historical mean $E\{\pi(\mathbf{Z}^*, \theta) | \mathcal{H}\} = (.59, .43)$ for a reference patient with $\mathbf{Z}^* = (48, \text{Intermed})$. The fitted target curve was $\pi_T = -.5605 + 2.1226\pi_E - .9591\pi_E^2$, and C was defined

to be the set of π satisfying this equation for $0.30 \leq \pi_E \leq 1$, as shown in Fig. 2.

Elicited prior means of $\pi(x_j, \mathbf{Z}^*, \theta)$ for the five dose levels were the historical means (0.59, 0.43), (0.70, 0.50), (0.75, 0.55), (0.85, 0.70), and (0.90, 0.90), for the doses 12, 15, 18, 21, and 24 mg/m². A least squares fit^[15,19] yielded prior means (0.773, 0.262, 0.0077) for $(\alpha_{E,0}, \alpha_{E,1}, \alpha_{E,2})$ and (0.098, 0.344, 0.1025) for $(\alpha_{T,0}, \alpha_{T,1}, \alpha_{T,2})$. The prior means of all $x - \mathbf{Z}$ interaction parameters γ_E and γ_T were set equal to 0, with a maximum ESS of 1.3 for any $\pi(x_j, \mathbf{Z}^*)$. Each parameter in $\alpha_E, \alpha_T, \gamma_E$ and γ_T was given prior standard deviation 1.33, except for those of the quadratic coefficients $\alpha_{E,2}$ and $\alpha_{T,2}$, which were set equal to 0.266.

The planned maximum sample size of the AML trial was $N = 60$ patients. Fig. 3 illustrates the posteriors of $\pi(x_j, \mathbf{Z}, \theta)$ and corresponding desirability $\delta_n(x_j, \mathbf{Z})$ for each combination of x_j and $\mathbf{Z} = (27, \text{Good}), (48, \text{Interm})$ or $(58, \text{Poor})$ based on the final data D_{60} of a hypothetical trial.

Table 2 summarizes four scenarios under which the trial was simulated. Three methods were studied in the simulations: (i) the covariate-adjusted method as described here including $x - \mathbf{Z}$ interactions, denoted by $\mathbf{Z}$ -INT; (ii) the

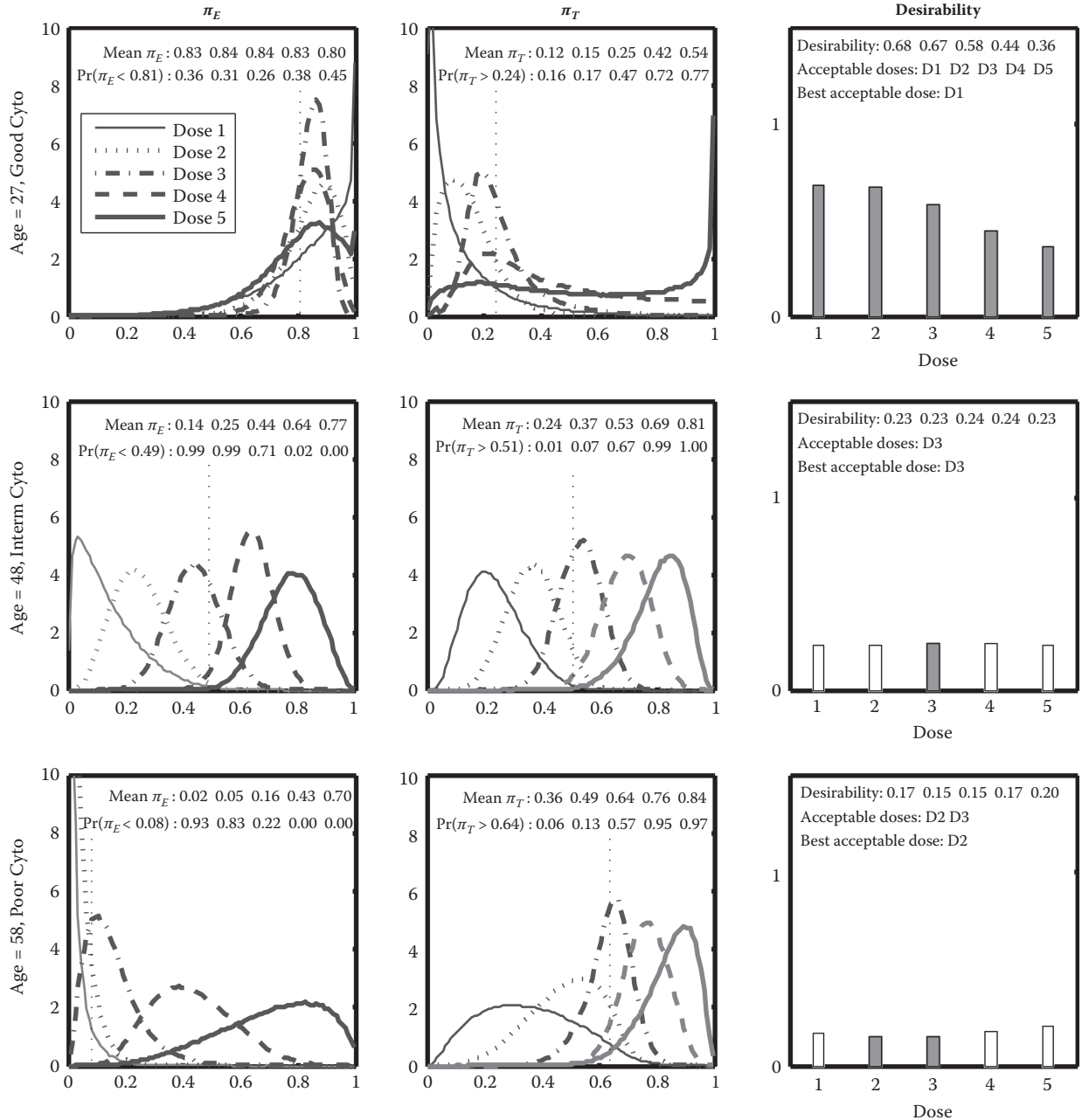


Fig. 3 Illustration of the posteriors of $\pi_E(X_j, \mathbf{Z})$ and $\pi_T(X_j, \mathbf{Z})$ corresponding desirability values $\delta(X_j, \mathbf{Z}) = r_c\{\pi(X_j, \mathbf{Z})\}$ for each dose and $\mathbf{Z} = (27, \text{Good})$, $(48, \text{Interm})$ and $(58, \text{Poor})$. In the third column, the desirabilities of acceptable and unacceptable doses are represented by shaded and unshaded rectangles, respectively

covariate-adjusted method assuming that no $x - \mathbf{Z}$ interactions, $\gamma = 0$, denoted by $\mathbf{Z}$ -NO INT; and (iii) the simplified version of the method that ignores covariates entirely, denoted by NO $\mathbf{Z}$. To compare the methods for each $\mathbf{Z}$ under each scenario, the followed statistic was used. Given $\mathbf{Z}$ and D_n , unless $A_n(\mathbf{Z})$ is the empty set, let $x(\mathbf{Z}, D_n)$ denote the covariate-specific dose that is selected. The true value of $\pi(x(\mathbf{Z}, D_n), \mathbf{Z})$ under the scenario and thus $d_c\{\pi(x(\mathbf{Z}, D_n), \mathbf{Z})\}$ at the selected dose for the given $\mathbf{Z}$ are known. A statistic that compares the selected $x(\mathbf{Z}, D_n)$ to

the best possible dose that could have been selected for a patient with covariates $\mathbf{Z}$ is

$$\rho(\mathbf{Z}) = \frac{d_c\{\pi(x(\mathbf{Z}, D_n), \mathbf{Z})\}}{\max_{j=1, \dots, 5} d_c\{\pi(x_j, \mathbf{Z})\}}. \quad (8)$$

This takes on values between 0 and 1, with $\rho(\mathbf{Z}) = 1$ if the best possible dose was selected.

Under Scenario 1, there are no $x - \mathbf{Z}$ interactions ($\gamma = 0$) the desirability $d_c(\pi_E, \pi_T)$ increases with x for each $\mathbf{Z}$,

and for each x the values of $d_c(\pi_E, \pi_T)$ change dramatically with $\mathbf{Z}$. This scenario is similar to what was seen in the historical data. Scenario 2 is obtained from Scenario 1 by adding the $x - \mathbf{Z}$ interactions $\gamma_{E,2} = \gamma_{E,Good} = .6, \gamma_{E,3} = \gamma_{E,Poor} = -.8, \gamma_{T,2} = \gamma_{T,Good} = -.6, \gamma_{T,3} = \gamma_{T,Poor} = .8$. This says that, for higher x , the $x - \mathbf{Z}$ interactions lead to larger $\pi_E(x, \mathbf{Z}, \theta)$ and smaller $\pi_T(x, \mathbf{Z}, \theta)$ for patients with Good Cyto, whereas the $x - \mathbf{Z}$ interactions have the opposite effect on patients with Poor Cyto. These effects are illustrated strikingly by Fig. 4, which shows that higher doses are more desirable for patients with Good or Intermediate Cyto but lower doses are more desirable for patients with Poor Cyto. In Scenario 3, $\pi_T(x, \mathbf{Z}, \theta)$ and $d_c(\pi_E, \pi_T)$ are non-monotone in x , with $d_c(\pi_E, \pi_T)$ largest at the middle

dose. Scenario 4 is obtained from Scenario 3 by adding the same $x - \mathbf{Z}$ interaction parameters as in Scenario 2.

The simulation results are presented graphically in Fig. 5. The $\mathbf{Z}$ -INT method does a very reliable job of selecting the most desirable doses within patient prognostic subgroups. The results for Scenarios 2 and 4 show that the method very reliably detects $x - \mathbf{Z}$ interactions and selects the doses that are most desirable within each subgroup. In contrast, under the $x - \mathbf{Z}$ interaction Scenarios 2 and 4, the $\mathbf{Z}$ -No INT method has ρ values substantially lower than those for $\mathbf{Z}$ -INT, and the ρ values for the No $\mathbf{Z}$ method are extremely low for (58, Poor) patients. Thus, either ignoring $x - \mathbf{Z}$ interactions or ignoring covariates produces a method with greatly inferior properties.

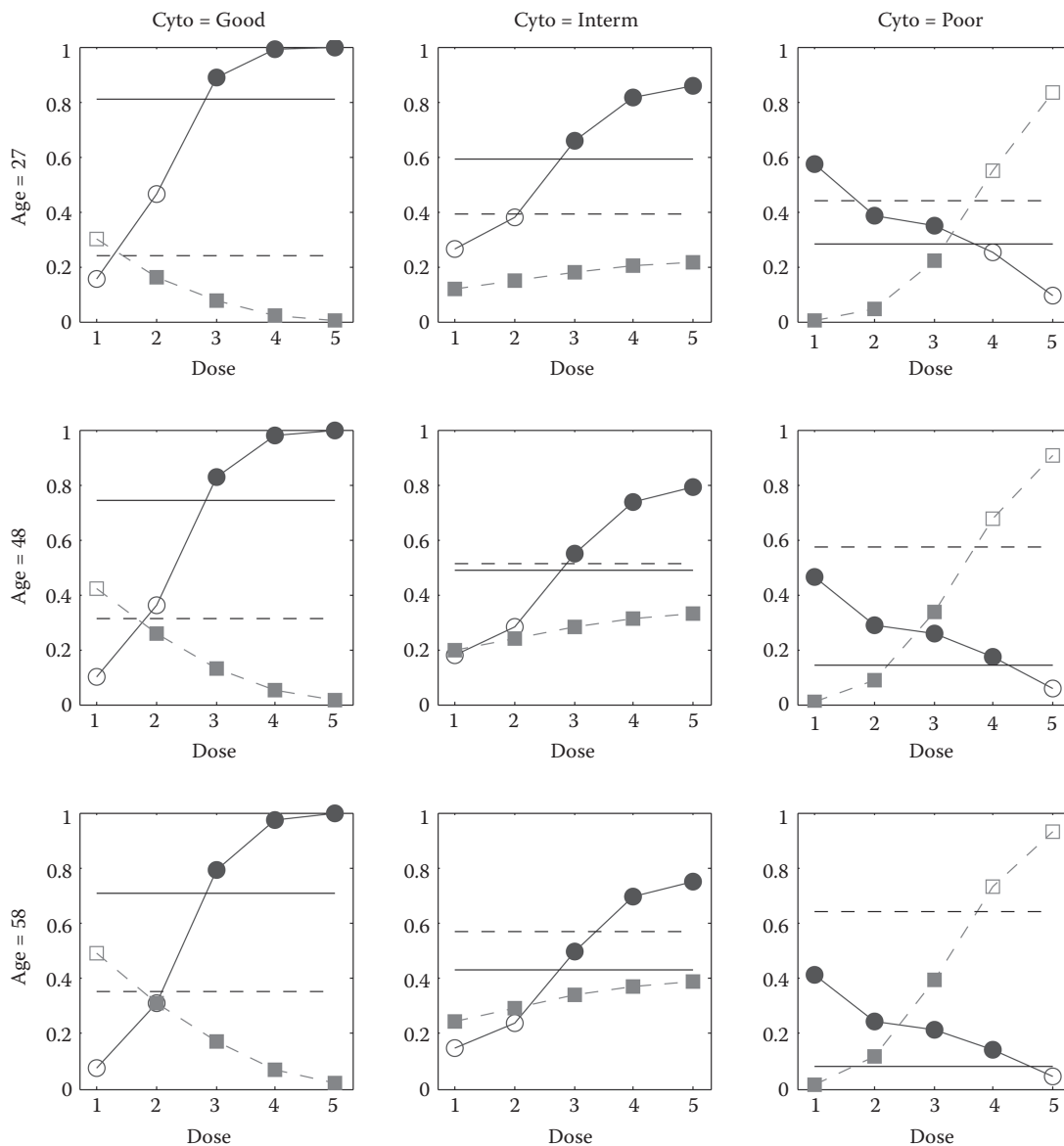


Fig. 4 Simulation Scenario 2 as represented by $\pi_E(X_j, \mathbf{Z})$ (shown as circles) and $\pi_T(X_j, \mathbf{Z})$ (shown as squares) for each of nine representative covariate vectors. Solid points denote acceptable doses and open points denote unacceptable doses

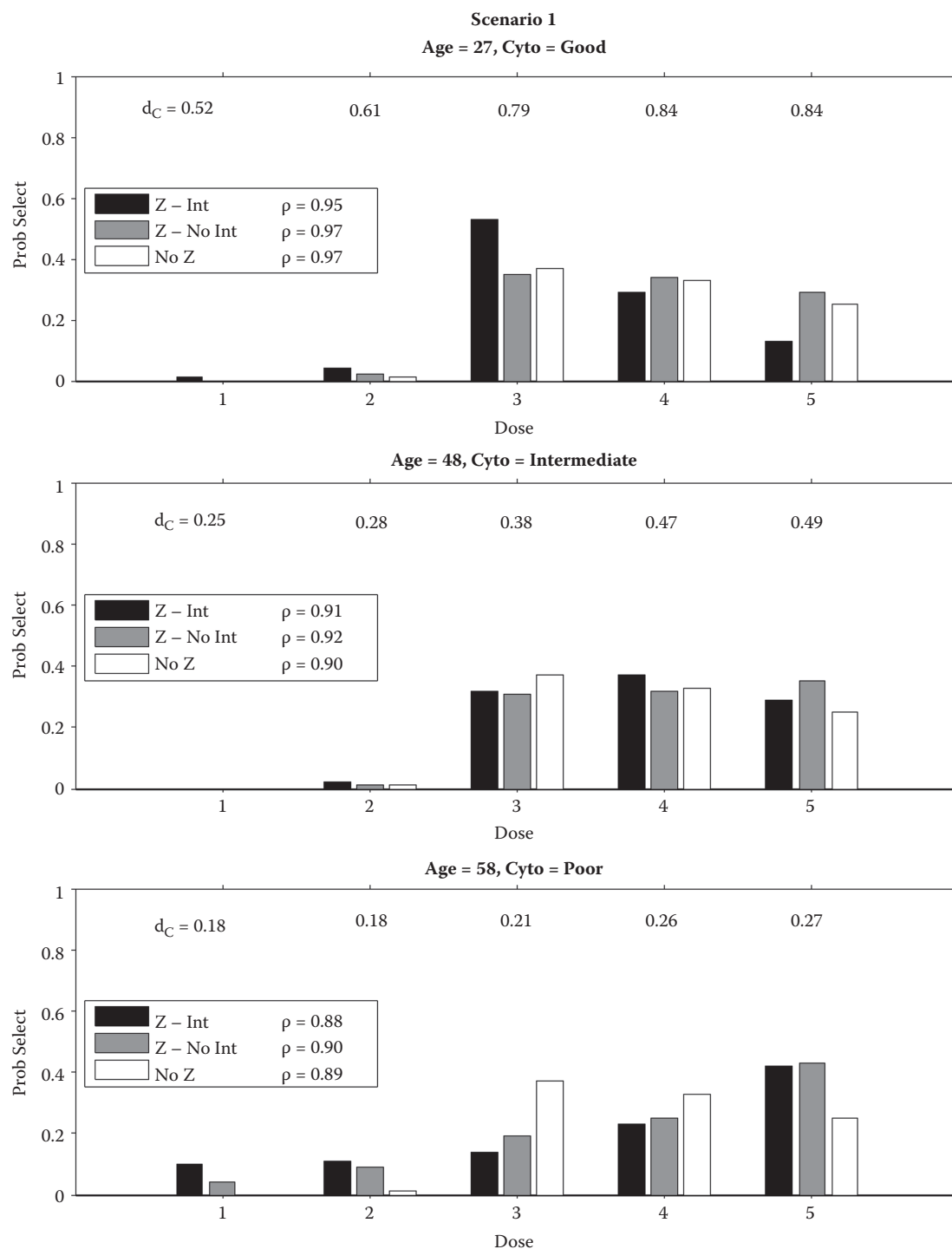


Fig. 5 (Continued) Simulation results. The desirability d_C of each dose for each Z under each scenario is given along the top of each graph. For each Z , the rectangles representing dose selection percentages are dark-shaded for the Z -INT method, light-shaded for the Z -NO INT method and open for the NO Z method

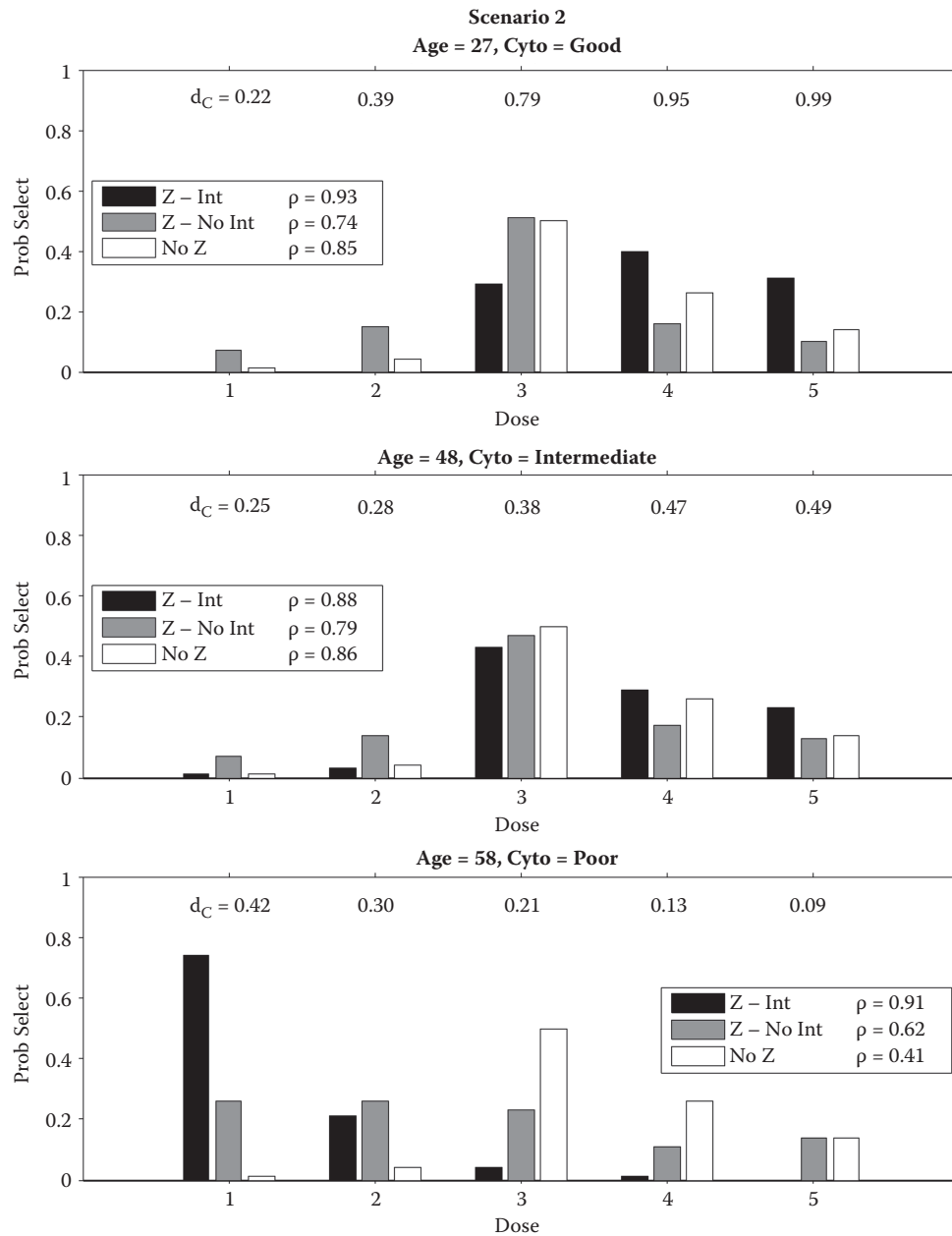


Fig. 5 (Continued) Simulation results. The desirability d_C of each dose for each Z under each scenario is given along the top of each graph. For each Z , the rectangles representing dose selection percentages are dark-shaded for the Z -INT method, light-shaded for the Z -NO INT method and open for the NO Z method

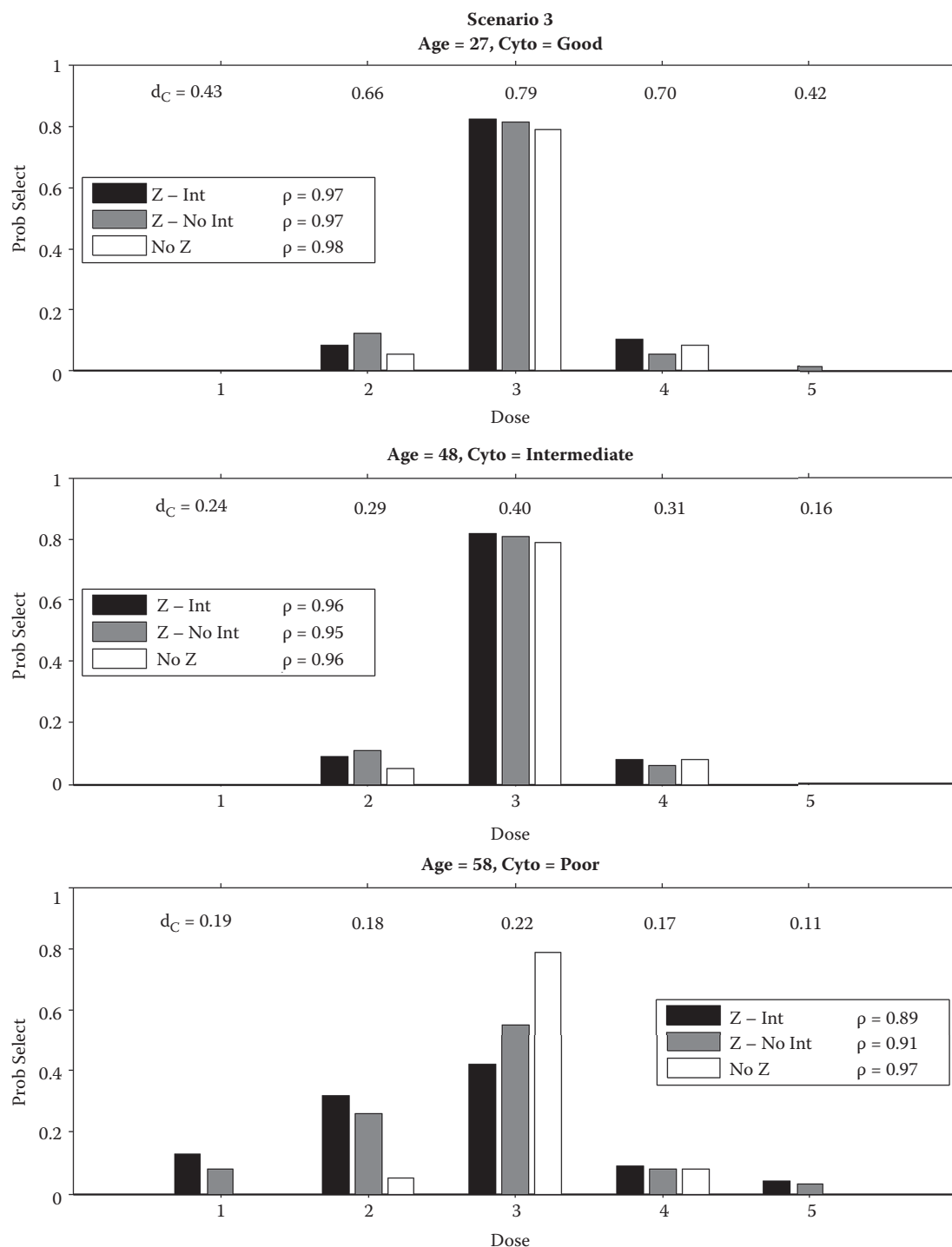


Fig. 5 (Continued) Simulation results. The desirability d_C of each dose for each Z under each scenario is given along the top of each graph. For each Z , the rectangles representing dose selection percentages are dark-shaded for the Z -INT method, light-shaded for the Z -NO INT method and open for the NO Z method

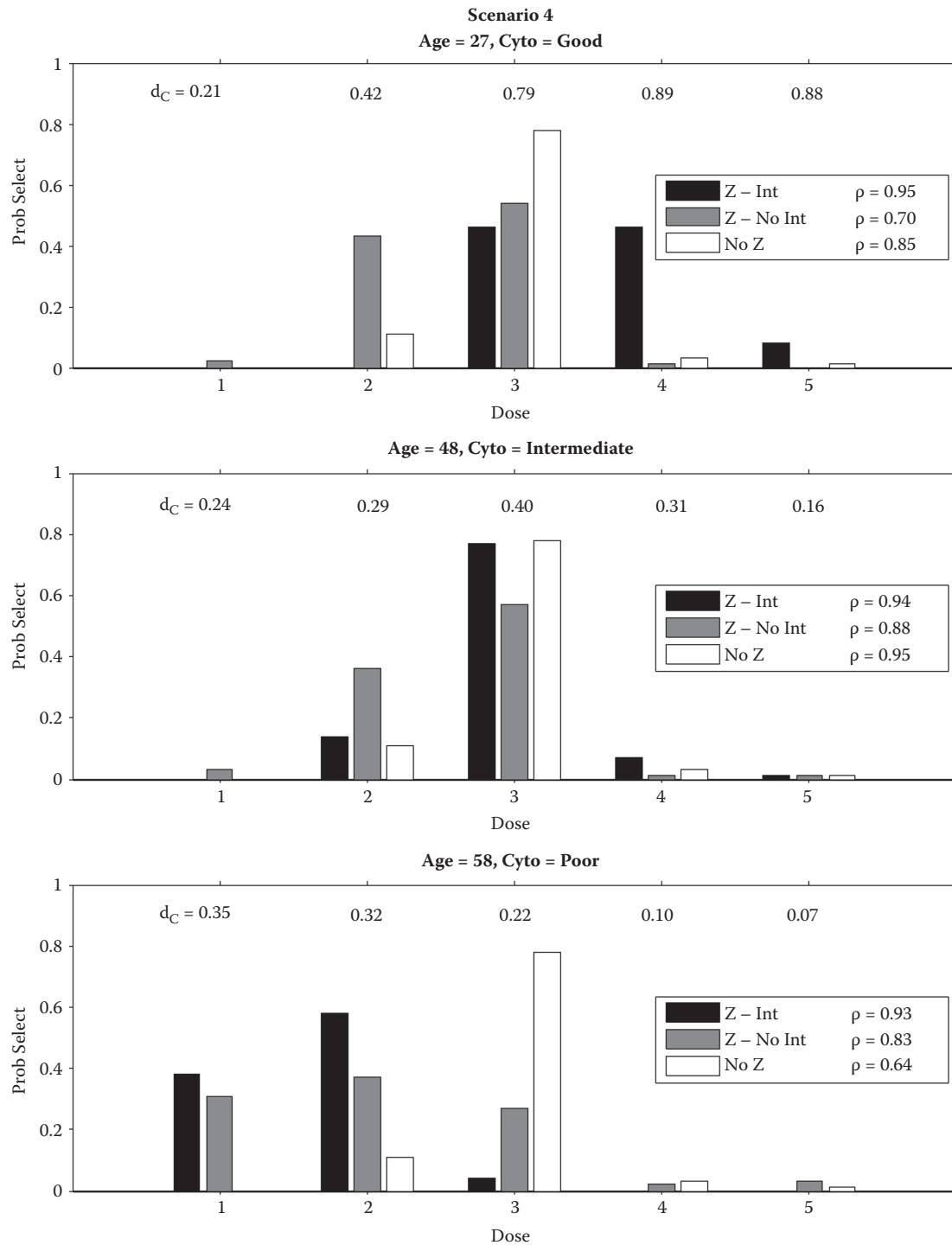


Fig. 5 (Continued) Simulation results. The desirability d_C of each dose for each Z under each scenario is given along the top of each graph. For each Z , the rectangles representing dose selection percentages are dark-shaded for the Z -INT method, light-shaded for the Z -NO INT method and open for the NO Z method

CONCLUSIONS

The phase I/II design described here chooses doses based on efficacy and toxicity while accounting for each patient's prognostic covariates. The method is very complex, and requires a substantial effort from both the statistician and the physicians. Computer simulations show that this complexity is justified, since the method provides substantial improvements over simpler versions that either ignore $x - Z$ interactions or ignore covariates entirely.

ARTICLES OF RELATED INTEREST

Adaptive Designs; Adaptive Dose-Finding Based On Efficacy-Toxicity Tradeoffs; Bayesian Statistics; Cancer Trials; Clinical Endpoint; Clinical Trial Simulation; Clinical Trials; Dose Response Study Design; Maximum Tolerable Dose for Cancer Chemotherapy; Multiple Endpoints; Phase I Cancer Clinical Trials.

ACKNOWLEDGMENT

This research was partially supported by NIH grant RO1 CA 83932.

REFERENCES

1. Storer, B.E. Design and analysis of phase I clinical trials. *Biometrics* **1989**, *45*, 925–937.
2. Lin, Y.; Shih, Weichung J. Statistical properties of algorithm-based designs for phase I cancer clinical trials. *Biostatistics* **2001**, *2*, 203–215.
3. O'Quigley, J.; Pepe, M.; Fisher, L. Continual reassessment method: A practical design for Phase I clinical trials in cancer. *Biometrics* **1990**, *46*, 33–48.
4. Babb, J.S.; Rogatko, A.; Zacks, S. Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Stat. Med.* **1998**, *17*, 1103–1120.
5. Mick, R.; Ratain, M.J. Model-guided determination of maximum tolerated dose in phase I clinical trials: Evidence of increased precision. *J. Natl. Cancer Inst.* **1993**, *85*, 217–223.
6. Wijesinha, M.C.; Piantadosi, S. Dose-response models with covariates. *Biometrics* **1995**, *51*, 977–987.
7. Ratain, M.J.; Mick, R.; Janisch, L.; Berezin, F.; Schilsky, R.L.; Vogelzang, N.J.; Kut, M. Individualized dosing of amonafide based on a pharmacodynamic model incorporating acetylator phenotype and gender. *Pharmacogenetics* **1996**, *6*, 93–101.
8. Babb, J.S.; Rogatko, A. Patient specific dosing in a phase I cancer trial. *Stat. Med.* **2001**, *20*, 2079–2090.
9. Cheng, J.D.; Babb, J.S.; Langer, S.; Aamdal, S.; Robert, F.; Engelhardt, L.P.; Fernberg, O.; Schiller, J.; Forsberg, G.; Alpaugh, R.K.; Weiner, L.M.; Rogatko, A. Individualized patient dosing in phase I clinical trials: the role of escalation with overdose control in PNU-214936. *J. Clin. Oncol.* **2004**, *22*, 602–609.
10. Gooley, T.A.; Martin, P.J.; Fisher, L.D.; Pettinger, M. Simulation as a design tool for phase I/II clinical trials: An example from bone marrow transplantation. *Control. Clin. Trials* **1994**, *15*, 450–462.
11. Thall, P.F.; Russell, K.T. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics* **1998**, *54*, 251–264.
12. O'Quigley, J.; Hughes, M.D.; Fenton, T. Dose-finding designs for HIV studies. *Biometrics* **2001**, *57*, 1081–1029.
13. Braun, T. The bivariate continual reassessment method: extending the CRM to phase I trials of two competing outcomes. *Control. Clin. Trials* **2002**, *23*, 240–256.
14. Ivanova, A. A new dose-finding design for bivariate outcomes. *Biometrics* **2003**, *59*, 1001–1007.
15. Thall, P.F.; Cook, J.D. Dose-finding based on efficacy-toxicity trade-offs. *Biometrics* **2004**, *60*, 684–693.
16. Bekele, B.N.; Shen, Y. A Bayesian approach to jointly modeling toxicity and biomarker expression in a phase I/II dose-finding trial. *Biometrics* **2004**, *60*, 343–354.
17. Thall, P.F.; Cook, J.D. Adaptive dose-finding based on efficacy-toxicity trade-offs; *Encyclopedia of Biopharmaceutical Statistics*, 2nd Ed.; 2006; 1–5.
18. Thall, P.F.; Cook, J.D.; Estey, E.H. Adaptive dose selection using efficacy-toxicity trade-offs: illustrations and practical considerations. *J. Biopharm. Stat.*; Taylor & Francis Group, LLC: London, **2006**, *16*, 623–638.
19. Thall, P.F.; Nguyen, H.; Estey, E.H. Patient-specific dose-finding based on bivariate outcomes and covariates. *Biometrics* **2008**, *64*, 1126–1136.
20. Nelsen, R.B. *An Introduction to Copulas*; Springer-Verlag: New York, 1999.
21. Morita, S.; Thall, P.F.; Mueller, P. Determining the effective sample size of a parametric prior. *Biometrics* **2008**, *64*, 595–602.

Covariates: Adjustment for

Thomas Permutt

U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.

Abstract

The techniques of analysis of covariance are employed in three mathematically similar but conceptually very different kinds of problem. Examples of all three kinds arise in connection with the development of pharmaceutical products. This entry is focused on the adjustment for covariates in an experimental setting.

INTRODUCTION

The techniques of analysis of covariance are employed in three mathematically similar but conceptually very different kinds of problem. Examples of all three kinds arise in connection with the development of pharmaceutical products.

In the first case, a regression model is expected to fit the data well enough to serve as the basis for prediction. In testing the stability of a drug product, for example, the potency may be modeled as a linear function of time, and the possibility of different lines for different batches of the product needs to be allowed for. The purpose of the statistical analysis is to ensure, with a stated degree of confidence, that the potency at a given time will be within given limits.

The second and perhaps widest application of analysis of covariance is in observational studies, such as arise in the postmarketing phase of drug development. It may be desired, for example, to study the association of some outcome with exposure to a drug. It is necessary to adjust for covariates that may be systematically associated both with the outcome and with the exposure and so induce a spurious relationship between the outcome and the exposure. In such studies the unexplained variation is typically high, so the model is not expected to fit the individual observations well. It must, however, include all the important potential confounders and must have at least approximately the right functional form, if a causal relationship, or the absence of one, between the outcome and the exposure is to be inferred.

The third kind of application of analysis of covariance, although the first historically,^[1] is to randomized, controlled experiments such as clinical trials of the efficacy of new drugs. In such experiments, adjustment for covariates is optional in a sense, because the validity of unadjusted

comparisons is ensured by randomization. Adjustments properly planned and executed, however, can reduce the probabilities of inferential errors and so help to control the size, cost, and time of clinical trials.

The modeling problem is straightforward, well covered in textbooks, and, strictly speaking, not a matter of “adjustment.” The observational problem, in contrast, is essentially intractable from the standpoint of formal statistical inference; but heuristic methods have had wide application and discussion. We focus here on the adjustment for covariates in the experimental setting. This problem has had relatively little attention in the literature, partly because early writings^[1] are largely complete, correct, and still sufficient. Unfortunately, the more recent literature on modeling and on observational studies has been misapplied to the experimental problem. Either a well-fitting model is thought to be required, as in the first problem, or the analysis is supposed to be heuristic, as in the second. In fact, a rigorous theory of analysis of covariance in controlled experiments can be developed, even in the absence of a good model for the covariate effects.

ADJUSTING FOR BASELINE VALUES

Consider the case of a randomized trial of two treatments, with a continuous measure of outcome (Y) that is also measured at baseline (X). If the populations are normal or the samples are large, the treatments might be compared by a two-sample t -test on the difference in mean outcome $\bar{Y}$. Alternatively, the change from baseline, $Y-X$, might be analyzed in the same way. The difference between groups in $\bar{Y}$ and the difference between groups in $\bar{Y}-X$ have the same expectation, because the expected difference between groups in $\bar{X}$ is zero. We therefore have two unbiased estimators of the same parameter. They have different variances, according to how well the baseline predicts the outcome. If the variances (within treatment groups) of baseline and outcome

The opinions expressed are those of the author and not necessarily of the U.S. Food and Drug Administration.

are the same and the correlation is ρ , then the standard errors are in the ratio $(2 - 2\rho)^{1/2}$. The adjusted estimator is better if $\rho > 0.5$.

Of course, there is no need to choose. The average of the two estimators has standard error proportional to $(1.25 - \rho)^{1/2}$, which is less than either of the two whenever $0.25 < \rho < 0.75$. This average can be written as the difference between treatment groups in $\bar{Y} - 0.5\bar{X}$. So $\bar{Y} - 0.5\bar{X}$ is a less variable measure of outcome than either the mean raw score $\bar{Y}$ or the mean difference from baseline $\bar{Y} - \bar{X}$, whenever the correlation is between 0.25 and 0.75. This can, but need not, be viewed as fitting parallel straight lines with slope 0.5 to the two groups and measuring the vertical distance between them.

Naturally, there is no need to choose 0.5 either. The difference in group means of any statistic of the form $Y - \beta X$ can be used to estimate the treatment effect. The smallest variance, and so the most sensitive test, is achieved when β happens to coincide with the least-squares common slope, but the variance does not increase steeply as β moves away from this optimal value. Thus, even a very rough a priori guess for β is likely to perform better than either of the special cases $\beta = 0$ (no adjustment) and $\beta = 1$ (subtract the baseline).

Finally, there is no need to guess. The least-squares slope, calculated from the data, can be used for β , without any consequences beyond the loss of a degree of freedom for error. Asymptotic theory for the resulting adjusted estimator of the treatment effect was given by Robinson,^[2] and an exact, small-sample theory by Tukey.^[3]

In general, then, the best way to adjust for a baseline value is neither to ignore it nor to subtract it, but to subtract a fraction of it. The fraction will be estimated from the data, simultaneously with the treatment effect, by analysis of covariance. There is no need to check the “assumption” that the outcome is linearly related to the baseline value, because this assumption plays no role in the analysis. If it did, not only the analysis of covariance would be tainted: After all, the unadjusted analysis also “assumes” a linear relationship, with slope 0, and the change-from-baseline analysis “assumes” a slope of 1.

OTHER COVARIATES

Any single, prespecified covariate can be adjusted for in much the same way as a baseline measurement of the outcome variable. That is, the mean of a linear function $Y - \beta X$ may be compared across treatment groups, the coefficient β being estimated, simultaneously with the treatment effect, by least squares. Again, the much-tested assumption of a linear relationship between Y and X is superfluous. Two other critical assumptions are sometimes neglected, however.

First, the covariate must be unaffected by treatment. While it is possible to give an interpretation of analysis of covariance adjusting for intermediate causes, this interpretation is not often useful in clinical trials. Any covariate measured before randomization is acceptable. With care, some covariates measured later may be assumed to be unaffected by treatment: the weather, for example, in a study of seasonal allergies. It may be noted that, while analysis of covariance is not usually appropriate for variables in the causal pathway, some of the advantages of analysis of covariance are shared by instrumental-variables techniques^[4] that are appropriate.

Second, the covariate is assumed to be prespecified. Model-searching procedures are unavoidable in observational studies, for there are typically many potential confounding variables whose effects must be considered and eliminated if necessary. Alarming little is known about the statistical properties of such procedures, however, and what is known is not generally encouraging. It is usual, although unjustifiable, to ignore the searching process in reporting the results, presenting simply the chosen model, its estimates, and its optimistic estimates of variability.

Randomized trials are radically different from observational studies in this respect. There is no confounding, because a covariate cannot be systematically associated with treatment if it is not affected by treatment and if treatment is assigned at random. The purpose of analysis of covariance in randomized studies is to reduce the random variability of the estimated treatment effects by eliminating some of what would otherwise be unexplained variance in the observations. This difference has implications for the choice of covariates, which will be discussed in the next section.

CHOICE OF COVARIATES

Whereas a confounder in an observational study is a variable correlated both with the outcome and with the treatment, a useful covariate in a randomized trial is a variable correlated just with the outcome. The greater the absolute correlation, the more the reduction in residual variance and so also in the standard error of the estimated treatment effect. This benefit is realized whether the treatment groups happen to be balanced with respect to the covariate or not. It is neither necessary nor useful, therefore, to choose covariates retrospectively, on the basis of imbalance.^[5]

It is accordingly safe to prespecify, in the protocol for a randomized trial, a covariate, or a few covariates, unaffected by treatment but likely to be correlated with the outcome. Analysis of covariance, adjusting for these covariates, may then be carried out and relied on, without

any justification after the fact. The probability of type I error will be controlled by significance testing, and the probability of type II error will be less than if covariates were not used.

The improvement, however, depends on the correlations (and partial correlations) between the covariates and the outcome, and these may not be perfectly known ahead of time. It might therefore seem advantageous to determine the correlations for some candidate covariates with the data in view, and select a subset that explains a high proportion of the variance of the outcome. With care, it is possible to specify an unambiguous algorithm for selecting a model and to control the probability of type I error.^[3] It is not known, however, whether such procedures have any advantage with respect to type II error over simply prespecifying the model. In practice, in critical efficacy trials the relevant covariates will often be apparent in advance; and when they are not, it may not be any easier or better to specify a set of candidates and an algorithm for choosing among them than to specify a single model.

The properties of models with large numbers of covariates are not well understood. Various rules of thumb relating the number of variables to the sample sizes have been given, but none has any compelling theoretical justification. Furthermore, searches in large sets of potential models probably share some of the defects of models with many covariates, even if the chosen model has only a few covariates.

NONLINEAR MODELS

The word *linear* in the context of the analysis of covariance may be understood in two senses. In many applications, the model is linear in the covariates. However, a model with polynomial or other nonlinear covariate effects is still linear in the *coefficients*, and the least-squares estimators are consequently linear functions of the outcome measurements, so the theory of the general linear model applies. In contrast, logistic, proportional-hazards, and Poisson regression models all involve covariates in a more fundamentally nonlinear way.

Nonlinear covariate effects can be added to an analysis of covariance without difficulty. The most common examples are the 1/0 variables used to represent categorical covariates, but polynomial, logarithmic, exponential, and other functions may sometimes be useful. It is important to bear in mind, however, that in randomized trials the purpose of the covariate model is to reduce unexplained variance. Thus, nonlinear terms should be introduced when they are expected to explain substantial variance in the outcome, and not simply because it is feared that the “assumption” of a linear relationship between the outcome and the covariate may be violated.

Conversely, trials with outcomes that are successes or failures, survival times, or small numbers of events are analyzed by methods that are nonlinear in the second sense. Recent theoretical developments (the generalized linear model) and computer programs have tended to emphasize the analogies between these methods and the linear model. Some of the same principles undoubtedly apply when such methods are used to analyze randomized trials. For example, if a model selection procedure is used, it is vital to understand the statistical properties of the procedure as a whole, rather than simply to report the nominal standard errors and *p*-values of the model that happens to be chosen. On the other hand, the similarity in form may conceal important differences in mathematical structure between linear and nonlinear models, and the linear results must not be casually assumed to have nonlinear analogs. It is not clear, for example, that the robustness of linear models against misspecification in randomized trials carries over to all the nonlinear cases.

INTERACTION

If the difference in mean outcome between treatments changes as a covariate changes, there is said to be a treatment-by-covariate interaction. In a drug trial, such a finding would have important implications. In the extreme case, the treatment effect might change direction as the covariate changed. That is, a drug that was beneficial in one subset of patients, identified by the covariate, would be harmful in a different subset. Clearly such a drug would be “effective.” Equally clearly, for such a drug to be useful, the populations in which it was beneficial and harmful would need to be characterized. In less extreme cases, where the magnitude but not the direction of the treatment effect changes, considerations of risk and benefit might also make it very desirable to estimate the effect in different subgroups.

The question of interaction often arises in connection with analysis of covariance, but it really has little to do with adjustment for covariates. Everything in the preceding paragraph is equally true whether the covariate in question is adjusted for, ignored, or even unmeasured. Furthermore, if the treatment main effect is to be estimated, it is still better to estimate it by analysis of covariance, even without an interaction term, than by the unadjusted difference in means. As with the “assumption” of linearity, the analysis of covariance is not invalidated by violation of the “assumption” of parallelism, for this assumption plays no role in the analysis. Also, as with linearity, if this assumption were crucial, its failure would taint as well the unadjusted analysis, which also “assumes” parallel regressions of the outcome on the covariate, but forces them to have slope 0.

The possibility of interaction should be taken into account whenever it appears at all probable that different groups may respond differently. The reason for this is practical and concerns the interpretation and application of the results of a successful trial. However, the presence of interaction or, what is more common, the inability to rule interaction in or out with confidence, should not be seen as invalidating analysis of covariance nor, especially, as a reason to prefer unadjusted analysis.

REFERENCES

1. Fisher, R.A. *Statistical Methods for Research Workers*, 14th Ed.; Oliver and Boyd: Edinburgh, 1970; 272–286.
2. Robinson, J. J. R. *Stat. Soc., Ser. B* **1973**, 35, 368–376.
3. Tukey, J.W. *Control. Clin. Trials* **1993**, 14, 266–285.
4. Angrist, J.D.; Imbens, G.W.; Rubin, D.B. *J. Am. Stat. Assoc.* **1996**, 91, 444–455.
5. Permutt, T. *Statist. Medd.* **1990**, 9, 1455–1462.

Crossover Design

Stephen Senn

Department of Statistics, University of Glasgow, Glasgow, U.K.

Abstract

Crossover trials are potentially extremely efficient designs for studying the effects of treatment. The focus of this entry is on the types of projects for which crossover design is an efficient method.

Keywords: Carryover; Crossover; Placebo-controlled trial.

INTRODUCTION

“A cross-over trial is one in which individual subjects are given sequences of treatments with the object of studying differences between individual treatments (or subsequences of treatments).”^[1]

Crossover trials are potentially extremely efficient designs for studying the effects of treatment. They are, however, not suitable for all disease indications and they can be affected by carryover, a phenomenon that may make it impossible to produce unbiased estimates except by sacrificing the efficiency crossover designs are employed to provide. This important issue will be dealt with extensively below. First, however, some general features of such trials, and possible uses, will be reviewed.

OVERVIEW

The most studied form of crossover is the so-called AB/BA crossover in which patients are treated in two periods and are randomized to receive either treatment A followed by treatment B or vice versa.

Example 1. A trial was carried out to compare the effects in Swedish schoolchildren suffering from asthma of a single dose of 200 μg salbutamol and 12 μg formoterol, both given by metered-dose inhaler.^[2] Fourteen children were randomized in equal numbers to one of two sequences: either formoterol followed by salbutamol or salbutamol followed by formoterol. A washout period of at least 2 days was observed between treatments. The principal outcome measure was peak expiratory flow in 1 second (PEF).^[1]

More complicated crossover designs are often used, however, and designs in which patients receive three, four, or five treatments in as many periods are not uncommon.

Example 2. A placebo-controlled trial in migraine was run in Sweden and Finland to estimate and compare the effects of two doses (50 and 100 mg) of the potassium salt of diclofenac.^[3] Each of 72 patients received each treatment

once and once only. Patients were randomized to receive one of the six possible sequences of the three treatments. Patients were not to use the first treatment until they had observed a treatment-free and attack-free period of at least 1 week. Thereafter they were to treat the first three attacks with the treatments provided in the order indicated. The main outcome measure was a visual analogue pain score two hours after treatment.^[1]

Crossover trials are of greater importance for the biopharmaceutical statistician than is generally the case for fellow medical statisticians working outside of the industry. Some typical applications in drug development are:

- Phase I
 - Pharmacokinetic studies to establish concentration time profiles.
 - Bioequivalence.^[4]
 - Food interaction studies.
 - Dose proportionality.
 - Dose-escalation studies for investigating maximum tolerated dose.
- Phase II
 - Studies of pharmacodynamic response.
 - Dose finding.
 - Parallel assay.
- Phase III
 - Specialist scientific studies.
 - Determination of individual response.
- Phase IV
 - Studies of patient preferences.

Example 3. This is an example of a pharmacodynamic parallel assay in asthma.^[5] Three single doses (6, 12, and 24 μg) for each of two formulations of formoterol to be given by inhalation in the form of a dry powder were compared to placebo. It was not possible to treat patients in more than five periods. Hence, an incomplete blocks design in five periods and 21 sequences was chosen, each patient receiving five of the seven treatments. The trial was

carried out in 15 centers in four countries and 161 patients were recruited. The main outcome variable was the area under the forced expiratory volume in 1 second (FEV₁) curve over 12 hours.

Crossover trials are also used in preclinical work involving animals. This topic is not covered here. Crossover trials are rarely used as pivotal phase III studies in support of a drug application. One reason is that drug regulatory agencies are wary of crossover trials because of the potential danger of carryover. This issue is dealt with in detail in the section “Carryover.” An even more important reason is that a principle advantage of crossover trials is to reduce the number of patients in a trial. However, for most drug developments, more patients are needed to satisfy the regulator as regards the tolerability of the product than are needed to demonstrate its efficacy, even if a parallel group trial is chosen. There is thus less incentive for the sponsor to use crossover trials in phase III. The exception as regards sample size requirements would be trials in which survival is the outcome, but these are, of course, in any case not suitably run as crossovers nor, indeed, are any trials in which cure or death is the outcome.

In fact, suitable indications for crossovers are chronic diseases in which the patients are relatively stable. Such indications include asthma, mild or moderate hypertension, rheumatism, migraine, sleep disturbances, angina, and epilepsy.

It is not the disease alone, however, that determines the suitability of a crossover. Also relevant is the nature of the treatment. Quickly reversible treatments are more suitable than those for which the effect is more persistent. Thus, for example, in asthma, beta-agonists are easier to study using crossover trials than are steroids. Other considerations also apply. An important distinction in drug development is between single-dose studies, which often involve pharmacodynamic measures and may study onset and duration of action, and multidose studies in which, more usually, the therapeutic steady-state effect of treatments is studied. Crossover trials may be used for both but are generally more suited to the former.

CARRYOVER

The outstanding problem of crossover trials is regarded by many as being carryover. Carryover has been defined as, “the persistence (whether physically or in terms of effect) of a treatment applied in one period in a subsequent period of treatment” (Ref. [1]). The simplest example of carryover is where the half-life of a treatment has been underestimated and some relevant concentration of the previous treatment is still present at a time when another is being measured. However, except for bioequivalence studies in which drug concentration itself is the outcome, it is the pharmacodynamic time course rather than the pharmacokinetic course that is relevant. Much play has also been

made in the literature of so-called psychological carryover, where perhaps some memory of the therapeutic experience under the previous treatment affects the patient’s current judgement. However, it should not be forgotten that in a double-blind trial, to the extent that blinding is successful, the origin of all carryover, however indirectly, must ultimately be pharmacological.

Where it occurs, the consequence of carryover is that the trialist will measure the combined (in some cases partial) effects of two or more treatments. (S)he may be unaware that this is happening and this may lead to biased assessment and even where detected the disentangling of individual effects may be difficult. The most appropriate way to deal with carryover is with washout periods. These may be “passive” in that no treatment is given or “active.” The latter strategy occurs quite commonly in multidose therapeutic trials. Here patients may be switched over almost immediately to the subsequent treatment. The previous treatment will, of course, be being eliminated by the patient’s body during the time that the new treatment is being given. Provided measurement of the effects of the new treatment is delayed until this process is complete, the problem of carryover is dealt with.

If, however, as is sometimes the case in single-dose pharmacodynamic studies, it is desired to study onset of action, then a passive washout is necessary. The issue of modeling carryover is considered in due course below.

ANALYSIS USING THE GENERAL LINEAR MODEL

In many cases the primary outcomes for a crossover trial will be continuous. For example, FEV₁ is commonly used in asthma and diastolic or systolic blood pressure in hypertension. As explained above, crossover trials are often more relevant to studying pharmacodynamic rather than therapeutic outcomes. In particular, except in rare cases to be discussed below, survival is not a relevant outcome for a crossover trial. A general linear model in which disturbance terms are assumed to be normally distributed will often be a suitable framework for analysis.

Crossover trials share a feature of fractional factorials that can make their representation in a linear model somewhat awkward. Consider an AB/BA crossover trial in two patients only. There are two patients, two periods, and two treatments but only four and not eight observations. A consequence of this is that two of the three factors patient, period, and treatment determine the third. Thus, if patient 1 is allocated to the AB sequence if the current observation is the observation in period 2, it must be under treatment B.

If we ignore the problem of carryover, then, following Jones and Kenward,^[6] a general model for observed values y_{ijk} of a random variable Y_{ijk} from a crossover trial in g sequences, p periods, N patients, and t treatments is given as follows

$$Y_{ijk} = \mu + s_{ik} + \pi_j + \tau_{d[i,j]} + \varepsilon_{ijk} \quad (1)$$

Here, μ is a general mean, s_{ik} is the effect of subject k in sequence i , $i = 1, 2, \dots, g$, $k = 1, 2, \dots, n_i$, π_j is the effect of period j , $j = 1, 2, \dots, p$, $\tau_{d[i,j]}$ is the direct effect of the treatment administered in period j to subjects (either patients or healthy volunteers) in group i , and ε_{ijk} is a stochastic disturbance term.

A number of points may be noted about this model.

- The ε_{ijk} terms are often assumed independent. For designs with three or more periods, however, it may be appropriate to allow for autocorrelation within patients over time.
- The model is overparameterized, but it is identifiable contrasts, in particular those of the treatment parameters, which are of interest.
- The treatment parameter is indexed by a functional subscript $d[i,j]$, which identifies it by position (sequence and period). This is because, as discussed above, crossover designs are a form of fractional design: There are t treatments, N patients, and p periods but only $N \times p$ observations. To complete the specification, the function $d[i,j]$ must be separately defined, say by a table that indicates which treatment is given to which sequence group in a given period.
- There is no term for carryover in the model. Jones and Kenward^[6] propose adding a term $\lambda_{d[i,j-1]}$. This form of carryover has been referred to as simple carryover^[1] and is appropriate if carryover lasts only for one period and depends on engendering treatment only (is *not* modified by the perturbed treatment).
- The term s_k can be treated as fixed or random depending on approach. This may or may not make a difference to the resulting inference regarding treatment contrasts: Whether it does so depends on whether interblock information is present and this in turn depends on the balance of the design, whether autocorrelation is allowed or not and whether a carryover term is fitted.^[7]
- Whether or not the *patient* effect is treated as fixed or random, the model does not allow for a true random effect of *treatment*: the possibility that the effect of treatment can vary from patient to patient. The fixed effect approach is extremely common but it can be of interest on occasion to use a random effects approach, in particular in designs where the number of periods exceed the number of treatments.

A popular approach to analyzing crossover trials within the pharmaceutical industry is to apply ordinary least squares using `proc glm`[®] of SAS[®] and treating the patient effect as fixed. Alternatively, the patient effects are sometimes allowed to be random, and this may be conveniently modeled using `proc mixed`[®] of SAS[®]. Sometimes, where this is done, a factor for the sequence itself is introduced to

force the treatment estimate to be a “within-patient” estimate. If this is done, however, between-patient information cannot be recovered.

This issue can be understood simply by considering an AB/BA crossover in which some patients have dropped out at random after the first period. If a fixed effect is included for patients, only patients who have completed both periods will contribute information. However, direct comparison of patients who have received A only with those who have received B only can be made if only the patient effects are allowed to be random (as they must be in any parallel group trial). This will generally contribute a very small amount of information but can in principle be combined with that from patients who have completed both periods using an appropriate approach to analysis as provided, say, by `proc mixed`[®] of SAS[®].

This is a simple example where so-called interblock information is available to be recovered. There are more complex cases. Usually, if carryover is introduced into the model, it means that interblock information can in principle be recovered. An exception is given by models in which simple carryover only is allowed for and the design chosen is such that simple carryover is orthogonal to the treatment effect. For example, such a design is the four-period, two-treatment design using the sequences AABB, BBAA, ABBA, and BAAA. For incomplete block designs, such as illustrated by example 3 above, interblock information is available to be recovered whether or not carryover is fitted.

A general feature of conventional approaches to analyzing crossover trials is that the residual degrees of freedom for error can exceed the number of patients.^[8] Consider, for example, the degrees of freedom for the analysis of variance for a complete blocks design in N patients, p periods, and p treatments. (This sort of design is very common. Examples 1 and 2 above are cases in point with $N = 14$ and $p = 2$ and $N = 72$ and $p = 3$, respectively.) The degrees of freedom for the analysis of variance may be partitioned as follows:

Source	Degrees of freedom		
	In general	Example 1	Example 2
Patients	$N - 1$	13	71
Periods	$p - 1$	1	2
Treatments	$p - 1$	1	2
Error	$Np - N - 2p + 2$	12	140
Total	$Np - 1$	27	215

Now, in the first example, the degrees of freedom for error are 12 and thus inferior to the number of patients. In the second, however, there are nearly twice as many degrees of freedom for error as patients and this signals that strong assumptions are involved.^[8] An alternative approach to using ordinary least squares under such cases can be to reduce the data for a given patient to a contrast of interest and then analyze these contrasts. A similar problem arises with any crossover trial with more than

two periods and where the same treatment, as may be the case in so-called, “*n* of 1 studies,” is repeatedly given to the same patient, this summary measures approach may be particularly valuable.^[1,8] Alternatively a true random effect (multilevel) model can be applied.

MODELING CARRYOVER

There is an enormous literature on adjusting the analysis of crossover trials to allow for the effect of carryover and a corresponding interest in appropriate efficient designs. Most authors have enthusiastically (and usually uncritically) adopted the simple carryover model^[6,9–11] whereby carryover depends only on the engendering treatment and last for one period only. Of course, for the AB/BA design, there are only two periods and the treatment that follows is completely determined by the treatment that precedes. Hence, the simple carryover model is the only one that needs to be considered. Approaches to dealing with carryover in the case of the AB/BA design will be considered below in due course, but in any case, those interested in designing supposedly more efficient crossover trials have usually considered more complex designs. For these more complex designs, simple carryover may not be realistic.

Consider, for example, the design in two periods in which patients are randomized in equal numbers to four sequences AB, BA, AA, and BB.^[12] We can summarize the responses using eight cell means defined by the product of the four sequences and two periods. If we fit a fixed effect for each patient any estimate of the treatment effect (A–B) can be expressed as a linear combination of the four contrasts produced by taking the period 2 cell mean from a given sequence from the corresponding period 1 value. If carryover is ignored altogether, the relevant weights for these four contrasts are 1/2, –1/2, 0, and 0. The patient effects are eliminated by virtue of having used within-cell differences, the period effects are eliminated because the second weight is the negative of the first and it can also be seen by inspection that the treatment difference is appropriately recovered. Let us call this contrast, C_1 .

In the above scheme, the two sequences AA and BB contribute nothing to the estimate of the treatment effect and would appear to be a complete waste of time and resources. If, however, carryover were present and that from A into B were not the same as from B into A, the treatment estimate so formed would have a bias, $\hat{\alpha}$. If it could be assumed that the carryover from A into A were the same as from B into B, this bias could be estimated using the weights 0, 0, 1/2, –1/2 on the four cell mean differences. Let us call this contrast C_2 . A little reflection shows that this second scheme of weights only estimates a carryover effect: patient, period, and treatment effects are eliminated from it. It thus follows that the difference between the first linear combination and the second, $C_1 - C_2$, provides an unbiased estimate in the presence of “simple” carryover.

This does not mean that such a design is useful. On the contrary it is *not* generally useful. The variance of the estimated treatment effect is 4 times what it would be if all patients had been allocated to the first two sequences only. This is a very high price to pay for unbiasedness and modern approaches to data analysis recognize the importance of variance bias trade-offs.^[13] There is, however, a more important criticism, namely, that if carryover occurred, it is extremely unlikely that the carryover from A into A would be the same as from A into B. For example, A might be an active treatment with a response either at steady state (in a multiple trial) or near the top of a dose-response curve. On the other hand B might be a placebo. The carryover from A into B would thus be more important than from A into A.^[1,14–16] There are, in fact, occasions where correcting for simple carryover can increase the bias of the resulting treatment estimate.^[1,16]

When applying the simple carryover model one can distinguish two rather different cases. The first occurs where the design is such that adjusting for simple carryover has no adverse effect on the variance. This is, for example, the case in the four-period design in two treatments using the four sequences AABB, BBAA, ABBA, and BAAB. Note that in this design A follows A twice and B follows A twice and now consider the 16 cell means for this design defined by the cross-classification of sequence and period. A natural and fully efficient estimate of the treatment effect A–B is obtained by multiplying each cell mean corresponding to an A by 1/8, each cell mean corresponding to B by –1/8, and forming the weighted sum. But if simple carryover occurs two of the cell means under A are affected by carryover from a preceding A and so are two of the cell means by B. The weights associated with these cell means add to zero and thus simple carryover is eliminated automatically.

Under such circumstances, fitting simple carryover in the model only reduces the error degrees of freedom and because these will generally be many, there is little reason not to do so. The danger here is rather in assuming that *because* simple carryover has been fitted that all forms of carryover have been eliminated.

The second case is where the design is not fully efficient in the presence of simple carryover. Here, fitting carryover will increase the variance of the treatment estimate. In some cases, it may even increase the bias. It is extremely doubtful whether the simple carryover model is of any value in such cases. A useful review of approaches to modeling carryover in multiperiod designs is given by Matthews.^[17]

An interesting alternative philosophy for adjusting for carryover involves using pharmacokinetic-pharmacodynamic modeling. This is to regard the carryover effect as being a phenomenon of the same sort as the direct effects of treatment and to use a single pharmacokinetic model for both, then linking this to drug response via a suitable dose response models such as, for example, the eMax model.^[18,19]

TESTING FOR CARRYOVER

Until fairly recently an alternative to adjusting for carryover was popular with applied statisticians. This was to perform a preliminary test of (simple) carryover, use an adjusted estimate (or test) of the treatment effect if this preliminary test was significant and an unadjusted test if (as would more usually be the case) this was not. This was a particularly popular approach to analyzing the AB/BA design and will now be described in that connection.^[20,21]

In the presence of carryover, the only unbiased estimate possible from the AB/BA design is that based on first-period data only. Clearly, if the second-period data are discarded, the remaining data have the structure of a parallel group trial and can be analyzed as such. Note that as discussed above, the implicit assumption is that the patient effect is random. An estimate produced under such circumstances will be referred to in the discussion that follows as PAR.^[22]

In the absence of carryover, an unbiased estimate of the treatment effect is provided by weighting the two cell means from the AB sequence with weights $1/2$ and $-1/2$ and the two cell means from the BA sequence with weights $-1/2$ and $1/2$ and then summing. In the discussion that follows, this particular contrast will be referred to as CROS. In the presence of carryover, however, the bias in CROS is $-\lambda/2$, where λ is the difference between the carryover from A into B and from B into A.

An estimate of the *semi* carryover effect, $\lambda/2$, is given by using weights $1/2$ for both cell means in the first sequence and $-1/2$ for both cell means in the second sequence. In other words, it corresponds to the difference between the mean response over both periods in the AB sequence and the corresponding mean response in the BA sequence. Like PAR, this estimate, which will be referred to as CARRY below, is a between-patient estimate, and so to base analyses upon it, a random patient effect must be assumed. It is unbiased for carryover since each sequence mean reflects both treatments and both periods so that the difference between means can only reflect either differences between patients, which are assumed random, or the order in which treatments were administered.

The general situation is given in Table 1. Note that these three estimates satisfy the relationship $PAR = CROS + CARRY$.^[22]

Table 1 Weights for Estimating Effects for an AB/BA Crossover

Sequence	AB		BA	
Treatment	A	B	B	A
PAR	1	0	-1	0
CROS	1/2	-1/2	-1/2	1/2
CARRY	1/2	1/2	-1/2	-1/2
PAR - (CROS + CARRY)	0	0	0	0

In 1965, in an influential paper, Grizzle^[20] proposed a scheme whereby a preliminary test for CARRY was performed at the 10% level. (This higher than usual nominal level was suggested in view of the low power of this test.) If the result was not significant CROS would be used as the basis for tests of the treatment effect at the 5% level. If it was significant PAR would be used, also at the 5% level. This proposal was also clearly described in a further influential paper by Hills and Armitage.^[21] This scheme was very popular for many years, and some pharmaceutical companies had even written SAS macros to perform this two-stage testing procedure automatically. However, in an extremely important paper published in 1989, Freeman^[23] showed that the procedure as a whole does not have the correct implicit nominal size of 5% but, in the case where there is no carryover, has a Type I error rate that lies between 7% and 9.5%.

The reason is the extremely high correlation between PAR and CARRY. In fact, if anything, Freeman's paper understates the problem with the two-stage approach because the pretest is either irrelevant or the conditional Type I error rate of PAR lies between 25% and 50%.^[1,22] In fact, although general medical statistics textbooks continue to recommend this procedure, none of three monographs devoted to the crossover does so^[1,6,9] and its use in drug development and regulation appears to be being abandoned.

It is possible to adjust the two-stage procedure in a number of ways, for example by carrying out the test using PAR at a lower nominal level: Performing this test at the 0.5% rather than the 5% level deals with the problem, for example.^[24,25] However, under such circumstances any power advantages in the face of carryover disappear when compared to the simpler strategy of always using CROS.

It seems plausible that the problems with the two-stage procedure in the case of the AB/BA designs are likely to be present with strategies involving pretesting for more complex designs. Either the carryover effect will be orthogonal to the treatment effect in which case the only loss involved in adjusting for carryover would be residual degrees of freedom, or the effects are not orthogonal in which case similar problems to that with the AB/BA design may occur. In an investigation of the pretesting strategy, Abeysekera and Curnow^[26] came to the conclusion that always adjusting was the best strategy. They implicitly assumed, however, that the simple carryover model would be correct. In a more general investigation allowing for the possibility of an alternative, steady-state model, whereby carryover from a treatment into itself is zero^[14,15] and also allowing for any arbitrary mixture of steady-state and simple carryover model, Senn and Lambrou^[27] came to the opposite conclusion: that never adjusting was the best policy. The field remains controversial and the best advice to the trialist seems to be not to rely on statistical modeling to eliminate carryover.

BAYESIAN APPROACHES

One interpretation of the difficulties that the above approaches to carryover have encountered is that they are too extreme. For example, in the case of the AB/BA design the choice is between an unbiased inefficient estimator PAR or an efficient but potentially biased estimator CROS. The former estimator would correspond roughly to a Bayesian analysis in which uninformative priors were used for both treatment and carryover effects. The latter would apply in a Bayesian analysis in which a completely informative prior was used for carryover, assigning it the value 0 with probability.^[1] (The two-stage procedure is, of course, completely incoherent in a Bayesian sense in that a choice is made between these two extremes on the basis of an amount of information that is quite inadequate to the task in hand.^[28]) A general advantage of Bayesian approaches, however, is that compromise positions are possible or, to reinterpret in a frequentist framework, that bias-variance trade-offs can be achieved naturally.^[13]

A series of papers by Grieve starting 1985 has developed a Bayesian approach to analyzing crossover trials.^[28–32] (An earlier paper by Selwyn et al. considered Bayesian approaches to bioequivalence.^[33]) The general difficulty is that of handling the prior for carryover appropriately. Grieve's approach establishes the appropriate posterior distribution under the assumption that carryover is zero and also making no such assumption. As discussed above, the former corresponds roughly to the CROS analysis and the latter to the PAR analysis, but for the fact that some variance information is recovered from the second period.^[34] However, prior odds for these cases are elicited and these can be updated to produce posterior odds via a Bayes factor. These can be used to mix over the two models and hence produce an integrated posterior distribution for the treatment effect.^[28]

OTHER OUTCOMES

The discussion above has been in terms of the general linear model. This has relatively more importance for the crossover trial than for parallel group trial as the former is restricted in use to chronic diseases for which continuous outcome measurements are common. It may be useful, however, to supplement such analyses with nonparametric techniques. A simple approach for the AB/BA design uses the Wilcoxon-Mann-Whitney test.^[35] A good review of approaches in more complex cases is given by Tudor and Koch.^[36]

Binary outcomes present particular problems in crossover trials because of the difficulty in handling dependence of observations on the same subject. A useful review of various approaches is given by Kenward and Jones^[37] who themselves have developed an approach based on log linear models. The matter is also treated extensively in their

book.^[6] An alternative approach, particularly useful in the more general case where ordered categorical variables are involved, is that of Ezzet and Whitehead.^[38] They generalize the proportional odds model to the crossover context by allowing the true log-odds to have a normal distribution over all patients. It is relatively simple to implement the binary version of this^[39] using *proc nlmixed*[®] of version 8 of SAS[®].

Although survival in the classic sense is not relevant for crossover trials, certain outcome measures may be appropriately analyzed using survival approaches. One example is that of exercise tolerance tests in angina. France et al.^[40] show how such data may be analyzed by establishing patient "preferences:" if a patient "survived" longer under A than B then A is said to be "preferred" for that patient. A tie occurs where both measurements are censored. An alternative approach is given by Feingold and Gillespie.^[41]

BASELINES

Baseline measurements in crossover trials can be of various sorts. There can be baseline measurements taken before the trial begins, before each period of treatment, or even at the end of the trial when it is assumed that all treatment effects have worn off. Various different recommendations have been made for using such measurements to improve inferences. Some authors have suggested using these as a way of adjusting for carryover^[6,9] but this depends on the rather unrealistic assumption that the carryover at the end of a period of treatment will be the same as it was at the start.^[22]

If, however, active washout periods have been used and it can be assumed that they are successful, then baselines may be useful to adjust for random differences between outcome measurements. An important article by Kenward and Roger describes how this may be done in view of the fact that there may be both within and between subject correlations that need to be taken into account.^[42]

HYBRID DESIGNS

Consider clinical trials for infertility in which two treatments are being compared. If a previously infertile couple conceive having been given one treatment, they are hardly likely to seek a second alternative treatment on another occasion. However, if they have failed to conceive in a first period of treatment, they may well be interested in receiving the alternative. It is thus possible to construct clinical trials in which some couples, namely those that fail to conceive in the first period, are crossed over to the alternative therapy. This is therefore an example in which, although it is not possible to cross over all patients, it is possible

to cross over some. Such designs can thus be regarded as hybrids between parallel groups and crossover trials. No doubt there are many other clinical situations that could yield similar data.

Many statisticians will be hesitant about recommending such trials. However, if they are regarded as parallel group trials with some extra data, rather than as crossover trials with missing data, they can be seen in a more positive light. It is crucially important to use a method of analysis that is appropriate for both crossover and parallel group trials. It is thus inappropriate, for example, to treat the patient effect as fixed.

Cohlen et al.^[43] describe an approach to analysis of such trials using time dependent covariates. Makubate, in his doctoral dissertation,^[44] adapted the approach of Ezzet and Whitehead to binary data. Various approaches are discussed by Nason and Follmann.^[45]

FURTHER READING

The books by Jones and Kenward,^[6] Ratkowsky et al.,^[9] and Senn^[1] all give very different perspectives on the crossover trial: The last of these deals most closely with the concerns of the biopharmaceutical statistician. The paper by Hills and Armitage^[19] is an excellent introduction to the AB/BA design although, for reasons explained above, its advice as regards carryover is not sound. The encyclopedia articles by Kenward and Jones^[46] and Senn^[47] may also be useful, as may be a web-based tutorial^[48] and a review of papers that have appeared in *Statistics in Medicine*.^[19] A special issue of *Statistical Methods in Medical Research* was devoted to the crossover trial and has articles on the AB/BA design,^[16] multiperiod designs,^[17] binary data,^[35] nonparametric,^[34] and Bayesian^[30] approaches.

CONCLUSION

Crossover trials are not suitable for all indications and even in indications in which they can sometimes safely be employed, will not be suitable for investigating all treatments. Their use is best restricted to early drug development or for answering specific scientific questions. They are almost never suitable for phase III therapeutic studies. Nevertheless, it would be a grave mistake to conclude, therefore, that they are not useful. Their extreme efficiency makes them very attractive for early phase drug development including pharmacokinetics, bioequivalence, dose-finding, single-dose pharmacodynamic studies and proof of concept. They are also particularly suited for investigating treatment by subject interaction and as such may have an important role to play in the developing field of pharmacogenomics.^[49,50]

REFERENCES

1. Senn, S.J. *Cross-over Trials in Clinical Research*; Wiley: Chichester, 1993, 3–8 (second edition 2002).
2. Graff-Lonnevig, V.; Browaldh, L. Clin. Exp. Allergy **1990**, 20, 429–432.
3. Dahlof, C.; Bjorkman, R. Cephalalgia **1993**, 13, 117–123.
4. Chow, S.C.; Liu, J. *Design and Analysis of Bioavailability and Bioequivalence Studies*, 2nd Ed.; Marcel Dekker: New York, 2000.
5. Senn, S.J.; Lillienthal, J., Patalano, F.; Till, D. *Cross-over Trials*; Fischer: Stuttgart, 1997; 3–26.
6. Jones, B.J.; Kenward, M.G. *Design and Analysis of Cross-Over Trials*; Chapman and Hall: London, 1989.
7. Chi, E.M. Stat. Med. **1991**, 10, 1115–1122.
8. Senn, S.J.; Hildebrand, H. Stat. Med. **1991**, 10, 1361–1374.
9. Ratkowsky, D.A.; Evans, M.A.; Alldredge, J.R. *Cross-over Experiments Design, Analysis and Application*; Marcel Dekker: New York, 1993.
10. Kershner, R.P.; Federer, W.T.J. Am. Stat. Assoc. **1981**, 76, 612–619.
11. Lasker, E.M.; Meisner, M.; Kushner, H.B. Biometrics **1983**, 39, 1089–1091.
12. Balaam, L.N. Biometrics **1968**, 24, 61–73.
13. Carlin, B.P.; Louis, T.A. *Bayes and Empirical Bayes Methods for Data Analysis*; Chapman and Hall: London, 1996.
14. Fleiss, J.L. Biometrics **1986**, 42, 449–450.
15. Fleiss, J.L. Control Clin. Trials **1989**, 10, 1121–1130.
16. Senn, S.J. Stat. Med. **1992**, 11, 715–726.—(Correction, Stat. Med. 11: 1619, 1992).
17. Matthews, J. Stat. Methods Med. Res. **1994**, 3, 383–405.
18. Sheiner, L.B.; Hashimoto, Y.; Beal, S.L. Stat. Med. **1991**, 10, 303–321.
19. Senn, S.J. Stat. Med. **2006**, 25, 3430–3442.
20. Grizzle, J.E. Biometrics **1965**, 21, 467–480.—(Corrigenda, J.E. Grizzle Biometrics **1965** 30, 727 and AP Grieve Biometrics **1982**, 38, 517).
21. Hills, M.; Armitage, P. Br. J. Clin. Pharm. **1979**, 8, 7–20.
22. Senn, S.J. Stat. Methods Med. Res. **1994**, 3, 303–324.
23. Freeman, P.R. Stat. Med. **1989**, 8, 1421–1432.
24. Senn, S.J. Stat. Med. **1997**, 16, 2021–2024.
25. Wang, S.J.; Hung, H.M. J. Biometrics **1997**, 53, 1081–1091.
26. Abeyasekera, S.; Curnow, R.N. Biometrics **1984**, 40, 1071–1078.
27. Senn, S.J.; Lambrou, D. Stat. Med. **1998**.
28. Grieve, A.P.; Senn, S.J. J. Biopharm. Stat. **1998**, 8, 191–233.
29. Grieve, A.P. Biometrics **1985**, 41, 979–990.
30. Grieve, A.P. *Statistical Methodology in the Pharmaceutical Sciences*; Marcel Dekker: New York, 1989.
31. Grieve, A.P. Stat. Med. **1994**, 13, 905–929.
32. Grieve, A.P. Stat. Methods Med. Res. **1994**, 3, 407–429.
33. Selwyn, M.R.; Dempster, A.R.; Hall, N.R. Biometrics **1981**, 40, 1103–1108.
34. Senn, S.J. Statistician **2000**, 49, 135–176.
35. Koch, G.G. Biometrics **1972**, 28, 577–584.
36. Tudor, G.; Koch, G.G. Stat. Methods Med. Res. **1994**, 3, 345–381.
37. Kenward, M.G.; Jones, B. Stat. Methods Med. Res. **1994**, 3, 325–344.

38. Ezzet, F.; Whitehead, J. A random effects model for ordinal response from a cross-over trial. *Stat. Med.* **1991**, *10*, 901–907.
39. Ezzet, F.; Whitehead, J. *Appl. Stat.* **1991**, *41*, 117–126.
40. France, L.A.; Lewis, J.A.; Kay, R.A. *Stat. Med.* **1991**, *10*, 1099–1161.
41. Feingold, M.; Gillespie, B.W. *Stat. Med.* **1996**, *15*, 953–967.
42. Kenward, M.; Roger, J. *Biostat* **2009**, *10*.
43. Cohlen, B.J.; Te Velde, E.R.; Van Kooij, R.J.; Looman, C.W.; Habbema, J.D. *Hum Reprod* **1998**, *13*, 1553–1558.
44. Makubate, *PhD thesis Department of Statistics*; University of Glasgow: Glasgow, 2009.
45. Nason, M.; Follmann, D. *Biometrics* **2010**, *66*.
46. Kenward, M.G.; Jones, B. *Encyclopedia of Statistics, Update*; Wiley: New York, 1998, Vol. 2, 167–175.
47. Senn, S.J. *Cross-over Trials*. In: *Encyclopedia in Biostatistics*, Vol. 2, Armitage, P., and Colton, T., Eds.; Wiley: New York, 2005, 2, 1033–1049.
48. Senn, S.J. *Symptom Research Methods and Opportunities*; —<http://symptomresearch.com/chapter6/index.htm>.
49. Kalow, W.; Tang, B.K.; Endrenyi, L. *Pharmacogenetics* **1998**, *8*, 283–289.
50. Senn, S.J. *Drug Inf. J.* **2001**, 1479–1494.

Crossover Design: Rank-Based Robust Analysis

M. Mushfiqur Rashid

U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

This entry develops two rank-based methods for AB/BA crossover designs with baseline values as covariates under the assumption of within-subject exchangeable errors.

Keywords: Covariance; Crossover design; Rank-based method; Rank estimate.

INTRODUCTION

In the simple two-period, two-treatment (A and B) crossover design, subjects are randomly allocated to one of the two treatment sequences AB or BA. Between the treatment periods, there is typically a washout period that is long enough to reduce the potential for carryover of treatment effects from the first period to the second period. It is sometimes possible to make baseline measurements in AB/BA crossover trials. Such baseline measurements are usually made with the object of giving general or background information rather than direct information on treatment effect. If the subjects are subject to individual trend over time and the washout period is long enough compared to the treatment period, the baseline measurements taken at the beginning of each treatment may contain useful information and may be used easily as covariates. Their use may, however, increase directly the precision of measurements made directly on the treatment. Note that baseline measurements are the only covariates that may be used in the analysis of covariance in crossover trials because they are the only covariates that change during the trial and may not therefore be identical for each treatment. In order to use the baseline measurements as the covariates, we need to assume that there are no interactions between the treatments and the periods (see, for example, Refs. [1,2] for discussions). It is worth mentioning that in repeated-measures crossover trials, the observations within each subject are dependent. In the absence of the covariates in the model, besides normal theory inference, there are several nonparametric approaches available in the literature. One can conduct the Mann-Whitney-Wilcoxon test,^[3] the Brown-Mood median test, and the sign test (adjusted) for testing the equality of treatment effects using period differences (see Refs. [1,4,5] for further discussions).

However, there do not appear to be any rank-based tests for the equality of treatment effects (or period effects) in repeated-measures AB/BA crossover trials in the presence of covariates in the model. One may be tempted to use the procedures of Jaeckel^[6] (fixed-effects model approach, based on the joint rankings of all the residuals), which are analogous to the ordinary least squares (OLS) approach for the mixed-effects model in normal theory inference. However, Jaeckel's procedures suffer from loss in efficiency (see "Comparisons of $\hat{\alpha}$ and $\hat{\alpha}^{(P)}$ with $\hat{\alpha}^{(P)}$ When the Baseline Values Are Absent") when the observations within a subject are correlated. This is because the fixed-effects model cannot take into account the dependence of the data within a subject. It is worth mentioning that the analysis of unbalanced mixed models by the OLS approach is not appropriate. For the dependent data, the generalized least squares (GLS) approach gives more efficient inferences than those of the OLS approach.

OVERVIEW

In this entry, we develop two rank-based methods for AB/BA crossover designs with baseline values as covariates under the assumption of within-subject exchangeable errors. In the first method, we will define an objective function for each subject based on the intra-subject ranks of the residuals. Then we will define an overall objective function by taking the sum of all the objective functions. We will obtain R estimates by minimizing this overall objective function. We will then make inferences for the treatment effects and the slope using this overall objective function and the R estimates. In the second method, we will define a single objective function based on period differences for all the subjects for the purpose of making inferences about treatment effects and the slope parameter.

The rest of the entry is organized as follows. In "Model and Assumptions," we discuss the model and its assumptions. "Advantages and Uses of AB/BA Crossover Designs"

Notes: The views expressed in this article are those of the author and not those of the U.S. Food and Drug Administration.

discusses the advantages and uses of crossover designs; “Reasons for Analysis of Covariance in AB/BA Crossover Designs in Clinical Research” discusses reasons for the use of analysis of covariance in AB/BA crossover designs in clinical research. In “Intra-subject Rank-Based Method,” rank-based procedures based on intra-subject ranks of the residuals are developed. In “Rank-Based Methods Using Period Differences,” rank-based methods based on the period differences are developed. “Asymptotic Comparisons” compares the performances of the two rank-based methods presented in this entry—Jaeckel’s method^[4] and normal theory methods. In “An Illustration,” we give an example. The summary is presented in the “Conclusion.”

Model and Assumptions

Suppose, in a two-period, two-treatment crossover study design, that there are n subjects. These n subjects are divided at random into two groups. In the first group, there are n_1 subjects; in the second group, there are n_2 subjects. Let Y_{ijk} and x_{ijk} be the response (outcome) and the baseline value, respectively, corresponding to the i th ($i = 1, 2$) treatment applied on the k th ($k = 1, 2, \dots, n$) subject in the j th ($j = 1, 2$) period. A model for the two-period, two-treatment crossover design in clinical research is described (assuming no treatment by period interaction) as follows:

$$Y_{ijk} = \alpha_i + \pi_j + \beta x_{ijk} + \varepsilon_{ijk} \quad (1)$$

where α_i ($\alpha_1 + \alpha_2 = 0$) is the effect of the i th treatment, π_j ($\pi_1 + \pi_2 = 0$) is the effect of the j th period, β is the slope parameter, and ε_{ijk} is the random error term. It is usually assumed that ε_{ijk} in Eq. 1 is the sum of δ_k ($k = 1, 2, \dots, n_1 + n_2$) and ε_{ijk}^* , where δ_k is the independent normal random variable with a mean of zero and a variance σ_b^2 ($0 < \sigma_b^2 < \infty$) and ε_{ijk}^* is the independent normal random variable with a mean of zero and a variance σ_e^2 . Note that δ_k and ε_{ijk}^* are independent. In this case, Eq. 1 can be termed as an unbalanced mixed model. One may use the SAS procedure PROC MIXED (GLS approach) to make inferences for treatment effects, period effects, and the slope. Note that if we assume that the δ_k is a fixed effect, then we need to use the OLS technique. Note that in the absence of baseline values in Eq. 1, both the OLS and the GLS techniques give identical estimates. In the presence of baseline values in Eq. 1, the GLS inferences differ from the OLS inferences because the OLS approach does not take into account the equicorrelation structure of the within-subject errors.

In the following, we reparameterize Eq. 1 to reflect the dependency of the data within each subject. Let $\theta_{ij} = \alpha_i + \pi_j$. For the k th subject in the sequence AB, Eq. 1 can be written as:

$$\begin{pmatrix} Y_{11k} \\ Y_{22k} \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_{11k} \\ 0 & 1 & x_{22k} \end{pmatrix} \begin{pmatrix} \theta_{11} \\ \theta_{22} \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_{11k} \\ \varepsilon_{22k} \end{pmatrix}$$

For the k th subject in the sequence BA, Eq. 1 can be written as:

$$\begin{pmatrix} Y_{21k} \\ Y_{12k} \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_{21k} \\ 0 & 1 & x_{12k} \end{pmatrix} \begin{pmatrix} \theta_{21} \\ \theta_{12} \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_{21k} \\ \varepsilon_{12k} \end{pmatrix}$$

Let $\varepsilon_k^{(1)} \begin{pmatrix} \varepsilon_{11k} \\ \varepsilon_{22k} \end{pmatrix}$ ($k = 1, 2, \dots, n_1$) and $\varepsilon_k^{(11)} \begin{pmatrix} \varepsilon_{21k} \\ \varepsilon_{12k} \end{pmatrix}$ ($k = n_1 + 1, n_1 + 2, \dots, n_1 + n_2$) be the error vectors for the k th subject in sequences AB and BA, respectively. Without a loss of generality, we suppress the superscripts on the error vectors. In parametric inference, it is assumed that the error vectors $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_{n_1}, \varepsilon_{n_1+1}, \varepsilon_{n_1+2}, \dots, \varepsilon_{n_1+n_2}$ are i.i.d. bivariate normal. If the error vectors are not bivariate normal, alternative techniques such as nonparametric procedures are required. The normal theory F -test for a fixed-effects model is robust against a moderate departure from normality if the design is balanced. However, the problem of robustness in the mixed models is more serious than in the fixed-effects model because there are more assumptions to be met (e.g., normality of the δ_k) in the mixed-effects models (see Ref. [7] for discussions). We assume that the elements of ε_k are exchangeable random variables, that the ε_k values are i.i.d. continuous random vectors with the joint c.d.f. $F(u_1, u_2)$, and that the bivariate density $f(u_1, u_2)$, corresponding to $F(u_1, u_2)$, is continuous in R^2 .

Here, we find R estimates of the model parameters and test the following hypotheses:

1. $H_0^1: \frac{\theta_{11} + \theta_{12}}{2} = \frac{\theta_{21} + \theta_{22}}{2}$ vs. $H_a^1: \frac{\theta_{11} + \theta_{12}}{2} \neq \frac{\theta_{21} + \theta_{22}}{2}$
2. $H_0^2: \frac{\theta_{11} + \theta_{21}}{2} = \frac{\theta_{12} + \theta_{22}}{2}$ vs. $H_a^2: \frac{\theta_{11} + \theta_{21}}{2} \neq \frac{\theta_{12} + \theta_{22}}{2}$
3. $H_0^3: \beta = 0$ vs. $H_a^3: \beta \neq 0$

Note that H_0^1 implies that there are no treatment effects ($\alpha_1 = \alpha_2$), H_0^2 implies that there are no period effects ($\pi_1 = \pi_2$), and H_0^3 implies zero slope.

Advantages and Uses of AB/BA Crossover Designs

Advantages of AB/BA Crossover Designs

In many situations, there are constraints arising from the need to obtain efficacy and safety information in a limited time frame and with limited availability of subjects. When

the variability among subjects is much greater than the variability within subjects, the use of crossover design is appropriate, given that the clinical condition for treatment has a recurrent nature that is essentially equivalent across periods (e.g., discomfort from symptoms from chronic health order, breathing problems in asthma episodes, and problems with hypertension). Note that in these diseases, each patient can be given both treatments in turn, as is the case in the study of certain long-term chronic diseases when the treatments to be evaluated are palliatives designed to lessen pain or to bring comfort or a relief for a particular season rather than to effect a complete cure. One can often argue that carryover effects, if they exist, will be small relative to the treatment effects when there is a suitably long washout period between the treatment periods and when the observations are made at the end of the treatment periods. Note that we will not be able to test for carryover effects under Eq. 1 because we are using baseline values as covariates. Even in the absence of baseline values, it is hard to establish a powerful test for carryover effects. In choosing a crossover design, one needs to consider which assumptions are realistic regarding the nature of the treatment effects. Senn^[1] mentions that no help concerning carryover is to be expected from the data. The solution to this problem lies in designing experiments. The scientist must only use crossover trials in appropriate indications, and must allow for adequate washout.

Note that an important advantage of crossover designs is the requirement of a smaller number of subjects than a design with separate groups of patients for each treatment. One reason for this is that each subject provides information for two treatment periods. Another reason is that the crossover design enables an assessment of treatment comparisons with respect to the usually more sensitive framework of within-subject variability as opposed to that of between-subject variability. By using the crossover design, we reduce or eliminate from treatment contrasts one of the largest uncontrollable sources of variation in a clinical trial, the between-subject variability, and hence we obtain a better interpretation of information obtained from each individual.

Uses of Crossover Designs

The conclusion of the Biometric and Epidemiology Methodology Advisory Committee (BEMAC) of the Food and Drug Administration with respect to the two-period crossover design is that it “is not the design of choice for clinical trials, where unequivocal evidence of treatment effect is required.” Rather, the committee recommended that “in most cases, the completely randomized (or randomized block) design with baseline measurements will be the design of choice because it furnishes unbiased estimates of treatment effects without appeal to modeling assumptions save those associated with the randomization procedure itself.”^[8,9] The BEMAC report encourages the

use of baseline information in the analysis, and recommends the use of change from baseline scores, percentage change from baseline scores, and analysis of covariance. The report states, “If the baseline values vary between sequences or between periods (or both), it is essential that they be included in any analyses” (see Smith^[10] for a further discussion about current regulatory experience concerning crossover trials).

Over half of the 80 crossover trials published in the *British Medical Journal* between January 1980 and April 1988 concerned hypertension, diabetes, angina, and asthma. The design used in most of these trials was the standard AB/BA design (see Jones and Kenward^[4] for further discussion). A survey of 12 large pharmaceutical companies in the United States^[11] revealed a not-too-dissimilar view. Out of 72 crossover trials conducted by these companies over a 5-year period, over half of the trials used the AB/BA design.

A 12-year-old view from the FDA^[12] suggests that the two-period crossover design has not always been viewed negatively. Indeed, of the 22 clinical trials reviewed by Dubey,^[12] crossover designs are recommended in five, permitted in three, permitted but discouraged in eight, and not recommended in six. The U.S. Food and Drug Administration^[13] recommends the use of AB/BA crossover designs for bioequivalence studies. Senn^[1] reports that crossover trials are also popular in single-dose pharmacokinetic and pharmacodynamic studies in healthy volunteers as well as in Phase I tolerability studies and trials for bioequivalence.

Smith^[10] reports that crossover trials play a smaller role as Phase III “pivotal” efficacy trials for most medical indications. However, crossover trials play a major role in the clinical development of a new drug. Phase I and Phase II studies often follow crossover designs.

Grieve^[14] noted, “The saga of the crossover designs for clinical trials is not yet over, and it is healthy that the debate over the applicability of these designs continues, and will continue in the future.”

Reasons for Analysis of Covariance in AB/BA Crossover Designs in Clinical Research

Jones and Wang^[15] point out that if the washout period is such that the investigator may confidently expect that the effect of the previous treatment has been eliminated by the beginning of the following treatment period, and if it is considered that patients may be subjected to individual and specific trends over time (independent of the treatments to which they are allocated), then there may be some value in fitting baseline values in an analysis of covariance.

It is worth mentioning that the analysis of covariance technique is an efficient procedure for addressing two issues in clinical trials. The first is the reduction of variability in treatment effect and thereby the production of narrower confidence intervals and more powerful tests. The second

is the clarification of the extent of treatment effects in the presence of any baseline imbalance between the treatment groups. If it is suspected that there are treatment-by-period interactions, the researcher could use the procedures of Koch,^[3] which are based on gain score (response minus baseline). However, the gain score method can sometimes give different results than those of an analysis of covariance (see Grieve^[14] for parametric analysis using the gain score method and for a related discussion).

INTRA-SUBJECT RANK-BASED METHOD

In this section, we consider R estimation and rank tests based on intra-subject ranks of the residuals.

R Estimation

In this subsection, we develop R estimators and their asymptotic distributions. We will use a sum of Jaeckel-type dispersion functions^[6] to develop the rank-based inference. The role of this dispersion function is similar to the sum of squared errors (SSE) in the least squares theory. We will use the R estimators to carry out tests and confidence intervals.

Let $\theta = (\theta_{11}, \theta_{22}, \theta_{21}, \theta_{12})'$. A measure of dispersion for the k th subject, following the idea of Jaeckel,^[6] is:

$$D_k(\underline{\Delta}) = \sqrt{12} \sum_{j=1}^2 \sum_{i=1}^2 n_{ij}^{(k)} \left[\frac{R_{ijk}}{3} - \frac{1}{2} \right] W_{ijk} \quad (2)$$

where $n_{ij}^{(k)} = 1$ if the i th treatment is applied at the j th period on the k th subject and $n_{ij}^{(k)} = 0$ otherwise, $W_{ijk} = Y_{ijk} - \theta_{ij} - \beta x_{ijk}$, and R_{ijk} is the intra-subject rank of W_{ijk} .

It can be shown that $D_k(\underline{\Delta})$ is non-negative, continuous, location-free, and convex in $\underline{\Delta}$. It is worth mentioning that the coefficient of the residuals W_{ijk} lies in the interval $(-1/2, 1/2)$, and this coefficient (of W_{ijk}) assigns a weight proportional to the difference between its rank (between two elements) and the average rank. Because $D_k(\underline{\Delta})$ is a piecewise linear function of the residuals, the outliers enter into the dispersion function in a linear fashion.

In practice, individual subjects are different (not similar). Therefore the subjects may differ in their responses to a drug. By using the intra-subject ranks of the residuals, we are removing the subject effect. As mentioned earlier, in crossover designs, it is usually assumed that the variation among subjects is much greater than the variation within subjects, and so the estimation of model parameters on using intra-subject ranks is appropriate. We will show that inferences based on intra-subject rank-based methods are more efficient than those based on normal theory methods (for heavy-tailed and heavier-tailed distributions) and rank-based methods (for heavier-tailed distributions) using period differences.

Using Eq. 2, we can construct a combined dispersion function for Eq. 1 as follows:

$$D(\underline{\Delta}) = \sum_{k=1}^n D_k(\underline{\Delta}) \quad (3)$$

$D(\underline{\Delta})$ in Eq. 3 is non-negative, continuous, location-free, and convex in $\underline{\Delta}$. Although the ranking of the residuals is done within each subject separately, there is a “borrowing of strength” from across the subjects because the R estimates are obtained by minimizing the combined dispersion function. Under the assumption that the subjects are similar, a model that pulls the estimates together is used. Note that the sum of Jaeckel-type dispersion function^[6] (based on within rankings) has been used to incorporate concomitant variables in one-way repeated-measures designs (see Ref. [16]), a nested error regression model (see Ref. [17]), and incomplete block clinical trials (see Ref. [18]).

We obtain an R estimator of $\underline{\theta}$ by minimizing $D(\underline{\Delta})$. Let $\hat{\underline{\Delta}} \in \Gamma$, where $\Gamma = [\underline{\Delta}: D(\underline{\Delta}) = \text{a minimum}]$. One can use the Nelder-Mead algorithm^[19] to minimize $D(\underline{\Delta})$ (see the NLPNMS procedure of PROC IML of SAS for a detailed discussion of the Nelder-Mead algorithm).

Recently, Kloke et al.^[20] have developed rank-based estimation and associated inferences for linear models with cluster correlated errors. Their method is based on the joint (single) ranking (JR) of residuals of all the subjects. In a future work, the JR method will be applied to 2×2 and higher order crossover designs.

Kloke et al.^[20] has also obtained the asymptotic relative efficiency (ARE) results (for balanced designs) for several common error distributions comparing the JR estimator with the intra-subject rank-based estimator, which is based on a Friedman-type of ranking within blocks (many rankings: MR) and was proposed by Rashid.^[21] Except in extreme cases the JR estimator does quite well. For example, the JR estimator is more efficient than the MR estimator at the normal distribution except when the block size is large or when the correlation is high. They confirmed these ARE results by a Monte Carlo simulation study. This simulation study also confirms that for heavier tailed distributions (contaminated normals) the JR estimator is more efficient than least-square (LS). It will be of interest to see how the two methods (JR and MR) perform in the context of cross-over designs.

In order to obtain the asymptotic distribution of the R estimators, we need the asymptotic distribution of the gradient of the dispersion function. In the following, we develop the asymptotic distribution of the gradient of the dispersion function.

Let

$$\underline{S}(\underline{\theta}, \underline{\beta}) = [s_{11}(\underline{\theta}, \underline{\beta}), s_{22}(\underline{\theta}, \underline{\beta}), s_{21}(\underline{\theta}, \underline{\beta}), s_{12}(\underline{\theta}, \underline{\beta}), s_5(\underline{\theta}, \underline{\beta})]' \quad (4)$$

be the negative of the gradient of $D(\underline{\theta}, \underline{\beta})$. The domain of $D(\underline{\theta}, \underline{\beta})$ consists of polygonal subsets, in each of which

$D(\underline{\theta}, \beta)$ is linear in $\underline{\theta}$ and β . Therefore the partial derivatives exist almost everywhere and are given by the coefficients of the parameters in the dispersion function.

The elements of the negative of the gradient $\underline{S}(\underline{\Delta})$ of $D(\underline{\Delta})$ are given by:

$$\begin{aligned} s_{11}(\underline{\theta}, \beta) &= \sqrt{12} \sum_{k=1}^{n_1} \left[\frac{R_{11k}}{3} - \frac{1}{2} \right] \\ s_{22}(\underline{\theta}, \beta) &= \sqrt{12} \sum_{k=1}^{n_2} \left[\frac{R_{22k}}{3} - \frac{1}{2} \right] \\ s_{12}(\underline{\theta}, \beta) &= \sqrt{12} \sum_{k=1}^{n_2} \left[\frac{R_{12k}}{3} - \frac{1}{2} \right] \\ s_{21}(\underline{\theta}, \beta) &= \sqrt{12} \sum_{k=1}^{n_2} \left[\frac{R_{12k}}{3} - \frac{1}{2} \right] \\ s_5(\underline{\theta}, \beta) &= \sqrt{12} \sum_{k=1}^n \sum_{i=1}^2 \sum_{j=1}^2 n_{ij}^{(k)} \left[\frac{R_{ijk}}{3} - \frac{1}{2} \right] x_{ijk} \end{aligned}$$

Under the true value $\underline{\Delta} = \underline{0}$ and exchangeability of errors, the ranks R_{ijk} within each subject are uniformly distributed over the permutations of the integers 1 and 2. Therefore under exchangeability of the errors:

$$E_0[R_{ijk}] = \begin{cases} \frac{3}{2} & \text{if } n_{ij}^{(k)} = 1 \\ 0 & \text{otherwise} \end{cases}$$

Also

$$\text{var}_0[R_{ijk}] = \begin{cases} \frac{1}{4} & \text{if } n_{ij}^{(k)} = 1 \\ 0 & \text{otherwise} \end{cases}$$

Further

$$\text{COV}_0[R_{ijk}, R_{i'j'k'}] = \begin{cases} -1/4 & \text{if } (i, j) = (1, 1) \text{ and } (i', j') = (2, 2) \\ -1/4 & \text{if } (i, j) = (2, 1) \text{ and } (i', j') = (1, 2) \\ 0 & \text{otherwise} \end{cases}$$

Thus $E_0[S(\underline{0})] = 0$

Let $\underline{\Sigma} = \text{cov}_0[(1/\sqrt{n})\underline{S}(\underline{0})]$.

Then

$$\underline{\Sigma} = (\sigma_{ii'}) = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} \begin{bmatrix} \frac{n_1}{n} & \frac{-n_1}{n} & 0 & 0 & \frac{\mu_1}{n} \\ \frac{-n_1}{n} & \frac{n_1}{n} & 0 & 0 & \frac{-\mu_1}{n} \\ 0 & 0 & \frac{n_2}{n} & \frac{-n_2}{n} & \frac{\mu_2}{n} \\ 0 & 0 & \frac{-n_2}{n} & \frac{n_2}{n} & \frac{-\mu_2}{n} \\ \frac{\mu_1}{n} & \frac{-\mu_1}{n} & \frac{\mu_2}{n} & \frac{-\mu_2}{n} & \frac{\eta}{n} \end{bmatrix} \quad (5)$$

where $\mu_1 = x_{11} - x_{22}$, $\mu_2 = x_{21} - x_{12}$.

$$\eta = 2 \sum_{j=1}^2 \sum_{i=1}^2 \sum_{k=1}^n n_{ij}^{(k)} (x_{ijk} - \bar{B}_{.k})^2 (> 0) \text{ and } \bar{B}_{.k} = (1/2) \sum_{i=1}^2 \Sigma_i = \frac{1}{2} \sum_{j=1}^2 n_{ij}^{(k)} x_{ijk}.$$

We assume

$$\lim_{n \rightarrow \infty} \text{cov}_0[n^{-1/2} \underline{S}(\underline{0})] = \lim_{n \rightarrow \infty} \left(\frac{1}{n} \right) \underline{\Sigma} = A = (a_{ii'}) \quad (6)$$

Let $x_{ijk}^4 < L (< \infty)$. Note that $\underline{S}(\underline{0})$ can be written as a sum of n bounded and independent random vectors. By a multivariate central limit theorem, it can be shown that under the true value $\underline{\Delta}^0 = \underline{0}$:

$$n^{-1/2} \underline{S}(\underline{0}) \xrightarrow{D} \text{MVN}(0, A) \quad (7)$$

as $n \rightarrow \infty$, where MVN stands for multivariate normal.

Note that the sum of the first three rows of A is zero. Therefore the matrix A is singular. Further, the diagonal elements are positive because they are variances of the components of $\underline{S}(\underline{\theta}, \beta)$ and we assume that $a_{ii} = \lim_{n \rightarrow \infty} (\sigma_{ii}/n) > 0$. Note that the rank of $\underline{\Sigma}$ is 3. The role of $\underline{S}(\underline{\Delta})$ in this entry is similar to the scores in maximum likelihood techniques.

Next, we derive the asymptotic distribution of $\hat{\underline{\Delta}}$. Using arguments similar to those given in Rashid and Nandram,^[16] it can be shown that:

$$\sqrt{n}(\hat{\underline{\Delta}}) \text{ and } \sqrt{n}\{(\tau/n)A^{-1}\underline{S}(\underline{0})\} \text{ are asymptotically equivalent} \quad (8)$$

where

$$\tau = \left[\frac{1}{\{\sqrt{12} \int_{-\infty}^{\infty} f(\varepsilon, \varepsilon) d\varepsilon\}} \right] < \infty \quad (9)$$

and $f(\cdot, \cdot)$ is the bivariate density of the two components of the error vector.

It follows from Eqs. 7 and 8, under the true value $\underline{\Delta}^0$, that:

$$\sqrt{n}[\hat{\underline{\Delta}} - \underline{\Delta}^0] \xrightarrow{D} \text{MVN}(\underline{0}, \tau^2 A^{-1}) \text{ as } n \rightarrow \infty \quad (10)$$

Note that there are model-checking procedures available for the nonparametric nested error regression model.^[17] Similar procedures can be applied to Eq. 1.

Rank Tests

In this section, we develop the drop-in-dispersion test based on the R estimators. We need an R estimator $\hat{\underline{\Delta}}_H$ of $\underline{\Delta}$ under the hypothesis $a'\underline{\Delta} = 0$. We also need the asymptotic distribution of the R estimator of $\hat{\underline{\Delta}}_H$. Let $\hat{\underline{\Delta}}_H$ minimize $D(\underline{\Delta})$ under $a'\underline{\Delta} = 0$. Using arguments similar to those given in Rashid,^[20] we have:

$$\sqrt{n} \hat{\underline{\Delta}}_H = \sqrt{n}[\hat{\underline{\Delta}} - A^{-1}a(a'A^{-1}a)^{-1}a'\hat{\underline{\Delta}}] + o_p(1) \quad (11)$$

Note that the reduced model R estimator $\hat{\Delta}_H$ has a representation (asymptotically) similar to reduced model estimators in the least squares analysis. For testing $\underline{a}'\underline{\Delta} = 0$, we compare $D(\hat{\Delta})$ with $D(\hat{\Delta}_H)$. It can be shown that under $\underline{a}'\underline{\Delta} = 0$:

$$D^* = \frac{2[D(\hat{\Delta}_H) - D(\hat{\Delta})]}{\hat{\tau}} \quad (12)$$

follows approximately a chi-square distribution with 1 *df* for large n . For $H_0^1: (\theta_{11} + \theta_{12})/2 = (\theta_{21} + \theta_{22})/2$, we take $\underline{a}' = (1/2, -1/2, -1/2, 1/2, 0)$.

The D^* test is analogous to $-2 \log \Lambda$ in maximum likelihood techniques. It is also analogous to F -tests in the normal theory inference (see Rashid and Nandram^[16] for the drop-in-dispersion test in one-way repeated-measures designs with a changing covariate). The proof of Eq. 12 goes along the same lines given in Rashid and Nandram^[16] (see also Rashid^[20]).

For testing a general linear hypothesis based on scores $\underline{S}(\underline{\Delta})$ evaluated at the reduced model estimate $\hat{\Delta}_H$, the reader is referred to Rashid.^[18] The scores test is analogous to Rao's scores test in the maximum likelihood technique.

Note that the results of this entry are derived under the assumption that the number of subjects is large. Rashid^[22] conducted a small-scale simulation study for the drop-in-dispersion test for multigroup repeated-measures designs (without covariates) with exchangeable errors within each subject. Based on this simulation study, it is recommended that the chi-square percentile be used for the D^* tests when $10 \leq n_1 \leq 15$ and $10 \leq n_2 \leq 15$.

One can construct a $100(1 - \gamma)\%$ confidence interval for any estimable function of $\underline{\Delta}$ using Eq. 9. Let $\underline{a}'\underline{\Delta}$ be a contrast in $\underline{\Delta}$. An approximate $100(1 - \gamma)\%$ confidence interval for $\underline{a}'\underline{\Delta}$ is:

$$\underline{a}'\hat{\Delta} \pm \left(z \frac{\gamma}{2} \right) \hat{\tau} \sqrt{\left(\underline{a}' \sum - \underline{a} \right) / n} \quad (13)$$

Estimation of τ

As mentioned earlier, the parameter $\tau = 1/[\sqrt{12} \{ \int_{-\infty}^{\infty} f(\varepsilon, \varepsilon) d\varepsilon \}]$ plays a role similar to $\sigma^2(1 - \rho)$ in parametric repeated-measures analysis. It can be shown that the parameter $1/[\sqrt{12\tau}]$ is the density of $d_k = [\varepsilon_{11k} - \varepsilon_{22k}]$ (or $d_k = [\varepsilon_{21k} - \varepsilon_{12k}]$) at zero for $k = 1, 2, \dots, n_1$ (or for $k = 1, 2, \dots, n_2$). Thus d_k ($k = 1, 2, \dots, n_1, n_1 + 1, \dots, n_1 + n_2$) are n i.i.d. random variables having continuous and symmetric distributions, each with the median at zero. Hence estimating τ involves estimating the inverse of the density function at zero. Following Bloch and Gastwirth,^[23] we can obtain a consistent estimator of the inverse of the density function at zero as:

$$g = \left\{ \frac{n}{(2m)} \right\} [d_{(n, [n/2]+m)} - d_{(n, [n/2]-m+1)}] \quad (14)$$

where $m = pn^{4/5}$ as $n \rightarrow \infty$. Thus a consistent estimator of τ is $\hat{\tau} = g / \sqrt{12}$. It is recommended that p be taken to be 0.5. In practice, we will replace ε_{ijk} s by the residuals.

RANK-BASED METHODS USING PERIOD DIFFERENCES

In this section, we develop the inferences for treatment effect based on the period differences. We will also discuss the estimation of a scale parameter. In normal theory inference, treatment effect is estimated using the period differences from each subject. The unbiasedness of this resulting estimate rests on the validity of the assumption that there is no treatment-by-period interaction. Our method in this section will also be based on the period differences. We also assume that there is no treatment-by-period interaction. First, we define an objective function for estimating the treatment effect and the slope. Then we develop the asymptotic distribution of these estimates under the full model and the reduced model (under the hypothesis). Then we develop rank tests.

Let $U_k^{(I)} = Y_{22k} - Y_{11k}$ and $U_k^{(II)} = Y_{12k} - Y_{21k}$ be the period differences of the responses for sequences I and II, respectively. Similarly, let $V_k^{(I)} = x_{22k} - x_{11k}$ and $V_k^{(II)} = x_{12k} - x_{21k}$ be the period differences of the baseline values for sequences I and II, respectively. Then Eq. 1 can be written separately for each sequence as:

$$\begin{aligned} U_k^{(I)} &= \alpha + \pi + \beta V_k^{(I)} + e_k^{(I)}, \quad k = 1, \dots, n_1 \\ U_k^{(II)} &= -\alpha + \pi + \beta V_k^{(II)} + e_k^{(II)}, \quad k = 1, \dots, n_2 \end{aligned}$$

where $\alpha = \alpha_2 - \alpha_1$ (treatment effect), $\pi = \pi_2 - \pi_1$ (period effect), $e_k^{(I)} = \varepsilon_{22k} - \varepsilon_{11k}$, and $e_k^{(II)} = \varepsilon_{12k} - \varepsilon_{21k}$. Note that, by using the period differences, we remove the subject effects. Without a loss of generality, we suppress the superscripts in $e_k^{(I)}, e_k^{(II)}, U_k^{(I)}, U_k^{(II)}, V_k^{(I)},$ and $V_k^{(II)}$. Note that e_k s are i.i.d. continuous random variables with a median of zero and a variance $2\sigma^2(1 - \rho)$. Using Wilcoxon scores, we define the Jaeckel dispersion function^[6] based on the joint ranks of the residuals $u_k - \psi_k \alpha - \pi - \beta v_k$ as follows:

$$D^{(P)}(\alpha, \beta) = \sqrt{12} \sum_{k=1}^n \left[\frac{R_k^{(P)}}{n+1} - \frac{1}{2} \right] W_k^{(P)} \quad (15)$$

where $W_k^{(P)} = u_k - \psi_k \alpha - \pi - \beta v_k$, $R_k^{(P)}$ is the rank of the k th residual $W_k^{(P)}$ among n residuals, and:

$$\psi_k = \begin{cases} 1 & \text{if } k = 1, 2, \dots, n_1 \\ -1 & \text{if } k = n_1 + 1, \dots, n_1 + n_2 \end{cases}$$

The term π (the period effect) has dropped out of the dispersion function because ranks of the residuals do not change by inclusion or exclusion of π . Thus by ranking

the residuals (over all the subjects) based on the period differences, we remove the period effect.

One may define two dispersion functions separately for Eq. 1 by using ranks (intraperson) of the n residuals corresponding to two periods, and then by adding the two dispersion functions to obtain a combined dispersion function. By doing this, we will be able to remove the period effects, but we will not be able to remove the subject effects from the estimates of the slope and treatment effects. The n subjects in Eq. 1 represent an effort to reduce experimental errors and prevent misleading comparisons of “apples and oranges.” In such a comparison, a difference in treatment effects would be confounded with basic characteristics (e.g., smoking, age) of the subjects, the latter being of little or no interest in a particular experiment. It would be foolish to compare the response of treatment 1 at period 1 of one subject with the response of treatment 2 at period 1 of another subject when the two subjects differ in a particular characteristic (see, for example, Hollander and Wolfe^[24] for a detailed discussion). Thus the use of intracomponent (e.g., intraperson) rankings for inferences for the repeated-measures designs might not be feasible.

Note that $D^{(p)}(\alpha, \beta)$ is non-negative, continuous, location-free, and convex in $(\alpha, \beta)'$. We can thus obtain an R estimator of $(\alpha, \beta)'$ by minimizing $D^{(p)}(\alpha, \beta)$. Let $(\alpha, \beta) \in \Gamma^{(p)}$, where $\Gamma^{(p)} = [(\alpha, \beta) : D^{(p)}(\alpha, \beta) = \text{a minimum}]$.

In order to obtain the asymptotic distribution of the R estimators, we need the asymptotic distribution of the gradient of the dispersion function. In the following, we develop the asymptotic distribution of the gradient of the dispersion function.

Let $\underline{S}^{(p)}(\alpha, \beta) = [s_1^{(p)}(\alpha, \beta), s_2^{(p)}(\alpha, \beta)]'$ be the negative of the gradient of $D^{(p)}(\alpha, \beta)$. Then:

$$\underline{S}^{(p)}(\alpha, \beta) = \left(\sqrt{12} \sum_{k=1}^n \psi_k \left[\frac{R_k^{(p)}}{n+1} - \frac{1}{2} \right], \sqrt{12} \sum_{k=1}^n v_k \left[\frac{R_k^{(p)}}{n+1} - \frac{1}{2} \right] \right)'$$

Under $(\alpha^0, \beta^0)' = \underline{0}$, $R_k^{(p)}$ s are uniformly distributed over the integer 1 through n . Therefore:

$$E_0[R_k^{(p)}] = 0, \quad \text{var}[R_k^{(p)}] = \frac{n^2 - 1}{12}, \text{ and} \\ \text{cov}[R_k^{(p)}, R_{k'}^{(p)}] = \frac{-(n+1)}{12} \quad (k \neq k')$$

It can be shown that under $(\alpha^0, \beta^0)' = \underline{0}$:

$$n^{-1/2} \underline{S}^{(p)}(0, 0) \xrightarrow{D} \text{BVN}(0, A^{(p)}) \quad (16)$$

as $n \rightarrow \infty$, where BVN stands for bivariate normal and:

$$A^{(p)} = (a_{ii}^{(p)}), \quad a_{11}^{(p)} = \lim_{n \rightarrow \infty} 4n_1 n_2 / \{n(n+1)\} \\ a_{12}^{(p)} = a_{21}^{(p)} = \lim_{n \rightarrow \infty} \frac{1}{n+1} \left(\sum_{k=1}^n \psi_k v_k - \frac{1}{n} \sum_{k=1}^n v_k \sum_{k=1}^n \psi_k \right)$$

$$a_{22}^{(p)} = \lim_{n \rightarrow \infty} \frac{1}{n+1} \sum_{k=1}^n (v_k - \bar{v})^2$$

We assume that the matrix $A^{(p)}$ is positive definite. Note that Eq. 16 can be proven using arguments similar to those given in Hettmansperger,^[25] and Hettmansperger and McKean.^[26]

Using arguments similar to those given in Jaeckel,^[6] it can be shown that:

$$\sqrt{n}(\hat{\alpha}^{(p)}, \hat{\beta}^{(p)})' \text{ and } \sqrt{n} \left\{ \left(\frac{\tau_1}{n} \right) (A^{(p)})^{-1} \underline{S}^{(p)}(0, 0) \right\} \text{ are} \\ \text{asymptotically equivalent} \quad (17)$$

where

$$\tau_1 = \frac{1}{\left\{ \sqrt{12} \int_{-\infty}^{\infty} f_1^2(e) de \right\}} \quad (18)$$

and $f_1(\cdot)$ is the density of $e_k = (\varepsilon_{22k} - \varepsilon_{11k})$ (or $\varepsilon_{12k} - \varepsilon_{21k}$) (see also Hettmansperger^[25] and Heiler and Willers^[27] for further details).

It follows from Eqs. 16 and 17 and under the true value $(\alpha^0, \beta^0)'$:

$$\sqrt{n}[(\hat{\alpha}^{(p)}, \hat{\beta}^{(p)})' - (\alpha^0, \beta^0)'] \xrightarrow{D} N[\underline{0}, \tau_1^2 (A^{(p)})^{-1}] \\ \text{as } n \rightarrow \infty \quad (19)$$

Now consider the estimation of α when $\beta = 0$. In this case, the dispersion function Eq. 15 will reduce to:

$$D^{(p)}(\alpha, 0) = \sqrt{12} \sum_{k=1}^n \left[\frac{R_k^{(p)}}{n+1} - \frac{1}{2} \right] W_k^{(p)}$$

where $W_k^{(p)} = u_k - \psi_k \alpha$ and $R_k^{(p)}$ is the rank of the k th residual $W_k^{(p)}$ among n residuals. Let $\hat{\alpha}_H$ minimize $D^{(p)}(\alpha)$. It follows from Eq. 19 that under the true value α^0 :

$$\sqrt{n}(\hat{\alpha}_H^{(p)} - \alpha^0) \xrightarrow{D} N \left[0, \frac{\tau_1^2}{a_{11}^{(p)}} \right] \text{ as } n \rightarrow \infty \quad (20)$$

Next, consider the estimation of β when $\alpha = 0$. In this case, the dispersion function Eq. 15 will reduce to:

$$D^{(p)}(0, \beta) = \sqrt{12} \sum_{k=1}^n \left[\frac{R_k^{(p)}}{n+1} - \frac{1}{2} \right] W_k^{(p)}$$

where $W_k^{(p)} = u_k - \beta$ and $R_k^{(p)}$ is the rank of the k th residual $W_k^{(p)}$ among n residuals. Let $\hat{\beta}_H^{(p)}$ minimize $D^{(p)}(\beta)$. It follows from Eq. 19 that under the true value β^0 :

$$\sqrt{n}(\hat{\beta}_H^{(p)} - \beta^0) \xrightarrow{D} N \left[0, \frac{\tau_1^2}{a_{22}^{(p)}} \right] \text{ as } n \rightarrow \infty \quad (21)$$

In the following, we develop rank tests for no treatment effects and zero slope. It can be shown that under $H_0^1: \alpha = 0$:

$$D_1^* = \frac{2[D^{(P)}(0, \hat{\beta}_H^{(P)}) - D(\hat{\alpha}^{(P)}, \hat{\beta}^{(P)})]}{\hat{\tau}_1} \quad (22)$$

follows approximately a chi-square distribution with 1 df. It can further be shown that under $H_0^3: \beta = 0$:

$$D_3^* = \frac{2[D^{(P)}(\hat{\alpha}_H^{(P)}, 0) - D^{(P)}(\hat{\alpha}^{(P)}, \hat{\beta}^{(P)})]}{\hat{\tau}_1} \quad (23)$$

follows approximately a chi-square distribution with 1 df.

The proofs of drop-in-dispersion tests D_1^* and D_3^* go along the lines of McKean and Hettmansperger^[28] (see also Hettmansperger^[25]).

Like the Wilcoxon-Mann-Whitney test, it is recommended that the chi-square percentile be used for the drop-in-dispersion test when both $n_1 \geq 10$ and $n_2 \geq 10$. In fact, the drop-in-dispersion test (D_1^*) is asymptotically equivalent to the test (the aligned rank test) based on the scores $s_1^{(P)}(0, \hat{\beta}^{(P)})$. The aligned rank test for testing $H_0^1: \alpha = 0$ is based on the reduced model residuals $W_k^{(P)} = u_k - \Psi - \hat{\beta}_H^{(P)} v_k$ (see Hettmansperger^[25] for aligned rank tests for the linear model with i.i.d. errors). In the absence of β in Eq. 1, the test ($H_0^1: \alpha = 0$), based on the scores $s_1^{(P)}(0, 0)$, reduces to the Wilcoxon-Mann-Whitney test proposed by Koch.^[3]

Because the distribution of e_k is symmetric about zero, π can be estimated by the median of the Walsh averages $(W_k^{(P)} + W_{k'}^{(P)})/2$, $k < k' = 1, 2, \dots, n$ (see Randles and Wolfe^[29]) where $W_k^{(P)} = u_k - \Psi - \hat{\beta}^{(P)} v_k$. Thus a nonparametric estimate of π is:

$$\hat{\pi} = \text{median}_{k < k'} \left\{ \frac{W_k^{(P)} + W_{k'}^{(P)}}{2} \right\}$$

Also we will be able to test $H_0^2: \pi = 0$ by the Wilcoxon signed rank test using the residuals $W_k^{(P)} = u_k - \Psi - \hat{\beta}^{(P)} v_k$. Because in crossover designs we are interested mainly in the treatment effect, we will not pursue the inferences concerning period effect.

Note that there are model-checking procedures available for the nonparametric linear model with i.i.d. errors. For example, see Refs. [30,31] for detailed procedures.

Estimation of τ_1

Note that $e_1, e_2, \dots, e_n$ are i.i.d. random variables. Let the density of e_k be $f_1(e)$. Then $\tau_1 = 1/(\sqrt{12}\gamma_1)$, where $\gamma_1 = \int_{-\infty}^{\infty} f_1^2(e) de$. The random variable e_k follows a continuous and symmetric distribution with median at zero. Hence estimating τ_1 involves estimating the inverse of the density function at zero. Following Bloch and Gastwirth,^[23] we can obtain a consistent estimator of the inverse of the density function at zero as:

$$\frac{1}{\hat{\gamma}_1} = \left\{ \frac{n}{(2m)} \right\} [e_{(n, [n/2]+m)} - e_{(n, [n/2]-m+1)}] \quad (24)$$

where $e_{(k)}$'s are the ordered e_k 's and $m = pn^{4/5}$ as $n \rightarrow \infty$. Thus a consistent estimator of τ_1 is $\hat{\tau}_1 = 1/[\sqrt{12}\hat{\gamma}_1]$. It is recommended that ρ be taken to be 0.5. In practice, we will replace the $e_{(k)}$'s by the residuals $\hat{e}_k^{(P)} = \hat{W}_k^{(P)} = u_k - \Psi_k \hat{\alpha}^{(P)} - \hat{\beta}^{(P)} v_k$. Thus the parameter τ_1 can be estimated by $\hat{\tau}_1 = 1/[\sqrt{12}\hat{\gamma}_1]$.

ASYMPTOTIC COMPARISONS

In this section, we compare the performances of R , $R^{(P)}$, and GLS methods in terms of asymptotic variances when the baseline values are absent. We also discuss the performances of R , $R^{(P)}$, and GLS methods on empirical grounds when the baseline values are present. We further compare both $\hat{\alpha}$ and $\hat{\alpha}^{(P)}$ with $\hat{\alpha}^{(J)}$, where $\hat{\alpha}^{(J)}$ is the R estimate of α using Jaeckel procedures,^[6] assuming that ε_{ijk} in Eq. 1 is the sum of δ_k (the fixed subject effect) and ε_{ijk}^* (the random error term with $\text{var}(y_{ijk}) = \text{var}(\varepsilon_{ijk}^*) \sigma^2$ when the baseline values are absent).

Comparisons of $\hat{\alpha}$, $\hat{\alpha}^{(P)}$, and $\hat{\alpha}_{\text{GLS}}$ Without the Baseline Values

In this section, we consider the asymptotic relative efficiencies of the three estimators $\hat{\alpha}$ (based on the within rankings), $\hat{\alpha}^{(P)}$ (based on the period differences), and $\hat{\alpha}_{\text{GLS}}$ in the absence of baseline covariate in Eq. 1.

It is interesting to note that it is the asymptotic distribution of the test under a sequence of alternatives that determines the asymptotic distribution of the estimator. In most cases, the empirical power is consistent with the results predicted by the asymptotic relative efficiency (Pitman efficiency).

Without a loss of generality, we assume that there is no covariate in Eq. 1. The asymptotic variances are as follows:

$$\begin{aligned} \text{var}(\sqrt{n}\hat{\alpha}) &= \frac{3n^2\tau^2}{4n_1n_2} \\ \text{var}(\sqrt{n}\hat{\alpha}^{(P)}) &= \frac{n(n+1)\tau_1^2}{4n_1n_2} \\ \text{var}(\sqrt{n}\hat{\alpha}_{\text{GLS}}) &= \frac{2n^2\sigma^2(1-\rho)}{4n_1n_2} \end{aligned}$$

Bivariate Normal Distribution

For the bivariate normal distribution with correlation coefficient ρ and variance σ^2 , we have:

$$\tau^2 = \frac{\pi\sigma^2(1-\rho)}{3} \quad \text{and} \quad \tau_1^2 = \frac{2\pi\sigma^2(1-\rho)}{3}$$

Therefore

$$\begin{aligned}\lim_{n \rightarrow \infty} \left\{ \frac{\text{var}(\sqrt{n}\hat{\alpha}^{(P)})}{\text{var}(\sqrt{n}\hat{\alpha})} \right\} &= 0.666 \\ \lim_{n \rightarrow \infty} \left\{ \frac{\text{var}(\sqrt{n}\hat{\alpha}_{\text{GLS}})}{\text{var}(\sqrt{n}\hat{\alpha})} \right\} &= 0.636 \\ \lim_{n \rightarrow \infty} \left\{ \frac{\text{var}(\sqrt{n}\hat{\alpha}_{\text{GLS}})}{\text{var}(\sqrt{n}\hat{\alpha}^{(P)})} \right\} &= 0.955\end{aligned}$$

We see that the F -test and $\hat{\alpha}_{\text{GLS}}$ are examples of procedures that have high power and efficiency for the normal model.

It should be pointed out that normal theory-based procedures are highly unstable. Only one observation is enough to alter the conclusions reached by an F -test. The $\hat{\alpha}$ and the corresponding drop-in-dispersion test are quite stable (see “An Illustration” for an example).

Bivariate t Distribution

The variance of t random variable with scale parameter σ and degrees of freedom η_1 is $\eta_1 \sigma^2 / (\eta_1 - 2)$. For the bivariate t -distribution with correlation coefficient ρ , scale parameter σ , and degrees of freedom η_1 , we get:

$$\begin{aligned}\tau^2 &= \left[\sqrt{\eta_1} \Gamma\left(\frac{\eta_1}{2}\right) / \Gamma\left(\frac{\eta_1 + 1}{2}\right) \right]^2 \frac{\pi \sigma^2 (1 - \rho)}{6} \\ \text{and } \tau_1^2 &= \left[\sqrt{\eta_1} \Gamma\left(\frac{\eta_1}{2}\right) / \Gamma\left(\frac{\eta_1 + 1}{2}\right) \right]^2 2\pi \sigma^2 \frac{(1 - \rho)}{6}\end{aligned}$$

Therefore

$$\begin{aligned}\lim_{n \rightarrow \infty} \left\{ \frac{\text{var}(\sqrt{n}\hat{\alpha}^{(P)})}{\text{var}(\sqrt{n}\hat{\alpha})} \right\} &= 0.666 \text{ for } \eta = 3, 4, 5, 8; \text{ and} \\ \lim_{n \rightarrow \infty} \left\{ \frac{\text{var}(\sqrt{n}\hat{\alpha}_{\text{GLS}})}{\text{var}(\sqrt{n}\hat{\alpha})} \right\} &= 1.62, 1.12, 0.96, 0.8 \\ \lim_{n \rightarrow \infty} \left\{ \frac{\text{var}(\sqrt{n}\hat{\alpha}_{\text{GLS}})}{\text{var}(\sqrt{n}\hat{\alpha}^{(P)})} \right\} &= 2.55, 1.77, 1.52, 1.26 \text{ for } \eta = 3, 4, 5, 8\end{aligned}$$

Thus both $\hat{\alpha}$ and $\hat{\alpha}^{(P)}$ do much better than the GLS estimator $\hat{\alpha}_{\text{GLS}}$ for the heavy-tailed distributions. However, $\hat{\alpha}^{(P)}$ does better than $\hat{\alpha}$ for the heavy-tailed distributions. It is expected that the R estimates obtained by minimizing $D(\Delta)$ and $D^{(P)}(\alpha, \beta)$ will be more robust than normal theory counterparts because the influence of the outliers is linear rather than quadratic.

Bivariate Double Exponential Distribution

Consider a bivariate double exponential distribution with $\sigma^2 = 2$ and $\rho = 0.5$. We can generate bivariate double exponential variables ε_1 and ε_2 with $\sigma^2 = 2$ and $\rho = 0.5$ by defining $\varepsilon_1 = x_1 - x_2$ and $\varepsilon_2 = x_1 - x_3$, where x_1, x_2, x_3 , and

x_4 are i.i.d. exponential random variables with a mean of one. Thus the p.d.f. of $\varepsilon_1 - \varepsilon_2$ is again double exponential. Therefore $\tau = 1/[\sqrt{12}/2]$ and $\tau_1 = 1/[\sqrt{12}/4]$. Thus:

$$\begin{aligned}\lim_{n \rightarrow \infty} \left\{ \frac{\text{var}(\sqrt{n}\hat{\alpha}^{(P)})}{\text{var}(\sqrt{n}\hat{\alpha})} \right\} &= 1.33 \\ \lim_{n \rightarrow \infty} \left\{ \frac{\text{var}(\sqrt{n}\hat{\alpha}_{\text{GLS}})}{\text{var}(\sqrt{n}\hat{\alpha}^{(P)})} \right\} &= 1.5 \\ \lim_{n \rightarrow \infty} \left\{ \frac{\text{var}(\sqrt{n}\hat{\alpha}_{\text{GLS}})}{\text{var}(\sqrt{n}\hat{\alpha})} \right\} &= 2\end{aligned}$$

Thus for the heavier-tailed distributions (e.g., double exponential), we prefer the tests of “Intra-subject Rank-Based Method” rather than those of “Rank-Based Methods Using Period Differences.” This is not surprising because the ranking is done within subjects; thus the ranks (in “Intra-subject Rank-Based Method”) are bounded and thereby again nullify the effects of outliers. We cannot discard the intra-subject rank-based tests because of its poor performance when the data are normally distributed. The intra-subject ranking is an attempt to eliminate a nuisance effect because of the subject’s characteristics. Usually the purpose of clinical trials is not to test for subject effects but to concentrate on testing treatment differences (see, for example, Senn and Auclair^[32] for further discussion).

Empirical Comparisons of $\hat{\alpha}$ and $\hat{\alpha}^{(P)}$ with $\hat{\alpha}_{\text{GLS}}$ in the Presence of Baseline Values

As to be expected, if the normal error structure holds, there are minor differences between the rank-based method and the GLS method. Recall that the GLS estimators are functions of σ_b^2 and σ_e^2 . Thus to obtain the point estimates, estimates of these variance components must be obtained. Thus while the GLS estimates require estimated variance components, they are not required for the rank-based estimates. Therefore the rank-based procedures are expected to show better performance than the GLS counterparts.

We know that the GLS estimators are no longer the best when the estimates of variance components (σ_b^2 and σ_e^2) are substituted in expressions for the estimates of α , π , and β . The R estimates, on the other hand, are not functions of the variance components. So the rank-based methods have the potential to perform better than the GLS methods even when the model based on the normal structure holds; and they (rank-based methods) do even better, especially when this structure does not hold.

Rashid and Nandram^[17] performed a small-scale simulation study for comparing the performances of the GLS procedures and the intracluster R estimator-based procedures for the nested error regression models in the context of small area estimation. They noted that there were gains in precision using the R estimates when normality is tenuous. The simulation study they conducted indicated

that if there were outliers occurring uniformly in either tails of the error distributions, the rank-based (intracluster) method would be superior to the GLS method. Similar results were indicated for skewed error distributions.

We expect that rank-based procedures based on the period differences would perform better than GLS procedures if there were outliers occurring uniformly in either tails of the error distribution. Similar conclusions would hold for skewed error distributions.

Comparisons of $\hat{\alpha}$ and $\hat{\alpha}^{(P)}$ with $\hat{\alpha}^{(J)}$ When the Baseline Values are Absent

First we discuss how to estimate $\hat{\alpha}^{(J)}$ where $\hat{\alpha}^{(J)}$ is the Jaekel's R -estimate of α based on fixed effects model. We assume that ε_{ijk} in Eq. 1 is the sum of δ_k (the fixed subject effect) and ε_{ijk}^* (the random error term) with $\text{var}(y_{ijk}) = \text{var}(\varepsilon_{ijk}^*) = \sigma^2$. Jaekel^[6] proposed an objective function, known as the dispersion function, to compute the R estimates of the fixed-effects model (i.e., the linear models with i.i.d. errors). The dispersion function of Jaekel^[6] is based on the joint ranking of all the residuals. Following Jaekel,^[6] we define a dispersion function for the fixed-effects version of Eq. 1 based on Wilcoxon scores as follows:

$$D(\alpha_1, \alpha_2, \pi_2, \pi_2, \underline{\delta}) = \sqrt{12} \sum_{k=1}^n \sum_{j=1}^2 \sum_{i=1}^2 n_{ijk} \times \left[\frac{R_{ijk}^{(J)}}{n+1} - \frac{1}{2} \right] W_{(ijk)}^{(J)}$$

where $W_{ijk}^{(J)} = y_{ijk} - \alpha_i - \pi_j - \delta_k - \beta x_{ijk}$, $R_{ijk}^{(J)}$ is the rank of the (ijk) th residual $W_{ijk}^{(J)}$ among $2n$ residuals, $\underline{\delta} = (\delta_1, \delta_2, \dots, \delta^n)'$ is a vector of fixed subject effects ($\sum_k \delta_k = 0$), and $n_{ij}^{(k)} = 1$ if the i th treatment is applied at the j th period on the k th subject and zero otherwise. The R estimates and drop-in-dispersion tests can be developed analogously as done in "Rank-Based Methods Using Period Differences."

Let $\alpha = \alpha_2 - \alpha_1$. It can be shown (see Ref. [25]) that when the covariate is absent in the model:

$$\text{var}(\sqrt{n}\hat{\alpha}^{(J)}) = \frac{n(n+1)(\tau_1^*)^2}{4n_1n_2}$$

Where

$$\tau_1^* = \frac{1}{\sqrt{12} \int_{-\infty}^{\infty} f_*^2(\varepsilon) d\varepsilon}$$

and $f_*(\cdot)$ is a probability density function of ε_{ijk}^* .

Comparisons of $\hat{\alpha}$ with $\hat{\alpha}^{(J)}$

It is sufficient to compare $\hat{\alpha}$ with $\hat{\alpha}^{(J)}$ in the absence of the covariate. Thus the ARE of $\hat{\alpha}$ with respect to $\hat{\alpha}^{(J)}$ is given by:

$$\lim_{n \rightarrow \infty} \left\{ \frac{\text{var}(\sqrt{n}\hat{\alpha}^{(J)})}{\text{var}(\sqrt{n}\hat{\alpha})} \right\} = \left(\frac{2}{3} \right) \left\{ \frac{\int_{-\infty}^{\infty} f_*^2(\varepsilon) d\varepsilon}{\int_{-\infty}^{\infty} f(\varepsilon, \varepsilon) d\varepsilon} \right\}^2$$

For the normal and t distributions, the preceding ARE reduces to $2/\{3(1-\rho)\}$. That is, when $\rho \geq 0.35$, the ARE of $\hat{\alpha}$ with respect to $\hat{\alpha}^{(J)}$ for the normal distribution is greater than one. For moderate and high values of ρ ($0.35 \leq \rho < 1$), the methods of "Intra-subject Rank-Based Method" are preferred over the methods of Jaekel.^[6]

For the double exponential distribution with $\rho = 0.5$ and $\sigma^2 = 2$, it can be shown that the ARE of $\hat{\alpha}$ with respect to $\hat{\alpha}^{(J)}$ is 2.67.

Comparisons of $\hat{\alpha}^{(P)}$ with $\hat{\alpha}^{(J)}$

It can be shown that the ARE of $\hat{\alpha}^{(P)}$ with respect to $\hat{\alpha}^{(J)}$ is given by:

$$\lim_{n \rightarrow \infty} \left\{ \frac{\text{var}(\sqrt{n}\hat{\alpha}^{(P)})}{\text{var}(\sqrt{n}\hat{\alpha}^{(J)})} \right\} = \frac{\left\{ \int_{-\infty}^{\infty} f_*^2(\varepsilon) d\varepsilon \right\}^2}{\left\{ \int_{-\infty}^{\infty} f_1(\varepsilon) d\varepsilon \right\}^2}$$

For the normal and t distributions, the preceding ARE reduces to $1/(1-\rho)$. Thus for $\rho > 0$ (which is expected in repeated-measures data), the ARE of $\hat{\alpha}^{(P)}$ with respect to $\hat{\alpha}^{(J)}$ for the normal distribution is greater than one. This result is analogous to the corresponding normal theory result, because in normal theory inference, the parameter $1/(1-\rho)$ also measures the efficiency of a paired design with respect to an unpaired design.

For the double exponential distribution with $\rho = 0.5$ and $\sigma^2 = 2$, it can be shown that the ARE of $\hat{\alpha}^{(P)}$ with respect to $\hat{\alpha}^{(J)}$ is 2.

In clinical crossover trials, the data are repeated measures; therefore the methods of this entry are very applicable. Therefore we will not pursue the method of Jaekel^[6] for repeated-measures data.

AN ILLUSTRATION

The analysis methods will be illustrated using the data set from a 2×2 crossover trial described in Patel.^[33] The trial included 17 subjects with mild to moderate bronchial asthma in which the acute effects of single doses of two active drugs were compared. We will label the treatments A and B. All medications are discontinued 12 hr prior to receiving the trial treatments. The response measurement used to compare the two treatment groups is forced expiratory volume in 1 sec (FEV₁). Subjects were divided at random into two groups, and those in the first group received the treatments in the order AB. It has been recently reported (see protocol 9 of the CIBA-GEIGY Brethine program, Basle, and Ref. [34]) that the baseline value for subject 11 prior to period 1 is 2.05. In fact, this subject did not improve at all after the medication. It has also been reported that the response for subject 8 on period 2 is 2.43 instead of 2.41. The observed responses are given in Tables 1 and 2. Because Patel found that there were no

Table 1 A 2×2 Crossover Trial of Single Oral Doses of Two Active Drugs (A and B) in Patients with Bronchial Asthma: Forced Expiratory Volume in 1 sec (FEV₁, L): Group I (Sequence AB)

Patient	x_{11k}, y_{11k}	x_{22k}, y_{22k}
1	1.09, 1.28	1.24, 1.33
2	1.38, 1.60	1.90, 2.21
3	2.27, 2.46	2.19, 2.43
4	1.34, 1.41	1.47, 1.81
5	1.31, 1.40	0.85, 0.85
6	0.96, 1.12	1.12, 1.20
7	0.66, 0.90	0.78, 0.90
8	1.69, 2.43	1.90, 2.79

Table 2 A 2×2 Crossover Trial of Single Oral Doses of Two Active Drugs (A and B) in Patients with Bronchial Asthma: Forced Expiratory Volume in 1 sec (FEV₁, L): Group II (Sequence BA)

Patient	x_{21k}, y_{21k}	x_{12k}, y_{12k}
1	1.74, 3.06	1.54, 1.38
2	2.41, 2.68	2.13, 2.10
3	2.05, 2.60	2.18, 2.32
4	1.20, 1.48	1.41, 1.30
5	1.70, 2.08	2.21, 2.34
6	1.89, 2.72	2.05, 2.48
7	0.89, 1.94	0.72, 1.11
8	2.41, 3.35	2.83, 3.23
9	0.96, 1.16	1.01, 1.25

interactions between the treatments and the periods, we use the baseline values as covariates.

We will consider two cases. In the first case, we use all the data points. In the second case, we delete subject 9, which is considered to be an outlier (mild) by Jones and Kenward^[4] and Kenward and Jones.^[35] In fact, Kenward and Jones^[35] deleted subject 9 from their analyses.

Patel's Data Without the Outlier

We consider two cases. The first case is the analysis in the presence of baseline values. The second case is the analysis without the baseline values.

Inferences in the Presence of Baseline Values

Using the subroutine NLPNMS of PROC IML of SAS, we compute the R estimates of the treatment effects and the slope. We have used PROC MIXED (PROC GLM) of SAS to compute the GLS (OLS) estimates.

First we consider the inferences of β . The estimate of β , its estimated standard error, and the value of the test statistic, along with the p value corresponding to each

Table 3 Patel's Data: Inference for the Slope

Method	β	SE	$F_{\beta = 0/D_{\beta = 0}}^*$	p value
OLS	1.3959	0.2953	22.34	0.0003
OLS _{WTO}	1.1075	0.1786	38.44	0.0001
R	1.1846	0.1304	27.07	< 0.0001
R_{WTO}	1.1833	0.2664	15.67	< 0.0001
GLS	1.1603	0.1017	130.03	0.0001
GLS _{WTO}	1.1207	0.0915	149.96	0.0001
$R^{(P)}$	1.2500	0.1115	50.23	< 0.0001
$R_{WTO}^{(P)}$	1.1823	0.1212	39.73	< 0.0001

WTO = without outlier; P = period.

Table 4 Patel's Data: Inference for the Treatment Effects with Baseline

Method	$\alpha_2 - \alpha_1$	SE	$F_{\alpha_1 = \alpha_2/D_{\alpha_1 = \alpha_2}}^*$	p value
OLS	0.2526	0.0761	11.02	0.0051
OLS _{WTO}	0.1943	0.0456	18.10	0.0009
GLS	0.2543	0.0755	11.32	0.0046
GLS _{WTO}	0.1946	0.0444	19.21	0.0007
R	0.2119	0.0450	19.87	< 0.0001
R_{WTO}	0.2115	0.0665	13.53	0.0002
$R^{(P)}$	0.2337	0.0289	64.64	< 0.0001
$R_{WTO}^{(P)}$	0.2117	0.0310	50.93	< 0.0001

WTO = without outlier; P = period.

method (OLS, GLS, R , and $R^{(P)}$), are given in Table 3. All methods show that there is a significant linear relationship between the response and the baseline values. However, the OLS method gave the highest standard error. The rest of the method gave comparable estimated standard errors. The OLS estimate is also much different from the GLS estimate. However, the GLS estimate and $\hat{\beta}$ are very close.

Second, we consider the inference for the treatment effect. We report the estimate, its estimated standard error, and the test statistic, along with the p value, in Table 4. We see that the OLS and the GLS estimates are very close. The $\hat{\alpha}$ and $\hat{\alpha}^{(P)}$ are very close but differ from both $\hat{\alpha}_{OLS}$ and $\hat{\alpha}_{GLS}$. The OLS estimate has the highest estimated standard error, whereas $\hat{\alpha}^{(P)}$ has the lowest estimated standard error. All the methods show that there is a significant difference between the two treatments.

Inferences Without the Baseline Values

We present the estimates of the treatment effect, its estimated standard errors, and the values of the test statistic, along with the p values for each method (GLS, R , and $R^{(P)}$), in Table 5 (assuming that there is no covariate effect in Eq. 1). The two R estimates are very close. However, $\hat{\alpha}^{(P)}$ has a smaller estimated standard error than $\hat{\alpha}$. The OLS/GLS estimate has the highest estimated standard

Table 5 Patel's Data: Inference for the Treatment Effects Without the Baseline

Method	$\alpha_2 - \alpha_1$	SE	$F_{\alpha_1 = \alpha_2} / D_{\alpha_1 = \alpha_2}^*$	p value
OLS/GLS	0.2552	0.1184	4.64	0.0478
OLS _{WTO} /GLS _{WTO}	0.1750	0.0873	4.01	0.0650
R	0.1521	0.0839	5.26	0.0217
R_{WTO}	0.1521	0.0798	3.68	0.0550
$R^{(P)}$	0.1772	0.0496	9.31	0.0022
$R_{WTO}^{(P)}$	0.1494	0.0542	6.61	0.0100

WTO = without outlier; P = period.

error. All the methods show that there is a significant difference between the two treatments.

Comparing Tables 4 and 5, we see that the inclusion of the baseline values in the model resulted in smaller standard errors (estimated) of the estimates. Thus we obtain more precise estimates of the effect when the baseline values are included in the model as covariates.

Patel's Data Without the Outlier

An investigation of the residuals (for period 1) from the OLS technique has shown that the residuals are not normally distributed. Also, subject 9 seems to be an outlier. We have therefore dropped this subject from the analysis.

Inferences in the Presence of Baseline Values

Using the subroutine NLPNMS of PROC IML of SAS, we compute the R estimates of the treatment effects and the slope. The estimates are presented under the subscript WTO (without outlier). First we consider the inferences of β . The estimate of β , its estimated standard errors, and the value of the test statistic, along with the p value for each method (OLS, GLS, R , and $R^{(P)}$), are given in Table 3. All the methods show that there is a significant linear relationship between the response and the baseline values. However, the intra-subject rank-based method (R) gives the largest standard error. The OLS estimate has the next largest standard error. The GLS estimate has the smallest estimated standard error, although it is very close to that of $\hat{\beta}^{(P)}$. Note that without the outlier, the normal probability plot of the OLS residuals (from the first period) shows that the data are normally distributed. However, all the estimators, except $\hat{\beta}$, are affected by the outlier. The most affected is the OLS estimate. However, the $\hat{\beta}^{(P)}$ is also affected by the outlier, but not as much as the OLS estimate. The GLS and OLS estimates are very close.

Second, we consider the inference for the treatment effect. The estimate of α , its estimated standard errors, and the value of the test statistic, along with the p value for each method (OLS, GLS, R , and $R^{(P)}$), are given in Table 4. Note that $\hat{\alpha}$ and $\hat{\alpha}^{(P)}$ are very close but differ from both the OLS and the GLS estimates. Note also that $\hat{\alpha}$ is not

affected by the outlier at all. The OLS/GLS estimate has the highest estimated standard error, whereas $\hat{\alpha}^{(P)}$ has the lowest estimated standard error.

Inferences Without the Baseline Values

The estimate of α , its estimated standard errors, and the value of the test statistic, along with the p value for each method (GLS, R , and $R^{(P)}$), are given in Table 5. Only the $R^{(P)}$ method shows that there is a significant difference between the two treatments. The intra-subject rank-based method shows borderline significance (p value of 0.055). The OLS/GLS method gives a p value of 0.0650, which shows that there is a numerical trend in favor of treatment 2 over treatment 1.

CONCLUSION

The results of this entry provide a robust alternative to the parametric analysis of repeated-measures AB/BA crossover over designs with baseline values as covariates. The rank-based tests have applications to biopharmaceutical studies, clinical trials, and other scientific experiments. One of the problems commonly encountered in bioavailability studies is that the data set sometimes contains some extremely large and/or small observations known as outliers. These outlying data may have dramatic effects on the bioequivalence test. Results with and without the outlying observations could be totally opposite. It is therefore appropriate to use rank-based methods, which are less affected by outliers.

The maximum likelihood and the GLS approaches are based on the assumption that the data follow a multinormal theory model. Clearly, this is a strong assumption and needs to be investigated for each case. One possibility would be to consider the use of transformations. But the transformation will distort the model and its assumptions. A nonparametric test would be required, for example, if the data could not be transformed to satisfy the usual (e.g., normality) assumptions. It is worth mentioning that both the GLS and the maximum likelihood approaches require the finiteness of the covariance matrix of the error vectors. In the rank-based methods, we do not need the finiteness of the covariance matrix of the error vectors. For this reason, the procedures proposed in this entry will have an attractive appeal to data analysts.

Because the estimates of the variances σ_e^2 and σ_b^2 are required in the GLS approach for the estimation of the fixed-effects part of the mixed model, it is important that they be non-negative. Also, the estimators of σ_e^2 and σ_b^2 will affect the variance of the estimators of the fixed effects. Searle^[36] noted, "These are difficulties of the mixed model: for unbalanced data, estimating fixed effects in the presence of variance components, and estimating variance components themselves" in the context of ANOVA-based estimates.

The default variance component estimation used by PROC MIXED of SAS is the restricted maximum likelihood estimators (REML) algorithm. Also, the computer algorithm BMDP-5V provides REML estimators of the variance components for a mixed model. For balanced experiments, REML estimates are identical to the estimates obtained from expected mean squares, provided all estimates are positive. It is important to obtain non-negative estimates of the variance components because they are used in the analysis of fixed effects. The REML estimation procedure does not, however, include estimating the fixed effects. Nevertheless, having obtained the estimators of the variance components using REML, one would undoubtedly use these to obtain maximum likelihood estimates for the fixed effects after replacing the variance components in the maximum likelihood equations by their estimates (see Searle^[36] for further discussion). On the other hand, the *R* estimates developed in this entry are easy to compute using the PROC IML of SAS. They do not suffer from the problems mentioned by Searle in the context of the GLS technique.

In this entry, we have assumed that there is no treatment-by-period interaction because we have used baseline values as covariates. In a future work, we will consider a rank-based approach to incorporate the baseline as responses in the presence of period-by-treatment interaction in the model (see Refs. [34,35] for normal theory approach). Koch^[37] suggests that the evaluation of treatment-by-interaction still requires attention, regardless of whether such interaction is because of carryover effects or not. It is also mentioned by Jones and Wang^[15] that carryover effects in AB/BA crossover designs should not always be ignored. Until now in the absence of baseline values, both in normal theory and in nonparametric inferences, the tests (the *t*-test and the Wilcoxon-Mann-Whitney test) for no period-by-treatment interaction are based on the subject totals. The tests based on the subject totals are not powerful because these procedures do not eliminate the subject effects from the test statistics.

ACKNOWLEDGMENTS

The author would like to thank Dr. A.J. Sankoh of the U.S. Food and Drug Administration for reading the manuscript and making comments that improved the quality of this entry.

REFERENCES

1. Senn, S.J. *Cross-Over Trials in Clinical Research*; Wiley: Chichester, 1993.
2. Senn, S.J. The use of baseline in clinical trials of bronchodilators. *Stat. Med.* **1989**, *8*, 1339–1350.
3. Koch, G.G. The use non-parametric methods in the statistical analysis of the two-period change-over design. *Biometrics* **1972**, *28*, 577–584.
4. Jones, B.; Kenward, M.G. *Design and Analysis of Cross-Over Trials*; Chapman and Hall: London, 1989.
5. Chow, S.-C.; Liu, J.-P. *Design and Analysis Bioavailability and Bioequivalence Studies*; Marcel Dekker: New York, 1992.
6. Jaeckel, L.A. Estimating regression coefficients by minimizing the dispersion of residuals. *Ann. Math. Stat.* **1972**, *43*, 1449–1459.
7. Lindman, H.R. *Analysis of Variance in Experimental Design*; Springer-Verlag: New York, 1992.
8. Food and Drug Administration. A Report on the Two-Period Cross-Over Design and Its Applicability in Trials of Clinical Effectiveness. Minutes of the Biometric and Epidemiology Methodology Advisory Committee (BEMAC) Meeting, Rockville, MD.
9. O'Neill, R.T. Subject-Own Control Designs in Clinical Drug Trials: Overview of the Issues with Emphasis on Two Treatment Problem. Presented at the Annual NCDEU Meeting, Key Biscayne, FL.
10. Smith, N.D. Discussion of "Estimating treatment effects in clinical cross-over trials": a regulatory perspective. *J. Biopharm. Stat.* **1998**, *8*, 243–247.
11. Fava, G.I.; Patel, H.I. *A Survey of Cross-Over Designs Used in Industry*; 1986. Unpublished manuscript.
12. Dubey, S.D. Current thoughts on cross-over designs. *Clin. Res. Pract. Drug Regul. Aff.* **1986**, *4*, 127–142.
13. U.S. Food and Drug Administration. Guidance: Statistical Procedures for Bioequivalence Studies Using a Standard Two-Treatment Cross-Over Design, Rockville, MD, 1992.
14. Grieve, A.P. Cross-Over Versus Parallel Designs. In *Statistical Methodology in the Pharmaceutical Sciences*; Berry, D.A., Ed.; Marcel Dekker: New York, 1990, 239–270.
15. Jones, B.; Wang, J. Comments on "Estimating treatment effects in clinical cross-over trials". *J. Biopharm. Stat.* **1998**, *8*, 235–238.
16. Rashid, M.M.; Nandram, B. A rank-based analysis of one-way repeated measures designs with a changing covariate. *J. Stat. Plan. Inference* **1995**, *46*, 249–263.
17. Rashid, M.M.; Nandram, B. A rank-based predictor for the finite population mean of a small area: An application to crop production. *J. Agric. Biol. Environ. Stat.* **1998**, *3*, 201–222.
18. Rashid, M.M. On the use of R-estimators in analyzing repeated measures incomplete block clinical trials with baseline values as covariates. *J. Ital. Stat. Soc.* **1996**, *5*, 261–284.
19. Nelder, J.A.; Mead, R. A simplex method for function minimization. *Comput. J.* **1965**, *7*, 308–313.
20. Kloke, J.D.; McKean, J.W.; Rashid, M.M. Rank-based estimation and associated inferences for linear models with cluster correlated errors. *J. Am. Statist. Assoc.* **2009**, *104*, 384–390.
21. Rashid, M.M. Robust analysis of two-way models with repeated measures on both factors. *Test* **1995**, *4*, 39–62.
22. Rashid, M.M. Inference based on ranks for two-way models with a grouping factor and repeated measures factor. *J. Nonparametr. Stat.* **1996**, *6*, 27–42.
23. Bloch, D.A.; Gastwirth, J.S. On a simple estimate of the reciprocal of the density function. *Ann. Math. Stat.* **1968**, *39*, 1083–1085.
24. Hollander, M.; Wolfe, D.A. *Nonparametric Statistical Inference*; Wiley: New York, 1973.

25. Hettmansperger, T.P. *Statistical Inference Based on Ranks*; Wiley: New York, 1984.
26. Hettmansperger, T.P.; McKean, J.W. *Robust Nonparametric Statistical Methods*; Arnold: London, 1998.
27. Heiler, S.; Willers, R. Asymptotic normality of R-estimates in the linear model. *Statistics* **1988**, *19*, 173–184.
28. McKean, J.W.; Hettmansperger, T.P. Tests of hypotheses on ranks in the general linear model. *Commun. Stat. Theory Methods* **1976**, *5*, 693–709.
29. Randles, R.H.; Wolfe, D.A. *Introduction to the Theory of Nonparametric Statistics*; Wiley: New York, 1979.
30. McKean, J.W.; Sheather, S.J.; Hettmansperger, T.P. Regression diagnostics for rank-based methods. *J. Am. Stat. Assoc.* **1990**, *85*, 1018–1028.
31. McKean, J.W.; Sheather, S.J.; Hettmansperger, T.P. The use and interpretation residuals based on robust estimation. *J. Am. Stat. Assoc.* **1993**, *88*, 1254–1263.
32. Senn, S.J.; Auclair, P. The graphical representation of clinical trials with particular reference to measurements overtime. *Stat. Med.* **1990**, *9*, 1287–1302.
33. Patel, H.I. Use of baseline measurements in the two-period cross-over designs. *Commun. Stat. Theory Methods* **1983**, *12*, 2693–2712.
34. Senn, S.J.; Grieve, A. Estimating treatment effects in clinical cross-over trials. *J. Biopharm. Stat.* **1998**, *8*, 191–233.
35. Kenward, M.G.; Jones, B. The analysis of data from 2×2 cross-over trials with baseline measurements. *Stat. Med.* **1987**, *6*, 911–926.
36. Searle, S.R. *Linear Models for Unbalanced Data*; Wiley: New York, 1987.
37. Koch, G.G. Discussion of “Estimating treatment effects in clinical cross-over trials”. *J. Biopharm. Stat.* **1998**, *8*, 239–242.

Cutoff Designs

Joseph C. Cappelleri

Pfizer Inc., New London, Connecticut, U.S.A.

William M. K. Trochim

Department of Policy Analysis and Management, Cornell University, Ithaca, New York, U.S.A.

Abstract

This entry discusses alternative design strategies that are intended to address ethical or practical concerns when it is deemed unethical or infeasible to randomize all subjects to study interventions.

INTRODUCTION

The randomized design is the preferred method for assessing the efficacy of treatments. Randomization of all subjects should be employed whenever possible. Randomization, in principle, serves at least three important purposes: (i) it avoids known and unknown biases on average; (ii) it helps convince others that the trial was conducted properly; and (iii) it is the basis for the statistical theory that underlies hypothesis tests and confidence intervals.^[1]

Randomization of all subjects has been criticized, however, because it may raise ethical concerns or practical limitations in certain situations. Ethical tensions may arise, for example, when strong a priori (though inconclusive) information favors the experimental treatment, when the disease is potentially life-threatening, and when randomization does not explicitly incorporate subjects' baseline clinical need or their willingness to incur risk.^[2,3] Examples that have stirred considerable debate about the ethics of the randomized design include the controversies about the release of drugs for AIDS,^[4] the availability of drugs for cancer treatment,^[5] and the use of extracorporeal membrane oxygenation (ECMO) for neonatal intensive care.^[6,7]

A second potential drawback of the randomized design occurs in instances when randomization is not feasible or practical. Such situations may arise in health services or outcomes research where, for example, a health education program is to be targeted only to people who need it.^[8] An evaluation and comparison between managed care and usual care could be made feasible if high users of health-care utilization receive managed care only and low users of health-care utilization receive usual care only. A study concerned with the effect of a letter as an intervention to control health-care costs could be made practical if the letter is sent only to physicians with high billed charges per subscriber, whereas those with lower billed charges per subscriber don't receive a letter.^[9] In these contexts, economic constraints and logistical barriers may dictate

that an experimental intervention is neither practical nor efficient for those who do not need it or who are not the targeted candidates. Moreover, treatment allocation that reflects actual practice allows for testing the effectiveness of the intervention—its benefit in a real-life setting, as opposed to its benefit in a controlled setting.

This entry discusses alternative design strategies that are intended to address ethical or practical concerns when it is deemed unethical or infeasible to randomize all subjects to study interventions. These design strategies may be called *cutoff* designs because they involve, at least in part, the assignment of subjects to treatments based on a cutoff score on a quantitative baseline variable that measures clinical need, severity of illness, or some other relevant measure. What follows is an overview of cutoff designs.

DESCRIPTION OF THE REGRESSION-DISCONTINUITY (RD) DESIGN

The most basic of cutoff designs is the regression-discontinuity (RD) design^[8,10–13] in which a baseline indicator, for example severity of illness, can be used to assign subjects to an intervention. All subjects below a cutoff point on the baseline indicator receive one treatment, while all subjects above it receive another treatment. The history of the RD design is found in the social sciences, specifically in program evaluation. It has been employed to evaluate the effects of compensatory education, being on the Dean's list, a criminal justice program, a health education program on serum cholesterol, accelerated math training, and the NIH Career Development Award.^[14] In these scenarios randomization of subjects was not a viable alternative. A comprehensive history of the regression-discontinuity design in three academic disciplines—psychology, statistics, economics—has been published.^[15]

The traditional RD design is a single-cutoff quasi-experimental design that involves no random assignment.

The RD design received its name from the “jump,” or discontinuity, at the cutoff in the regression line of baseline and outcome (follow-up) scores that occurs when there is a treatment effect. Figure 1 depicts a RD design with a hypothetical 10-point treatment effect (reduction). All subjects with scores above 20 on the baseline assignment indicator are most in the need of the intervention and hence are automatically assigned to the test (experimental) treatment, whereas those with scores of 20 or less (those less in need) are automatically assigned to control treatment.

As Fig. 1 illustrates, the outcome scores of the test treatment group (those scoring above the cutoff) are lowered by an average of 10 points from where they would be expected in the absence of a treatment effect. The solid lines show the predicted regression lines for a 10-point effect, and the dashed lines show the expected regression lines for patients in a treatment group if they were given the other intervention instead.

The baseline assignment covariate should be measured on at least an ordinal scale; it is more desirable, though, to have a continuous (ratio-level or interval-level) baseline assignment variable. Baseline and outcome may be the same or different, the cutoff can be placed anywhere along the baseline measure (as long as there are sufficient numbers in the control group), the direction of improvement can be positive or negative for either variable, the treatment groups could have more than two levels, and the response variable can be discrete or continuous.

VALIDITY OF THE RD DESIGN

Under the assumption that the outcome-baseline functional form is correctly specified, the RD design results in an unbiased estimate of treatment effect. An unbiased estimate of treatment effect is obtained because the assignment process is known perfectly and controlled for

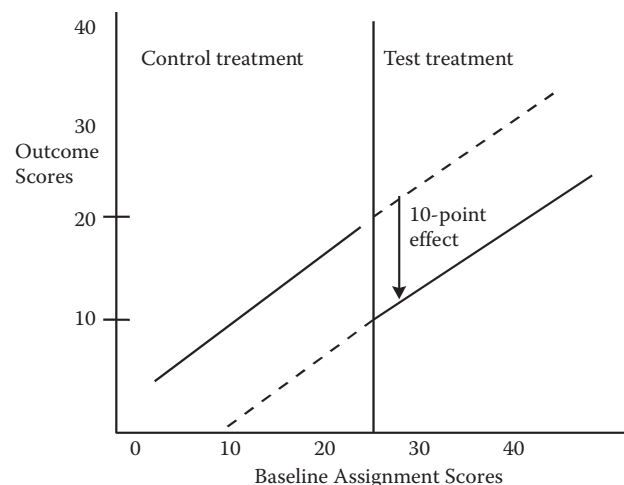


Fig. 1 Regression-discontinuity design with a 10-point treatment effect

in the analysis.^[10] Formal statistical derivations proving this unbiasedness are found elsewhere.^[14–18] Like the randomized experimental (RE) design, the RD design gives a known probability of assignment to treatments. It is imperative, though, that the cutoff assignment rule be followed strictly. If subjects are misclassified, then the treatment effect is likely to be biased.

It can also be demonstrated that the estimate of treatment effect in the RD design, like the RE design, remains unbiased when random measurement error in the observed, fallibly measured baseline covariate is considered.^[14–18] The reason for this is, once the fallibly measured observed baseline scores are known, treatment assignment is completely determined and hence independent of anything else, including the perfectly measured true baseline scores, in the RD design. Similarly, in the RE design, treatment assignment is completely determined by a randomization scheme and hence independent of anything else.

Regression to the mean, which naturally emanates from random measurement error in the observed baseline covariate, does not therefore affect the estimate of treatment effect in both the RD design and the RE design. Fig. 2 graphically shows the impact of regression to the mean, or, equivalently, random measurement error in the observed covariate, in the case of no treatment effect when the same variable is measured at baseline and follow-up. In the absence of a treatment effect, and with no other effects that may change a subject's score at follow-up, the true regression line should be a 45-degree line beginning at the origin. Regression to the mean causes the fitted regression line to be attenuated by an amount proportional to the reliability coefficient of the baseline covariate; therefore, the sample regression coefficient of the baseline covariate on the outcome measure is biased, but the sample regression coefficient of the treatment effect is not.^[14,16]

The RE design is robust in giving unbiased estimates of treatment effect when the true functional form between the baseline covariate and the outcome measure is not correctly specified. On the other hand, the RD design is not robust here. The most critical step in obtaining an unbiased estimate of effect in the RD design lies in modeling

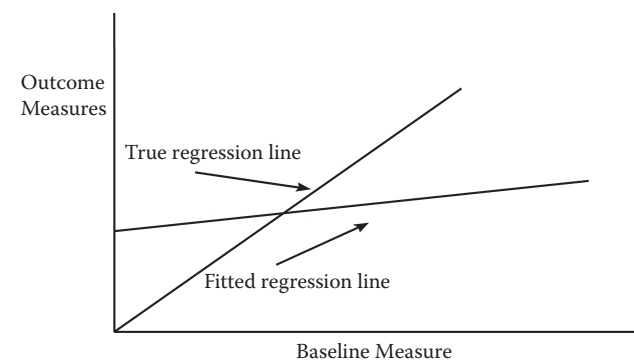


Fig. 2 Regression to the mean: randomized and regression-discontinuity designs

this true functional form correctly. The true functional form, however, is not known in the RD design because of missing data. As shown in Fig. 1, which assumes a linear functional form, the extrapolated regression line of the control group (dashed line, right) if this group's subjects were given treatment instead is assumed to continue in the same linear way as its fitted line (solid line, left). The extrapolated regression line of the test-treated group (dashed line, left) if this group's subjects were given control treatment is assumed to continue in the same way as its fitted line (solid line, right). There is no way to know prospectively whether the form or the slope of the lines in the region of missing data will be the same as that in the region of observed data.

SPECIFYING THE FUNCTIONAL FORM

One suggestion for helping to arrive at the correct functional form is to use a polynomial backward elimination regression approach.^[19] Another suggestion uses empirical Bayesian methods to overcome situations when the outcome and baseline relationship may not be linear, as when true baseline scores are not normally distributed.^[20,21] A third approach, which can be used with either of the other two approaches, is to fit a regression line over a wider range of the baseline-outcome distribution, resulting in less extrapolation and hence a more valid fit. This last approach can be achieved by combining the RD design with the RE design, resulting in a cutoff design with randomization.

COMBINING RD AND RE DESIGNS

A regression-discontinuity design can be described as a cutoff design without randomization. This design can also be coupled with a randomized design. For instance, patients who score within the middle range of scores on a baseline severity-of-illness indicator (e.g., those moderately ill) are randomized to either one of two treatments, while patients who score below a given cutoff value on this indicator (e.g., those most ill) are automatically assigned to the novel treatment and patients who score above another, higher cutoff value (e.g., those least ill) are assigned to the control treatment. Another type of cutoff design, for instance, would have subjects below the single cutoff point (e.g., the most ill) randomized to either treatment, whereas those above it (e.g., the least ill) are automatically assigned to control treatment. These are only two possible design variations that can combine cutoff assignment and random assignment. Other variations are mentioned elsewhere.^[22,23]

Combining the RE design and the RD design may give advantages over either design alone.^[22–24] Relative to the RE design, this hybrid design may be better suited to address ethical or practical concerns, may result in a larger eligible and diverse sample, and may better address

the effectiveness (as opposed to the efficacy) of interventions in particular circumstances. Compared with the RD design, RD-RE design has enhanced validity and improved statistical power.

ILLUSTRATION: COCAINE PROJECT

To illustrate the combined design, we describe a cocaine project, conducted by Havassy and colleagues at the University of California at San Francisco, that applied the RD-RE design instead of the completely randomized design, which was considered neither ethical nor feasible. The study included about 500 patients with cocaine addiction. The objective of the study was whether inpatient (intensive) rehabilitation showed better improvement, and by how much, over outpatient rehabilitation. The baseline assignment covariate was based on a weighted composite of four scales: (i) employment and legal status, (ii) family relationship and recovery, (iii) alcohol and drug history, and (iv) psychological status. Higher scores indicated more clinical need for the more intensive (inpatient) rehabilitation. The primary outcome variable was the same variable measured at follow-up.

Figure 3 portrays how patients may be allocated into inpatient or outpatient rehabilitation in this setting. All patients who score above 60—those most severely ill or most in need—are automatically assigned to inpatient rehabilitation; all patients who score below 40—those least ill or least in need—are automatically assigned to outpatient rehabilitation; and patients who score in between 40 and 60, inclusive—those moderately ill or in need—are randomized to either inpatient rehabilitation or outpatient rehabilitation. Note that it is this cutoff interval of randomization that distinguishes the RD-RE design from the RD design, which instead has a cutoff point(s) with no randomization.

Like Figs. 1 and 3 has solid lines representing the predicted regression lines and dash lines representing the extrapolated regression lines, showing a constant improvement from inpatient rehabilitation over outpatient rehabilitation. An analysis of covariance model, with the baseline assignment measure and the treatment group variable as predictors, would be a correct model to fit the fitted lines in Figs. 1 and 3. An analysis of variance model, which excludes the baseline assignment variable, should not be fit as it would result in a biased estimate of treatment effect. Although linear relations are highlighted in these two figures, cutoff designs are not restricted to a linear baseline-outcome relationship; higher-order terms (e.g., quadratic or cubic terms), transformations on baseline or outcome variables, and interaction terms may also be fitted.

In a simulation study, several RD-RE design variations, of which the basic design in Fig. 3 is the simplest one, were evaluated and compared among themselves, along with the traditional RD design and the traditional RE design.^[22,23] An unbiased main treatment effect was found for all these designs.

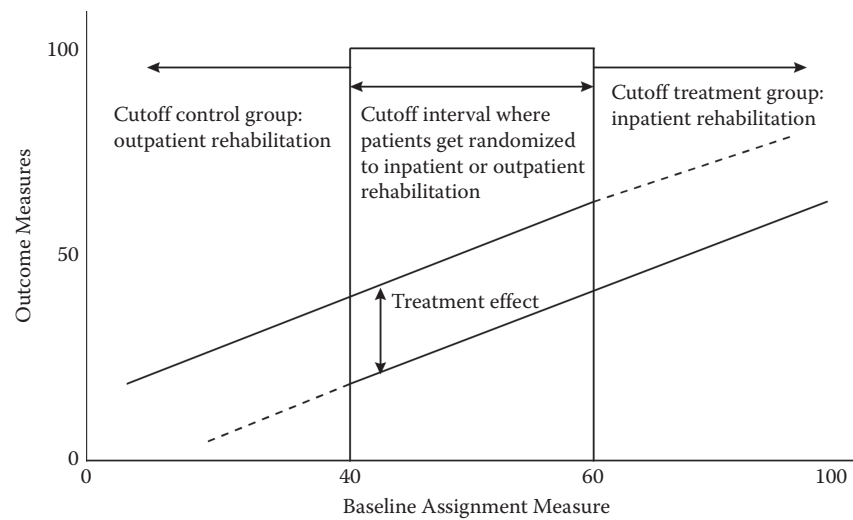


Fig. 3 Illustration of a combined randomized and regression-discontinuity design

Figure 4 shows one of the more advanced RD-RE designs that may be useful for accommodating varying amounts of resources. One cutoff interval has its bounds at 45 and 55; the other cutoff interval has its bounds at 40 and 60. Both intervals are symmetric around 50. Because the two intervals have different widths, they include different numbers of randomized patients, with the wider interval containing more randomized subjects. As subjects accrue into a study, investigators of a clinical site may favor one interval of randomization over the other in order to address the cost implications of having a shortage or surplus of hospital beds for inpatient rehabilitation. Or one interval may be preferred because it is more commensurate with a hospital's level of resources and expertise with respect to a given treatment.

MODELING AND ANALYZING CUTOFF DESIGNS

The RD-RE combination can be modeled and analyzed with the polynomial backward elimination approach mentioned in section “Specifying the Functional Form” for the RD design. Specifically, the initial model equation is

$$y + b_{int} + (b_{trt})^*z + (b_{xcut})^*xcut + (b_{xcut})^*(xcut)^2 + (b_{xcut}3)^*(xcut)^3 + (b_{linint})^*(z^*xcut) + (b_{linquad})^*(z^*xcut^2) + (b_{lincub})^*(z^*xcut^3) + error$$

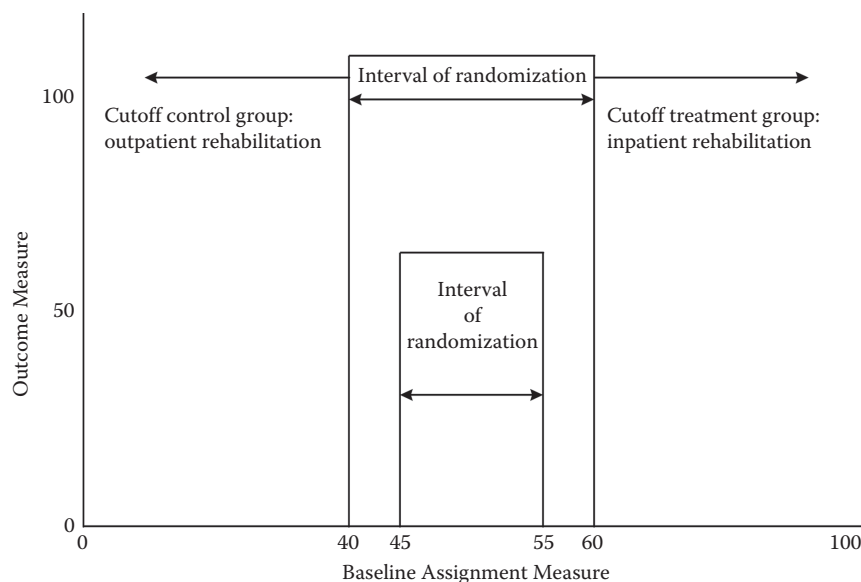


Fig. 4 Randomized and regression-discontinuity design with two cutoff intervals

where

y = outcome measure

$xcut$ = baseline assignment covariate minus a baseline value at which to measure the treatment effect (e.g., the middle value in a cutoff interval in a RD-RE design or the cutoff value itself in a RD design)

z = binary treatment group variable

$bint$ = intercept estimator

$btrt$ = treatment effect estimator

$bxcut$ = linear slope estimator

$blincut$ = linear interaction estimator

$error$ = sample regression error term.

The other regression coefficients are the coefficients for powers of “ $xcut$ ” higher than 1 and for their corresponding higher-order interactions. The same set of assumptions that apply to linear regression (for continuous responses) and to logistic regression (for discrete responses) also applies here.

The modeling strategy first tests the significance of each regression coefficient separately beginning with the higher-order interaction terms (i.e., the cubic interaction is tested first, followed by the quadratic interaction, and then linear interaction); interaction terms are tested before main effect terms. All significant terms and their lower-order counterparts are retained. The baseline covariate term and the treatment group variable are always kept in the final model.

RELATIVE SAMPLE SIZES NEEDED IN CUTOFF DESIGNS

The simulation study mentioned in section “Illustration: Cocaine Project” also showed that, everything else the same, more randomization resulted in lower standard errors of the treatment effect estimate and therefore increased precision. It can be shown that the amount of this precision is completely determined by the multicollinearity or correlation (R) between the baseline assignment covariate and the treatment group variable as expressed by the Variance Inflation Factor:^[16]

$$VIF = \frac{1}{(1 - R^2)}.$$

Suppose that there is a binary treatment group variable and a normally distributed baseline covariate. Table 1 provides the correlation between these two variables (R) and

the accompanying variance inflation factor (VIF) in symmetric cutoff designs with varying amounts of randomization and with 50% of the subjects within the interval assigned randomly to either treatment. The VIF can be interpreted as the design effect of how many more subjects are need in a given cutoff design relative to the completely randomized design in order to achieve the same level of statistical power, everything else the same.

Table 1 shows that, to achieve the same level of statistical power as the RE design, 2.75 times more subjects are needed in a RD design; 2.48 times more subjects are needed in a RD-RE design with 20% of all subjects randomized (i.e., 20% randomization); 1.96 times more subjects are needed in a RD-RE design with 40% randomization; 1.46 times more subjects are needed in a RD-RE design with 60% randomization; and 1.14 times more subjects are needed in a RD-RE design with 80% randomization. Derivations for the efficiency of such a cutoff design using an analogous approach, which gives essentially the same results, are published elsewhere.^[25]

SOME ADDITIONAL RESEARCH

Cutoff designs are certainly not without limitations. As mentioned earlier, an unbiased estimate of treatment effect requires that the functional relationship between outcome and baseline covariate be correctly modeled. Finkelstein et al.^[20,21] proposed a mathematical and statistical foundation, illustrated with examples, for how to analyze the RD design and to draw valid statistical conclusions about treatment efficacy. The authors discussed and illustrated their empirical Bayes methodology, which they mention can be used in a variety of circumstances, as a way to overcome restrictive assumptions about the functional form between outcome and baseline covariate. In other research, Hahn et al.^[26] have proposed a way of non-parametrically estimating treatment effects and offered an interpretation of the Wald estimator as an estimator of effect.

Table 1 Correlations and Variance Inflation Factors of Designs with Varying Amounts of Randomization

Percentage of all subjects within the interval of randomization	Correlation coefficient ^a	Variance inflation factor
0 (Regression-Discontinuity Design)	0.79	2.75
20	0.77	2.48
40	0.70	1.96
60	0.56	1.46
80	0.35	1.14
100 (Randomized Design)	0.00	1.00

^aExpected correlation between a binary treatment variable and normally distributed baseline covariate.

Another reservation about cutoff designs is that they preclude any serious attempt at complete blinding of treatment, making them similar to non-randomized designs in this regard. A further drawback of cutoff designs is they are less efficient (precise) than completely randomized designs in terms of their estimates of treatment effects. According to Senn,^[27] the considerable excess of patients treated on the inferior treatment in cutoff designs (especially the RD design) relative to the RE design is likely to undermine the ethical argument that favors cutoff designs. Although it is also true that more patients will receive the superior treatment in cutoff designs, regardless of which treatment it is, researchers are urged to consider Senn's position^[27] before abandoning randomization as a perceived ethical problem in a clinical trial.

Another variation of the RD design is the clustered RD design where groups (rather than individuals) are assigned to an intervention. The clustered RD design was the primary design to evaluate the federal education program prescribed in the No Child Left Behind Act of 2001. Schochet^[28] examined statistical power under clustered RD designs (without randomization) using techniques from the causal inference and hierarchical linear modeling literature. The main conclusion is that three to four times larger samples are typically required under the clustered RD design than the clustered RE design to achieve estimates of effect with the same level of precision.

Published studies using the RD design have focused primarily on linear regression applied to a categorical indicator and an interval-level response. Berk and de Leeuw^[29] formalized a generalization of the usual RD design to a wider range of situations. They focused on the use of a categorical treatment and response variables, but considered the more general case of any regression relationship. In addition, a resampling sensitivity analysis was shown as a way to address the credibility of the assumed assignment process. The broader formulation is applied to an evaluation of California's inmate classification system, which is used to allocate prisoners to different kinds of confinement.

CONCLUSIONS

Randomization should be employed whenever possible. Cutoff designs should not replace the completely randomized design in the majority of circumstances, usually involving a drug intervention, when no appreciable logistical barriers preclude all subjects from being randomized to interventions. Cutoff designs are an alternative design when circumstances in health services research or outcomes research warrant that randomization of all subjects cannot be undertaken, for whatever reason. Cutoff designs are much more likely to be relevant and appropriate in studies on program evaluation that involve educational or behavioral interventions, such as in disease management programs where randomization is not feasible ethically

or logistically for the entire sample,^[30] than in traditional Phase III studies on drug interventions, but cutoff designs may have some potential in Phase II therapeutic trials as well. When compared with non-randomized designs, the RD design (a cutoff design with no randomization) is an attractive alternative. When some subjects can be randomized, coupling the RD design with the RE design is even a more attractive alternative than the RD design.^[22–24,31]

REFERENCES

1. Green, S. Patient heterogeneity and the need for randomized clinical trials. *Control. Clin. Trials* **1982**, *3*, 189–198.
2. Beecher, H. Ethics and clinical research. *N. Engl. J. Med.* **1966**, *274*, 1354–1360.
3. Parker, L.S.; Arnold, R.M.; Meisel, A.; Siminoff, L.A.; Roth, L.H. Ethical factors in the allocation of experimental medical therapies. *Clin. Res.* **1990**, *3*, 537–544.
4. Marshall, E. Quick release of AIDS drugs. *Science* **1989**, *245*, 346–347.
5. Marx, J.L. Drug availability is an issue for cancer patients, too. *Science* **1989**, *245*, 345–346.
6. Ware, J.H. Investigating therapies of potentially great benefit: ECMO (with discussion). *Stat. Sci.* **1989**, *4*, 298–340.
7. Truog, R.D. Randomized controlled trials: Lessons from ECMO. *Clin. Res.* **1992**, *38*, 537–544.
8. Trochim, W.M.K. The regression-discontinuity design. In *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data*; Sechrest, L.; Perrin, P.; Bunker, J.; Eds.; Agency for Health Care Policy and Research, U.S. Public Health Service: Rockville, MD, 1990; 119–130.
9. Williams, S.V. Regression-discontinuity design in health evaluation. In *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data*; Sechrest, L.; Perrin, P.; Bunker, J.; Eds.; Agency for Health Care Policy and Research, U. S. Public Health Service: Rockville, MD, 1990, 145–149.
10. Shadish, W.R.; Cook, T.D.; Campbell, D.T. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*; Houghton Mifflin Company: Boston, MA, 2002, 207–245.
11. Trochim, W.M.K. *Research Design for Program Evaluation*; Sage Publications: Beverly Hills, CA, 1984.
12. Panel on the Evaluation of AIDS Interventions. In *Evaluating AIDS Prevention Programs*, Expanded Ed.; Coyle, S.L.; Boruch, R.F.; Turner, C.F.; Eds.; National Academy Press: Washington, DC, 1991, 144–159.
13. Mohr, L.B. *Impact Analysis for Program Evaluation*, 2nd Ed.; Sage Publications: Newbury Park, CA, 1995, 133–155.
14. Cappelleri, J.C.; Trochim, W.M.K.; Stanley, T.D.; Reichardt, C.S. Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: I. The case of no interaction. *Eval. Rev.* **1991**, *15*, 395–419.
15. Cook, T.D. Waiting for life to arrive: A history of the regression-discontinuity design in psychology, statistics, and economics. *J.Econom.* **2008**, *142*, 636–654.

16. Goldberger, A.S. *Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations*; Institute for Research on Poverty: Madison, WI, 1972; Discussion Paper #123.
17. Rubin, D.B. Assignment to treatment groups on the basis of a covariate. *J. Educ. Stat.* **1977**, *2*, 1–26.
18. Reichardt, C.S.; Trochim, W.M.K.; Cappelleri, J.C. Reports of the death of regression-discontinuity analysis are greatly exaggerated. *Eval. Rev.* **1995**, *19*, 39–63.
19. Cappelleri, J.C.; Trochim, J.C. An illustrative statistical analysis of cutoff-based randomized clinical trials. *J. Clin. Epidemiol.* **1994**, *47*, 261–270.
20. Finkelstein, M.O.; Levin, B.; Robbins, H. Clinical and prophylactic trials with assured new treatment for those at greater risk: I. A design proposal. *Am. J. Public Health* **1996**, *86*, 691–695.
21. Finkelstein, M.O.; Levin, B.; Robbins, H. Clinical and prophylactic trials with assured new treatment for those at greater risk: II. Examples. *Am. J. Public Health* **1996**, *86*, 696–705.
22. Trochim, W.M.K.; Cappelleri, J.C. Cutoff assignment strategies for enhancing randomized clinical trials. *Control. Clin. Trials* **1992**, *13*, 190–212.
23. Cappelleri, J.C.; Trochim, W.M.K. Ethical and scientific features of cutoff-based designs of clinical trials: A simulation study. *Med. Decis. Making* **1995**, *15*, 387–394.
24. Boruch, R.F. Coupling randomized experiments and approximations to experiments in social program evaluation. *Sociol. Methods Res.* **1995**, *4*, 31–53.
25. Senn, S.J. *Statistical Issues in Drug Development*, 2nd Ed.; Wiley: Chichester, 2007, 88–93.
26. Hahn, J.; Todd, P.; Van Der Klaauw, W. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* **2001**, *69*, 201–209.
27. Senn, S.J. A personal view of some controversies in allocating treatment to patients in clinical trials. *Stat. Med.* **1995**, *14*, 2661–2674.
28. Schochet, P.Z. Statistical power of regression discontinuity designs in education evaluations. *J. Educ. Behav. Stat.* **2009**, *34*, 238–266.
29. Berk, R.A.; de Leeuw, J. An evaluation of California's innate classification system using a generalized regression discontinuity design. *J. Am. Stat. Assoc.* **1999**, *94*, 1045–1052.
30. Linden, A.; Adams, J.L.; Roberts, N. Evaluating disease management programme effectiveness: an introduction to the regression discontinuity design. *J. Eval. Clin. Pract.* **2006**, *12*, 124–131.
31. Mandell, M.B. Having one's cake and eating it, too: combining true experiments with regression discontinuity designs. *Eval. Rev.* **2008**, *32*, 415–434.

Data Mining and Biopharmaceutical Research

Patricia B. Cerrito

University of Louisville, Louisville, Kentucky, U.S.A.

Abstract

In this entry, an overview will be given of the data-mining process and categories, including five data-mining methods: link analysis, kernel density estimation, text analysis, genetic programming, and artificial neural networks.

INTRODUCTION

Outcomes research requires routine data collection and the examination of information. Physicians should have a prominent role in decisions involving data collection, storage, and retrieval. However, physicians still have difficulty in adequately retrieving relevant medical information,^[1,2] and they require training to become proficient in this process.^[3] Methods that aid the physician in retrieving information need to be developed and disseminated to encourage evidence-based medicine.^[4] Improvement in information-retrieval tools is still needed,^[5–7] and data collection is still an important issue independent of data analysis.^[8–10] The collected clinical data require the use of data-mining tools to analyze the information.^[11–13] Data-mining techniques are ideal when data are collected in large databases.^[14–16]

Here an overview will be given of the data-mining process and categories, including five data-mining methods: link analysis, kernel density estimation, text analysis, genetic programming, and artificial neural networks. The application of these data-mining tools to clinical research will be illustrated by an analysis of MedPar data, a study of heparin dosing practices by two physicians, a quality-of-life analysis, and a study of polypharmacy. The examples will demonstrate the potential of the data-mining tools to find relevant information from clinical data.

THE DATA-MINING PROCESS

Data-mining tools come in three general categories:^[17] query-and-reporting tools, multidimensional analysis tools, and intelligent agents. Query-and-reporting tools include all the traditional statistical methods. They require close interaction between the investigator and the data in a specialized database or spreadsheet. Multidimensional analysis tools include the recently developed artificial neural networks and the Bayesian decision trees and require relatively structured data are structured by field and

observation. Text tends to be unstructured with sentences, phrases, and codes. Text uses more loosely defined terminology with different terms to define relatively the same concept, and similar terms with different meanings. The intelligent agents can investigate unstructured data and are the tools used to examine texts. These three categories of data-mining tools can be used in sequence to solve problems.

The process of data mining includes the following steps:^[18]

- Developing an understanding of the application domain, the relevant prior knowledge, and the goals of the end user.
- Creating a target data set, selecting a data set, or focusing on a subset of variables or data samples on which discovery is to be performed while considering the homogeneity of the data, change over time, and sampling strategy.
- Data cleaning and preprocessing, which includes removing noise and outliers and deciding on strategies for handling missing data.
- Data reduction and transformation, finding useful features to represent the data, depending on the goal of the task by using dimensionality reduction or transformation methods. In other words, the data need to be transformed into more useful forms from the original.
- Choosing the data-mining task by deciding whether the goal is classification, regression, clustering, summarization, or modeling.
- Data mining, performing the actual analysis, is automated and most often done using computer software.
- Evaluating output by determining what is actually knowledge and what is “fool’s gold.” Knowledge must be filtered from other outputs by using statistical checks, visualization, or expertise.
- Incorporating knowledge into the performance system and resolving potential conflicts with previously extracted knowledge.

DATA-MINING METHODS

There are many effective methods of data mining that are rarely used in medical studies:

- Link analysis (0 papers found).
- Kernel density estimation (5 papers found).^[19–23]
- Rule induction (7 papers found).
- Text analysis (15 papers found).^[24–25]
- Genetic programming (19 papers found).^[26–28]
- Artificial neural networks (909 papers found).^[29–31]

Note that only artificial neural networks are used with any regularity in medical studies. However, each method can be used to provide substantial information about drug actions and adverse events in clinical data. Image processing uses artificial neural networks. Many of the papers found dealt with the use of neural networks to automate the process of reading images that are recorded digitally.

Link Analysis

Link analysis explores the associations between large numbers of objects of different types.^[32] One of the principal uses of link analysis in medical research is to link primary effects to secondary effects. Link analysis is also a way to examine the adverse effects of medications during postmarketing surveillance. In particular, link analysis can determine whether a particular adverse event can be linked to just one medication or to a combination of medications. Further, link analysis can examine specific patient characteristics to determine whether they might correlate to both the adverse event and to the medication.

Recently developed software can examine thousands of objects and links simultaneously. For example, link analysis was applied to Medpar data, data that each hospital is required to report to the federal government for all Medicare patients. For a fee, the Medpar data are publicly available. Secondary diagnoses were linked based on a common primary diagnosis (Fig. 1) (SAS Institute, Cary, NC). For the Medpar data, it was possible to label up to nine different secondary diagnoses. The links were among the secondary DRG codes, limited to the most prominent 100 links. Code 4111 = intermediate coronary (the largest square in the diagram) was the code used most often in the patient data set. One of the strongest connections was to secondary code #4 = 4019 (hypertension).

Kernel Density Estimation

When the data are not normally distributed, it is possible to estimate the population distribution by using the kernel density estimate,^[33,34] defined by

$$f(x) = \frac{1}{nh_n} \sum_{j=1}^n K\left(\frac{x - X_j}{h_n}\right)$$

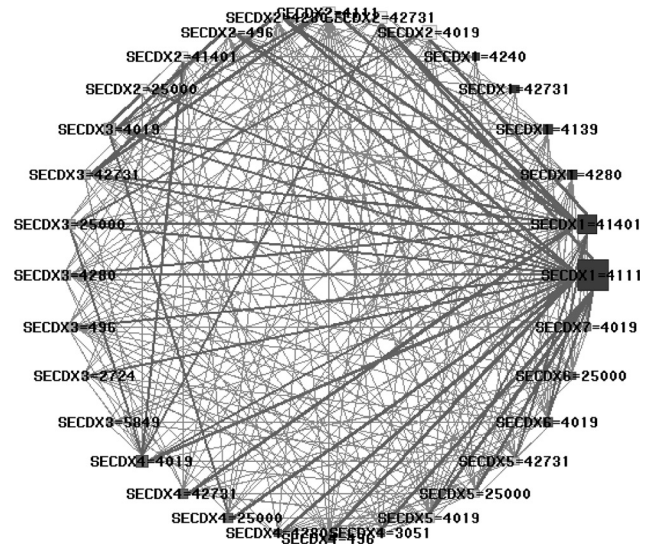


Fig. 1 Diagram of a link analysis

where K is a known density function, n is the sample size, and h_n is a parameter depending upon n and the population density function. It is known that the choice of K is unimportant to the relative efficiency of the estimator. There are several algorithms used to optimize the choice of h_n . The choice of bandwidth in this example is given by solving the fixed-point equation:

$$h_n = \left[\frac{R(\varphi)}{nR(\hat{f}_{g(h)}) \left(\int x^2 \varphi(x) dx \right)^2} \right]^{1/5}$$

The fixed-point equation can be used to accurately estimate drug levels over time, or the actual time between monitoring drug levels.

Although these techniques are still relatively new and inconstant development, they are not generally used in applications.^[35,36] Nevertheless, some studies have used the technique of kernel density estimation to classify data.^[19,22] Kernel density estimation provides more information than the standard statistical techniques, which rely on averages. Consider Fig. 2 where it is clear that one physician

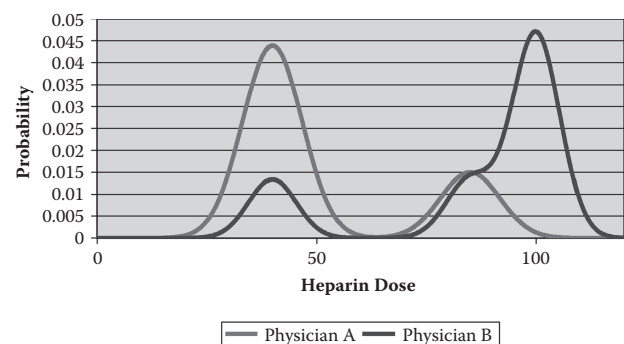


Fig. 2 Heparin dose by physician

typically uses a higher dose of heparin per unit weight than does a second physician. Although a standard analysis of variance would show that the two physicians are different, the graph demonstrates the extent of the difference. A conference with Physician B resulted in a change in practice to lower heparin-dosing habits.

Rule Induction

Rule induction takes the form IF condition THEN conclusion. For a rule under consideration, the following sets of records are defined:

A = {r|condition of rule is satisfied (i.e., the rule classifies a record)}.

B = {r|conclusion of rule is satisfied (e.g., the patient has died)}.

C = {condition and conclusion of rule are satisfied (i.e., the rule correctly classifies a record)}.

If a,b,c, respectively, are the number of records in the sets A,B,C, then the accuracy is equal to c/a and the sensitivity is equal to c/b . The objective then is to produce rules that maximize c and minimize the differences $a - c$ and $b - c$. The basic format of rule induction is modified to account for missing values, and multiple nodes of IF-THEN rules are added until all patient records are classified.

It is also possible to examine the issue of which secondary factors are associated with higher predicted risk using rule induction with the Medpar data. Different levels of complexity in the model can be used to increase the level of accuracy in predictions. In this way, it is possible to find the risk factors that need to be the focus to improve hospital ranking. This can be done by increasing the proportion of patients with the needed risk factors. The major leaves of the Bayesian tree were used to predict the risk of severity. Using the first five recorded secondary diagnoses as input variables, the results are summarized in the table below:

0389 20280 25002 2763 2765 2866 2874 29570 3441 34510 3688 40391
41011 41091 42090 4230 4241 42610 42653 4270 42741 4275 43491 436
4589 4822 4824 5070 5185 51881 51882 53540 53541 55320 5715 5750
5770 5789 5845 5849 585 7070 78009 78551 78791 78830 7885 7889
7991 99701 99702 9973 9975 9983 99859 V420

Node	Number of patients	Classified in risk category 1-3 (%)	Classified in risk category 4 (%)
2	137	18.98	81.02
4	84	32.14	67.86
6	31	22.58	77.42
7	1779	98.20	1.80

Node 2 is identified solely based upon the secondary diagnoses in column 5.

TEXT ANALYSIS

Text analysis examines unstructured texts for meaningful information. Using standard clinical data that are routinely collected, often in paper format, text analysis can be used to find standardized “natural language” pieces of information. Once the text has been organized, statistical methods for clustering and classification can be used. It is possible, for example, to construct patient profiles based upon the clustering of secondary diagnoses. The use of natural language allows for equivalent phrasing to classify patients based on comorbidities. For example, a Medpar database containing 4814 Medicare patients was examined using text analysis on the fields of secondary diagnoses given a specific primary diagnosis. The resulting text analysis was grouped into 10 separate clusters that had the following numbers of patients (Table 1) (SAS Institute, Cary, NC). The top three clusters were identified by DRG codes (Table 2).

Cluster 4 is focused primarily on mental problems, including drug use; cluster 9 is focused on urinary and digestive problems; and cluster 7 is focused on anemia, fluid depletion, and chronic obstructive bronchitis.

Genetic Programming

Genetic programming extracts ideas from the growth and evolution of living creatures and applies them to practical problems. It has been used occasionally to facilitate difficult diagnoses.^[27,28] The basic unit of work in genetic programming is the expression. An expression is a dendrite structure built from a predetermined alphabet of basic program elements. Program elements include simple arithmetic functions, Boolean operators, and terminal symbols such as random numbers, constants, or problem data. Genetic programming is used by building an initial population of random expressions, each of which attempts to solve the problem, usually in a series of object cases or

Table 1 Number of Patients in Each Cluster Identified by Secondary Diagnoses

Cluster	Patient total
1	75
2	30
3	19
4	732
5	68
6	25
7	1612
8	57
9	905
10	73

Table 2 Most Reported Secondary Diagnoses Per Cluster

Rank appearance	Cluster 4	Cluster 7	Cluster 9
1	2721 (lipoid metabolism disorder)	2851 (acute posthemorrhagic anemia)	43330 (occlusion and stenosis)
2	V1048 (history, malignant neoplasm)	2767 (fluid disorder)	5997 (hematuria)
3	V4502 (cardiac device in situ)	3963 (mitral and aortic valves, rheumatic fever)	9975 (urinary surgical complication)
4	V1301 (history of other diseases)	2765 (volume depletion)	2749 (unspecified gout)
5	V5861 (long-term drug use)	49120 (chronic obstructive bronchitis)	4401 (renal artery atherosclerosis)
6	7802 (syncope, collapse)	4589 (hypotension)	4270 (paroxysmal supraventricular tachycardia)
7	311 (depressive disorder)	4260 (complete atrioventricular block)	44020 (native arteries atherosclerosis)
8	V145 (history of allergy)	4275 (cardiac arrest)	53390 (peptic ulcer)
9	4242 (other disease, endocardium)	78039 (convulsions)	V103 (history of malignant neoplasm)
10	V1271 (history of certain other diseases)	49121 (chronic obstructive bronchitis)	42613 (atrioventricular block)

trials. Some expressions are more successful than others in solving various cases.

Genetic operations of selection, mutation, and reproduction are applied to the initial expressions, leading to the construction of a new generation of expressions that generally shows improved performance over the initial population. The expressions in each succeeding generation continue to attempt to establish solutions and breed high-performing expressions. Each succeeding generation builds on the best efforts of the last. In every generation, the expressions are ranked according to their ability to make a prediction. The ablest expression is designated the “best of generation.” Eventually, the problem is solved by the emergence of an expression with the highest possible predictive rate. The best of the final generation is designated the “best of run” and becomes the final solution.

Artificial Neural Networks

A neural network is a connected network of processing elements called neurons. A neuron occupies the computational ground between simple on/off memory bits and complex data processing units. It receives input values, represented typically as real numbers, on multiple input lines and computes a weighted sum based on these values. The neuron then uses the sum to compute an activation or threshold value. The activation value is passed to an output line that may in turn be passed as input to another neuron or may represent the final output of the network. A group of neurons, arranged in interconnected layers with judiciously chosen interconnection weights, can perform useful computations. Some neurons accept input

from the environment, others perform internal (hidden) calculations, and still others present the final result of the network's cogitation (Fig. 3).

Neural networks have several advantages. They are able to accommodate and represent a range of data types (continuous, binary, and categorical values) and they can be made relatively insensitive to missing data. They are ideal for classification studies and consistently outperform both logistic regression and discriminant analysis (both parametric and nonparametric). Neural networks have been used to examine heart function. They excel in classification, far exceeding the predictive ability of the more standard logistic regression.^[20,21] Consider an analysis comparing the quality of life for patients with two competing treatments (Table 3). The question to be answered was whether patients who had a lower quality of life routinely took the experimental treatment. Based upon the results of the quality-of-life scale, the model attempted to identify whether the patient was taking the experimental treatment, or was in the control group. Overall, the accuracy

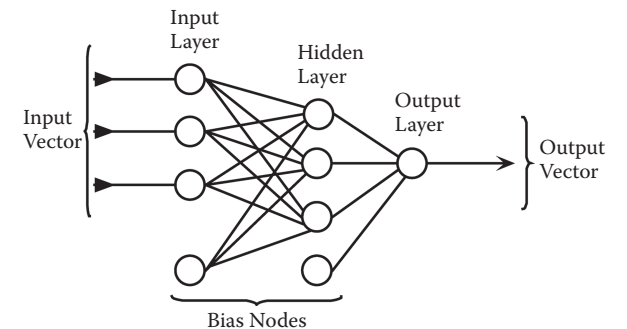


Fig. 3 Diagram of a neutral network

Table 3 Prediction of Treatment Group Based on Quality of Life Responses

Value	Training set neural network	Validation set neural network	Training set genetic programming	Validation set genetic programming
Sensitivity (%)	72	63	59	53
Specificity (%)	79	79	96	84
False-negative rate (%)	3.5	9	23	21
False-positive rate (%)	7	62	7	38
Accuracy (%)	76	72	81	74

of prediction was equal to 70% using logistic regression compared to 72% for neural network analysis and 74% for genetic programming.

Logistic regression, without splitting the data into training and validation sets, achieved sensitivity of 74% and a specificity of 51%, whereas the neural network achieved a sensitivity of 63% on the validation set (72% on the training set) but a 79% specificity, indicating a more accurate model.

CASE STUDY: DATA MINING AND POLYPHARMACY

Given many possible combinations of drugs and the high incidence of polypharmacy, data mining is particularly useful in examining the problem of polypharmacy, defined as when a patient takes more medications than are needed and often suffers from drug interactions.^[37] Many people are taking multiple medications for comorbid problems. It becomes difficult if not impossible to examine the problem of adverse effects without the investigation of drug interactions and the complex combinations of drugs. Data-mining techniques were initially designed to deal with issues of complexity; however, data-mining techniques need to be incorporated into investigations of polypharmacy and drug interactions. Antipsychotic prescribing was examined from March 1 to May 31, 2000. There were 26 nursing facilities serviced and 270 physicians treating patients. The data were collected by the consulting pharmacist who did not have access to any identifying information. The physicians were identified by number only, and only the total number of prescriptions for each physician identification number was made available to the investigator.

The number of prescriptions issued is given in Table 4. Of 2876 patients in 26 facilities, 520 patients (18%) were prescribed antipsychotics. For the 520 patients, 34 (6%) were prescribed two different antipsychotics. It is only when looking at global prescribing practices that abuses can be uncovered, and improved.

CONCLUSION

The techniques of data mining—link analysis, kernel density estimation, rule induction, text analysis, genetic programming, and artificial neural networks—can greatly

Table 4 Physician Prescribing Patterns

Drug	Number of prescriptions
Chlorpromazine	7
Clozapine	1
Fluphenazine	4
Haloperidol	49
Mesoridazine	4
Molindone	2
Olanzapine	62
Perphenazine	8
Prochlorperazine	14
Quetiapine	30
Risperidone	97
Thioridazine	28
Thiothixene	7
Trifluoperazine	1
Triflupromazine	1

facilitate the study of patient outcomes, particularly when it comes to investigating quality-of-life issues and polypharmacy. The techniques were designed to deal with large databases of clinical information routinely collected in the course of practice. The text mining techniques can also deal with the collection of paper in each patient chart. Therefore, surveillance of drugs and drug interactions can be much facilitated by the use of these techniques.

REFERENCES

1. Buckley, J.S., Jr. Planning for effective use of on-line systems. *J. Chem. Inf. Comput. Sci.* **1975**, 15 (3), 161–164.
2. Fozi, K.; Teng, C.L.; Krishnan, R.; Shajahan, Y. A study of clinical questions in primary care. *Med. J. Malays.* **2000**, 55 (4), 486–492.
3. Hersh, W.R.; Crabtree, M.K.; Hickam, D.H.; Sacherek, L.; Rose, L.; Friedman, C.P. Factors associated with successful answering of clinical questions using an information retrieval system. *Bull. Med. Lib. Assoc.* **2000**, 88 (4), 323–331.
4. Karanikolas, N.N.; Skourlas, C. Computer assisted information resources navigation. *Med. Inform. Internet Med.* **2000**, 25 (2), 133–146.

5. Rodgers, R.P. Searching for biomedical information on the World Wide Web. *J. Med. Pract. Manage.* **2000**, *15* (6), 306–313.
6. Brown, P.J.; Sonksen, P. Evaluation of the quality of information retrieval of clinical findings from a computerized patient database using a semantic terminological model. *J. Am. Med. Inform. Assoc.* **2000**, *7* (4), 392–403.
7. Beahler, C.C.; Sundheim, J.J.; Trapp, N.I. Information retrieval in systematic reviews: Challenges in the public health arena. *Am. J. Prev. Med.* **2000**, *18* (4 Suppl), 6–10.
8. Delaney, C.; Reed, D.; Clarke, M. Describing patient problems and nursing treatment patterns using nursing minimum data sets (NMDS and NMMDS) and UHDDS repositories. In *Proceedings/AMIA Annual Symposium*; 2000, 176–179.
9. Kaplan, B.; Brennan, P.F.; Dowling, A.F.; Friedman, C.P.; Peel, V. Toward an informatics research agenda: Key people and organizational issues. *J. Am. Med. Inform. Assoc.* **2001**, *8* (3), 235–241.
10. Humphreys, B.L. Electronic health record meets digital library: A new environment for achieving an old goal. *J. Am. Med. Inform. Assoc.* **2000**, *7* (5), 444–452.
11. Silver, M.; Sakata, T.; Su, H.C.; Herman, C.; Dolins, S.B.; O'Shea, M.J. Case study: How to apply data mining techniques in a healthcare data warehouse. *J. Healthc. Inf. Manag.* **2001**, *15* (2), 155–164.
12. Richards, G.; Rayward-Smith, V.J.; Sonksen, P.H.; Carey, S.; Weng, C. Data mining for indicators of early mortality in a database of clinical records. *Artif. Intell. Med.* **2001**, *22* (3), 215–231.
13. Spence, M. Visualization and interactive analysis of blood parameters with InfoZoom. *Artif. Intell. Med.* **2001**, *22* (2), 159–172.
14. Hobbs, G.R. Data mining and healthcare informatics. *Am. J. Health Behav.* **2001**, *25* (3), 285–289.
15. Helma, C.; Gottmann, E.; Kramer, S. Knowledge discovery and data mining in toxicology. *Stat. Methods Med. Res.* **2000**, *9* (4), 329–358.
16. Smyth, P. Data mining: Data analysis on a grand scale? *Stat. Methods Med. Res.* **2000**, *9* (4), 309–327.
17. Watterson, K. A data miner's tools. *Byte* **October 1995**, *20*, 91–92.
18. Elder, J.; Pregibon, D. A Statistical Perspective on Knowledge Discovery in Databases. In *Advances in Knowledge Discovery and Data Mining*; Fayyad, W. Piatetsky-Shapiro, G. Smyth, G. Uthurusamy, P., Eds.; MIT Press: Cambridge, MA, 1996.
19. Chen, S.X. Estimation in independent observer line transect surveys for clustered populations. *Biometrics*. **1999**, *55* (3), 754–759.
20. West, D.; West, V. Model selection for a medical diagnostic decision support system: A breast cancer detection case. *Artif. Intell. Med.* **2000**, *20* (3), 183–204.
21. West, D.; West, V. Improving diagnostic accuracy using a hierarchical neural network to model decision subtasks. *Int. J. Med. Inform.* **2000**, *57* (1), 41–55.
22. Burkhauser, R.V.; Cutts, A.C.; Lillard, D.R. How older people in the United States and Germany fared in the growth years of the 1980s: A cross-sectional versus a longitudinal view. *J. Geront., Ser. B Psychol. Sci. Soc. Sci.* **1999**, *54* (5), S279–S290.
23. Saint, D.A.; Chung, S.H. Variability of channel subconductance states of the cardiac sodium channel induced by protease. *Recept. Channels*. **1999**, *6* (4), 283–294.
24. Stephens, M.; Palakal, M.; Mukhopadhyay, S.; Raje, R.; Mostafa, J. Detecting gene relations from Medline abstracts. *Pacific Symp. Biocomput.* **2001**, 483–495.
25. Brebilla-Perrot, B.; Houriez, P.; Claudon, O.; Preiss, J.P.; de la Chaise, A.T. Evolution of QRS duration after myocardial infarction: clinical consequences. *Pacing Clin. Electrophysiol.* **1999**, *22* (10), 1466–1475.
26. Koza, J.R.; Mydlowec, W.; Lanza, G.; Yu, J.; Keane, M.A. Reverse engineering of metabolic pathways from observed data using genetic programming. *Pacific Symp. Biocomput.* **2001**, 434–445.
27. Malolepszy, A.; Kacki, E.; Dogdanik, T. Application of genetic programming for the differential diagnosis of acid–base and anion gap disorders. *Stud. Health Technol. Inform.* **2000**, *77*, 388–392.
28. Bojarczuk, C.C.; Lopes, H.S.; Freitas, A.A. Genetic programming for knowledge discovery in chest-pain diagnosis. *IEEE Eng. Med. Biol. Mag.* **2000**, *19* (4), 38–44.
29. Snow, P.B.; Kerr, D.J.; Brandt, J.M.; Rodvold, D.M. Neural network and regression predictions of 5-year survival after colon carcinoma treatment. *Cancer*. **2001**, *91* (8 Suppl), 1673–1678.
30. Bostwick, D.G.; Burke, H.B. Prediction of individual patient outcome in cancer: comparison of artificial neural networks and Kaplan–Meier methods. *Cancer*. **2001**, *91* (8 Suppl), 1643–1646.
31. Sargent, D.J. Comparison of artificial neural networks with other statistical approaches: Results from medical data sets. *Cancer*. **2001**, *91* (8 Suppl), 1636–1642.
32. Goldberg, H.G.; Senator, T.E. Restructuring databases for knowledge discovery by consolidation and link formation. In *AAAI Fall Symposium*; 2001. Available at: eksl-www.cs.umass.edu/aila/link-analysis.html. Accessed November 2001.
33. Silverman, B. *Density Estimation for Statistics and Data Analysis*; CRC Press: Boca Raton, FL, 1986.
34. Scott, D. *Multivariate Density Estimation: Theory, Practice, and Visualization*; John Wiley and Sons: New York, 1992.
35. Hedberg, S.R. Is AI going mainstream at last? A look inside Microsoft research. *IEEE Intell. Syst. Their Appl.* March/April **1998**, *13* (2), 21–25.
36. Weiss, S.M.; Apte, C.; Damerau, F.J. Maximizing text-mining performance. *IEEE Intell. Syst. Their Appl.* **July/August 1999**, *14* (4), 63–69.
37. Coulter, D.M.; Bate, A.; Meyboom, R.H.; Lindquist, M.; Edwards, I.R. Antipsychotic drugs and heart muscle disorder in international pharmacovigilance: data mining study. *BMJ, Br. Med. J.* **2001**, *322* (7296), 1207–1209.

Data Monitoring Committees (DMC)

Jay Herson, PhD

Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, U.S.A.

Abstract

This article describes the objectives, membership, creation, and organization of a data monitoring committee in industry-sponsored clinical trials. Both safety and efficacy monitoring are included as are a description of the charter, conflicts of interest, clinical, regulatory and statistical issues/methods, multiregional trials, use of MedDRA and standard MedDRA queries, sources of bias, and types and methods of decisions made by data monitoring committees including interim analyses for superiority and futility. The article concludes with emerging issues including adaptive designs, novel designs in oncology, centralized risk-based monitoring, fraud detection, and training of committee members.

Keywords: Adaptive designs; Charter; Clinical trials; Data monitoring committee; Futility analysis; Interim analysis; MedDRA; Patient safety; Safety analysis.

INTRODUCTION

A data monitoring committee (DMC) in an industry-sponsored clinical trial is an independent panel charged with preserving the integrity of the trial by protecting patient safety. The DMC must also ensure that interim analyses of superiority or futility of an experimental agent are performed correctly. If there needed to be only one word to describe the DMC's responsibility, it would be *stewardship*. The DMC might be thought of as the board of directors of the trial and the sponsor study team as the officers of the trial. We will see later that there are other types of independent committees that also interact with the sponsor on operational issues.

We will use the terms “treatment” or “drug” in various places, but these terms could apply easily to pharmaceuticals, biologics, or medical devices. The term “pharma” could apply to trial sponsors in either of the three previously mentioned industries. The terms “study drug” and “study treatment” denote either the experimental treatment or the control group treatment.

History

In the late 1960s, the National Institutes of Health (NIH) concerned that clinical trials under their sponsorship should have some form of independent review and felt that there was a need for some definition of best practices. The National Heart Institute, as it was known at the time, commissioned what has become known as the “Greenberg Report” as a way of beginning to deal with

these issues. The report was completed in May 1967 but not published until 1988.^[1]

The 1970s saw a considerable growth in pharmaceutical industry clinical trials. Although the NIH-sponsored trials were research oriented using already approved drugs, the pharmaceutical industry trials were more development or product oriented. They were designed to generate efficacy and safety data for submission to a regulatory agency for approval of an experimental treatment. One of the first DMCs in a pharmaceutical industry trial was the cimetidine stress ulcer clinical trial in 1988–1989.^[2] This trial was conducted in intensive care units and compared cimetidine to placebo for prophylaxis of upper gastrointestinal bleeding due to stress.^[3] The sponsor decided that an independent committee was needed to review the bleeding data in a masked (i.e., blinded to treatment assignment) manner rather than the sponsor so as to minimize the perception of bias. Because these were patients treated in emergency rooms, the DMC was also needed to determine whether deaths on treatment were disease related. This was the beginning of the DMC era for industry-sponsored trials.

As DMCs began to appear in pharmaceutical industry trials, regulatory agencies such as FDA and the International Conference on Harmonisation created guidance documents.^[4–7] These documents mentioned the advisability of independent review of safety and efficacy data. Ellenberg et al.^[8] and DeMets et al.^[9] wrote the first two books that dealt specifically with DMCs. These books drew on the authors' experience in NIH-sponsored trials. Herson's first and second editions deal with the evolving best practices for DMCs in the pharmaceutical industry.^[10,11]

Other Vehicles for Patient Protection

DMCs are not the only source of protection of patient safety. Each clinical site comes under the auspices of an Institutional Review Board (IRB), in some countries called an Ethics Committee, which reviews protocols and their amendments and receives periodic safety reports. Each sponsor has an internal masked safety review system, and some sponsors even have internal safety review committees. Some large research hospitals have their own internal DMC-like committees that review safety on ongoing trials regardless of sponsorship.

The FDA recently issued a guidance calling for a Safety Assessment Committees (SAC).^[12] This is a term likely to be confused with DMCs. The SAC, consisting of both internal and external experts in medicine, biostatistics, and epidemiology, has broader responsibilities than the DMC. The SAC will review data from a specific clinical trial but also from other trials on the same or similar drugs, epidemiologic studies, insurance claims data, and so on in order to get an overall impression of the safety profile of the experimental treatment. SAC implementation is not clear at this time and may be an activity more likely to be supported by only the largest firms in the industry.

Place of DMCs in the Development Cycle

DMCs, as described in this article, have become commonplace in Phase III/confirmatory trials. Earlier phase clinical trials (Phases I and II) might have some form of independent review but would not usually be as independent of the sponsor as the DMC for a confirmatory trial. Several papers in the literature give guidance on how independent review can be incorporated into early-phase trials if sponsors want to establish this type of input.^[13,14]

A current concern among regulatory agencies is the occasional blurring of lines between development phases. A recently approved oncology drug began as a Phase I trial, but, as favorable results were observed, more patients were added. Eventually, there were over 1200 patients on this protocol, and the 179 patients with advanced melanoma were considered sufficient for FDA approval.^[15] There was no DMC involvement.

Sponsor Size

We are assuming that the sponsor in the trials we discuss is a company—pharmaceutical, biotechnology, or medical device. The size of the sponsor drives many activities of the DMC.

Herson^[11] defines *Big Pharma* as a public company with many products on the market with a goal in the current trial to expand its product line, *Middle Pharma* as a public company with a few products on the market with a goal of expanding the product line and forming corporate partnerships or mergers with Big Pharma, and *Infant Pharma* as a

company that could be public but likely surviving on venture capital investments, no products on the market with a goal of showing positive progress of a successful product. Big Pharma and Middle Pharma have DMC procedures in place. This is usually not the case for Infant Pharma, and, indeed, procedures are often written when the trial is in progress and often at the advice of experienced DMC members. Infant Pharma is much more vulnerable to DMC decisions than Big Pharma or Middle Pharma, and thus, there is increased tension in the DMC–sponsor relationship in Infant Pharma trials.

ORGANIZATION OF A SAFETY MONITORING PROGRAM IN CONFIRMATORY TRIALS

Most of the data collected in confirmatory trials are safety related. As we will see later, DMCs are involved in safety assessment at and between periodic meetings but with efficacy analysis only at specified interim analyses if such decision points are specified in the protocol. In this section, we will describe how the DMC fits into the safety monitoring process.

Members of the Safety Monitoring Team

The members of the safety monitoring team in addition to the DMC will be the sponsor staff headed by the Sponsor Representative, the sponsor employee in charge of the clinical trial, the Data Analysis Center (DAC), and the IRB.

The DAC will prepare the tables, listings, and graphs for DMC review. The DAC representative will be a biostatistician unmasked to treatment assignment. The DAC biostatistician will attend DMC meetings. The DAC might be a division of the sponsor with an effective firewall between this DAC and the sponsor staff. However, it is considered better practice for the DAC to be an independent contract research organization (CRO) retained by the sponsor for this purpose.

The IRB or Ethics Committee exists at every institution participating in the trial. The IRB protects the patients of their own institution by inspecting the trial protocol, informed consent, investigators brochure, and so on. They will review safety data within their institution and only trial-wide safety data on adverse events that are serious, treatment related, and unexpected. All IRB members are masked to treatment assignment.

There are two other committees that may be involved in clinical trials. The *endpoint adjudication committee* will be masked and make decisions on efficacy endpoints that require interpretation. This will often entail reviewing images, EKGs, hematology reports, and so on. The *steering committee* advises the sponsor on the conduct of the trial. Although DMCs exist in virtually all confirmatory trials, the endpoint adjudication and steering committees exist in only a few trials.

DMC Creation

It is the sponsor's responsibility to recruit DMC members. A typical industry DMC consists of three members—two physicians and one biostatistician. These three are voting members and are generally recruited from academia and government. The two physician members are generally experts in the indication under investigation, e.g., oncologists, cardiologists, and rheumatologists. If certain adverse events are anticipated, such as cardiotoxicity in an oncology trial, a cardiologist might be added to the DMC. For some adverse event types, it might be advisable to enroll ad hoc consultants who would not be the members of the DMC but could advise the DMC when necessary. It is considered good practice not to disclose the names of DMC members to investigators or the public until after the trial is completed.

Conflicts of Interest

Each sponsor should have standard operating procedures (SOPs) for DMC creation and operations. These SOPs must define what would constitute a conflict of interest for a DMC member. The criteria would usually consist of consulting for or serving on a DMC for a competing company working on therapeutics for the same indication as the trial addresses. There would also be limits as to how much a DMC member could earn annually from the sponsor since the DMC member could also be advising the sponsor in other ways. Indeed if a DMC member received high compensation from the sponsor, the independence of this member would come under question. As a result, DMC members receive modest compensation. DMC service should always be seen as service to science and not a lucrative consulting opportunity. DMC members will be asked at every meeting if there have been any changes that might constitute a conflict of interest since the previous meeting.

Liability and Indemnification

There must be clearly a way to protect DMC members against liabilities that might occur in the trial for which they are not responsible. It is considered the best practice for the sponsors to include an indemnification clause in DMC member contracts. Although some contracts call for mutual indemnification, it makes little sense for DMC members to indemnify the sponsor. DeMets et al.^[16] provide useful guidelines for what should be included in an indemnification clause. Fleming et al.^[17] discuss the many issues involved in sponsor indemnification of DMC members.

MEETINGS

DMC Charter

The DMC Charter issued by the sponsor, with possible modifications by the DMC members, governs all meetings

and operations of the DMC. The Charter will specify DMC membership and responsibilities and conflict of interest policy. Additional items would include whom the DMC reports to, serious adverse event (SAE) data flow, DAC responsibilities, meeting minutes, and retention policy. Eventually, the DMC, sponsor, and DAC will agree on a Data Review Plan (DRP)—templates of tables, listings, and graphs to be reviewed at DMC meetings. The final version of this plan will be appended to and considered part of the Charter.

Orientation Meeting

Before the trial begins, the sponsor staff will convene a kickoff meeting with the DMC members and the DAC biostatistician. At this meeting, the sponsor will present the Charter and possibly preliminary drafts of the DRP. The orientation meeting is often face to face.

The sponsor will have already chosen the DMC Chair, and this person will cochair the orientation meeting, but the DMC Chair will run all subsequent meetings. The DMC members will select a recording secretary among themselves. This person will record the minutes of closed meetings.

Masking policy is an important component of the orientation meeting. The sponsor staff will certainly be masked to treatment assignment during the trial. DMC members will be either unmasked or partially unmasked. In the latter case, the treatments are referred to only as A and B throughout the trial. At any meeting after the orientation meeting, the DMC members may ask the DAC biostatistician to reveal the identity of A and B. This is done because many DMC members do not feel that they can do their job properly without knowing treatment assignment. If they remain only partially unmasked, they will inevitably try to guess the identity of A and B as adverse event reports accumulate. However, if they guess wrong, this could adversely affect decision-making throughout the trial. If DMC members decide to unmask, they do not have to inform the sponsor of this decision.

At some point, DMC members might have a safety concern that they would like to discuss with the sponsor. They cannot discuss this directly with the sponsor team because it might unmask them. Big Pharma and Middle Pharma sponsors will usually have a pharmacovigilance (PV) group separated from the sponsor team. Infant Pharma would not have such a group so the orientation meeting is a good time for the Infant Pharma sponsor to identify an independent consultant who could act as a PV. The PV would take on the task deciding what should be done and what information should be conveyed to the study team. It is good practice for all documents that circulate during a trial to put an asterisk after those persons who are known to be unmasked so as to minimize accidental unmasking.

The DMC members will also be asked to review the investigator brochure and the protocol at the orientation

meeting. The DMC is free to recommend the changes to the protocol such as adding or changing the frequency of stress echocardiography, modifying eligibility requirements, or inquiring about plans to ensure enrollment of minorities. Although each institution participating in the trial and its IRB has the responsibility to prepare its own informed consent agreements, the DMC members may also review and comment on the trial-specific expected adverse event information the sponsor has sent to the institutions for inclusion in their informed consent.

Agreements will also be made of the SAE data flow—72 h reporting adverse events and periodic reports of newly occurring adverse events. Usually, these reports will only go to the DMC Chair and he/she will alert other members as needed.

As noted previously, it is the Charter and DRP that will guide DMC operations. The sponsor will have already planned a statistical analysis plan and perhaps have followed the SPERT Committee guidelines^[18] and prepared a Program Safety Analysis Plan, but these plans refer to planned interim and final analyses of efficacy and safety. The DRP is a subset of these documents relating to the ongoing data review of the DMC. Plans would be made at the orientation meeting for sponsor–DMC–DAC collaboration to complete the DRP.

Herson^[11] provides an appendix with a complete list of potential DMC responsibilities and DRP contents.

Data Review Meetings

The first data review meeting will be scheduled after X patients have received Y cycles of treatment or after Z months from trial commencement, depending on which comes first. X , Y , and Z are the numbers to be specified based on the specifics of a trial protocol. This policy ensures that early-enrolled patients are not unduly at risk for lack of data review because enrollment is slow. There are usually two to three data review meetings each year, but the ultimate frequency depends on the specific clinical trial. Most data review meetings are by teleconference/WebEx.

The data review meeting has an open and a closed session. The sponsor team, DMC members, and DAC biostatistician attend the open session. The sponsor team will report the study progress and compare the projected and actual enrollment explaining the discrepancy, if there is one. The data cutoff for data compiled for this meeting will be announced. It may be 6–8 weeks prior to the meeting due to the need for data cleaning. A complete safety report will be presented with all treatment groups combined because the sponsor staff must remain masked. The sponsor staff will report on pending action items from previous meetings. This usually entails tasks with enrollment by center, noncompliance, and so on. Herson^[11] presents a sample agenda for an open session.

Only the voting DMC members and the nonvoting DAC biostatistician attend the closed session. The DMC members will go over the tables, listings, and graphs prepared by the DAC. They will follow up on previous safety issues and seek new concerns. If a planned interim analysis of efficacy is due at this meeting, this will be done as well. The precise methods to be used for these activities will be reported later in this article.

At the conclusion of the open session, the DMC Chair will report to the sponsor representative whether the trial can continue with or without protocol amendments.

CLINICAL ISSUES

Goals of Safety Analysis

The safety analyses undertaken by the DMC seek to separate the signal from noise, i.e., to distinguish adverse events that are part of the disease process, or preexisting or concurrent conditions, and those related to concomitant medication from those that are related to study treatment. We call the latter *treatment-emergent adverse events* but will use the shortened term *adverse events*. The latter can be defined as any unfavorable and unintended sign (e.g., including an abnormal laboratory finding), symptom or disease temporally associated with the use of a treatment, or whether considered related to this intervention.

A SAE is any untoward medical occurrence that, at any dose, results in death, is life threatening, requires inpatient hospitalization or prolongation of existing hospitalization, and results in disability/incapacity. This is a clinical/regulatory term, and the various regulatory agencies have expedited reporting requirements for SAEs. Sponsors and DMC members agree on a SAE data flow at their orientation meeting.

A purely clinical term that must not be confused with SAE is the *severe adverse event*. The latter refers to the intensity of the event but not necessarily the seriousness. A severe headache would usually not qualify as serious, but a mild case of dehydration would be considered a SAE. There are various ways to define the severity of adverse events. We will provide a brief description later.

Another category of adverse event is the serious unexpected suspected adverse reaction (SUSAR). This new definition comes from an FDA guidance communication to industry on safety reports,^[19] which has become known as the “Final Rule.” The SUSAR is an adverse event that the sponsor, not necessarily the DMC, believes is causally related to the experimental treatment. Closely related is the adverse events of special interest (AESI). The sponsor may prepare an AESI list prior to the beginning of the trial, but the DMC will suggest additions to both the SUSAR and the AESI list as the trial progresses.

Safety Data

The safety data originate on case report forms and are transmitted by the investigator sites to the sponsor or CRO where data cleaning takes place. Eventually, the coding of the adverse events will be through an adverse event dictionary such as *Medical Dictionary for Regulatory Activities (MedDRA)*.^[20] MedDRA is a hierarchical coding system beginning with a system organ class (SOC) and down to a preferred term for a condition within the SOC. Often there are many preferred terms within a SOC, and this contributes to the granularity of the SOC. DMCs must deal with granularity because similar adverse events might be spread over several preferred terms and, thus, clouding analysis. We will return to the issue of granularity later, but it is encouraging to report that progress has been made recently in the definition of *standardized MedDRA queries (SMQs)*. The SMQs were developed by the MedDRA organization using the advice of expert committees. An example of an SMQ would be major adverse cardiac event (MACE), which combines MedDRA terms such as “all-cause mortality,” “myocardial infarct,” and “target vessel revascularization.”^[21] In addition, some sponsors have created proprietary SMQs, and the DMC members often suggest additional SMQs as they review data.

Adverse event severity must also be coded. The common severity levels are “mild,” “moderate,” “severe,” or “life threatening.” Adverse events are coded into these categories using various dictionaries like *MedDRA* or *Common Terminology Criteria for Adverse Events*,^[22] the latter being the preferred dictionary for oncology trials.

Another important source of safety data for DMCs are the narratives. These are written descriptions of the adverse events prepared by masked members of the sponsor’s PV staff. The narratives are compiled after reviewing all information available including interviews with the investigator. The content of the narratives is often discussed during DMC closed sessions. Often the narratives are delayed because some text must be translated to English before the PV staff can prepare the narrative. A common format for narratives was developed by the Council for International Organizations of Medical Science and is often referred to as the CIOMS form.^[23]

Clinical Issues in Multiregional (Global) Trials

In the past few decades, there has been a considerable increase in conducting clinical trials on a global scale. Inclusion of India, China, and Eastern European sites is now common. Binkowitz and Ibia review problems encountered in multiregional trials for efficacy endpoints.^[24] What follows is a review of issues encountered by DMCs with safety endpoints.

There are geographic patterns of genetic variation that may affect adverse reactions. Different cultures vary in their propensity to self-report adverse events. In some

state-run European health systems, physicians have a financial incentive to hospitalize patients who experience adverse events. These patients would not usually be hospitalized in other countries. Hospitalization qualifies as a SAE, and, thus, the SAE listing is a pool of “real” SAEs and those generated by a financial incentive. Similarly, some European countries require sponsors to unmask whenever a SAE is reported. This can introduce bias in conduct of the trial so most sponsors would just have the PV group unmask rather than the study team.

If surgery is involved in the clinical trial, it can become problematic when surgical and anti-infective practices vary among countries. Similarly, standard of care is often a control group treatment in clinical trials, and the standard may vary among regions. Even when an already approved drug is the control treatment (active control), different control treatments may need to be used in different regions because some active control candidates have not been approved or are not available because of cost in some regions.

Clearly, the issues described previously can differentially affect the incidence of adverse events seen by DMC members. Some DMCs will request physician consultants from the different regions to help with interpretation of data in multiregional trials. Herson^[11] presents a comprehensive description of safety issues faced by DMCs in multiregional trials.

STATISTICAL ISSUES

Goal of Statistical Analysis

The goal of safety analysis is to separate signal from noise, i.e., treatment-emergent adverse events that are related to study treatment rather than the disease process, or preexisting or concurrent conditions. DMC members will review tables, listings, and graphs prepared by the DAC and occasionally use statistical methods. Unlike efficacy analysis, safety analysis does not hinge on the statistical significance of a single endpoint. Indeed, there might be statistical significance in the difference in the incidence of mild transient headache between treatments, but the DMC members would likely not consider this of clinical significance. Similarly, there might be two cases of renal failure in the experimental treatment group and none in the control group. Although this difference might not be of statistical significance, DMC members would likely consider it of clinical significance. In this section, we will show the important ways that DMC members can use statistical methods.

Useful Data Displays

DMC members will see many useful displays in the open session where there is no distinction between treatment groups. This would include enrollment by center, ethnicity,

and stage of disease. These tables will indicate where patients are coming from and the underproductive sites, which may need attention. A graph of cumulative patient enrollment by month because trial commencement vs. projected enrollment is useful to indicate trial progress. At one time, the projected enrollment was usually presented as a straight line, but, more recently, computer simulations taking sponsor experience with investigator site start-up and time to steady-state enrollment generate the projections which are usually not linear.

In the closed session, the first useful graph is the cumulative distribution of patient exposure to study treatment by month. The Kaplan–Meier graphical method,^[24] to be described next, is useful for this display. The *Y*-axis is the number of patients, and the *X*-axis is the number of months after treatment start. The curves are monotonically increasing and show the cumulative number of patients still on therapy by treatment arm and month. If the curves diverge, it may be evidence of difference between treatments in toxicity incidence, disease progression or mortality.

Of course, the major source of closed session DMC deliberations is the adverse event tables. These tables come with the highly detailed MedDRA hierarchy and are supplemented by the SMQs discussed previously and by treatment group denoted A or B. Sponsors usually offer a table classified by investigator's judgment of relatedness to study treatment, but most DMCs ignore this table preferring the conservative approach to assuming all adverse events could be study treatment related. The first tables are usually for all adverse events, and then come tables for all SAEs and then tables for adverse events with severity moderate or higher. For each of these tables, DMC members will first look at the comparison of treatment groups in the first section consisting of all adverse event types combined and later explore the individual adverse event types. Early in the trial, there will be few safety signals because of low frequencies in most of the cells.

Statistical Description

In this section, it is assumed that the reader is already familiar with most of the common statistical methods described. Thus, we will not provide detailed formulas. A useful reference is the textbook by Rosner.^[25]

Data reduction begins with the incidence calculation. The incidence of any adverse event frequency is the frequency divided by the number of patients in the treatment group. This is the most common calculation used by DMCs, but more useful is the rate/100 patient-years of exposure. DACs currently always report the total sample size at the top of each treatment group column, but they should also report the number of patient-years of exposure. The latter is simply the sum of exposure time to study treatment converted to years. The rate per 100 patient-years is the frequency multiplied by 100 divided by the number of patient-years. For example, suppose the frequency of moderate or higher

nausea in the experimental treatment is 55 patients, and the total number of months of exposure of this treatment group is 912. We convert months to years $912/12 = 76$ years. The rate per 100 patient-years then is $55 \times 100/76 = 72.4$. We will see in Section "Hypothesis Testing" how this experimental treatment rate can be compared to that of the control group.

For adverse event types that the DMC feels need closer inspection or the declared AESI described in above, the DMC might request a Kaplan–Meier graph of time to the first occurrence of the event. The Kaplan–Meier life table method is usually used to compare two treatment groups, and we will see how in Section "Hypothesis Testing". The one-sample use of Kaplan–Meier generates the landmark estimate. If the DMC wants an estimate of the cumulative adverse event incidence at 12 months, it would be improper to divide by the number of patients still on the trial at 12 months, but the Kaplan–Meier 12-month estimate read from the Kaplan–Meier curve is the landmark estimate. It provides a 12-month incidence that includes the exposure times of all patients who began the trial including those no longer being followed in the trial.

Hypothesis Testing

Tables prepared by the DAC for DMC use in closed sessions generally have frequencies and incidences computed but not *p*-values for treatment comparisons. The DMC biostatistical member might look over the tables and compute the summary statistics and *p*-values of observed differences (computed by chi-square test or Fisher's exact test) as a way of calling potential differences to the attention of the physician members. Of course, these comparisons are not planned, and multiplicity issues might cloud the inference. In addition, we often do not find statistical significance in these comparisons because the sample sizes are too small to yield the power to detect a difference. These are all points that the biostatistical member should explain to the physician members. Calculations that might be made with or without associated *p*-values would be relative risk and odds ratios. The rate per 100 patient-years described previously follows the Poisson distribution, and the ratio of two rates is called the Poisson rate ratio. This is a useful method. Of course, the Kaplan–Meier table method can be used to compare the treatments corrected for exposure, and a log-rank test can be used for statistical significance. Herson^[11] illustrates all of these methods with numerical examples from a high-profile clinical trial. He also presents the likelihood and Bayesian methods although these are rarely used in DMC work.

BIAS AND PITFALLS IN INTERPRETATION

What Is Bias?

Piantadosi^[26] defines bias as a systematic (nonrandom) error in the estimate of a treatment difference, i.e., effects

other than the treatments themselves. DMC members are concerned with sources of bias that can create distorted treatment differences in safety. There are other pitfalls to interpretation that do not meet the precise definition of bias, but we will discuss them in this section as well.

Knowledge of Treatment Assignment

The sponsor staff should make every effort to avoid accidental unmasking to treatment assignment. The PV staff, or a designated consultant acting in this capacity, is a big help in shielding study staff from accidental unmasking. The DMC members themselves may have decided to unmask themselves, and many believe this is the only way the DMC can protect patient safety. There is no evidence that DMC unmasking has contributed serious bias to the DMC operations.

Reporting Bias

If investigators are unmasked or de facto unmasked because they can distinguish treatments by the adverse event footprints they leave, there might be an underreporting of adverse events in the active control group because investigators feel that the adverse event profile of the active control treatment is well known. Clearly, historical data on adverse event levels for the historical control are available for comparison. Fay et al.^[27] have developed a method to test the statistical significance of SAE incidence in a clinical trial compared to a historical control. This method may be useful to DMCs who suspect underreporting.

When a trial is terminated early due to treatment superiority or non-inferiority in efficacy, there may be incomplete follow-up on adverse events after the termination because investigators have moved on to other tasks. DMCs must insist on complete follow-up and not relinquish their responsibilities just because the trial has ended.

Adverse events may be collected in a spontaneous or solicited manner.^[28] In the former, the study merely asks a patient what side effects, if any, the patient has experienced since the last visit. For solicited adverse events, the nurse will go over a list of adverse events, which are either on the AESI list or on the ones that are often forgotten on underreported. The solicited route would reduce underreporting of these adverse event types. DMC members can advise the sponsor on which approach appears most appropriate for a given clinical trial.

Bias at the Analysis Level

There are several sources of bias that can creep in at the time of data analysis, and DMC members should be aware of these possibilities. When a clinical trial is terminated early due to efficacy, there may be lingering safety questions that cannot be answered because of small numbers of patients. If the sponsor is providing continuing treatment for experimental

patients and crossover treatment for control patients, the DMC should insist on the remaining intact to analyze the accumulating data for this and other safety issues.

The issue of granularity (numerous preferred terms for an organ class) has been introduced in Section “Clinical Issues.” There we saw that the SMQs such as MACE in cardiovascular disease are a big help. Still there are many diseases where safety signals may be coded across preferred terms and organ classes. For example, congestive heart failure can be missed if adverse event reports are spread across terms such as “dyspnea,” “orthopnea,” and “fatigue.” Indeed a practice of intentional splitting can keep a SAE off the product label. In addition to the use of SMQs, when available, bias can be controlled by DMCs adopting a rule that if an entire organ class has only 1% incidence, then only the organ class will be analyzed as a whole. If the organ class has 5% incidence, then the preferred term analysis can be undertaken. It should be understood that interpretation of the MedDRA table is not strictly a statistical issue.

Competing risks is a long-standing analytic issue for biostatisticians and should not be overlooked by DMCs. In a clinical trial, there may be low incidence of an adverse event in one treatment group only because patients have discontinued treatment due to toxicity, progression, or death. These events are the competing risks to adverse event observation, and their presence means this treatment group cannot necessarily be declared at low risk to this adverse event. The low incidence could be merely because patients in that group have not been in the trial long enough to experience this adverse event. This underscores the importance of always considering exposure in analyses and for the DMC to always get reports from the DAC of patient-years exposure in each treatment group. In the section entitled “Hypothesis Testing” above it was suggested that treatment group differences might be analyzed by the Poisson rate ratio or the Kaplan–Meier life table method. These methods take exposure into account and can avoid biased inference due to competing risks.

DMC DECISIONS

Types of DMC Decisions

When trialists think of DMC decisions, they invariably concentrate on trial termination due to an efficacy endpoint demonstrating sufficient superiority or futility at a planned interim analysis. We will certainly describe interim analysis decisions below but will also review decisions the DMC might make regarding safety other than trial termination. These would include writing a “Dear Investigator Letter” informing investigators of potential risk and recommending changes in the informed consent document. Decisions of the DMC are advisory to the sponsor. It is the sponsor who will make the final decision taking the DMC opinion and other factors into account.

Decision-Making Environment

It is important to note that at most of their closed session meetings, the DMC is making decisions on safety without the knowledge of efficacy. Efficacy knowledge comes only at the planned interim efficacy analyses. Schactman and Wittes^[29] contrast the difference in safety analysis at the end of the trial by the sponsor, then unmasked, and analyses provided by the DMC with limited resources and, perhaps, only early efficacy data. Physician members of DMCs will often claim that they will use predefined efficacy metrics to follow the protocol on interim analysis decisions, but they do not consider this of use to them in deciding if patients are benefitting. The metrics are part of the clinical trial construct but are not the way they evaluate patients in practice. For example, a visual acuity measure would not persuade an ophthalmologist that a patient has benefitted from therapy for age-related macular degeneration if the patient still complains he/she cannot read the morning crossword puzzle.

When a Safety Issue Arises

As has been said previously, DMC concerns about differential SAE incidence will most often not be a result of computing a *p*-value. The clinical implications for exposing further patients to this treatment and the public health concerns of marketing this product will be paramount. Actions such as writing the Dear Investigator Letter, modifying the informed consent, modifying protocol for dose, eligibility, and schedule all have their unmasking potential, and care must be made in how this information is provided to the sponsor. If the DMC recommends trial termination due to safety, this should be taken as the beginning of a dialog with the sponsor rather than an absolute decision. The sponsor and PV unit may have further relevant information about the SAEs or other options.

In the event of safety concern, the biostatistical DMC member might want to warn other members of the “bad news travels first” phenomenon whereby SAE reports must be submitted within a short time frame, often 48 h, due to regulations, but reports of patients who do not experience this SAE float in at a slower pace. This could give a distorted view of the safety risk. It is often prudent to wait until more data arrive in order to avoid a hasty irreversible decision.

Planned Interim Analyses Regarding Efficacy

Many confirmatory trial protocols will have a planned interim analysis at which point the trial can be terminated due to demonstrated superiority of experimental treatment to control or by futility. These terminations would be justified by ethical and efficiency reasons if continuing the trial would not be likely to lead to a positive outcome (futility) or if there is already considerable evidence of

experimental treatment superiority. These analyses require unmasking, and thus, administering this interim analysis is an important DMC responsibility.

A hypothesis will be tested at the interim analysis and there will be a Type I error associated with this test. The significance level required for this test will have to be strategically selected so that, if the trial continues, the overall Type I error will be close to the customary 0.05. The process of multiple testing of a hypothesis is called *alpha spending*. The interim analysis will take place at an intermediate point(s) in the trial. As the analysis will be done in groups, we call this type of trial *group sequential*. In practice, there are at most two interim analyses and a final analysis in the trial. Two closely related methods for alpha spending are those of O'Brien-Fleming^[30] and Lan-DeMets.^[31] As an example of alpha spending, if there were to be one interim analysis halfway through the trial, the significance level for determining superiority might be 0.003 and 0.048 at the end of the trial, allowing for an overall Type I error of 0.05. The smaller significance level at the earlier time point makes sense because we would want only an extreme treatment difference to indicate the ethical necessity for terminating the trial.

Futility testing is also performed at interim analyses. This analysis computes an assessment of how likely we are to reject the null hypothesis of no treatment difference at the end of the trial under several assumptions of “drift” or trajectory of future efficacy results. The method most often used by DMCs for futility analysis is *conditional power*. Just as the term implies, this method considers the data accumulated at interim as given and then looks at several drifts, e.g., null hypothesis prevails, alternative hypothesis prevails, or current efficacy level prevails, and the power for null hypothesis rejection under each. No alpha spending adjustment is needed for futility analysis because the assessment is spending Type II error not Type I. Futility analysis should always have the modifier *binding* or *non-binding*. A common criterion for termination due to low conditional power (CP) is $CP < 0.20$. If a futility analysis is binding, then the trial will terminate if the condition is met. If nonbinding, a dialog with the sponsor will commence. The sponsor might want to continue the trial in order to get approval in other countries or to gather more data to plan a future trial. A review of methods and rationale for futility analysis in clinical trials is given by Herson et al.^[32]

Another assessment often made at interim analysis is *sample size reestimation*. We will review that method in Section “Adaptive Designs.”

Herson^[11] provides a numerical example illustrating the interim analysis methods described previously for a hypothetical rheumatoid arthritis trial.

Final Meeting

When the trial goes to completion, there are several potential DMC responsibilities such as reviewing safety

data in a final manuscript, in the integrated summary of safety, or on the product label. In practice, these tasks are rare because they take place so long after trial completion when all concerned have moved on to other responsibilities.

Interim Analysis for Infant Pharma Sponsors

Infant Pharma sponsors are especially vulnerable to trial termination or even to protocol modification or Dear Investigator Letters. They must report these events to their investors and often remind their DMCs of the seriousness of negative interim analysis results. Nevertheless, DMCs not only realize that their responsibility is to patients but are also aware of making hasty decisions.

EMERGING ISSUES

The art and science of clinical trials is constantly changing, and DMCs must reconsider the existing practices and invent new practices in order to continue to protect patient safety. The issues to be discussed herein concern novel trial designs, new data collection methods, training of DMC members, and administrative challenges.

Adaptive Designs

Surely, the most popular of the new clinical trial designs are the so-called *adaptive designs*. Dragalin^[33] and Gallo and Krams^[34] have provided a good overview of the types of designs proposed. Adaptive designs seek to accelerate drug development by allowing for using data collected during the trial to change some aspects of the design without changing the overall Type I error of the trial. The DMC will administer the adaptive design process described in the protocol without unmasking the sponsor and, using their knowledge of the safety data, not allowing a change that would place patients at greater risk of adverse events. This responsibility represents a considerable challenge to DMCs, and the best practices are only now emerging.^[11,35]

If a trial has more than two treatment arms representing different doses or drugs, an adaptive design might call for dropping an arm at an interim analysis. Similarly, in a two-arm trial, an interim analysis might call for changing the randomization from 1:1 to, e.g., 2:1 based on interim results. These changes cannot go into effect immediately or automatically. The DMC must consider the safety consequences of these changes. The treatment groups that have already presented safety concerns to the DMC might be now enrolling more patients than originally planned as a result of the adaptation scheme. The DMC will need some “white space” to talk to the pharmacovigilance group about the advisability of this type of change.

Another type of adaptive design is *sample size reestimation* and *seamless transition* from Phase II to Phase III. The methods of sample size reestimation are

summarized by Chuang-Stein et al.^[36] The methods use protocol-defined algorithms to increase the sample size to preserve the power in the face of increased variability or varying efficacy in the control group. Similarly, if Phase II data pass a protocol-defined efficacy criterion, the trial may proceed to Phase III and begin enrolling additional clinical sites and patients. Here again the DMC must decide if there is a safety justification for this change.

A popular sample size reestimation strategy is the Cui-Hung-Wang (CHW) method.^[37] It must be understood that, unlike other sample size reestimation methods, CHW is reestimating the sample size to deliver the desired power to the effect size observed at interim. This effect size will be smaller than the one deemed clinically significant at the start of the trial. This brings up considerable regulatory and ethical issues. It is the role of the DMC biostatistician to explain to the clinical DMC members that they must consider if they were comfortable with the observed safety profile at the original effect size are they still comfortable with toxicity when the effect size is as observed at interim. The clinicians must also be reminded that this effect size is smaller than what was originally considered clinically significant. This will take some thought on the part of the DMC and require communication with the PV unit. Surprisingly, many CHW designs are employed where DMC members are unaware that the method implies a change in effect size.

Novel Designs in Oncology

The recent interest in precision medicine in cancer treatment has brought the *umbrella* and *basket* trial designs to actual protocols. These are master protocol designs where treatments rotate in and out of the protocol, and statistical methods either drop treatments for lack of efficacy or promote them to Phase III trials. Here DMCs will have to have rotating members because the protocols never close. Perhaps each time a treatment is promoted to Phase III one master protocol, DMC member can leave that DMC to chair a DMC for the new Phase III trial. Of course, the DMCs will still have their safety monitoring responsibilities and will have to be careful to differentiate between the various treatments rotating through the trial.

Biosimilar Designs

Biosimilars are follow-on drugs for branded biologics just like generics are follow-ons for chemical-based drugs.^[38] Due to the manufacturing of biologics, biosimilars cannot be precise copies of the branded biologic as generics are considered for the chemical-based branded product. Thus, full-scale bioequivalence trials will often be required for regulatory approval. DMCs will have to concentrate on adverse events observed in the two treatment groups. This will not always be straightforward. In a breast cancer bioequivalence trial for the biosimilar of trastuzumab, e.g., patients in both arms will also

receive taxanes, but some investigator sites might use docetaxel and some paclitaxel. DMC members will have to be knowledgeable of which adverse events are related to which drugs, and there will not always be unanimous agreement on the differentiation.

Centralized Risk-Based Monitoring

For many decades, data quality and protocol adherence clinical research associates monitored investigator sites as employees of the sponsor or a contractor. The monitoring was accomplished by making multiple visits to the sites. The FDA recently issued a guidance document, indicating that site monitoring could now be centralized, and statistical and graphical methods could replace routine site monitoring.^[39] Sophisticated software is becoming available to accomplish this type of monitoring.^[40] As a result, DMCs will be receiving data quality metrics by site. Thus, keeping track of data quality and recommending site improvements will be another important responsibility of DMCs.

Fraud Detection

Closely related to centralized risk-based monitoring is the topic of detection of fraud in clinical trials. The data coming from the sophisticated software used might, occasionally, detect the evidence of possible fraud in the data submitted. Indeed, Herson^[11,41] has proposed a *Fraud Recovery Plan* to be written by sponsors before the onset of a trial and to be administered by a DMC whenever the possibility of fraud appears.

Training of DMC Members

Various methods for training and certification of DMC members prior to service have been proposed—workshops at professional meetings, online courses, apprenticeships, and so on. None of these methods have gained traction because of the belief that people with the required clinical and statistical expertise will always be available to advise sponsors. It would still make sense to have a ready pool of people who have never served on a DMC and allocate them only to DMCs with experienced members. It is not uncommon for vacancies to occur on DMCs for trials that have already passed the interim analysis stage. Surely, in this case, people from the inexperienced professional pool could be assigned as replacements for these vacancies as a means of training for future trials.

Mergers and Licensing

It is not uncommon for a sponsor to merge or be acquired by another company while a clinical trial is in progress. When this happens, there may be uncertainty as to who is in charge at the sponsor, and indeed, several key

sponsor study team personnel may leave the company. DMC members must insist on doing their job during this transition and not surrender their responsibilities until the successor organization has found a suitable replacement. In most cases, the same DMC will continue to operate under the new ownership.

CONCLUSION

The art and science of clinical trials is constantly changing, and this article has attempted to illustrate the traditional and emerging responsibilities of DMCs. In the future, data on adverse events and patient compliance may be arriving in real time from patient-wearable technologies and/or sensors in the patient's home. DMC members may be provided access to sponsor clinical trial databases and statistical software so that they can continuously perform analyses of their own rather than rely only on the periodic structured data displays coming from the DAC. Much work must be done in postmarket safety, and the emerging surveillance technologies will help with that. However, our understanding of new treatments begins with the original efficacy and safety profile, and DMCs will always have a vital role in this process.

Herson^[11] provides extensive tables on the content of a DRP and on DMC responsibilities as well as numerous case studies from actual DMC–sponsor interactions on clinical trials.

REFERENCES

1. Greenberg, B. Organization, review and administration of cooperative studies (Greenberg Report): A report from the Heart Special Project Committee to the National Advisory Heart Council, May 1967. *Control. Clin. Trials* **1988**, 9 (2), 137–148.
2. Herson, J.; Ognibene, F.P.; Peura, D.A. et al. The role of an independent data monitoring board in a clinical trial sponsored by a pharmaceutical firm. *J. Clin. Res. Pharmacoevidiol.* **1992**, 6 (2), 285–292.
3. Martin, L.F.; Booth, F.V.; Karlstadt, R.G. et al. Continuous intravenous cimetidine decreased stress-related upper gastrointestinal hemorrhage without promoting pneumonia. *Crit. Care Med.* **1993**, 21 (1), 19–30.
4. U.S. Food and Drug Administration. *Guidance for Clinical Trial Sponsors: Establishment and Operation of Clinical Trial Data Monitoring Committees*; 2006.
5. International Conference on Harmonisation. *E3: Structure and Content of Clinical Study Reports*; 1995.
6. International Conference on Harmonisation. *E6: Guideline for Good clinical Practice*; 1996.
7. International Conference on Harmonization *E9: Statistical Principles for Clinical Trials*; 1998.
8. Ellenberg, S.S.; Fleming, T.R.; DeMets, D.L. *Data Monitoring Committees in Clinical Trials: A Practical Perspective*; John Wiley & Sons: New York, 2002.

9. DeMets, D.L.; Furberg, C.D.; Friedman, L.M. *Data Monitoring in Clinical Trials: A Case Studies Approach*; Springer: New York, 2006.
10. Herson, J. *Data and Safety Monitoring Committees in Clinical Trials*; Chapman and Hall/CRC Press: Boca Raton, FL, 2009.
11. Herson, J. *Data and Safety Monitoring Committees in Clinical Trials*, 2nd Ed.; Chapman and Hall/CRC Press: Boca Raton, FL, 2017.
12. U.S. Food and Drug Administration. *Safety Assessment for IND Safety Reporting—Draft Guidance for Industry*; 2015.
13. Dixon, D.O.; Freedman, R.S.; Herson, J. et al. Guidelines for data and safety monitoring for clinical trials not requiring traditional data monitoring committees. *Clin. Trials* **2006**, *3* (3), 314–319.
14. Hibberd, P.L.; Weiner, D.L. Monitoring participant safety in Phase I and II interventional trials: Options and controversies. *J. Investig. Med.* **2004**, *52* (7), 446–452.
15. Patnaik, A.; Kang, S.P.; Rasco, D. et al. Phase I study of pembrolizumab in patients with advanced solid tumors. *Clin. Cancer Res.* **2015**, *21* (19), 4286–4293.
16. DeMets, D.L.; Fleming, T.R.; Rockhold, F. et al. Liability issues for data monitoring committee members. *Clin. Trials* **2004**, *1* (6), 525–531.
17. Fleming, T.R.; DeMets, D.L.; Roe, M.T. et al. Data monitoring committees: Promoting best practices to address emerging challenges. *Clin. Trials* **2017**, *14* (2), 115–123.
18. Crowe, B.J.; Xia, H.A.; Berlin, J.A. et al. Recommendations for safety planning, data collection, evaluation and reporting during drug, biologic and vaccine development: A report of the safety planning, evaluation and reporting team. *Clin. Trials* **2009**, *6* (5), 430–440.
19. U.S. Food and Drug Administration. *Guidance for Industry and Investigators: Safety Reports Required for INDs and BA/BE Studies*; 2012.
20. MedDRA Maintenance Support and Services Organization. 2016. www.meddra.org.
21. MedDRA Maintenance Support and Services Organization. *Introductory Guide for Standardized MedDRA Queries*, Version 16. 2013.
22. U.S. National Cancer Institute Cancer Therapy Evaluation Program. *Common Terminology Criteria for Adverse Events v3.0*. 2016.
23. Council for International Organizations of Medical Sciences. *Development Safety Update Report. Harmonizing the Format and Content of Periodic Safety Reporting during Clinical Trials: Report of the CIOMS Working Group VII*. Geneva, CIOMS, 2006.
24. Binkowitz, B.; Ibia, E. Multiregional clinical trials: An introduction from an industry perspective. *Drug Inf. J.* **2011**, *45* (5), 569–573.
25. Rosner, B. *Fundamentals of Biostatistics*, 8th Ed.; Cengage: Boston, MA, 2015.
26. Piantadosi, S. *Clinical Trials: A Methodological Perspective*, 2nd Ed.; John Wiley & Sons: Hoboken, NJ, 2005.
27. Fay, M.P.; Huang, C.-Y.; Twum-Danso, N.A. Monitoring rare serious adverse events from a new treatment and testing for a difference from historical controls. *Clin. Trials* **2007**, *4* (6), 598–610.
28. Wernicke, J.F.; Faires, D.; Milton, D. et al. Detecting treatment-emergent adverse events in clinical trials—A comparison of spontaneously reported and solicited collection methods. *Drug Safety* **2005**, *28* (11), 1057–1063.
29. Schactman, M.; Wittes, J. Why a DMC safety report differs from a safety section written at the end of a trial. In *Quantitative Evaluation of Safety in Drug Development: Design, Analysis and Reporting*; Jiang, Q., Xia, H.A., Eds.; Chapman and Hall/CRC Press: Boca Raton, FL, 2015, Chapter 5.
30. O'Brien, P.C.; Fleming, T.R. A multiple testing procedure for clinical trials. *Biometrics* **1979**, *35* (3), 549–586.
31. Lan, K.; DeMets, D. Discrete sequential boundaries for clinical trials. *Biometrika* **1983**, *70* (3), 659–663.
32. Herson, J.; Buyse, M.; Wittes, J. On stopping a randomized clinical trial for futility. In *Designs for Clinical Trials: Perspectives on Current Issues*; Harrington, D. Ed.; Springer: New York, 2012, Chapter 5.
33. Dragalin, V. Adaptive designs: Terminology and classification. *Drug Inf. J.* **2006**, *40* (4), 425–435.
34. Gallo, P.; Krams, M. PhRMA working group on adaptive designs: Introduction to the full white paper. *Drug Inf. J.* **2006**, *40* (4), 421–423.
35. Herson, J. Coordinating data monitoring committees and adaptive clinical trial designs. *Drug Inf. J.* **2008**, *42* (4), 297–301.
36. Chuang-Stein, C.; Anderson, K.; Gallo, P. et al. Sample size estimation: A review and recommendations. *Drug Inf. J.* **2006**, *40* (4), 475–484.
37. Cui, L.; Hung, M.H.J.; Wang, S.J. Modification of sample size in group sequential clinical trials. *Biometrics* **1999**, *55* (3), 853–857.
38. Chow, S.-C. *Biosimilars: Design and Analysis of Follow-On Biologics*; Chapman and Hall/CRC Press: Boca Raton, FL, 2014.
39. U.S. Food and Drug Administration. *Guidance for Industry. Oversight of Clinical Investigations. A Risk-Based Approach to Monitoring*; 2013.
40. Venet, D.; Doffagne, E.; Burzykowski, T. et al. A statistical approach to central monitoring of data quality in clinical trials. *Clin. Trials* **2012**, *9* (6), 705–713.
41. Herson, J. Strategies for dealing with fraud in clinical trials. *Int. J. Clin. Oncol.* **2016**, *21* (1), 23–37.

Diagnostic Device Evaluation

Meijuan Li, Tinghui Yu, Qin Li, Gene Pennello, R. Lakshmi Vishnuvajjala,
Bipasa Biswas, Changhong Song, and Lilly Q. Yue

*Division of Biostatistics, Center for Devices and Radiological Health,
U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.*

Abstract

A diagnostic medical device is used alone or with other information to help assess a past, present, or future health condition of a subject. This entry provides general considerations for measurement validation and clinical validation of diagnostic devices from a statistical perspective. This entry also discusses issues with diagnostic device training and validation as well as decision analysis.

Keywords: Clinical performance; Analytical performance; ROC; Classification; Decision threshold; Diagnostic accuracy; Predictive value; Decision theory; Training and validation.

INTRODUCTION

A diagnostic medical device is an instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article, including a component part or accessory that is recognized in the official National Formulary, the United States Pharmacopoeia, or any supplement to them, intended for use in the diagnosis of disease or other conditions.^[1]

Diagnostic medical devices provide results that can be used alone or with other information to help assess a patient's past, present, or future health condition of interest (also known as target condition). Diagnostic devices encompass in vitro and in vivo diagnostic devices. In vitro diagnostic (IVD) devices are intended for use in the collection, preparation, processing, and examination of a biological sample taken from the human body, such as blood tests for signs of infections. In vivo diagnostic devices are intended for use within a whole and living organism, which provides an anatomical measurement (e.g., bone density, brain volume, and retinal thickness) or a physiological measurement (e.g., heart rate, uterine contractions). Although the term diagnostic is often associated with assessing the presence or absence of a disease/condition at the time of testing, the term as used in this entry is much broader than that. Diagnostic devices can also provide results useful for predicting a future clinical event, outcome, or condition. Intended uses for diagnostic devices include:

- diagnosis of symptomatic subjects
- screening of asymptomatic subjects
- monitoring patient status over time continuously (e.g., monitoring heart rhythms) or intermittently (e.g., monitoring cardiac health by electrocardiogram)

- monitoring patient treatment response (e.g., in leukemia patients with minimum residual disease)
- predicting a future state of health or clinical outcome in diseased subjects (prognosis) or disease-free subjects (risk prediction), and
- identifying patients who are most likely to respond to a treatment (companion diagnostic device)

The validation of a new diagnostic device includes both measurement validation and clinical validation. Measurement validation involves characterization of various aspects associated with a device's ability to detect or measure qualitative or quantitative attributes (measurand). Clinical validation refers to characterization of a device's ability to demonstrate clinically meaningful results, e.g., diagnose and monitor the target condition (condition of interest), or to predict the onset of a future condition or a treatment response. It is important to validate the measurement performance of the device before conducting the clinical validation study. For example, if measurement validation indicates poor measuring ability, then the developer may need to consider modifying the device before conducting a clinical validation study. Clinical validation involves two types of studies—diagnostic clinical performance study and clinical outcome study.^[2] A diagnostic clinical performance study of a medical test generates evidence on diagnostic accuracy, e.g., characterized by sensitivity and specificity, diagnostic positive and negative likelihood ratio (PLR/NLR), or positive and negative predictive values (PPV/NPV) in an intended use population.^[2,3] The intended use describes the clinical purpose, type of test, measurand, on what it measures (e.g., specimen type), where it measures (e.g., anatomical location), operation setting (e.g., laboratory, home use), and the population(s)

the test is intended for. For example, a clinical performance study may be a study comparing the result from the device with the result from the clinical reference standard (also commonly referred as “gold standard”) that is used to assess subjects for the target condition of interest.^[3] In contrast, a clinical outcome study generates evidence on whether the incorporation of a diagnostic device improves clinical outcomes on subjects.^[2] In a clinical outcome study, the device result is used during a treatment or management intervention. A clinical outcome study might be a randomized trial where the intent is to assess the impact of test results on clinical outcome by comparing clinical outcomes in one group in which the test results are used against another group in which the test is not used at all. Device performance is assessed in part by the intervention’s effect on a subject’s outcome.^[2] For example, a clinical outcome study is used to evaluate an implantable wireless monitoring system designed to provide pulmonary artery pressure of patients with heart failure. The clinical validity is assessed by comparing the heart failure-related hospitalization rate between the treatment group of patients who were managed by device results and the control group of patients who were not managed by device results. Diagnostic devices produce quantitative, semiquantitative, or qualitative (binary, nominal) results.

This entry provides general considerations of measurement validation and clinical validation performed after the diagnostic device has been finalized (i.e., the design parameters and the cutoff(s) have been finalized and locked down). Additional topics discussed in the entry are benefit risk analysis, training, and validation.

MEASUREMENT VALIDATION

Measurement validation of a device output is the characterization of various aspects associated with a device’s ability to detect or measure a qualitative, semiquantitative, or quantitative attributes (e.g., measurand^[4]). Throughout this section, the emphasis is on estimating and validating the measuring capability of a device.

For IVD devices, measurement validation is typically referred to as analytical validation. A device’s measurement performance can be impacted by preanalytical, analytical, and postanalytical factors. Preanalytical factors include specimen collection (timing, aliquoting, pipetting, retrieval, processing, etc.) as well as handling and storage (time, temperature, humidity, volume, etc.). Postanalytical factors include data entry and calculations by laboratory staff, interpretation of the result, data transfer, and the method used to report the results (electronic, paper, or telephone), etc.

Measurement validation for IVD devices consists of a range of analytical studies, including, but not limited to, an assessment of measurement bias, measurement precision (repeatability, intermediate precision, and reproducibility),

limits of detection and quantitation, limit of blank, reference intervals, linearity, matrix effect, stability (of the device and the measurand), interferences, sample commutability, cross-hybridization, and carryover, etc. The types of measurement validation studies required for a diagnostic device may depend on its intended use. The definitions of the terminology in this section are provided in literatures.^[5–13]

Limit of detection is defined as the lowest amount of analyte in a sample that can be detected with (stated) probability.^[8] Determining the limit of detection can be different for a device utilizing polymerase chain reaction and other types of tests.^[8] Linearity is the ability (within a given range) to provide results that are directly proportional to the concentration (amount) of the analyte in the test sample.^[9] Linearity assessment is an important part of measurement validation study for devices with quantitative results; while it may not be applicable for devices with qualitative results. The other important aspects of measurement validation commonly assessed are measurement bias and precision.^[5–7] Measurement bias for quantitative tests is the difference between the expectation of test results and the true value. Assessment of measurement bias involves the comparison between the new test and a reference standard for the true value. In a method comparison study, regression method can be used to estimate bias at any point along the measuring interval.^[4] When an imperfect comparator subject to measurement error is used, then systematic difference between the new device and the comparator will be evaluated.

Measurement precision is the closeness of agreement between replicate measurements on the same or similar object (e.g., sample) under specified testing conditions. The specified testing conditions can be repeatability conditions, reproducibility conditions, or some other set of intermediate conditions. The variables that may impact the device’s precision are testing site, reagent lot, operator, instrument, day, run, replicate, etc. When the device results are interpreted by readers, it may be important to evaluate the reader-to-reader variability when possible. For example, for immunohistochemistry (IHC) test systems and diagnostic imaging devices, reader is required to obtain final results. For categorical or binary measurements, precision is typically assessed by a measure of agreement of test results among the replicates.

Both bias and precision can be constant or vary across the measuring interval (low to high) of the device. Therefore, it is important to evaluate the measurement performance of a device at medical decision points for non-qualitative devices and near the device cutoff for qualitative devices. The cutoff here is defined as the threshold at which a qualitative device differentiates a positive (or disease) from a negative (non-disease) outcome. Evaluation of measurement validation may depend on the intended use and the technological characteristics of the test.

For tests that measure and record anatomical or physiological characteristics in vivo (e.g., electrocardiogram, imaging), the concept of measurement validation is important but may be difficult to assess in practice due to practical logistics or ethical concerns. For these devices, measurement validation may include animal studies, sometimes called test–retest reliability or bench testing. Numerous consensus standards for the analytical evaluation of medical tests are available to assist laboratories and manufacturers to evaluate medical tests.^[5–15]

CLINICAL PERFORMANCE

Clinical validation of a diagnostic device is the process through which one examines and provides objective evidence to confirm that the test results are clinically meaningful in the intended use population of the device. The study design, performance measures, and hypothesis testing as well as the statistical analyses are usually based on the intended use of the device. For example, a device used for disease diagnosis is evaluated for its diagnostic accuracy (true positive and negative fractions). In the evaluation of a screening test, the true-negative fraction may be emphasized. The validation of a monitoring device involves repeated testing on a subject over a specified time period, with the relationship between the target condition and the temporal pattern of the test results being of primary interest.

A diagnostic device with qualitative results is designed to determine whether a target condition is present or absent in a subject from the intended use population. The results of a diagnostic device can be binary (e.g., diseased vs. non-diseased), or ordinal (e.g., scores on IHC stained slides 0/1+/2+/3+), or multicategorical (nominal, e.g., genotype A/C/T/G). Analyses developed for binary outcome variables can be generalized to multinomial distributions. This entry will only focus on diagnostic devices with binary outcome. The clinical performance evaluation involves assessment of agreement between the new device output and the underlying “clinical truth,” determined by a clinical reference standard (Gold standard). The output of the clinical reference standard, whether the condition is present or absent, needs to be independent of the output of the new device under evaluation.

For a binary-valued test with possible results $T = 0, 1$ denoting test negative (0) and test positive (1) results, test accuracy is commonly evaluated with measures of classification and prediction of the absence or presence of disease $D = 0, 1$.^[11] Measures of classification involve the conditional probabilities.

$$\tau_d = P(T = 1 | D = d) \quad (1)$$

A common pair of classification measures is the *true-positive fraction (sensitivity)* τ_1 and the *true-negative*

fraction (specificity) $1 - \tau_0$, which are respectively the proportion of disease positive subjects that test positive and disease negative subjects that test negative. In order for a test to be informative, the test performance needs to satisfy $\tau_1 > \tau_0$.

Another important pair of classification measures is the positive likelihood ratio (PLR) and NLR τ_1/τ_0 and $(1 - \tau_1)/(1 - \tau_0)$, respectively. For example, if $PLR = 2$, then a test positive result is twice as likely in a diseased subject than a non-diseased subject. If $NLR = 0.5$, then a test negative result is twice as likely in a non-diseased subject than a diseased subject.

Measures of prediction involve the conditional probabilities:

$$p_i = P(D = 1 | T = i) \quad (2)$$

A common pair of prediction measures is the positive predictive value (PPV) p_1 and the NPV $(1 - p_0)$. Predictive values can be of prime importance to subjects and to their attending physician who is making a management or treatment decision based on the test result. However, PPV and NPV depend on the pretest probability (prevalence) of disease p . If disease is rare, sometimes studies are enriched for subjects with disease, which distorts the observed prevalence, and thus invalidating PPV and NPV estimated from the enriched study.

Fundamental relationships between prediction measures PPV and NPV and classification measures PLR and NLR are:

$$\begin{aligned} PPV &= \frac{p}{\left\{ p + \frac{1-p}{PLR} \right\}} \\ NPV &= \frac{1-p}{\left\{ (1-p) + p * NLR \right\}} \end{aligned} \quad (3)$$

In particular, PPV and NPV are monotone functions of PLR and NLR, respectively. Thus, even in studies enriched for disease, PPV and NPV can be estimated for a range of prevalence using the above relation.

If a clinical reference standard is not available, the true status of the subject may not be known with certainty. Thus, sensitivity and specificity of the new device cannot be estimated. Instead, one may validate a new device by comparing the new device to an established comparator, and such a study could involve a performance based on the new device’s positive percent agreement (PPA) and negative percent agreement (NPA) with the comparator. The estimations of the PPA and NPA are exactly the same as those of the sensitivity and specificity except that the basis of comparison is defined by the comparator, which has its own bias and imprecision when being compared to the clinical truth if there is any.^[3]

A summary performance measure of diagnostic accuracy is the area (AUC) under the receiver operating characteristic

curve (ROC).^[14,16,17] A ROC plot is a graphical description of test performance representing the relationship between the true-positive fraction (sensitivity) and false-positive fraction (1-specificity) mapped to the complete spectrum of assay cutoffs. A ROC analysis can be used for determination of the optimal cutoff value of a test. It is also frequently seen in comparing two diagnostic devices with continuous or ordinal measurement. The AUC represents the average device performance in ranges of sensitivity/specificity, which may have limited clinical interpretability.

Some devices are designed to provide prognostic or risk prediction rather than to diagnose any specific clinical conditions. For example, detection of certain gene expressions may be linked to higher risk of recurrence in diagnosed and treated breast cancer patients.^[18] The true patient status in this case is defined by the clinical outcome observed through follow-up over a period of time. The risk prediction given by the diagnostic device at the baseline or through monitoring can be compared to the so-called truth observed at follow-up time.

Another type of diagnostic device produces quantitative/continuous output (interval scaled) value, e.g., a quantitative measurement [e.g., B-type natriuretic peptide (pg/mL) in whole blood or plasma as an aid in diagnosing severity of congestive heart failure], a ratio (e.g., D-dimer/fibrinogen ratio to diagnose pulmonary embolism), and a compositional result (e.g., percent of tumor-infiltrating immune cells exhibiting Programmed death-ligand 1 (PD-L1) expression, as detected in IHC staining, as a biomarker for likelihood of response to cancer immunotherapy). The validation studies for such devices usually focus on the quantitative relationship between the device output and clinical outcome. For instance, the efficacy of a medicine targeting patients with specific genetic mutations may be highly correlated with the gene expression scores reported by the medicine's companion diagnostic. In this case, the researchers may want to demonstrate the effectiveness of the diagnostic device using survival analysis. The connection between the patients' survival hazard and the device output can be established by fitting a Cox proportional hazards model to the collected time-to-event data with the gene expression scores as a major covariate. An alternative option is to model the occurrence of the event of interest (e.g., death of the patient) using a generalized linear model involving the gene expression scores. Various study designs and statistical modeling can be used to validate a quantitative device, depending on its intended use.

In contrast, some quantitative devices are aimed to measure quantity of a measurand, e.g., a blood glucose measurement or a red blood cell count obtained from a patient's capillary blood sample. In this case, an appropriate measurement validation study can constitute the clinical validation study by comparing the measurement to a reference method or any established comparator. The developer is expected to collect sufficient number of clinical samples/device measurements from subjects covering the claimed measuring

interval of the new device. Equivalence of the quantitative measurement between two devices is usually demonstrated using the agreement between the results reported by the new device and the reference/comparator device. Agreement can be assessed via a linear regression model allowing for measurement error in measurements from both devices.^[4] In addition, bias (difference) plots are good complements to the regression analyses.^[19] Guidelines and recommendations for study designs and data analyses used to validate quantitative diagnostics are available from many sources.^[4]

Often diagnostic devices involve human readers to make interpretation and assign a diagnosis, e.g., mammography for breast cancer, whole slide imaging for primary pathological diagnoses, etc. The setup of laboratory environment, readers' skills, and their operating scales are known to have large variability and can play an important role in the performance of the imaging devices.^[20,21] To evaluate the clinical performance of such devices, it is best to consider incorporating the readers' variability in addition to the case, system, and site variability in the study design and analysis. A study design often used in this area is the so-called multireader multicase (MRMC) study in which every reader reads all modalities (e.g., an investigational technology vs. the current practice) for every case. It is important for the reading schemes in MRMC studies (e.g., sequential or crossover design) to be aligned with the intended use of the device (e.g., adjunctive, concurrent use, or replacement). To avoid bias in the estimation of the device performance, the readers need to be properly blinded to any information related to the identity and reference diagnosis of the enrolled samples. Reading schemes involving multiple sessions, randomized reading orders, and washout periods of sufficient length can be helpful for removal of the memory effect and reading session effect.

The heterogeneity of the study populations is of concern. To evaluate whether the device performance is dependent on the patient profiles, the environment, and setup of the medical practice, it may be desirable to perform the validation studies at multiple geographic locations and enroll patients of different age, gender, and race to cover the complete spectrum of the intended use population. In case some of the covariates have significant effects on the device performance, one may choose to report the average device performance estimated from the pooled data together with the estimates stratified by any control factors of interest. Some alternative statistical methods such as the Cochran-Mantel-Haenszel test can also be used to evaluate the association of test result with disease status after adjustment for strata (categorical covariates) that may be correlated with both variables. Stratified analysis can be helpful in examining the device performance in various subgroups. Any multiple testing needs to be prespecified with appropriate allocation of alpha in order to control type I error.

The validation studies are considered successful if the performance metrics/end points can be shown to exceed prespecified acceptance criteria dictating that the study

results have clinical significance. The desired performance level varies with the intended use of the new device. A new device intended to replace an older technology may require evidence of non-inferiority or even superiority when they are both compared to the clinical reference standard. A particularly interesting example rises in the development of companion diagnostics such as the mutation tests for epidermal growth factor receptor (EGFR) exon 19 deletions and exon 20 (T790M) substitutions. These tests target the mutations known to have significant impact on the treatment effect of certain EGFR tyrosine kinase inhibitors on diagnosed non-small-cell lung cancer patients.^[22]

Studies evaluating the performance of diagnostic devices could result in missing data, in particular missing test results and/or the clinical reference standard. Reasons for missing data may include invalid or uninterpretable test results, insufficient or unevaluable specimen, lack of informed consent, lost follow-up, etc. Simply reporting the naive estimates of diagnostic performance only on subjects with complete data can be biased and misleading. The *intent-to-diagnose* (ITD) principle illustrates not to discard subjects with missing data from an analysis of diagnostic performance.^[23] ITD is analogous to the well-known intent-to-treat principle in therapeutic randomized trials to analyze patients as randomized to preserve the integrity of the randomization. Some diagnostic devices provide an indeterminate or equivocal result. While these results are not missing, they too are sometimes discarded from analysis, which can similarly bias diagnostic performance. Missing data can frequently be reduced if types of potential missing data can be identified at the planning stage and study procedures can be developed to prevent or obtain missing data. Missing data may not be rare even in a very well-designed study with strict compliance. Reporting the percentage and reasons of missingness provides relevant information for further investigation. When meaningful for analysis, multiple imputation of the missing data can be considered. When appropriate, robustness analyses using proper imputation techniques can be indispensable for the unbiased evaluation of the device performance.

DECISION THEORY FOR DIAGNOSTIC TESTS

To facilitate decision-making in the medical management of subjects, one or more thresholds or cutoffs in a continuous test result are often applied. For a single cutoff c , subjects can be defined as test positive or test negative based on the cutoff, $t = \phi(x) = I(x > c)$. Without loss of generality, we will denote the binary test result $t = \phi(x) = I(x > c)$. Test negative and positive results may indicate, respectively, likely absence or presence of disease.

An important question is how to set the cutoff to meet clinical objectives. Statistical decision theory can be used to select a cutoff that minimizes the expected loss associated with test classifications, according to a loss function.

Suppose the clinical “loss” incurred from a false-positive result relative to a false-negative test result is r and that true-positive and true-negative losses are 0. If $d = 0, 1$ indicates absence and presence of a disease with prevalence p , then the loss function is:

$$L(c, d) = t(1 - d)r + (1 - t)d \quad (4)$$

which has expectation:

$$E[L(c, d)] = \tau_0(c)(1 - p)r + (1 - \tau_1(c))p \quad (5)$$

where,

$$\tau_d(c) = \Pr(t = 1 | d) = \Pr(x > c | d) \quad (6)$$

i.e., $\tau_0(c)$ is the false-positive fraction ($1 - \text{specificity}$) and $\tau_1(c)$ is the true-positive fraction (sensitivity) of the test at threshold c . Considering the ROC plot of $(\tau_0(c), \tau_1(c)) = (s, \text{ROC}(s))$ for every cutoff c in x , expected loss may be reexpressed as:

$$E[L(s, d)] = s(1 - p)r + (1 - \text{ROC}(s))p \quad (7)$$

Setting to zero the derivative of expected loss with respect to false-positive fraction s , we find it is minimized when:

$$\frac{\partial \text{ROC}(s)}{\partial s} = \frac{r}{\theta} \quad (8)$$

where $\theta = p/(1 - p)$ is the pretest odds of disease.^[17] In words, the optimal cutoff in terms of minimizing expected loss corresponds to that point on the ROC plot at which the derivative (slope of tangent line) is r/θ . This theoretical result can be extremely useful not only in the development of an operating point (cutoff) for a test but also for comparing the operating points of two tests with the same intended use.

For a point on the ROC plot corresponding to cutoff c , the derivative is the diagnostic likelihood ratio $\tau'_1(c)/\tau'_0(c)$, where $\tau'_d(c) = \Pr(x = c | d)$. Thus, when deciding if disease is present or absent, the Bayes decision rule minimizing expected loss given test result x declares disease to be present if:

$$p_x/(1 - p_x) \geq r \quad (9)$$

or equivalently,

$$p_x \geq r/(1 + r) \quad (10)$$

where $p_x = \Pr(d = 1 | x)$ is known as the *risk score* or predictive value of x for disease.^[24] This Bayes rule holds in general when the risk score is a function of a vector of measurements x . The risk score may be used to set the optimal cutoff. It is appropriate even if based on a retrospective case-control study in which the prevalence is

incorrect because the risk score is a monotonic function of the likelihood ratio, which is independent of prevalence.^[24]

More generally, non-zero losses may also be assigned to true-positive and negative test results.^[24,25] For example, while a true positive correctly classifies a subject's disease status, it nonetheless may lead to an invasive medical procedure with attendant risks of adverse events.

Decision theory for diagnostic tests may be formulated equivalently with utilities rather than losses. For risk prediction, expected utility of the test relative to a perfect risk prediction may be plotted as a function of threshold in risk at which a patient is willing to be treated.^[26] Statistical decision theory appears to be underutilized in the development of diagnostic tests and the evaluation of their benefit–risk trade-off.

TRAINING AND VALIDATION

Diagnostic devices (tests) may involve a model or algorithm that separates patients into classes with or without the condition or into different grades based on relevant parameters. The reason for developing classifiers is to allow reliable classification of patients. A classification model is usually developed on one or more training data sets. However, it may not be sufficient to show that the model/algorithm predicts outcome in the data used to develop it. If a classifier is evaluated on the same data set(s) on which it was trained, its performance (e.g., diagnostic accuracy) can be vastly overstated, leading to the so-called resubstitution bias. The bias can be especially acute if the number of variables considered for inclusion into the model approaches or exceeds the training set sample size.^[25,27]

Generally, building and validating a classifier involve three steps. The first step is building the classifier. Then, it is fine-tuned to fit the device's intended use. Finally, it is validated on a different, independent data using subjects/specimens from the intended use population to assure that the study results can be generalized.

The training data set is the data on which the test or classifier is developed. Internal validation, which is sometimes done on all or part of the training data set, is used to fine-tune the model. External validation on a separate independent data set is generally referred to as testing. During the development of the classifier, one needs to be blinded to the test data. The classifier needs to be finalized or fixed, before one can look at the testing data set.

Quite often, issues arise with validating the classifiers. Some of the issues are: 1) assuming internal validation is external validation when it is part of training; 2) looking at the test data before the classifier is fixed; 3) using data that have been previously used to support a different/ modified classifier; and 4) using parts of the same data set to build and validate the classifier.

For internal validation, it is fairly common to use different parts of the same data set. Investigators may use one

part to build the classification model and the other part to validate (test) the model. They may use all the data to develop and validate the model (e.g., bootstrap method or k-fold cross validation). Both of them, however, are internal validation, and cannot substitute for external validation, which is generally done on an independent data set to which people developing the model are blinded until the model is fixed. Cross-validation estimates of performance, even if properly carried out, may not generalize out of the training data set. A reason is that the classifier may have been fitted to patient characteristics, specimen (or imaging) characteristics, and measurement errors (e.g., batch effects) that were peculiar to the training data set.

External (out-of-sample) validation refers to evaluating a “locked-down” classifier on a cohort of subjects that is completely independent of the training set(s). External validation mimics how the classifier will be used in practice, namely, to classify subjects. A well-designed and conducted external validation study provides estimates of classifier performance that are generalizable to the intended use population.

Gary Collins^[28] described this in a review of articles on multivariable prediction from 2010. They conclude that a vast majority of studies describing some form of external validation were poorly reported and state that this is consistent with other reviews of prediction models. They proposed an initiative, called Transparent Reporting of a multivariate model for Individual Prognosis Or Diagnosis (TRIPOD). These are essential guidelines for reporting, rather than developing and validating the model, but planning for appropriate reporting should also lead to an appropriate study design.

The TRIPOD initiative started by developing an extensive list of items based on a review of the literature that was reduced after a web-based survey and revised during a 3-day meeting in June 2011 in Oxford, United Kingdom, with methodologists, health-care professionals, and journal editors. The list consists of a checklist of 22 items, deemed essential for transparent reporting of prediction model studies.^[29]

Guidelines for classifier development and validation are available regarding best practices^[30] and reporting.^[29]

CONCLUSION

This entry discusses the evaluation of diagnostic devices, covering clinical and measurement validation, risk and benefit analysis, training, and validation. These general considerations are applicable to a variety of medical tests. This entry does not discuss other types of diagnostic device evaluation such as the clinical validation of companion diagnostic devices. For study design and analysis of companion diagnostic devices and precision medicine, the chapters on “companion diagnostic” and “precision medicine,” respectively, are referred.

REFERENCES

1. Campbell, G.; Irony, T.Z.; Lao, C.S.; Li, H.; Pennello, G.; Vishnuvajjala, R.L.; Yue, L.Q. Medical devices. *Enc. Biostat.* **2005**, *5*, 3119–3137.
2. Food and Drug Administration. *Design Considerations for Pivotal Clinical Investigations for Medical Devices - Guidance for Industry, Clinical Investigators, Institutional Review Boards and Food and Drug Administration Staff*; United States Food and Drug Administration: Rockville, MD, 2013.
3. Food and Drug Administration. *Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests*; United States Food and Drug Administration: Rockville, MD, 2007.
4. Clinical and Laboratory Standards Institute. *Method Procedure Comparison and Bias Estimation Using Patient Samples; Approved Guideline—Third Edition. CLSI Document EP9-A3*; Clinical and Laboratory Standards Institute: Wayne, PA, 2013.
5. International Organization for Standardization. *Accuracy (Trueness and Precision) of Measurement Methods and Results—Part 1: General Principles and Definitions. ISO 5725-1*; International Organization for Standardization: Geneva, 1994.
6. International Organization for Standardization. *Accuracy (Trueness and Precision) of Measurement Methods and Results—Part 2: Basic Method for the Determination of Repeatability and Reproducibility of a Standard Measurement Method. ISO 5725-2*; International Organization for Standardization: Geneva, 1994.
7. Clinical and Laboratory Standards Institute. *Evaluation of Precision of Quantitative Measurement Procedures; Approved Guideline—Third Edition. CLSI Document EP5-A3*; Clinical and Laboratory Standards Institute: Wayne, PA, 2014.
8. Clinical and Laboratory Standards Institute. *Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures; Approved Guideline—Second Edition. CLSI Document EP17-A2*; Clinical and Laboratory Standards Institute: Wayne, PA, 2012.
9. Clinical and Laboratory Standards Institute. *Evaluation of the Linearity of Quantitative Measurement Procedures: A Statistical Approach; Approved Guideline. CLSI Document EP6-A*; Clinical and Laboratory Standards Institute: Wayne, PA, 2003.
10. Clinical and Laboratory Standards Institute. *Interference Testing in Clinical Chemistry; Approved Guideline—Second Edition. CLSI Document EP7-A2*; Clinical and Laboratory Standards Institute: Wayne, PA, 2005.
11. Clinical and Laboratory Standards Institute. *User Protocol for Evaluation of Qualitative Test Performance—Second Edition. CLSI Document EP12-A2*; Clinical and Laboratory Standards Institute: Wayne, PA, 2008.
12. Clinical and Laboratory Standards Institute. *Estimation of Total Analytical Error for Clinical Laboratory Methods; Approved Guideline. CLSI Document EP21-A*; Clinical and Laboratory Standards Institute: Wayne, PA, 2003.
13. Clinical and Laboratory Standards Institute. *Defining, Establishing, Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline—Third Edition. CLSI Document EP28-A3c*; Clinical and Laboratory Standards Institute: Wayne, PA, 2010.
14. Clinical and Laboratory Standards Institute. *Assessment of the Diagnostic Accuracy of Laboratory Tests using Receiver Operating Characteristic Curves; Approved Guideline—Second Edition. CLSI Document EP24-A2*; Clinical and Laboratory Standards Institute: Wayne, PA, 2011.
15. Food and Drug Administration. *Standards (Medical Devices)*. Available at <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/Standards/default.htm> accessed February 2011; United States Food and Drug Administration: Silver Spring, MD, 2011.
16. Metz, C.E. Basic principles of ROC analysis. *Semin. Nucl. Med.* **1978**, *8* (4), 283–298.
17. Zweig, M.H.; Campbell, G. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clin. Chem.* **1993**, *39*, (4), 561–577.
18. Drukker, C.A.; Bueno-de-Mesquita, J.M.; Retèl, V.P.; van Harten, W.H.; van Tinteren, H.; Wesseling, J.; Roumen, R.M.; Knauer, M.; van 't Veer, L.J.; Sonke, G.S.; Rutgers, E.J.; van de Vijver, M.J.; Linn, S.C. A prospective evaluation of a breast cancer prognosis signature in the observational RASTER study. *Int. J. Cancer.* **2013**, *133*, 929–936.
19. Bland, J.M.; Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* **1986**, *1* (8476), 307–310.
20. Beam, C.; Layde, P.M.; Sullivan, D.C. Variability in the interpretation of screening mammograms by US radiologists. *Arch. Intern. Med.* **1996**, *156*, 209–213.
21. Wagner, R.F.; Metz, C.E.; Campbell, G. Assessment of medical imaging systems and computer aids: A tutorial review. *Acad. Radiol.* **2007**, *14*, 723–748.
22. Tan, C.S.; Gilligan, D.; Pacey, S. Treatment approaches for EGFR-inhibitor-resistant patients with non-small-cell lung cancer. *Lancet. Oncol.* **2015**, *16* (9), e447–e459.
23. Campbell, G.; Pennello, G.; Yue, L. Missing data in the regulation of medical devices. *J. Biopharm. Stat.* **2011**, *21*, 180–195.
24. McIntosh, M.W.; Pepe, M.S. Combining several screening tests: Optimality of the risk score. *Biometrics.* **2002**, *58*, 657–664.
25. Gail, M.H.; Pfeiffer, R.M. On criteria for evaluating models of absolute risk. *Biostatistics.* **2005**, *6*, 227–239.
26. Baker, S.G. Putting risk into perspective: Relative utility curves. *J. Natl. Cancer I* **2009**, *101*, 1538–1542.
27. Vickers, A.J.; Elkin, E.B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making* **2006**, *26*, 565–574.
28. Gary, S.C.; de Groot, J.A.; Dutton, S.; Omar, O.; Shanyinde, M.; Tajar, A.; Voysey, M.; Warton, R.; Yu, L.-M.; Moons, K.G.; Altman, D. External Validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Med. Res. Method.* **2014**, *14*, 40.
29. Gary, S.C.; Johannes, B.R.; Douglas, G.A.; Karel, G.M.M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Ann. Intern. Med.* **2015**, *162* (1), 55–63.
30. Institute of Medicine. *Evolution of Translational Omics: Lessons Learned and the Path Forward*; The National Academies Press: Washington, DC, 2012.

Diagnostic Imaging

Suming W. Chang

Berlex Laboratories, Inc., Montville, New Jersey, U.S.A.

Annpey Pong

Merck Research Laboratories, Summit, New Jersey, U.S.A.

Abstract

Diagnostic imaging refers to a variety of new innovative techniques to identify disease or injury via images representing the internal anatomic structure of the human body. This entry

INTRODUCTION

In drug research and development, diagnostic imaging refers to a variety of new innovative techniques to identify disease or injury via images representing the internal anatomic structure of the human body. Physicians use the resulting images to diagnose various medical illness or injury so that patient treatment and therapy can be specifically planned and implemented. The history of diagnostic imaging refers to the radiology, which began as a medical subspecialty in the first decade of the 1900s after the discovery of X-rays by Professor Roentgen. However, the radiology has grown extensively since World War II, where X-ray imaging was used broadly. In the past 25 yr, the digital computer and the new imaging modalities such as ultrasound and magnetic resonance imaging (MRI) have combined to create an explosion of the diagnostic imaging techniques increasingly. In today's health care system, the diagnostic imaging is a key tool in the early diagnosis and prevention of disease.

OVERVIEW

In general, the medical imaging drug products use medical imaging methods to provide information on anatomy, physiology, and pathology. The commonly used methods are radiology, computed tomography (CT), ultrasonography, MRI, and radionuclide imaging. The term *images* can be used as films, likenesses, or other renderings of the body, body parts, organ systems, body functions, or tissues. For example, an image obtained with an MRI contrast agent may refer to a set of images acquired with different pulse sequences and interpulse delay times. An illustration of the medical imaging, as shown in [Fig. 1](#), uses MRI to provide remarkably clear and detailed pictures of the living internal body. Usually, images are created from the computerized acquisition of digital signals.

There are two types of medical imaging drugs: contrast drug products and diagnostic radiopharmaceuticals. The

intentions of a medical imaging drug are twofold: (1) delineate nonanatomic structures such as tumors or abscesses and (2) detect disease or pathology within an anatomic structure. The differences of these two types of medical imaging drugs are described as follows.

Contrast Drug Products

Contrast drug products are used to increase the relative difference of the signal intensities and to provide the additional information in combination with an imaging device beyond the device alone. That is, imaging with the contrast drug product should provide more information than that without the contrast drug product. The most commonly used contrast drug products that are in combination with medical imaging devices are listed in [Table 1](#).

Diagnostic Radiopharmaceuticals

The diagnostic radiopharmaceuticals are radioactive drugs that contain a radioactive nuclide that may be linked to a ligand and a carrier. These products are used in planar imaging, single-photon emission computed tomography, positron emission tomography, or with other radiation detection probes.

Diagnostic trials are clinical trials with diagnostic outcomes needed to establish for therapeutic or diagnostic products under study. Because of the different considerations in designs, the techniques in evaluating the performance of diagnostic medical products are much different from those in evaluating therapeutic pharmaceuticals or from nondiagnostic devices.

REGULATORY REQUIREMENTS

Medical imaging drugs are generally governed by the same regulations as other drug and biological products.^[1]



Fig. 1 Frontal and side view of human body by MRI
(From Imaginis Corporation, sponsored by SIEMENS.)

Table 1 The Combination of Contrast Drug Products with Medical Imaging Devices

Modality	Contrast drug products
X-ray and CT	Iodine agents (photon scattering)
MRI	Gadolinium, dysprosium, and helium
Ultrasound	Liposomes and microbubbles
Suspensions nuclear	Tc-99m, TI-201, indium, and samarium

However, medical imaging drugs have their special characteristics in drug development. In 1998, Food and Drug Administration (FDA) proposed a rule to amend the drug and biologics regulations and add provisions for the evaluation and approval of *in vivo* radiopharmaceuticals used in the diagnosis or monitoring of disease (63 FR 28301, 1998).

The proposed rule was a response to the requirements of the FDA Modernization Act of 1997. The most current guidance for developing medical imaging drugs and biological products in planning and coordinating their clinical investigations and the FDA submission are shown in the draft guidance in June 2000. This draft guidance elaborates on the concepts contained in the proposed rule on radiopharmaceutical diagnostic products entitled *Developing Medical Imaging Drugs and Biological Products*.^[1]

STUDY DESIGN CONSIDERATIONS

Good diagnostic trials have diagnostic endpoints consistent with a well-planned clinical setting. Ideally, the trial is sufficiently controlled to show that the added value of the diagnostic agent and endpoints are objectively defined. That is, good diagnostic trials should be able to translate the effect of a new agent in terms of clinical utility. In efficacy evaluation of imaging studies, it is very often that the study design is based on the results from two individual components. The first one is the clinical component, which is often designed as an open label, nonrandomized, and with noncontrast image as the “control.” In addition to the clinical component, imaging trials may also have the second component that is an off-site “blinded-read” component.

The designs of both components are shown in “Clinical Component” and “Blinded-Read Component.” Potential bias of each component is discussed in “Bias Considerations.” Moreover, the commonly used methods to measure the reader agreement for blinded-read studies are given in “Assessing Reader Agreement for a Blinded-Read Trial.”

Clinical Component

There are many types of designs in clinical component of diagnostic trials. The most commonly used designs are the parallel group design and the crossover design. The choice of clinical designs for a diagnostic trial is often determined by patient population and disease type.

In the parallel group design, the subjects are randomized to one of two or more arms. A different treatment has been allocated to each arm. These treatments include the investigational product at one or more doses and one or more control treatments, such as placebo and/or active comparator. The assumptions underlying this design are less complex than most of other designs. However, similar to other designs, there may be additional features of the study design that complicate the analysis and interpretation (e.g., covariates, repeated measurements over time, interactions between design factors, dropouts, and withdrawals). Parallel group trials generally require large sample sizes to show noticeable effect and may mask important differences in test capabilities. Parallel designs are not practical for hard-to-enroll populations either.

In the crossover design, each subject is randomized to a sequence of two or more treatments, and hence acts as its own control for treatment comparison. This simple maneuver is attractive primarily because it reduces the number of subjects and usually the number of assessments needed to achieve a specific power, sometimes to a marked extent. In the simplest 2×2 crossover design, each subject receives either one of the two treatments in a randomized order in two successive treatment periods, often separated by a washout period. The most common extension of this entails comparing n treatments in n periods, each subject receiving all n treatments. Crossover designs enable better measuring effect of a new test. However, such designs may not be practical either and often involve ethical issues. When the crossover design is used, it is critical to avoid carryover, i.e., the residual influence of treatment in subsequent treatment periods. Furthermore, the crossover design should generally be restricted to situations when loss of subjects from the trial is expected to be small.

Blinded-Read Component

In addition to a clinical component, imaging trials may also have an off-site “read” component. The off-site read has its own design and is often called “blinded-read.” The goal of the blinded-read is to control for investigator bias. A good image viewing strategy guards against bias and avoids introducing effects confounding the clinical component. Therefore the readers who read the images are those who are *independent* and *blinded*. The *independent* are those who have not participated in studies and who are not affiliated with the sponsor or with institutions in which the studies were conducted. The meaning of *blinding* differs from the common way the term is used in therapeutic clinical trials. *Blinded* are those who are unaware of (1) treatment identity used to obtain a given image and (2) patient-specific clinical information or study protocol. However, the specific information provided may depend on the modality under study, the type of agent, and the proposed indication. Three commonly used methods for blinded-read of imaging evaluations are described as follows:

- i. Fully blinded image evaluation—readers have no knowledge of treatment identity, results of the evaluation with the true standard, final diagnosis, or any patient-specific information (e.g., history, laboratory results, and results of other imaging study).
- ii. Imaging evaluation blinded to outcome—readers have no knowledge of the results of the evaluation with the true standard, treatment identity, the final diagnosis, or patient outcome. However, the readers may have knowledge of particular elements of patient-specific information, such as the results of other imaging studies.
- iii. Sequential unblinding—readers evaluate images progressively on each read. The typical evaluation of a

three-step process is: Step 1, perform a fully blinded image evaluation, and record and lock the data. No data can be changed when additional information is available; Step 2, perform an image evaluation which is blinded to outcome, and recorded and locked in the data set; Step 3, compare the result of the above two blinded evaluations to the true standard (or the final diagnosis of the patient outcome). If sequential unblinding is considered, the primary one for determining efficacy should be specified before the study begins.

Bias Considerations

In clinical trial, almost all endpoints in a study may be subject to bias because of different sources. Bias means that the results observed reflect other factors in addition to the effect of the treatment. Similarly, there are many sources of systematic and random errors that are unique to the imaging studies. For example, by using a new ultrasound contrast agent, the bias may come from the poor training and inadequate skill of operators; unclear specifications and settings for device; uninformative response in patient’s characteristics; misinterpretation of the results from the radiologist/clinician; and the information from the gold standard is not the truth. In summary, the commonly encountered bias in study design of diagnostic imaging is listed below for open-label trial and blinded-read trial, respectively.

Bias for an Open-Label Trial

The major disadvantage of the open label is the influence of the response to treatment because of the knowledge of treatment identity. For example, patients who know that they have been assigned to receive the new ultrasound contrast agent may have favorable expectations or increased anxiety, and patients who are not on the new or experimental treatment may become dissatisfied and may drop out of the trial in large numbers. Therefore the open-label trial is particularly prone to bias when the outcome measures are subjective. Other primary biases for an open-label trial in diagnostic imaging studies can be:

- Knowledge of patient clinical information including results from other diagnostic tests (truth standard result and/or comparator result) may affect the way the assessor assesses the outcomes.
- Knowledge of trial protocol and study objectives (entry criteria and effects on pretest probabilities) may affect the way the investigator deals with the patient.
- Image collection (device operation/imaging window) may be influenced in data compilation if the knowledge of treatment identity is not covered.

Bias for a Blinded-Read Trial

It is obvious that the blinded image evaluation can minimize the risk for outcome assessment. Particularly, blinding is useful in reducing bias when determining the performance characteristics for a new treatment compared with the comparator product. However, there are still many situations that biases for blinded-read trial are unavoidable in diagnostic imaging, which are listed below:

- Test drug or imaging modality may be obvious to an examiner. For example, the instructions for image evaluations may differ according to treatment. Therefore blinding to treatment identity may be infeasible.
- The reader can recall image placement and reading order. For example, in an independent evaluation of the unenhanced and enhanced images, the reader may remember the images that were obtained at different times used to a specific tumor. Therefore the effect of randomizing the images during the study is no longer effective for those readers with long-term memory. In a combined image evaluation, the reader often evaluates unenhanced and enhanced images concurrently. The likelihood of bias may increase because of the confounding effects from multiple images (e.g., by systematic overreading or underreading particular findings on images).
- Missing information was concluded for nondiagnostic or uninterpretable images. Therefore the imaging data may be lost if there was no additional information available. However, this would not be more reflective of the clinical setting for the use of contrast agent.

Assessing Reader Agreement for a Blinded-Read Trial

As indicated in the draft guidance,^[1] at least two independent, blinded readers (and preferably three or more) are recommended for each study that is intended to demonstrate efficacy. This allows an evaluation of the reproducibility of the readings (interreader variability). The purpose is to provide a better basis for the subsequent generalization of the findings in the studies. Ideally, each reader should review all the images to give an unbiased evaluation of images. However, this is impractical for large studies.

To facilitate the consistency among readers, the common approach is to assess the reader agreement quantitatively. Suppose that n_{ij} and p_{ij} are the number of patients and the proportion of patients in the i th category by the first reader and j th category by the second reader from a total of k categories, respectively, where $i, j = 1, 2, \dots, k$. The most commonly used statistical method to measure the consistency among readers is the kappa statistic,

which is based on the count data to measure the reader agreement, and was introduced by Cohen.^[2] It is developed as a measurement of nominal scale agreement among readers, particularly applied to the problems of the reliability or reproducibility of the diagnostic tests. The underlying assumption of the kappa statistic, κ , is that the two response variables are independent ratings from N subjects, where

$$N = \sum_{i=1}^k \sum_{j=1}^k n_{ij}$$

$$\text{Kappa Statistics : } \kappa = \frac{p_o - p_e}{1 - p_e}$$

where

$$p_o = \sum_{i=1}^k p_{ii} = \sum_{i=1}^k \frac{n_{ii}}{N}$$

= proportion of observed agreement

$$p_e = \sum_{i=1}^k \sum_{j=1}^k p_{ig} p_{gj}$$

= proportion of expected agreement

The meaning of strength of reader agreement was defined by Landis and Koch^[3] based on the magnitude of kappa statistics. It is shown as below:

Kappa statistic range	Strength of reader agreement
0.81–1.00	Almost perfect
0.61–0.80	Substantial
0.41–0.60	Moderate
0.21–0.40	Fair
0.00–0.20	Slight
<0.00	Poor

The variance of κ is given by

$$\text{Var}(\kappa) = \frac{A + B + C}{N(1 - p_e)^2}$$

where

$$A = \sum_{i=1}^k p_{ii} [1 - (p_{ig} + p_{gj})(1 - \kappa)]^2$$

$$B = (1 - \kappa)^2 \sum_{i,j} p_{ij} (p_{ig} + p_{gj})^2$$

$$C = [\kappa - p_e(1 - \kappa)]^2$$

The kappa value is +1 for complete agreement and zero if the agreement is the same as by chance alone; on the other hand, the value becomes -1 for complete disagreement.

Furthermore, Brenner and Kliebsh^[4] proposed a weighted kappa coefficient on the number of categories as shown below.

$$\kappa_w = \frac{p_{0w} - p_{ew}}{1 - p_{ew}}$$

where

$$p_{0w} = \sum_{i=1}^k w_{ij} p_{ii}$$

and

$$p_{ew} = \sum_{i=1}^k \sum_{j=1}^k w_{ij} p_{ig} p_{gj}$$

The variance of the weighted kappa yields

$$\text{Var}(\kappa_w) = \frac{1}{n(1 - p_e)^2} \left(\sum_{i=1}^k \sum_{j=1}^k p_{ij} [w_{ij} - (w_i + w_j)(1 - \kappa_w)]^2 - [\kappa_w - p_{ew}(1 - \kappa_w)]^2 \right)$$

The common kappa weights for categorical data are:

1. Cicchetti–Alison weight, which is inversely proportional to the distance from the diagonal, such that:

$$w_{ij} = 1 - \frac{|C_i - C_j|}{C_k - C_1}$$

where C_i and C_j are the category of response for readers 1 and 2, respectively, and C_k and C_1 are the maximal and minimum category of response, respectively, and

2. Fleiss–Cohen weight, which is inversely proportional to the squared distance from the diagonal, such that

$$w_{ij} = 1 - \frac{(C_i - C_j)^2}{(C_k - C_1)^2}$$

In general, kappa statistics are suitable for the agreement in qualitative measurements. To measure the interrater agreement in quantitative data, the intraclass correlation is useful. The intraclass correlation coefficients (ICCs) are designs used to assess consistency or conformity between two or more quantitative measurements. They can be used for investigating the properties of reliability, validity, and reproducibility. The commonly used method to estimate the ICC is formulated by the analysis of variance (ANOVA) model. For example, assume that each of the N patients is rated by each of the same two raters and the

data for agreement is normally distributed. Then, it yields a two-way ANOVA model illustrated below:

Source of variation	Degrees of freedom	Mean squares
Between patients	$N - 1$	MSB
Within patients	N	MSW
Between raters	1	MSR
Residual	$N - 1$	MSE
Total	$2N - 1$	

The ICC is given by:

$$\text{ICC} = \frac{\text{MSB} - \text{MSE}}{\text{MSB} + \text{MSE}}$$

For number of rater $k > 2$, the ICC yields

$$\text{ICC} = \frac{\text{MSB} - \text{MSE}}{\text{MSB} + (k - 1)\text{MSE}}$$

Additional information on the intraclass correlation can be found in Snedecor and Cochran.^[5]

DIAGNOSTIC PERFORMANCE MEASURES

To determine how well a diagnostic imaging agent can distinguish between diseased subjects and nondiseased subjects, the outcome may often be classified into one of the four groups depending on (1) whether a disease is present and (2) the results of the diagnostic test of interest (positive or negative). The outcomes are shown in the following 2×2 contingency table:

Test result	Disease		Total
	Present	Absent	
Positive	a	b	$a + b = m_1$
Negative	c	d	$c + d = m_2$
Total	$n_1 = a + c$	$n_2 = b + d$	$N = a + b + c + d$

In the above table, the number in each cell is denoted as the frequency count, which are called the true positive (denoted by a), false positive (denoted by b), false negative (denoted by c), and true negative (denoted by d). The further notations are defined as below:

Total with positive test: $m_1 = a + b$.
 Total with negative test: $m_2 = c + d$.
 Total with disease: $n_1 = a + c$.
 Total without disease: $n_2 = b + d$.
 Total in study: $N = a + b + c + d$.

To demonstrate the efficacy of the agent, the commonly used measurements in diagnostic accuracy are illustrated in the following table.

<i>Diagnostic accuracy</i>	<i>Formula</i>	<i>Interpretation</i>
Accuracy	$(a + d)/N$	The proportion of cases that considers both positive and negative test results for which the test results are correct.
Likelihood ratios	$LR(+)= (a/n_1)/(b/n_2)$ $LR(-)= (c/n_1)/(d/n_2)$	• $LR(+)$ is the probability of a positive test result in subjects with the disease compared with the probability of a positive result in subjects without the disease. • $LR(-)$ is the probability of a negative test result in subjects with the disease compared with the probability of a negative result in subjects without the disease.
Negative predictive value	d/m_2	The probability that a subject does not have the disease given that the test result is negative.
Positive predictive value	d/m_1	The probability that a subject has disease given that the test result is positive.
Odds	$\frac{(a/m_1)/(b/m_1)}{(c/m_2)/(d/m_2)} = \frac{(a/n_1)/(b/n_1)}{(c/n_2)/(d/n_2)} = \frac{ad}{bc}$	The probability that an event occurs compared with the probability that the event will not occur.
Posttest odds of disease	For positive test result, posttest odds of disease = a/b For negative test result, posttest odds of disease = c/d	The posttest odds of disease is the odds of disease in a subject after the diagnostic test results are known. It should be noted that the posttest odds of disease is equal to the pretest odds of disease multiplied by the likelihood ratio.
Posttest probability of disease	For positive test result, posttest probability of disease = a/m_1 For negative test result, posttest probability of disease = c/m_2	The posttest probability of disease is the probability of disease in a subject after the diagnostic test results are known.
Sensitivity	a/n_1	The probability that a test result is positive given that the subject has a disease.
Specificity	d/n_2	The probability that a test result is negative given that the subject does not have a disease.

It is to be noted that the accuracy of the information determined from a diagnostic test requires a judicious choice of a comparator agent or procedure. In other words, the presence of disease is often determined by a truth standard or gold standard. A true standard or gold standard is an independent method of measuring the same variable being measured by the investigational drug that is known or believed to give the truth state of a patient or true value of a measurement. The true standards are used to demonstrate that the results obtained by the medical imaging drug are valid and reliable. The potential gold standard for a diagnostic clinical trial includes:

- Histopathology.
- Therapeutic response.
- Clinical outcome.
- Another diagnostic test.
- *In vitro* tests.
- Animal autoradiography.
- Autopsy.

For example, for a MRI contrast agent intended to visualize the number of lesions in liver or determine whether a mass is malignant, the gold standard might include results from the pathology or long-term clinical outcomes.

The common methods used to test for the differences in diagnosis include the McNemar test, Stuart Maxwell test, and the confidence intervals for sensitivity and specificity. In addition, the receiver operating characteristic analysis is another alternative that can be used to evaluate diagnostic accuracy.^[6]

CONCLUSION

Although kappa statistics are appropriate for testing whether the agreement exceeds the chance levels for binary and nominal ratings, kappa is an omnibus index of agreement. It has no distinctions among various types and sources of agreement. Also, kappas are not comparable across studies, procedures, or populations.^[7,8] The major limitation of the kappa statistic is to require that an observation be an assignment of a subject to only one response category; also, it requires equal number of observations per subject. Kraemer^[9] proposed an extension of the kappa coefficient, which is applied for multiple responses per observation and which allows multiple and not necessarily an equal number of observations per subject.

Another controversial issue for kappa statistics is to use kappa to quantify the levels of agreement. The kappa

was calculated based on the proportion of chance agreement, which is the proportion of times that raters would agree by chance alone. However, this is under the assumption of the independence of raters. The appropriateness of the kappa statistics is questionable when raters are not independent in many applications. In addition, kappa may be low even when there are high levels of agreement between raters. Whether a given kappa value implies a good or bad diagnostic method depends on what model one assumes about the decision making of the raters.^[10]

The sensitivity and specificity are commonly performed for the diagnostic accuracy of the binary data. However, if the correlated samples of the binary data are presented, the true variance of sensitivity and specificity will be underestimated if a conventional standard test is used. The underlying assumption for sensitivity and specificity was for independent observations. To prevent this from happening, the within-subject correlation should be taken into account to estimate the sensitivity and specificity. The examples for dependent observations are: (1) the detected multiple lesions within the same patient and (2) the diagnostic tests to a tooth surface within a patient's mouth. Moreover, these correlated samples of binary data appear in many fields of applications.

Several alternative approaches were proposed for dependent observations in the literature. Sternberg and Hadgu^[11] used a generalized estimating equation (GEE) to estimate the sensitivity and specificity for corrected observations. Lee and Dubin^[12] and Lee^[13] proposed the weighted estimators to calculate the accuracy, sensitivity, specificity, and positive predictive probability for correlated observations. In addition, Ahn^[14] introduced several statistical methods for correlated diagnostic tests based on the assumption that any two observations in the same patient have the same correlation value. Furthermore, the performance of ratio estimator,^[15,16] the intracluster correlation estimator,^[17] the weighted estimator, and GEE estimator were compared in terms of their biases, observed variances, MSE, and relative efficiencies of their variances; also, 95% coverage proportions were compared by Ahn^[14] based on the simulation study.

REFERENCES

1. FDA. *Draft Guidance for Industry: Developing Medical Imaging Drugs and Biological Products*; 2000.
2. Cohen, J. A coefficient of agreement of nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46.
3. Landis, R.J.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *50*, 183–193.
4. Brenner, H.; Kliebsh, U. Dependence of weighted kappa co-efficients on the number of categories. *Epidemiology* **1996**, *7* (2), 199–202.
5. Snedecor, G.W.; Cochran, W.G. *Statistical Methods*, 6th Ed.; The Iowa State University Press: Ames, Iowa, 1967.
6. Chow, S.C.; Shao, J. *Statistics in Drug Research—Methodology and Recent Development*; Marcel Dekker, Inc.: New York, 2002.
7. Thompson, W.D.; Walter, S.D. A reappraisal of kappa coefficient. *J. Clin. Epidemiol.* **1988**, *41*, 969–970.
8. Feinstein, A.R.; Cicchetti, D.V. High agreement but low kappa: I. The problems of two paradoxes. *J. Clin. Epidemiol.* **1990**, *43* (6), 543–549.
9. Kraemer, H.C. Extension of the kappa coefficient. *Biometrics* **1980**, *36* (2), 207–216.
10. Uebersax, J.S. Diversity of decision-making models and the measurement of interrater agreement. *Psychol. Bull.* **1987**, *101*, 140–146.
11. Sternberg, M.R.; Hadgu, A. A GEE approach to estimating sensitivity and specificity and coverage properties of the confidence intervals. *Stat. Med.* **2001**, *20*, 1529–1539.
12. Lee, E.; Dubin, Estimation and sample size considerations for clustered binary responses. *Stat. Med.* **1994**, *13*, 1241–1252.
13. Lee, E. Two sample comparison for large groups of correlated binary responses. *Stat. Med.* **1996**, *15*, 1187–1197.
14. Ahn, C. An evaluation of methods for estimation of sensitivity and specificity of site-specific diagnostic tests. *Biom. J.* **1997**, *39* (7), 793–807.
15. Cochran, W.G. *Sampling Techniques*, 3rd Ed.; Wiley: New York, 1997.
16. Rao, J.K.; Scott, A. A simple method for the analysis of clustered binary data. *Biometrics* **1992**, *48*, 577–585.
17. Donner, A.; Klar, N. Confidence interval construction for effect measures arising from cluster randomization trials. *J. Clin. Epidemiol.* **1993**, *46*, 123–131.

Diagnostic Procedure: Sensitivity and Specificity

Jiayin Zheng and Shein-Chung Chow

Department of Biostatistics and Bioinformatics, Duke University School of Medicine,
Durham, North Carolina, U.S.A.

Abstract

In biopharmaceutical studies, diagnostic tests are routinely used to screen for, diagnose, grade, and monitor the progression of diseases. To validate and evaluate the performance of a test, it is often compared with a reference standard that reflects the “truth,” and then different measures of accuracy are calculated to see how good is the test at diagnosis of the disease. In this article, we provide a brief introduction of two well-known and widely used measures, sensitivity and specificity.

Keywords: Accuracy; Likelihood ratio; Reference standard; True positive; True negative.

INTRODUCTION

The use of diagnostic tests is common in a physician’s daily practice, aiming to confirm or reject the presence of a disease or health event. Diagnostic information is obtained from a variety of sources, including imaging and biochemical techniques, pathological and psychological assessments, observing signs and symptoms, medical history, and clinical examinations.^[1,2] Ideally, such tests are expected to correctly identify subjects with and without the disease with 100% accuracy. A perfect test will show a negative result for all disease-free subjects and a positive result for all diseased subjects. However, tests are prone to errors. To evaluate a test, multiple aspects are taken into consideration, including reproducibility and accuracy.^[3–5] Reproducibility refers to the ability of producing the same result if the same test is done again.^[6] Accuracy refers to the degree of agreement between the results from the diagnostic test under study and the “truth” (or those from a reference standard).^[7] In this article, we focus on the accuracy and introduce two important measures: sensitivity and specificity.

The definitions of sensitivity and specificity consist of the following four fundamental terms. The terms positive and negative refer to the presence and absence of the condition of interest, respectively.

1. True positive: the test is positive for the diseased patient.
2. False positive: the test is positive for the disease-free patient.
3. True negative: the test is negative for the disease-free patient.
4. False negative: the test is negative for the diseased patient.

Table 1 shows the results of a diagnostic test, with rows summarizing the test results and columns summarizing the “truth.” The number of diseased subjects being positive and negative are denoted by a and c . The number of disease-free subjects being positive and negative are denoted by b and d .

SENSITIVITY AND SPECIFICITY

The sensitivity quantifies the ability of the test to correctly identify the diseased subjects and is defined as the proportion of true positives among the diseased subjects.

$$\text{Sensitivity} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} = \frac{a}{a + c}$$

The specificity quantifies the ability of the test to correctly identify the disease-free subjects and is defined as the proportion of true negatives among the disease-free subjects.

$$\text{Specificity} = \frac{\text{True negatives}}{\text{True negatives} + \text{False positives}} = \frac{d}{b + d}$$

The confidence intervals of sensitivity and specificity can be calculated using standard methods for proportions.^[8]

A test with 80% sensitivity correctly detects 80% of diseased subjects, with the remaining 20% undetected. Likewise, a test with 80% specificity correctly reports 80% of disease-free subjects as negative, with the remaining 20% incorrectly identified as positive. In general, the higher the sensitivity, the lower the specificity, and vice versa.

When a test aims at identifying a serious but treatable disease, a high sensitivity is desirable.^[9] However, a high

Table 1 Results of a diagnostic test

Results of the diagnostic test	The “truth” or results of the reference test	
	Disease present	Disease absent
Positive	True positive (a)	False positive (b)
Negative	False negative (c)	True negative (d)

sensitivity alone does not make a good test, as the test also needs to be negative for all disease-free subjects. In other words, high sensitivity is required for a test to be useful at ruling out a disease, and high specificity is required for it to be useful at confirming a disease.^[1] In general, whether one prefers a higher sensitivity but a lower specificity or vice versa is not straightforward. A diagnostic test with a lower sensitivity and specificity could be preferred when the test with better accuracy has a high risk of complications such as some invasive methods with potential complications. In this situation, one may attempt to balance the desirable and undesirable consequences of performing different diagnostic tests.^[10]

Unlike some other measures of accuracy, the prevalence of the disease does not affect the sensitivity and specificity of a diagnostic test.^[11] It is because they are calculated from subgroups of subjects with and without the disease, respectively.

Two other terms, positive likelihood ratio and negative likelihood ratio defined as follows, describe the discriminatory properties of positive and negative results, respectively.

$$\text{Positives likelihood ratio} = \frac{\text{Sensitivity}}{1 - \text{Specificity}} = \frac{a/(a + c)}{b/(b + d)}$$

$$\text{Negative likelihood ratio} = \frac{1 - \text{Sensitivity}}{\text{Specificity}} = \frac{c/(a + c)}{d/(b + d)}$$

They indicate the number of times more likely particular test results are in diseased subjects than in disease-free subjects.^[1] Positive likelihood ratios above 10 and negative likelihood ratios below 0.1 have been noted as providing convincing diagnostic evidence, whereas those above 5 and below 0.2 have given strong diagnostic evidence.^[12] A likelihood ratio close to 1 indicates that the test provides little additional information regarding the presence or absence of the disease.

SOME REMARKS ON PRACTICE

Although the sensitivity and specificity of diagnostic tests are not affected by the disease prevalence and thus can be compared across study populations with different prevalence rates, they can be influenced by differences in disease characteristics (such as disease severity) and characteristics of subjects such as age and gender.^[4] Many

diagnostic tests have been introduced with great enthusiasm because of their high sensitivity and specificity, but have nevertheless been rejected at a later stage. One of the reasons for this rejection was that measures of accuracy of a diagnostic test have often been presented as inherent and fixed properties, whereas in reality, they can be affected by different sources of variation and bias.^[13,14]

Thus, it is critical to evaluate the group of subjects to whom the diagnostic test has been applied when the sensitivity and specificity of two diagnostics tests are compared. Systematic reviews of evaluations of tests are required to produce the estimates of test performance and impact based on all available evidence, to evaluate the quality of published studies, and to account for variation and bias in findings between studies.^[15,16] The important aspects of study quality include the selection of a clinically relevant cohort, the consistent use of a single good reference standard, and the blinding of results of experimental and reference tests.^[1,12] Unless the study is of high quality, its results cannot be included and compared with those from other studies. If the results of studies are reasonably homogeneous, then sensitivities, specificities, and likelihood ratios can be combined directly. Otherwise, appropriate statistical methods for pooling the results should be used to properly address multiple heterogeneity sources between studies and notably variation in diagnostic thresholds.

In recent years, several guidelines have been published on how to report on diagnostic tests, which are recommended for further reading.^[7,10,17]

REFERENCES

1. Deeks, J.J. Systematic reviews of evaluations of diagnostic and screening tests. *BMJ* **2001**, 323 (7305), 157–162.
2. Sackett, D.L. Clinical epidemiology: What, who, and whither. *J. Clin. Epidemiol.* **2002**, 55 (12), 1161–1166.
3. Altman, D.G.; Bland, J.M. Diagnostic tests. 1: Sensitivity and specificity. *BMJ* **1994**, 308 (6943), 1552.
4. Zhou, X.H.; McClish, D.K.; Obuchowski, N.A. *Statistical Methods in Diagnostic Medicine*, Vol. 569; John Wiley & Sons: Hoboken, NJ, 2009.
5. Loong, T.W. Understanding sensitivity and specificity with the right side of the brain. *BMJ* **2003**, 327 (7417), 716–719.
6. Van Stralen, K.J.; Stel, V.S.; Reitsma, J.B.; Dekker, F.W.; Zoccali, C.; Jager, K.J. Diagnostic methods I: Sensitivity, specificity, and other measures of accuracy. *Kidney International* **2009**, 75 (12), 1257–1263.
7. Bossuyt, P.M.; Reitsma, J.B.; Bruns, D.E.; Gatsonis, C.A.; Glasziou, P.P.; Irwig, L.M.; Moher, D.; Rennie, D.; de Vet, H.C.W.; Lijmer, J.G. The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Ann. Intern. Med.* **2003**, 138 (1), W1–W12.
8. Altman, D.; Machin, D.; Bryant, T.; Gardner, M. *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*; John Wiley & Sons: Hoboken, NJ, 2009.
9. Lalkhen, A.G.; McCluskey, A. Clinical tests: Sensitivity and specificity. *CEACCP* **2008**, 8 (6), 221–223.

10. Schnemann, H.J.; Oxman, A.D.; Brozek, J.; Glasziou, P.; Jaeschke, R.; Vist, G.E.; Williams, J.W Jr; Kunz, R.; Craig, J.; Montori, V.M.; Bossuyt, P.; Guyatt, G.H. Rating quality of evidence and strength of recommendations: GRADE: Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* **2008**, *336* (7653), 1106.
11. Wong, H.B.; Lim, G.H. Measures of diagnostic accuracy: Sensitivity, specificity, PPV and NPV. *Proc. Singapore Healthcare* **2011**, *20* (4), 316–318.
12. Jaeschke, R.; Guyatt, G.H.; Sackett, D.L.; Guyatt, G.; Bass, E.; Brill-Edwards, P.; Browman, G.; Cook, D.; Farkouh, M.; Gerstein, H.; Haynes, B.; Hayward, R.; Holbrook, A.; Juniper, E.; Lee, H.; Levine, M.; Moyer, V.; Nishikawa, J.; Oxman, A.; Patel, A.; Philbrick, J.; Richardson, W.S.; Sauve, S.; Sackett, D.; Sinclair, J.; Trout, K.; Tugwell, P.; Tunis, S.; Walter, S.; Wilson, M. Users' Guides to the Medical Literature: III. How to Use an Article about a Diagnostic Test B. What Are the Results and Will They Help Me in Caring for My Patients? *JAMA* **1994**, *271* (9), 703–707.
13. Whiting, P.; Rutjes, A.W.; Reitsma, J.B.; Glas, A.S.; Bossuyt, P.M.; Kleijnen, J. Sources of variation and bias in studies of diagnostic accuracy: A systematic review. *Ann. Intern. Med.* **2004**, *140* (3), 189–202.
14. Leeflang, M.M.; Rutjes, A.W.; Reitsma, J.B.; Hooft, L.; Bossuyt, P.M. Variation of a tests sensitivity and specificity with disease prevalence. *Can. Med. Assoc. J.* **2013**, *185* (11), 537–544.
15. Irwig, L.; Tosteson, A.N.; Gatsonis, C.; Lau, J.; Colditz, G.; Chalmers, T.C.; Mosteller, F. Guidelines for meta-analyses evaluating diagnostic tests. *Ann. Intern. Med.* **1994**, *120* (8), 667–676.
16. Irwig, L.; Macaskill, P.; Glasziou, P.; Fahey, M. Meta-analytic methods for diagnostic test accuracy. *J. Clin. Epidemiol.* **1995**, *48* (1), 119–130.
17. Bossuyt, P.M.; Reitsma, J.B.; Bruns, D.E.; Gatsonis, C.A.; Glasziou, P.P.; Irwig, L.; Lijmer, J.G.; Moher, D.; Rennie, D.; de Vet, H.C.W.; Kressel, H.Y.; Rifai, N.; Golub, R.M.; Altman, D.G.; Hooft, L.; Korevaar, D.A.; Cohen, J.F.; Kressel, H.Y. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *Radiology* **2015**, *277* (3), 826–832.

Disease Modification Effects: Design and Analysis

Djokouri A. Kouassi and Annpey Pong

Merck Research Laboratories, Summit, New Jersey, U.S.A.

Abstract

This entry provides an overview of candidate study designs by reviewing examples of study designs that could prove useful in assessing disease-modifying effects of new drug compounds as well as discussing methods of data analysis.

INTRODUCTION

In drug development, when the cause or etiology of a disease is known, a suitable treatment can be developed to alter its underlying pathogenic process and thereby abate its symptoms. However, for some chronic illnesses such as rheumatoid arthritis, depression, Parkinson's, and Alzheimer's diseases, the etiology is not known and there are as of yet no valid markers for disease progression. Consequently, no therapies that will specifically target the root causes of these maladies have been instituted thus far. In fact, the existing treatments for these disorders have been approved solely on the basis of their symptomatic effects. In other words, these medicines have been proven to relieve the symptoms rather than to disrupt the pathophysiological processes of these diseases. Of course symptomatic treatments or medicines that address only the symptoms are quite useful in providing patients with some degree of comfort. But they also mask the degenerative process of the illness which goes on unchecked for years and ultimately leads to death.

In recent years, the aim of many drug developers has been to provide these growing patient populations with drug products capable of affecting the pathogenic process of these disorders. If developed, such disease-modifying therapies can reasonably be expected to change the natural course of these illnesses, curb their progression, or outright cure them. Unfortunately, the development of such drugs has been thwarted by the lack of knowledge for the etiology of these diseases coupled with the absence of reliable markers for disease progression. Metaphorically, the challenge facing drug makers is nothing less than the daunting task of finding a needle in a haystack. How can a disease modification claim be made in the treatment of an illness about which very little to nothing is known? This is precisely the question that the scientific and regulatory communities have been scrambling to find an answer for. Some scientists have been exploring plausible hypotheses for the etiologic factors involved in the genesis of these diseases.

Others are searching for biological markers that could serve as valuable hints for the progression of the structural decay caused by the diseases, and yet other researchers are investigating clinical trial design strategies that could have the potential to demonstrate disease modification effects. In this chapter, we provide an overview of candidate study designs. Note that none of the design techniques that are discussed here have been used in clinical trials that have led to a successful demonstration of disease modification effects. The section "Design Considerations" will present three study designs that could be useful in assessing disease-modifying effects, if any, of new drug compounds. The section "Methods for Data Analysis" will be devoted to the discussion of likely methods of data analysis, and our concluding remarks will be presented in the "Conclusion."

DESIGN CONSIDERATIONS

Randomized Start/Withdrawal Designs

Randomized start/withdrawal designs were proposed in 1996 and 1997 by Paul Leber,^[1,2] a medical reviewer from the U.S. Food and Drug Administration (FDA), for demonstrating disease-modifying effects of new therapeutic compounds in the treatment of Alzheimer's disease. They are crossover type studies conducted over two periods. In the first period, patients are randomly assigned to either an experimental drug or placebo and treated in a double-blinded fashion. After the first period, all patients receive either the experimental drug (randomized start designs) or placebo (randomized withdrawal designs). The treatment effects are typically assessed at the end of each period. The key rationale behind these designs is that if the drug compound being tested has disease-modifying effects, any treatment difference observed at the end of the first treatment period will be sustained throughout the second. However, if it has only symptomatic effects, then the observed difference will fade away (see [Figs. 1 and 2](#)).

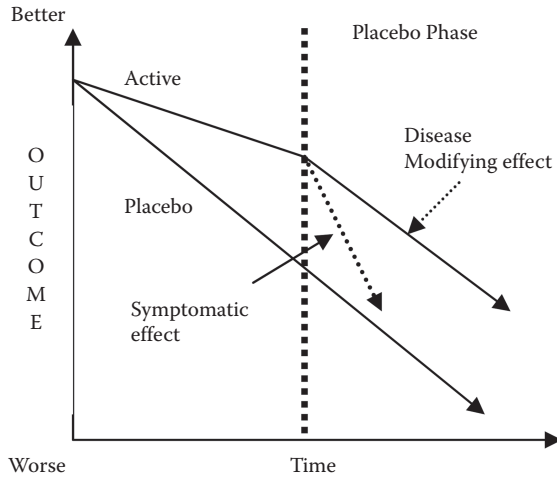


Fig. 1 Randomized withdrawal

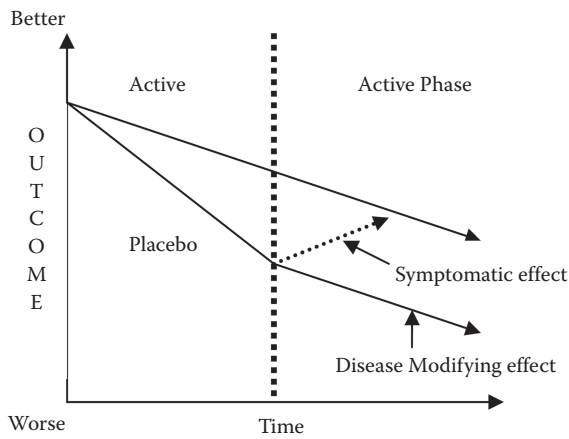


Fig. 2 Randomized start

This reasoning is predicated on the assumption that patients randomized to placebo will suffer more structural degradation than their counterparts in the active treatment group during the first period. As a result, they are expected to lag behind during the second treatment period. But, if the test drug has only symptomatic effects, the extent of deterioration in the disease condition of patients in both treatment groups would be similar in the first period. Therefore, any difference observed between the two treatment groups at the end of the first period is expected to vanish when all the patients are administered the same treatment during the second period.

Let $\mathcal{T}_1$ be the actual treatment difference at the end of the first treatment period. This difference could be expressed as the sum of the disease-modifying effect $\mathcal{T}_d$ and symptomatic effect $\mathcal{T}_s$, that is,

$$\mathcal{T}_1 = \mathcal{T}_s + \mathcal{T}_d \quad (1)$$

If $\bar{x}_{1a}$ and $\bar{x}_{1p}$ are the estimated means response at the end of the first period for patients treated with the test drug

and placebo, respectively, a natural estimator of $\mathcal{T}_1$ is the observed treatment difference

$$\hat{\mathcal{T}}_1 = \bar{x}_{1a} - \bar{x}_{1p} \quad (2)$$

Assuming that the outcome measure is normally distributed, a $100(1 - \alpha)\%$ two-sided confidence interval for $\mathcal{T}_1$ is

$$\left[\hat{\mathcal{T}}_1 - SE_1 * t_{\left(\frac{\alpha}{2}; \nu\right)}; \hat{\mathcal{T}}_1 + SE_1 * t_{\left(\frac{\alpha}{2}; \nu\right)} \right], \quad (3)$$

where SE_1 is the standard error of $\hat{\mathcal{T}}_1$ and $t_{\left(\frac{\alpha}{2}; \nu\right)}$ the upper $\frac{\alpha}{2}$ point of the central Student's t distribution with ν degrees of freedom. Clearly, if $\mathcal{T}_1$ is not statistically significant, neither will its components. Since the main purpose of randomized start/withdrawal designs is the estimation of the disease-modifying effect $\mathcal{T}_d$, the treatment difference observed at the end of the first period $\mathcal{T}_1$ must be statistically meaningful to warrant further investigation of the disease-modifying effect.

McDermott et al.^[3] have indicated that $\mathcal{T}_d$, the disease-modifying component of the treatment effect is estimated by

$$\hat{\mathcal{T}}_d = \bar{x}_{2a} - \bar{x}_{2p}, \quad (4)$$

where $\bar{x}_{2a}$ and $\bar{x}_{2p}$ are the estimated mean response at the end of the second treatment period for patients previously treated with the active drug and placebo, respectively. The reliability of this estimate depends upon which one of the two designs (i.e. randomized start and randomized withdrawal designs) is used. For instance, in a randomized withdrawal trial, the symptomatic component of the treatment effect is expected to disappear during the course of the second period when all patients are treated with placebo. Hence, it is conceivable to attribute any treatment difference lingering at the end of the second period to the disease modification activity of the active drug. In other words, in this case, $\mathcal{T}_d$ can reliably be estimated by $\hat{\mathcal{T}}_d$. In a randomized start trial, however, the reliability of $\hat{\mathcal{T}}_d$ hinges on the validity of the assumption that the total incremental change induced by treatment during the second period is the same in both treatment groups. If this assumption does not hold, $\hat{\mathcal{T}}_d$ could lead to erroneous and misleading inferences.

There is no doubt that patients treated the longest with a disease-modifying drug will exhibit, at the end of the trial, the greatest gain in terms of altering the degenerative process of the disease. However, due to the lack of sensitivity of some measuring instruments used to assess change in patients' outcomes, such a benefit may not necessarily translate into a treatment difference sufficiently large to

support a disease modification claim. In other words, $\hat{T}_d$ may be so small that one cannot infer from it a statistically meaningful disease-modifying effect.

To illustrate this point, consider an hypothetical clinical study designed as a randomized start trial for assessing whether a new drug compound has disease-modifying effects in the treatment of patients with mild to moderate Alzheimer's disease (Mini-Mental State Examination (MMSE) scores between 16 and 26 inclusive, with higher scores indicating less severe disease condition). Using a stratified randomization method with MMSE score (<20 ; ≥ 20) as the stratification factor, patients were randomly assigned to either the active drug or placebo. For simplicity, the primary outcome measure was the Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-cog), a 70-point scale instrument that evaluates cognitive change in patients (higher scores of ADAS-cog are indicative of greater cognitive impairment). The durations of both treatment periods were "adequate" for the disease-modifying effect to appear in the first period and the symptomatic effect to disappear in the second. Both treatment groups were comparable at baseline with respect to ADAS-cog and MMSE scores (see Table 1).

Now, suppose that at the end of the first period, patients in the active drug group showed some improvement in their cognitive ability while those in the placebo group worsened. The means change from baseline in ADAS-cog were 7.5 and 9.3 for the active drug and placebo groups, respectively. From the results in Table 2, the means MMSE scores were 23.8 for the active drug group versus 17.4 for the placebo group.

During the second treatment period however, less change was noted in the ADAS-cog scores for patients initially randomized to the active drug group as opposed to those previously treated with placebo (see table below). At the end of the second period, the treatment difference was so small that for all intended purposes, the two treatment groups were not statistically significantly different from one another as shown in Table 3.

Table 1 Mean Baseline Characteristics of Patients

Endpoints	Treatment groups	
	Active Drug	Placebo
ADAS-cog	23.3	22.9
Disease Severity MMSE	20.6	20.8

Table 2 Mean Cognitive and Disease Severity Results at the End of the First Period

Endpoints	Treatment groups		<i>p</i> -value
	Active Drug	Placebo	
Change from Baseline in ADAS-cog	7.5	9.3	0.034
Disease Severity MMSE	23.8	17.4	

Table 3 Mean Cognitive and Disease Severity Results at the End of the Second Period

Endpoints	Treatment groups		<i>p</i> -value
	Active Drug	Placebo	
Change from Baseline in ADAS-cog	7.1	8.6	0.417
Disease Severity MMSE	24.1	23.9	

Clearly, the trial data failed to provide sufficient evidence to support the claim of disease modification activity. But for the sake of argument, one could think that the small incremental cognitive changes observed during the second period in patients originally randomized to the active drug was due to a lack of sensitivity of the ADAS-cog. Harrison et al.^[4] reported that ADAS-cog is not sensitive for measuring cognitive changes in patients with MMSE scores greater than 18. Because the average MMSE score for patients randomized to the active group was 23.8 versus 17.4 in the placebo group at the end of the first period, it is questionable whether the treatment difference observed at the end of the second period is a true reflection of the disease-modifying effect of the active drug. Hence, in a randomized start design, failure to show a statistically significant treatment difference at the end of the second period is not necessarily indicative of lack of disease modification activity.

As pointed out by Mani,^[5] randomized start/withdrawal designs do have the potential to discriminate between symptomatic and disease-modifying effects of new drug compounds. However, they present a number of serious challenges:

- It is unclear how long the first period must be for the disease-modifying effects to take hold, and how long the second period must be for the symptomatic effects to vanish completely (randomized withdrawal) or to fully appear (randomized start).
- The second treatment period in both designs is not blinded and could be prone to assessment bias.
- It is unclear what effect size for disease modification should be postulated for sample size computation.
- In randomized withdrawal designs, replacing the active drug with placebo could present an ethical issue, especially when the patients are responding to the therapy. This practice has the potential of compromising patient accrual.
- Patients may drop out as their condition worsens. As a result, the missing data mechanism may be not missing at random (NMAR) which complicates further the analysis of the data.

Because of these issues, the implementation of these design techniques by clinical trialists has been very slow and as a result, alternative designs are being sought.

Parallel Group Designs

In a parallel group design, patients are randomly assigned to receive either the active drug or placebo and remain on that treatment for the entire duration of the trial. Most clinical scientists are familiar with parallel group designs; therefore, they are less vulnerable to procedural errors and are more likely to provide better precision for parameter estimation than the new and untested randomized start/withdrawal designs. Parallel group designs are blinded throughout the duration of the treatment. Thus, they are less prone to assessment bias. Assuming that the duration of the trial is long enough for the disease-modifying effects to take hold and the symptomatic effects to vanish, any observed treatment difference at the end of the study can conceivably be attributed to the disease-modification ability of the active drug. But how long a treatment duration is long enough? This question has been and continue to be unanswerable, at least for the time being. Furthermore, it is unclear what effect size for disease modification should be postulated for sample size computation.

METHODS FOR DATA ANALYSIS

Survival Analysis

Survival analysis with death as the endpoint of interest could prove an adequate method for demonstrating disease-modification effect as the treatment associated with the longest median survival time can be assumed to have slowed the progression of the illness. The drawback of this method is that the chronic diseases being discussed progress slowly in time. Thus, clinical trials with death as the primary endpoint will require an excessively long duration in order to observe an adequate number of events. Therefore, in the absence of variables that could be used as valid surrogates for death, survival analysis, though attractive from a theoretical standpoint, is not a practical analysis tool for assessing disease-modifying effects.

Slope Analysis

Slope analysis is a method for comparing the rate of decline of patients in the active group against that of patients in the placebo arm. Although the goal of slope analysis is to demonstrate disease-modifying effect, the interpretation of the difference between the rates of decline in the two treatment groups could significantly vary from one study design to another. In randomized start/withdrawal trials, “similar” rates of decline in both treatment groups during the second treatment period is viewed as a proof of disease-modifying effect. Hence, proving disease modification activity in randomized start/withdrawal trials is tantamount to demonstrating that the difference in the slopes of decline in the two treatment groups is small enough to be considered

scientifically and clinically irrelevant. In other words, randomized start/withdrawal trials will have to be designed as non-inferiority trials which will require large sample sizes. In parallel group trials however, divergent rates of decline with patients randomized to the active group expected to declines lower than their counter parts in the placebo group, are indicative of disease-modifying effect. As a result, parallel group trials will be designed as superiority trials. The duration of a parallel group study has to be long enough so that during the course of the trial disease-modifying effects could take hold while symptomatic effects wear off.

Slope analysis is often conducted in a number of ways. One way is to estimate the slope of decline of each individual patient and use these slope as the dependent variable in appropriate statistical models (see Ref. [6]). Another way is to use the primary endpoint in linear models such as a mixed model for repeated measurements (MMRM), and use the time by treatment interaction to test the equality of the slope of decline in both treatment groups. The main problem is that comparing the slopes of decline between the treatment groups may lead to erroneous conclusions because individual patients often tend to decline at different rates independent of the therapy they are assigned to. Furthermore, patients may drop out as their condition worsens. As a result, the missing data mechanism may be not missing at random (NMAR) which complicates further the analysis of the data.

CONCLUSION

In the treatment of a disease, the observed effect of a medicine could be an expression of its impact on either the root cause of the illness or the symptoms associated with it. If the medicine affects only the symptoms of the disease without any meaningful impact on its root cause, its effects are temporary and last as long as it is administered. However, if the drug alters the root cause of the disease, it will not only relieve the symptoms but its effects will be permanent. Hence, the goal of drug developers is to develop drugs capable of impacting the mechanism responsible for the destructive effects of the illness under study. This goal is attainable when the etiology of the disease is known and the effects of the drug on the key etiologic factors can easily be measured. When the etiology and reliable markers of the disease are unknown however, measuring the impact of the drug on the root cause of the disease is a formidable task, a task in which researchers have to find innovative ways to demonstrate the impact of the drug on the pathophysiological process of the disease. It is in this context that the concept of “disease modification” discussed in this chapter was born. The concept is still evolving, as is the regulatory guidance pertaining to methods for demonstrating disease modification.

Indeed, there is no consensus among the health authorities. As pointed out by Vellas et al.,^[7,8] in order to approve

a disease-modifying therapy, the FDA requires a new drug product to show meaningful effects on the mechanism of action of the target disease. For the European Medicines Agency (EMA) on the other hand, the approval is based on the demonstration of sound clinical outcomes. Hence, the FDA tends to favor randomized start/withdrawal clinical trials for demonstrating disease-modifying effects, whereas the EMA favors the traditional parallel group trials. The statistical methods presented here are consistent with the current understanding of the notion of disease modification. As this concept continues to evolve and more information becomes available on the diseases, the design techniques as well as the analysis methodologies described here will be adapted accordingly.

REFERENCES

1. Leber, P. Observations and suggestions on antidementia drug development. *Alzheimer. Dis. Assoc. Disord.* **1996**, *10* (suppl 1), 31–35.
2. Leber, P. Slowing the progression of Alzheimer disease: methodologic issues. *Alzheimer. Dis. Assoc. Disord.* **1997**, *11* (suppl 5), S10–S21.
3. McDermott, P.M.; Hall, W.J.; Oakes, D.; Eberly, S. Design and analysis of two-period studies of potentially disease-modifying treatments. *Control. Clin. Trials* **2002**, *23*, 635–649.
4. Harrison, J.; Minassian, S.L.; Jenkins, L.; Black, R.S.; Koller, M.; Grundman, M. A Neuropsychological test battery for use in Alzheimer disease clinical trials. *Arch. Neurol.* **2007**, *64* (9), 1323–1329.
5. Mani, R.B. The evaluation of disease modifying therapies in Alzheimer's disease. *Stat. Med.* **2004**, *23*, 305–314.
6. Thomas, R.G.; Berg, J.D.; Sano, M.; Thal, L. Analysis of longitudinal data in Alzheimer's disease clinical trial. *Stat. Med.* **2000**, *19*, 1433–1440.
7. Vellas, B.; Andrieu, S.; Sampaio, C.; Wilcock, G. Disease-modifying trials in Alzheimer's disease: A European task force consensus. *Lancet Neurol.* **2007**, *6*, 56–62.
8. Vellas, B.; Andrieu, S.; Sampaio, C.; Coley, N.; Wilcock, G. Endpoints for trials in Alzheimer's disease: A European task force consensus. *Lancet Neurol.* **2008**, *7*, 436–450.

Dissolution Profile Comparison: Statistical Methods

Harry Yang

MedImmune, LLC, Gaithersburg, Maryland, U.S.A.

Abstract

In vitro dissolution testing is a critical analytical methodology used for monitoring product quality and assessing performance characteristics after scale up and post-approval changes. In this article we review published statistical work on dissolution profile comparison and related regulatory requirements. A Bayesian approach that overcomes many of the statistical issues of the current methods such as f_2 statistic is described in detail.

Keywords: Bayesian model; Dissolution profile; f_2 statistic; *In vitro* release; Multivariate statistical distance; Similarity.

INTRODUCTION

The effectiveness of a drug in a solid oral dosage form, such as a tablet or capsule, relies on the release of the drug substance from the drug product, the dissolution of the drug under physiological conditions, and diffusion of the drug through the gut wall into the body. *In vitro* dissolution testing is used to mimic this *in vivo* process, even acting as a surrogate measure of *in vivo* performance if a strong *in vitro* and *in vivo* correlation is established. Since dissolution testing can be carried out using specialized laboratory equipment, it is relatively fast and inexpensive when compared to *in vivo* studies.

In vitro dissolution testing is used to achieve different objectives throughout the lifecycle of a drug. In the early stages of drug development, it is used to guide formulation development. In process development, it is used to assess the effects of critical manufacturing variables on drug performance. For commercial manufacturing, it can also be used for quality control of scale-up and commercial batches. Furthermore, it is a useful technique for assessing the impact of certain post-approval changes, as described in several regulatory guidelines published since 1997. These numerous regulatory documents cover topics including general recommendations for dissolution testing, approaches for setting dissolution specifications, statistical methods for comparing dissolution profiles, and conditions under which a waiver for an *in vivo* bioequivalence study is granted.

It is of great scientific and regulatory interest to assess the similarity between dissolution profiles. For example, the demonstration of similarity of dissolution profiles before and after a manufacturing or formulation change is a regulatory requirement. Various methods for comparing dissolution profiles have been suggested in both scientific literature and regulatory documents. They fall into two broad categories, namely, model-independent and

model-dependent methods. Among the model-independent approaches, of note are f_2 statistic and multivariate statistical distance (MSD) test. Various model-dependent methods are also used for dissolution profile comparison, including ones dependent on mechanistic models of drug dissolution processes and others on empirical models such as Weibull. More recently, a Bayesian test procedure was suggested that overcomes many of the statistical issues of the current methods.

In this article we review published statistical work on dissolution profile comparison and related regulatory requirements. A detailed critique of the current methods is provided, followed by a Bayesian approach that overcomes many of the issues of the current methods. We also note directions for future research on dissolution profile comparisons.

DISSOLUTION PROFILE COMPARISON

Dissolution Profile

Consider a dissolution study with times $t_i, i = 1, \dots, k$. Let $R_j = (R_{j1}, \dots, R_{jk})'$, $j = 1, \dots, n_1$, and $T_l = (T_{l1}, \dots, T_{lk})'$, $l = 1, \dots, n_2$, be the dissolution measurements of reference tablet j and test tablet l , respectively. For simplicity, we assume $n_1 = n_2 = n$. Of interest is to compare the dissolution profiles of the reference and test products characterized by R_j and T_l . When designing a dissolution experiment, consideration needs to be given to dissolution media, apparatus type, and hydrodynamics such as agitation rate appropriate for the product. A dissolution testing run typically involves dissolving 6–12 dosage units in agitated media, measuring API content by HPLC or UV spectroscopy at set time points. Figure 1 shows a typical dissolution profile which is the curve of measured percent of API dissolved plotted against time.

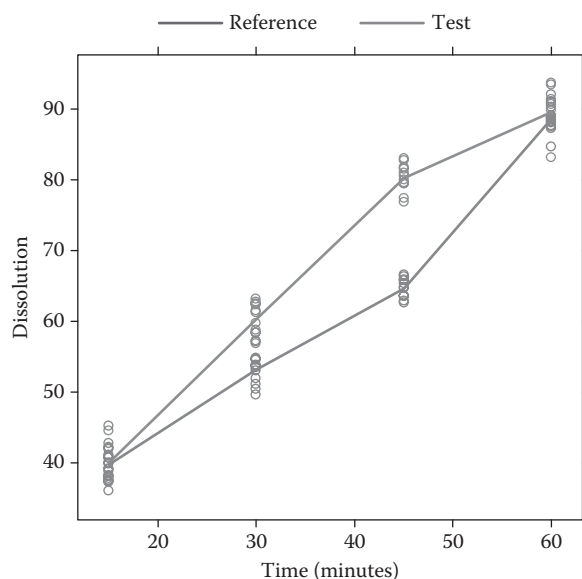


Fig. 1 Dissolution profiles of test and reference drugs (adapted from Novick et al.^[30])

Statistical Methods for Assessing Similarity

As previously discussed, dissolution profile comparisons can be carried out using either model-independent or model-dependent methods. Here the term “model” refers to the function used to describe the dissolution profile, not the sampling distribution of the response values.

Model-Dependent Approaches

Univariate Methods

Moore and Flanner^[1] proposed a model-independent approach to measure the similarity between drug dissolution profiles of a test and reference formulation. The method includes the calculation of two mathematical indices:

$$f_1 = \frac{\sum_{i=1}^k |\bar{R}_i - \bar{T}_i|}{\sum_{i=1}^k \bar{R}_i} \times 100$$

$$f_2 = 50 \log_{10} \left\{ 100 \left(1 + \frac{1}{k} \sum_{i=1}^k w_i w_i (\bar{R}_i - \bar{T}_i)^2 \right)^{-1/2} \right\}$$

where $\bar{R}_i = \frac{1}{n} \sum_{j=1}^n R_{ij}$ and $\bar{T}_i = \frac{1}{n} \sum_{j=1}^n T_{ij}$ are the sample means of the reference and test results at time t_i , $i = 1, \dots, k$, respectively, and w_i is an optional weighting factor.

The factor f_1 , generally referred to as the difference factor, measures the relative difference in percent (%) dissolution between the two curves by time point. It is equal to zero when the test and reference profiles are identical and becomes larger when the difference between the two profiles increases. The similarity factor f_2 is a measurement of the similarity in the percent (%) dissolution between the two curves. It assumes a value between 0 and 100, with $f_2 = 100$ when the two dissolution curves are identical and decreasing to zero as the dissimilarity between the two profile increases. A similarity test procedure based on similarity factor f_2 with $w_i = 1$ either in combination with the difference factor f_1 or alone, has been recommended in various regulatory publications.^[2–8] In general, it is required that $f_2 \geq 50$ in order to deem two dissolution profiles similar. Thus, while not explicitly stated, inference based on f_2 is seemingly intended to test the hypotheses:

$$H_0 : F_2 \leq 50 \quad \text{versus} \quad H_a : F_2 > 50,$$

where

$$F_2 = 50 \log_{10} \left\{ 100 \left(1 + \frac{1}{k} \sum_{i=1}^k (E[R_i] - E[T_i])^2 \right)^{-1/2} \right\}$$

and, $E(R_i)$ and $E(T_i)$ are the expected values, across the population of dosage units, of the dissolution measurements at time t_i ($i = 1, \dots, k$), from the reference and test batches, respectively.

As noted by Shah et al.^[9] and empirically evaluated by Ma, et al.,^[10] f_2 is a biased estimator of the true similarity measure F_2 . A method of calculating f_2 using a bias correction was suggested by Shah et al.^[9] and Ma, et al. (2000) explored the use of a bootstrap approach to construct a confidence interval on F_2 since the sampling distribution of f_2 is difficult to derive analytically.

A few other model-independent similarity factors have also been suggested in the literature. One based on the index of Rescigno^[11] was originally used to compare plasma drug concentration curves and mean dissolution profiles. As with f_2 , these methods do not explicitly define the statistical hypotheses to be tested for similarity, nor can the operating characteristics of the tests (such as Type I and II errors) be determined at all.

Multivariate Statistical Distance

Another class of model-independent methods hinges on the normality assumption underlying the *in vitro* release values observed at an individual time point and constructs a measure of the distance between two sets of multivariate random variables.^[12] It is assumed that

$$R_{ij} \sim N(\mu_R, \Delta) \quad \text{and} \quad T_{ij} \sim N(\mu_T, \Delta).$$

Let $\bar{R} = (\bar{R}_1, \dots, \bar{R}_k)$, $\bar{T} = (\bar{T}_1, \dots, \bar{T}_k)$ be the sample mean vectors comprised of sample mean values at times t_i , $i = 1, \dots, k$, S_R and S_T the sample variance-covariance matrices for the reference and test lots, respectively, and S_p the pooled sample variance-covariance matrix.

$$S_p = \frac{(k-1)S_R + (k-1)S_T}{2k-2} = \frac{1}{2}(S_R + S_T).$$

Thus $T^2 = n_1 n_2 (n_1 + n_2)^{-1} [(\bar{R} - \bar{T}) - (\mu_R - \mu_T)]' S_p^{-1} [(\bar{R} - \bar{T}) - (\mu_R - \mu_T)]$ is the Hotelling's T^2 statistic. Recall that

$$\frac{n_1 + n_2 - k - 1}{(n_1 + n_2 - 2)k} T^2 \sim F_{k, n_1 + n_2 - k - 1}.$$

Then, a 95% confidence region for $\mu_R - \mu_T$ is obtained as follows:

$$C(\bar{R}, \bar{T}, S_p) = \left\{ T^2 \leq \frac{(n_1 + n_2 - 2)k}{n_1 + n_2 - k - 1} F_{k, n_1 + n_2 - k - 1; 0.05} \right\}.$$

Tsong et al.^[12] proposed a method that compares the maximum Mahalanobis distance

$$D_{M,\alpha} = \sup_{\{(\mu_R, \mu_T) \in C(\bar{R}, \bar{T}, S_p)\}} \sqrt{(\mu_R - \mu_T)' S_p^{-1} (\mu_R - \mu_T)}$$

with the tolerance limit

$$TL = \sqrt{\delta^2 J' S_p^{-1} J},$$

where δ is the maximum allowable dissimilarity (i.e. difference) at any sampling time, and J denotes the $k \times 1$ vector of unities. Similarity is claimed if $D_{M,\alpha} \leq TL$.

However, although $|(\mu_R - \mu_T)_i| \leq \delta$, for $i = 1, \dots, k$, implies $D_{M,\alpha} \leq TL$, the converse is not true. Therefore it is conceivable that the two dissolution curves may have a meaningful difference at one or more individual sampling times, and still have the maximum Mahalanobis distance remain acceptable. Notice also that TL , as defined above, depends on the observed pooled sample variance S_p , calculated from the same data used for the similarity assessment. Thus, the definition of similarity – and the inference about similarity for the two products being compared – will vary from data set to data set.

One issue associated with the above multivariate distance methods, as with the univariate f_1 and f_2 approaches, is the difficulty in assessing their operating characteristics. Saranadasa and Krishnamoorthy,^[13] in an effort to address this shortcoming, developed a multivariate test of size α for assessing the similarity of two dissolution profiles assuming parallelism. Chow and Ki^[14] used a time series approach to account for the correlation between

consecutive time points and overcome the drawback of the multivariate distance methods.

Model-Dependent Methods

Model-dependent methods are also listed in regulatory guidance documents as alternative procedures for dissolution profile similarity testing for both IR and MR products. These rely on describing dissolution profiles through mathematical functions, based on understanding of the dissolution processes underlying the dissolution measurements over time. Because biological systems and related physio-chemical processes involved in drug dissolution are exceedingly complex, both mechanistic models^[15–19] and empirical models have been proposed to describe dissolution concentration-time profiles. Empirical functions that can be used to fit the dissolution profiles include Weibull,^[20] exponential, probit, Gompertz, and logistic.^[21] In general, the Weibull is considered to be the most flexible able to describe a wide variety of profile shapes.^[22] More recently, linear mixed effects and non-linear mixed effects models have been suggested.^[23,24] One advantage of these methods is that they permit an explicit specification of the covariance structure of the data. Potential pitfalls of these methods are: 1) the model parameters may be biased or not estimable if the sampling points are not appropriately chosen, and 2) the inferences may be affected by the choices of time points.

Merits and limitations of the aforesaid methods are briefly summarized in Table 1.

REGULATORY GUIDANCE

FDA Guidelines

As discussed previously, an important application of dissolution testing is to address issues related to post-marketing changes. Depending on the nature and extent of the changes, dissolution testing may and may not be needed or sufficient. To help sponsors, FDA guidance^[5] clearly defines extent of dissolution testing needed to support postapproval changes.

Nature of Product or Process Changes

Since 1995, the U.S. FDA has issued several guidelines concerning approval of solid oral dosage form generic drugs or drugs that have undergone post approval changes.^[2–7] In the guidances, both the type and level of change are defined, as shown in Table 2.

Depending on the level and type of change, detailed recommendations on CMC tests (in particular, dissolution testing), *in vitro* and *in vivo* studies, and the necessary documentation that support the change are provided. For detail please refer to relevant FDA guidance.

Table 1 Merits and limitation of current methods for dissolution profile comparison

Model	Approach	Advantage	Disadvantage
Independent	f_2	• Intuitive • Easy to calculate	• Lack of statistical justification
	MSD	• Does not rely on specific model • Account for variances and co-variances among observations	• Lack of statistical rigor
Dependent	Weibull	• Empirical	• Model parameters may not be estimable or estimates are biased
	Exponential	• Permit specification of covariance structure	• Inferences may depend on time points
	Probit		
	Gompertz		
	Logistic		
	Higuchi	• Mechanistic	• Model parameters may not be estimable or estimates are biased
	Korsmeyer-Peppas	• Permit specification of covariance structure	• Inferences may depend on time points
	Hixson-Crowell		

Dissolution Testing Based on Biopharmaceutical Classification System

The extent of dissolution testing required by regulatory agencies depends on the solubility, permeability, dissolution, and pharmacokinetics of the drug product. The Biopharmaceutics Classification System (BCS), originally proposed by Amidon et al.^[25] adopted by the FDA, classifies drug substances into four groups based on solubility and permeability, as shown in Table 2. The objective of the BCS is to define the conditions under which an *in vitro* - *in vivo* correlation (IVIVC) is expected. In a 1997 FDA guideline^[3], IVIVC is defined as “[a] predictive mathematical model describing the relationship between an *in vitro* property of an extended release dosage form (usually the rate or extent of drug dissolution or release) and a relevant *in vivo* response, e.g., plasma drug concentration or amount of drug absorbed.” When an IVIVC is established, the dissolution test may be used as a surrogate for *in vivo* studies for certain scale-up and post approval changes, reducing the number of bioequivalence studies

required. Table 3 lists the four BCS classes and, their associated IVIVC expectations.

The FDA guidance presents a comprehensive perspective on 1) methods of developing an IVIVC and evaluating its predictability; 2) using an IVIVC to set dissolution specifications; and 3) applying an IVIVC as a surrogate for *in vivo* bioequivalence when it is necessary to document bioequivalence during the initial approval process, scale-up, or because of certain pre- or post-approval changes.

Dissolution Test Specifications

FDA 1997 guidance^[4] describes three categories of dissolution test specifications for immediate release drug products, along with their intended purposes. They are summarized in Table 4.

Dissolution Profile Comparison

FDA guidance^[3] requires that dissolution profile comparisons be carried out using model independent or model

Table 2 Levels and types of changes

Change	Description	
Level	1	Changes that are unlikely to have any detectable impact on formulation quality and performance
	2	Changes that could have a significant impact on formulation quality and performance
	3	Changes that are likely to have a significant impact on formulation quality and performance
Type	1	Components and composition
	2	Site of manufacture
	3	Scale of manufacture
	4	Manufacturing process and equipment

Table 3 Biopharmaceutical classification

Class	Solubility	Permeability	IVIVC expectation
I	High	High	IVIV correlation if dissolution rate is slower than gastric emptying rate. Otherwise, limited or no correlation is expected
II	Low	High	IVIV correlation expected if <i>in vitro</i> dissolution rate is similar to <i>in vivo</i> dissolution rate, unless dose is very high
III	High	Low	Absorption is rate-determining and limited or no IVIV correlation with dissolution rate
IV	Low	Low	Limited or no IVIV correlation expected

Table 4 Categories of dissolution test specifications

Type	Purpose
One point	As a routine quality control test (for highly soluble and rapidly dissolving drug products)
Two-point	- For characterizing the quality of the drug product - As a routine quality control test for certain types of drug products (e.g., slow dissolving or poorly water soluble product like carbamazepine)
Dissolution profile	- For accepting product sameness under SUPAC-related changes - To waive bioequivalence requirements for lower strengths of a dosage form

dependent methods. Descriptions of the methods noted in the guidance are provided in Table 5.

In the FDA 1997 guidance for SUPAC-MR: Modified Release Solid Oral Dosage Forms,^[5] the similarity factor f_2 is specifically recommended along with the requirement that the average difference at any dissolution sampling time point should not be greater than 15% between the pre- and post-change products, though other model independent or model-dependent methods can also be used. In addition, it is also stated that an f_2 value below 50 does not necessarily indicate lack of similarity, though

if $f_2 < 50$ a justification needs to be submitted to substantiate a similarity claim.

Differences between FDA and Other Regulatory Guidelines

As with the FDA, use of f_2 to assess profile similarity is recommended by the European Medicines Agency (EMA),^[8] again with similarity declared if $f_2 > 50$. However, EMA specifies that for calculation of f_2 , the following conditions need to be met: 1) a minimum of three time points (excluding zero) are needed; 2) the time points should be the same for the reference and test products; 3) twelve (12) individual values are required for every time point; 4) there can be no more than one mean value of $> 85\%$ dissolved; and 5) the coefficient of variation (CV) should be less than 20% for the first time point and less than 10% for all subsequent time points. If more than 85% of the drug is dissolved within 15 minutes for both products, dissolution profiles may be accepted as similar without further mathematical evaluation. Whereas the FDA guidance states that model-dependent methods can be used for dissolution profile similarity assessment, the EMA guideline makes no explicit mention of methods other than f_2 .

As noted by Diaz et al.,^[26] there has been a proliferation of divergent similarity requirements from regulatory authorities around the world. The differences not only reside in statistical considerations but also aspects of *in vitro* dissolution testing study design, such as the minimum

Table 5 Test procedures for dissolution comparison for immediate release solid oral dosage forms

Method Type	Conditions	Description
Model-Independent	f_2 Within batch variation $\leq 15\%$	- Estimate f_1 and f_2 - Similarity is claimed if $f_1 \leq 15$ and $f_2 > 50$
	MSD Within batch variation $> 15\%$	- Estimate MSD between test and reference mean dissolutions - Estimate 90% CI of true MSD - Similarity is achieved if the upper limit of the 90% CI is no larger than the similarity limit
Model-Dependent		- Select the most appropriate model and fit the data to the model Calculate MSD in model parameters and 90% CI of the true MSD - Similarity is achieved if the upper limit of the 90% CI is no larger than the similarity limit

number of time points and selection of the last time point, as well as a maximum %CV at individual time points. For example, FDA requires a minimum of three time points excluding zero but the Brazilian regulatory authority recommends five time points (excluding zero).

CRITIQUES OF CURRENT MODEL-INDEPENDENT METHODS

Requirements for Statistical Tests

Various critiques of the statistical approaches described above for demonstrating similarity for *in vitro* dissolution profiles have appeared in the literature. Thoughtful consideration was provided by Eaton et al.,^[27] who defined two normative requirements (NR) for a statistically rigorous dissolution profile similarity test:

- NR1: A specified function of the population parameters (not involving data) should be used to define dissolution profile similarity.
- NR2: No matter what testing procedure is used, there needs to be sufficiently detailed knowledge about the power function to allow and assessment of the probabilities of Type I and Type II errors.

NR 1 ensures that the assessment of dissolution profile similarity can be made using statistical hypothesis testing. NR 2 warrants that the performance of the test procedure can be evaluated. Three additional NRs were suggested by LeBlond et al.^[28] as follows:

- NR3: The parametric definition of similarity should be independent of the statistical methodology used to test for similarity.
- NR4: The test for similarity should make clear the inference space for the conclusion. For instance, does the conclusion apply to the populations of test and reference batches or only to those batches providing data for the comparison?
- NR5: A minimum confidence level should be included which properly accounts for estimation errors.

Taken together, these requirements can be used to assess the validity of statistical test procedures for dissolution profile comparison.

Statistical Issues with f_2 and MSD

f_2 Statistic

As originally described by Moore and Flanner,^[1] the f_2 acceptance criterion was defined in terms of the observed data. The similarity criterion of $f_2 > 50$, was justified on the basis that it corresponded to a constant difference of

10 between the observed test and reference results, which seems intuitively reasonable when test and reference dissolution profiles are parallel. This definition provided an easily calculated statistic and decision criterion. Dissolution similarity tests using f_2 have thus been conducted for the past two decades without consideration of underlying population parameters or other modeling considerations, such as correlation among the response measured at different time points. The use of f_2 has been criticized on this and other bases by various authors. Most problematic is that it lacks statistical justification, as it does not meet any of the above normative requirements. It is likely $f_2 > 50$ while the two dissolution profiles are dissimilar, and vice-versa.^[29]

MSD

FDA agencies provide for the use MSD in cases where the dissolution measurement is highly variable. The MSD approach is an improvement over the current practice using f_2 as recommended in guidance documents. First, MSD does rely on an underlying statistical model. Second, the MSD takes into account variances and covariances among the different dissolution time points, whereas f_2 does not. However, there is no recommended decision criterion for MSD as there is with f_2 . While some researchers have proposed similarity criteria for MSD or MSD-like statistical procedures, there currently is no industry standard and users of MSD are left to justify their own criteria for similarity.

ALTERNATIVE METHOD

Over the last decade, the importance of risk-based decision making in the pharmaceutical industry has been increasingly emphasized by global regulatory agencies. Risk-based decisions oftentimes require an assessment of decision error probabilities, thus proper statistical modeling of the underlying data-generation mechanism. To align dissolution similarity practice with current regulatory thinking, it seems imperative to introduce additional statistical rigor into dissolution similarity comparisons. A first step in this direction could be to re-define the f_2 “fit factor” parametrically. To this end, Novick et al.^[30] introduced the parametric function F_2 .

Model

In the following discussion, the problem of interest is to compare the dissolution from a single reference batch with those from a single test batch. Let $Y_{br} = (Y_{b1r}, Y_{b2r}, \dots, Y_{bpr})^T$ be the vector of dissolution measurements from vessel r ($r = 1, 2, \dots, R$) from batch b ($b = \text{reference, test}$) at time points t_p , $i = 1, 2, \dots, k$. It is assumed that

$$Y_{br} = \mu_b + \epsilon_b$$

where $\mu_b = (\mu_{b1}, \mu_{b2}, \dots, \mu_{bk})^T$ is the mean of Y_{br} and ϵ_b is the vector of measurement errors, Y_{br} and ϵ_b following a multi-variate normal distribution, $N(\mu_b, V_b)$.

The structure of the variance-covariance matrices V_b may take on various forms, allowing for modeling potential correlations among measurements taken at different time points. For instance, if Y_{br} follows an auto-regressive model AR(1), V_b may be expressed as $(\sigma^2 \rho^{|i-j|})$. Table 6 lists some examples from SAS/STAT 9.2 User's Guide.^[31]

It is worth noting that the variance can also be modeled as a function of the mean. For example, one may use $\text{Var}(Y_{br}) = \sigma_1^2 + \sigma_2^2(|\mu_{br}/100|)(|1 - \mu_{br}/100|)$ to capture the relationship between the mean response values and corresponding variances at various time points. One may also assume $V_{\text{Reference}} = V_{\text{Test}}$ if supported by historical data, which simplifies model fitting.

Hypotheses for Dissolution Profile Comparison

Novick et al.^[30] suggested establishing dissolution profile similarity through testing the following hypotheses:

$$H_0: \left| \mu_{\text{Ref},t} - \mu_{\text{Test},t} \right| \geq 15 \text{ for some } t = 1, 2, \dots, k \text{ OR } F_2 \leq 50$$

$$H_a: \left| \mu_{\text{Ref},t} - \mu_{\text{Test},t} \right| < 15 \text{ for all } t = 1, 2, \dots, k \text{ AND } F_2 > 50$$

Bayesian Test Procedure

Novick et al.^[30] also suggested a Bayesian procedure for testing the above hypotheses. Let

$$\delta(p) = \max_{t=1, \dots, k} \left| \mu_{\text{Ref},t} - \mu_{\text{Test},t} \right| \text{ and}$$

$$q = \Pr(\delta(k) < 15 \text{ and } F_2 > 50 | \text{data})$$

A decision can be made to accept H_a if $q > q_0$, where q_0 is a pre-specified value between 0 and 1 but close to 1; otherwise H_0 is accepted. Calculation of the above probability can be carried out in a couple of ways: 1) If the joint posterior distribution of the model parameters can be analytically derived, the probability in (4) can be estimated by simulating random samples from the posterior distribution and calculating the percent of simulated samples for which the conditions $\delta(k) < 15$ and $F_2 > 50$ are met; or 2) when a closed-form solution for the posterior distribution is unavailable, the Markov Chain Monte

Carlo (MCMC) method can be used to estimate the probability q . This method indirectly draws random samples from the joint posterior distribution by constructing a Markov chain that has the joint posterior distribution as its limiting distribution. The MCMC method can be computationally intensive, however with the availability of MCMC software packages such as OpenBUGS,^[32] JAGS,^[33] and SAS Proc MCMC^[31], the computations can be readily carried out.

The use of a Bayesian approach for assessing dissolution similarity has several advantages. First, it allows for incorporating the uncertainty in the model parameters in the posterior probability calculation. Second, it provides a level of assurance to the similarity claim by way of calculating posterior probability for meeting the similarity acceptance criteria. Third, the availability of MCMC software packages makes it more straightforward to make statistical inferences about functions of model parameters when compared to frequentist methods. Lastly, Bayesian inference is not as limited by small amounts of data as the frequentist methods are.

Posterior Distribution

Let $Y_b = (Y_{b1}, Y_{b2}, \dots, Y_{bk})$, $\bar{Y}_b = R^{-1} \sum_{r=1}^R Y_{br}$, and $S_b^2 = \sum_{r=1}^R (Y_{br} - \bar{Y}_b)(Y_{br} - \bar{Y}_b)^T$. Consider the parameters μ_b and V_b to follow the conjugate prior distributions:

$$\mu_b | V_b \sim N(\mu_0, V_b / k_0)$$

$$V_b \sim \text{IW}(\Lambda_0^{-1}, \nu_0)$$

where μ_0 , k_0 , Λ_0^{-1} , and ν_0 are hyper-parameters, of the Normal and inverse Wishart distributions. It is well-known that

$$V_b | Y_b \sim \text{IW}(\Lambda_R^{-1}, \nu_R)$$

$$\mu_b | (V_b, Y_b) \sim N(\mu_R, V_b / k_R),$$

where $\nu_R = \nu_0 + R$, $k_R = k_0 + R$, $\nu_R = \nu_0 + S_b^2 + \left(\frac{k_0 R}{k_0 + R} \right) (\bar{Y}_b - \mu_0)(\bar{Y}_b - \mu_0)^T$, and

$$\mu_R = \left(\frac{k_0}{k_0 + R} \right) \mu_0 + \left(\frac{R}{k_0 + R} \right) \bar{Y}_b.$$

Table 6 Some examples of covariance structures

Covariance Structure	Covariance Element	Description
AR(1)	$\sigma^2 \rho^{ i-j }$	Equally-spaced time points, homogeneous variance
ARH(1)	$\sigma_i \sigma_j \rho^{ i-j }$	Equally-spaced time points, heterogeneous variance
SP(POW)	$\sigma^2 \rho^{ t_i - t_j }$	Unequally-spaced time points
UN	σ_{ij}	Unstructured

With a Jeffrey's prior ($k_0 = 0$, $\nu_0 = -1$, $|A_0| = 0$), samples from the posterior distribution may be drawn with $\Lambda_R = S_b^2$ and $\mu_R = \bar{Y}_b$ with $\nu_R = R-1$ and $k_r = R$. Assuming that $V_{\text{Ref}} = V_{\text{Test}} = V$, the posterior distributions are:

$$V | Y_b \sim \text{IW}(\Lambda_R^{-1}, \nu_0 + 2R)$$

$$\mu_b | (V, Y_b) \sim N(\mu_R, V / k_r),$$

where $\Lambda_R = \Lambda_0 + \sum_b \left\{ S_b^2 + \left(\frac{k_0 R}{k_0 + R} \right) (\bar{Y}_b - \mu_0)(\bar{Y}_b - \mu_0)^T \right\}$,

and $\mu_R = \left(\frac{k_0}{k_0 + R} \right) \mu_0 + \left(\frac{R}{k_0 + R} \right) \bar{Y}_b$. With a Jeffrey's prior, samples from the posterior distribution may be drawn with $\Lambda_R = \sum_b S_b^2$ and $\mu_R = \bar{Y}_b$ with $\nu_R = 2R-2$ and $k_r = R$.

For dissolution profile comparisons, other prior distributions than the ones described above may be used, if appropriate, based on knowledge of the model parameters. It is worth pointing out that use of the inverse Wishart distribution as a prior for V_b has been criticized for being too sensitive to the chosen degrees of freedom.^[34] In the literature, several remedies such as the scaled inverse Wishart have been proposed.^[34,35] For example, Gelman and Hill^[35] method begins with decomposing the covariance matrix into the form:

$$\Sigma = \text{Diag}(\xi) \mathbf{Q} \text{Diag}(\xi),$$

where $\text{Diag}(\xi)$ and $\mathbf{Q}$ are diagonal and correlation matrices, respectively. Such an re-expression of the covariance matrix makes it possible assign priors to the variances and correlation separately.

The use of prior information is a critical step in statistical inference based on a posterior distribution, and the selection of prior distributions has been much discussed in the literature. An informative prior may be established from early experiments, which may render more power for the test procedures than the use of non-informative priors. In the absence of relevant data, subjective prior information from expert opinions may be used.

Example

In the following example, we illustrate the use of the Bayesian method discussed above, using one of the four examples by Novick et al.^[30] The AR(1) covariance structure in Table 6 is assumed with $\sigma^2 = 6.25$ and $\rho = 0.4$, respectively. It is further assumed that the test and reference products were tested at times 15, 30, 45, 60, and 75 minutes, with 12 vessels tested at each time point. The mean values at each time point are provided in Table 7; the dissolution profiles are shown in Fig. 2. The posterior probability q previously defined was calculated for three choices of prior distributions: 1) Jeffrey's prior; 2) Scaled inverse Wishart

with $\mu_{\text{bt}} \sim N(50, 30^2)$, diagonal elements of $\text{Diag}(\xi)$ equal to $U(0, S_i)$ where S_i is a frequentist upper 99% confidence limit for the i^{th} diagonal element of V , and $\mathbf{Q} \sim \text{IW}(\mathbf{I}, p+1)$, where $\mathbf{I}$ is the identity matrix; and 3) AR(1) with $\mu_{\text{bt}} \sim N(50, 30^2)$, $\sigma \sim U(0, 100)$ and $\rho \sim U(0, 1)$.

For the Jeffrey's prior, twenty thousand (20,000) independent posterior samples were drawn. For the scaled inverse Wishart and AR(1) priors, four independent chains were run in OpenBUGS, each with burnin = 5,000 and posterior samples = 10,000 after thinning by 25, for a total of 40,000 posterior draws; effective sample sizes for all parameters were above 35,000. From the model parameters, posterior distributions for F_2 and $\delta(p)$ were generated and posterior probabilities were calculated for $\Pr(F_2 > 50)$, $\Pr(\delta(p) < 15)$ and $q = \Pr(\delta(p) \text{ and } F_2 > 50)$.

The true values, based on the true dissolution profiles, were $F_2 = 51.3$ (pass) and $\delta(p) = 16$ (fail). For this data set, the observed statistic $f_2 = 52.3$, from which one would declare that the two dissolution profiles are similar using the FDA-recommended approach, yet one of the

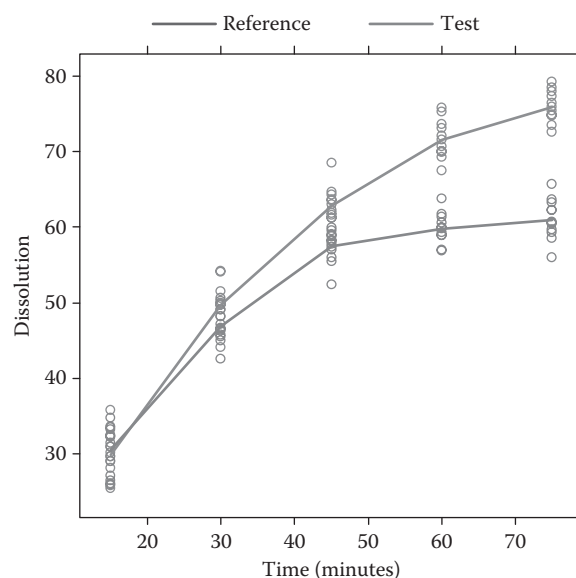


Fig. 2 Dissolution profiles from the simulated example. The response represents percent of dissolution (adapted from Novick et al.^[30])

Table 7 Means and mean differences at each time point for reference and test batches

Time (minute)	Reference	Test	Difference
15	30	30	0
30	50	48	2
45	64	58	6
60	72	60	8
75	76	60	16

mean differences between batches is quite high (=16%); thus, adding the criterion on $\delta(p)$ provides some protection against declaring in such a scenario. The results from the Bayesian analyses using the three prior-distribution methods were all similar; the results are provided in Table 8.

Because the data were generated with an AR(1) covariance structure, it is reasonable to expect that the AR(1) assumption in the posterior sampling would result in more precise estimates of the parameters than the other choices of priors. From the three data examples in section 2, it appears that the three choices of priors result in similar performance.

DISCUSSION

Dissolution testing is an important component of pharmaceutical development. It can be used for guiding the development of new formulations, monitoring drug product lot-to-lot quality, and ensuring drug product performance after certain post approval changes. Despite the importance of dissolution testing, there are inconsistencies in regulatory requirements for dissolution testing and dissolution similarity profile comparisons. Although the f_2 similarity factor approach suffers many deficiencies, it is recommended by many regulatory authorities for assessing dissolution profile similarity. Likewise, despite lack of statistical robustness, the MSD method is also adopted by some regulatory agencies including FDA.

This article provides an overview of regulatory requirements for dissolution similarity assessment, critiques of the f_2 and MSD approaches, and discussion of published statistical methods proposed as remedies for the deficiencies of the f_2 and MSD methods. A Bayesian alternative for dissolution profile comparisons is presented in detail, which is intended to be an improvement over current methods. The Bayesian approach not only accounts for uncertainties in unknown parameters, but also allows for the incorporation of correlations among responses measured at different time points. The approach can be easily implemented through the use of Bayesian MCMC simulation. As pointed out by Novick et al. (2015), the approach can be extended in two aspects: one is to compare dissolution profiles which are

measured in a continuous temporal manner and the other is to quantify dissolution equivalences for two processes based on dissolution testing of multiple batches from each process. For the latter, one would need to use a hierarchical model which contains effects that represent batch-to-batch variations.

ACKNOWLEDGEMENT

The author would like to thank Ms. Katherine Giacoletti and Dr. Lorin Roskos for their helpful review of the article.

REFERENCES

1. Moore, J. W. Mathematical comparison of dissolution profiles. *Pharm. technol.* **1996**, *20*, 64–75.
2. FDA. Guidance for industry: immediate release solid oral dosage forms. scale-up and postapproval changes: chemistry, manufacturing and controls, *in vitro* dissolution testing and *in vivo* bioequivalence documentation. U.S. Department of Health and Human Services, FDA. Center of Drug Evaluation and Research (CDER): Rockville, MD, 1995. <http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm070636.pdf>.
3. FDA. Guidance for industry: extended release oral dosage forms: development, evaluation, and application of *in vitro*/*in vivo* correlations. U.S. Department of Health and Human Services, FDA. Center of Drug Evaluation and Research (CDER): Rockville, MD, 1997a. <http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm0702.pdf>.
4. FDA. Guidance for industry: dissolution testing of immediate release solid oral dosage forms. U.S. Department of Health and Human Services, FDA. Center of Drug Evaluation and Research (CDER): Rockville, MD, 1997b. <http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm070237.pdf>.
5. FDA. Guidance for industry: SUPAC-MR: modified release solid oral dosage forms. Scale-up and postapproval changes: chemistry, manufacturing and controls; *in vitro* dissolution testing and *in vivo* bioequivalence documentation. U.S. Department of Health and Human Services, FDA. Center of Drug Evaluation and Research (CDER): Rockville, MD, 1997c. <http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm070640.pdf>.
6. FDA. Guidance for industry: waiver of *in vivo* bioavailability and bioequivalence studies for immediate-release solid oral dosage forms based on biopharmaceutics classification system. U.S. Department of Health and Human Services, FDA. Center of Drug Evaluation and Research (CDER): Rockville, MD, 2000. <http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm070246.pdf>.
7. FDA. Guidance for industry: bioavailability and bioequivalence studies for orally administered drug products – general considerations, 2003.

Table 8 Results of Bayesian analysis of the simulated example

Parameter	Prior Distribution		
	Jeffrey's	SIW	AR(1)
F_2 (lower 95% CI)	52.2 (49.9)	52.3 (54.4)	52.3 (50.1)
$\delta(\delta)$ (upper 95% CI)	14.9 (16.7)	14.9 (16.5)	14.9 (16.6)
$\Pr(F_2 > 50)$	0.94	0.97	0.96
$\Pr(d(p) < 15)$	0.53	0.54	0.53
$\Pr(F_2 > 50, \delta(p) < 15)$	0.53	0.54	0.53

8. EMA (2008). Guideline on the Investigation of Bioequivalence; CPMP/ EWP/QWP/1401/98 Rev. 1; Committee for Medicinal Products for Human Use (CHMP), European Medicines Agency: London, 2008.
9. Shah, V.P., Tsong, Y., and Sathe, P. "In Vitro dissolution profile comparison—Statistics and analysis of the similarity factor, f_2 ," Pharm. Res. **1998**, *16* (6), 889–896.
10. Ma, M.C., Wang, B.B.C., Liu, J.-P. "Assessment of similarity between dissolution profiles", J. Biopharm. Stat. **2000**, *10* (2), 229–249.
11. Rescigno. Bioequivalence. Pharm. Res. **1992**, *9* (7), 925–928.
12. Tsong, Y., Hammerstrom, T., Sathe, P., Shah, V. P. Statistical assessment of mean differences between two dissolution data sets. Drug Inform. J. **1996**, *30* (4):1105–1112.
13. Saranadasa, H., Krishnamoorthy, K. A multivariate test for similarity of two dissolution profiles. J. Biopharm. Stat. **2005**, *15* (2), 265–278.
14. Chow, S.C. and Ki, F.Y.C. Statistical comparison between dissolution profiles of drug products. J. Biopharm. Stat. **1997**, *7*, 241–258.
15. Gibaldi, M.; Feldman, S. Establishment of sink conditions in dissolution rate determinations. Theoretical considerations and application to nondisintegrating dosage forms. J. Pharm. Sci. **1967**, *56* (10), 1238–1242. doi: 10.1002/jps.2600561005.
16. Wagner, J. G. Interpretation of percent dissolved–time plots derived from in vitro testing of conventional tablets and capsules. J. Pharm. Sci. **1969**, *58* (10), 1253–1257. doi: 10.1002/jps.2600581021.
17. Higuchi, T. Rate of release of medicaments from ointment bases containing drugs in suspension. J. Pharm. Sci. **1961**, *50* (10), 874–875. doi: 10.1002/jps.2600501018.
18. Higuchi, T. Mechanism of sustained-action medication. Theoretical analysis of rate of release of solid drugs dispersed in solid matrices. J. Pharm. Sci. **1963**, *52* (12), 1145–1149. doi: 10.1002/jps.2600521210.
19. Hixson, A. W.; Crowell, J. H. Dependence of Reaction Velocity upon Surface and Agitation. I-Theoretical Consideration. Ind. Eng. Chem. **1931**, *23* (8), 923–931. doi: 10.1021/ie50260a018.
20. Langenbucher, F. Letters to the Editor: Linearization of dissolution rate curves by the Weibull distribution. J. Pharm. Pharmacol. **1972**, *24* (12), 979–981. doi: 10.1111/j.2042-7158.1972.tb08930.x.
21. Tsong, Y.; Sathe, P. M.; Shah, V. P. In Vitro Dissolution Profile Comparison. In *Encyclopedia of Biopharmaceutical Statistics*, 2nd Ed.; Chow, S.-C., Ed.; CRC Press: Boca Raton, FL, 2003, 456–462. doi: 10.1201/b14760-68.
22. Polli, J. E.; Rekhi, G. S.; Augsburger L. L.; Shah, V. P. Methods to Compare Dissolution Profiles and a Rationale for Wide Dissolution Specifications for Metoprolol Tartrate Tablets. J. Pharm. Sci. **1997**, *86* (6), 690–700. doi: 10.1021/js960473x.
23. Adams, E.; De Maesschalck, R.; De Spiegeleer, B.; Vander Heyden, Y.; Smeyers-Verbeke, J.; Massart, D. L. Evaluation of dissolution profiles using principal component analysis. Int. J. Pharm. **2001**, *212* (1), 41–53. doi: 10.1016/S0378-5173(00)00581-0.
24. Adams, E.; Coomans, D.; Smeyers-Verbeke, J.; Massart, D. L. Non-linear mixed effects models for the evaluation of dissolution profiles. Int. J. Pharm. **2002**, *240* (1–2), 37–53. doi: 10.1016/S0378-5173(02)00127-8.
25. Amidon G.L., Lennernäs H., Shah V.P., Crison JR. A theoretical basis for a biopharmaceutical drug classification: the correlation of in vitro drug product dissolution and in vivo bioavailability". Pharm. Res. **1995**, *12* (3), 413–20.
26. Diaz, D.A., Colgan, S.T., Langer, S.S., Bandi, N.T., Likar, M.D., and Van Alstine, L. Dissolution similarity requirements: how similar or dissimilar are the global regulatory expectations? AAPS J. **2016**, *18* (1), 15–22.
27. Eaton, M. L.; Muirhead, R. J.; Steeno, G. S. Aspects of the Dissolution Profile Testing Problem. Biopharm. Rep. **2003**, *11* (2), 2–7.
28. LeBlond, D., Altan, S., Novick, S., Peterson, J., Shen, Y., Yang, H. In vitro dissolution curve comparisons: A critique of current practice. J. Biopharm. Stat. February **2016**, 14–23.
29. Chow, S.-C and Shao, J. *Statistics in Drug Research: Methodologies and Recent Development*. Marcel Dekker, Inc.: New York, 2002.
30. Novick, Shen, Y., Yang, H., Peterson, J., LeBlond, D., and Alton, S. Dissolution curve comparisons through the F_2 parameter, a Bayesian extension of the f_2 statistic. J. Biopharm. Stat. **2015**, *25*, 351–371.
31. SAS Institute Inc. (2008). SAS/STAT® 9.2 User's Guide. SAS Institute: Cary, NC.
32. Thomas, A., O'Hara, B., Ligges, U., Sturtz, S. (2006). Making BUGS open. R News *6*, 12–17.
33. Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, March 20–22, Vienna, Austria. ISSN 1609-395X. Available at: <http://mcmc-jags>.
34. Huang, A., Wand, M. P. Simple marginally noninformative prior distributions for covariance matrices. Bayesian Anal. **2013**, *8* (2), 439–452.
35. Gelman, A., Hill, J. *Data Analysis Using Regression and Multilevel Hierarchical Models*. Cambridge University Press: Cambridge, 2006.

Dissolution Profiles Comparison: Model-Independent Approaches

Yanbing Zheng and Lanju Zhang

Data and Statistical Sciences, AbbVie Inc., North Chicago, Illinois, U.S.A.

Abstract

Dissolution profiles comparison has become an important regulatory requirement for drug manufacturing process changes and has received intensive scientific interests in pharmaceutical industry. In this entry, we review the model-independent approaches based on multivariate statistical distance tests to test the similarity between dissolution profiles and evaluate their operating characteristics through a simulation study.

Keywords: Dissolution profile; Equivalence; Model-independent approaches; Multivariate statistical distance; Similarity.

INTRODUCTION

Dissolution profile of a drug product describes its release pattern under certain specified conditions. Compared to a single-point dissolution test, dissolution profile characterizes the drug product more completely. Comparison between dissolution profiles is performed when a change takes place in, for instance, scale of manufacture, manufacturing site, component and composition, equipment, and process.^[1] Dissolution profiles comparison is important in understanding the effect of manufacturing process changes on drug quality.

In general, the methods for dissolution profiles comparison can be categorized into three groups: 1) model-dependent methods; 2) model-independent methods based on a similarity factor; and 3) model-independent methods using multivariate statistical distance (MSD) tests.^[2] A model-dependent approach models dissolution profiles through mathematical functions and considers the time variable as a predictor to the amount (percentage) dissolved. For example, Langenbucher^[3] proposed to use Weibull distribution to characterize the dissolution rate curve, which is flexible in describing a wide variety of dissolution profile shapes. Alternatively, other mathematical models such as the first-order kinetic model, Hixson–Crowell model, Higuchi model, probit function, exponential function, Gumpertz function, and logistic function have been studied in literature to fit dissolution data (refer, e.g., literatures^[4–7]).

The U.S. Food and Drug Administration (FDA) guidance for industry on dissolution testing of immediate release solid oral dosage forms^[8] provides a model-independent approach using a similarity factor (f_2), which was proposed by Moore and Flanner.^[9] The similarity factor f_2 is defined as:

$$f_2 = 50 \times \log_{10} \left\{ \left[1 + (1/T) \sum_{t=1}^T (R_t - T_t)^2 \right]^{-0.5} \times 100 \right\} \quad (1)$$

where T is the number of dissolution sample time points and R_t and T_t are the mean percent dissolved at each time point, t , for the reference and test dissolution profiles, respectively. FDA guidance suggests using only one measurement after 85% dissolution of both products. When the two profiles are identical, $f_2 = 100$. An average difference of 10% at all measured time points results in an f_2 value very close to 50. FDA has suggested an f_2 value between 50 and 100 to indicate similarity between two dissolution profiles.

The factor f_2 measures the closeness between two profiles with emphasis on the larger difference among all time points and is referred to as a similarity factor. The f_2 approach is the simplest approach for dissolution profiles comparison. However, this criterion only takes into account the sample mean difference at sampled time points without incorporating the variation information in the dissolution data. As a result, the FDA guidance recommends that, to implement the f_2 approach, the percent coefficient of variation at the earlier time points should not be more than 20%, and at other time points should not be more than 10%. A multivariate model-independent procedure is suggested when one of these two requirements fail. LeBlond et al.^[2] gives a detailed discussion of the deficiencies of the f_2 approach.

In this entry, we focus on model-independent MSD test approaches. These approaches rely on the assumption of multivariate normality of amount (%) of drug release at sampled time points and consider different distance measurements to quantify the similarity between

the two dissolution profiles. In section “Model-Independent MSD Methods for Dissolution Profile Comparison,” we review three model-independent MSD approaches—two Hotelling’s T^2 -based approaches developed by Tsong et al.^[10] and Saranadasa,^[11] respectively, and a two one-sided test (TOST)-based approach developed by Saranadasa and Krishnamoorthy.^[12] In section “Simulation Study,” we perform simulation studies to compare the performance of the three methods under different scenarios, followed by section “Discussion and Conclusions.”

MODEL-INDEPENDENT MSD METHODS FOR DISSOLUTION PROFILE COMPARISON

Hotelling’s T^2 -Based Approaches

Suppose $x_{ij} \sim N_p(\mu_i, \Sigma)$, where i denotes the population number, $i = 1, 2$, j is the replicate number, $j = 1, 2, \dots, n_i$, and p is the number of time points that are measured. Note that x_{1j} and x_{2j} are p -dimensional random vectors with the same covariance matrix, Σ . The sample mean vector and covariance matrix for each profile are denoted as $\bar{x}_i$ and S_i , respectively. The pooled covariance matrix, an estimate of Σ , is:

$$S_{\text{pooled}} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \quad (2)$$

The null hypothesis based on Hotelling’s T^2 is $H_0: \mu_1 - \mu_2 = \mu$. The $(1 - \alpha)100\%$ confidence region of the true mean difference μ can be derived from:

$$T^2 = (\bar{x}_1 - \bar{x}_2 - \mu)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{\text{pooled}} \right]^{-1} (\bar{x}_1 - \bar{x}_2 - \mu) \leq \frac{(n - 2)p}{(n - p - 1)} F_{p, n-p-1}(\alpha) \quad (3)$$

where $F_{p, n-p-1}(\alpha)$ is $(1 - \alpha)100\%$ percentile of the central F distribution with p and $n - p - 1$ degrees of freedom with $n = n_1 + n_2$.

Tsong et al.^[10] proposed to use the Mahalanobis distance^[13] to measure the similarity of two dissolution profiles:

$$D_M = \sqrt{(\bar{x}_1 - \bar{x}_2)' S_{\text{pooled}}^{-1} (\bar{x}_1 - \bar{x}_2)} \quad (4)$$

Using the Lagrange multipliers approach, they calculated the $(1 - \alpha)100\%$ confidence limit of the true distance:

$$D_M^\mu = \sqrt{\mu' S_{\text{pooled}}^{-1} \mu} \quad (5)$$

based on the confidence region of μ (Eq. 3) and compared the confidence limit of true D_M^μ with the tolerance limit

$\sqrt{\theta^2 J_p' S_{\text{pooled}}^{-1} J_p}$ for equivalence determination, where θ is the difference specified as the global similar limit and J_p is a vector of 1s. Tsong et al.^[10] suggested to specify the similarity limit taking into account the variation of dissolution data.

Alternatively, Saranadasa^[11] assumed the parallelism of the two dissolution profiles and then derived the maximum mean difference d , for which the two dissolution profiles can be declared equivalent with $(1 - \alpha)100\%$ confidence. Based on the confidence region (Eq. 3), d is the solution to the equation:

$$T^2 = (\bar{x}_1 - \bar{x}_2 - dJ_p)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{\text{pooled}} \right]^{-1} (\bar{x}_1 - \bar{x}_2 - dJ_p) = \frac{(n - 2)p}{(n - p - 1)} F_{p, n-p-1}(\alpha) \quad (6)$$

The closed expression for d can be obtained as follows. By solving:

$$T_0^2 = (\bar{x}_1 - \bar{x}_2)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{\text{pooled}} \right]^{-1} (\bar{x}_1 - \bar{x}_2) = \frac{(n - 2)p}{(n - p - 1)} F_{p, n-p-1, \lambda}(\alpha) \quad (7)$$

for λ , the non-centrality parameter, the maximum mean distance can be calculated as:

$$d = \left[\frac{\lambda n(n - 2)}{(n - p - 3)n_1 n_2 \text{sum}(S_{\text{pooled}}^{-1})} \right]^{\frac{1}{2}} \quad (8)$$

where $\text{sum}(S_{\text{pooled}}^{-1})$ is the sum of the elements of S_{pooled}^{-1} . If,

$$T_0^2 < \frac{(n - 2)p}{(n - p - 1)} F_{p, n-p-1}(\alpha) \quad (9)$$

then d is defined as 0. When the estimated maximum mean distance, d , is less than a prespecified threshold, d_0 , the two dissolution profiles are deemed equivalent. The threshold, d_0 , depends on the confidence level. With a 90% confidence level, Saranadasa^[11] suggests $d_0 = 6\%$.

Saranadasa and Krishnamoorthy’s (SK) Method

The SK multivariate method used to determine the equivalence of two dissolution profiles is described in the study by Saranadasa and Krishnamoorthy.^[12] Similar to Saranadasa,^[11] the SK method assumes the parallelism of two dissolution profiles. Let $\bar{x}_d = \bar{x}_1 - \bar{x}_2$ follow a multivariate normal distribution:

$$\bar{x}_d \sim N_p \left(\delta J_p, \Sigma \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right) \quad (10)$$

where δ is a constant. Consider the hypotheses:

$$H_0: \delta > \delta_0 \text{ or } \delta < -\delta_0 \text{ versus } H_A: -\delta_0 \leq \delta \leq \delta_0 \quad (11)$$

If H_A is true, then the dissolution profiles are deemed equivalent. Following the description in the study by Saranadasa and Krishnamoorthy,^[12] δ can be estimated as:

$$\hat{\delta} = \frac{J_p' S_p^{-1} \bar{x}_d}{J_p' S_p^{-1} J_p} \quad (12)$$

H_0 is rejected if:

$$P_1 = P\left(t_{n-p-1} > \frac{\hat{\delta} + \delta_0}{SE}\right) < \alpha \text{ and} \\ P_2 = P\left(t_{n-p-1} < \frac{\hat{\delta} - \delta_0}{SE}\right) < \alpha \quad (13)$$

or

$$p \text{ value} = \max\{P_1, P_2\} < \alpha \quad (14)$$

where SE is the standard error of $\hat{\delta}$ and is calculated as described in the study by Saranadasa and Krishnamoorthy.^[12] This test is known as a TOST.^[14,15] This α level test is equivalent to calculating a $(1 - 2\alpha)100\%$ confidence interval for δ and checking that the interval is contained in $(-\delta_0, \delta_0)$.

Saranadasa and Krishnamoorthy^[12] recommended using $\delta_0 = 10\%$ with $\alpha = 0.05$. If $\delta = 10\%$, then the similarity factor f_2 is approximately 50%, which is the threshold recommended by the FDA guidance^[8] for the comparison procedure.

SIMULATION STUDY

A simulation study is conducted to evaluate the performance of the three model-independent MDS approaches. We assume the dissolution profile of a drug multivariate distribution tablet measured at four time points ($p = 4$) follows a multivariate normal distribution with a covariance matrix following the compound symmetry heterogeneous structure:

$$\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \cdots & \rho\sigma_1\sigma_p \\ \vdots & \ddots & \vdots \\ \rho\sigma_1\sigma_p & \rho\sigma_2\sigma_p \cdots & \sigma_p^2 \end{bmatrix} \quad (15)$$

We consider the cases where $\rho = 0$ and $\rho = 0.5$, $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (10, 10, 5, 5)$, $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (5, 5, 5, 5)$

and $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (2, 2, 2, 2)$. Dissolution profiles for 12 tablets ($n_1 = n_2 = 12$) are generated, respectively, for a reference batch and a test batch. Regarding the selection of similarity limit, we use $\theta = 10\%$ as suggested in the study by Tsong et al.,^[10] $d_0 = 6\%$ as suggested in the study by Saranadasa^[11] and $\delta_0 = 10\%$ as suggested in the study by Saranadasa and Krishnamoorthy.^[12]

We evaluate the performance of MDS approaches with/without the parallelism assumption of two dissolution profiles satisfied. For the cases where the parallelism assumption is not satisfied, we consider the scenarios with parallelism assumption marginally satisfied $\mu_1 - \mu_2 = \mu = (10, 10, 3, 3)$, $\mu_1 - \mu_2 = \mu = (20, 10, 3, 3)$, and parallelism assumption completely failed $\mu_1 - \mu_2 = \mu = (40, 20, 10, 3)$. A total of 1000 simulation runs were conducted under each simulation setting and the probability of claiming equivalence for each method was reported.

Figs. 1–3 show the simulation results when the assumption of parallelism between two dissolution profiles is satisfied. From the figures, we can see that the type I error rate of the SK method is roughly controlled at 0.05 while the f_2 method is very liberal. The Hotelling's T^2 -based method proposed in the study by Tsong et al.^[10] is relatively conservative especially when the variance of dissolution data is large. Another Hotelling's T^2 -based method proposed in the study by Saranadasa^[11] is liberal when the dissolution data have large variance and relatively conservative when the variance is small.

Tables 1–4 show the simulation when the assumption of parallelism between two dissolution profiles is not satisfied. When the two dissolution profiles are partially dissimilar or dissimilar [$\mu = (20, 10, 3, 3)$ or $\mu = (40, 20, 10, 3)$ in the simulation study], the f_2 factor and the approach proposed by Tsong et al.^[10] have good performance on not claiming equivalence. However, the approach proposed in the study

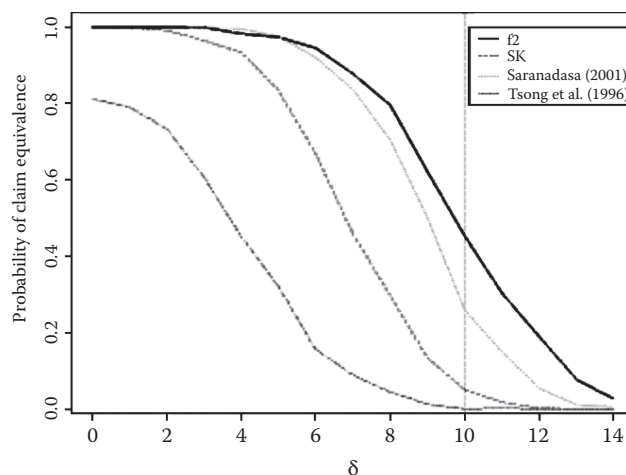


Fig. 1 The probability of claiming equivalence for the simulation setting with constant mean difference between dissolution profiles ($\mu_1 - \mu_2 = \mu = \delta J_4$), $\sigma_1, \sigma_2, \sigma_3, \sigma_4 = (10, 10, 5, 5)$, and $\rho = 0.5$

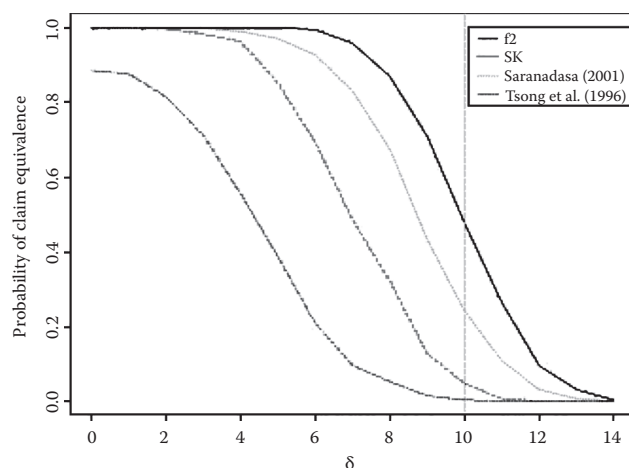


Fig. 2 The probability of claiming equivalence for the simulation setting with constant mean difference between dissolution profiles ($\mu_1 - \mu_2 = \mu = \delta J_4$), $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (5, 5, 5, 5)$, and $\rho = 0.5$

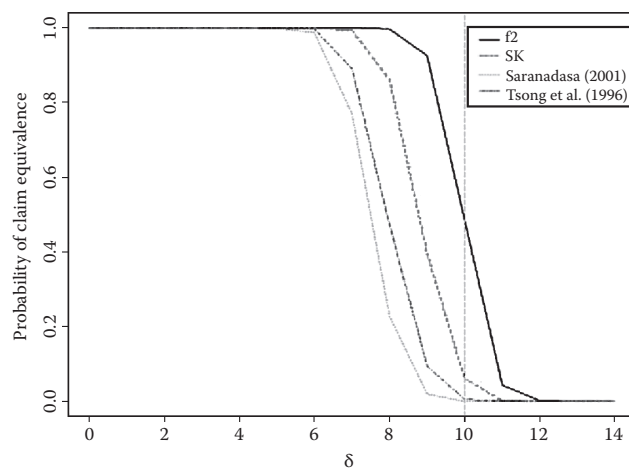


Fig. 3 The probability of claiming equivalence for the simulation setting with constant mean difference between dissolution profiles ($\mu_1 - \mu_2 = \mu = \delta J_4$), $\sigma_1, \sigma_2, \sigma_3, \sigma_4 = (2, 2, 2, 2)$, and $\rho = 0.5$

Table 1 Probability of claiming equivalence for the simulation setting $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (10, 10, 5, 5)$ and $\rho = 0$

Method	Mean difference = (10, 10, 3, 3)	Mean difference = (20, 10, 3, 3)	Mean difference = (40, 20, 10, 3)
f_2	0.865	0.186	0
Tsong et al., ^[10] $\theta = 10\%$	0.601	0.201	0
Saranadasa, ^[11] $d_0 = 6\%$	0.971	0.67	0
SK, $\delta_0 = 10\%$	0.964	0.771	0.024

Table 2 Probability of claiming equivalence for the simulation setting $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (10, 10, 5, 5)$ and $\rho = 0.5$

Method	Mean difference = (10,10,3,3)	Mean difference = (20,10,3,3)	Mean difference = (40,20,10,3)
f_2	0.825	0.247	0
Tsong et al., ^[10] $\theta = 10\%$	0.359	0.054	0
Saranadasa, ^[11] $d_0 = 6\%$	0.951	0.514	0.001
SK, $\delta_0 = 10\%$	0.958	0.895	0.302

Table 3 Probability of claiming equivalence for the simulation setting $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (5, 5, 5, 5)$ and $\rho = 0$

Method	Mean difference = (10, 10, 3, 3)	Mean difference = (20, 10, 3, 3)	Mean difference = (40, 20, 10, 3)
f_2	0.984	0.076	0
Tsong et al., ^[10] $\theta = 10\%$	0.349	0.009	0
Saranadasa, ^[11] $d_0 = 6\%$	0.648	0.023	0
SK, $\delta_0 = 10\%$	0.782	0.164	0

Table 4 Probability of claiming equivalence for the simulation setting $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (2, 2, 2, 2)$ and $\rho = 0$

Method	Mean difference = (10, 10, 3, 3)	Mean difference = (20, 10, 3, 3)	Mean difference = (40, 20, 10, 3)
f_2	1	0	0
Tsong et al., ^[10] $\theta = 10\%$	0.904	0.049	0
Saranadasa, ^[11] $d_0 = 6\%$	0.507	0.002	0
SK, $\delta_0 = 10\%$	0.968	0.142	0

by Tsong et al.^[10] could be very conservative when the variance of dissolution data is large. The SK method and the approach in the study by Saranadasa^[11] could incorrectly result in high probability of claiming equivalence especially when there is relatively high variation in dissolution data and high correlation between time points. Relatively, the f_2 approach has good power of claiming equivalence and good type I error control.

DISCUSSION AND CONCLUSIONS

The model-independent approaches for dissolution profile comparison using MSD tests developed in the studies by

Tsong et al.,^[10] Saranadasa,^[11] and Saranadasa and Krishnamoorthy,^[12] take into account the mean difference between dissolution profiles while adjusting for the covariance structure of dissolution data. They are all equivalence test approaches with prespecified tolerance limits. But the literature implementations of the MSD test methods have been controversial.^[2]

According to the simulation study, when the parallelism assumption is satisfied, the SK method has relatively good power of claiming equivalence with type I error under control. However, this method fails when the parallelism assumption is not satisfied. The performance of the method in the study by Saranadasa^[11] is highly affected by the covariance structure of dissolution data. The method proposed in the study by Tsong et al.^[10] is relatively conservative. The f_2 method recommended by FDA has good power to determine similarity but it could be very liberal when the two dissolution profiles are parallel. The performance of these methods is not satisfying and more research is needed for better approaches.

REFERENCES

1. Shah, V.P.; Tsong, Y.; Sathe, P.; Liu, J.-P. In vitro dissolution profile comparison—statistics and analysis of the similarity factor, f_2 . *Pharm. Res.* **1998**, *15* (6), 889–896.
2. LeBlond, D.; Atlan, S.; Novick, S.; Peterson, J.; Shen, Y.; Yang, H. In vitro dissolution curve comparisons: A critique of current practice. *Dissolut. Technol.* **2016**, *23* (1), 14–23.
3. Langenbucher, F. Letters to the editor: Linearization of dissolution rate curves by the Weibull distribution. *J. Pharm. Pharmacol.* **1972**, *24* (12), 979–981.
4. Sathe, P.M.; Tsong, Y.; Shah, V.P. In-vitro dissolution profile comparison: Statistics and analysis, model dependent approach. *Pharm. Res.* **1996**, *13* (12), 1799–1803.
5. Polli, J.E.; Rekhi, G.S.; Augsburger, L.L.; Shah, V.P. Methods to compare dissolution profiles and a rationale for wide dissolution specifications for metoprolol tartrate tablets. *J. Pharm. Sci.* **1997**, *86* (6), 690–700.
6. Tsong, Y.; Sathe, P.M.; Shah, V.P. In vitro dissolution profile comparison. In *Encyclopedia of Biopharmaceutical Statistics*. 2nd Ed.; Chow, S.-C., Ed.; CRC Press: Boca Raton, FL, 2003; 456–462.
7. Yuksel, N.; Kanik, A.E.; Baykara, T. Comparison of in vitro dissolution profiles by ANOVA-based, model-dependent and -independent methods. *Int. J. Pharm.* **2000**, *209* (1–2), 57–67.
8. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER). *Guidance for Industry, Dissolution Testing of Immediate Release Solid Oral Dosage Forms*. U.S. Government Printing Office: Washington, DC, 1997.
9. Moore, J.W.; Flanner, H.H. Mathematical comparison of curves with an emphasis on in-vitro dissolution profiles. *Pharm. Technol.* **1996**, *20*, 64–74.
10. Tsong, Y.; Hammerstrom, T.; Sathe, P.; Shah, V.P. Statistical assessment of mean differences between two dissolution data sets. *Drug Inf. J.* **1996**, *30*, 1105–1112.
11. Saranadasa, H. Defining the similarity of dissolution profiles through Hotelling's T^2 statistic. *Pharm. Technol.* **2001**, *25*, 46–54.
12. Saranadasa, H.; Krishnamoorthy, K. A multivariate test for similarity of two dissolution profiles. *J. Biopharm. Stat.* **2005**, *15* (2), 265–278.
13. Mahalanobis, P.C. On the generalised distance in statistics. *Proc. Natl. Inst. Sci. India.* **1936**, *2*, 49–55.
14. Schuirmann, D.J. On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics.* **1981**, *37*, 617.
15. Schuirmann, D.J. Comparison of two one-sided procedures and power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biop.* **1987**, *15* (6), 657–680.

Dosage Units Using Small and Large Sample Sizes: Uniformity

Meiyu Shen and Yi Tsong

Center for Drug Evaluation and Research, U.S. Food and Drug Administration,
Silver Spring, Maryland, U.S.A.

Abstract

The purpose of uniformity of dosage unit test is to determine the degree of uniformity in the amount of drug substance among dosage units in a batch. Content uniformity test is based on the assay of the individual content of drug substance(s) in a number of dosage units. Content uniformity test is required by the regulatory authorities such as US Food and Drug Administration to confirm that the unit dose of a drug product is consistent with the label claim. Although USP <905> is not intended for batch releases as it is clearly declared in the general notices and requirements of the USP pharmacopeia, many pharmaceutical companies still propose to implement USP <905> as their batch release specification. We present the details of USP <905> and propose a two one-sided parametric tolerance interval for small samples. We also present the details of parametric and nonparametric methods by European Pharmacopoeia 8.1 and propose a parametric tolerance interval method for large samples. Furthermore, we discuss the performances of these acceptance sampling plans.

Keywords: Compendia test; Content uniformity; Counting test; Tolerance interval; Uniformity of dosage units.

INTRODUCTION

The purpose of the uniformity of dosage unit (UDU) test is to quantify the variability in the amount of drug substance between dosage units in a batch and to compare that variability to a corresponding specification.^[1,2] In this chapter, the main interest is to discuss the compendia UDU test and to develop a parametric UDU test for batch release. Dosage units are defined as dosage forms containing a single dose or a part of a dose of drug substance in each unit. In most cases, the UDU test measures the amount of drug substance contained within the unit (e.g., tablets and capsules). In some cases UDU measures the amount of drug delivered via a dose metering device component (e.g., metered dose inhalers abbreviated as MDIs). In these cases, the test is referred to as uniformity of delivered dose (UDD). In a few instances, both the UDU and UDD tests may apply to the same product. For example, in a dry powder inhaler (DPI) that uses individual capsules or blisters containing the formulation to be inhaled. In such cases, UDU testing is needed to control for the variability in the individual capsules or blisters, and UDD testing is needed to control for the delivered dose ex-device used for inhalation.

Content uniformity test is based on the assay of the individual content of drug substance(s) in a number of dosage units. Content uniformity test is required by the regulatory authorities such as US Food and Drug Administration to confirm that the unit dose of a drug product is consistent with the label claim. ICH guideline Q4B, Annex 6^[3] recommends that a Pharmacopeia procedure is used to assess the uniformity of dosage units. For example, the United States Pharmacopoeia (USP) publishes USP <905>.^[1] USP <905> is used in the United States, the European Union, and Japan as the harmonized content uniformity test based on tolerance interval and indifference zone concepts. Although USP <905> is not intended for batch releases as it is clearly declared in the general notices and requirements of the USP pharmacopeia, many pharmaceutical companies still propose to implement USP <905> as their batch release specification. Since USP <905> uses an indifference zone concept (M), USP <905> is biased in favor of off-target products. Here M is defined by the sample mean of the tested n dosage units, $\bar{X}$, as: $M = 98.5\%$ if $\bar{X} < 98.5\%$, $M = 101.5\%$ if $\bar{X} > 101.5\%$, and $M = \bar{X}$ otherwise. To address this issue, Tsong and Shen^[4] propose a parametric tolerance interval test (PTIT) to test a two-sided specification that is equivalent to the test of two one-sided hypotheses. Testing against a lower specification is to assure that the drug product is not under-dosed for the sake of efficacy. On the other hand, testing against an upper

This article represents only the authors' opinion and not necessarily the official position of the US Food and Drug Administration.

specification is to assure that the drug product is not overdosed for the sake of safety.

With the development of analytic technology such as near-infrared spectroscopy, testing on large sample sizes becomes feasible. As a compendia test, the USP <905>, a harmonized content uniformity test, has been developed for a small fixed sample size, and it provides no guidance on how to conduct the test at different sample sizes. There have been several proposals in the literature for content uniformity using large sample sizes. In 2006, Sandell, Vukovinsky, Diener, Hofer, Pazdan, and Timmermans^[5] proposed a one-tiered nonparametric test which counted the number of tablets with content outside (85%, 115%) LC. The batch complies if no more than c units are outside (85%, 115%) LC. They also provided values of c for a selection of potential sample sizes. This proposed counting test has operating characteristic (OC) curves intersecting with the OC curve of the USP <905> around 45% probability of accepting a batch with the batch mean of 100% LC in^[5] (see Figure 3). This counting test has been proposed as a batch-release test when a large number of dosage units are tested.

In 2010, Bergum and Vukovinsky^[6] proposed a modified version of the counting test by Sandell et al.^[5] They propose to set the integer part of $0.03 \times n$ equal to c^* , which is the acceptance limit for the maximum number of tablets outside (85%, 115%) LC, here n is the number of dosage units tested (For example, with $n = 260$, c^* is 7). Thus, a batch complies if the number of tablets outside (85%, 115%) LC is no more than c^* .

In 2012, the Council of Europe^[7] published two options for uniformity of dosage units using large sample sizes in European Pharmacopoeia 7.7. Option 1 is a parametric two-sided tolerance interval based method modified with an indifference zone and counting units outside of (0.75M, 1.25M). Recall M is a function of data defined above. Option 2 is a nonparametric counting method with an additional indifference zone concept. Option 2 in European Pharmacopoeia 7.7 was intended to revise the original proposal of large sample UDU test whose OC curve intersects with OC curve of USP <905> around 45% passing probability.^[8] However, mathematical and statistical methodologies to derive the revised Option 2 were not discussed in.^[8] It would result in an unanswered question how to determine $c1$ and $c2$ for sample sizes not included in Table 2.9.47.2.^[7]

In April 2014, the Council of Europe^[9] published the corrected version of EU Option 2 in which the indifference zone is removed for uniformity of dosage units using large sample sizes in European Pharmacopoeia 8.1. With Option 2, a batch is accepted if no more than $c1$ units lie outside (0.85T, 1.15T) and no more than $c2$ units lie outside (0.75T, 1.25T), where T is defined as 100. The counts $c1$ and $c2$ for the selected sample sizes are given in Table 2.9.47-2 of European Pharmacopoeia 8.1. European Pharmacopoeia 8.1 provides no explanation how the counts

were derived. But the counts $c2$ are almost identical to Option 1.

Note that the approaches described earlier provide some assurance that a batch passing the release test may comply with USP <905>, the content uniformity sampling test using small number of dosage units.

In 2014, Shen, Tsong, and Dong^[10] studied the statistical properties of the large sample tests for content uniformity. They propose a large sample acceptance sampling method based on parametric two one-sided tolerance intervals. In particular, this proposed method is designed to have its OC curve for any given sample size intersect with the OC curve of the harmonized USP <905> at the acceptance probability of 90% for a batch whose individual tablets follow normality with the mean of 100% LC. They denote this proposed method as PTIT_matchUSP90 method. They compare the acceptance probabilities of PTIT_matchUSP90 method with those of the two procedures recommended in European Pharmacopoeia 7.7 when the unit dose is assumed to be a continuous random variable, e.g., normal or mixture of normal distributions. They also show that with two one-sided tolerance intervals criteria, the proposed test can accept or reject a batch based on the percentage outside either 85% or 115% LC regardless the distribution is normal or not. Such statistical property assures high acceptance probability of samples from the same batch when repeatedly tested with USP <905>, a small sample size test, since OC curve (acceptance probability versus standard deviation) of PTIT_matchUSP90 for any sample size intersects with OC curve of USP <905> at 90% of acceptance probability when the batch mean is on the target (100% LC). OC curves of others^[5,6] intersect with OC curve of USP <905> at <90% (e.g., 45%) of acceptance probability provide low assurance that samples in a batch pass compendia standard when repeatedly subjected to USP<905> UDU acceptance sampling plan.

The structure of this chapter is arranged as follows. In Section 2, we present the details of USP <905> and PTIT of Tsong and Shen^[4] for small samples. We also present the details of EU Option 1 and EU Option 2 (the counting method) proposed in European Pharmacopoeia 8.1 and the PTIT_matchUSP90 method proposed by Shen, Tsong, and Dong for large samples^[10] since European Pharmacopoeia 8.1 is most updated version so far. In Section 3, we discuss the performance of USP <905> and PTIT for small samples and the performance of PTIT_matchUSP90 method, EU Option 1, and the EU Option 2.

ACCEPTANCE SAMPLING PLANS

In this section, we are going to present the details of USP <905> UDU acceptance sampling plan and PTIT method for dose content uniformity using small sample sizes and then present the details of EU Option 1, EU Option 2, and

PTIT_matchUSP90 for UDU acceptance sampling plans using large sample sizes. Small sample sizes are referred to 10 samples at Stage 1 and 30 samples at Stage 2, which are required in USP <905>. Large sample sizes are referred to at least hundreds of samples as in EU Option 1 and EU Option 2.

Acceptance Sampling Plans for Small Samples

USP <905> UDU Acceptance Sampling Plan

The USP <905> consists of a tolerance interval test and two indifference zones. Let $\bar{X}$ and s be respectively the sample mean and sample variability of a random sample. The 1st indifference zone is for samples with mean value $\bar{X} \leq 98.5\%$ and the 2nd indifference zone is for sample with mean value $\bar{X} \geq 101.5\%$. For the 1st indifference zone, 98.5% will be used instead of $\bar{X}$ for the tolerance limit calculation. For the 2nd indifference zone, 101.5% will be used instead of $\bar{X}$ for tolerance interval calculation. The USP <905> can be described as follow.

At the 1st Tier, collect a random sample of 10 tablets from the batch and assay these 10 tablets. Let $M = 98.5\%$ if $\bar{X} < 98.5\%$, $M = 101.5\%$ if $\bar{X} > 101.5\%$, and $M = \bar{X}$ otherwise. The batch compiles if $M - 15\% < \bar{X} - 2.4 \cdot s < \bar{X} + 2.4 \cdot s < M + 15\%$ and all 10 measurements are within 75–125% M . Otherwise go to the 2nd tier.

At the 2nd Tier, assay the remaining 20 tablets and use the total 30 measurements to update $\bar{X}$ and s . Let $M = 98.5\%$ if $\bar{X} < 98.5\%$, $M = 101.5\%$ if $\bar{X} > 101.5\%$, and $M = \bar{X}$ otherwise. The batch compiles if $M - 15\% < \bar{X} - 2.0 \cdot s < \bar{X} + 2.0 \cdot s < M + 15\%$ and all 30 measurements are within 75–125% M .

PTIT

The objective of a PTIT with 95% confidence and 87.5% coverage for UDU acceptance sampling test is to ensure with 95% confidence that the percentages of tablets below 85% and above 115% LC are both less than 6.25% of the batch. The PTIT acceptance sampling plan is proposed to ensure that the percentage of over-filled tablets and the percentage of under-filled tablets are both below a specific limit.^[4,11] The PTIT acceptance sampling plan consists of two one-sided tests. The first one-sided test is applied to ensure that the percentage of over-filled tablets is below 6.25%. The test is carried out by comparing the upper limit of a left one-sided 95% confidence bound and 93.75% coverage tolerance interval with 115%. Recall that $\bar{X}$ and s be respectively the sample mean and sample variability of a random sample. The one-sided 95% confidence level tolerance interval $(-\infty, \bar{X} + Ks)$ with 93.75% coverage probability is calculated by solving for K in the following

$$\text{equation: } \Pr \left\{ \int_{-\infty}^{\bar{X} + Ks} f(x : \mu, \sigma) dx \geq (1 + 87.5\%)/2 \right\} = 0.95,$$

where μ and σ are the population mean and the standard deviation of random variable X , respectively, and $f(x : \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. K is often called the one-sided tolerance factor. Faulkenberry and Daley^[12] showed that $K\sqrt{n}$ is the 95% percentile of a noncentral t distribution with $n - 1$ degrees of freedom and a noncentrality parameter, $-Z_{93.75\%}\sqrt{n}$.

When $\bar{X} + Ks < 115\%$ LC, it ensures that the percentage of over-filled tablets is less than 6.25%. Similarly, when $\bar{X} + Ks > 85\%$ LC, it ensures that the percentage of under-filled tablets is below 6.25%.

The two-stage PTIT acceptance sampling plan can be described as follows. In the first tier, assay a random sample of 10 tablets from a batch and calculate $\bar{X}$ and s . Accept the batch if $85\% < \bar{X} - K_1s < \bar{X} + K_1s < 115\%$, where $K_1\sqrt{n} = t(n - 1, -\sqrt{n}Z_{(1+P)/2}, 1 - \alpha_1)$, where $t(n - 1, -\sqrt{n}Z_{(1+P)/2}, 1 - \alpha_1)$ is the $100 \cdot (1 - \alpha_1)^{\text{th}}$ percentile of a noncentral t distribution with $n - 1$ degrees of freedom and the noncentral parameter $-\sqrt{n}Z_{(1+P)/2}$, P is the coverage within (85, 115)% LC, e.g., $P = 0.875$. Note: α is the overall type 1 error rate; α_1 is the type 1 error rate for first tier when we control the nominal value α of 2-tier sequential test, e.g., at 0.05.^[13]

In the second tier, assay a random sample of an additional 20 tablets from the batch and use the total 30 measurements to update $\bar{X}$ and s . Accept the whole batch if $85\% < \bar{X} - K_2s < \bar{X} + K_2s < 115\%$, where $K_2\sqrt{n} = t(n - 1, -\sqrt{n}Z_{(1+P)/2}, 1 - \alpha_2)$, $100 \cdot (1 - \alpha_2)^{\text{th}}$ percentile of a noncentral t distribution with $n - 1$ degrees of freedom, and the noncentral parameter $-\sqrt{n}Z_{(1+P)/2}$. Note α_2 is the type 1 error rate for the second tier when we control the overall type 1 error rate of the two-tier sequential test as α , e.g., 0.05.

Acceptance Sampling Plans for Large Samples

Although there are several versions of the counting method^[5–9], we will focus on EU Option 2 in European Pharmacopoeia 8.1 since all counting methods share certain statistical properties as EU Option 2.

EU Option 1

EU Option 1 is based on two-sided tolerance interval modified with an indifference zone. For normally distributed variable, the two-sided tolerance interval with $1 - \alpha$ confidence level and P coverage is commonly represented in the form $(\bar{X} - K^*s, \bar{X} + K^*s)$. Here, K^* can be numerically calculated by solving the equation below.

$$\Pr \left\{ \int_{\bar{X} - K^*s}^{\bar{X} + K^*s} f(x : \mu, \sigma) dx \geq P \right\} = 1 - \alpha$$

The acceptance sampling test is implemented as follows:

With a sample of n (≥ 100) units, we calculate the acceptance value (AV) using the following formula:

$$|M - \bar{X}| + K^*s$$

Here, K^* is sample size-dependent tolerance factor and is defined in Table 2.9.47-1 of European pharmacopoeia 8.1; M is defined by the value of sample mean $\bar{X}$, such that $M = 98.5\%$ if $\bar{X} < 98.5\%$; $M = 101.5\%$ if $\bar{X} > 101.5\%$; and $M = \bar{X}$ otherwise.

The following criteria are applied, unless otherwise specified.

The requirements for dosage form uniformity are met if:

1. the acceptance value (AV) is less than or equal to $L1$ ($= 15.0$ unless other specified), and
2. in the calculation of acceptance value (AV) under content uniformity or under mass variation, the number of individual dosage units outside $(1 \pm L2 \times 0.01)M$, where $L2 = 25$ unless otherwise specified, is less than or equal to $c2$, as defined for a given sample size n in Table 2.9.47-1 of European pharmacopoeia 8.1.

EU Option 2

EU Option 2 is a nonparametric acceptance sampling test defined as follows:

With a sample of n (≥ 100) units, count the number of individual dosage units with a content outside $(1 \pm L1 \times 0.01)100$ LC and the number of individual dosage units with a content outside $(1 \pm L2 \times 0.01)100$ LC. Here $L1 = 15$ and $L2 = 25$. A batch passes the test for the uniformity of dosage units if

1. The number of individual dosage units outside $(1 \pm L1 \times 0.01)100$ LC is less than or equal to $c1$; and
2. The number of individual dosage units outside $(1 \pm L2 \times 0.01)100$ LC is less than or equal to $c2$.

Here, $c1$ and $c2$ are listed in Table 2.9.47-2 of European pharmacopoeia 8.1.

PTIT_matchUSP90

Again recall that $\bar{X}$ be the sample mean and s the sample standard deviation of contents of n units. Unlike USP <905> with a two-tier procedure, PTIT_matchUSP90 is a single tier content uniformity test with large sample size of n . A batch passes the PTIT_matchUSP90 test if $(\bar{X} - K(n)s, \bar{X} + K(n)s) \in (85\%, 115\%)LC$, where $K(n)$ is the tolerance factor for the PTIT_matchUSP90 sampling test using n units and is derived as follows.

The PTIT_matchUSP90 test consists of two one-sided hypothesis tests which ensure that the proportions of

over-filled and under-filled tablets are both under control. Let $p(n)^{[10]}$ be the desired proportion of the population within the interval (85%, 115%) LC for the PTIT_matchUSP90 sampling test using n units. In general, the acceptance probability of a batch with a given standard deviation when the mean is 100% LC increases with an increase of n for a fixed coverage within the interval (85%, 115%) LC. Since the PTIT_matchUSP90 test is designed to have 90% acceptance probability for any n at the standard deviation at which USP <905> has 90% acceptance probability when the batch mean is 100% LC, then we must increase $p(n)$ within the interval (85%, 115%) LC. The first one-sided test is to ensure that the proportion of over-filled tablets ($>115\%$ LC) is less than $(1 - p(n))/2$ and the second one-sided test is to ensure the proportion of under-filled tablets ($<85\%$ LC) is also less than $(1 - p(n))/2$. The first test is carried out by comparing the upper one-sided tolerance interval with 95% confidence level and $(1 + p(n))/2$ coverage with the upper limit 115% LC. More specifically, this one-sided tolerance interval is in the form of $(-\infty, \bar{X} + K(n)s)$, where the tolerance factor $K(n)^{[12]}$ is solved by the following equation for given values of μ and σ :

$$\Pr \left\{ \left[\int_{-\infty}^{\bar{X} + K(n)s} f(x; \mu, \sigma) dx \right] \geq \frac{1 + p(n)}{2} \right\} = 0.95$$

Where $f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. It can be easily showed that for normally distributed random variable, $K(n) = t_{n-1, 1-\alpha} \left(Z_{(1+p(n))/2} \sqrt{n} \right) / \sqrt{n}$ where $Z_{(1+p(n))/2}$ is the 100 $(1 + p(n))/2\%$ -th percentile of the standard normal distribution and $t_v(\gamma)$ denotes the non-central t -distribution with degree of freedom v and non-centrality parameter γ . If $\bar{X} + K(n)s < 115\%$ LC, it ensures that the percentage of over-filled tablets is less than $(1 - p(n))/2$. Similarly, if $\bar{X} - K(n)s > 85\%$ LC, the second test ensures that the percentage of under-filled tablets ($<85\%$ LC) is less than $(1 - p(n))/2$.

The algorithm to obtain $K(n)$ and $p(n)$ is briefly described below under normality and the detail was referred.^[14] First, $K(n)$ is determined for any n such that the probability of the event of $(\bar{X} - K(n)s, \bar{X} + K(n)s) \in (85\%, 115\%)LC$ is 90% at the standard deviation at which USP <905> has 90% acceptance probability when the batch mean is 100% LC. Second, $p(n)$ is determined by satisfying the equation $K(n) = t_{n-1, 1-\alpha} \left(Z_{(1+p(n))/2} \sqrt{n} \right) / \sqrt{n}$. Once we have $K(n)$, the acceptance probability of a batch with an individual table content as a random variable can be obtained from Monte Carlo simulations.

DISCUSSIONS

In Section 3.1, we discuss bias of USP <905> based on the comparison of acceptance probabilities between the

USP <905> acceptance sampling plan and PTIT acceptance sampling plan for small samples. In Section 3.2, we discuss the difference between EU Option 1 and PTIT_matchUSP90 for large samples. In Section 3.3, we discuss the difference between EU Option 2 and PTIT_matchUSP90 for large samples.

USP <905> and PTIT for Small Samples

The USP <905> acceptance sampling plan has often been modified through the years based on the capability of manufacturing without proper statistical support. As a result, the USP <905> often turns out to be both biased and incapable of inferring the characteristics of the population (the batch). Shen and Tsong^[11] showed the USP harmonized plan is biased in favor of batches with off-target means because of its indifference zones. The bias of the procedure is displayed with the plots of the acceptance probability against the coverage percentage (inside 85%–115% LC range) of the batch (See Figures 1–8, Shen and Tsong^[11]). The USP <905> has much higher acceptance probability for batches with off-target means (i.e., the batch mean $\neq$ 100% LC) than batches with on-target means when the coverage percentage within the interval 85%–115% is the same. In contrast, the acceptance probability curves of the PTIT do not differ much for batches with different target means. Furthermore, the PTIT has a slightly lower acceptance probability for the batches with off-target means than batches with on-target means when the coverage within the 85%–115% interval is greater than 0.91.

Hence, Shen and Tsong^[11] recommend that USP remove the indifference zone and replace the current USP harmonized content uniformity test with PTIT.

EU Option 1 and PTIT-matchUSP90 for Large Samples

As shown earlier, the tolerance factor K is free of batch mean and standard deviation. We calculated k values using $K = t_{n-1, 1-\alpha} \left(\frac{Z_{1+p(n)}}{2} \sqrt{n} \right) / \sqrt{n}$, where $\alpha = 0.05$ and $p(n)$ is coverage for sample size n . EU calculated K for various sample sizes n by adopting $\gamma = 0.84$ and $\beta = 0.91$:

$$K = Z_{\frac{1-\beta}{2}} \sqrt{\frac{n-1/n}{\chi_{\gamma, n-1}^2}}, \text{ where } Z_{\frac{1-\beta}{2}} \text{ is the critical value of the}$$

normal distribution and $\chi_{\gamma, n-1}^2$ is the critical value of the chi-square distribution with $n-1$ degrees of freedom.

From Table 1, it can be seen that the K values for PTIT_matchUSP90 are smaller than those for EU option 1 for each corresponding sample size.^[10] Hence it is expected that the acceptance probability of PTIT_matchUSP90 should be always higher than that of EU option 1 for a given population if EU option 1 uses a tolerance interval as the only decision rule. In fact, this is not always true because EU option 1 is two-sided tolerance interval

Table 1 Comparison of K values between PTIT_matchUSP90 and EU Option 1

Sample Size	PTIT_matchUSP90	Option 1
100	2.0475	2.15
1000	2.2321	2.27

method modified with an indifference zone and counting outside of $(0.75M, 1.25M)$.

Differences in acceptance probabilities between EU Option 1 and PTIT-matchUSP90 can be summarized as follows.^[10]

First, there is a large difference in acceptance probability among EU option 1 and PTIT_matchUSP90 when the batch mean is off-target. More specifically, for a batch with off-target mean with data either of a single normal distribution or of a mixture of two normal distributions, EU option 1 has much higher acceptance probability than PTIT_matchUSP90. For instance, the OC curve of PTIT_matchUSP90 is left to that of EU option 1 with 102% LC at any given large sample size, e.g., 100 and 1000, respectively as shown in Figures 2, 7, and 8.^[10] As the sample size increases, the differences in the OC curve between PTIT_matchUSP90 and EU option 1 both become larger. Moreover the acceptance probability of PTIT_matchUSP90 is almost zero for a batch with standard deviation 6.6%; however, the acceptance probability is about 27% for EU option 1. In addition, the counting number outside $(75\%M, 125\%M)$ LC criterion in EU option 1 has little impact on the acceptance probability under normality assumption. For instance, adding counting test in EU option 1 only reduces the acceptance probability by 0.23% for $\mu = 102$, $\sigma = 6$, and $n = 100$.

Second, there is no apparent difference in acceptance probability among EU Option 1 and PTIT_matchUSP90 for a batch with on-target mean (100% LC). According to our simulation study, EU option 1 has a slightly lower acceptance probability than PTIT_matchUSP90 at a given sample size and standard deviation when data are either taken from normal distribution or from a mixture of two normal distributions.

Third, there is bias in EU option 1. The acceptance probability of EU option 1 at mean of 102% LC is much larger than that of PTIT_matchUSP90 at mean of 102% LC under normality assumption. This contradicts the observation that the acceptance probability of EU option 1 at mean of 100% LC is smaller than that of PTIT_matchUSP90 at mean of 100% LC under normality assumption because K value in EU 1 is a slight larger than K value in PTIT_matchUSP90 for sample size of 1000. This contradiction is most likely resulted from indifference zone in EU1.

Fourth, it is also clearly shown (see Figure 9)^[10] that the difference in OC curves of EU option 1 between batches with mean of 102% and of 100% LC is much smaller than the difference in OC curves of PTIT_matchUSP90 between batches with mean of 102% and of 100% LC. This means that the PTIT_matchUSP90 has more power in discriminating off-target batches from on-target batches.

As discussed in,^[11] the indifference zone criterion used in USP harmonization method induce bias that results in favoring batches with off-target mean. The undesirable behavior of EU option 1 is also observed in.^[10] Hence, we recommend that European Pharmacopoeia investigate the impact of indifference zone and consider removal of the indifference zone criterion from EU option 1 and replacement of EU option 1 with two one-sided parametric tolerance intervals.

EU Option 2 and PTIT-matchUSP90 for large samples

It is well understood that content values may not be normally distributed and a non-parametric content uniformity acceptance sampling plan may need to be developed. However, when developing a new large sample non-parametric approach, we need to provide assurance of proper power to satisfy small sample compendia method USP <905> applied to samples at any time. Any method should not fail the basic requirement that samples on shelf should pass USP <905> at any time. For a lack of publication of mathematical or statistical derivation of the approach, we can only evaluate the EU Option 2 in European pharmacopoeia 8.1 through simulation studies.^[15] PTIT_matchUSP90 is developed to assure 90% acceptance probability at the standard deviation where USP <905> has 90% acceptance probability for batches with 100% LC mean. The following are the summaries of such evaluations.^[15]

Under normality, the acceptance probability of EU option 2 is slightly higher than that of PTIT_matchUSP90 for batches with a large variability. However, under normality, EU Option 2 is not sensitive to the off target mean since the acceptance probability of EU Option 2 is much larger than that of PTIT_matchUSP90 for batches with large variances. Furthermore, EU Option 2 is not sensitive to large variances and a mean shift under a mixture of two normal variables as the PTIT_matchUSP90 is.

Given that 97% of tablets have content values falling within the range from 85% LC to 115% LC, clearly EU Option 2 is not sensitive to a large variability in contents from a uniform distribution; on other hand, EU Option 2 is over-sensitive to a small percent of extreme observations. As a result, EU Option 2 provides low assurance to be accepted by USP <905> in the subsequent compendia testing. The EU Option 2 is not sensitive to the mean shift of the majority population (97%) from 100% LC to 90%

LC. In addition, the EU Option 2 is not sensitive to low assay values (about 90% LC). Together, it shows that it may be hard to declare the EU Option 2 is a good sampling test in the sense of providing high probability assurance of acceptance by the subsequent USP<905> compendia test. Certainly, the examples simulated are extremes for any batch with consistent good quality. However, they are useful to illustrate the potential problems of interpreting a variable acceptance sampling test (e.g., tolerance interval method) with an acceptance sampling by attribute test (e.g., counting test).

In conclusion, the counting method such as EU Option 2 performs insensitively for a batch with a large variability and an off target mean even under normality and a mixture of two normal variables, and performs over-sensitively for a batch with small percent of low extreme values. For content uniformity assessment, information such as variability can be lost when EU Option 2 is used.

RECCOMENDATION

Parametric two one-sided tolerance interval test is recommended for batch release based on small samples. PTIT_matchUSP90 is recommended for batch release based on large samples.

REFERENCES

1. The United States Pharmacopeia Convention, General Chapter <905> Uniformity of dosage units, United States Pharmacopeia (USP) 38: 675–679 (2015). Rockville, MD.
2. Food and Drug Administration, Guidance for Industry, Nasal Spray and Inhalation Solution, Suspension, and Spray Drug Products—Chemistry, Manufacturing, and Controls Documentation, 2002
3. Food and Drug Administration, Guidance for Industry, ICH Q4B Evaluation and Recommendation of Pharmacopoeial Texts for Use in the ICH Regions, Annex 6 Uniformity of Dosage Units General Chapter, June 2014.
4. Tsong Y., M. Shen. Parametric two-stage sequential quality assurance test of dose content uniformity. *Journal of Biopharmaceutical Statistics* **2007**, *17*, 143–157.
5. Sandell D., K. Vukovinsky, M. Diener, J. Hofer, J. Pazdan, J. Timmermans, Development of a content uniformity test suitable for large sample sizes. *Drug Information Journal* **2006**, *40* (3), 337–344.
6. Bergum J. and K.E. Vukovinsky, A proposed content-uniformity test for large sample sizes. *Pharmaceutical Technology* **2010**, *34* (11), 72–79.
7. Council of Europe, Uniformity of dosage units using large sample sizes. Chapter 2.9.47 of European Pharmacopoeia 7.7: 5142–5145, Renouf Pub Co Ltd; PhEur 7th edition (16 Oct 2012).
8. Holte Ø. and M. Horvat, Uniformity of dosage units using large sample sizes. *Pharmaceutical Science Technology* **2012**, *36* (10), 118–122.

9. Council of Europe, Uniformity of dosage units using large sample sizes. Chapter 2.9.47 of European Pharmacopoeia 8.1: 3669–3671, Renouf Pub Co Ltd; PhEur 8th edition (April 2014).
10. Shen, M.; Tsong, Y.; Dong, X. Statistical properties of large sample tests for dose content uniformity. *Therapeutic Innovation & Regulatory Science* **2014**, *48* (5), 613–622.
11. Shen, M.; Tsong, Y. Bias of the USP harmonized test for dose content uniformity, Stimuli to the Revision Process. *Pharmacopeial Forum* **2011**, *37*.
12. Faulkenberry, G.D.; Daley, J.C. Sample size for tolerance limits on a normal distribution. *Technometrics* **1970**, *12* (4), 813–821.
13. Jennison, C.; Turnbull, B.W. Group sequential methods with applications to clinical trials. CRC Press: Boca Raton, FL, 1999.
14. Tsong, Y.; Dong, X.; Shen, M.; Lostritto, R.T. Quality assurance test of delivered dose uniformity of multiple-dose inhaler and dry powder inhaler drug products. *Journal of Biopharmaceutical Statistics* **2015**, *25* (2), 328–338.
15. Shen, M.; Tsong, Y. Counting test and parametric two one-sided tolerance interval test for content uniformity using large sample sizes, *Journal of Biopharmaceutical Statistics*, submitted, **2016**.

Dose Finding Clinical Trials Using R

Naitee Ting

Biostatistics and Data Sciences, Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, Connecticut, U.S.A.

Ding-Geng (Din) Chen

Department of Biostatistics, Gillings School of Global Health, University of North Carolina, Chapel Hill, North Carolina, U.S.A.; Department of Statistics, University of Pretoria, Pretoria, South Africa; School of Social Work, University of North Carolina, Chapel Hill, North Carolina, U.S.A.

Abstract

In this entry, we illustrate the three approaches in analysis of dose-ranging trials which include the multiple comparison procedure (MCP), the modelling (MOD) approach, and the combined MCP-Mod approach. We illustrated these procedures with implementation in R with a real clinical trial data set.

Keywords: Clinical trials; Dose-finding; Multiple comparison; Family-wise error-rate (FWER); Dose-response modeling; Emax model; MCP-Mod.

INTRODUCTION

Traditionally, proof of concept (PoC) study and dose-ranging trials are conducted separately in a sequential fashion. A two-group PoC design involves a comparison between MTD (Maximally Tolerable Dose) of test drug and a placebo control, a comparison that is conducted first. Only after the concept is proven, then dose-ranging trials are designed to study the dose-response relationship. In recent years, more sponsors are combining these two steps into one – using one trial with multiple doses to serve both the PoC and dose-ranging purposes. The focus of this entry is to discuss data analysis of dose-ranging trials – either this dose-ranging trial is designed after PoC or in combination with the PoC objective.

In the analysis of dose-ranging clinical trials, there are typically three approaches commonly used to help determining the dose-response relationship: the first is a multiple comparison procedure (MCP) approach, the second is to use a dose-response modeling (Mod) approach and the third to combine both the MCP and Mod as a newer approach, which is commonly referred as “MCP-Mod” approach (Bretz et al. 2005^[1]; Pinheiro et al. 2006^[2]).

In the MCP approach, the dose is treated as a qualitative factor with a set of dose levels from the dosages available for testing in the clinical trial. In this approach, there are usually no assumptions about the underlying dose-response relationship and the statistical inferences to find doses of interest are restricted to this set of studied doses. In contrast to the MCP approach, the “Mod” approach typically assumes a parametric dose-response

relationship between the dose and the efficacy response in the trial. Then the doses are treated as a quantitative continuous variable that would allow more flexibility for dose finding. Therefore, the validity of this “Mod” approach would depend strongly on the assumed dose-response model.

Combining both the MCP and Mod approaches, Bretz et al. (2005^[1]) developed an MCP-Mod approach to unify the MCP and Mod approaches, which is to assume the existence of several candidate parametric dose-response models and use multiple comparison techniques to choose the one most likely to represent the true underlying dose-response curve. At the same time, the hybrid MCP-Mod approach preserves the family-wise error rate. The selected model is then used to provide inference on adequate doses. Based on this novel approach, Bretz and colleagues developed an R package (i.e. “MCP-Mod”) which is publicly available at <http://cran.r-project.org/web/packages/MCPMod/index.html> with detailed description in Bornkamp et al. (2009^[3]). This package later became part of the new package in “DoseFinding” which can be freely accessed at <https://cran.r-project.org/web/packages/DoseFinding/index.html>.

This entry is then organized as following sections. Section 2 provides the data set and some preliminary analyses. Section 3 presents analysis for making a PoC decision. The multiple comparison procedures (MCP) are introduced in Section 4 and the modeling (Mod) approach in Section 5. In Section 6, we introduce the combined MCP and Mod approach (MCP-Mod) and conclude the entry with a discussion in Section 7.

DATA AND PRELIMINARY ANALYSIS

Let's first describe the data. As reported in Bretz et al. (2005^[1]), this study was a randomized double-blind parallel group trial with a total of 100 patients being allocated to either placebo or one of four active doses coded respectively as 0.05, 0.20, 0.60, and 1, with $n = 20$ per group. The response variable was assumed to be normally distributed with larger values for a better outcome. The data is publicly available in R packages "MCPMod" and "DoseFinding" as a data frame named "biom" with 2 columns named "dose" for dose levels and "resp" for responses. There are 20 observations for each of the 5 dosages. The data are available from both R packages. The data distributions at each dose can be graphically illustrated in Figure 1 with a boxplot and it can be seen from Figure 1 that the response distributions are overlapped indicating large variations existed in this data. Also as seen from Figure 1, there is an observed monotonic increasing dose-response relationship with sample means of 0.345, 0.457, 0.810, 0.934 and 0.949, respectively.

ESTABLISHING POC WITH A TREND TEST

The PoC can be established based on the contrast test for statistical significance. Let μ_i be the population mean for dose group i ($i = 0, 1, \dots, k$), where μ_0 is the mean from the placebo group. For example, k is 4 corresponding to the "biom" data in Section 2 and the estimated sample means of μ_i ($i = 0, 1, 2, 3, 4$) are 0.345, 0.457, 0.810, 0.934 and 0.949, respectively.

To construct a contrast for PoC, let c_i be the corresponding contrast coefficients with the condition that $\sum_{i=0}^k c_i = 0$.

Then the null hypothesis (H_0) of no treatment effect versus the one-sided alternative hypothesis (H_A) of significant treatment effect (PoC) can be written as follows:

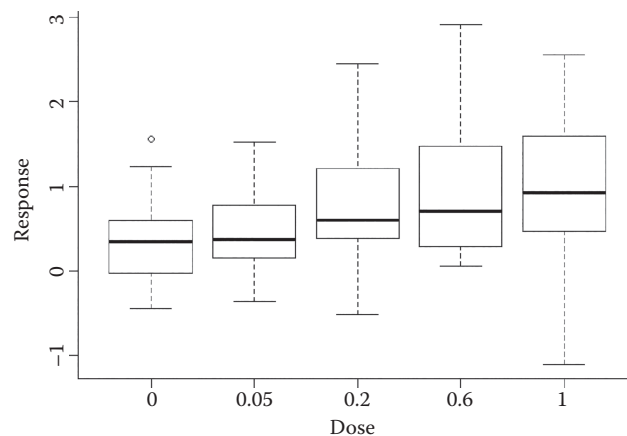


Figure 1 Boxplot for Dose-Response Distribution for dataframe "biom"

$$H_0 : L(\mu) = \sum_{i=0}^k c_i \mu_i \leq 0 \text{ vs. } H_A : L(\mu) = \sum_{i=0}^k c_i \mu_i = \delta > 0 \quad (1)$$

If at the design stage, the planned PoC step is based on a trend test, then, corresponding to "biom" data with $k = 4$, the trend contrast coefficients are $c = (c_0, c_1, c_2, c_3, c_4) = (-2, -1, 0, 1, 2)$, which can be found in Table 2 in Wang and Ting (2012^[4]), and the above hypotheses become as follows:

$$H_0 : L(\mu) = (-2)\mu_0 + (-1)\mu_1 + 0\mu_2 + 1\mu_3 + 2\mu_4 \leq 0, \text{ vs.}$$

$$H_A : L(\mu) = (-2)\mu_0 + (-1)\mu_1 + 0\mu_2 + 1\mu_3 + 2\mu_4 = \delta > 0.$$

The test statistic can then be constructed as a one-sided t-test as follows:

$$T = \frac{\widehat{L(\mu)}}{SE(\widehat{L(\mu)})} = \frac{\sum_{i=0}^k c_i \bar{x}_i}{SE\left(\sum_{i=0}^k c_i \bar{x}_i\right)} \quad (2)$$

where $\bar{x}_i$ is the sample mean of μ_i and SE denotes its standard error.

We can fit an analysis of variance (ANOVA) model to estimate the sample means and their associated standard errors using R function "lm" (i.e., linear model) to fit the ANOVA model with summary table as shown below:

Estimate	Std. Error	t value	Pr(> t)	
dose=0	0.3449	0.1593	2.165	0.0329
dose=0.05	0.4568	0.1593	2.867	0.0051
dose=0.2	0.8103	0.1593	5.087	1.83e-06
dose=0.6	0.9344	0.1593	5.866	6.47e-08
dose=1	0.9487	0.1593	5.956	4.35e-08

Residual standard error: 0.7124 on 95 degrees of freedom
Multiple R-squared: 0.5336, Adjusted R-squared: 0.509
F-statistic: 21.74 on 5 and 95 DF, p-value: 1.897e-14

We can see that the ANOVA model is statistically significant with p-value = 1.897e-14 and a $R^2 = 53.36\%$. The estimated means are labeled under the column "Estimate" for all the doses which are all statistically significant as shown for their associated p-values. The estimated standard deviation, also known as the residual standard error, is 0.7124. We can extract these values for PoC calculations.

Under the same study setting, the design could have pre-specified a PoC just like the traditional high dose against placebo comparison –

$$H_0 : L(\mu) = -\mu_0 + \mu_4 \leq 0, \text{ vs.}$$

$$H_0 : L(\mu) = -\mu_0 + \mu_4 > 0.$$

In this case, the t -test as given in (2) can still be applicable, where the contrast coefficients are now $-1, 0, 0, 0, 1$. The study result indicates a p -value of 0.00867, indicating the null hypothesis is rejected and the concept is considered proven.

The statistical test for PoC in equation (2) can then be easily calculated with these estimates and shown to be true. This established the PoC for these data and the next step is to identify the dose from the dose range where the statistically significant difference occurs with MCP, Mod, and MCP-Mod approaches.

MULTIPLE COMPARISON PROCEDURE (MCP) APPROACH

By nature of design, dose-response trials typically include multiple dose groups plus a control. Therefore a multiple comparison procedure should be utilized to adjust for multiplicity. For example, in this data, there are four doses of 0.05, 0.2, 0.6, 1 plus the placebo dose. There are many procedures in multiple comparison adjustment and we illustrate the most commonly used procedures, such as the Fisher's protected least significant difference (LSD), Bonferroni procedure, Dunnett's procedure, Holm's procedure, Hochberg, and gate-keeping procedures.

Fisher's Protected LSD (Fixed Sequence Test)

As a comparison, we first illustrate the naïve MCP procedure which was termed as Fisher's protected LSD (least significant difference). This procedure is also known as the "fixed sequence test" in Westfall and Krishen (2001^[5]) and in Bretz et al. (2005^[1]) to determine the dose using a 5% one-sided level. The first step is to perform an F -test to see if there is any difference among all treatment groups. Only after this test is significant, then to perform pairwise comparisons of each dose against placebo. This can be easily conducted using the R function " lm " for all pairwise comparison to placebo as summarized as follows:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.3449	0.1593	2.165	0.03287
Dose=0.05	0.1118	0.2253	0.497	0.62068
Dose=0.2	0.4654	0.2253	2.066	0.04155
Dose=0.6	0.5895	0.2253	2.617	0.01032
Dose=1	0.6038	0.2253	2.680	0.00867

Residual standard error: 0.7124 on 95 degrees of freedom

Multiple R-squared: 0.1153, Adjusted R-squared: 0.07808
F-statistic: 3.096 on 4 and 95 DF,
p-value: 0.01921

It can be seen that the overall model fitting is statistically significant with two-sided p -value = 0.019 from the F -test statistic of 3.096 with degrees of freedom of 4 and 95. The values in the column "Estimate" corresponding to "dose" 0.05, 0.2, 0.6 and 1 are the estimated mean difference to the placebo dose, that is, $\mu_i - \mu_0$, which are 0.1118, 0.4654, 0.5895 and 0.6038, respectively. Examining the corresponding p -values at the column "Pr(>|t|)", we can see that they are all statistically significant as two-sided t -test except the dose at 0.05. As a default setting in R function " lm ," the column "Pr(>|t|)" reported the 2-sided p -values which can be converted a one-sided p -value. The one-sided raw p -values (i.e. raw.pval) can then be calculated as follows:

Dose=0.05	dose=0.2	dose=0.6	dose=1
0.310339936	0.020774426	0.005160541	0.004334960

This reproduced the results at their Table 2 (page 743) from Bretz et al. (2005^[1]). Therefore, this fixed sequence test concluded that the top three doses (i.e. 0.2, 0.6 and 1) were statistically significantly different to placebo.

Bonferroni Correction

As a way to adjust for multiple comparisons, the classical Bonferroni correction (Bonferroni 1935^[6] 1936^[7]) adjusts the raw p -values, ensuring strong family-wise error-rate (FWER) control under arbitrary dependence of the raw p -values. It simply multiplies each raw p -value by the total number of hypotheses. We make use of the R package of "mutoss" which is maintained by the MUTOSS (multiple hypotheses testing in an open software system) team and can be downloaded from <http://mutoss.r-forge.r-project.org/>. This package is designed to unify the application and comparison of multiple testing procedure. Most of the MCP procedures are implemented in this package and we only need to call the corresponding function. The Bonferroni correction procedure can be done by calling the R function Bonferroni to the raw p -value where the "raw.pval" is the vector of unadjusted p -values for the pairwise comparison (i.e., $\mu_i - \mu_0$) which is from the ANOVA model fitting above. The output of this Bonferroni correction with the adjusted p -values and rejected at $\alpha = 0.05$ can be shown as follows:

dose=0.05	dose=0.2	dose=0.6	dose=1
adjPValues:			
1.00000000	0.08309770	0.02064217	0.01733984
Rejected:			
FALSE	FALSE	TRUE	TRUE

From this output, we can find the adjusted p -values associated with each comparison of the specific dose to placebo. With alpha (α) = 0.05, it also showed whether the comparison was rejected or not. Among the four pairwise multiple comparisons, the doses of 0.05 and 0.2 were not rejected, whereas the top two doses of 0.6 and 1 were rejected.

Dunnett's Test

Dunnett's test is also to compare each of the dose treatments to the placebo treatment. This test was developed in Dunnett^[8] and updated in Dunnett^[9]. The procedure takes advantage of positive correlations among pairwise comparisons. The fact that each dose is compared with the same placebo control makes all pairwise comparisons to be positively correlated. Dunnett's test was widely used in multiple comparison procedure for simultaneously comparing, by interval estimation or hypothesis testing, all active dose treatments with placebo when the outcome is sampled from a distribution where the normality assumption is reasonable. This test is designed to hold the familywise error rate at or below α when performing multiple comparisons of treatment group with control. To implement the Dunnett's test, we make use of the R package "DoseFinding". The "MCTtest" function in this package gave the Dunnett's contrast to compare all the dose treatments to the placebo and the test-statistic with the adjusted p -values for each dose treatment as follows:

Multiple Contrast Test		
	t-Stat	adj-p
Dose=1	2.680	0.0153
Dose=0.6	2.617	0.0178
Dose=0.2	2.066	0.0653
Dose=0.05	0.497	0.6033

It can be seen from the multiple contrast test that the doses of 0.05 and 0.2 were not statistically significant and the other top two doses of 0.6 and 1 were statistically significant (adjusted p -value less than the nominal alpha of 0.05).

Holm's Step-down Procedure

The Holm's step-down-procedure was developed in Holm^[10] and detailed in Huang and Hsu^[11]. The terms "step-down" or "step-up" are in reference to the absolute values of t -test. "Step-down" implies the procedure starts with the pairwise t -test with the largest absolute value, and then steps down to the smaller absolute value of t . It controls the family-wise error rate (FWER) and offers a simple test uniformly more powerful than the Bonferroni correction. The Holm's procedure is implemented in the R package "mutoss" with function "holm." The function

"holm" is called to adjust the raw p -values "raw.pval" to control FWER and the output is produced as follows:

dose=0.05	dose=0.2	dose=0.6	dose=1
adjPValues:			
0.31033994	0.04154885	0.01733984	0.01733984
Rejected:			
FALSE	TRUE	TRUE	TRUE

The Holm's test showed that the dose of 0.05 was not rejected. However, the doses of 0.2, 0.6 and 1 were rejected as each of them is statistically significantly different from placebo.

Hochberg Step-up Procedure

The Hochberg procedure (Hochberg^[12]) is based on marginal p -values. It controls the FWER in the strong sense under joint null distributions of the test statistics that satisfy Simes' inequality as summarized in Sarkar and Chang^[13]. Huang and Hsu^[11] compared this Hochberg procedure with Holm's procedure in Section 9.4.4 and concluded that the Hochberg procedure is more powerful than Holm's^[10] procedure, but the test statistics need to be independent or have a distribution with multivariate total positivity of order two or a scale mixture thereof for its validity. Whereas Holm's procedure is a step-down version of the Bonferroni test, Hochberg's is a step-up version of the Bonferroni test. Note that Holm's method is based on the Bonferroni inequality and is valid regardless of the joint distribution of the test statistics.

The Hochberg procedure is implemented in R function "hochberg" in R package "mutoss" which produces following output:

dose=0.05	dose=0.2	dose=0.6	dose=1
adjPValues:			
0.31033994	0.04154885	0.01548162	0.01548162
criticalValues:			
0.01250000	0.01666667	0.02500000	0.05000000
Rejected:			
FALSE	TRUE	TRUE	TRUE

With the same conclusion reached as the Holm's procedure, the Hochberg's procedure showed that the dose of 0.05 was not rejected and the doses of 0.2, 0.6 and 1 were rejected and hence statistically significantly different from placebo.

Gate-keeping Procedure

Gate-keeping procedures are proposed to test hierarchically ordered hypotheses from clinical trials. The hypotheses are grouped into ranked families, such that the hypotheses in higher order ranked families are tested first and the lower ranked families can be tested only if rejections achieved for higher ranked families. In other words,

the higher ranked families serve as gatekeepers for lower ranked ones.

In general, there are two types of gatekeeping procedures known as serial and parallel gate-keeping procedure. As discussed in Bauer et al.^[14] and Dmitrienko et al.^[15], the serial gatekeeping requires all hypotheses in one family be rejected before testing the next family, whereas parallel gatekeeping as discussed in Dmitrienko et al.^[16] only requires at least one hypothesis be significant in order to test the next one. The gatekeeping approaches were first proposed as closed tests as discussed in Marcus, Peritz and Gabriel^[17] by considering all possible configurations of the null hypotheses where in general, the number of computational steps grows exponentially with n (about 2^n calculations need to be performed to test n null hypotheses). Under the assumption of the nondecreasing dose response with respect to increasing doses, the multiple testing of individual doses could fit into the serial gate-keeping procedure by ordering the testing sequence from the highest to the lowest dose. A lower dose is tested only if all preceding higher doses are tested and rejected at level of significance (α , α). The testing stops once a hypothesis fails to be rejected.

This gate-keeping procedure can be easily implemented to start with the highest dose and then goes down to find the minimal dosage where dosage to placebo is statistically significant at this dosage, all higher doses should be significant which can be seen as follows:

Dose=0.05	dose=0.2	dose=0.6	dose=1
FALSE	TRUE	TRUE	TRUE

Same as the Holm's and Hochberg's procedures, the gate-keeping procedure showed that the dose of 0.05 was not rejected and the doses of 0.2, 0.6 and 1 were rejected, with each being statistically significantly different from placebo at the 0.05 level of significance.

MODELING APPROACH (MOD)

Different from the MCP approach discussed in the Section 4, the modeling approach (Mod) treats dose as a continuous variable and assumes a functional relationship between the response and the dose. Sometimes, the multiplicity issue is avoided as a result of the assumed functional relationship (e.g., non-decreasing dose–response relationship).

Dose-response Models

The dose-response models are used to describe certain functional relationship between the observed responses y_{ij} and the corresponding dose d_i , where i indicates the dose-level from 1 to k ($k = 5$ in the “biom” data) and j indicates the observations in each dose level from 1 to n_i ($n_i = 20$ in the “biom” data for all 5 dosages). This model can be mathematically expressed as follows:

$$y_{ij} = f(d_i, \theta) + \varepsilon_{ij}.$$

where ε_{ij} is the error term. The mean response $\mu_i = E(y_{ij})$ and the corresponding dose d_i can then be expressed as

$$\mu_i = f(d_i, \theta).$$

where $f(\cdot)$ is a linear or nonlinear function of dose d and the vector of modeling parameters θ . There are potentially many types of dose-response models, such as the linear model, linear in log-dose model, the exponential model, the quadratic model, the logistic model, the Emax model, the sigmoid Emax model and the beta model. These models are briefly summarized in Table 1 with their functional form and the associated model parameters.

Table 1 Commonly Used Dose-Response Models

Model	Functional form	Parameters
Linear	$E_0 + \delta d$	E_0, δ
Linear in log-dose (linlog)	$E_0 + \delta \log(d)$	E_0, δ
Exponential	$E_0 \exp(d/\delta)$	E_0, δ
Quadratic	$E_0 + \beta_1 d + \beta_2 d^2$	E_0, β_1, β_2
Logistic	$E_0 + \frac{E_{max}}{1 + \exp((ED_{50} - d)/\delta)}$	E_0, E_{max}, δ
Emax	$E_0 + \frac{E_{max} d}{ED_{50} + d}$	E_0, E_{max}, ED_{50}
Sigmoid Emax	$E_0 + \frac{E_{max} d^\lambda}{ED_{50}^\lambda + d^\lambda}$	$E_0, E_{max}, ED_{50}, \lambda$

The graphical presentation of these models can be shown in Figure 2.

In this presentation, we specified these models with some selected parameters and also we illustrated two Emax models and two logistic models. As seen from Figure 2, all these models except for the quadratic model are monotonic increasing as doses increase.

Among these models, a very popular dose-response model is the sigmoid Emax model as seen in Table 1. This sigmoid Emax model is a four-parameter model where E_0 is the effect at dose 0 (the placebo response); E_{max} is the maximum effect attributable to the drug; ED_{50} is the dose, which produces half of E_{max} ; and λ is the slope factor (Hill factor). In practice, the three-parameter Emax model is typically considered, assuming the Hill factor $\lambda = 1$. As discussed in Thomas^[18], the Emax model can fit very well for most of the actual dose-response data and we illustrate this Emax model in this entry for the Mod approach.

R Implementations

We fit the Emax dose-response model to the “biom” data using the R function “fitMod” as in follow R code: fitMod(dose, resp, data = biom, model = “emax”). With the

Emax model, the model fitting can be graphically illustrated in Figure 3. It can be seen from Figure 3 that the Emax model fitted the “biom” data very satisfactorily. In Figure 3, the dashed vertical lines at each observed dosage indicated the confidence intervals for the means at those observed dosages. The solid curve in the middle is the fitted Emax model, which is closed to the observed mean responses at each observed dosages. The confidence interval for the fitted Emax model can be seen from the upper and lower curved lines about the fitted Emax model.

It is important to note that the modelling approach is, by nature, an estimation process and hence p-values are not directly interpretable. In order to help readers who are used to p-values, these numbers are still presented below. We calculate the t-statistics and the associated p-values at dose-levels of 0.05, 0.2, 0.6 and 1 are:

Dose=0.05	dose=0.2	dose=0.6	dose=1
0.1216920183	0.0207664633	0.0007128031	0.0003555304

It can be seen that treatment effects start from dose 0.20 and indicates that the top three doses of 0.2, 0.6 and 1 are can be considered as different from placebo.

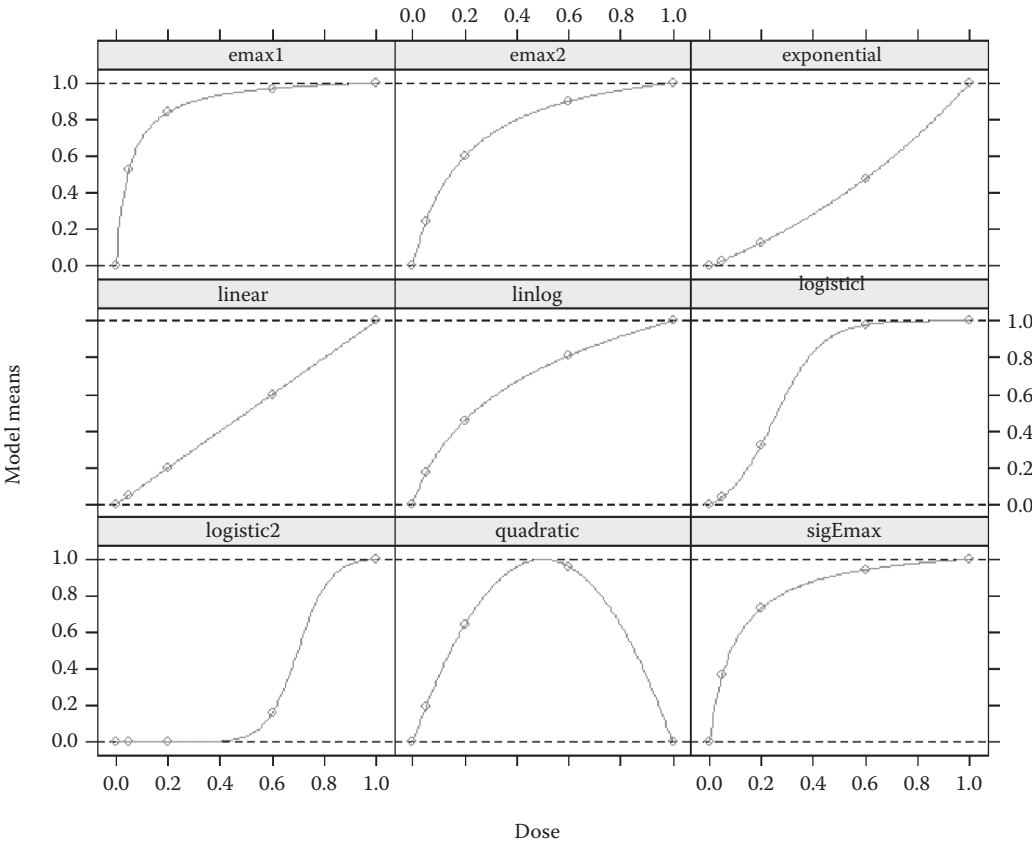


Figure 2 Shapes of dose-response models in Table 2

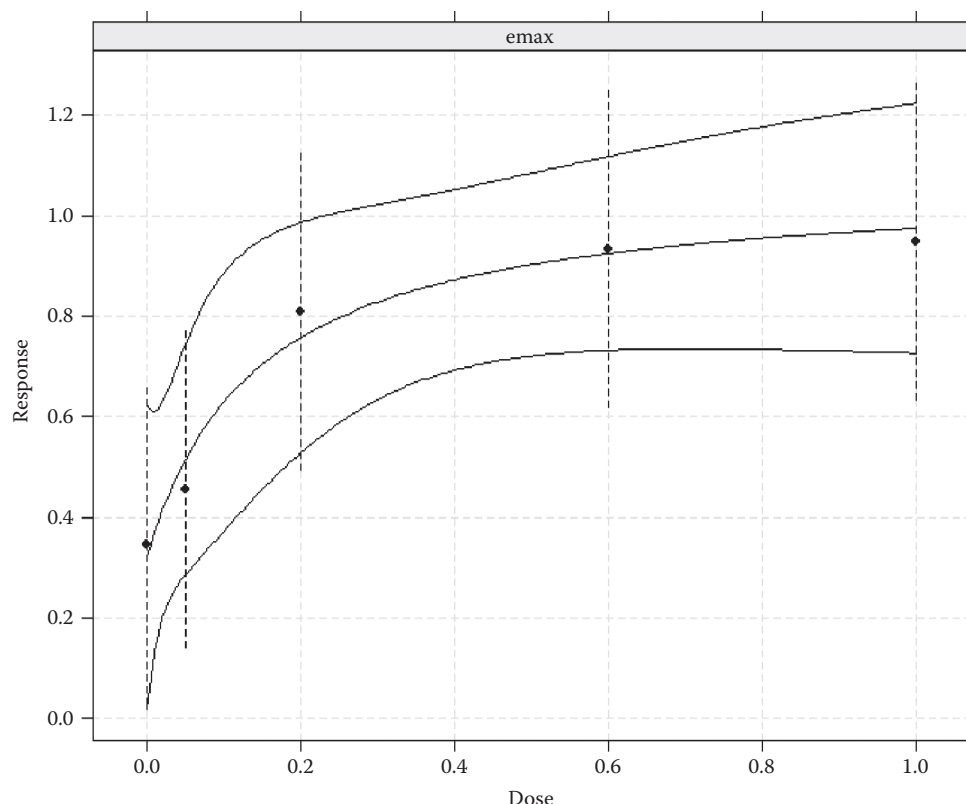


Figure 3 Model fitting to “biom” using Emax model where the curves as Upper 95% CI, Predicted, Lower 95% (from top to bottom)

MCP-MOD APPROACH

As the third approach in analyzing dose-response trials, the MCP-Mod method is to combine the MCP approach in Section 4 and Mod approach in Section 5 into one unified approach by first selecting the dose-response model using MCP to protect the FWER; then, the selected model is used to estimate the dose-response relationship (readers are encouraged to refer to the MCP-Mod entry of this Encyclopedia). This approach is flexible in the sense of model selection; on the other hand, it pays the price of splitting α at the model selection step.

This MCP-Mod approach was developed by Bretz et al.^[1] and they further developed an R package (i.e. “MCPMod”) (Bornkamp et al.^[3]) which is available at <http://cran.r-project.org/web/packages/MCPMod/index.html>. We do not intend to replicate the theory in this entry but instead concentrate on illustration of this package for real data analysis. We again make use of the “biom” data for this illustration. This illustration is largely described in Bornkamp et al.^[3] and this entry is a further application of this package.

Implementations in R package “MCPMod”

The first step to use MCPMod is to select a candidate set of models for approximating the underlying dose-response

relationship. The package gave a list of candidate models to be used, such as linear model, log-linear model, quadratic model, logistic model, exponential model, beta-model, Emax model, and sigmoid Emax model. In addition to these models, the users can also define their own non-linear models. The candidate models should be specified as a list with the list elements named according to the underlying dose-response model function, along with guesstimates or NULL for the model parameters. For illustration purpose, let’s specify the candidate model lists for the “biom” dose-response data as follows:

```
# Specify the candidate model list
> modlist = list(linear=NULL,
  linlog=NULL,
  betaMod = c(0.5, 1), logistic=c(0.25,
    0.09),
  quadratic=-1, exponential=1,
  emax=c(0.05,0.2), sigEmax=c(0.1,1))
```

We deliberately select the candidate model list, which is different from the models used in Bretz et al.^[1], to show flexibility of the MCPMod to perform in model fitting and model selection. Bretz et al.^[1] used linear model, log-linear model, two exponential models, two quadratic models and the Emax model, whereas we select the linear model,

log-linear model, beta model, logistic model, quadratic model, exponential model, Emax model and sigmoid Emax model, respectively, as seen in the above R code chunk.

We named the selected model list as an R object “modlist” and the numerical values are the guestimated parameter initial values. It can be seen that this candidate model list includes some monotonic increasing functions (such as the linear model, log-linear model, logistic model, exponential model, Emax model and sigmoid Emax model) as well as other shapes of models (such as beta model and quadratic model).

The main function in this package is also named as “MCPMod.” It provides all the functionalities of the full MCP-Mod approach and includes two key steps:

1. MCP-step to calculate the optimal contrasts, critical value, contrast test statistics and possibly p -values, and selection of the set of significant models,
2. Modeling step for model fitting, model selection/model averaging, and dose estimation.

We call this function to analyze the “biom” and the following simple R code chunk is all that is needed for MCP-Mod analysis:

```
# Fit the MCP-Mod
>fit.MCPMod = MCPMod(biom, modlist,
  alpha = 0.05, pVal = TRUE,
  selModel = "maxT", doseEst = "MED2",
  clinRel = 0.4, dePar = 0.05)
# Print the summary of model fitting
>summary(fit.MCPMod)
```

We now explain some of the most important arguments used for the MCPMod function. It can be seen from the above R code chunk that the first argument is the dose-response data set (i.e., biom in this case) followed by the candidate model list (i.e., modlist defined previously). The alpha is the level of significance for the multiple contrast test, specified as $\alpha = 0.05$. The pVal argument is to determine whether multiplicity adjusted p -value for the multiple contrast test should be calculated or not and we specified it as $pVal = TRUE$. The selModel argument is a dose-estimation criterion to determine how to select a dose-estimation model from the set of significant models (if there are any significant models). The argument doseEst is to determine the minimum effective dose (MinED) once the overall dose-response relationship is established from an adequate model fitting. The MinED is defined as the smallest dose that demonstrates a clinically relevant and a statistically significant effect (denoted by Δ), specified as “clinRel.” Four dose-estimation options can be selected. One option is called “ED” to estimate the dose that gives a pre-specified percentage of the maximum effect. Specifically, the target dose ED_p is defined as the smallest dose that gives a certain percentage p of the maximum effect

Δ observed in the dose range of $(d_1, d_k]$. The other three options are defined in Bretz et al. (2005) as “MED1,” “MED2” and “MED3” with default selection of “MED2.” The dePar argument is used to specify additional parameters for the four dose estimators. For ED-type estimators, it determines which effective dose is estimated with the default 0.5 to estimate ED50. For MED-type estimators, it determines the confidence level γ used in the estimator. The used confidence level is given by $1 - 2\gamma$ in corresponding to $1 - 2\gamma$. In the R code chunk, we specified doseEst = “MED2” with dePar = 0.05 and clinRel = 0.4 to use the MED2 estimator with $\gamma = 0.05$ and the clinical threshold $\Delta = 0.4$ to estimate the MED. This configuration means that we wish to determine the smallest dose such that the lower limit of the 95% confidence interval for the predicted response is greater than the predicted response for placebo and the point estimate is 0.4 above the predicted placebo level.

With all those settings, the summary of the MCPMod model fitting is given by print(fit.MCPMod) as follows:

Input parameters:

```
alpha = 0.05 (one-sided)
model selection: maxT
clinical relevance = 0.4
dose estimator: MED2 (gamma = 0.05)
```

Multiple Contrast Test:

	Tvalue	pValue
emax2	3.464	0.001
sigEmax	3.462	0.001
linlog	3.364	0.002
emax1	3.339	0.003
logistic	3.235	0.003
linear	2.972	0.007
exponential	2.752	0.012
betaMod	2.402	0.030
quadratic	1.850	0.093

Critical value: 2.156

Selected for dose estimation:

emax

Parameter estimates:

```
emax model:
  e0  eMax  ed50
0.322 0.746 0.142
```

Dose estimate

```
MED2,90%
0.17
```

where:

1. Input parameters for some information about important input parameters that were used for MCPMod modelling.
2. “Multiple Contrast Test” and “Critical value” are the contrast test statistics, the

multiplicity adjusted p -values from the multivariate t -distribution in Bretz et al.^[1], and the critical value associated with the candidate models specified in the modlist. It should be noted in this part that all contrast tests with an adjusted p -value less than 0.05 or equivalently with t -value > 2.18 (the critical value) can be declared as statistically significant, with FWER maintained at 5% level. Of these nine candidate models only the quadratic model is not statistically significant, which is then not used in the reference set.

3. “Selected for dose estimation” and “Parameter estimates” for the best fitted dose-response model from the candidate models and its parameter estimates. For these data, the best model is found to be the “Emax” model with estimated parameters as $E_0 = 0.322$, $E_{\max} = 0.746$ and $ED_{50} = 0.142$.
4. “Dose estimate” for the target dose estimate from this MCP-Mod modeling, which is 0.17 based on the settings in this modeling and close to the estimate from MoD approach at Section 4.

In summary for the MCP-Mod approach, the best model selected for dose estimation from the model list is the Emax model and the estimated significant dosage is 0.17.

DISCUSSION

In this entry, we used a real data set “biom” to illustrate the three approaches in analysis of dose-ranging trials, namely, the MCP approach, the Mod approach, and the combined MCP-Mod approach. We illustrated these procedures with implementation in R and the R program can be obtained from the authors upon request.

With the established PoC using a trend test (or the two group comparison) in Section 3, we introduced six MCP procedures in Section 4 to control the FWER to identify the dosage which are statistically significantly different from placebo. These six approaches are the Fisher’s protected LSD (fixed sequence test), the Bonferroni test, the Dunnett test, the Holm’s test, the Hochberg test, and gate-keeping test. Among the six MCP tests, the Bonferroni test and the Dunnett test identified that the top two doses (i.e., 0.6 and 1) were statistically significantly different from placebo and the rest of four tests (i.e., the fixed sequence test, the Holm’s test, the Hochberg test and gatekeeping test) identified the top three doses (i.e., 0.2, 0.6 and 1) were statistically significantly different from placebo.

The Mod approach in Section 5 used the Emax model to estimate the dose-response relationships. With the estimated dose-response Emax model, we further illustrated the estimation of doses of interest along with their confidence intervals. We also investigated other models, such as

the log-linear and the sigmoid Emax model (as an extension of Emax model to include another parameter) and found that both models gave the same conclusions (not shown here due to the space limitations).

The combined MCP-Mod approach in Section 6 selected the best model as the Emax model from a list of dose-response models that included the linear model, log-linear model, logistic model, exponential model, Emax model, the sigmoid Emax model, the beta model, and the quadratic model. With the best model being Emax, the statistically significant dose was identified as 0.17.

REFERENCES

1. Bretz, F.; Pinheiro, J.C.; Branson, M. Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* **2005**, *61* (3), 738–748.
2. Pinheiro, J.C.; Bretz, F.; Branson, M. Analysis of dose-response studies – modeling approaches, In: Ting, N. (Ed.). *Dose Finding in Drug Development*. Springer: New York, 2006, 146–171.
3. Bornkamp, B.; Pinheiro, J.C.; Bretz, F. MCPMod: An R package for the design and analysis of dose-finding studies. *Journal of Statistical Software* **2009**, *29* (7), 1–23.
4. Wang, X.; Ting, N. A proof-of-concept clinical trial design combined with dose-ranging exploration. *Pharmaceutical Statistics*, wileyonlinelibrary.com. DOI 10.1002/pst. 1525 (2012).
5. Westfall, P.; Krishen, A. Optimally weighted, fixed sequence and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference* **2012**, *99*, 25–40.
6. Bonferroni, C.E. Il calcolo delle assicurazioni su gruppi di teste. ‘In Studi in Onore del Professore Salvatore Ortu Carboni. Rome, Italy, **1935**, 13–60.
7. Bonferroni, C.E. Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze **1936**, *8*, 3–62.
8. Dunnett, C.W. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **1955**, *50*, 1096–1121.
9. Dunnett C.W. New tables for multiple comparisons with a control. *Biometrics* **1964**, *20*, 482–491.
10. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **1979**, *6* (2), 65–70.
11. Huang, Y.; Hsu, J. Hochberg’s step-up method: cutting corners off Holm’s step-down method. *Biometrika* **2007**, *94*, 965–975.
12. Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **1988**, *75* (4), 800–802.
13. Sarkar, S.K.; Chang, C.-K. The Simes method for multiple hypothesis testing with positively dependent test statistics. *J. Amer. Statist. Assoc.* **1997**, *92*, 1601–1608.
14. Bauer, P.; Röhm, J.; Maurer, W.; Hothorn, L. Testing strategies in multidose experiments including active control. *Statistics in Medicine* **1998**, *17* (18), 2133–2146.

15. Dmitrienko, A.; Tamhane, A.C.; Wang, X.; Chen, X. Step-wise gatekeeping procedures in clinical trial applications. *Biometrical Journal* **2006**, *48* (6), 984–991.
16. Dmitrienko, A.; Offen, W.; Westfall, P. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine* **2003**, *22* (15), 2387–2400.
17. Marcus, R.; Peritz, E.; Gabriel, K.R. On closed testing procedure with special reference to ordered analysis of variance. *Biometrika* **1976**, *63* (3), 655–660.
18. Thomas, N. Hypothesis testing and Bayesian estimation using a sigmoid Emax model applied to sparse dose–response designs. *Journal of Biopharmaceutical Statistics* **2006**, *16* (5), 657–677.

Dose Proportionality

C. Gordon Law

Pfizer Inc., Groton, Connecticut, U.S.A.

Abstract

In drug development, it is essential to manage a safe and effective dosing. This entry gives a brief overview of statistical models for assessing dose proportionality.

INTRODUCTION

In drug development, it is essential to manage a safe and effective dosing. While a complete pharmacokinetic (PK) profile for all doses is impossible to establish, prediction of PK effects in a certain dose range can easily be made if the compound possesses the property of dose proportionality (DP). This property, also called dose independence or linear pharmacokinetics, implies that rates of absorption, distribution, metabolism, and elimination remain constant over the dose range. With this property, plasma concentration, area under the plasma concentration–time curve (AUC), maximum plasma concentration ($C_{\max}$), and urinary recovery should increase in direct proportion to dose.

OVERVIEW

Fig. 1 depicts a typical pattern of a PK profile with DP. It is assumed that dose 1 < dose 2 < dose 3. The left panel of the figure shows that the concentration–time curve shifts upward as the dose increases. The shapes of the three concentration curves at the absorption and the excretion stages are nearly identical. As shown in the middle panel, if the concentrations are divided by the corresponding doses, the difference will be indistinguishable. In the right panel, the three AUC values almost fall on the same line when they are plotted against dose. Similarly, a linear relationship for $C_{\max}$ can be observed.

In case of repeated dosing, the concentration after multiple dosing can be well projected if the pharmacokinetics of the drug are time invariant and are dose independent. Based upon the superposition principle,^[1] the concentrations over time can be predicted by laying the profile of the next dose over the residual concentrations resulting from all previous doses. Rowland and Tozer^[1] describe the circumstances when DP cannot be established.

For the statistical methods to assess DP, Ezzet and Spiegelhalter^[2] and Hutchison and Keene^[3] provide excellent reviews. They proposed a power model to estimate

the degree of proportionality. Besides this approach, two popular models—linear regression and analysis of variance (ANOVA)—are also used to assess DP. All of these models will be reviewed and compared in the following section. Some limitations in these models will be examined. In the last section, the design aspects of the DP study will be discussed. Some interesting works by Ezzet and Spiegelhalter^[2] will be illustrated.

STATISTICAL MODELS FOR ASSESSING DOSE PROPORTIONALITY

Linear Regression

The classical approach to test DP is to fit a few regression models using the key PK parameters, denoted by Y , as the response and the dose level as the covariate. Because heterogeneity in variances is often found in DP studies, the models are usually fitted with weights equal to the reciprocal of the dose level or with a transformation for Y . The model for initial assessment can be:

$$Y = a + b(\text{dose}) + c(\text{dose})^2$$

When the coefficient for the quadratic term c is not significantly different from zero, dose linearity is claimed. The next model of interest is:

$$Y = a + b(\text{dose})$$

If the intercept a is not significant, then DP can be declared. In other words, it requires dose linearity and zero intercept to assess DP.

The main drawback of the regression approach is the lack of a measure that can quantify DP. If the measure existed, we could have derived a confidence interval describing the uncertainty of the estimate. Also, when the quadratic term is significant or when the intercept is not close to zero, we are unable to estimate the magnitude of departure from DP.

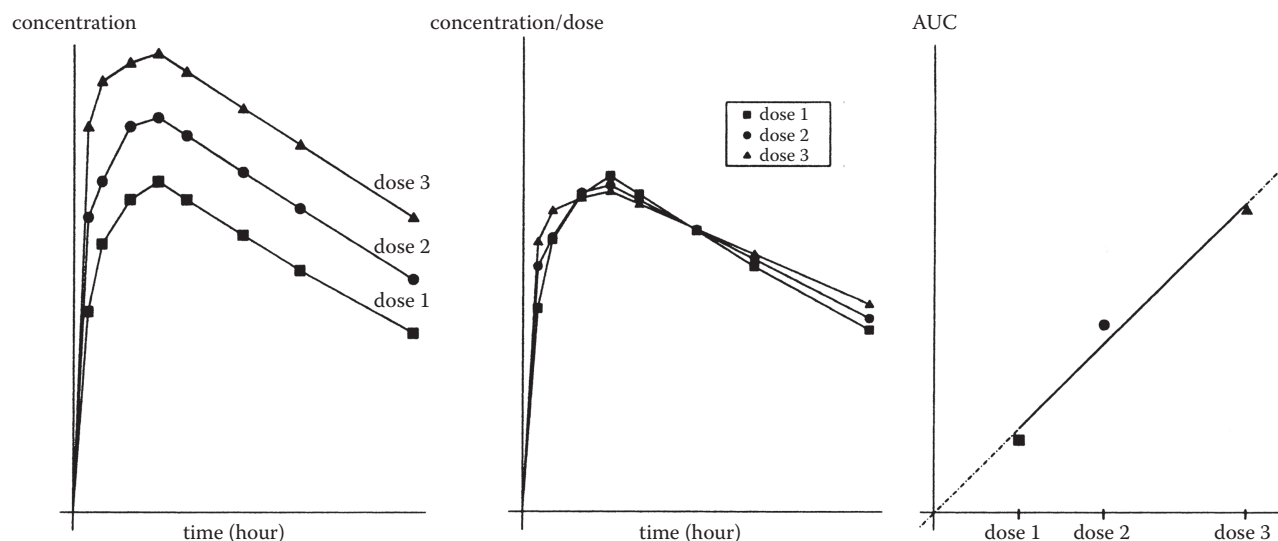


Fig. 1 Pharmacokinetic profile under dose proportionality

Besides the classical approach, Haynes and Weiss^[4] proposed the following biphasic model to assess DP:

$$Y = \begin{cases} b_1 \text{ dose} & 0 < \text{dose} < x_0 \\ b_2 \text{ dose} & x_0 < \text{dose} < x_{\max} \end{cases} \quad \text{if}$$

They modeled DP as a change-point problem and considered the slope ratio b_2/b_1 to be a measure of DP. When the slope ratio is greater than 1, it may reflect capacity-limited elimination or first-pass effects or both. If the ratio is less than 1, it may reflect increased clearance or reduced absorption or both.

Analysis of Variance

The second popular approach to analyze DP studies is to use ANOVA. Before analysis, correction to the response is made by dividing the response by dose. If the response is proportional to dose, such adjustment will make the corrected response constant. Each dose level is assumed to be associated with a treatment effect. Thus the basic model can be written as:

$$\frac{Y_j}{(\text{dose})_j} = \mu + \alpha_j$$

where j is the index for dose level, μ denotes the grand effect, and α denotes the treatment effect. If the hypothesis $H_0: \alpha_1 = \alpha_2 = \dots = 0$ is not rejected with an overall F -test, then DP is claimed. Otherwise, statistical methods for multiple comparison can be employed to test if any individual's α_j s are close to zero or to assess the differences among dose levels. Yuh et al.^[5] proposed a step-down method using Helmert contrasts to assess DP.

Unlike the regression procedure, the basic model of this approach does not need to assume any relationship among doses. However, because it ignores the ordering of doses, the approach is not as efficient as the regression. Given the estimates of the model, it is difficult to predict the response for the doses not actually studied.

Power Model

According to an empirical relationship between AUC of $C_{\max}$ and dose, Troy et al.^[6] used the following power model to assess DP:

$$Y = a(\text{dose})^\beta$$

In this model, the exponent of dose, β , is a measure of DP. When $\beta = 0$, it implies that the response Y is independent of dose. When $\beta = 1$, DP is declared. The effect of changing dose can be evaluated by:

$$\frac{Y_{j'}}{Y_j} = \left(\frac{(\text{dose})_{j'}}{(\text{dose})_j} \right)^\beta$$

Thus when the dose is doubled, the response is expected to increase by a factor of 2^β .

It should be noted that, after log transformation, the model will become linear:

$$\log Y = \alpha + \beta \log(\text{dose}) \quad \text{where } \alpha = \log a$$

The usual linear regression procedures can then be applied in this situation. Because the log transformation tends to stabilize the variance, we do not need to select any weights to correct the heterogeneity in variance. In fact, following FDA guidelines,^[7] the assumption on the normal error term can be well justified under the transformation.

The primary objective of a DP study is to examine the relationship between the key PK parameters and dose. Rather than test DP solely, it is desirable to estimate the degree of DP or the effects of dose change and to describe the uncertainty. For obtaining confidence intervals, the analysis of the raw ratio in the linear regression model relies on the use of Fieller's method, whereas it is straightforward in the power model. Also, the confidence intervals based on the power model are relatively narrower than with the ANOVA model because dose order and values are ignored in the latter.

DESIGN OF THE DOSE-PROPORTIONALITY STUDY

Recall that the DP study purports to compare doses and to estimate their effects on the key PK parameters. It should not be conducted with any designs where background factors may confound with dose. Usually, Latin squares or a balanced incomplete block design is considered. However, when these crossover designs are used, it is necessary to justify additional sources of variation in models. For example, if the power model is used to fit the PK data, we may consider the following model:

$$\log Y_{ijk} = \alpha_i + \beta \log(\text{dose}_j) + P_k$$

where α_i is the i th subject effect and P_k is the k th period effect.

Dose selection is often dictated by safety concerns, which involve the results of other PK studies. From a statistical perspective, the optimal design for estimating β (or b in linear regression) is to select the lowest and highest doses with equal observations. However, it is unwise to assess DP with only two doses because the relationship between the PK parameters and the dose can be quadratic or in other nonlinear patterns. In practice, at least three doses should be used in a DP study.

When more than three doses are available, one can make the dose selection by comparing the design matrix of

our model. To illustrate the idea, Ezzet and Spiegelhalter^[2] considered the following model

$$\log \text{AUC}_{ij} = \alpha_i + \beta \log(\text{dose}_j) + \gamma [\log(\text{dose}_j)]^2$$

They specified five doses—25, 50, 100, 200, and 400 mg—and assumed a maximum of five doses per subject. Using the criterion of minimizing the standard error of model estimates, they compared the diagonal element in $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to β and γ , where $\mathbf{X}$ is the design matrix. For a three-period crossover study, the optimal dose combination was found to be (25, 100, 400). Similar comparisons were also made for crossover design with four and five periods.

REFERENCES

1. Rowland, M.; Tozer, T. *Clinical Pharmacokinetics—Concepts and Applications*, 3rd Ed.; Williams and Wilkins: Baltimore, MD, 1995, 394–423.
2. Ezzet, F.; Spiegelhalter, D.J. Pharmacokinetic Dose Proportionality: Practical Issues on Design, Sample Size and Analysis. In *The Second International Meeting on Statistical Methods in Biopharmacy*, Paris; 1993.
3. Hutchison, M.; Keene, O. *Assessment of Dose Proportionality*. Report from the PSI/PK–UK Joint Working Party, 1994.
4. Haynes, J.; Weiss, A. Modeling Pharmacokinetic Dose-Proportionality Data. In *ASA 1989 Proceedings of the Biopharmaceutical Section*; 85–89.
5. Yuh, L.; Eller, M.; Ruberg, S. A Stepwise Approach for Analyzing Dose Proportionality Studies. In *ASA 1990 Proceedings of the Biopharmaceutical Section*; 47–50.
6. Troy, S.M.; Cevallos, W.H.; Conrad, K.A.; Chiang, S.T.; Latts, J.R. The absolute bioavailability and dose proportionality of intravenous and oral dosage regimens of Recainam. *J. Clin. Pharmacol* **1991**, *31*, 433–439.
7. Office of Generic Drugs, FDA. *Statistical Procedures for Bioequivalence Studies Using a Standard Two-Treatment Crossover Design*; 1988.

Dose Response Analysis in Clinical Trials

James MacDougall

Genzyme, Cambridge, Massachusetts, U.S.A.

Abstract

Dose response analysis plays an integral role in the process of finding an optimal dose or range of doses for physicians to prescribe. This entry explores the factors that influence dose response analysis and discusses different trial designs for testing these factors.

INTRODUCTION

A primary objective of clinical trials in the drug development process is the selection of an optimal dose or range of doses for physicians to prescribe. Dose–response analysis plays an integral role in achieving this objective. Ruberg^[1] summarizes the objectives of dose–response analysis into the following four questions.

1. Is there any drug effect?
2. What doses exhibit a response different from control?
3. What is the nature of the dose–response relationship?
4. What is the optimal dose?

The design and analysis of a dose–response study should be in accordance with the objectives of the study. The choice of study design for a dose–response study depends on the therapeutic area, phase of drug development, and the endpoint(s) of interest. Ruberg,^[2] Ting,^[3] and Sheiner^[4] review the clinical trial design aspects of dose–response studies.

REGULATORY GUIDELINES

The ICH E-4 guideline “Dose–Response Information to Support Drug Registration”^[5] provides broad guidelines on the design and analysis of dose–response studies. The ICH E-9 guideline “Statistical Principles for Clinical Trials”^[6] also briefly discusses trials to show dose–response relationships. The following is a summary of the guidelines relative to dose–response analysis.

- Assessment of the dose response should be an integral part of establishing the safety and efficacy of the drug.
- When available, dose concentration data are useful and should be incorporated into the dose–response analysis.

- Dose–response information should be derived from well-controlled studies based on accepted statistical methodologies. In addition to the formal dose–response analysis, the entire database should be examined for dose–response information.
- A parallel dose–response study design provides group mean (population average) dose response, not individual dose–response information.
- A titration design can have difficulties from a dose–response analysis standpoint, with the potential of overestimating the optimal dose.
- A forced titration design has the disadvantage of not being able to distinguish between a response to increased dose from a response to increased time on drug.
- Crossover designs have potential drawbacks, but have the advantage that each individual receives several different doses.
- Regulatory agencies and sponsors should be open to new approaches and receptive to reasoned exploratory data analysis in analyzing and describing dose–response data.
- A well-controlled dose–response study is also a study that can serve as primary evidence of effectiveness.
- Depending on the objective, the use of confidence intervals and graphical methods may be as important as the use of statistical tests.

CHARACTERIZING THE DOSE RESPONSE

In general, a dose–response curve can be characterized by its minimum effect, maximum tolerability, variability of response at a particular dose, rate of change in response as the dose changes, and potency. Determining the minimum effective dose is an important part of an efficacy trial. Finding the maximum tolerable dose (the highest dose with an acceptable safety profile) is critical for safety considerations.

Tests for Determining Drug Effect

An initial objective of dose–response analysis is determining if there is any drug effect. Hamaski,^[7] Tamhame,^[8] and Phillips^[9] review various methods of testing for a drug effect. The following are brief nontechnical descriptions of common methods for detecting a drug effect.

Linear Trend

A widespread and straightforward approach to dose–response analysis is to apply regression methods to determine if there is a linear dose response.^[10] The dose or concentration is a covariate and regression methods are used to determine if there exists a linear trend.

Overall *F*-Test

In the analysis of variance (ANOVA) setting, test the null hypothesis that all the means are equal. Variations of this method include testing only the highest dose vs. placebo and pooling all active doses and testing vs. placebo.^[11,9]

Contrasts

In the ANOVA setting, linear contrasts can provide powerful methods for detecting dose response. Stewart and Ruberg^[11] and Ruberg^[1] review methods of creating linear contrast tests powered for estimated dose–response alternatives.

William's Test

Based on a monotonic dose–response alternative, the estimate of the highest group mean is compared to the control group.^[12,13] William's testing methodology can also be used in estimating the minimally effective dose MED.

Bartholomew's Test

This is a modification of the *F*-test based on applying a monotonic order restriction to the likelihood ratio principle.^[14–16]

Jonckheere's Test

This is a rank-based method utilizing an ordered alternative comparing the number of times an observation from a higher-dose group is larger than an observation from the lower-dose group.^[17]

Cochran–Armitage Test

This method is used for detecting a linear trend across dose groups in proportions of events when the outcome is binomial.^[18,19]

Although no one test can be universally recommended for all scenarios, the results of Phillips^[9] indicate that tests incorporating the ordering of the dose groups are powerful against a variety of dose–response patterns and should be considered when the objective is to determine if there is a dose response. Further information on order-restricted inference is given in Barlow et al.^[20]

Minimally Effective Dose and Maximum Tolerated Dose

A common objective in characterizing the dose response of a compound is determining the MED. Ruberg,^[1] Källén and Larsson,^[21] and Senn^[10] review the concept of MED and methods of determining the MED. A key issue surrounding the MED is that there is no universally accepted definition for MED. Ruberg^[1] proposed the following definition: “The minimally effective dose is the smallest dose producing a clinically important response that can be declared statistically, significantly different from the placebo response,” incorporating both clinical and statistical aspects MED.

The maximum tolerated dose (MTD) is an important concept in dose–response clinical trials as it defines an upper limit in determining an optimal dose. Similar to the MED, there is no universally accepted definition of MTD for clinical trials. Culter et al.^[22] discuss various issues surrounding the MTD.

Sigmoid Models

It is often reasonable to assume that the dose–response relationship will be sigmoidal in nature. In pharmacokinetics/pharmacodynamics, the E_{MAX} model is commonly used to characterize the dose–response curve.^[21,23,24] This model is represented by the following equation: $R = E_0 + (D^N \times E_{MAX}) / (D^N + ED_{50}^N)$. Where R is the response, E_0 is the Baseline Effect, D is the dose, E_{MAX} is the maximum drug effect, ED_{50} is the dose at half-maximal effect, and N is the slope factor (often called the Hill factor). The E_{MAX} model can estimate useful parameters in determining an optimal dose; for example, the ED_{90} , the dose that yields 90% of the maximum effect could be used as an estimate for the optimal dose. This model is most appropriate when there is sufficient number of different doses (or concentration range) such that the dose response can be accurately characterized. This model can also be extended to a mixed model when subjects are assessed at more than one dose level.

Mixed-Effects Models

If the design of the trial is such that different doses are measured within the same subject, it may be beneficial to utilize mixed-effects modeling to characterize the dose–response curve. This method has the advantage

that it is based on dose response at the subject level and not the population. Mixed-effects modeling for dose–response analysis can be either linear or nonlinear. Information on mixed-effects models with applications to dose–response analysis can be either linear or nonlinear. Information on mixed-effects models with applications to dose–response analysis and clinical trials is given in Zhang and Gould,^[25] Källen and Larsson,^[21] and Davidian and Giltinan.^[26]

Orthogonal Polynomials

Orthogonal polynomials can be used to characterize the shape of the dose response by determining if there are polynomial trends in the dose response.

Nonparametric Regression

As mentioned in regulatory guidelines, the appropriate use of graphical methods may be as important as statistical tests. Nonparametric regression can be a useful tool in dose–response visualization and informally assessing goodness of fit for a parametric dose–response model. Nonparametric regression estimates a smooth function based on the data. Two popular methods are kernel regression and smoothing splines. More detailed information on nonparametric regression is given in Hazelton^[27] and Hardle.^[28]

DETERMINING WHICH DOSES ARE EFFECTIVE

In the later-phase clinical trials, the objective of the study is to determine which doses demonstrate a response different than control. Testing multiple doses vs. control inherently raises the issue of multiplicity. It is anticipated by regulatory agencies that adjustment for multiplicity in a confirmatory clinical trial will be addressed and documented.^[6] One method of addressing the multiplicity issue is to control the overall familywise error rate of the hypotheses at a predefined level α by use of a multiple comparison procedure (MCP).

Descriptions of the theory and applications of MCPs can be found in Hochberg and Tamhane,^[29] Hsu,^[30] and Miller.^[31] A review of various MCPs applied to the clinical trial setting can be found in Tamhane and Dunnett,^[32] Zhang et al.,^[33] Lakshminarayanan,^[34] and Westfall et al.^[35] The following are brief nontechnical descriptions of common MCPs applied in dose–response analyses.

Bonferroni Adjustment

The standard Bonferroni adjustment tests each of k hypotheses at level α/k to control the familywise error rate at α .

Fisher's Least Significant Difference

Fisher's least significant difference test is a two-step method for controlling the familywise error rate in the ANOVA setting. The first step is an overall test of equality of treatment means at level α . If that test is rejected, the second step tests the individual hypotheses each at level α .

Bonferroni–Holm Procedure

Holm^[36] developed a uniformly more powerful sequential version of the single-step Bonferroni procedure based on ordering the p values. This procedure is referred to as a step-down sequential method as the p values are tested from smallest to largest.

Hochberg Sequential Procedure

Hochberg^[37] developed a sequential method of testing starting with the least significant p value. This procedure is referred to as a step-up sequential method as the p values are tested from largest to smallest. Any hypothesis rejected by the Bonferroni–Holm procedure will also be rejected by the Hochberg procedure. The Hochberg sequential procedure does not guarantee control of the familywise error rate for all types of dependence among p values. However, simulation studies have shown that the method is conservative for most scenarios that would arise in a dose–response setting.^[38]

Dunnett's Procedure

Dunnett^[39] developed a single-step MCP for testing multiple treatments against a control incorporating the normality of the observations and the correlation structure. This method can be extended to both step-down and step-up sequential MCP tests.^[40]

Fixed Sequential Testing Procedure

Fixed sequential testing predefines the sequence of the hypothesis tests. The first hypothesis is tested at level α . If the first hypothesis is rejected, test the second hypothesis at level α . If the second hypothesis is rejected, test the third at level α and so on.^[1,35,41] In contrast to other sequential methods, the hypothesis testing order is defined a priori, not determined by the magnitude of the corresponding p values. This method has also been referred to as the “step-down” method because in studies comparing multiple doses to a control in a monotonic dose setting, the a priori ordering would generally be from highest dose to lowest.

In general, if your primary interest is in hypothesis testing, sequential MCPs provide better power than their single-step counterparts. The choice of which particular MCP to utilize depends on the endpoint, the hypotheses

being tested, the underlying assumptions of the MCP, and the alternatives you wish to be powered for.

ANALYSIS CONSIDERATIONS BASED ON TRIAL DESIGN

Three common dose–response trial designs are the parallel, crossover, and dose-escalation (titration) designs. The following table taken from Sheiner^[4] contrasts various design and analysis aspects of the three different dose–response designs.

The following is a brief summary of study design specific issues in dose–response analysis.

Parallel Group Designs

Randomized parallel-group dose–response studies are widely used in clinical trials. They have the advantage of being relatively straightforward in design and analysis. A disadvantage of the parallel-group design is that as a subject receives only one level of the treatment, the characterization of the dose response is at the population level not the individual level.

<i>Problem</i>	<i>Parallel dose</i>	<i>Crossover</i>	<i>Dose escalation</i>
Ethical	++	++	0
Problems in study execution			
Number of patients	<i>N</i>	<i>N/m</i>	<i>N/m</i>
Time	1 unit	<i>m</i> units	<i>m</i> units
Need for many centers	++	+	+
Complexity of protocol	0	+	+/++
Problems in study analysis			
Complexity	0	+	+
Reliance on modeling assumptions	0	+	++
Carryover effects	0	++	+
Period effects	0	++	+
Center effects	++	+	+
Few doses examined	++	++	0
Biased parameter estimates	+++	0	+
Poor/no estimates of variability	+++	+	+
Problems in extrapolation			
Not representative of individual response	+++	0	0
Not representative of clinical practice	++	+++	0
Not representative of patient group	+	+++	0

m = number of escalations/crossovers; 0 = problem absent; + = minor problem; ++ = moderate problem; +++ = major problem.

Crossover Designs

Senn^[42] reviews the design, analysis, and issues pertaining to crossover design clinical trials. Crossover designs have the advantage of a subject being assessed at multiple dose levels, allowing for the characterization of subject level dose response. Disadvantages of crossover designs are concerns surrounding carryover effects and treatment withdrawals.^[5,43] The analysis of crossover designs dose–response data should incorporate the repeated-measures aspect of the design.^[4] Further information on the analysis of crossover designs can be found in Senn^[42] and Rashid.^[44]

Titration Designs

Titration designs (or dose-escalation designs) are described in Refs.^[3–5,45] Titration designs have the advantage that they can be used to study a wide range of doses, the subject is assessed at multiple dose levels, and the design can provide clear evidence of dose effect.^[5] Disadvantages of this design are the potential confounding between dose and duration of treatment and concerns surrounding overestimation of the optimal dose.^[4–5,45–46] Further information on the analysis of titration designs can be found in Agin and Pun,^[45] Pun,^[47] Chuang,^[48] and Chuang-Stein and Shih.^[49]

ORDINAL ENDPOINTS

Most of the focus on dose–response analysis methods assumes that the endpoint is either continuous or binary. Chuang-Stein and Agresti^[50] review tests for detecting a dose–response relationship with ordinal response data.

SAFETY ENDPOINTS

Many of the dose–response analysis methods can be used for safety as well as efficacy endpoints. The following references focus on dose response for safety endpoints: Cutler et al.,^[22] Thall and Russell,^[51] O'Neill,^[52] and Hsu and Laddu.^[53]

COMBINATION-TREATMENT STUDIES

In some studies, the objective is to investigate dose–response relationships over a combination of therapies. In this type of study, special consideration must be given to compound hypotheses and multiple comparisons. Raghuvaran and Altan^[54] and Chow and Liu^[55] review response surface methodology. Further details on the analysis of combination-treatment studies are given in Hafner et al.,^[56] Phillips et al.,^[57] Hung,^[58] and Tangen.^[59]

CONCLUSION

Dose–response analysis is a fundamental part of the drug development process both in the early and later phases of clinical trials. Dose–response analysis determines if a drug effect exists, characterizes the dose–response curve, provides necessary information for choosing doses for future studies, and demonstrates safety and efficacy at a particular dose.

Because of the variety of objectives, designs, and endpoints that arise in the clinical trial setting, many dose–response analysis methods exist. There is no one dose–response analysis method that can be universally recommended. The choice as to which methods are most appropriate to apply should be based on the objectives and design of the trial.

REFERENCES

- Ruberg, S.J. Dose–response studies. II. Analysis and interpretation. *J. Biopharm. Stat.* **1995**, 5 (1), 15–42.
- Ruberg, S.J. Dose–response studies. I. Some design considerations. *J. Biopharm. Stat.* **1995**, 5 (1), 1–14.
- Ting, N. Dose Response Study Designs. In *Encyclopedia of Biopharmaceutical Statistics*; Chow, S., Ed.; Marcel Dekker, 2003.
- Sheiner, L.B.; Beal, S.L.; Sambol, N.C. Study designs for dose–ranging. *Clin. Pharmacol. Ther.* **1989**, 46, 63–77.
- ICH-E4 Guideline, *Dose–Response Information to Support Drug Registration, Step 4*; 1998.
- ICH-E9 Guideline, *Statistical Principles for Clinical Trials, Step 4*; 1998.
- Hamaki, T.; Tatsuya, I.; Mitsumasa, B.; Masahi, G. Statistical approaches to detecting dose–response relationships. *Drug Inf. J.* **2000**, 34, 579–590.
- Tamhane, A.; Hochberg, Y.; Dunnett, C. Multiple test procedures for dose finding. *Biometrics*. **1996**, 52, 21–37.
- Phillips, A. A review of the performance of tests used to establish whether there is a dose effect in dose–response studies. **1998**, 32, 683–692.
- Senn, S. *Statistical Issues in Drug Development*; John Wiley and Sons: Chichester, 1997.
- Stewart, W.H.; Ruberg, S. Detecting dose response with contrasts. *Stat. Med.* **2000**, 19, 913–921.
- Williams, D.A. A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*. **1971**, 27, 103–177.
- Williams, D.A. The comparison of several dose levels with a zero dose control. *Biometrics*. **1972**, 28, 519–531.
- Bartholomew, D.J. A test of homogeneity for ordered alternatives. *Biometrika*. **1959**, 46, 36–48.
- Bartholomew, D.J. A test of homogeneity for ordered alternatives II. *Biometrika*. **1959**, 46, 328–335.
- Bartholomew, D.J. A test of homogeneity of means under restricted alternatives. *J. R. Stat. Soc., Ser. B.* **1961**, 23, 239–281.
- Jonckheere, A.R. A distribution-free k sample test against ordered alternatives. *Biometrika*. **1954**, 41, 133–145.
- Cochran, W.G. Some methods for strengthening the common 2 tests. *Biometrics*. **1954**, 10, 417–451.
- Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics*. **1955**, 11, 375–386.
- Barlow, R.E.; Bartholomew, D.J.; Bremner, J.M.; Brunk, H.D. *Statistical Inference Under Order Restrictions*; John Wiley and Sons: New York, 1972.
- Källén, A.; Larsson, P. Dose response studies: How do we make them conclusive? *Stat. Med.* **1999**, 18, 629–641.
- Cutler, N.R.; Sramek, J.J.; Greenblatt, D.J.; Chaikin, P.; Ford, N.; Lesko, L.J.; Davis, B.; Williams, R.L. Defining the maximum tolerated dose: Investigator, academic, industry and regulatory perspectives. *J. Clin. Pharmacol.* **1997**, 37 (9), 767–783.
- Holford, N.H.G.; Sheiner, L.B. Understanding the dose–effect relationship: Clinical application of pharmacokinetic–pharmacodynamic models. *Clin. Pharmacokinet.* **1981**, 6, 429–453.
- Machaco, S.G.; Miller, R.; Hu, C. A regulatory perspective on pharmacokinetic/pharmacodynamic modelling. *Stat. Methods Med. Res.* **1999**, 8, 217–245.
- Zhang, D.; Gould, A.L. Mixed Effects Models. In *Encyclopedia of Biopharmaceutical Statistics*; Chow, S., Ed.; Marcel Dekker: New York, 2000.
- Davidian, M.; Giltinan, D.M. *Nonlinear Models for Repeated Measurement Data*; Chapman & Hall: London, 1995.
- Hazelton, M.L. Nonparametric Regression. In *Encyclopedia of Biostatistics*; Armitage, P. Colton, T., Ed.; John Wiley & Sons: New York, 1998.
- Hardle, W. *Applied Nonparametric Regression*; Cambridge University Press: Cambridge, 1990.
- Hochberg, Y.; Tamhane, A.C. *Multiple Comparison Procedures*; John Wiley and Sons: New York, 1987.
- Hsu, M. *Multiple Comparisons*; Chapman and Hall: London, 1996.
- Miller, R. *Simultaneous Statistical Inference*; Springer-Verlag: New York, 1981.
- Tamhane, A.C.; Dunnett, C. Stepwise multiple test procedures with biometric applications. *J. Stat. Plan. Inference*. **1999**, 82, 55–68.
- Zhang, J.; Quan, H.; Ng, J.; Stepanavage, M. Some statistical methods for multiple endpoints in clinical trials. *Control. Clin. Trials*. **1997**, 18, 204–221.
- Lakshminarayanan, M. Multiple Comparisons. In *Encyclopedia of Biopharmaceutical Statistics*; Chow, S., Ed.; Marcel Dekker: New York, 2000.
- Westfall, P.; Tobias, R.; Rom, D.; Wolfinger, R.; Hochberg, Y. *Multiple Comparisons and Multiple Tests using the SAS® System*; SAS Institute: Cary, NC, 1999.
- Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **1979**, 6, 65–70.
- Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. **1988**, 75, 800–802.
- Hochberg, Y.; Hommel, G. Multiple Hypotheses, Simes' test; Kotz, S., Read, C.B., Banks, D.L., Ed.; *Encyclopedia of Statistical Sciences Update*, John Wiley & Sons: New York, 1998, Vol. 2.
- Dunnett, C. A multiple comparison procedure for comparing several treatments to a control. *JASA*. **1955**, 50, 1096–1121.

40. Dunnett, C.; Tamhane, A. A step-up multiple test procedure. *JASA*. **1992**, *87*, 162–170.
41. Maurer, W.; Hothorn, L.; Lehmacher, W. Multiple comparisons in drug clinical trials and preclinical assays: A-priori ordered hypotheses. *J. Biom. Chem.-Pharm. Ind.* **1995**, *6*, 3–18.
42. Senn, S. *Cross-Over Trials in Clinical Research*; John Wiley and Sons: Chichester, England, 1993.
43. Phillips, A. Design and analysis of dose–response studies: Reality versus regulatory requirements. *Drug Inf. J.* **1997**, *31*, 737–744.
44. Rashid, M.M. Robust Analysis for Crossover Design. In *Encyclopedia of Biopharmaceutical Statistics*; Chow, S., Ed.; Marcel Dekker, 2000.
45. Agin, M.A.; Pun, E.F.C. Titration Design. In *Encyclopedia of Biopharmaceutical Statistics*; Chow, S., Ed.; Marcel Dekker: 2000.
46. Temple, R. Government viewpoint of clinical trials of cardiovascular drugs. *Cardiovasc. Pharmacother.* **1989**, *73*, 495–509.
47. Pun, E.F.C. A Parallel Dose-Titration Design to Determine Dose Range; In *ASA Proc. Biopharm. Sect.*; **1990**; 159–164.
48. Chuang, C. The analysis of a titration study. *Stat. Med.* **1987**, *6*, 583–590.
49. Chuang-Stein, C.; Shih, W.J. A note on the analysis of titration studies. *Stat. Med.* **1991**, *10*, 323–328.
50. Chuang-Stein, C.; Agresti, A. A review of tests for detecting a monotone dose–response relationship with ordinal response data. *Stat. Med.* **1997**, *16*, 2599–2618.
51. Thall, P.F.; Russell, K.E. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics*. **1998**, *54*, 251–264.
52. O'Neill, R.T. Statistical concepts in the planning and evaluation of drug safety from clinical trials in drug development: Issues of international harmonization. *Stat. Med.* **1995**, *14*, 1117–1127.
53. Hsu, P.; Laddu, A.R. Analysis of adverse events in a dose titration study. *J. Clin. Pharmacol.* **1994**, *34*, 136–141.
54. Raghavaro, D.; Altan, S. Response Surface Methodology. In *Encyclopedia of Biostatistics*; Armitage, P.; Colton, T., Ed.; John Wiley & Sons, 1998.
55. Chow, S.; Liu, J. *Design and Analysis of Clinical Trials: Concepts and Methodologies*; John Wiley and Sons: New York, 1998.
56. Hafner, K.B.; Ruberg, S.J. Factorial dose–response studies using frequency and magnitude of dose. *J. Biopharm. Stat.* **1996**, *6*, 253–262.
57. Phillips, J.A.; Cairns, V.; Koch, G.G. The analysis of a multiple-dose, combination-drug clinical trial using response surface methodology. *J. Biopharm. Stat.* **1992**, *2*, 49–67.
58. Hung, H.M.J. On identifying a positive dose–response surface for combination agents. *Stat. Med.* **1992**, *11*, 703–711.
59. Tangen, C.M.; Koch, G.G. Non-parametric analysis of covariance for confirmatory randomized clinical trials to evaluate dose–response relationships. *Stat. Med.* **2001**, *20*, 2585–2607.

Dose Response Study Design

Naitee Ting

Pfizer Global Research and Development, New London, Connecticut, U.S.A.

Abstract

Choosing the best designs for dose-response studies will help reduce the risk of recommending an incorrect dose for a new drug. The purpose of this entry is to review the considerations necessary to a successfully designed dose response study.

INTRODUCTION

For every drug that is available at the pharmacy, there is information regarding how much drug should be used by a patient. In many over-the-counter drugs, such information may also be available for adults and children separately.

From a physician's point of view, patients may respond to a drug differently according to age, gender, race, metabolism, or other characteristics. Hence, in practice, a physician may first prescribe a drug with a given dose based on the patient's background, and then the physician may adjust dosage depending on the patient's response to the prescribed drug.

All dosing information is based on the dose-response relationship obtained from the drug. Understanding the dose-response relationship is one of the most challenging tasks in drug development.

OVERVIEW

Conceptually, the efficacy of a drug increases as the dose increases (Fig. 1). The left curve plotted in Fig. 1 can be viewed as an efficacy-response curve; that is, the response (y-axis) represents some measure of the drug efficacy. In addition to efficacy response, there are also toxicity dose-response curves. We will assume, as shown in Fig. 1, when dose increases, both efficacy and toxicity increase. Based on these curves, the maximum effective dose (MaxED) and the maximum tolerable dose (MTD) can be defined: MaxED is the dose above which there is no clinically significant increase in pharmacological effect or efficacy, and MTD is the maximal dose acceptably tolerated by a particular patient population. Another dose parameter of interest is the minimum effective dose (MinED). Ruberg^[1] defines the MinED as "the lowest dose producing a clinically important response that can be declared statistically, significantly different from the placebo response."

In certain drugs, the efficacy curve and the toxicity curve are widely separated. When this is the case, there is

a wide range of doses for patients to take; that is, as long as a patient receives a dose between MaxED and MTD, the patient can benefit from the efficacy and at the same time avoid toxicities from the drug. However, for other drugs, the two curves may be very close to each other. Under this situation, physicians have to dose patients very carefully so that while benefiting the efficacy, patients do not have to be exposed to potential toxicity from the drug. The area between the efficacy and the toxicity curve is known as the "therapeutic window." One way to measure therapeutic window is to use the "therapeutic index (TI)." TI is defined as the ratio of MTD/MaxED. Clearly, a drug with a wide therapeutic window (or a high TI) tends to be preferred by both physicians and patients. If a drug has a narrow therapeutic window, then the drug will need to be developed carefully, and physicians will prescribe the drug with caution.

Oftentimes, drugs developed for life-threatening diseases (e.g., cancer or AIDS) tend to have narrow therapeutic windows. On the other hand, drugs developed to treat chronic diseases (e.g., high blood pressure, depression, high cholesterol, or arthritis) are required to have wide therapeutic windows because these drugs are frequently used in older patients, or patients with other concurrent diseases, and sometimes used in children.

The next section introduces the process of selecting doses in developing a new drug. The "General Design Considerations" section discusses some general considerations in designing dose-response clinical trials. "Drug/Disease Specific Considerations" presents a few special conditions, and the "Discussion" section covers the overall strategy in dose selection.

DOSE-SELECTION PROCESS FOR A NEW DRUG

In the drug development process, a new drug candidate would have to be studied in the laboratories first before it could be tested in humans. Drug testing that is performed in humans is called clinical trial, and tests out of humans

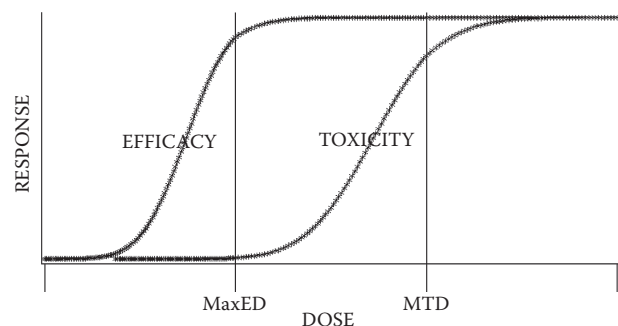


Fig. 1 Dose–response curves for efficacy and toxicity

are called nonclinical studies. Nonclinical studies are performed to evaluate various properties of the new drug. The knowledge obtained from these studies would then be used to help design and conduct clinical trials.

In nonclinical studies, pharmacokinetic (PK) and pharmacodynamic (PD) properties (details of PK and PD can be found in Vol. 1 of *Encyclopedia of Biopharmaceutical Statistics*) are studied to help identify dose–response relationships for the new drug. These data are useful in understanding how frequent and how much drug should be used to dose animals. Another important area of the nonclinical studies is toxicology. Drug toxicity is studied in a number of experiments. Results of these studies will help identify the toxic doses and types of toxicity to be anticipated in clinical trials. This information is useful in estimating the highest dose to be tested in humans.

There are four phases (phase I, II, III, and IV) in clinical development. The purposes of Phase I clinical trials usually include testing the PK properties and studying the safety of the test drug. These studies are mostly performed on normal volunteers. Sometimes surrogate markers are used in Phase I trials to study PD properties of the test drug. At the early stage of clinical testing, single-dose, dose-escalation designs are often used to study tolerability of the new drug. In a single dose, dose-escalation design, a few subjects are randomized to placebo and low dose of the test drug. If there are no adverse events, then another group of subjects is randomized to placebo and the next higher dose of test drug. This process is continued until either some adverse events are observed or until the top designed dose is reached. The information collected from these studies can be used to estimate the MTD.

In general, phase II clinical trials are designed to study the dose–response relationship. Usually, this is the first time the study drug is tested in patients with the disease under study. A typical phase II dose–response design randomizes patients into a few treatment groups, which include placebo and some low or high doses of the test drug. Fig. 2 shows the graph of a set of data simulated from a dose–response study with placebo and three active doses.

Sometimes it may take several phase II studies before a set of doses can be recommended for phase III clinical development. When in phase III, we usually assume that

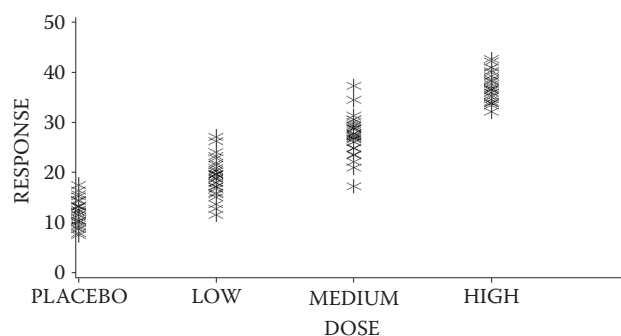


Fig. 2 Observations from a simulated dose–response study

the doses are known. Phase III clinical trials are designed to confirm these findings. However, in some cases, results from phase III studies may still lead to dose changes. During or after the phase III program, a sponsor would put together a new drug application (NDA), which includes results and conclusions made from all of the nonclinical and clinical studies. The NDA is then submitted to regulatory agencies (e.g., Food and Drug Administration [FDA] in the United States) for review and approval. One important component of an NDA is the drug label. Dosing information should be well specified in the label. The dosing information is mainly obtained from phase II and phase III clinical trials.

After a new drug is approved, phase IV clinical studies are conducted. Some of these studies are referred to as “post-marketing surveillance” trials. One important purpose of these trials is to monitor the drug safety in the general patient population (as opposed to clinical trial patient population). Occasionally in postmarketing studies, we may see a drug is efficacious at a lower dose than the dosage recommended in drug label. This lower dose tends to provide a better safety profile. When this is the case, the drug label could be changed to include the lower dose as one of the recommended doses.

Based on the dose-selection process described above, it is obvious that selecting the right dose for a new drug is a very important process. Without dose information, it is not possible for a physician to prescribe the drug to patients. One of the regulatory documents that describes the importance and practical difficulties in the study of the dose–response relationship is International Conference on Harmonization (ICH) E4^[2] Guidance. Readers are encouraged to refer to this document for some of the regulatory viewpoints.

GENERAL DESIGN CONSIDERATIONS

Fig. 2 presents an example of data collected from a typical dose–response study design, which is sometimes referred to as a “randomized, double-blind, placebo-controlled, fixed-dose, parallel dose–response design.” In such a study, patients are randomized into a few predetermined dose groups (often include the placebo, a low dose, a medium dose, and a high dose), and they take the randomized dose for a period of time (usually several weeks to several months). After the study is

completed, the efficacy and safety data obtained from these patients are analyzed to study the dose–response relationship. Note that “parallel,” “fixed-dose,” and “placebo-control” are some important features of this design.

A parallel design is in contrast with a crossover design. In a 2×2 crossover design, a subject is randomized into one of two sequences. For one sequence, the subject takes treatment A for a period of time, and then takes treatment B after washout of treatment A. The treatment order is reversed for the other sequence. Sometimes more complicated crossover designs are utilized. In a parallel design, the subject is randomized to receive a treatment and stays on the same treatment until the end of study.

A fixed-dose design is in contrast with a dose-titration design. In a fixed-dose design, once a patient is randomized to a dose group, the patient would take the same dose of study drug throughout the complete dosing period. In a dose-titration design, a patient is randomized to a dose regimen with a starting dose, then the dose for a patient can be changed over time depending on the design, or on the patient’s response to the dose.

Fixed-Dose Vs. Dose-Titration Designs

In an escalating dose-titration study, subjects are randomized to start with a low dose, and depending on either patient’s response to the drug or a predetermined schedule, the dose is gradually increased until a suitable dose level is found. Each subject can receive more than one dose. A patient who responds to a low dose may stay on this low dose, and a patient who does not respond to a low dose after a fixed treatment period may receive the next highest dose of the drug. This procedure is repeated until some designed criteria are satisfied.

There are some advantages of a titration design. For example, a study with this design will allow a patient to be treated at the optimum dose for the patient; this dose-allocation feature reflects the actual medical practice. However, the disadvantage of a titration design is the difficulty in data analysis. For example, if a patient responded to the test drug after doses are escalated, it is unclear whether the higher dose or the accumulation of the lower dose causes the response. In titration designs with multiple treatment groups, there may be overlapping doses; for example, one treatment group is 10 mg escalating to 20 mg, and another group is 20 mg escalating to 40 mg. When this is the case, it is difficult to make inferences about the 20-mg-dose group.

In some cases, instead of a dose–response study, a concentration–response study is designed. A concentration–response study assesses efficacy and safety measurements observed from subjects according to the plasma concentration of the study drug, but not the doses of the study drug. There are many practical concerns in using this type of designs, e.g., how to blind the patient and the physician, how and when to measure the blood concentration, etc.

Because of the issues with dose-escalation designs and concentration–response designs, the parallel, fixed-dose designs are, in general, the preferred design for dose–response studies. Therefore, in most of the dose–response studies, patients are randomized to a few fixed-dose groups and are compared to a control treatment.

Selection of Control

Two types of controls can be considered in dose-finding designs: placebo control or active control. In some situations, an active control group may be used in dose–response study designs. The advantages and disadvantages of using an active control group depend on the disease under study, characteristics of the drug candidate being developed, and specific clinical inferences of interest (ICH E10).^[3] Studies with an active control, but without a placebo group, suffer from the additional burden of demonstrating that the treatment groups are effective, either through superiority to the active control or on the basis of some type of historical control information.^[4,5] An active control, however, may provide a reference against a treatment of “known” effectiveness.

In the early stage of clinical development of a new drug, it is a common practice (if deemed ethical) to compare the test drug with a placebo. This is important because detecting positive signals of effectiveness beyond that achieved with placebo is an important milestone for the continuing development of this drug. Accordingly, placebo plays an important role in a dose–response study—it represents a zero dose in the study. Patient response at zero dose is a basic standard for comparison with active doses. In typical dose–response studies, a few fixed doses (usually two, three, or four) would be chosen. These doses plus placebo make up the treatment groups for a randomized, dose–response trial.

Number of Doses to be Tested

Because there are difficulties in finding the right doses for a new drug, it is desirable to study as many doses as we can. However, the number of doses that can be tested in a given study is limited. There are practical constraints in determining the number of treatment groups. Most trials with sufficient number of patients are multicenter trials. Different centers may have different recruitment rates. As a result, there is an imbalance in the number of patients recruited by these centers. If this happens, some centers may fail to provide enough patients to be randomized to each treatment group, and the treatment by center interaction becomes nonestimable. Therefore, we have to limit the number of treatment groups in order to minimize the imbalance problem.

Another practical issue is dosage form. Sometimes there are only limited dosage forms available in the early stage of clinical development. When this is the case, the number

of doses to be used in a study may also be restricted. For some studies, the technique to achieve blinding is to produce matching placebos for each treatment group. The more doses that are included in a study, the more “matching placebos” may be needed. The number of dose groups may also be limited by practical considerations of how many pills we might reasonably expect a subject to take for any given dose. For these reasons and others, clinical studies designed with more than six or seven treatment groups are rare.

Sample Size Considerations

The general issue regarding sample sizes can be found in Vol.1 of *Encyclopedia of Biopharmaceutical Statistics*. When designing a dose–response study, the main concern in sample size calculation is how to handle multiple group comparisons. This issue is dictated by the objective of the study. A few examples of the objective of a dose–response study may be:

- Testing to see if a specific dose of test drug is different from placebo.
- Finding the minimum effective dose.
- Differentiating activities between active doses.
- Checking to see if there is an increasing trend.
- Demonstrating noninferiority between a particular dose and the active control.

As a good clinical practice, it is important to keep a single and clear primary objective for a single study. Hence, the above examples are mutually exclusive. For each of the objective in a given trial, the appropriate statistical method should be used to perform data analysis. Sample-size estimation should be consistent with the method used for data analysis.

In the first example, if the trial is designed to differentiate a specific dose of test drug from placebo, then the sample-sizing method would be similar to performing a two-sample *t*-test comparing the test drug against placebo. In most situations when analyzing dose–response studies, multiple comparison adjustments will be needed. Depending on the type of multiple comparison to be used for data analysis, sample sizes should be estimated accordingly. For example, if a stepwise multiple comparison adjustment will be used for analysis, then again, the two-sample *t*-test method is appropriate. On the other hand, if the Bonferroni adjustment is proposed for data analysis, then the same adjustment will have to be used in the sample-size-estimation procedure.

Dose Allocation, Dose Spacing

Dose–response studies are usually carried out in phase II. At this point, there is often considerable uncertainty regarding any hypothetical dose–response curve. It is typical at this time that the MTD is known from phase I studies, and it may

also be assumed that efficacy is nondecreasing with increasing dose. Even so, the underlying dose–response curve can still take many possible shapes. Given these assumed curves, there are various strategies of allocating doses. For example, if a drug is very potent and the efficacy curve rises rapidly at low dose, then we need to allocate doses at the low end to capture the drug activity. On the other hand, if the new drug is less potent and it takes some high doses in order to reach the desired efficacy response, then we need to choose a few high doses in the design. We can see that the dose-allocation strategy can be very different depending on the underlying dose–response curve.

After the number of dose groups is chosen, it is still a challenge to determine the high dose, low dose, and spacing between test doses. Typically, the high dose is a dose selected around or below the MTD, but choices of lower doses are often challenging. Wong and Lachenbruch^[6] introduce cases using equal dose spacing from low to high doses, i.e., divide the distance from placebo to highest dose by the number of active doses, then use that divided distance as the space between two consecutive doses (e.g., 20, 40, 60 mg). Others may consider some type of log dose spacing; e.g., 1, 2, 4, 8 mg for the design.

Optimal Designs

The dose levels and the number of subjects at each level can be chosen mathematically (using mathematical theory, simulation, or some other tools) in order to optimize a statistical criterion such as small errors of estimation. This set of dose levels and number of subjects at each level is called the statistically optimal experimental design, and the design depends on the chosen criterion and the underlying model for the dose–response curve.

The statistical principles of optimal experimental design can be applied to many of the study designs, and optimal design techniques can be used with various statistical models. Optimization techniques allow one to determine the statistically best set of doses and number of subjects to be used at each dose in order to estimate the parameters of the model, e.g., slope and ED50 (the dose that achieves 50% of efficacy response) in logistic regression, intercept and slope in linear regression. The doses used might not necessarily be those intended for use in the label, but they provide a basis for estimation of the dose–response curve so that the response at any dose can be predicted with good accuracy and precision. Depending on the optimality criterion chosen, the doses studied may not necessarily be equally spaced or have equal numbers of subjects at each dose. Information on the shape of the dose–response curve should be attained where possible from PK–PD studies and early phase II studies. This can then help the design of the late phase II study using the above principles.

A statistical approach to optimal experimental design is given in Pukelsheim.^[7] Wong and Lachenbruch^[6] review dose–response designs and use optimal design criteria for

linear and quadratic regression. They also use simulation to illustrate the effect of optimal design criteria on spacing of doses and the numbers of subjects at each dose.

DRUG/DISEASE SPECIFIC CONSIDERATIONS

The preceding section provides a general overview of considerations about dose–response study design. However, in some cases, there are special concerns; for example, for some diseases, it is not ethical to use placebo control. In this case, we need special designs to study drug efficacy. We present a few specific situations in this section, and discuss some of the concerns under these situations.

Drugs with Narrow Therapeutic Window

As indicated earlier, one characteristic of a good drug is that it is associated with a wide therapeutic window, i.e., the drug shows good efficacy at low doses, and it is safe even when doses are very high. Unfortunately, most drugs under development are associated with only a reasonable therapeutic window, and some drugs are with a very narrow therapeutic window. Doses for those drugs would have to be selected very carefully. In some cases, the appropriate doses may have to be determined according to individual patient characteristics. Titration designs and concentration–response designs may be reasonable approaches to deal with this issue.

Life-Threatening Diseases

In the development of drugs to treat certain life-threatening diseases (such as cancer or AIDS), the dose-selection process can be very different from what was described in the section “General Design Considerations.” Take cancer as an example. If a patient is diagnosed with cancer, the most important task is to help the patient survive. When survival is the top priority, patients tend to tolerate more drug toxicity, and a drug with narrow therapeutic window may be acceptable for these patients.

Cancer treatments include surgical operations, radiation therapy, and chemotherapy. Cancer drugs are often developed as a type of chemotherapy. Chemotherapy is usually dosed according to a patient’s body size, and is injected into the patient. Under this circumstance, it is difficult to conduct blinded studies. Hence, most chemotherapy clinical trials are not blinded. Another problem with cancer is that the use of placebo control is not ethical. Therefore, cancer trials typically include active agents, either as a control, or as background therapies. Generally speaking, chemotherapy is toxic, and because of high toxicity, it is not ethical to recruit normal volunteers for Phase I trials. Thus, in most of these drugs, cancer patients, instead of normal volunteers, are recruited for Phase I trials.

Drugs Developed for Prevention

Some drugs are developed for prevention purposes. One example is the use of female hormone supplement to prevent bone loss. In this case, the true clinical variable (clinical endpoint) of interest is bone fracture, which is very difficult to measure. If we design a study to observe bone fracture, we would need to follow many patients for many years. Hence, we use surrogate markers (e.g., bone mineral density) as the clinical endpoint in study designs.

However, the major challenge in developing drugs for prevention is that we can only propose very few doses, often with a single dose, on the drug label. The reason is that when a physician prescribes a drug for a patient to control his/her disease, the physician makes a dose adjustment according to the patient’s response to the prescribed dosage. If a drug is used for prevention, then it takes a very long time before any symptoms can be observed, and it is almost impossible for the physician to adjust the dosage. Hence, for this type of drug, the best recommendation is a single dose for prevention. The challenge in drug development becomes how to find this single dose for the drug label.

Drugs Need New Formulation

Some drugs are developed with TID (3 times a day) dosing or QID (4 times a day) dosing. In many cases, this dosing schedule can be inconvenient for patients, especially older patients. It is desirable to develop some type of slow-release formulation for these drugs so that a once-daily dosing is available. Doses used in the new formulation can be different from doses used in TID or QID dosing. Oftentimes a new formulation is developed as a new drug, and the dose-selection process will have to be repeated for the new formulation.

Drugs for Special Patient Population

Some drugs are developed for patients with chronic diseases. The population with chronic diseases generally includes older patients. Older patients tend to suffer from multiple diseases. Some of them may have liver or renal problems, in addition to the disease the drug is developed for. Doses used for this population need to be different from doses used for patients without these additional problems. In the development of these drugs, one problem is that not many such patients satisfy trial inclusion criteria, and hence there are only limited data in the entire clinical trial database. In most cases, dosing information for these patients is not available.

In the development of drugs for infectious diseases, one special population of interest is children. Doses used for children should be lower than doses used for adults. Hence, dose–response relationship established from phase II program for adults is not appropriate for pediatric use. Separate development programs are necessary to establish pediatric doses.

DISCUSSION

Dose selection in new drug development is a constant learning process. From nonclinical studies, we learn some preliminary dose information to help design early phase clinical trials. From phase I, we try to establish some PK information and to estimate the MTD. These data are used to design phase II studies. The key dose–response knowledge should be established during phase II. In phase III, we confirm these early findings. In the early stages, estimation is more important than hypothesis testing. We estimate the dose frequency, the efficacy dose–response curve, the toxic doses, and other parameters. In later phases, we use hypothesis testing to confirm these results.

In the phase II dose-selection process, it usually takes several dose–response studies before a final set of doses can be recommended for phase III. In order to speed up this process, some authors suggested sequential designs (Sitter and Wu^[8]) or adaptive designs (Palmer and Rosenberger^[9]) for dose selection. Hence, in an early phase II program, it is very important to choose the best design for dose response. Because dose selection is a critical task in drug development, we should consider planning for several dose–response studies at the beginning of phase II. The first study should cover a wide dose range in a hope that this study will help identify the doses where most of the activities are. The next study will then be designed to capture the dose–response relationship using information obtained from the first study. At early phase II, sequential designs or adaptive designs may not be able to provide the most reliable information to estimate the dose–response relationship.

CONCLUSION

Dose–response studies are critical because if we choose the wrong dose(s), the consequences are very expensive.

If wrong doses are brought into Phase III, all the Phase III studies designed to confirm these incorrect doses are wasted, and the development time line is severely delayed. If a drug is approved with a wrong dose in the label, there would be millions of patients taking the drug at a suboptimal dose; this could create a heavy burden on health care. Choosing the best designs for dose–response studies will help reduce the risk of recommending an incorrect dose for a new drug.

REFERENCES

1. Ruberg, S.J. Dose response studies I. Some design considerations. *J. Biopharm. Stat.* **1995**, 5 (1), 1–14.
2. *ICH-E4 Harmonized Tripartite Guideline*; Dose–Response Information to Support Drug Registration, U.S. FDA: 1994.
3. *ICH-E10 Harmonized Tripartite Guideline*; Choice of Control Group and Related Issues in Clinical Trials, U.S. FDA, 2001.
4. Temple, R.; Ellenberg, S.S. Placebo-controlled trials and active-control trials in the evaluation of new treatments (Part 1). *Ann. Intern. Med.* **2000**, 133 (6), 455–463.
5. Ellenberg, S.S.; Temple, R. Placebo-controlled trials and active-control trials in the evaluation of new treatments (Part 2). *Ann. Intern. Med.* **2000**, 133 (6), 464–470.
6. Wong, W.K.; Lachenbruch, P.A. Tutorial in biostatistics: Designing studies for dose response. *Stat. Med.* **1996**, 15, 343–359.
7. Pukelsheim, F. *Optimal Design of Experiments*; John Wiley & Sons: New York, 1993.
8. Sitter, R.R.; Wu, C.F.J. Two-stage design of quantal response studies. *Biometrics.* **1999**, 55 (2), 396–402.
9. Palmer, C.R.; Rosenberger, W.F. Ethics and practice: Alternative designs for Phase II randomized clinical trials. *Control. Clin. Trials.* **1999**, 20, 172–186.

Dose-Ranging Studies: Design Considerations

Qiqi Deng and Naitee Ting

Department of Biostatistics and Data Sciences, Boehringer-Ingelheim Pharmaceuticals, Inc., Ridgefield, Connecticut, U.S.A.

Abstract

In the process of drug discovery and drug development, understanding the dose response relationship is one of the most challenging tasks. It is also critical to identify the right range of doses in early stages of clinical development so that Phase III trials can be designed to confirm some doses within this dose range. Usually at the beginning of Phase II, there is not a lot of available information to help guiding the study design. At this stage, Phase II clinical studies are needed to establish proof of concept (PoC), to identify a set of potentially effective and safe doses, and to estimate dose-response relationships. Challenges in designing these studies include: selection of the dose frequency and the dose range, choice of clinical endpoints or biomarkers, and use of control(s), among others. Consequences of bad Phase II study designs may lead to the delay of the entire clinical development program or the waste of R&D investment. Misleading results obtained from poor designs could cause a Phase III program to confirm a wrong set of doses, or to stop developing a potentially useful drug. Therefore, it is critical to consider an entire drug development plan, to make best use of all the available information, and to include all relevant experts in designing Phase II dose response clinical trials. This presentation discusses some of these considerations.

Keywords: Clinical trial; Design; Dose ranging; Dose response; Dose spacing; Phase II.

1 BACKGROUND

Typically, Phase I clinical trials are designed to evaluate escalating doses in order to find the maximum tolerable dose (MTD) so as to establish the upper anchor of doses for future development of new medicinal products. After Phase I, one of the most important objectives for Phase II clinical development is to answer the “Go/No Go” question – proof of concept. For new drugs treating non-life-threatening diseases, the next important objective is to establish the “dose range.” A range of doses can be identified by the high dose and the low dose, and the high dose is limited by MTD. As such, a key Phase II objective is to find a low dose of this dose range. Usually this low dose is considered as minimum effective dose (MinED). Simply put, Phase I is moving doses upward to find MTD; Phase II is moving doses downward to find MinED.

After the concept is proven, the most important next step is to explore and recommend the commercial doses to be tested in large phase III confirmatory clinical trials. PoC is usually faster and cheaper as it only requires a well-tolerated dose of test therapy plus the placebo (control) group. A dose-ranging study, however, needs multiple doses of the test therapy to characterize the dose-response relationship. Therefore, a classical phase II development program usually consists of small scale PoC trials followed by moderately sized dose-ranging studies.

Conceptually, the efficacy of a candidate product increases as the dose increases (Fig. 1). The left curve plotted in Fig. 1 can be viewed as an efficacy–response curve, with the response (y-axis) representing some measure of product efficacy. In addition to efficacy response, toxicity dose–response curves should be considered. We will assume, as shown in Fig. 1, when dose increases, both efficacy and toxicity increase. Based on these curves, the maximum effective dose (MaxED) and the MTD can be found: MaxED is the dose above which there is no clinically significant increase in pharmacological effect or efficacy; MTD is the maximal dose acceptably tolerated by a particular patient population. Ideally speaking, it is desirable that MTD is much higher than MaxED. In this case, any dose that is at or above MaxED, but much lower than MTD can be recommended as a labelled dose. Unfortunately, this is not the common case in most of drug development programs. When MaxED is above MTD, and that if the MTD was correctly estimated, the decreased dose from MTD usually lead to decreased treatment effect on efficacy, but also likely better tolerability, then the recommendation for phase III doses need to be based on a careful evaluation of benefit and risk.

In certain products, the efficacy curve and the toxicity curve are widely separated. When this is the case, there is a wide range of doses for patients to take; that is, as long as a patient receives a dose that is greater than

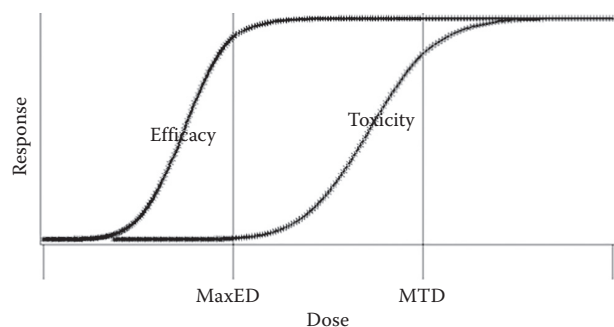


Fig. 1 Efficacy and toxicity dose response curves

the minimum effective dose, and below a toxic level, the patient can benefit from the efficacy while avoiding certain toxicities from the test product. However, for other products, the two curves may be very close to each other. Under this situation, physicians have to dose patients very carefully so that while benefitting in efficacy, patients do not have to be exposed to potential toxicity from the candidate product.

The area between the efficacy and the toxicity curve is known as the “therapeutic window.” One way to measure the therapeutic window is to use the “therapeutic index (TI).” TI is thought of as the ratio of TD_{50}/ED_{50} , where TD_{50} is the dose that produces 50% of maximum toxicity and ED_{50} is the efficacious dose that produces 50% of maximum treatment effect. Clearly, a product with a wide therapeutic window (or a high TI) tends to be preferred by both physicians and patients. We want the drug to have sufficient efficacy with the smallest dose in order to lessen or minimize side effects (even when the dose is on the low side). If a test product has a narrow therapeutic window, then it will need to be developed carefully. Caution is needed should the medicine be approved and prescribed in the target population.

Frequently, products developed for life-threatening diseases (e.g., cancer) tend to have narrow therapeutic windows. In contrast, products developed to treat chronic diseases (e.g., high blood pressure, depression, high cholesterol, or arthritis) are required to have wide therapeutic windows, because these products are frequently used in older patients with other concurrent diseases, and sometimes used in children.

In clinical development, Phase III trials are designed to confirm product efficacy. Results obtained from Phase III studies provide evidence to help regulatory agencies make decisions to approve or not to approve the candidate product. Therefore, it is critical that dose-ranging trials from Phase II deliver the information to help with Phase III design. In other words, if the appropriate dose or doses for product registration is not included in Phase III trials, it is likely that the compound will miss the opportunity of being approved for marketing. Hence information obtained from Phase II dose-ranging trials need to provide sufficient support for dose selection in Phase III study designs.

2 FINDING MINIMUM EFFECTIVE DOSE (MINED)

Minimum effective dose (MinED) is a crucial concept in product development. Unfortunately, there is no universally accepted definition of this term (Thomas and Ting, 2009).^[1] Finding MinED is critical because it is generally believed that toxicity increases as dose increases. Hence a lower dose that is effective implies it could be safer than higher doses.

Clinical results obtained from Phase III studies are used to support product registration. Based on these data, if a dose is efficacious and safe, then that dose can be approved, and the product label specifies the dose or the range of doses the product is approved for. Nevertheless, after approval, if excessive adverse events are observed from the low dose of the product, then a question as to whether this low dose is low enough could be raised. At this time, it would be very challenging to ask what is the MinED for this product. Therefore, it is critical to find MinED in Phase II, before it is too late.

Drug price in US is not directly associated with dosage. However, in many other countries, the price of a medicinal product is linked to its dose. In other words, if the dose of the product is higher, then it is more expensive. Now suppose a certain product is approved in such a country, and the price is set based on the lowest approved dose. If later it was detected that the MinED for the product is, in fact, lower than the approved dose, then a dose reduction is needed at a post-market stage. Accordingly, the price of this product at a lower dose may have to be further lowered as well. When this happens, there is a huge impact on the earnings from such a product. Therefore, from a marketing strategy point of view, it would be very critical to identify the MinED before the product enters its Phase III development.

MinED Finding is performed in the practice of Phase II trial. As discussed previously, if MinED was not identified at Phase II, it would be very difficult and very expensive to find the MinED at a later stage of product development. In fact, Phase II studies may be the only opportunity to characterize the relationship between doses and product efficacy. Therefore, it is vital to pay more attention to Phase II and to carefully design dose-ranging clinical trials during this critical phase of clinical development for a new product.

Although there is a definition of MinED from ICH E4 (1996)^[2] – “MinED is the **smallest dose with a discernible useful effect**”, the interpretation and implementation of such a definition varies a lot in practice. The first question is the understanding of “a useful effect.” It is difficult to achieve consensus about the smallest magnitude of an effect that is useful. One approach is to follow a well-established treatment effect such as lowering 5 mmHg of blood pressure in developing an antihypertensive product. Another is to consider the concept of minimal clinically

important difference (Farrar et al 2001,^[3] Katz et al, 2015,^[4] Osoba et al, 1998,^[5] and Rosen et al. European Urology 2011^[6]), the magnitude of the treatment effect above and beyond placebo effect is large enough to be perceived by a subject.

The implementation of a “discernible” effect could be even more difficult. Typically a “discernible” effect could be referred as a “statistically significant” difference from control. Under this interpretation, various authors recommend different ways of calling a dose as MinED based on the difference in treatment effect from a given dose as compared with placebo. Some researchers suggest the use of a dose-response model to help estimate MinED (Fillon, 1995^[7]; Pinheiro, et al, 2006^[8,9]). In the modeling approach, a treatment difference (δ) against placebo is proposed and from the model, if there is a dose achieves such a δ , then this dose is considered as the MinED.

Others (Hochberg and Tamhane, 1987,^[10] Wang and Ting, 2012^[11]) suggest the use of pairwise comparisons. Under pairwise comparisons of each dose vs placebo, a point estimate and an interval estimate can be calculated. Another commonly considered approach is William’s test based on isotonic regression (William, 1971,^[12] 1972^[13]). In William’s test, the mean of each dose group will be estimated under the monotonic order restriction. And the lowest dose, from where on the estimated effect at or above are all statistically significant from placebo, is considered as the MinED. Depending on the definition of MinED, these estimates are used to help proposing a dose as MinED. Even among those authors recommending use of pairwise comparisons to find MinED, there could be many differences, too. For example, some may argue that multiple comparison adjustment could be necessary, others may not. Whether or not multiple comparisons are needed, the understanding of “a dose that delivers treatment effect which is different from placebo” could still vary substantially, depending on individual’s point of view.

For example, let the useful effect be denoted as δ ; then, for a given dose, a point estimate and a confidence interval can be constructed for δ . If a positive value indicates a benefit effect, then the treatment response of the given dose of the test product subtracted from the placebo response is expected to be positive. Hence it is hoped that the point estimate of the treatment difference could be as much as δ . This leaves a wide range of interpretations – for MinED, should that dose be such that the point estimate is greater than δ ? Should that dose be such that the lower confidence limit be greater than 0? Should that dose be such that the upper confidence limit be greater than δ ? Or should that dose be such that the lower confidence limit be greater than δ (McLeod et al. 2016^[14])?

The implementation of “the **smallest dose with a discernible useful effect**” is fraught with various interpretations and complexity, making the definition difficult to implement unambiguously. In practice, one feasible way of finding a MinED in a dose-ranging clinical trial

is to find a dose that is lower than MinED. Regardless of which practical definition the project team uses to identify MinED, suppose there is a dose deemed to be “not efficacious,” and another dose that is higher could be considered to be “efficacious.” Then it can be considered that the non-efficacious dose is below MinED and the efficacious dose is above MinED. From a practical point of view, this information would usually be sufficient to help the team to design Phase III clinical trials.

Of course in the selection of Phase III doses, many other aspects will also have to be considered—such as safety or toxicity findings, pharmacokinetic properties, formulations, and so on. But the information about MinED would be one of the most critical deliverables from Phase II in supporting any Phase III development plans.

3 A MOTIVATING EXAMPLE

In a clinical development program for a test drug to treat osteoarthritis, the first dose response study included placebo, 80 mg, 120 mg, and 160 mg of the test drug (Ting, 2009^[15]). Results from this study indicate that all three doses are efficacious (Fig. 2a). These results also show that 80 mg may have already been at the high end of the efficacy dose-response curve. A second study was designed to explore the 40 mg dose, and results from this second study are given in the middle panel of the Fig. 2 (Fig. 2b). With these findings, the project team started Phase III studies using a dose range of 40 mg to 120 mg to confirm the long-term efficacy of this test drug.

An End-of-Phase II meeting with the Food and Drug Administration (FDA) was held after the long-term Phase III study started. FDA commented that the minimum effective dose (MinED) was not yet found. Hence the project team designed a third dose-response study to explore the lower dose range. Doses included in the third study are 2.5 mg, 10 mg, and 40 mg. As shown in the bottom panel of the Fig. 2 (Fig. 2c), the third study successfully established the dose-response relationship. Based on these results, it is clear that the Phase III studies were designed with a range of doses that are too high. The impact of this process can cause a major delay in the development program and considerable waste of resources.

4 HOW WIDE A RANGE OF DOSES TO STUDY?

The key lesson learned from this osteoarthritis drug development program is that a very low dose was not explored in early Phase II. In other words, if a dose at or below 10 mg was explored in the first study, a narrower dose range could have been established using a second study. Then the Phase III studies could have been designed with the appropriate dose range.

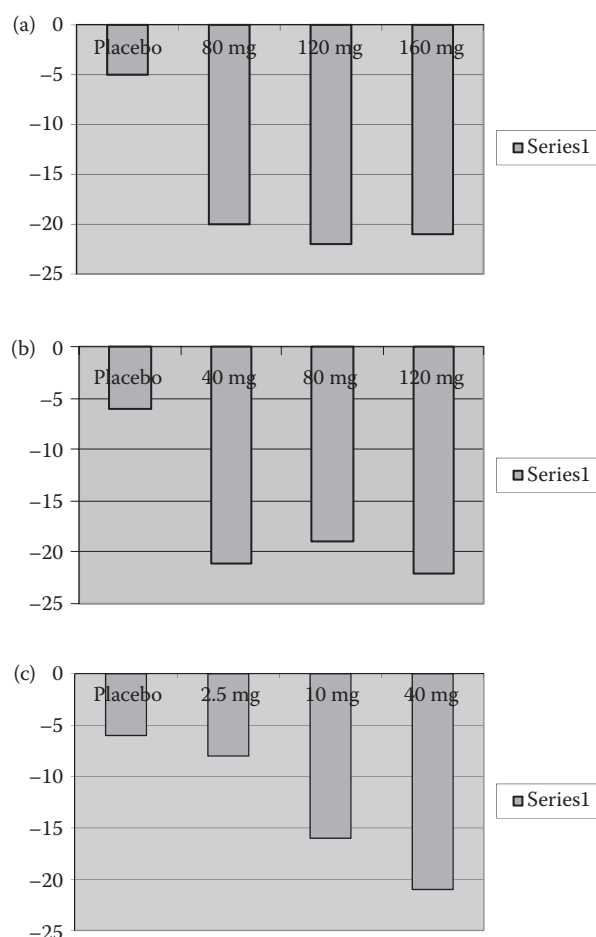


Fig. 2 Dose Response Results from the Three Studies of Osteoarthritis Drug

In general, the most difficult challenge in designing the first dose ranging trial in developing a new intervention is the selection of doses to be included in this design. The dose-ranging study is designed with two very important assumptions:

1. The maximal tolerated dose (MTD) was correctly estimated from earlier trials and
2. Efficacy response is monotonic to ascending doses.

However, it is unclear what range of doses to choose for efficacy activities. At this stage, the best guidance available is data obtained from pre-clinical experiments or from Phase I biomarkers. Even under the assumptions that the MTD was correctly estimated and there is a monotonic efficacy dose-response relationship, there can still be many possible shapes of dose-response curves.

Figure 3 presents an example of three dose-response curves. The three curves on this figure represent dose projections based on three animal models from non-clinical studies. One curve represents a dog model, another is a mouse model, and the third is from a rabbit model. The vertical dashed grey line (at the right side of Fig. 3)

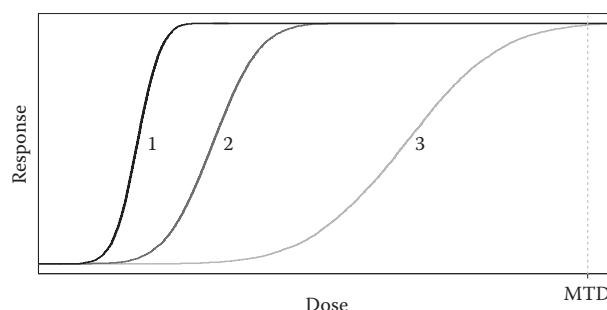


Fig. 3 Several Possible Dose Response Curves

represents the maximum tolerated dose (MTD). The question then becomes, what range of doses should be considered in the upcoming dose-ranging clinical trial? From this perspective, the primary challenge in the dose-ranging trial design is to consider which range of doses needs to be studied.

4.1 Definition of Dose Range in a given study

Dose range in a clinical trial design is defined as the ratio calculated as the highest dose divided by the lowest dose included in the same design. For example, if the doses in trial A are placebo, 20 mg, 40 mg, and 80 mg, then the dose range is 4 ($= 80/20$). If the doses in trial B are placebo, 0.1 mg, 1 mg and 10 mg, then the dose range is 100 ($= 10/0.1$). Although the doses used in trial A are higher than doses used in trial B, trial B has in fact a much wider dose range – 25 times wider than trial A.

Given the discussion in the previous sections about MinED, it is critical to realize that in the first dose-ranging design, a wide dose range need to be studied. The high dose is typically set at MTD or a dose below MTD to avoid any potential toxicity. The real design challenge is to choose the lowest dose for this dose-ranging trial. If there is an interest to identify a dose that is below MinED, then the difficult question would be “how low would it be low enough?” Once the low dose is selected for the design, a dose range can be calculated as the ratio of the highest dose divided by the lowest dose.

4.2 Binary dose spacing

Hamlett *et al* (2002)^[16] propose to use a binary dose spacing (BDS) design for dose allocation. If the trial includes 2 dosing groups and a placebo, then the BDS design identifies a mid-point between the placebo and the MTD and allocates one dose above and another below the midpoint. If the trial includes 3 dosing groups and a placebo, then the BDS design identifies the high dose as described in the 2 dose case. A second midpoint is then identified between the placebo and the first midpoint, and a low dose is selected below the second midpoint and the medium dose is selected between the two mid-points. The BDS design

iteratively follows this strategy with more dosing groups. The BDS design employs a wide dose range, helps identify the MinED, employs a log-like dose spacing strategy avoiding allocating doses near the MTD (i.e., identifies more doses near the lower end), and is flexible and easy to implement.

5 PARALLEL CONTROLLED FIXED DOSE DESIGNS

Typical dose ranging trials are designed with parallel dose groups plus a placebo control. Patients are randomized to each treatment group and receive study medications for the entire duration of the study. For example, in a four treatment group design with four weeks duration, the doses are placebo, low dose, medium dose and high dose, then a patient is randomized into one of these four groups and will receive the same dosage of study medication from after randomization to week 4. A parallel design indicates that at the time a patient is recruited, the opportunity of randomizing into each dosing group is dictated by the study design, and that there is no requirement for any study dose to be before, or after another study dose.

Dose escalation designs are frequently used in Phase I. In a dose-escalation design, a group of subjects started with the lowest designed dose. After these subjects complete the lowest dose and there is no safety concern, then a different group of subjects will receive a higher dose. If this dose is well tolerated, then a third group of subjects will receive an even higher dose. This process continues until the stopping rule is met. A typical stopping rule would be some kind of pre-specified adverse events. In oncology, the stopping rule could include a dose limiting toxicity (DLT). In other words, if several subjects experiencing DLT's then it can be claimed that the MTD was exceeded. Dose escalation designs are very useful in finding maximally tolerable dose (MTD). However, because the duration of each treatment period is relatively short, these designs are not used in clinical efficacy trials. Phase II dose ranging trials are designed to study drug efficacy and, in most of therapeutic areas, drug efficacy can only be observed after a reasonable treatment period. Hence dose escalation designs are not used in dose-ranging trials.

A fixed-dose design and a dose-titration design are in contrast to each other. In a fixed-dose design, once a patient is randomized to a dose group, the patient would take the same dose of study product throughout the complete dosing period. In a dose-titration design, in contrast, a patient is randomized to a dose regimen with a starting dose, where the dose for a patient can be changed over time. In a titration design, patients are randomized to regimens, instead of doses. The regimen could be a 10 mg–20 mg test drug vs placebo. In this case, a patient could be randomized to test drug or placebo. If the patient is randomized to test drug, he or she would be dosed at 10 mg or 20 mg

of test drug, depending on study design. If a patient is randomized to placebo, then this patient may receive 10 mg matching placebo, or 20 mg matching placebo, according to the double-blind study design.

Dose-titration designs include forced titration, and response guided titration. For example, in an eight-week study with forced titration of test drug against placebo, with 10 mg and 20 mg dosages, then the design may dictate that every patient starts with 10 mg of test drug or matching placebo. Then after two weeks of 10 mg (or placebo) double-blind treatment, every patient will be titrated up to 20 mg, until end of treatment (week eight). On the other hand, a response guided titration design (again, two treatment groups with test drug against placebo, and eight weeks duration) would randomize patients into one of the two treatment groups, and every patient starts with 10 mg (or matching placebo). Then, depending on patient response, if the efficacy is not sufficient, then the treating physician may consider titrate the dose up to 20 mg. However, after the patient on the 20 mg dose, if there is safety concern at this dose, then the physician may down titrate the dose back to 10 mg. Hence dose-titration designs can be viewed as two different types – forced titration or response guided titration.

There are some advantages of a titration design. For example, a study with this design will allow a patient to be treated at the best dose for that specific patient; this dose allocation feature reflects the actual medical practice. However, the disadvantage of a titration design is the difficulty in data analysis. For example, if a patient responded to the test drug after doses are titrated up, it is unclear whether the higher dose delivered the efficacy or the accumulation of lower doses caused the response. In titration designs with multiple treatment groups, there may be overlapping doses as when, for example, one treatment group is 10 mg escalating to 20 mg, while another group is 20 mg escalating to 40 mg. When this is the case, it is difficult to make inferences about the 20 mg dose group.

In some rare cases, instead of a dose-response study, a concentration response study could be considered. A concentration response study assesses efficacy and safety measurements observed from subjects according to the plasma concentration of the study drug but not the doses of the study drug. There are many practical limitations in using this type of designs. These limitations include, among others, how to blind the patient and the physician and also how and when to measure the blood concentration. Because of the issues with dose-titration designs and concentration response designs, the parallel, fixed dose designs are, in general, the more commonly used designs for dose-ranging studies. Therefore, in many of the dose-response studies, patients are randomized to a few fixed-dose groups and are compared with one or more control treatment groups.

In Phase II dose-ranging clinical trials, the primary objective is to evaluate drug efficacy of each studied dose, and hence the most useful design is the parallel, controlled,

fixed dose design. In these designs, each patient is randomized to a pre-specified dose and followed for the designed study duration. During the entire study period, patients received the same dosage of the double-blind medication. Results obtained from these studies can be used to help understanding drug efficacy associated with each studied dose. It is important to note that dose-escalation design, dose-titration design, and concentration response design, cannot be used for efficacy dose-response purposes.

6 NUMBER OF DOSES AND CONTROL GROUPS

As discussed in the previous sections, a wide dose range is critical in the phase II dose-ranging studies. Once MTD is estimated from earlier studies and a wide dose range is considered, the natural questions to think about is how many doses should be included in the study and whether a placebo control group is sufficient or an active control group is needed.

A single high dose plus a placebo arm design may be able to demonstrate the proof-of-concept but cannot adequately characterize the dose-response relationship. Any attempt to interpolate a dose-response relationship between the placebo and the test dose would require very strong assumptions, which often times are not realistic. Therefore, a dose-ranging study typically needs several test dose levels. Generally speaking, it would be desirable to have more than two test doses plus one placebo arm in defining the dose-response relationship and estimating the difference of the test doses vs. placebo. One very commonly used design is a four-group study with placebo, low-dose, medium-dose and high-dose groups of the drug candidate under development.

Some authors suggest that more doses could help (Krams et al 2003)^[17]—that is, in the first dose-ranging study, adding more doses to the study design. However, the number of doses to be used in a dose-ranging trial is usually limited by practical and logistical considerations, such as available formulations of test doses, dosing frequencies, convenience for outpatients to use on their own, and blinding complexity. In some situations, an active control is also employed in a dose-ranging trial. Given the multiple treatment groups already included in a study, it may not be practical to add very many test doses into the same study.

In an early Phase II dose-ranging trial design, typically the total sample size is limited depending on budget, ethical concerns, availability of patients, and difficulty in recruitment. Under this circumstance, more treatment groups in a design imply fewer patients per treatment group. A smaller sample size of each treatment group provides a less precise estimate of treatment effect. Hence merely adding dose groups to a study design does not necessarily make the design to deliver more informative results. A well-thought through design strategy is needed before the number of dose groups can be determined.

In our opinion, for the first-dose ranging study design, it is more important to cover a wide dose range, than simply adding more doses to cover a narrow range of doses. This suggestion is both practical and sound. In practice, a trial with 4 to 5 test doses, plus a placebo control (a total of 5 or 6 treatment groups), will deliver a good understanding of where the test medication is most active, if the dose range is wide enough and the dose spacing is reasonable. Some simulation studies (Yuan and Ting 2014^[18]) suggest that, among the number of treatment arms (4, 5, 6, or 7, including placebo), it looks like three test doses (the 4-arm design) could be insufficient in some cases. The performance increases when more than three doses are studied. Similar conclusion was made in PhRMA adaptive dose-finding working group white paper (Bornkamp et al 2007^[19]). Given the same sample size, although four-arm studies with three active doses and placebo give more information for each pairwise comparison in secondary analysis, the power for the primary analysis based on contrast test provided in section 7 are usually comparable. In order to provide adequate coverage with reasonable spacing for a wide dose range, three-arm study is discouraged for dose-finding trials, and 4–5 doses are recommended.

On the other hand, six test doses (the 7-arm design) may not necessarily deliver better results than the four test doses (the 5-arm design) or five test doses (the 6-arm design) comparisons. When the total sample size is fixed, the 7-arm design offers a smaller sample size per group when equal sample size is used, and the precision on the treatment effect would be sacrificed. In this case, it may be a good idea to consider an unbalanced sample size allocation, for example by allocating more subjects to highest dose and placebo arm. Therefore, from a practical point of view, a dose-ranging design including four to five test doses (in addition to placebo), which cover a wide dose range, may be very useful in designing the first-dose ranging clinical trial.

7 SAMPLE SIZE CONSIDERATIONS

Whenever sample size calculation is needed for a given study, the first point is to understand the hypothesis to be tested. In the design of a Phase II dose-ranging clinical trial, the statistical hypothesis associated with such a trial is typically a proof-of-concept (PoC) hypothesis. In the case when a Phase II PoC study designed with only two treatment groups, a placebo group is compared with the MTD of study product, the hypothesis can be expressed in a straightforward fashion – responses from the test treatment is compared against responses from placebo. If the comparison is based on treatment means, then the hypothesis can be written as

$$H_0: \mu_T = \mu_P \text{ vs } H_1: \mu_T > \mu_P,$$

where μ_T is the test treatment group mean and μ_P is the placebo treatment group mean.

In other words, if mean responses from the test treatment group is statistically significantly improved than the mean responses from the placebo group, then the concept is considered proven.

However, in a dose-ranging trial design, the primary hypothesis is not straightforward. In this case, the trial includes more than two treatment groups. A typical dose-ranging design consists of the highest dose of the test product (could be the MTD) on the upper end, and of the zero-dose group in the form of placebo. Additionally, such a design includes one or more different doses in between these two extreme treatment groups. Under this circumstance, two possible options could be considered in specifying the primary hypothesis. One option is to keep the hypothesis the same as that in the two-group setting – namely, only test the highest dose (may be the MTD) against placebo.

Another option is to use a contrast which includes more than two treatment groups in order to perform the primary hypothesis test. If the second option is selected, then there could be many varieties in writing such a contrast.

For example, in a four-group design with placebo, low dose, medium dose and high dose, let μ_L be the mean response of low dose, μ_M be the mean response of the medium dose, and μ_H be the mean response of high dose. The contrast can be a trend test based on all doses (Wang and Ting, 2012^[11]; Ting, 2009^[15]) assuming monotonic dose-response relationship:

$$H_0: -3\mu_P - \mu_L + \mu_M + 3\mu_H = 0 \text{ vs } H_0: -3\mu_P - \mu_L + \mu_M + 3\mu_H > 0$$

Another possible proof-of-concept contrast may also include (which assumes all test doses are equally effective) is

$$H_0: -3\mu_P + \mu_L + \mu_M + \mu_H = 0 \text{ vs } H_0: -3\mu_P + \mu_L + \mu_M + \mu_H > 0$$

Alternatively, it can be assumed that only the high dose is effective:

$$H_0: -\mu_P - \mu_L - \mu_M + 3\mu_H = 0 \text{ vs } H_0: -\mu_P - \mu_L - \mu_M + 3\mu_H > 0$$

As can be found from these sets of hypotheses, if a contrast (which includes more than two treatment groups) is used for the statistical hypothesis, there could be many possible ways of writing such a contrast. As long as the contrast coefficients sum up to zero, (e.g., in the first example, $-3 - 1 + 1 + 3 = 0$), the rejection from the contrast test mentioned above will imply the rejection of the null hypothesis that all treatment arms are equal.

Once the statistical hypothesis, the clinical difference (after adjusted for standard deviation), and the corresponding alpha (α) and beta (β) have been determined, sample size for a given dose-ranging trial can be calculated.

Based on the above discussion, the sample size for a given contrast test can be calculated using the following formula (Chang and Chow, 2006^[20]).

$$n = \left[\frac{(z_{1-\alpha} + z_{1-\beta})\sigma}{\delta} \right]^2 \sum_{i=0}^k \frac{c_i^2}{f_i}$$

where

n is the total sample size of all treatment groups combined,

$z_{1-\alpha}$ and $z_{1-\beta}$ are the critical values of the standard normal distribution corresponding to Type I and Type II error, respectively,

δ is the treatment difference between test product and placebo,

σ is the standard deviation,

f_i is the fraction of number of patients in treatment group i ,

c_i is the contrast coefficient that is associated with treatment group i , it is assumed to have k dose groups plus placebo, where i is denoted as the 0 dose group.

In the calculation of sample sizes for a dose-ranging study, two popular types of patient allocation are possible: (1) balanced allocation where each treatment group receives an equal number of patients and (2) imbalanced allocation where the highest dose and the placebo groups receive more patients and other dose groups receives fewer patients per group. The sample size formula given here works well with either type of patient allocation. Examples of these types of allocation can be found in Yuan and Ting (2014),^[18] as well as Deng and Ting (2016).^[21] The above formula applies to cases where the primary endpoint is a continuous variable. In fact, sample size formula based on other types of response variables such as binary or time-to-event can also be found in Chang and Chow (2006).^[20]

Contrast test can be viewed as a test on the weighted average of all treatment arms, where the weight (i.e., the contrast coefficient) can be optimized for a certain known dose-response curve to maximize the power. However, a set of weights that is optimal for one dose-response curve may perform poorly for another. At the design stage of the first dose-ranging study, very little is known for the dose response shape and therefore it is important that the method for PoC is robust to model uncertainty. There are two robust approaches can be suggested to consider.

7.1 Ordinal Linear Contrast Test (OLCT)

An OLCT is expressed in terms of an ordinal linear contrast under a statistical null hypothesis (Wang and Ting 2012, Deng and Ting 2016). For trials including placebo and three doses of test drug, the contrast coefficients chosen are $(-3, -1, 1, 3)$. For trials with different number of treatment groups, the associated contrast coefficients are provided in Table 1. Since OLCT combines all the data

Table 1 Coefficients to be used in contrast for OLCT

Number of Doses plus Placebo	Coefficients						
	Placebo	Lowest Dose	Doses increase from left to right				Highest Dose
Two Doses	−1	0					1
Three Doses	−3	−1	1				3
Four Doses	−2	−1	0	1			2
Five Doses	−5	−3	−1	1	3		5
Six Doses	−3	−2	−1	0	1	2	3

available, and allocates the entire alpha to a single contrast test, it is in general a powerful test. In addition, OLCT can be viewed as a model-free test based on ranks. The weight for each arm is set to be proportional to the order of the expected responses for the different treatment groups. This method is robust to various dose-response shapes under monotonic assumption and also mild non-monotonic curves. For binary response, OLCT is very similar to the Cochran-Armitage trend test (Neuhauser and Hothorn 1999^[22]). It is important that an equally-spaced ordinal score should be used instead of the actual dose.

7.2 MCPMod

MCPMod stands for Multiple comparison procedures and modeling (Pinheiro et al 2006^[9]). The procedure starts with a specification of a set of different candidate models that covers a wide range of dose-response shapes. For each candidate model, the optimal contrast can be calculated. Since there will be a number of contrast tests, one for each candidate model, the multiplicity will be subsequently adjusted using multivariate *T* distribution. This multiplicity adjustment takes into account of the similarities between candidate models, which is reflected by correlation of contrast coefficient. Therefore, when similar dose response models are included, the multiplicity penalty is small. MCPMod will reject the null hypothesis of a flat dose response curve when at least one of the candidate models is significant after multiplicity adjustment. Extensions of MCPMod beyond normally distributed endpoints are available, including binary, count, longitudinal and time-to-event data (Pinheiro et al., 2014^[23]).

Under MCPMod, doses and patient allocation to different arms can be optimized according to different criteria: 1) D-optimal design minimizes the variance of the model parameters and 2) TD-optimal design minimizes the variance of estimation for target dose (Dette et al 2008^[24]). For more details regarding MCPMod, readers are encouraged to refer to the chapter “Dose Response Clinical Trials – MCPMod Approach” in this book.

In general, the contrast based approach including OLCT and MCPMod provide greater power advantages comparing to pairwise approach when the number of doses increase. Details of sample size calculation based on other

methods, for example William’s test for Minimum effective dose, are provided in Chang and Chow (2006,^[20]), and Chow et al (2003,^[25]).

8 DISCUSSION

There are many challenges in designing the first dose-ranging clinical trial. Two main problems are apparent. First, there is not much of information about the efficacy of the candidate product available. Second, too many answers are expected for this trial to deliver. Project team members tend to anticipate such a study will establish the PoC, identify the MinED, characterize the dose-response relationship, and recommend doses for Phase III designs. Fundamental statistical training points out that the best design is that one study answers one question. Hence the nature of a dose-ranging trial makes it a very difficult design in order to address many scientific questions, and with sparse information available before the study design stage.

In fact, there should be a logic sequence of thinking in addressing these relevant questions. The first step is to consider how wide a dose range should be – which is the most critical step. As a way to approach this question, information about the candidate under development, other marketed drugs in similar class, the indications this study is designed for, the target product profile, and characteristics of the primary endpoint, as well as other considerations, can all contribute to help with a better understanding of the potential dose range need to be explored. In practice, it is also important to have a good understanding of dosing frequency (from Phase I results), formulation (from manufacturing), and other relevant issues.

After the dose range is determined, the next step would be to consider the number of treatment groups. Will there be a need of an active control? How many doses should be included in this design? What dose spacing should be considered? Basically speaking, a trial with 4 to 5 test doses, plus a placebo control, will deliver a good understanding of where the test medication is most active, if the dose range is wide enough, and the dose spacing is reasonable. Extensive simulations indicate that six or more test doses (in addition to placebo) design may not

necessarily deliver better results than the four test doses or five test doses (Yuan and Ting, 2014^[18]). A study design with too many doses may introduce additional complexities from practical and logistical points of view (e.g., formulation, blinding).

Sample size calculation is needed for almost every clinical trial. Considerations in sample size calculations for a dose-ranging trial are very similar to those for other trials. However, because the nature of a dose-ranging trial includes multiple treatment groups, this feature may cause some level of confusion, both to statisticians and to non-statisticians. Nonetheless, if the decision is to only use a single degree of freedom contrast test as the primary comparison, then the situation can be simplified and addressed successfully. In the case of one contrast, the two major challenges would be what assumptions of dose-response relationship the team is willing to make and sample size allocation, either a balanced approach (an equal number of patients per group) or an imbalanced approach (an unequal number of patients per group). After these decisions are made, the statistician would then follow up and calculate the sample size for the study.

Suppose a contrast with one degree of freedom is selected and the sample size allocation is determined, the typical sample size considerations such as choice of treatment difference and choices of alpha and beta can then be specified. With this information, sample size can be calculated using existing formula. The sample size using MCPMod approach can also be calculated using contrast test, with an adjustment for multiplicity introduced by different candidate models.

Finally, although the sample size is usually selected based on PoC part of dose-ranging studies, it is worth noting that estimating dose response, or identifying appropriate dose for phase III is considerably more difficult than proof-of-concept test (Bornkamp et al 2007^[21]). Therefore, dose-ranging study should include an assessment on the precision of target dose estimate. Carrying more than one dose into confirmatory phase may be advised to ensure a reasonable chance of covering an appropriate dose range in phase III.

REFERENCES

1. Thomas, N.; Ting, N. Minimum Effective Dose. *Encyclopedia of Clinical Trials*, Wiley-Blackwell: Hoboken, NJ, 2009.
2. ICH-E4 Harmonized Tripartite Guideline. Dose-response information to support drug registration, 1996.
3. Farrar, J.T.; Young, J.P.; LaMoreaux, L.; Werth, J.L.; Poole, R.M. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain* **2001**, *94* (2) 149–158.
4. Katz, N.P.; Paillard, F.C.; Ekman, E. Determining the clinical importance of treatment benefits for interventions for painful orthopedic conditions. *Journal of Orthopaedic Surgery and Research* **2015**, *10*, 24.
5. Osoba, D.; Rodrigues, G.; Myles, J.; Zee, B.; Poter, J. Interpreting the significance of changes in health-related quality-of-life scores. *Journal of Clinical Oncology* **1998**, *16* (1), 139–144.
6. Rosen, R.C.; Allen, K.R.; Ni, X.; Arujo, A.B. Minimal clinically important differences in the erectile function domain of the international index of erectile function scale. *European Urology* **2011**, *60* (5), 1010–1016.
7. Filloon, Thomas G. Estimating the minimum therapeutically effective dose of a compound via regression modeling and percentile estimation (Disc: p933-933). *Statistics in Medicine* **1995**, *14*, 925–932.
8. Pinheiro, J.C.; Bretz, F.; Branson, M. Analysis of dose-response studies – modeling approaches, *Dose Finding in Drug Development*, Springer, New York, 2006, 146–171.
9. Pinheiro, J.C.; Bornkamp, B.; Bretz, F. Design and analysis of dose-finding studies combining multiple comparisons and modeling procedures. *Journal of Biopharmaceutical Statistics* **2006**, *16* (5), 639–656.
10. Hochberg, Y.; Tamhane, A. *Multiple Comparison Procedures*, New York: Wiley, 1987.
11. Wang, X.; Ting, N. A Proof-of-concept Clinical Trial Design Combined with Dose- Ranging Exploration. *Biopharmaceutical Statistics*, **2012**. wileyonlinelibrary.com doi: 10.1002/pst.1525.
12. Williams, D.A. A test for difference between treatment means when several dose levels are compared with a zero dose control. *Biometrics* **1971**, *27*, 103–117.
13. Williams, D.A. Comparison of several dose levels with a zero dose control. *Biometrics* **1972**, *28*, 519–531.
14. McLeod, L.D.; Cappelleri, J.C.; Hays, R.D. Best (but oft-forgotten) practices: expressing and interpreting associations and effect sizes in clinical outcome assessments. *American Journal of Clinical Nutrition*. **2016**, *103* (3), 685–693.
15. Ting, N. Practical and Statistical Considerations in Designing an Early Phase II Osteoarthritis Clinical Trial: a Case Study. *Communications in Statistics – Theory and Methods* **2009**, *38* (18), 3282–3296.
16. Hamlett, A.; Ting, N.; Hanumara, C.; Finman, J.S. Dose Spacing in Early Dose Response Clinical Trial Designs. *Drug Information Journal* **2002**, *36* (4), 855–864.
17. Krams, M.; Lees, K.R.; Hacke, W.; Grieve, A.P.; Orgogozo, J.M.; Ford, G.A., and for the ASTIN Study Investigators. ASTIN: An adaptive dose-response study of UK-279, 276 in acute ischemic stroke. *Stroke* **2003**, *34* (11), 2543–254.
18. Yuan, G.; Ting, N. First Dose Ranging Clinical Trial Design – More Doses? Or a Wider Range? *Clinical Trial Biostatistics and Biopharmaceutical Applications* (2014), Taylor & Francis.
19. Bornkamp, B.; Bretz, F.; Dmitrienko, A.; Enas, G.; Gaydos, B.; Hsu, C.H.; König, F.; Krams, M.; Liu, Q.; Neuenschwander, B.; Parke, T.; Pinheiro, J.; Roy, A.; Sax, R.; Shen, F. Innovative Approaches for Designing and Analyzing Adaptive Dose-Ranging Trials. *Journal of Biopharmaceutical Statistics* **2007**, *17* (6), 965–995.
20. Chang, M.; Chow, S.C. “Power and sample size for dose response studies” *Dose Finding in Drug Development*, Springer: New York, 220–241.

21. Deng, Q.; Ting, N. Sample size allocation in a dose-ranging trial combined with PoC, to appear in *Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*, Springer: New York, 2016.
22. Neuhauser, M.; Hothorn, L.A. An exact Cochran-Armitage test for trend when dose-response shapes are a priori unknown. *Computational Statistics & Data Analysis* **1999**, *30*, 403–412.
23. Pinheiro, J.; Bornkamp, B.; Glimm, E.; Bretz, F. Model-based dose finding under model uncertainty using general parametric models. *Statist. Med.* **2014**, *33*, 1646–1661. doi:10.1002/sim.6052.
24. Dette, H.; Bretz, F.; Pepelyshev, A.; Pinheiro, J. Optimal designs for dose-finding studies. *Journal of the American Statistical Association* **2008**, *103*, 1225–1237.
25. Chow, S.C.; Shao, J.; Wang, H. *Sample Size Calculation in Clinical Research*, Marcel Dekker: New York, 2003.

Dropout

Cindy Rodenberg and Dan Schnell

The Procter & Gamble Co., Cincinnati, Ohio, U.S.A.

Cameron S. Liu

Ortho Biotech Oncology Research & Development, a Unit of Cougar Biotechnology, Los Angeles, California, U.S.A.

Abstract

It is frequently the case in clinical research that subjects drop out of a study before completion, especially in the case of long trials. This entry discusses ways in which to plan studies that will minimize the dropout factor, and reviews methods of compensating for the missing-data problem that subjects dropping out can create in a clinical study.

Keywords: Dropout; Longitudinal studies; Missing data.

INTRODUCTION

In clinical studies, the length of time for a study can vary tremendously, ranging from a few days or weeks to possibly a few years. For example, in an osteoporosis trial, to illustrate efficacy, subjects must be followed for two to three years. During this time, it is often of interest to track subjects periodically in order to characterize their response to treatment throughout the study as well as to obtain subject information in case the subject should discontinue (or, in other words, drop out). Thus, most primary endpoints of interest are collected repeatedly. Studies that follow subjects over periods of time are referred to as “longitudinal studies” and the corresponding data as “longitudinal data.”

It is frequently the case in clinical research that subjects drop out of a study before completion, especially in the case of long trials. Reasons for dropout vary, from relocation to another city, to lack of efficacy, to having a safety concern. For subjects dropping out of the study, responses that were to be captured at time points after which they dropped out are of course lost. Because the objective of most analyses in clinical trials is to estimate or test for differences in treatment effects, the problem then becomes one of estimating or testing for these effects, taking into consideration subject dropout. In this case, there are two sets of variables to be considered in a trial: the clinical response variables (involving the parameters of interest) and the variables describing the dropout process (involving nuisance parameters).

Note: Material presented is based in part on the efforts of a working group at Pfizer Central Research, which produced the technical report “Selected Approaches for the Analysis of Longitudinal Data with Dropouts”.^[1]

OVERVIEW

Dropout is a special case of the missing-data problem. In the case of “dropout” in longitudinal studies, it is assumed that no measurements post-dropout are observed. As a result of this property, dropout is also sometimes viewed as a “monotone” or “truncation” missing data pattern. Although this article focuses on dropout, other patterns of missing data may be present in any given longitudinal study and the statistical methods discussed herein will remain applicable with perhaps some modification or supplementation. Work concerning inference in the presence of incomplete data in more general situations can be found in Rubin^[2] and Little and Rubin.^[3] References dealing specifically with the problem of missing data in longitudinal studies include Laird,^[4] Wu and Carroll,^[5] Wu and Bailey,^[6] Diggle and Kenward,^[7] Little,^[8] Hogan et al.,^[9] and Molenberghs and Kenward.^[10]

Before proceeding to the statistical frameworks and methods for handling dropout, it is important to highlight some broader considerations regarding missing data due to dropout in longitudinal studies. First, it is important to plan and conduct studies so as to minimize the potential for dropout. Components of studies that can have a significant bearing on the frequency of missing data due to dropout and can that be addressed before a study begins are inclusion and exclusion criteria, visit schedules, data capture methods, and the definition of endpoints themselves. In some situations provisions can be made for collection of efficacy or safety endpoints even after a subject has dropped out of the study proper. Some examples are a search of records for vital status or hospitalizations in cardiac studies and follow-up directly with subjects via a telephone call or written request for endpoint information.

Second, pre-specification of the method(s) to be used in the presence of missing data, and sensitivity analysis to assess the potential impact on the results, has become an expectation of regulatory agencies.^[11,12] Third, the ability to obtain unbiased estimates depends upon the analysis method in addition to the underlying dropout process. For example, the observed responses and covariates may theoretically contain all the information necessary to conduct an MAR analysis, but the ability to actually produce unbiased estimates depends upon identifying an analysis methodology/model to actually incorporate that information appropriately.

Lastly, the data that are observed might provide some information about the dropout process and perhaps about the data unobserved due to dropout, but it should not be assumed that such data will provide complete or conclusive information. This point will be revisited in the section “Other Methods.” For example, the observed data may indicate that the probability of dropping out is related to observed responses earlier in the study and thus indicate that the assumption of missing “completely at random” would not be appropriate. However, failure to identify a relationship between observed data and the probability of dropping out should not be taken as conclusive evidence that the dropout process is that of “missing completely at random” or “missing at random.”

In the next section we will explore various relationships between the measurement and dropout processes. Discussion of some general analysis methods currently used in the pharmaceutical industry and the implications of missing data in characterizing the response distribution (for example, estimating treatment effects) will be given in the section “Analysis Populations and Accompanying Methods.” Two parametric methods developed specifically for this problem will be discussed in the section “Other Methods.” The main focus of this entry is on the analysis of continuous data as the outcome variables, with emphasis on a univariate response measurement. Extensions will be mentioned in the “Discussion” section.

THE DROPOUT PROCESS

If all subjects remain in the study, the only process of interest is the distribution of responses, in other words, the measurement process. For example, bone mineral density may be the primary response variable of interest in an osteoporosis trial or glycosylated hemoglobin (HbA_{1c}) in a trial for diabetes. However, when subjects drop out we also need to consider the dropout process and its relation to the measurement process. Unfortunately, many of the current statistical procedures available for analyzing longitudinal data do not suitably incorporate the missing-data process and thus may lead to biased results.

In general, by the “missing-data mechanism,” we mean the distribution of the variable indicating whether each response is observed or missing. In the case of dropout, this sequence of variables remains constant (say, 1, indicating an observation as being observed) until the subject discontinues, at which point it changes to a new value (say, 0, indicating an observation as being unobserved). Therefore, this process can be characterized by the time a subject drops out. The effect this has on analyzing longitudinal data depends on the relationship between the dropout process and the subject response measurements. The reasons for the dropout of a subject may not depend on the response measurements, may depend only on the observed measurements, or may depend on the unobserved measurements. Different terminology is used by various authors. We will follow the terminology of Diggle and Kenward.^[7]

When the dropout is independent of the measurement process, then the dropout mechanism is termed “missing completely at random” (MCAR). An example would be a subject who drops out because he or she moves to a different city for reasons that have nothing to do with the disease or treatment. The case of missing completely at random also applies when the reason for dropout depends only on observable covariables. In other words, given covariables, subjects remaining in the study are a random subset of those dropping out of the study. Thus, any analysis that involves conditioning on these covariables (such as in a regression) will provide valid inferences with respect to the whole population being sampled. An important special case of this is differing dropout rates among treatment groups being compared, but the dropout being unrelated to the observed or unobserved responses.

When the dropout depends only on the observed portion of the measurement process (and covariables) but does not depend on the unobserved portion then the corresponding mechanism is termed “missing at random” (MAR). For instance, a subject who experienced less relief in prior treatment of a migraine headache may be more inclined to terminate a study than those who experienced greater relief. Again, in this case, the distribution of responses (adjusted for prior relief) for unobserved subjects is assumed to be the same as for observed subjects.

The preceding two types of dropouts are referred to as “noninformative” or “ignorable” dropout. When the dropout depends on the unobserved cases of the response measurements, the dropout is said to be “informative” or “nonignorable.” This is the case when dropout cannot be completely characterized by the observed information. Suppose a subject takes the study medication but continues to experience disease deterioration. Furthermore, the subject calls the office, briefly describes the problem, and never returns. This can be considered as a dropout for lack of efficacy, without corresponding response data collected. It is an example of dropout that depends on an unobserved response, or an informative dropout.

Analysis Populations and Accompanying Methods

It is customary to analyze data with dropouts based on specific subgroups of subjects. One simple subgroup is defined as such that data analysis is performed based on those subjects with the complete sequences of data collected over the study period. This is usually referred to as the “completers” subgroup analysis. Those subjects who drop out and therefore do not have data after dropping out will be excluded from the analysis. Inferences based on this population apply to the original population only in the case where subjects who have dropped out (dropped out completely at random) and any covariables affecting the likelihood of dropout are included in the analysis. This case is, in most circumstances, unlikely. Even in the case where missing completely at random holds estimates from a complete case analysis will be less efficient than when analyzing all the data.

Another approach is to base the analysis on all enrolled subjects, with any unobserved data being filled in (imputed) using some algorithm. Analyses involving all randomized subjects are usually referred to as intention-to-treat (ITT) analyses. The simplest imputation technique is to carry forward an observation obtained during the study such that every subject has a complete sequence of data values for every time point. Algorithms for carrying forward an observation include filling in missing values with the last observed response for a subject [last observation carried forward (LOCF)], with the worst observed response, or in some cases with the best observed response. (See the chapters “Carry-Forward Analysis” and “Imputation in Clinical Research” for more details.)

Obviously there are statistical concerns associated with these approaches. It is possible that subjects who dropped out were those who did not benefit from the treatment, did not tolerate the treatment well, or experienced serious side effects. Those who completed the study may have either had better health conditions or experienced greater benefit from the treatment to which they were assigned. In this case, a completers subgroup analysis is likely to err toward indicating an exaggerated positive effect for one or more of the treatments under study. In the case of an ITT analysis using longitudinal regression modeling paired with carried forward imputation, a further concern is the reduced variability estimates from imputing a single, fixed value.^[13] Observations to be collected after a subject’s dropout would have followed a certain distribution with a certain mean and variance. Imputing a constant value over time would reduce the underlying variance for each subject. Finally, it can be argued that imputation methods such as LOCF, baseline observation carried forward, and worst observation carried forward in effect change the hypothesis that is actually being tested inasmuch as the parameters being estimated have

a different interpretation from those being estimated if dropout does not occur.

A third approach is to perform statistical analyses based on the observed data, i.e., analyzing all of the subjects with any available data. The linear random-effects modeling method fits into this class. Under the linear random-effect model, each subject’s set of serial measurements is assumed to follow a linear function to time, with normally distributed errors. It is further assumed that the underlying slope and intercept of this individual linear function are bivariate normal random variables. The treatment effect is determined by the comparison of the slopes.^[14]

This procedure produces unbiased estimates in the case of ignorable dropout. However, the problems of bias and inefficiency in estimating treatment effect still exist when the unobserved data (i.e., after dropping out) contains information of the unknown treatment effect parameter; i.e., the dropout process is informative. For example, consider an analgesic study in which many subjects in a placebo group drop out due to inadequate pain relief, but a few remain because of a placebo effect or a higher tolerance for pain. In this case the placebo group efficacy may be overstated because the subjects who dropped would likely have had poorer outcomes; this in turn hinders the ability of the study to find a difference between an active agent and placebo.

Another all observed data approach is the likelihood-based mixed effects model repeated measures (MMRM) approach. The MMRM method uses an unstructured approach (e.g., serial measurements not assumed to follow a linear relationship) to modeling both the treatment-by-time means and the variances and covariances. This method has been shown to perform well assuming the dropout mechanism is MAR and has been argued to be preferable to LOCF analyses.^[15–17]

The applicability of the aforementioned methods depends on the missing-data process. Because none of the methods mentioned simultaneously consider the dropout process, none are suitable when the dropout mechanism is informative. Methods that model the joint distribution of response and dropout in terms of the marginal distribution of the response and the conditional distribution of dropout given response are called “selection model” methods.^[8] A variation of this model is the random coefficient-selection model or shared parameter model, where the response and dropout processes are conditioned on shared within-subject random effects.

Two such procedures will be described briefly in the next section: Diggle and Kenward^[7] and Wu and Bailey^[6] (for specific details the reader is referred to the corresponding papers). Diggle and Kenward’s approach is an example of the traditional selection model approach. Wu and Bailey’s methodology is an approximation of an approach developed by Wu and Carroll,^[5] which is an example of a shared parameter model.

OTHER METHODS

Diggle and Kenward

Diggle and Kenward proposed a likelihood-based method to adjust for informative dropout in analyzing continuous longitudinal data. As mentioned earlier, in order to adjust for dropout, Diggle and Kenward consider a selection model approach to model the joint distribution of response and dropout. Their approach can be broken down into three basic components: an underlying multivariate normal linear model to model the longitudinal response vector; a logit model to model the likelihood of dropping out as a function of the longitudinal responses; and the likelihood function of the observed data, which ties the first two components together. An application of the Diggle and Kenward approach to data from a clinical trial with dropout can be found in the text by Molenberghs and Kenward.^[10] We will describe these components in general next.

The longitudinal responses of individuals across time are assumed to have a multivariate normal distribution. In clinical trials, the primary interest is whether or not there is a treatment effect. The treatment effect as well as other covariates, such as the corresponding measurement time for a particular response, baseline status, and centers, are incorporated as parameters in a linear function describing the relationship with the mean response vector. For the covariance structure, Diggle and Kenward allow for three independent variance components: a random intercept component between individuals, a serially correlated within-subject component, and a sampling or measurement error component that is uncorrelated across individuals.

The dropout process is modeled by implementing a linear logistic model. The logit of the probability that the subject, at time d , drops out is modeled as a linear function of the responses prior to dropout and the unobserved response at the time of dropout. Assuming the underlying dropout and response models are correct, submodels can be fit and tested using the likelihood ratio statistic to determine if the dropout process is informative. Using the definitions for the various dropout mechanisms given in "The Dropout Process," the full model corresponds to a case of nonignorable dropout (dropout depends on unobserved data). The reduced model, where the coefficient associated with the unobserved response at the time of dropout is zero, corresponds to the case of ignorable dropout. In this case, the dropout process is independent of the missing response. Lastly, completely random dropout corresponds to the model where all coefficients except the intercept are zero. In this case, the dropout process is independent of both observed responses and missing responses.

For each subject, the observed data consist of the observed longitudinal responses prior to dropout and the time of dropout. Therefore, each subject contributes the following to the likelihood: the marginal distribution of the multivariate linear model corresponding to the responses

up to the time prior to dropout and a sequence of terms corresponding to the probability that a subject drops out at the time he or she did based on the observed responses.

For each subject, the log-likelihood function for (θ, Σ, ϕ) given the observed can be partitioned into three components: $L(\theta, \Sigma, \phi) = L_1(\theta, \Sigma) + L_2(\phi) + L_3(\theta, \Sigma, \phi)$

The first likelihood component is

$$L_1(\theta, \Sigma) = \log \{ \phi_{d-1}(y) \} \quad (1)$$

where $\phi_{d-1}(y)$ is the joint marginal multivariate normal density for the first $d-1$ observed longitudinal responses, where d is the time of dropout for the particular subject.

The second and third components correspond to the probability that a subject drops out at time d given the $d-1$ observed longitudinal responses. This consists of a sequence of terms corresponding to the probability that a subject does not drop out before time d and a term for the probability that a subject drops out at time d . This involves the following two terms:

$$L_2(\phi) = \sum_{j=2}^{d-1} \log \{ 1 - \Pr(j | \dots, y_{j-1}, y_j; \phi) \} \quad (2)$$

and

$$L_3(\theta, \Sigma, \phi) = \log(\Pr(d | y_1, \dots, y_{d-1}; \theta, \Sigma, \phi)) \quad (3)$$

where $\Pr(j | y_1, \dots, y_{j-1}, y_j; \phi)$ in Eq. (2) has the following form based on the logistic regression model for modeling the dropout process:

$$\Pr(j; \phi) = \text{logit}^{-1} \left(\phi_0 + \phi_1 y_j + \sum_{k=2}^j \phi_k y_{j+1-k} \right) \quad (4)$$

Only subjects who drop out of the study contribute to Eq. (3). Because the response value y_d at time d (time of dropout) is unobserved, it is not conditioned upon Eq. (30). Eq. (3) is evaluated by integrating Eq. (4) for $j = d$ over the univariate normal conditional distribution of y_d given $y_1, \dots, y_{d-1}$.

Parameters are estimated iteratively via maximum likelihood (ML) using the simplex algorithm of Nelder and Mead.^[18] Using maximum likelihood, the parameters of the complete-data model and dropout process are estimated simultaneously, and thus parameter estimates corresponding to the response distribution, such as treatment effects, are obtained adjusting for dropout (assuming suitable models for response and dropout have been found). In addition, it is possible to test certain hypotheses about the dropout mechanism by testing whether coefficients associated with various response measurements given in Eq. (4) are zero. As mentioned above, it is important to understand that such tests might indicate whether or not the *observed* data are consistent with a certain type of dropout mechanism, but fundamentally they cannot be used to conclude

that unobserved measurements are or are not related to what has been observed.

As mentioned earlier, a generalization of the selection model is the random coefficient-selection model. In this case rather than the marginal response and dropout process, we have the response and dropout process conditional on within-subject random effects. A special case of this is the shared parameter model, where the dropout process possibly depends on unobserved within-subject random effects underlying the response distribution and is thus informative.

Wu and Bailey

Wu and Bailey give a methodology to approximate the approach developed by Wu and Carroll. Wu and Carroll use a random-effects linear model and discrete time survival model to model the response and dropout process, respectively, where the likelihood of dropout during a specific time interval is a function of the dropout time and on the subject-specific intercept and slope. We begin with a specific form of general random-effects linear model of Laird and Ware.^[19]

Assume that for an individual of a particular group k , conditional on the underlying vector of coefficients β_k , the series of measurements $\mathbf{Y}_i$ follow a multivariate normal distribution with mean $\mathbf{X}_i\beta_k$, and variance σ^2 , where σ^2 is an unknown scalar residual variance, and β_k are random variables that are multivariate normally distributed with mean B_k and covariance matrix Σ_β , where $\mathbf{Z}_i$ is a $(q \times p)$ between-subjects design matrix, β is an unknown $(p \times 1)$ vector of slopes, and Σ_β is an unknown $(q \times q)$ covariance matrix. The linear random-effects model implies that marginally $\mathbf{Y}_i$ are independent multivariate normal with mean $\mathbf{X}_i\beta_i$ and covariance matrix $\mathbf{K}_i\Sigma_\beta\mathbf{K}_i^T + \sigma^2\mathbf{I}$, where $\mathbf{X}_i = \mathbf{K}_i\mathbf{Z}_i$.

Wu and Carroll consider the case where dropout is caused by a subject's death or withdrawal and refer to it as the "right-censoring" process. They assume that the series of measurements follows a linear random-effects model and that the right-censoring process and the observed data share the random-effects parameters β_k are also related to the probability of being right censored through a regression function such as probit or logistic. The corresponding marginal likelihood for B_k , the true group effect parameter, consists of the joint distribution of the simple least-squares estimates of individual intercept and slope, the unobserved true individual random effect parameters, and the right-censoring process, which is then integrated over with respect to the individual random effects parameters β_k . To obtain parameter estimates, the unknown scalar variance and the covariance matrix of the random intercept and slope are substituted with the unbiased estimates of them in the likelihood function, and with the observed survival times and censoring times the marginal likelihood is called a pseudo-likelihood.

Under noninformative censoring, the maximum-likelihood estimates of the random slopes β_i are the weighted least-squares estimates (WLE), which are just a weighted average of the individual simple least-squares estimated slopes, with $[\Sigma_\beta + (\mathbf{X}_i^T\mathbf{X}_i)^{-1}\sigma^2\mathbf{I}]^{-1}$, the inverse of the variance of the random slopes, as the weight. In this case, the WLE are consistent and unbiased.

However, when censoring is informative, for example, censoring depends on the individual slope so that subjects who have more serious outcomes (say, steeper rate) are more likely to drop out early, then the WLE are biased. This is because the weights used in the weighted least-squares estimation are functions of the within-subject design matrix, $\mathbf{K}_i$, which is a function of the number of repeated measures within the subject. Subjects who drop out early have smaller weights than those who drop out late or who complete the study, resulting in estimates that tend to reflect only the treatment effect of those with less serious conditions, which are thus biased. Note that because the individual ordinary least-squares estimates for slope are unbiased, if an unweighted average is used (UWLE) to estimate the random slope effect, then this estimate is unbiased, although it can be inefficient.

Wu and Bailey provided a simple approximate method for the informative censoring problem. They observed that under the random-effects linear model with a probit censoring process, the conditional expectation of the random-effect slope (given the censoring time) is a monotonic function of the censoring time. Thus, instead of modeling the censoring process, they assume that, given the censoring time, the individual least-squares estimated slopes are linear functions of the censoring time with normal errors. Specifically, the conditional linear model assumes that, for each individual, conditional on the censoring time, the expected random slope has mean $\gamma_0 + \gamma_1 E_k(T_i)$ and conditional variance σ_{kTi}^2 , where $E_k(T_i)$ is the expected value of censoring time for subject i in treatment group k . The conditional variance σ_{kTi}^2 is an unknown scalar residual variance, which depends on censoring time and thus on the number of measurements before dropout. Under the random-effects linear model, the σ_{kTi}^2 equals $\Sigma_\beta + (\mathbf{X}_{Ti}^T\mathbf{X}_{Ti})^{-1}\sigma^2\mathbf{I}$, and when Σ_β and σ^2 are unknown, they can be estimated using the restricted maximum-likelihood method (REML) method.

Two methods to estimate the random effect parameters under the conditional linear model are proposed. These are referred to as the linear minimum variance unbiased (LMVUB) and the linear minimum mean squared error (LMMSE) estimates, respectively. Because of the good properties associated with the first method, we will discuss it next and refer the reader to Ref. [6] for discussion of the second.

The LMVUB procedure estimates the conditional linear coefficients γ_0 and γ_1 by weighted least squares and then substitutes the estimates of γ_0 and γ_1 into the conditional linear function to obtain the LMVUB of B_k . The sample

mean censoring time is used as an estimate for $E_k(T_i)$, the expected value of censoring time for subject i in group k . In practice, the LMVUB estimator can be derived following a sequence of steps. First, the ordinary least-squares estimators of slope for each individual are obtained, and the conditional variances of the estimated conditional slopes given the design matrix of observation times are estimated. Then, weighted analyses of covariance based on these individual slopes, using the inverse of the conditional variances as weights, with dropout times as covariates in the linear model, are used to estimate the group slopes. The difference between the slopes is then tested for significance.

In a simulation study, using a probit model for the right-censoring process, the performance of the LMVUB procedure was competitive when compared to the pseudomaximum-likelihood estimation of Wu and Carroll. As also shown in a Monte Carlo simulation study by Wang-Clow et al.,^[20] which compares the performance of several estimators (UWLE, WLE, LMVUB, weighted mean, complete-case, and MLE), when the missing process is informative but the same in both treatment and control groups, the LMVUB is the only unbiased estimator for the difference of the slopes. When the dropout process differs for the two groups, the UWLE has the smallest bias, but it is noncompetitive in terms of mean squared error. The LMVUB has the smallest mean squared error and the lowest bias of all the estimators.

DISCUSSION

A key assumption of the dropout model proposed by Diggle and Kenward is that the process depends at most on the responses up to and including the unobserved response at the time of dropout but not the latter unobserved responses. Little^[8] mentions that such a dropout model is more plausible when dropout is linked directly to the response being measured (such as withdrawal to avoid a particularly invasive procedure) rather than to some underlying process (such as weight loss, nausea, somnolence, etc., because of treatment with a medication).

In the second case, the dependence of dropout on future unobserved responses is a function of how accurately the response variable measures the underlying process. In this case, the framework underlying the random coefficient-selection model may be better satisfied.

An alternative approach to the selection model is the pattern mixture model. In this case, the joint distribution of response and dropout is modeled as the marginal distribution of dropout and the conditional distribution of response given dropout. Because the data are not observed when missingness occurs, parameters underlying the response distribution given dropout are linked by specifying prior information. Little^[21] considers this type of model for analyzing incomplete multivariate data.

The text by Molenberghs and Kenward^[10] provides an introduction to and examples of selection and pattern mixture models; a Bayesian approach to estimating pattern mixture models in clinical trials with dropout has been explored by Demirtas.^[22]

EXTENSIONS AND SOFTWARE

In this entry we discussed the effect of dropout primarily in analyzing continuous longitudinal data. An example of a shared parameter approach when the longitudinal data are binary is given in Ten Have et al.^[23] An example of a random coefficient-selection model approach for ordinal longitudinal data, where dropout is a function of the response variable and within-subject random effects, is given in Sheiner et al.^[24] For extensions involving nonparametric analysis of incomplete repeated measures, the reader is referred to Wei and Johnson.^[25]

Lastly, software for running Diggle and Kenward's procedure using Splus (called "Oswald") can be found on the Internet;^[26] the use of the software was recently illustrated in the context of a psychiatric clinical trial with informative dropout.^[27] As far as we know, no available software exists for the Wu and Bailey procedure, but the steps outlined under "Wu and Bailey" can easily be implemented using existing statistical software. In addition, single imputation methods, such as LOCF, and multiple imputation methods can be carried out using the software package SOLAS.^[28]

REFERENCES

1. Adler, R.; Balagtas, C.; Cappelleri, J.; Cheng, S.; Finman, J.; Lan, G.; Lan, S.; Li, M.; Pickering, E.; Rodenberg, C.; Ting, N. Selected approaches for the analysis of longitudinal data with dropouts; *Biometrics Technical Report*, Groton, CT; 1997.
2. Rubin, D. B. *Biometrika* **1976**, *63*, 581–592.
3. Little, R.; Rubin, D. B. *Statistical Analysis of Missing Data*; Wiley: New York, 1987.
4. Laird, N.M. *Stat. Med.* **1988**, *7*, 305–315.
5. Wu, M.; Carroll, R. *Biometrics* **1988**, *44*, 175–188.
6. Wu, M.; Bailey, K. *Biometrics* **1989**, *45*, 939–955.
7. Diggle, P.; Kenward, M.G. *Appl. Stat.* **1994**, *43* (1), 49–93.
8. Little, R. J. *Am. Stat. Assoc.* 1995, *90*, 1112–1121.
9. Hogan, J.W.; Roy, J.; Korkontzelou, C. *Stat. Med.* **2004**, *23*, 1455–1497.
10. Molenberghs, G.; Kenward, M.G. *Missing Data in Clinical Studies*; John Wiley & Sons Ltd.: Chichester, 2007.
11. International Conference on Harmonisation E9 Expert Working Group. *Stat. Med.* **1999**, *18*, 1905–1942.
12. Committee for Proprietary Medicinal Products. *Points to consider on missing data*; The European Agency for the Evaluation of Medicinal Products: London, <http://www.emea.eu.int>. 2001.

13. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; Wiley: New York, 1987.
14. Verbeke, G.; Molenberghs, G. *Linear Mixed Models for Longitudinal Data*; Springer: New York, 2000.
15. Mallinckrodt, C.H.; Kaiser, C.J.; Watkin, J.G.; Molenberghs, G.; Carroll, R. *Pharm. Stat.* **2004**, *3*, 171–186.
16. Lane, P. *Pharm. Stat.* **2008**, *7*, 93–106.
17. Mallinckrodt, C.H.; Lane, P.; Schnell, D.; Peng, Y.; Mancuso, J.P. *Drug Inf. J.* **2008**, *42*, 303–319.
18. Nelder, J.A.; Mead, R. *Comput. J.* **1965**, *7*, 303–313.
19. Laird, N.M.; Ware, J.H. *Biometrics* **1982**, *38*, 963–974.
20. Wang-Clow, F.; Lange, M.; Laird, N.M.; Ware, J.H. *Stat. Med.* **1995**, *14*, 283–297.
21. Little, R. J. *Am. Stat. Assoc.* **1993**, *88*, 125–134.
22. Demirtas, H.J. *Biopharm. Stat.* **2005**, *15*, 383–402.
23. Ten Have, T.R.; Kunselman, A.R.; Pulkstenis, E.P.; Landis, J.R. *Biometrics* **1998**, *98*, 367–383.
24. Sheiner, L.B.; Beal, S.L.; Dunne, A.J. *Am. Stat. Assoc.* **1997**, *92*, 1235–1244.
25. Wei, L.J.; Johnson, W.E. *Biometrika* **1985**, *72*, 359–364.
26. <http://www.maths.lancs.ac.uk/Software/Oswald/>.
27. Begley, A.E.; Tang, G.; Mazumdar, S.; Houck, P.R.; Scott, J.; Mulsant, B.H.; Reynolds III, C.F. *Comput. Methods Programs Biomed.* **2007**, *85*, 109–114.
28. *SOLAS for Missing Data Analysis 3.0.*; Statistical Solutions, Ltd.: Cork, 2001.

Drug Development

Naitee Ting

Pfizer Global Research and Development, New London, Connecticut, U.S.A.

Abstract

The drug-development process can be broadly classified into two major components: nonclinical development and clinical development. This entry breaks down these two phases of drug development and considers regulatory requirements and marketing of a new drug.

INTRODUCTION

Most of the drugs available in pharmacies started out as a chemical compound or a biologic discovered in laboratories. When first discovered, each new compound or biologic is denoted as a drug candidate. The drug candidate has to demonstrate some activities in the laboratory to indicate that it has the potential of treating certain diseases in humans. Drug development is a process that starts when the drug candidate is first discovered and continues until it is available to be prescribed by physicians to treat patients with the disease. A potential new drug or a drug candidate can generally be classified into one of two categories: a synthetic chemical compound or a biologic. A compound is usually a new chemical entity synthesized by scientists from pharmaceutical companies (often referred to as sponsors), universities, or research institutes. A biologic can be a protein, a part of a protein, DNA, or a different form either extracted from tissues of another live body or cultured by some type of bacteria. In any case, this new compound or biologic will have to go through the drug-development process before it can be consumed by the public.

OVERVIEW

The drug-development process can be broadly classified into two major components: nonclinical development and clinical development. Nonclinical development includes all drug testing performed outside of the human body, and clinical development is based on experiments conducted in the human body. Nonclinical development can be further divided into pharmacology, toxicology, and formulation. In these processes, experiments are performed in laboratories or pilot plants. Observations from cells, tissues, or animal bodies are collected to derive inferences for potential new drugs. Chemical processes are involved in formulating the new compound into drugs to be delivered into the human body. Clinical development can be further divided into Phases 1, 2, 3, and 4. Clinical studies are designed

to collect data from normal volunteers or patients who participate in these studies.

Throughout the whole drug-development process, two scientific questions are constantly being addressed: Does the drug candidate work? Is it safe? Starting from the laboratory where the compound or biologic is first discovered, the candidate has to go through many tests to see if it demonstrates both efficacy (the drug works) and safety. Only candidates that pass all those tests can progress to the next step of development. In the United States, after a drug candidate passes all of the nonclinical tests, an investigational new drug (IND) document is filed with the Food and Drug Administration (FDA). After the IND is approved, clinical trials (tests on the human body) can then be performed for the drug candidate. If this drug candidate is shown to be safe and efficacious through phases 1, 2, and 3 clinical trials, the sponsor will then file a new drug application (NDA) with the FDA in the United States. The drug can be made available for general public consumption only if the NDA is approved. Oftentimes the new drug continues to be developed even after it is on the market. These postmarketing studies are known as phase 4 clinical trials.

A great many statistical methods are used in both clinical and nonclinical drug-development processes. Details of these processes and statistical applications are discussed in the following sections: The next section introduces the regulatory requirements. The following sections discuss “Nonclinical Development” and “Premarketing Clinical Development.” “Postmarketing Clinical Development” is presented next.

REGULATORY REQUIREMENTS

In most modern countries, a drug has to be approved by drug regulatory agencies before it can be marketed. For example, in the United States this agency is the FDA, in Canada it is the Health Protection Branch (HPB), etc. There are guidelines from these agencies to regulate various products and processes in drug development.

These guidelines provide a consistent framework for sponsors (pharmaceutical companies) to develop a new drug and for reviewers in these agencies to evaluate documents submitted to them by the sponsors.

The guidelines cover many aspects of drug research and development. For example, the guidance documents from the FDA regulate various steps of new drug development, including chemistry, pharmacology/toxicology, INDs, clinical/medical, information technology, labeling, generics, advertising, and other areas of drug regulation. Among all these guidance documents, the one that is most relevant to biopharmaceutical statistics is the “Guideline for the Format and Content of the Clinical and Statistical Sections of an Application,”^[1] within the clinical/medical guidelines. This document regulates the format, contents, data sets, and statistical methods used for both efficacy and safety analysis in an NDA report. For details of these documents, interested readers may refer to the FDA guidelines. These guidelines can be ordered from the Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402. They can also be found from the following Web page: (<http://www.fda.gov/cder/guidance/index.htm>).

There has been a continuous effort to harmonize regulation guidelines from various countries. A new organization—the International Conference on Harmonization (ICH)—has arisen, with members from regulatory agencies and drug developers (or sponsors) of various countries (including the United States, Canada, Japan, and many other European and Asian countries), to generate a set of guidelines for drug developers and regulatory agencies to use. As of 1998, a number of ICH guidelines had been developed and finalized; other guidelines were being drafted. The ICH guidelines can be broadly divided into the areas of safety, efficacy, and quality. Statistics is an important component in these guidelines. In particular, ICH Efficacy Guideline E9—“Statistical Principles for Clinical Trials”—was developed specifically for statistical applications in the clinical development of a new drug.

NONCLINICAL DEVELOPMENT

A new chemical compound or biologic can be a drug candidate because it demonstrates some desirable pharmacological activities in the laboratory. At such an early stage of drug development, the focus is mainly on cells, tissues, organs, or animal bodies. Experimentation on human beings is performed after the candidate passes these early tests and looks promising. Hence, nonclinical development may also be referred to as “preclinical” development because these experiments are performed before clinical trials are performed on the human body.

One of the most important properties to be studied at the nonclinical testing stage is whether there is a dose–response relationship for the drug candidate. Fig. 1 shows

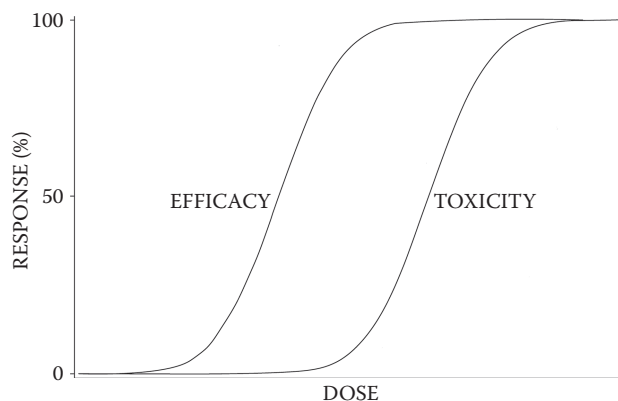


Fig. 1 Theoretical dose–response curves

two theoretical curves describing these relationships. The x -axis represents the doses of the drug candidate (can be linear scale or logarithmic scale), and the y -axis represents the responses. Suppose the y -axis allows 0% responses at the bottom and 100% responses (both efficacy response and toxicity response) at the top, then as dose increases, the responses increase. The curve on the left represents the desirable response (potency, efficacy); the curve on the right represents the undesirable response (toxicity). The area in between those two curves would be the doses at which the drug works and is safe. Theoretically speaking, the best drug candidate would be one for which the drug potency (efficacy) increases very rapidly while the toxic response does not increase much as the dose increases. Under this ideal situation, a wide range of doses can be prescribed to treat patients. However, in reality these two curves can be close to each other. When this is the case, there is a narrow range of doses that will be both potent and safe, and dose selection must be very careful. In the process of developing a certain drug candidate, should these two curves become too close to each other, or should the toxic response curve locate to the left of the efficacy curve, then the strategy may be to stop development for this candidate.

At the preclinical stage, the study of drug potency or efficacy is part of pharmacology, whereas the study of toxicity is part of drug safety. In addition to pharmacological activity and toxicity, another important process at this preclinical development stage is the formulation of the compound or biologic into a form that can be consumed by humans (for example, tablet, capsule). This process is generally referred to as “drug formulation.” Nonclinical development processes include pharmacology, toxicology/safety, and drug formulation. These processes are described in the following subsections.

Pharmacology

Pharmacology is the study of the selective biological activity of chemical substances on living matter.^[2] A substance has biological activity when, in appropriate doses, it causes

a cellular response. It is selective when the response occurs in some cells and not in others. Therefore, a compound or a biologic has to demonstrate these activities before it can be further developed. In the early stages of drug testing, it is important to differentiate an “active” candidate from an “inactive” candidate. There are screening procedures to select these candidates. Two properties of particular interest are sensitivity and specificity. Given that a compound is active, “sensitivity” is the conditional probability that the screen will classify it as positive. “Specificity” is the conditional probability that the screen will call a compound negative given that it is inactive.^[3] In the ideal case, both of these values should be close to 1.

The level of these pharmacological activities may be viewed as the drug potency or strength. The estimation of drug potency by the reactions of living organisms or their components is known as bioassay. According to Hubert et al.,^[2] “the experimental determination of the potency or strength of a chemical or biological substance based on the response observed after its administration to living matter (animal or animal tissue) constitutes a biological assay.”

As discussed previously, one of the most important relationships needed to be studied for pharmacological activities is the dose–response relationship. In these experiments, several doses of the drug candidates are selected, and the responses are measured for each corresponding dose. After response data are collected, regression methods may be applied to analyze the results. Regression modeling is a very useful statistical method applied in the analysis of pharmacological activities.

Toxicology/Drug Safety

Drug safety is one of the most important concerns throughout all stages of drug development. In the preclinical stage, drug safety needs to be studied in a variety of animals. Studies are designed to observe adverse drug effects or toxic events experienced by animals treated with different doses of the drug candidate. Animals are also exposed to the drug candidate for various lengths of time to see if there are adverse effects caused by cumulative dosing over time. These results are summarized and analyzed using statistical methods. When the results of animal studies indicate potentially serious side effects, drug development is either terminated or suspended pending further investigation of the problem.

Depending on the duration of exposure to the drug candidate, animal toxicity studies are classified as acute, subacute, chronic, or reproductive.^[4] Usually the first few studies are “acute studies”; i.e., the animal is given one or a few doses of the drug candidate. If only one dose is given, it can also be called a “single-dose study.” Only those drug candidates demonstrated to be safe in single-dose studies progress into “multiple-dose studies.” Acute studies are typically about 2 weeks in duration. Repeat-dose studies of 30–90 days’ duration are called “subacute

studies.” Chronic studies are usually designed with more than 90 days of duration. These studies are conducted in rodents and in at least one nonrodent species. Chronic studies may also be viewed as carcinogenicity studies because the rodent studies consider tumor incidence as an important endpoint. “Reproductive studies” are carried out to assess the drug's effect on fertility and conception; they can also be used to study the drug's effect on the fetus and developing offspring.

Statistical tools are very useful in designing and analyzing animal toxicity studies. Experimental design techniques are used in designing these studies; hypothesis tests are performed in comparing treatment group differences. Categorical data analysis is used to assess binary responses or count data. Life table techniques and survival analyses are very useful in analyzing chronic and carcinogenicity studies.

Drug Formulation Development

As discussed earlier, a potential new drug can be either a chemical compound or a biologic. If the drug candidate is a biologic, then the formulation is typically a solution that contains a high concentration of such a biologic, and the solution is injected into the subject. On the other hand, if the potential drug is a chemical compound, then the formulation can be tablets, capsules, solution, patches, suspension, or other forms. There are many formulation problems that require statistical analyses. The formulation problems that stem from chemical compounds are more likely to involve widely used statistical techniques, but the problems involving formulation of biologics are less likely to succumb to standard statistical solutions. The paradigm of a chemical compound is used here to illustrate some of these formulation-related problems and how statistical tools can be useful.

A drug is the mixture of the synthesized chemical compounds (active ingredients) with other, inactive ingredients designed to improve the absorption of the active ingredients. How the mixture is made depends on the results of a series of experiments, called optimization experiments. Usually the optimization experiments are performed under some physical constraints, e.g., the amount of raw materials, the capacity of the container, and the size and shape of the tablets. Some of the samples from the batch-processed tablets are stored to study the stability, or shelf life. These stored samples are exposed to various environmental conditions, e.g., freezing, extra sunlight, very dry or very humid conditions. Part of these stored samples are reanalyzed after half a year, 1 year, and 2 or more years. Such sample properties as potency, dissolution rate, and color, are measured at these time points to study the stability of the drug.

A few branches of statistics can be helpful in designing and analyzing these experiments. Optimization designs, factorial designs, or fractional factorial designs can be used

in designing these experiments. After the experiments are over and data are collected, statistical tools are used to analyze these data. Variable selection techniques of multiple regression help select the best subset of independent variables. Analysis of variance (ANOVA) is used to perform data analysis and to test various statistical hypotheses. In cases where multiple dependent variables are to be optimized, multivariate analysis (including MANOVA) may be used to analyze these data. Response surface methodology provides ways to estimate the optimal combination of a set of independent variables. From these results, pharmacists are able to operate under the physical constraints and to select the appropriate materials and processes to produce the tablets for a compound so that the tablets possess some optimal properties.

PREMARKETING CLINICAL DEVELOPMENT

If a compound or a biologic passes the selection process from animal testing and is shown to be safe and efficacious to be tested in human, it progresses into clinical development. As discussed earlier, the major distinction between clinical trials and nonclinical testing is the experimental unit. In clinical trials, the experimental units are human beings, whereas the experimental units in nonclinical testing are nonhuman subjects. Before a drug candidate can be tested in humans, the pharmaceutical company (sponsor) has to submit an IND to the FDA for review. If there is no response from the FDA after 30 days following IND submission, the sponsor can start clinical testing for this drug candidate. This compound or biologic may be referred to as the “new drug.”

An IND is a document that contains all the information known about the new drug up to the time the IND is prepared. Generally, the IND includes the name and description of the drug (chemical structure, other ingredients, etc.); how the drug is processed, manufactured, and packaged; information about any preclinical experiences relating to the safety of the drug; marketing information; past experiences or future plans for investigating the drug in foreign countries. In addition, it also contains a description of the clinical development plan. Such a description should contain all of the informational materials supplied to clinical investigators, the educational and scientific background of these investigators, a signed agreement from the investigators, and the initial protocols for clinical investigation.

Clinical development is broadly divided into four phases, namely, phases 1, 2, 3, and 4. Phase 1 trials are designed to study short-term effects of the new drug, e.g., dose range (What range of doses should be tested in humans?), pharmacokinetics (PK; What does a human body do to the drug?), and pharmacodynamics (PD; What does a drug do to the human body?). Phase 2 trials are designed to study the efficacy of the new drug in well-defined subject populations. Dose–response relationships are also studied during

phase 2. Phase 3 studies are usually long-term, large-scale studies to confirm findings established from earlier trials. These studies are also used to detect toxic effects caused by cumulative dosing. If a new drug is found to be safe and efficacious during the first three phases of clinical testing, an NDA is filed for regulatory agencies to review (e.g., FDA, in the United States). Note that for marketing approval of a new drug product, the U.S. FDA requires that substantial evidence of the effectiveness of the drug product be provided through the conduct of at least two adequate and well-controlled clinical trials. However, under certain circumstances, the U.S. FDA Modernization Act of 1997 includes a provision to allow data from one adequate and well-controlled clinical trial investigation and confirmatory evidence to establish effectiveness for risk/benefit assessment of drug and biological candidates for approval.^[5] Once the drug is approved by the FDA, phase 4 (postmarketing) studies are planned and carried out. Many of the phase 4 study designs are dictated by the FDA to examine safety questions; some are used to establish new uses. The following sections discuss “Phase 1 Clinical Trials,” “Phase 2/3 Clinical Trials,” and “New Drug Application.”

Phase 1 Clinical Trials

In a Phase 1 pharmacokinetics (PK) study, the purpose is usually to study the bioavailability of a drug or the bioequivalence among different formulations of the same drug. Bioavailability means “the rate and extent to which the active drug ingredient or therapeutic moiety is absorbed and becomes available at the site of drug action.”^[6] Experimental units in Phase 1 studies are mostly normal volunteers. Subjects recruited for these studies are generally in good health.

A bioavailability or bioequivalence study is carried out by measuring drug concentration in human blood or serum over time and then analyzing summary data from these measurements. Fig. 2 presents a drug concentration–time curve; data on this curve are collected at discrete time points. Typical variables used for analysis of PK activities

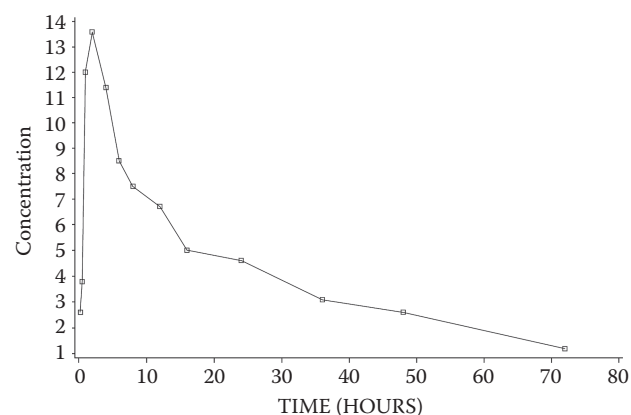


Fig. 2 Drug concentration in blood samples collected over time

include area under the curve (AUC), maximum concentration ($C_{\max}$), time to maximum concentration ($T_{\max}$), among others. These variables are computed from drug concentration levels, as shown in Fig. 2. Suppose AUC is the variable to be analyzed, then these discretely observed points are connected (for each subject) and the area under the curve is estimated using a trapezoidal rule. For example, AUC up to 24 hr for this curve is computed by adding up the area of the triangle between hour 0 and hour 0.25, the area of the trapezoid between hour 0.25 and hour 0.5, and so on through the area of the triangle between hour 16 and hour 24.

Statistical designs used in phase 1 bioavailability studies are often crossover designs; that is, a subject is randomized to be treated with drug A first and then treated with drug B after a “washout” period, or randomized to drug B first and then treated with drug A. In some complicated phase 1 studies, two or more treatments may be designed to cross several periods for each subject. The advantages and disadvantages of crossover designs are discussed in Chow and Liu.^[6] Response variables, including AUC, $T_{\max}$, and $C_{\max}$, are usually analyzed using ANOVA models. Random-effects and mixed-effects linear/nonlinear models are also commonly used in the analysis of Phase 1 clinical studies. In certain designs, model terms considered in these models can be very complicated.

Phase 2/3 Clinical Trials

Phase 2/3 trials are designed to study the efficacy and safety of a test drug. Unlike phase 1 studies, subjects recruited in Phase 2/3 studies are patients with the disease that the drug is developed for. Response variables considered in phase 2/3 studies are mainly efficacy variables and safety variables. For example, in a trial for the evaluation of hypertension (high blood pressure), the efficacy variables are blood pressure measurements. For an anti-infective trial, the response variables can be the number of subjects cured or the time to cure for each subject. Phase 2/3 studies are designed mostly with parallel treatment groups; i.e., if a patient is randomized to receive treatment A, then this patient would be treated with drug A throughout the whole study.

Phase 2 trials are often designed to compare a test drug against placebo. These studies are usually short term (a few weeks) and designed with a small or moderate sample size. Often, phase 2 trials are exploratory trials. One very important type of phase 2 study is the dose–response study. As in the left curve of Fig. 1, drug efficacy may increase as dose increases. In a dose–response study, the following fundamental questions need to be addressed:^[7]

Is there any evidence of a drug effect?

What doses exhibit a response different from the control response?

What is the nature of the dose–response relationship?

What is the optimal dose?

Typical dose–response studies are designed with fixed-dose, parallel treatment groups. For example, in a four-treatment group trial designed to study dose–response relationships, three test doses—low, medium, and high—are compared against a placebo. In this case, results may be analyzed using multiple comparison techniques.

Phase 3 trials are long term (can be a few years) and large scale, with less restricted patient populations, and often compared against a known active drug for the disease to be studied. Phase 3 trials tend to be confirmatory trials designed to verify findings established from early studies.

Statistical methods used in phase 2/3 clinical studies can be very different. Statistical analyses are selected based on the distribution of the variables and the objectives of the study. Categorical data analyses are frequently used in analyzing count data (e.g., number of subjects who responded, number with side effects, or number of subjects improved from “severe symptom” to “mild symptom”). Survival analyses are commonly used in analyzing time to an event (such as time to discontinuation of the study medication, time to the first occurrence of a side effect, or time to cure). *T* tests, regression analyses, analyses of variance (ANOVA), analyses of covariance (ANCOVA), and multivariate analyses (MANOVA) are useful in analyzing continuous data (e.g., blood pressure, grip strength, forced expiration volume, number of painful joints, area under the curve). In many cases, nonparametric analytical methods are selected because the data do not fit any known parametric distribution well. In other cases, the raw data are transformed (log-transformed, ranked, centralized, combined) before a statistical analysis is performed. A combination of various statistical tools may sometimes be used in a drug-development program. Hypotheses tests are often used to compare results obtained from different treatment groups. Point estimates and interval estimates are also frequently used to estimate subject responses to a study medication or to demonstrate equivalence between two treatment groups.

After a clinical study is completed, all of the data collected from the study are to be stored in a database; statistical analyses are performed on data sets generated from the database. A study report is prepared for each completed clinical trial. It is a joint effort to prepare such a study report. Statisticians, data managers, and programmers work together to produce tables, figures, and statistical report. Clinicians and technical writers will then put together clinical interpretations of these results. All of these are incorporated into a study report. Study reports from individual clinical trials are eventually put together as part of an NDA.

New Drug Application

When there is sufficient evidence to demonstrate a new drug that works and is safe, a new drug application is put together by the sponsor. An NDA is a huge document describing all of the results obtained from both nonclinical experiments and clinical trials. A typical NDA contains

sections on proposed drug labeling, pharmacological class, foreign marketing history, chemistry, manufacturing and controls, nonclinical pharmacology and toxicology summary, human pharmacokinetics and bioavailability summary, microbiology summary, clinical data summary, results of statistical analyses, benefit/risk relationship, and so on. If the sponsor intends to market the new drug in other countries, then separate packages will need to be prepared for submission to each corresponding country, too. For example, a new drug submission (NDS) needs to be filed with Canada's regulatory agency, a marketing authorization application (MAA) needs to be filed with the U.K. regulatory agency.

Oftentimes an NDA is filed while the phase 3 studies are ongoing. Sponsors need to be very careful in selecting the data cutoff date, because all of the clinical data in the database up to the cutoff date need to be frozen and stored so that NDA study report tables and figures can be produced from them. The data sets stored in such an NDA database may have to be retrieved and reanalyzed after filing in order to address various queries from regulatory agencies. After these data sets are created and stored, ongoing clinical data can then be entered into the regular database.

An NDA package usually includes not only individual clinical study reports, but also combined study results. These results may be summarized using meta-analyses or pooled-data analyses. Such analyses are performed on efficacy data to produce the "integrated summary of efficacy" and on safety data to produce the "integrated summary of safety." These summaries are important components of any NDA. In some cases, electronic submissions are filed as part of the NDA. Electronic submissions usually include individual clinical data, programs to process these data, and software/hardware to help FDA reviewers review the individual data as well as the whole NDA package.

POSTMARKETING CLINICAL DEVELOPMENT

An NDA serves as a landmark of the drug-development process. The development process does not stop when an NDA is submitted or approved. But the objectives of the process are changed after the drug is approved and is available on the market. Studies performed after the drug is approved are typically called postmarketing studies or phase 4 studies.

One of the major objectives in postmarketing development is to establish a better safety profile for the new drug. Large-scale, drug-safety surveillance studies are very common in phase 4. Subjects/patients recruited in phases 1, 2, and 3 are typically somewhat restricted (patients would have to be within a certain age range, gender, disease

severity, etc.). However, after the new drug is approved and is available to the general patient population, every patient with the indicated disease can be exposed to this drug. Problems related to drug safety that have not been detected from the premarketing studies (phases 1, 2, and 3) may now be observed in this large, general population.

Finally, another type of study frequently found in the postmarketing stage is to use the new drug for additional indications (symptoms or diseases). A drug developed for disease A may also be useful for disease B, but the sponsor may not have sufficient resources (budget, manpower) to develop the drug for both indications at the same time. In this case, the sponsor may decide first to develop the drug for disease A and obtain approval for the drug to be on the market and then to develop it for disease B. Phase 4 studies designed for "new indications" are also very common.

CONCLUSION

In the process of discovering and developing new drugs, much scientific work is involved. Potential drug candidates progress from preclinical experiments to clinical trials. During the clinical development process, drugs are tested at phases 1, 2, and 3 before they are submitted to the FDA. They are then further tested at phase 4 after approval for marketing. This is a large-scale, time-consuming, and expensive process in which a large number of scientists with various training and experiences is involved. This is why it takes substantial investments and a long time to discover and develop new drugs.

REFERENCES

1. U.S. FDA. *Guideline for the Format and Content of the Clinical and Statistical Sections of an Application*; 1988.
2. Hubert, J.J.; Bohidar, N.R.; Peace, K.E. Assessment of Pharmacological Activity. In *Biopharmaceutical Statistics for Drug Development*; Peace, K.E., Ed.; Marcel Dekker: New York, 1988.
3. Redman, C.E. Screening Compounds for Clinically Active Drugs. In *Statistics in the Pharmaceutical Industry*; Buncher, C.R., Tsay, J.Y., Eds.; Marcel Dekker: New York, 1981.
4. Selwyn, M.R. Preclinical Safety Assessment. In *Biopharmaceutical Statistics for Drug Development*; Peace, K.E., Eds.; Marcel Dekker: New York, 1988.
5. U.S. FDA. *The Modernization Act*; The United States Congress: Washington, DC, 1997.
6. Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*; Marcel Dekker: New York, 1992.
7. Ruberg, S.J. Dose response studies: I. Some design considerations. *J. Biopharm. Stat.* **1995**, 5 (1), 1–14.

Drug Interchangeability: Biosimilar Products

Laszlo Endrenyi

Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Laszlo Tothfalusi

Department of Pharmacodynamics, Semmelweis University, Budapest, Hungary

Abstract

When the patent of a biological drug expires, alternative, biosimilar products are often developed. They are expected to have highly similar clinical effects with the original drug. Still, even after the approval of the biosimilarity of two drug products, their interchangeability within subjects remains to be determined. Regulatory authorities in the United States, Canada and the European Union follow differing approaches to the approval of interchangeability. The disbursement of funds for drugs and, consequently, the decisions for interchangeability are generally left for the states, provinces and constituent states. This leads to widely diverging practices and, in the United States, possible legal complications. Statistical approaches are presented for the determination of the switching, alternation and interchangeability of biological products and for the design of the corresponding studies.

Keywords: Biosimilarity; Interchangeability; Switching; Alternation; Regulatory approaches; Statistical evaluation.

INTRODUCTION

Following the expiry of the patent of an innovative biological drug, alternative, biosimilar products are often developed. Biological drugs are therapeutic agents which are produced using a living system or organism. However, the size and complexity of biological molecules are much larger, their manufacturing processes are more complicated and difficult than those of chemistry-based drugs, and they may exhibit immunogenicity (unwanted immune response when the immune system recognizes a large molecule as a foreign invader). Consequently, an alternative biological product cannot be identical but only similar (i.e., biosimilar) to the drug of an originator. The two products should be highly similar so that their clinical effects should also be very close. The issue of what is “highly similar” is a matter of judgment and is based on the totality-of-the-evidence.^[1] This involves several considerations including structural, functional, analytical, immunological, pharmacokinetic, pharmacodynamics, manufacturing and clinical similarities.

Even if two products are judged and declared to be biosimilar, their interchangeability requires much more convincing evidence which can be obtained only from additional investigations and calls for further, careful judgment. For example, the Biologics Price Competition and Innovation (BPCI) Act of the United States sets

biosimilarity as a precondition for interchangeability and imposes additional requirements.^[2] These will be discussed later.

Thus, at least in the United States, approved biosimilar preparations may, without further permission, not be switched from and to their reference drug. There is a clear separation between their *prescribability* and *switchability*. Anderson and Hauck^[3] actually distinguished between these two terms even for small-molecule generic products. Two situations can be considered.

In the first situation, a patient had no prior exposure to a drug in any of its forms, i.e., s/he has been naïve to the drug. Under such conditions, any of the approved formulations are *prescribable* provided that their safety and efficacy are acceptable and, therefore, they have been approved by regulators. However, after the patient has started to receive one of the formulations, it may not be substituted by another product.

In the second situation, the patient is already taking the drug but opts for a cheaper alternative for economic reasons. In this case, the different products must be therapeutically closely similar. Therefore, the products will be *switchable* within individuals. The switch usually occurs from the reference drug to the test drug, which is denoted by “R → T”.

Switching between marketed drug products can become more general, and they can be entirely *interchangeable*.

Directions of *switches* are possible like the T→R and T→T' switch; here T' is a second biosimilar test product. Consequently, a more general *alternation* can be envisaged between the various biosimilar products as well as with the original drug. The products R, T and T' must be interchangeable in any possible sense to achieve a closely similar therapeutic effect after a switch and alternation.

Practices of and issues on the interchangeability of biological products in the United States, Canada and Europe will be described. Statistical approaches will then be presented. A few remarks will conclude this entry.

INTERCHANGEABILITY OF BIOSIMILARS

General Questions and Concerns

Klein et al.^[4] noted several questions for the careful consideration of substitution and interchangeability between biological products. They observed that the “complexity and impurity profile of an SEB means that automatic interchangeability of SEBs, or even of originator biologicals, has a real potential for different, and sometimes unexpected, clinical consequences”.

Another possible concern is that switching between biological products can exacerbate immunogenicity, with potentially negative effects. Klein et al.^[4] note that the immunogenicity of biosimilars “cannot be fully predicted using preclinical/clinical studies as the process is abbreviated and the number of patients/subjects in clinical studies is relatively limited.” For example, loss of tolerance and loss of efficacy was observed within 1 year when patients, having Crohn's disease and maintained with infliximab, were switched to adalimumab as compared to patients who remained on infliximab therapy.^[5]

Professional organizations raised concerns about switching patients from an effective and safe biological drug to a biosimilar. For instance, the British Society of Rheumatology^[6] advised against the summary switching of such patients. A committee of the European Society for Paediatric Gastroenterology, Hepatology, and Nutrition recommended that children with inflammatory bowel disease not be switched at present to a biosimilar.^[7]

Another concern is that the various responses of the biosimilar and also of the reference drug could drift in time.^[8] They form the basis of the totality-of-the-evidence for the judgment on biosimilarity. As a document of Health Canada observes, over time the biosimilar and reference products “make their own independent manufacturing changes, differences could be introduced that affect the drug products”.^[9]

United States

The Biologics Price Competition and Innovation (BPCI) Act of 2009 in the United States, clearly defines, separately, both biosimilarity and interchangeability.^[2]

The BPCI Act defines the term biosimilar (or biosimilarity) as when the biological product is *highly similar* to the reference product (notwithstanding minor differences in clinically inactive components), and also when there are *no clinically meaningful differences* between the biosimilar product and the reference product in terms of safety, purity and potency.

The BPCI Act of the United States provides explicit definitions and conditions for interchangeability. “The terms *interchangeable* and *interchangeability* mean that: (1) the biological product is *biosimilar* to the reference product; (2) the biological product can be *expected* to produce the *same clinical result* as the reference product *in any given patient*; (3) for a product administered more than once, the *risks of safety and reduced efficacy of alternating and switching* are not greater than with the use of the reference product without alternating or switching.” (Emphasis has been added.)

There is a clear and strong distinction between *biosimilarity* and *interchangeability*. Biosimilarity is a precondition of interchangeability. The regulatory approval of the biosimilarity of two products does not imply at all their interchangeability. If two biological products have already been approved and are marketed, an additional application must be submitted which would demonstrate that conditions for their interchangeability are also satisfied.

The strong distinction and hierarchy between the biosimilarity and interchangeability of biological products is in sharp contrast to the close parallel between the bioequivalence and interchangeability of small-molecule drug products.

However, while the BPCI Act distinguishes between biosimilarity and interchangeability, such a distinction is still to be seen in the forthcoming FDA guidelines. This has already led to legal controversies.

FDA published in January, 2017 a draft guidance on the interchangeability.^[10] It defines switching and alternating as changes occurring between the biosimilar and reference products. It describes a dedicated switching study design in which a lead-in period of treatment with the reference product is followed by a randomized two-arm period. One arm incorporates switchings between the proposed interchangeable product and the reference product whereas subjects in the other arm continue to receive the reference drug. An integrated study design could assess both the high similarity of the two drug products and their interchangeability.

FDA strongly recommends that patients be used in switching studies. If a sponsor intends to involve healthy subjects, the possible risks should be discussed with the agency. FDA strongly recommends also that U.S.-licensed reference products be used in switching studies.

Automatic Substitution in the United States

The BPCI Act of the United States decrees also important consequences for the approval of interchangeability

between biological drug products. The BPCI Act states that the interchangeable “biological product may be substituted for the reference product without the intervention of the health care provider who prescribed the reference product”.^[12]

Consequently, according to the BPCI Act, if a test product is judged to be interchangeable with the reference product then it may be substituted, even alternated, without a possible intervention, or even notification, of the prescribing physician.

This is a very permissive stipulation and has raised concerns at the level of state legislations. The basis of these concerns is that substitution of a drug product by another could give rise to changes of the safety and efficacy in individuals. It is assumed, however, that risks associated with these changes are well controlled.

For instance, when a patient is switched from a reference formulation to a test product, $R \rightarrow T$, then the associated risk should not be higher than when the subject is exposed to the reference product twice: $R \rightarrow R'$. This was the sense of the earlier considerations of individual bioequivalence.

Scenarios of switching and alternating have not been stated yet by the FDA. On the other hand, as of January 4, 2016, at least 31 states have considered legislation establishing state standard for substitution of biosimilar prescription product to replace the original biological product.^[11] It was signed into law in 18 states and Puerto Rico.

The language of the legislations can be confusing. They generally refer to the substitution of a prescribed biological product by an alternative biological product, as in California. This may mean alternative prescription but may imply also interchanging. It is therefore reassuring that the legislations generally contain various safeguards and conditions.^[11] For example, they often refer to the need of approval of interchangeability by FDA. Some require that the prescriber and/or the patient be notified. The prescriber's active or passive (“does not object”) approval of substitution is generally sought. Some provide liability protections for pharmacists who substitute biosimilars.

However, there is pressure to extend automatic substitution. For example, the Generic Pharmaceutical Association stated in 2015: “The Generic Pharmaceutical Association (GPhA) and its Biosimilars Council applaud the enactment of legislation in five states to allow automatic substitution for Food and Drug Administration (FDA) approved interchangeable biologic products.”^[12] The statement appears to reach somewhat beyond the intent and content of the legislations: interchangeability is not automatic but subject to conditions and constraints.

The conditions and constraints of the state legislations are more cautious (for instance, when they refer to the approval of substitution by the prescriber) than the federal BPCI Act which, as seen above, calls for substitution by biosimilars without the intervention of the prescriber. The differing views may require legal resolution.

Canada

In Canada, the term “substitutability” is applied to two products which can both be used in lieu of the other during and within the same treatment period, i.e. for interchangeability as discussed in this paper. The general guidance of Health Canada on biosimilars does not consider interchangeability.^[13] Still, the position of Health Canada on the substitution of biological products is clear but rather soft: “Health Canada does not support the automatic substitution of a subsequent-entry biologic for its reference biologic drug. Health Canada therefore recommends that physicians make only well-informed decisions regarding therapeutic interchange”.^[9]

The position of Health Canada does not have legal authority. Funds for the reimbursement of pharmaceutical expenses are dispensed by the provinces. They will be keen to promote introduction of new biosimilar products, i.e. their prescribability. The interchangeability of biologic products, and the related conditions, will be diverse and controversial.

European Union

The EU was the first region in the world to have set up legal framework and regulatory pathways for biosimilars. In fact, the word “biosimilars” was coined during this legislative process.

The concept of a “similar biological medicinal product” was adopted in EU pharmaceutical legislation in 2004 and came into effect in 2005. The EMA was the first to lay down an abbreviated regulatory pathway for biosimilars. As of April 2016, 23 biosimilars were approved within the product classes of human growth hormone, granulocyte colony-stimulating factor, erythropoietin, follitropin, TNF-inhibitor and insulin for use in the EU; some products were withdrawn.^[14] The European Medicines Agency has issued many guidelines for biosimilars, including several product-class specific guidelines. A general guideline on biosimilars was issued in 2014.^[15]

Only EMA can grant marketing authorization for biosimilars in the European Union. More precisely, the European Commission issues the decisions concerning the authorization of these medicinal products on the basis of the scientific opinions from the EMA. The resulting marketing authorization is valid in all EU Member States.

For granting approval for a biosimilar product, EMA generally requires clinical trials, including comparability studies to the originator product, to demonstrate safety and efficacy. Unlike generics, demonstration of bioequivalence is not enough for biosimilars. Additionally, the prospective market authorization holder also must demonstrate lack or at least comparable immunogenicity in long-term clinical trials. Guidelines were established to provide further details on specific needs for demonstrating biosimilarity for nine primary product classes.^[16]

The EMA has made it clear that a biosimilar is not the same as a generic drug and handed the interchangeability issue over to individual countries. As an EMA document states:^[17] “The EMA evaluates biosimilar medicines for authorization purposes. The Agency’s evaluations do not include recommendations on whether a biosimilar should be used interchangeably with its reference medicine. For questions related to switching from one biological medicine to another, patients should speak to their doctor or pharmacist.”

In the absence of central European directives, the 28 member states follow diverse directions. In several countries, either legislation has been passed or regulations prohibit automatic substitution by pharmacists.^[17] Given that most currently approved biosimilars are applicable only in hospital settings, the impact of such a strict prohibition is moderate. At the other extreme, some countries (e.g., Poland and Bulgaria) have no relevant law or guidelines and permit automatic substitution. Substitution is conditional in several other countries.^[18]

Equally important is the reimbursement policy^[19] where there are many national variations. One of the options is that new patients need to be treated with a biosimilar product if it is available, otherwise the treatment is not reimbursed. But patients already treated with a brand-name biological have the right to continue receiving the same biological brand. Consequently, the prescribability but not the interchangeability of biosimilar products is recommended.

There has been considerable experience in Europe with switching. No evidence has been found from either clinical trials or post-marketing surveillance that switching to and from different biopharmaceuticals, including biosimilar medicines, leads to safety concerns.^[20]

To describe the different scenarios, a European Consensus Group on Biosimilars proposed to make distinction between interchangeability, switching and substitution. According to the consensus document:^[21]

- Interchangeability means changing one medicine with the agreement of the prescriber.
- Switching means a decision by the treating physician to exchange one medicine for another.
- Substitution means of the practice of dispensing one medicine instead of another equivalent and interchangeable medicine at the pharmacy level without consulting the prescriber.

Making such distinctions between the post-approval use of biosimilars may be needed to describe the different possibilities. But reloading the meaning of these common English words which in some countries have been used as synonyms, could lead to confusion. The word switching appears to be a particularly bad choice of wording because some English speakers seem to associate switching with the inadvertent exchange of medicines. Moreover, confusion could arise, and has arisen, because the terms could be interpreted for the conditions of both prescribability (between subjects) and switch ability (within subjects).^[22]

STATISTICAL CONSIDERATIONS ON THE INTERCHANGEABILITY OF BIOLOGICAL PRODUCTS

A Possible Resolution of the Problematic Definition of Interchangeability

The definitions of the American BPCI Act for the interchangeability of biologicals were outlined earlier. One of them is particularly forbidding, apparently categorical: “the biological product can be *expected* to produce the *same* clinical result as the reference product in *any* given patient”. Responses of patients are variable and the same clinical response may not be anticipated. Furthermore, the expectation of the *same* result in *any* given patient could be interpreted to mean that if an individual experiences an unexpected (not to say, not the “same”) therapeutic or adverse effect s/he may well seek recourse. Lawyers, notably in the United States, would undoubtedly encourage this.

However, the word “*expected*” in the above phrase could be interpreted that the “same clinical result ... in any given patient” is not a categorical rule but only an expectation. “Expectation” has also a probabilistic meaning which involves some stated statistical assurance. It is hoped that this interpretation will be duly considered and adopted.

Study Designs for Interchangeability

While many biologicals have long terminal half-lives, some don’t. With these, it is possible to contemplate cross-over investigations which would enable the assessment of features of interchangeability. Notably, switching between two biological products, either reference (R) or test (T), such as R→T, T→R, R→R’, T→T’, could be investigated by Balaam’s 4 × 2 crossover design with 4 sequences and 2 periods,^[23,24] including the sequences of RT, TR, RR’, TT’. Here, e.g., RT refers to a sequence in which first the reference and then, after a sufficient washout period, the test product is provided to each individual. R and R’ denote two administrations of the reference product. Similarly, alternating between the test and reference products, i.e., R→T→R’, T→R→T’, could be studied by a 2-sequence, 3-period design with the sequences of RTR’, TRT’. Finally, both switching and alternating could be investigated by modifying Balaam’s design in 4 sequences and either 2 or 3 periods: TT’, RR’, TRT’, RTR’.

Statistical Assessment of Switching and Alternation

Assessment of Biosimilarity

In order to quantify switching and alternating and to provide statistical measures for their evaluation, Chow^[25] developed relevant measures. First, however, indices are considered for the quantitative assessment of biosimilarity.

For a given criterion of biosimilarity and under a valid study design, a *local biosimilarity index* can be obtained for a given functional area or domain by the following steps:^[26]

- Step 1: Assess average biosimilarity based on a given criterion, e.g. (80%, 125%) based on log-transformed data;
- Step 2: Calculate the *local biosimilarity index* (i.e., reproducibility) based on the observed ratio and variability;
- Step 3: Claim *local biosimilarity* if the 95% confidence lower bound of p is larger than p_0 , a pre-specified number, where p_0 can be obtained based on an estimated of reproducibility probability for a study comparing a reference product to itself (the reference product), i.e., an R-R study.

A *totality biosimilarity index* can be derived across all functional areas or domains can be obtained by the following steps:

- Step 1: Obtain $\hat{p}_i$, the biosimilarity index for the i -th domain;
- Step 2: Define the *totality biosimilarity index* as $\hat{p}_T = \sum_{i=1}^n w_i \hat{p}_i$ where w_i is the weight for the i -th domain, where $i = 1, 2, \dots, K$ (the number of domains or functional areas);
- Step 3: Claim biosimilarity if the 95% confidence lower bound of p_T is larger than a pre-specified value p_0 , which can be determined based on an estimated of *totality biosimilarity index* for studies comparing a reference product to itself (the reference product).

This *totality biosimilarity index* has the advantages that (1) it is robust with respect to the selected study endpoint, biosimilarity criteria, and study design, (2) it takes variability into consideration (one of the major criticisms in the assessment of average bioequivalence), (3) it allows the definition and assessment the degree of similarity (in other words, it provides partial answer to the question that “how similar is considered similar?”), and (4) the use of the biosimilarity index or totality biosimilarity index will reflect the sensitivity of heterogeneity in variance.

Assessment of Switching and Alternation

A similar idea can be applied to develop a *switching index* under an appropriate study design such as the crossover 4×2 Balaam design.^[27] Thus, the issue of switching needs to be addressed for the switches of $R \rightarrow T$, $T \rightarrow R$, $T \rightarrow T'$ and $R \rightarrow R'$.

Define $\hat{p}_T$ the totality biosimilarity index for the i -th switch, where $i = 1$ (the $R \rightarrow R'$ switch), 2 (for $T \rightarrow T'$), 3

(for $R \rightarrow T$), and 4 (for $T \rightarrow R$). As a result, a *switching index* (SI) can be calculated as follows:

- Step 1: Obtain $\hat{p}_T$; $i = 1, \dots, 4$;
- Step 2: Define the switching index as $SI = \max_i \{\hat{p}_T\}$, $i = 1, \dots, 4$, which is the largest order of the biosimilarity indices;
- Step 3: Claim switchability if the 95% confidence lower bound of p_s , $i = 1, \dots, 4$, is larger than a pre-specified value.

An *alternating index* can be developed similarly under an appropriate study design. Under the modified crossover Balaam design of (TT, RR, TRT, RTR), alternation can be evaluated with the $R \rightarrow T \rightarrow R$ and $T \rightarrow R \rightarrow T$ designs. For example, the assessment of differences between the switches of $R \rightarrow T$ and $T \rightarrow R'$ for the alternating of $R \rightarrow T \rightarrow R$ needs to be evaluated in order to determine whether the drug effect has returned to the baseline after the second switch. The *alternating index* can be calculated as follows:

Define p_{Ti} as the totality biosimilarity index for the i -th switch, where $i = 1$ refers to the switch $R \rightarrow R'$, $i = 2$ to the switch $T \rightarrow T'$, 3 to $R \rightarrow T$, and 4 to $T \rightarrow R$. As a result, the *alternating index* (AI) can be calculated as follows:

- Step 1: Obtain $\hat{p}_{Ti}$, $i = 1, \dots, 4$;
- Step 2: Define the range of these indexes, $AI = \max_i \{\hat{p}_{Ti}\} - \min_i \{\hat{p}_{Ti}\}$, $i = 1, \dots, 4$, as the alternating index;
- Step 3: Claim alternation if the 95% confidence lower bound of $\max_i \{p_{Ti}\} - \min_i \{p_{Ti}\}$, $i = 1, \dots, 4$, is larger than a pre-specified value p_{A0} .

Scaled Criterion for Drug Interchangeability (SCDI)

A scaled criterion proposed for drug interchangeability is based on the model for individual bioequivalence^[28]. Its features, as described below, include (i) the criterion for average bioequivalence adjusted for the within-subject variability of the reference product, and (ii) a correction for the variability due to subject-by-product variability (i.e. σ_D^2). The proposed criterion for assessing interchangeability (i.e., switching and alternating) is briefly derived below.

The model for average BE expects that the confidence interval around the difference between the logarithmic means of the test and reference products (μ_T and μ_R , respectively) should be between the upper and lower confidence limits ($-\theta$ and θ):

$$-\theta \leq \mu_T - \mu_R \leq \theta \quad (1)$$

The model for individual bioequivalence (IBE) was implemented in the statistical guidance of FDA^[29]. The model contrasts not only the squared deviation between the two logarithmic means, $\delta^2 = (\mu_T - \mu_R)^2$, but also the

difference between the corresponding within-subject variances, $(O_{WT}^2 - O_{WR}^2)$. It contains also an additional term for the variance component of the subject-by-formulation interaction, O_D^2 :

$$\theta_1 = (\delta^2 + O_{WT}^2 - O_{WR}^2 + O_D^2) / \max\{O_{W0}^2, O_{WR}^2\} \quad (2)$$

Here O_{W0}^2 is a constant the magnitude of which is set by regulators. IBE can be claimed if the one-sided 95% upper confidence bound for θ_1 is less than a pre-specified bioequivalence limit.

For the evaluation of SCDI, the model for average BE is first adjusted for the within-subject variability:

$$-\theta' \leq \frac{\mu_T - \mu_R}{\sigma_W} \leq \theta', \quad (3a)$$

or

$$-\theta' \sigma_W \leq \mu_T - \mu_R \leq \theta' \sigma_W, \quad (3b)$$

where σ_W^2 is a within-subject variation and θ' is the limit for scaled bioequivalence. In practice, σ_{WR}^2 , the within-subject variation of the reference product is often considered.

Scaled average bioequivalence (SABE) has been applied for the assessment of BE between highly variable drug products.^[30,31]

Next consider the first two components of the individual bioequivalence model (Eq. 2):

$$\frac{(\mu_T - \mu_R)^2 + \sigma_D^2}{\sigma_W^2} = \frac{2\delta\sigma_D + (\delta - \sigma_D)^2}{\sigma_W^2} \quad (4)$$

where $\delta = \mu_T - \mu_R$. When δ and σ_D are close, we observe that

$$\frac{(\delta + \sigma_D)^2}{\sigma_W^2} \approx \frac{2\delta\sigma_D}{\sigma_W^2}. \quad (5)$$

The assumption is reasonable when both δ and σ_D are small.

Thus, the proposed scaled criterion for drug interchangeability (SCDI) is:

$$-\theta' \leq \left(\frac{\mu_T - \mu_R}{\sigma_W} \right) \frac{2\sigma_D}{\sigma_W} \leq \theta' \quad (6a)$$

Now, let $f = \sigma_W / (2\sigma_D)$, a correction factor for drug interchangeability. Then, the proposed SCDI criterion is given by:

$$-\theta' f \sigma_W \leq \mu_T - \mu_R \leq \theta' f \sigma_W \quad (6b)$$

Following the concept of the criterion for individual bioequivalence and the idea of SABE, the proposed SCDI

criterion is developed in order to adjust the usual one-size-fits-all approach for both intra-subject variability of the reference product and the variability due to subject-by-product interaction. As compared to SABE, SCDI may result in wider or narrower limits depending upon the correction factor f which is a measure of the relative magnitude between σ_{WR} and σ_D .

The proposed SCDI criterion for drug interchangeability depends upon the selection of regulatory constants for σ_{WR} and σ_D . In practice, the observed variabilities may deviate far from the regulatory constants. Thus, it is suggested that the following hypotheses be tested before the use of SCDI criterion:

$$H_{01} : \sigma_{WR} > \sigma_{W0} \quad \text{vs.} \quad H_{a1} : \sigma_{WR} > \sigma_{W0}, \quad (7a)$$

and

$$H_{02} : \sigma_D > \sigma_{D0} \quad \text{vs.} \quad H_{a2} : \sigma_D > \sigma_{D0} \quad (7b)$$

If we fail to reject the null hypotheses H_{01} or H_{02} then we will stick with the suggested individual regulatory constants; otherwise estimates of σ_{WR} and/or σ_D should be used in the SCDI criterion.

However, it should be noted that statistical properties and/or the finite sample performance of SCDI with estimates of σ_{WR} and/or σ_D are not well established. Further research is needed. Preliminary analyses of statistical power and required sample size have already been conducted.^[32]

CONCLUDING REMARKS

The important concepts of prescribability and switchability should be kept in mind when pharmacokinetic features of two drug products are compared. Regulatory approval of both small-molecule generics and biological biosimilars means directly an agreement only that they are prescribable, i.e., that the drug products may be provided to naïve patient who have not taken the drug until this point in any form.

The terminology for the interchangeability of biological products has become regrettably confusing. In the United States, the BPCI Act introduced the terms of switching and alternating without providing clear interpretation. Switching presumably refers to changing between the reference and test products, R→T and T→R, primarily the former. Alternating could involve multisource products, thus also a change of T→T'. However, a recent interpretation limited alternation to between switches between the biosimilar and reference drug products.^[10] Interchangeability would cover both switching and alternating. All these changes are conducted within subjects. It is important to emphasize this because the terms are used at times, in the literature and regulations of the states, incorrectly, when they

imply changes between subjects, i.e., when they refer to the condition of prescribability.

The situation is different, but not less confusing in the European Union. The intent of the proposed terms of switching, interchanging and substitution may be that they be considered as changes within patients.^[21] However, they are often used and implemented for changes between subjects, i.e. again in the sense of prescribability.

The European Medicines Agency has issued product-related guidelines for biosimilars, as well as a general guideline in 2014, and approved a large number of biosimilars. In some cases, these were multisource products referring to the same reference drug. No safety issues were noted following interchanging among them.^[20]

However, EMA left the decision on interchanging to member states. In some member states, prescribability and switchability are governed by the rules of health insurance policy and not by formal law. As a consequence, the regulations have spanned a wide range from those enabling automatic substitution to laws forbidding automatic substitution.^[18] This is a very regrettable situation.

FDA has followed a more cautious approach. Relying on the BPCI Act, it has issued general guidelines on various aspects on the development and regulation of biosimilars. The consideration of interchangeability requires additional demonstration; the conditions for these have not been clarified yet. FDA approved the first product (Zarxio) in March, 2016 for which the condition of interchangeability was not requested and not granted.

A situation could arise in the United States where there will be a number of biosimilar products which have been shown to be biosimilar to the same innovative (reference) drug product. It is almost impossible for a sponsor to claim or demonstrate interchangeability according to the definition as given in the BPCI Act. Alternatively, it is possible that a meta-analysis that combines all data given in the submissions be conducted (by the regulatory agency) to evaluate the relative risks of switching and alternating of drug interchangeability. In practice, meta-analysis can be conducted for safety monitoring of approved biosimilar products. This is extremely important especially when there are a number of biosimilar products in the marketplace.

REFERENCES

1. FDA. Guidance on Scientific considerations in demonstrating biosimilarity to a reference product. The United States Food and Drug Administration, Silver Spring, MD, 2015.
2. Federal Register. Biologics Price Competition and Innovation Act. §7002(a)(2), 2009. Available from: <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/ucm216146.pdf>.
3. Anderson, S.; Hauck, W.W. Consideration of individual bioequivalence. *J. Pharmacokin. Biopharm.* **1990**, *18* (3), 259–273.
4. Klein, A.V.; Wang, J.; Bedford, P. Subsequent entry biologics (biosimilars) in Canada: approaches to interchangeability and the extrapolation of indications and uses. *Generics Biosim. Initiative J.* **2014**, *3* (3), 150–154.
5. Van Assche, G.; Vermeire, S.; Ballet, V.; Gabriels, F.; Noman, M.; D'Haens, G.; Claessens, C., et al. Switch to adalimumab in patients with Crohn's disease controlled by maintenance infliximab: Prospective randomised SWITCH trial. *Gut* **2012**, *61* (2), 229–234.
6. British Society for Rheumatology. Position statement on biosimilar medicines. 2015. http://www.rheumatology.org.uk/about_bsr/press_releases/bsr_supports_the_use_of_biosimilars_but_recommends_measures_to_monitor_safety.aspx.
7. Ridder, L.; Waterman, M.; Turner, D.; Bronsky, J.; Hauer, A.C.; Dias, J.A.; et al. Use of biosimilars in paediatric inflammatory bowel disease: A position statement. *J. Pediatr. Gastroenterol. Nutr.* **2015**, *61* (4), 503–508.
8. Schiestl, M.; Stangler, T.; Torella, C.; Čepeljnik, T.; Toll, H.; Grau, R. Acceptable changes in quality attributes of glycosylated biopharmaceuticals. *Nature Biotechnol.* **2011**, *29*, 310–312.
9. Health Canada: Questions and answers to accompany the final Guidance for Sponsors: Information and Submission Requirements for Subsequent Entry Biologics (SEBs). Question 15. Ottawa, ON, 2010. http://www.hc-sc.gc.ca/dhp-mps/brgtherap/applic-demande/guides/seb-pbu/notice-avis_seb-pbu_2010-eng.php.
10. FDA. Draft guidance for industry: Considerations in demonstrating interchangeability with a reference product. Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER). January 12, 2017. <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM537135>.
11. Cauchi, R. State laws and legislation related to biological medications and substitutions of biosimilars. National Conference of State Legislatures (NCSL): Washington, DC, 2016. <http://www.ncsl.org/research/health/state-laws-and-legislation-related-to-biologic-medications-and-substitution-of-biosimilars.aspx#Mandatory>.
12. GPhA. GPhA and Biosimilars Council praise new state laws to allow automatic substitution for interchangeable biologics. Press release of Generic Pharmaceutical Association, May 13, 2015. <http://www.gphaonline.org/gpha-media/press/gpha-and-biosimilars-council-praise-new-state-laws-to-allow-automatic-substitution-for-interchangeable-biologics>.
13. Health Canada. Guidance document: Information and submission requirements for biosimilar biologic drugs. Ottawa, ON. November 14, 2016. <http://www.hc-sc.gc.ca/dhp-mps/brgtherap/biosimilars-biosimilaires-index-eng.php>.
14. Declerck, P.; Farouk-Rezk; Rudd, P.M. Biosimilarity versus manufacturing change: two distinct concepts. *Pharm. Res.* **2016**, *33*, 261–268.
15. EMA. Guideline on similar biological medicinal products. European Medicines Agency: London, October 23, 2014. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/10/WC500176768.pdf.

16. Weise, M.; Bielsky, M.C.; De Smet, K., et al. Biosimilars: what clinicians should know. *Blood*. **2012**, *120*, 5111–5117.
17. EMA. Questions and answers on biosimilar medicines (similar biological medicinal products). European Medicines Agency: London, 2009. http://www.ema.europa.eu/docs/en_GB/document_library/Medicine_QA/2009/12/WC500020062.pdf.
18. Thimmaraju, P.K.; Rakshambikai, R.; Farista, R.; Jurulu, K. Legislations on biosimilar interchangeability in the US and EU – developments far from visibility. *Generics Biosim. Initiative GaBI Online*, 2015. <http://www.gabionline.net/Sponsored-Articles/Legislations-on-biosimilar-interchangeability-in-the-US-and-EU-developments-far-from-visibility>.
19. Declerck, P.J.; Simoons, S. European perspective on the market accessibility of biosimilars. *Biosimilars*. **2012**, *2*, 33–40.
20. Ebberts, H.C.; Muenzberg, M.; Schellekens, H. The safety of switching between therapeutic proteins. *Expert Opin. Biol. Ther.* **2012**, *12* (11), 1473–1485.
21. Ebberts, H.C.; Chamberlain, P. Interchangeability. An insurmountable fifth hurdle? *Generics and Biosim. Initiative J.* **2014**, *3* (2), 88–93.
22. Tothfalusi, L.; Endrenyi, L.; Chow, S.C. Statistical and regulatory considerations in assessments of interchangeability of biological drug products. *Eur. J. Health Econ.* **2014**, *15* (Suppl. 1), S5–S11.
23. Endrenyi, L.; Chang, C.; Chow, S.C.; Tothfalusi, L. On the interchangeability of biologic drug products. *Statistics in Medicine* **2013**, *32* (3), 434–441.
24. Lu, Y.; Chow, S.C.; Zhang, Z.Z. Statistical designs for assessing interchangeability of biosimilar products. *Drug Designing* **2013**, *2* (3), 2–6.
25. Chow, S.C. Quantitative evaluation of bioequivalence/biosimilarity. *J. Bioeq. Bioavil.* **2011**, Suppl. 1–002, 1–8.
26. Hsieh, T.C.; Chow, S.C.; Yang, L.Y.; Chi, E. The evaluation of biosimilarity index based on reproducibility probability for assessing follow-on biologics. *Stat. Med.* **2013**, *32* (3), 406–414.
27. Chow, S.C.; Yang, L.Y.; Starr, A.; Chiu, S.T. Statistical methods for assessing interchangeability of biosimilars. *Statistics in Medicine* **2013**, *32* (3), 442–448.
28. Chow, S.C.; Xu, H.L.; Endrenyi, L.; Song, F.Y. A new scaled criterion for drug interchangeability. *Chinese J. Pharm. Anal.* **2015**, *35* (5), 844–848.
29. Food and Drug Administration. *Guidance on statistical approaches to establishing bioequivalence*. U.S. Food and Drug Administration, Center for Drug Evaluation and Research (CDER): Rockville, MD, 2001. <http://www.fda.gov/downloads/Drugs/Guidances/ucm070244.pdf>.
30. Haidar, S.H.; Davit, B.; Chen, M.L.; et al. Bioequivalence approaches for highly variable drugs and drug products. *Pharm. Res.* **2008**, *25*, 237–241.
31. Tothfalusi, L.; Endrenyi, L.; Garcia Areta, A. Evaluation of bioequivalence for highly variable drugs with scaled average bioequivalence. *Clin. Pharmacokin.* **2009**, *48*, 725–743.
32. Chow, S.C.; Song, F.; Chen, M. Some thoughts on drug interchangeability. *J. Biopharm. Stat.* **2016**, *26* (1), 178–186.

Drug Interchangeability: Criteria and Design

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Fuyu Song

Peking University Health Science Center, Peking University Clinical Research Institute, Beijing, China

Meng Chen

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

Current regulation for generic approval is based on the assessment of average bioequivalence. As indicated by the United States Food and Drug Administration (FDA), an approved generic drug can be used as a substitute for the innovative drug. The FDA does not indicate that two generic copies of the same innovative drug can be used interchangeably even though they are bioequivalent to the same brand-name drug. In practice, bioequivalence between generic copies of an innovative drug is not required. However, as more generic drug products become available, it is a concern whether the approved generic drug products have the same quality and therapeutic effect as the brand-name drug product and whether they can be used safely and interchangeably. In this article, several criteria including a newly proposed criterion for assessing drug interchangeability are studied. In addition, comments on possible study designs and power calculation for sample size under a specific design are also discussed.

Keywords: Biosimilars; Generic drugs; Individual bioequivalence; Population bioequivalence; Scaled average bioequivalence; Scaled criterion for drug interchangeability.

INTRODUCTION

In the United States, when a drug product is going off-patent, a generic company can file an abbreviated new drug application (ANDA) for generic approval. For approval of a generic drug, the Food and Drug Administration (FDA) requires that evidence of average bioequivalence in drug absorption be provided through the conduct of a bioequivalence study on healthy volunteers. As indicated by the FDA, an approved generic (test) drug of an innovative (reference) drug can serve as a substitute for the reference product. The FDA, however, did not indicate that the approved generic drug and the innovative drug can be used interchangeably.

As indicated earlier, when a generic drug is claimed bioequivalent to an innovative drug product, it is assumed that they are therapeutically equivalent. A generic drug can be used as a substitute for the innovative drug if it has been shown to be bioequivalent to the innovative drug. The FDA does not indicate that two generic copies of the same brand-name drug can be used interchangeably even though they are bioequivalent to the same brand-name drug. In practice, bioequivalence between generic copies of a brand-name drug is not required. However, as more generic drug products become available, it is a concern whether the approved generic drug products have the same quality

and therapeutic effect as the brand-name drug product and whether they can be used safely and interchangeably.

The issue of drug interchangeability has been discussed tremendously among the regulatory agency, pharmaceutical industry, and academia since the early 1990s. Several criteria for the assessment of drug interchangeability have been proposed since then. These criteria include the criteria for population bioequivalence (PBE) and individual bioequivalence (IBE),^[1,2] the potential use of variability due to the subject-by-formulation interaction,^[2,3] and a scaled criterion for drug interchangeability (SCDI).^[5] In this entry, several criteria and study designs for drug interchangeability are discussed.^[6] Under these criteria, formulas for sample size calculation can be derived.

In Section “Drug Interchangeability,” the concept of drug interchangeability in terms of drug prescribability and drug switchability is introduced. Also included in this section are the criteria for PBE, IBE, and SCDI, and a newly proposed criterion. Section “Design for Drug Interchangeability” provides some common study designs considered for addressing drug interchangeability. Formulas derived under these criteria are given in Section “Power Analysis for Sample Size.” Some comments on the study designs for drug interchangeability are provided in Section “Concluding Remarks.”

DRUG INTERCHANGEABILITY

The concept of drug interchangeability for small-molecule drug products includes drug prescribability and drug switchability. Drug prescribability is referred to as the physician's choice for prescribing an appropriate drug for his/her patients between the brand-name drug and its generic copies. Drug switchability, on the other hand, is referred to as the switch from a drug (e.g., a brand-name drug or its generic copies) to another (e.g., a generic copy) within the same patient whose concentration of the drug has been titrated to a steady, efficacious, and safe level. In what follows, criteria for assessment of PBE for addressing drug prescribability and IBE for addressing switchability,^[1,2] a possible criterion based on the variability due to the subject-by-treatment interaction,^[3,4] an SCDI,^[5] and a newly proposed criterion for drug interchangeability are briefly introduced.

CRITERIA FOR PBE AND IBE

Population Bioequivalence

To address drug prescribability, the FDA recommends that PBE be assessed. In addition to the average of bioavailability, PBE focuses on the variability of bioavailability. The 2001 FDA guidance recommends the following criterion to be used for assessing PBE:

$$\theta_p = (\delta^2 + \sigma_{TT}^2 - \sigma_{TR}^2) / \max\{\sigma_{T0}^2, \sigma_{TR}^2\}, \quad (1)$$

where $\delta = \mu_T - \mu_R$, μ_T and μ_R are the mean responses of the test product and the reference product, respectively; $\sigma_{TT}^2, \sigma_{TR}^2$ are the total variances for the test product and the reference product, respectively; and σ_{T0}^2 is the scale parameter specified by the regulatory agency or the sponsor. PBE can be claimed if the one-sided 95% upper confidence bound for θ_p is less than a prespecified bioequivalence limit. In view of the above PBE criterion, PBE can be claimed if the null hypothesis in

$$H_0 : \lambda \geq 0 \text{ vs. } H_a : \lambda < 0$$

is rejected at the 5% level of significance and the observed geometric means ratio (GMR) is within the limits of 80% and 125%, where

$$\lambda = \delta^2 + \sigma_{TT}^2 - \sigma_{TR}^2 - \theta_p \max\{\sigma_{TR}^2, \sigma_{T0}^2\}$$

and θ_p is a constant specified in the 2001 FDA draft guidance.

Individual Bioequivalence

To address drug switchability, the FDA suggests that IBE be assessed. The 2001 FDA guidance recommends the following criterion to be used for assessing IBE:

$$\theta_l = \frac{(\mu_T - \mu_R)^2 + \sigma_D^2 + (\sigma_{WT}^2 - \sigma_{WR}^2)}{\max\{\sigma_{w0}^2, \sigma_{WR}^2\}},$$

where $\delta = \mu_T - \mu_R$, σ_D^2 is the variance component due to the subject-by-treatment interaction between drug products, and σ_{WR}^2 and σ_{WT}^2 are intra-subject variances for the reference product and the test product, respectively. Thus, we have

$$\theta_l = (\delta^2 + \sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2) / \max\{\sigma_{w0}^2, \sigma_{WR}^2\}, \quad (2)$$

where σ_{w0}^2 is the scale parameter specified by the regulatory agency or the sponsor. In view of the above IBE criterion, IBE can be claimed if the null hypothesis in

$$H_0 : \gamma \geq 0 \text{ vs. } H_a : \gamma < 0$$

is rejected at the 5% level of significance, and the observed GMR is within the limits of 80% and 125%, where

$$\gamma = \delta^2 + \sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2 - \theta_l \max\{\sigma_{WR}^2, \sigma_{w0}^2\}$$

and θ_l is a constant specified in the 2001 FDA draft guidance.

CRITERION FOR DRUG INTERCHANGEABILITY BASED ON σ_D

For addressing drug interchangeability, the FDA suggests that PBE and IBE be assessed under replicated crossover designs such as a replicated 2×2 crossover design, i.e., (TRTR, RTTR) or a 2×3 two-sequence dual design, i.e., (TRT, RTR). Under the 2×2 replicated crossover design, the one-sided 95% upper confidence bound for θ_l can be obtained under the following statistical model:

$$y_{ijk} = \mu + F_l + W_{ijk} + S_{ikl} + \epsilon_{ijk}, \quad (3)$$

where μ is the overall mean, F_l is the fixed effect of the l th drug product, W_{ijk} 's are fixed period, sequence, and interaction effects, S_{ijk} is the random effect of the i th subject in the k th sequence under the l th drug product, and ϵ_{ijk} 's are independent random errors distributed as $N(0, \sigma_{wi}^2)$. It is assumed that S_{ijk} 's and ϵ_{ijk} 's are mutually independent. Under model Eq. 3, σ_D^2 is given by

$$\sigma_D^2 = \sigma_{BT}^2 + \sigma_{BR}^2 - 2\rho\sigma_{BT}\sigma_{BR}, \quad (4)$$

which is variance of $S_{ikT} - S_{ikR}$. Note that σ_D^2 is usually referred to as the variance due to the subject-by-drug interaction.

Chen et al.^[3] indicated that IBE is to address drug switchability and variability due to the subject-by-treatment interaction is an indicator of drug switchability.

Further, Chen et al.^[3] pointed out that a value of 0.15 for the estimation of the standard deviation due to the subject-by-drug product may be considered to be important. In other words, we can test as to whether σ_D is greater than 0.15 for drug interchangeability. However, Endrenyi et al.^[4] pointed out that positive bias observed for the estimation of the variability due to the subject-by-treatment interaction—in other words, about a quarter to one-third of estimates greater than 0.15. These estimates do not result from the true existence of the subject-by-formulation interaction rather than the large intra-subject variability of the reference formulation. The results from the FDA data sets exhibit almost an identical pattern.

SCALED CRITERION FOR DRUG INTERCHANGEABILITY

While the criterion for the determination of average biosimilarity is based on the BE requirement of (80.00%, 125.00%) using log-transformed data, the criterion for the assessment of interchangeability is not clear. As indicated by Chen et al.,^[3] variability due to the subject-by-drug formulation interaction will have an impact on drug interchangeability. Let σ_D^2 be the variance component due to the subject-by-formulation interaction. We would like to propose a criterion to adjust for both intra-subject variability and the variability due to subject-by-formulation variability. The current BE criterion adjusted for intra-subject variability leads to so-called scaled average bioequivalence (SABE) criterion, which is considered suitable for highly variable drug products. In addition to adjusting intra-subject variability, following the idea of IBE, we may further adjust the BE criterion with respect to the variability due to subject-by-formulation variability in order to have a more accurate and reliable assessment of interchangeability. As indicated in the 2001 FDA guidance on bioequivalence, the criterion for assessing IBE is given in Eq. 2.

Chow et al.^[5] proposed an SCDI based on the first two components of the criterion for IBE, which consists of (i) the criterion for average biosimilarity adjusted for intra-subject variability of the reference product (i.e., SABE) and (ii) the correction for variability due to subject-by-product variability (i.e., σ_D^2). The proposed SCDI for assessing interchangeability (i.e., switching and alternating) is briefly derived as follows:

Step 1: Unscaled ABE criterion—Let BEL be the BE limit, which generally equals 1.25. Thus, biosimilarity requires that $\frac{1}{\text{BEL}} \leq \text{GMR} \leq \text{BEL}$. This implies

$$-\log(\text{BEL}) \leq \log(\text{GMR}) \leq \log(\text{BEL})$$

or

$$-\log(\text{BEL}) \leq \mu_T - \mu_R \leq \log(\text{BEL}),$$

where μ_T and μ_R are the logarithmic means.

Step 2: SABE criterion—Difference in logarithmic means is adjusted for intra-subject variability as follows:

$$-\log(\text{BELS}) \leq \frac{\mu_T - \mu_R}{\sigma_w} \leq \log(\text{BELS})$$

or

$$-\log(\text{BELS})\sigma_w \leq \mu_T - \mu_R \leq \log(\text{BELS})\sigma_w,$$

where σ_w^2 is a within-subject variation and BELS is the BE limit for SABE. In practice, σ_{wR}^2 , the within-subject variation of the reference product, is often considered.

Step 3: Proposed SCDI—Considering the first two components of the IBE criterion, we have the following relationship:

$$\frac{(\mu_T - \mu_R)^2 + \sigma_D^2}{\sigma_w^2} = \frac{2\delta\sigma_D + (\delta - \sigma_D)^2}{\sigma_w^2},$$

where $\delta = \mu_T - \mu_R$. When δ and σ_D are close, we observe that

$$\frac{\delta^2 + \sigma_D^2}{\sigma_w^2} \approx \frac{2\delta\sigma_D}{\sigma_w^2}.$$

The assumption is reasonable when both δ and σ_D are small.

Thus, the proposed SCDI is

$$-\log(\text{BELS}) \leq \left(\frac{\mu_T - \mu_R}{\sigma_w} \right) \left(\frac{2\sigma_D}{\sigma_w} \right) \leq \log(\text{BELS}).$$

Now, let $f = \sigma_w/(2\sigma_D)$, a correction factor for drug interchangeability. Then, the proposed SCDI is given by

$$-\log(\text{BELS})f\sigma_w \leq \mu_T - \mu_R \leq \log(\text{BELS})f\sigma_w. \quad (5)$$

Note that statistical properties and finite sample performance need further research.

ALTERNATIVE CRITERION BASED ON REVERSE OF TEST AND REFERENCE PRODUCT

As indicated by the FDA, an approved generic (test) drug product can serve as a substitute for the innovative (reference) drug product, provided the generic drug product has been shown to be bioequivalent to the innovative drug product. A test product is said to be bioequivalent to a reference product if the 90% confidence interval for the GMR (i.e., μ_T/μ_R), say (L_1, U_1) , is totally within the bioequivalence limit (0.8, 1.25). Based on the similar idea, a reference product can serve as a substitute for a test product if the reference product is shown to be bioequivalent

to the test product (i.e., the test product is now considered the *reference* product in the *new* bioequivalence assessment). In other words, the 90% confidence interval for the GMR of μ_R/μ_T , say (L_2, U_2) , is totally within the bioequivalence limit (0.8, 1.25). In this case, we conclude that the test (*T*) product and the reference (*R*) product can be used interchangeably.

Let *A* be the event that $(L_1, U_1) \subset (0.8, 1.25)$ and *B* be the event that $(L_2, U_2) \subset (0.8, 1.25)$. Define $p_1 = P(\text{Switch from } T \text{ to } R) = P(A)$ and $p_2 = P(\text{Switch from } R \text{ to } T) = P(B)$. In practice, we would expect that $p_1 > p_2$. p_2 is expected to be low at extreme cases, i.e., $(L_2, U_2) \subset (0.8, 1.25)$ at either lower end or upper end. For interchangeability, we would expect that $p_3 = P(\text{Switch from } T \text{ to } R \text{ and switch from } R \text{ to } T) = P(A \text{ and } B)$ is greater than $\max(p_1, p_2)$. For a given bioequivalence study, point estimates and the corresponding confidence intervals for p_i , $i = 1, 2, 3$ are necessarily obtained for the evaluation of the proposed criterion for drug interchangeability.

DESIGN FOR DRUG INTERCHANGEABILITY

Balaam's Design

Balaam's design is a 4×2 crossover design, denoted by (*TT*, *RR*, *TR*, *RT*). Under a 4×2 Balaam's design, qualified subjects will be randomly assigned to receive one of the four sequences of treatments: *TT*, *RR*, *TR*, and *RT*. For example, subjects in sequence 1 of *TT* will receive the test (biosimilar) product first and then crossed over to receive the reference (innovative biological) product after a sufficient length of washout. In practice, a Balaam's design is considered the combination of a parallel design (the first two sequences) and a crossover design (sequences 3 and 4). The purpose of the part of parallel design is to obtain independent estimates of intra-subject variabilities for the test product and the reference product. In the interest of assigning more subjects to the crossover phase, an unequal treatment assignment is usually employed. For example, we may consider a 1:2 allocation to the parallel phase and the crossover phase. In this case, for a sample size of $N = 24$, 8 subjects will be assigned to the parallel phase and 16 subjects will be assigned to the crossover phase. As a result, four subjects will be assigned to sequences 1 and 2, while eight subjects will be assigned to sequences 3 and 4, assuming that there is a 1:1 ratio treatment allocation within each phase. Under the 4×2 Balaam design, the following comparisons are usually assessed:

1. Comparisons by sequence
2. Comparisons by period
3. *T* vs *R* based on sequences 3 and 4—this is equivalent to a typical 2×2 crossover design
4. *T* vs *R* given *T* based on sequences 1 and 3
5. *R* vs *T* given *R* based on sequences 2 and 4

6. Comparison between (1) and (3) for assessment of the treatment-by-period interaction

It should be noted that the interpretations of the above comparisons are different. More information regarding statistical methods for data analysis of Balaam's design can be found in the work of Chow and Liu.^[7]

Two-Sequence Dual Design

Two-sequence dual design is a 2×3 higher order cross-over design consisting of two dual sequences, namely, *TRT* and *RTR*. Under the two-sequence dual design, qualified subjects will be randomly assigned to receive either the sequence of *TRT* or the sequence of *RTR*. Of course, there is a sufficient length of washout between dosing periods. Under the two-sequence dual design, we will be able to evaluate the relative risk of alternating between the use of the biological product and reference products and the risk of using the reference product without such alternating.

Note that expected values of the sequence-by-period means, analysis of variance table, and statistical methods (e.g., the assessment of average biosimilarity, the inference on carryover effect, and the assessment of intra-subject variabilities) for analysis of data collected from a two-sequence dual design are given in the work of Chow and Liu.^[7] In case there are missing data (i.e., incomplete data), statistical methods proposed by Chow and Shao^[8] are useful.

Modified Balaam's design

As indicated earlier, Balaam's design is useful for addressing switching, whereas two-sequence dual design is appropriate for addressing alternating. In the interest of addressing both switching and alternating in a single trial, we may combine the two study designs as follows: (*TT*, *RR*, *TRT*, *RTR*), which consists of a parallel design (the first two sequences) and a two-sequence dual design (the last two sequences). Data collected from the first two dosing periods (which are identical to the Balaam design) can be used to address switching, whereas data collected from sequences 3 and 4 can be used to assess the relative risks of alternating.

Complete Design

As it can be seen that the modified Balaam design is not a balanced design in terms of the number of dosing periods. In the interest of balance in dosing periods, it is suggested the modified Balaam design be further modified as (*TTT*, *RRR*, *TRT*, *RTR*). We will refer to this design as a complete design. The difference between the complete design and the modified Balaam design is that the treatments are repeated at the third dosing period for sequences 1 and 2. Data collected from sequence 1 will provide a more

accurate and reliable assessment of intra-subject variability, whereas data collected from sequence 2 is useful in establishing baseline for the reference product. Note that statistical methods for analysis of data collected from the complete design are similar to those under the modified Balaam design.

Alternative Designs

For assessment of IBE under a replicated design, Chow et al.^[9] indicated that the optimal design among 2×3 crossover designs is the so-called extra-reference design, which is given by (TRR, RTR). Thus, an alternative design is to combine a parallel design (TTT, RRR) and a 2×3 extra-reference design for addressing both switching and alternating. The resultant study design is then given by (TTT, RRR, RTR, TRR).

Adaptive Designs

In recent years, the use of adaptive design methods in clinical research has become very popular due to its flexibility and efficiency for identifying any (or optimal) clinical benefits of the test treatment under investigation (see, e.g., the work of Chow and Chang^[10]). Similar ideas can be applied for the assessment of biosimilarity and interchangeability of biosimilar products. For example, a two-stage adaptive design that combines two independent studies into a single trial may be useful. Some adaptations (modifications or changes) can be implemented at the end of the first stage after the review of accumulated data collected from the first stage. More information regarding various adaptive trial designs can be found in the work of Chow and Chang.^[10]

POWER ANALYSIS FOR SAMPLE SIZE

Sample size determination for the assessment of drug interchangeability is an important issue in bioequivalence assessment for generic approval regardless of the criteria used as described in Section “Design for Drug Interchangeability.” In what follows, sample size calculation under the criteria described previously are briefly discussed.

Testing for PBE/IBE

For assessment of PBE and IBE, the FDA recommends that a replicated crossover design (e.g., a 2×3 dual crossover design or a 2×4 replicated crossover design), which provides independent estimates of intersubject variability, intra-subject variability, and variability due to subject-by-treatment interaction, be used for achieving a desired power. As an example, for the assessment of PBE under a 2×4 replicated crossover design, when $n_1 = n_2 = n$, an approximate formula for determination of n (assuming that $n_1 = n_2 = n$) is given by

$$n \geq \frac{2\delta^2\sigma_{1,1}^2 + \sigma_{TT}^2 + c^2\sigma_{TR}^4 - 2c\rho^2\sigma_{BT}^2\sigma_{BR}^2}{\lambda^2} (z_{1-\alpha} + z_\beta)^2, \quad (6)$$

where $\delta = \mu_T - \mu_R$, $(\sigma_{WT}^2, \sigma_{WR}^2)$, and $(\sigma_{BT}^2, \sigma_{BR}^2)$ are within-subject variabilities for the test and reference products and between-subject variabilities for the test and reference products, respectively; $\sigma_{1,1}^2 = \sigma_{BT}^2 + \sigma_{BR}^2 - 2\rho\sigma_{BT}\sigma_{BR} + \sigma_{WT}^2 + \sigma_{WR}^2$; and

$$c = \begin{cases} 1 + \theta_{PBE} & \text{if } \sigma_{TR}^2 \geq \sigma_0^2 \\ 1 & \text{if } \sigma_{TR}^2 < \sigma_0^2 \end{cases}$$

for a set of given values of δ , $\sigma_{1,1}^2$, σ_{TT}^2 , σ_{TR}^2 , σ_{BT}^2 , σ_{BR}^2 , and ρ , where z_t is the t th quantile of the standard normal distribution, and $1 - \beta$ is the desired power. Similarly, under a replicated crossover design (e.g., a 2×3 dual design or a 2×4 replicated crossover design), sample size for the assessment of IBE can be determined (see, e.g., the work of Chow et al.^[9]).

Sample Size Determination

For average bioequivalence assessment under a standard 2×2 standard crossover design, Chow and Liu^[7] provided a formula for determination of the required sample size for achieving a $1 - \beta$ power at the α nominal level (assuming that $n_1 = n_2 = n$) as follows:

$$n \geq 2 \left[t(\alpha, 2n-2) + t(\beta, 2n-2) \right]^2 \left(\frac{\hat{\sigma}_d}{\Delta - \delta} \right)^2, \quad (7)$$

where $\delta = \mu_T - \mu_R$, Δ is the equivalence limit, and $\hat{\sigma}_d$ is given by

$$\hat{\sigma}_d^2 = \frac{1}{n_1 + n_2 - 2} \sum_{k=1}^2 \sum_{i=1}^{n_k} (d_{ik} - \bar{d}_k)^2,$$

where $d_{ik} = \frac{1}{2}(Y_{i2k} - Y_{i1k})$, $i = 1, \dots, n_k$, $k = 1, 2$, and $\bar{d}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} d_{ik}$, $k = 1, 2$, in which Y_{ijk} is the response of the i th subject

in the k th sequence at the j th period. For the multiplicative model, we consider the (0.8, 1.25) as the bioequivalence limit for $\delta = \mu_T/\mu_R$ and μ_T and μ_R are the median bioavailabilities of the test and reference formulations, respectively. Similarly, the sample size n required to achieve a $1 - \beta$ power at the α nominal level after the logarithmic transformation is determined by the following equations:

$$n \geq 2 \left[t(\alpha, 2n-2) + t(\beta, 2n-2) \right]^2 [CV/\log(1.25)]^2 \quad \text{if } \delta = 1,$$

$$n \geq 2 \left[t(\alpha, 2n-2) + t(\beta, 2n-2) \right]^2 [CV/(\ln(1.25) - \log\delta)]^2 \quad \text{if } 1 < \delta < 1.25,$$

and

$$n \geq 2 \left[t(\alpha, 2n-2) + t(\beta, 2n-2) \right]^2 \left[CV / (\ln(0.8) - \log \delta) \right]^2 \\ \text{if } 0.8 < \delta < 1$$

where $\ln$ denotes the natural logarithm and $CV = \sqrt{\exp(\sigma^2) - 1}$, the coefficient of variation in the multiplicative model, and σ^2 is the residual (within-subject) variance on the log scale. As a result, for sample size determination based on SCDI, the required sample size for achieving a $1 - \beta$ power at the α level of significance can be obtained by simply replacing Δ given in Eq. 7 with $\Delta' = \Delta f \sigma_w$.

SAMPLE SIZE BASED ON REVERSE OF T AND R

As discussed in Section “Sample Size Determination,” under a standard 2×2 crossover design, by formula (7) the sample size required for switching from T to R can be obtained, say n_1 . Using the same formula (7), the sample size required for switching from R to T can be similarly obtained, say n_2 . As a result, under the proposed p_3 criterion for drug interchangeability, the sample size required for achieving a $1 - \beta$ power at the α level of significance is then given by $n = \max(n_1, n_2)$.

CONCLUDING REMARKS

In practice, as indicated by regulatory agencies such as the FDA, approved generic drug products can be used as a substitute for the innovative drug product; it is a concern whether these approved generic drug products can be used interchangeably in terms of their quality, safety, and efficacy compared to the innovative drug product. It also recognized that bioequivalence assessment based on average bioavailability by ignoring the heterogeneity in variabilities between the test and reference products does not guarantee drug interchangeability. Although some criteria for drug interchangeability such as PBE and IBE^[1,2] have been proposed in the last decade, none of these criteria work satisfactorily for addressing drug prescribability and switchability. On the other hand, the SCDI criterion adjusted for the intra-subject variability and the variability due to subject-by-treatment interaction proposed by Chow

et al.^[5] seems reasonable. The newly proposed criterion by reversing the test and reference products may provide a simple answer to a complicated problem.

The proposed criteria for drug interchangeability for generic approval of small-molecule drug products can be directly applied to the assessment of interchangeability for biosimilar products. However, further evaluation of their statistical properties and/or performances are necessarily conducted.

REFERENCES

1. FDA. Guidance for industry: Statistical approaches to establishing bioequivalence. Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Rockville, MD, 2001.
2. FDA. Guidance on bioavailability and bioequivalence studies for orally administered drug products—general considerations. Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Rockville, MD, 2003.
3. Chen, M.L.; Patnaik, R.; Hauck, W.W.; Schuirmann, D.F.; Hyslop, T.; Williams, R. An individual bioequivalence criterion—Regulatory considerations. *Stat. Med.* **2000**, *19*, 2821–2842.
4. Endrenyi, L.; Taback, N.; Tothfalusi, L. Properties of the estimated variance component for subject-by-formulation interaction in studies of individual bioequivalence. *Stat. Med.* **2000**, *19*, 2867–2878.
5. Chow, S.C.; Xu, H.L.; Endrenyi, L.; Song, F.Y. A new scaled criterion for drug interchangeability. *Chin. J. Pharm. Anal.* **2015**, *35* (5), 844–848.
6. Chow, S.C.; Song, F.Y.; Chen, M. Some thoughts on drug interchangeability. *J. Biopharm. Stat.* **2016**, *26* (1), 178–186.
7. Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*, 3rd Ed.; Chapman and Hall/CRC Press, Taylor & Francis: New York, 2008.
8. Chow, S.C.; Shao, J. Statistical methods for two-sequence dual crossover designs with incomplete data. *Stat. Med.* **1997**, *16*, 1031–1039.
9. Chow, S.C.; Shao, J.; Wang, H. Individual bioequivalence testing under 2×3 designs. *Stat. Med.* **2002**, *21*, 629–648.
10. Chow, S.C.; Chang, M. *Adaptive Design Methods in Clinical Trials*, 2nd Ed.; Chapman and Hall/CRC Press, Taylor & Francis: New York, 2011.
11. Chow, S.C.; Shao, J.; Wang, H. Statistical tests for population bioequivalence. *Stat. Sin.* **2003**, *13*, 539–554.

Drug Interchangeability: Meta-Analysis for Safety Monitoring

Wen-Wei Liu and Shein-Chung Chow

Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

When an innovative (brand-name) drug is going off patent protection, pharmaceutical or generic companies may file an abbreviated new drug application for generic approval. As indicated by the U.S. Food and Drug Administration (FDA), an approved generic drug can be used as a substitute for the brand-name drug. FDA, however, does not indicate that approved generic drugs of the same brand-name drug can be used interchangeably. As more generic drugs become available in the marketplace, it is a concern whether the approved generic drugs are safe and can be used interchangeably. In this entry, we propose two safety margins as new bioequivalence limits for monitoring drug interchangeability based on a meta-analysis of data obtained from regulatory submissions which have been approved by FDA. In addition to the monitoring of drug interchangeability of generic drugs, the proposed margins can also be extended to address drug interchangeability of biosimilars.

Keywords: Drug interchangeability; Safety margins for safety monitoring; Switchability; Prescribability.

INTRODUCTION

When an innovative (brand-name) drug is going off patent protection, the innovative drug companies and/or generic drug companies may file an abbreviated new drug application (ANDA) for generic approval through the conduct of a bioequivalence study. Bioequivalence testing for generic approval is based on the Fundamental Bioequivalence Assumption that when two drug products have similar drug absorption profiles (or equivalent in average bioavailability), it is assumed that they will reach similar therapeutic effects or they are therapeutic equivalent. For an approved generic drug product, the United States Food and Drug Administration (FDA) indicates that it can be used as a substitute of the brand-name drug.

As more generic drug products are available in the marketplace, it is a concern whether the approved generic drugs work as well as the brand-name drug in terms of quality and safety. In practice, it is very common that patients may switch from one generic to another depending upon the availability and inexpensiveness of the generic drugs. Thus, it is a concern that such a switch from a generic drug (e.g., approved on the lower end of bioequivalence limit) to another (e.g., approved on the higher end of bioequivalence limit) could cause a drastic change in blood concentration and consequently cause a safety concern. This issue is critical not only for generic drugs of small molecular innovative drugs but also for biosimilars, which are (large molecular) similar biologic drug products.

In this entry, safety monitoring procedures based on meta-analysis of data available in approved regulatory submissions (e.g., ANDA for generic approval) is proposed. The proposed margins can be extended to monitoring the safety of drug interchangeability of biosimilar drug products.^[1] The entry is organized as follows. The section “Bioequivalence and Drug Interchangeability” provides outline of the bioequivalence and the most common decision rule and then describes why interchangeability might be the issue. The section “Meta-Analysis for Safety Monitoring” briefly reviews what general meta-analysis should consider and how to fit it to our simulation study mathematically. The proposed margins and simulated results are given in the section “Safety Margins for Monitoring Drug Interchangeability and Simulation Study.” There are some differences and similarities between generic drug products and biosimilars, and application of the results to biosimilars is discussed in section “Application to Biosimilars” and section “Conclusion” gives the concluding remarks.

BIOEQUIVALENCE AND DRUG INTERCHANGEABILITY

Bioequivalence

When the innovative drugs want to get the approval from FDA, the pharmaceutical companies have to file new drug applications (NDA). However, for the generic drugs which

are identical to the innovative (brand-name) drugs, they only need to file ANDA which lack the process of animal studies as well as clinical safety and efficacy trials compared to NDA.^[2] The U.S. Federal Food, Drug, and Cosmetic Act (FDCA), enacted by Congress in 1938, gives authority to FDA, and provisions for generic drugs were added in 1984 as *Drug Price Competition and Patent Term Restoration Act*. It states that rate and extent of drug absorption must be compared to establish bioequivalence between two products. Bioavailability of a drug can be described as the extent and rate of the absorption of active drug ingredient from the drug product to the site intended to treat, and it is usually measured by the area under the blood or plasma concentration-time curve (AUC) and the maximum concentration ($C_{\max}$), respectively. The FDA's regulations are codified in Title 21 of the Code of Federal Regulations (21 CFR) which clearly defines $C_{\max}$ as index of rate of absorption and AUC as index of extent of absorption. If the drug is not absorbed from bloodstream, bioavailability can also be assessed by measurements intended to reflect the extent and rate. A comparative bioavailability study indicates the comparison of bioavailabilities of different formulations of the same drug or different drug products, and they are claimed bioequivalent when assuming that they will provide the same therapeutic effect or that they are therapeutically equivalent.

"If two drug products are shown to be bioequivalent, it is assumed that they will generally reach the same therapeutic effect or they are therapeutically equivalent" stated by Chow and Liu^[3] is known as the Fundamental Bioequivalence Assumption, and bioequivalence assessment for generic drug approval should be conducted under this assumption. However, there are four possible scenarios in practice and only one indicates this assumption. The second scenario indicates that the two drug absorption profiles are not similar but they are therapeutically equivalent, and this is the case that generic drug companies might argue when their generic products fail to meet regulatory requirements. The third scenario indicates that the two drug absorption profiles are similar but they are actually not therapeutically equivalent, and innovative drug companies usually argue the inefficacy of the generic copies under this scenario. The fourth scenario is the worst one which indicates that the two drug absorption profiles are not similar and they are not therapeutically equivalent. Therefore, Fundamental Bioequivalence Assumption has been criticized that it is not based on scientific considerations.

80/125 Decision Rule

There are many decision rules for the evaluation of average bioequivalence. The ± 20 rule was applied to assess average bioequivalence in the early 1980s. During the same time, it was suggested that the 80/20 rule be used as the secondary analysis to support the ± 20 rule. However, since the two rules may result in different conclusions with some

possible practical issues, FDA recommended the 80/125 rule be used based on the log-transformed data.^[3]

FDA guidance^[4] suggested that the log-transformation on AUC and $C_{\max}$ be considered. Thus, if the 90% confidence interval (CI) for the geometric mean ratio (GMR) for average bioequivalence of the test formulation over the reference formulation is within (80.00%, 125.00%), then the bioequivalence can be concluded. The range for the criterion is $\ln(0.80) = -0.2231$ to $\ln(1.25) = 0.2231$, and it is symmetric about $\ln(1.00) = 0$, where the test formulation is totally equivalent to the reference formulation.^[3]

Drug Interchangeability

When the generic drugs get the approvals from FDA, it does not indicate that approved generic drugs and the innovative drugs can be used interchangeably. The FDA only indicates that approved generic drugs can be used as a substitute to the innovative drugs. Drug interchangeability for generic drugs can be classified as prescribability or switchability.^[5,6] However, we used average bioequivalence in this entry to simplify this problem since it is the usual method to assess the bioequivalence.

Drug prescribability is described as the physician's decision to prescribe the brand-name drug or its generic copies under the condition that they can be used interchangeably in terms of efficacy and safety. If brand-name drug and its generic copies are bioequivalent in both average and variability, it can be claimed as *population bioequivalence* (PBE) which is usually used to clarify drug prescribability.^[6] PBE is usually applied to new formulations, additional strengths, or new dosage forms in NDAs. Drug switchability is indicated as the exchangeability within the same subject. It can be referred as from brand-name drug to generic copy or from a generic drug to the other generic copies. Drug switchability is ensured by *individual bioequivalence* (IBE) which describes that the majority of subjects satisfy the criteria.^[5] IBE should be considered for ANDA or abbreviated antibiotic drug application for generic drugs. However, there has been a debate on the criteria recommended by FDA; so FDA suggests the method of small-sample CI approach proposed by Hyslop, Hsuan, and Holder be considered for the assessment of PBE/IBE for statistical analysis.^[7]

META-ANALYSIS FOR SAFETY MONITORING

In general, meta-analysis combines small studies for a systematic review. In practice, the sample sizes of some studies might be relatively small. In addition, some trials may fail to show statistically significant treatment effects. Thus, by combining these studies, meta-analysis can provide a more accurate and reliable assessment of the treatment effect under investigation.

In order to achieve the ultimate goals of better accuracy and reliability of the treatment effect, the following

general principles need to be taken into consideration when performing a meta-analysis.

1. It is often a concern that only positive clinical studies are included in the meta-analysis, which may have introduced selection bias. To avoid this, criteria for selection of studies and the time period to be included in the meta-analysis need to be clearly stated in the study protocol. If there is a potential trend over time, a statement need to be provided for scientific justification for inclusion of the studies selected.
2. It is critical to assess similarity and dissimilarity among studies to reduce variability for a more accurate and reliable assessment of the test treatment under investigation, because different studies may be conducted with similar but different study protocols, drug products and doses, patient populations, sample sizes, evaluability criteria, and so on.
3. Before the data sets obtained from different studies can be combined for a meta-analysis, it is suggested that a statistical test for probability be performed to determine whether there is significant treatment-by-study interaction.

Meta-Analysis

For each study trial, 90% CI for log-transformed data can be constructed. If it is within the range of 80% to 125%, then bioequivalence is claimed. However, generic approval criteria based on the average bioequivalence has the limitation that cannot guarantee the drug interchangeability. If any trial has lower bound L^* which is lower than L (lower margin) but higher than 80% or the upper bound U^* is higher than U (upper margin) but lower than 125%, then it might have the safety concern since the relative change of concentration of the drug in blood is dramatic. Assume the bioequivalence of the generic drug, a patient used for a long period was close to 125%, and a physician switches to another generic drug which with bioequivalence close to 80%. We may wonder if the new generic drug can still be used to heal the patient as efficient as the old one because the concentration of the drug in blood decreases $\frac{(125-80)}{125} = 36\%$ to the original. Correspondingly, the concentration of the drug in blood may also increase $\frac{(80-125)}{80} = 56\%$ to the original. By doing the meta-analysis in different clinical trials for the same brand-name drug, we will have several margins as the monitoring tool to evaluate all generic drugs and to calculate $P(L^* < L)$ and $P(U^* > U)$ as the probability of the improper rate.

From the previous study by Chow and Liu,^[8] a meta-analysis for a systemic overview of bioequivalence was proposed. Suppose there are K different trials for the same brand-name drug. Follow the assumptions below:

1. The same standard two-sequence, two-period crossover design was constructed with n_{k1} and n_{k2} at sequence 1 and 2, where $k = 1, \dots, K$.

2. The same statistical model for 2×2 crossover design was used for log-transformed data to assess bioequivalence.
3. Intersubject variabilities are the same for all subjects. ($\sigma_{sk}^2 = \sigma_s^2$ for all k .)
4. Intrasubject variabilities are the same for all subjects. ($\sigma_{ek}^2 = \sigma_e^2$ for all k .)

To combine the data from the K studies, we first test for the homogeneity of the reference products among the K studies. We can derive the following χ^2 distribution with $K - 1$ degrees of freedom to test homogeneity:

$$\chi_R^2 = \sum_{k=1}^K \frac{1}{C_k} (\bar{Y}_{Rk} - \hat{\mu}_R)^2 \quad (1)$$

where adjusted parameter,

$$C_k = \frac{1}{4} \left(\frac{1}{n_{k1}} + \frac{1}{n_{k2}} \right) \quad (2)$$

and unbiased estimator,

$$\hat{\mu}_R = \sum_{K=1}^K \left(\frac{\frac{1}{C_k(\sigma_e^2 + \sigma_s^2)} \bar{Y}_{Rk}}{\sum_{k=1}^K \frac{1}{C_k(\sigma_e^2 + \sigma_s^2)}} \right) = \sum_{K=1}^K \left(\frac{\frac{1}{C_k} \bar{Y}_{Rk}}{\sum_{k=1}^K \frac{1}{C_k}} \right) \quad (3)$$

and $\bar{Y}_{Rk}$ is the least squares mean of the reference product of the k^{th} study.

Reject the null hypothesis of homogeneity at the α level of significance if:

$$\chi_R^2 > \chi^2(\alpha, K - 1) \quad (4)$$

If we fail to reject the null hypothesis of homogeneity, we can combine drug effect and get the corresponding CIs. Let,

$$\hat{d}_k = \bar{Y}_{Tk} - \bar{Y}_{Rk}, k = 1, \dots, K \quad (5)$$

be the difference of the least squares means between the test product and the reference product of the k^{th} study. If we assume that μ_{Tk} , $k = 1, \dots, K$, are the same for all studies ($\mu_{Tk} = \mu_T$ for all k), a combined estimate for $\mu_T - \mu_R$ is the weighted means of $\hat{d}_k$, i.e.,

$$\bar{d} = \sum_{k=1}^K \left(\frac{\frac{1}{C_k} \bar{d}_k}{\sum_{k=1}^K \frac{1}{C_k}} \right) \quad (6)$$

where,

$$\text{Var}(\bar{d}) = \frac{\sigma_e^2}{\sum_{k=1}^K \frac{1}{C_k}} \quad (7)$$

Then a $(1 - 2\alpha) \times 100\%$ CI, denoted by (L_m, U_m) , can be obtained as follows:

$$(L_m, U_m) = \bar{d} \pm t\left(\alpha, \sum_{k=1}^K (n_{k1} - n_{k2} - 2)\right) \sqrt{\hat{V}(\bar{d})} \quad (8)$$

SAFETY MARGINS FOR MONITORING DRUG INTERCHANGEABILITY AND SIMULATION STUDY

The studies in meta-analysis are all approved generic drugs, so the 90% confidence intervals are all located within (80%, 125%). However, as described in section “Meta-analysis,” it might have some concerns while switching from one generic drug to another. Therefore, safety margin for interchangeability cannot use (80%, 125%), and if applying (L_m, U_m) , drug concentration in blood will not change so much, so it could be selected as a margin. However, it is too narrow and strict, for example, $(L_m, U_m) = (95\%, 105\%)$, so several margins to monitor the drug interchangeability are proposed. The hypothesis can be expressed as follows:

$$\begin{aligned} H_0: \frac{(|L - U|)}{L} \leq \tau \text{ and } \frac{(U - L)}{L} \leq \tau \text{ versus} \\ H_1: \frac{(|L - U|)}{L} > \tau \text{ or } \frac{(U - L)}{L} > \tau \end{aligned} \quad (9)$$

Safety Margins for Monitoring Drug Interchangeability

Margin1 (M1). Let (L_m, U_m) be the 90% CI obtained from a meta-analysis. Set $L_1 = L_m - \delta_L$ and $U_1 = U_m - \delta_M$, where $\delta_L = (L_m - 0.8)/2$ and $\delta_M = (1.25 - U_m)/2$. Actually, L_1 is in the middle of 0.8 and L_m , and U_1 is in the middle of 1.25 and U_m . (L_1, U_1) is our first set of margins.

Margin2 (M2). For each meta-analysis, we can get the lower margin (L) which is smaller than 95% of the lower bounds (L^*) from the 90% CI of different studies, and it can be expressed as $P(L^* \geq L_2) = 95\%$. Correspondingly, the upper margin (U) is larger than 95% of the upper bound (U^*) from the 90% CI of different studies, and it can be expressed as $P(U^* \leq U_2) = 95\%$. Therefore, L_2 will be like the 5th percentile of the boundaries, and U_2 will be the 95th percentile of the boundaries. Since we will simulate 20 data sets for a meta-analysis, L_2 is the second smallest lower bound and U_2 is the second largest upper bound.

Margin3 (M3) and **Margin4 (M4)** are similar to M2 but with different probability. For **Margin3**, it is $P(L^* \geq L_3) = 90\%$, $P(U^* \leq U_3) = 90\%$. For **M4**, it is $P(L^* \geq L_4) = 85\%$, $P(U^* \leq U_4) = 85\%$ (Fig. 1).

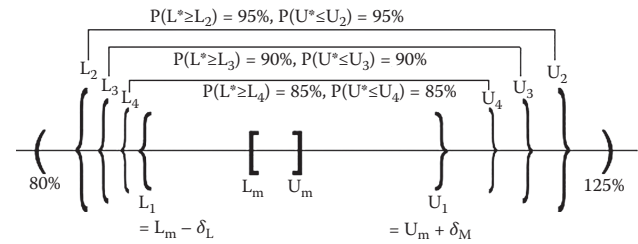


Fig. 1 Relationships among BE limits, meta-analysis confidence limits, M1, M2, M3, and M4

Simulation Study

All the studies in meta-analysis are approved generic drugs, so the data are regulated by FDA and we cannot get these real data. Therefore, the simulation study is conducted. In this study, we simulate 24 subjects in a data set and 20 data sets as one meta-analysis study, and we simulate 1000 meta-analysis studies from one scenario. There are six scenarios in this study (Table 1). For each scenario, there is the average meta-analysis 90% CI derived by the formula in section “Meta-Analysis” (Table 2). The margins (M1–M4) for each scenario (S1–S6) can also be obtained by the definition in section “Safety Margins for Monitoring Drug Interchangeability.” The probability of the 90% CI for a study below the lower margin (below L) or above the upper margin (above U) can be calculated. Safety impact is when the efficacy of original drug is close to the lower margin and suddenly change to the drug with efficacy close to the upper margin, which is calculated as $\frac{(|L - U|)}{L}$ (low to high), correspondingly, efficacy impact is calculated as $\frac{(U - L)}{U}$ (high to low).

The 90% CI obtained from meta-analysis is usually narrower than the CI from one study, so we adjust it by δ . From Table 2, we can see that the CIs are similar when the true variances (σ_s^2, σ_e^2) are the same, and they get larger when the true variances become larger. In addition, the probability of the CI for a study lower than the lower margin (below L) or higher than the upper margin (above U) are all 0.050 for M2, 0.100 for M3, and 0.150 for M4 since it is the definition for these margins. However, the safety/efficacy impact seems to be smaller when the true mean (μ) is larger. It is not surprising that when the true variance gets larger, the safety/efficacy impact also gets larger. In addition, the below L and above U have the same situation as the safety/efficacy impact, and they dramatically increase for M1 when variance gets larger. M1, M2, M3, and M4 are similar when variances are small. However, when variances get larger, they perform differently and

Table 1 Parameter specification for the simulation studies

Scenarios	S1	S2	S3	S4	S5	S6
μ	100	100	100	105	105	105
$\sigma_s = \sigma_e$	10	20	30	10	20	30

Table 2 Safety and efficacy margin for interchangeability

						Safety/efficacy impact	
		(Lm, Um)	(L, U)	Below L	Above U	Low to high	High to low
S1	M1	(0.991, 1.012)	(0.896, 1.131)	0.178	0.104	0.263	0.208
	M2		(0.882, 1.135)	0.050	0.050	0.287	0.223
	M3		(0.892, 1.121)	0.100	0.100	0.258	0.205
	M4		(0.900, 1.111)	0.150	0.150	0.235	0.190
S2	M1	(0.981, 1.024)	(0.891, 1.137)	0.649	0.586	0.277	0.217
	M2		(0.814, 1.228)	0.050	0.050	0.508	0.337
	M3		(0.821, 1.218)	0.100	0.100	0.484	0.326
	M4		(0.827, 1.208)	0.150	0.150	0.461	0.315
S3	M1	(0.973, 1.027)	(0.887, 1.139)	0.909	0.877	0.284	0.221
	M2		(0.805, 1.242)	0.050	0.050	0.543	0.352
	M3		(0.808, 1.238)	0.100	0.100	0.532	0.347
	M4		(0.811, 1.234)	0.150	0.150	0.522	0.343
S4	M1	(0.990, 1.010)	(0.895, 1.130)	0.143	0.082	0.263	0.208
	M2		(0.888, 1.126)	0.050	0.050	0.269	0.212
	M3		(0.897, 1.114)	0.100	0.100	0.242	0.194
	M4		(0.905, 1.105)	0.150	0.150	0.222	0.181
S5	M1	(0.983, 1.024)	(0.892, 1.137)	0.614	0.541	0.276	0.216
	M2		(0.816, 1.224)	0.050	0.050	0.500	0.333
	M3		(0.824, 1.213)	0.100	0.100	0.473	0.321
	M4		(0.831, 1.202)	0.150	0.150	0.446	0.308
S6	M1	(0.972, 1.025)	(0.886, 1.138)	0.889	0.860	0.284	0.221
	M2		(0.806, 1.241)	0.050	0.050	0.541	0.351
	M3		(0.809, 1.236)	0.100	0.100	0.529	0.346
	M4		(0.812, 1.232)	0.150	0.150	0.518	0.341

M1 is usually the narrowest one. From regulatory point of view, if we choose the criterion for safety/efficacy impact is 30%, that is, $\tau = 0.3$ in the hypothesis, then we can see when the variance is small, all margins can meet this criterion. However, the variance in real data is usually located between 20 and 30, and we can see that only M1 meets this criterion. Therefore, the proposed M1 is a useful tool for safety monitoring for interchangeability.

APPLICATION TO BIOSIMILARS

Biosimilars are the biological drug products that are made from living organisms by means of recombinant DNA or controlled gene-expression methods, which is also known as follow-on biologics; as indicated by Chow,^[9] there are fundamental differences between generic drugs and biosimilars (Table 3). As can be seen in Table 3, generic drug products have a well-defined structure and can be characterized, while biosimilars are difficult to characterize due to their heterogeneous and complicated structures with mixtures of various molecules. Thus, generic drug products are usually relatively stable. However, since

Table 3 Fundamental differences between generic drug products and biosimilars

Generic drug products	Biosimilars
Made by chemical synthesis	Made by living cells or organisms
Defined structure	Heterogeneous structure mixtures of related molecules
Easy to characterize	Difficult to characterize
Relatively stable	Variable and sensitive to environmental conditions, such as light and temperature
No issue of immunogenicity	Issue of immunogenicity
Usually taken orally	Usually injected
Often prescribed by a general practitioner	Usually prescribed by specialists

biosimilars are derived from living cells or organisms, they have the biological property which is variable and sensitive to environmental conditions such as light and temperature. A slight change or variation during the manufacturing process may result in different function, efficacy, and safety.

In addition, unwanted immune response is the worst result of biosimilars which may cause a loss of efficacy or cause some safety issues.

However, the basic concept of statistical method used to assess bioequivalence for generic drugs and biosimilarity for biosimilars are consistent (ex. the 80/125 decision rule), so we could apply the results in this entry to biosimilars. In Europe, generic versions of biological drug products are administrated by European Medicines Agency (EMA), and are defined as a “version of an already authorized biological medicinal product with demonstrated similarity in quality characteristics, biological activity, efficacy and safety, based on a comprehensive comparability exercise.”^[10] They have issued several guidelines describing overarching and specific biosimilar. In the United States, the approval for generic versions of biological drug products is based on *Biologics Price Competition and Innovation (BPCI) Act*, which was originally sponsored and introduced in June 26, 2007 by Senator Edward Kennedy (D-MA) and written into law on March 23, 2010. As indicated in the *BPCI Act*, a biosimilar product is defined as a product that is “highly similar” to an already-approved biological product clinically in terms of safety, purity, and potency, notwithstanding minor differences in clinically inactive components. Here, purity may be related to some important quality attributes at critical stages of the manufacturing process, and potency has something to do with the stability and efficacy of the biosimilar product.

The *BPCI Act* seems to suggest that a biosimilar product should be highly similar (sufficiently close) to the reference drug product in all spectrums of good drug characteristics such as identity, strength (potency), quality, purity, safety, and stability as described in the United States Pharmacopeia and National Formulary (USP/NF). In addition, FDA draft guidance^[11] suggests “stepwise approach” and the concept of “totality-of-the-evidence.” This indicates that the FDA is interested in demonstrating global similarity in all aspects related to safety, purity, and potency of the biosimilar products, but it is almost impossible to demonstrate that a biosimilar product is highly similar to the reference product in all aspects of good drug characteristics in a single study. Thus, to ensure that a biosimilar product is highly similar to the reference product in terms of these good drug characteristics, biosimilar studies for different aims may be required. For example, if safety and efficacy are the concern, then a clinical trial must be conducted to demonstrate that there are no clinically meaningful differences in terms of safety and efficacy between a biosimilar product and the innovator biological product. On the other hand, to ensure that important quality attributes are highly similar, critical stages of the manufacturing process, assay development/validation, process control/validation, and product specification of the reference product should be necessarily established through the conduct of relevant studies.

Similar to the Fundamental Bioequivalence Assumption, Chow et al.^[12] proposed the following Fundamental Biosimilarity Assumption for follow-on biologics—When a biosimilar product is claimed to be biosimilar to an innovator’s product based on some well-defined product characteristics, it is therapeutically equivalent, provided that the well-defined product characteristics are validated and are reliable predictors of safety and efficacy of the products. Unlike the Fundamental Bioequivalence Assumption that assumes that equivalence in the exposure measures implies therapeutical equivalence, Fundamental Biosimilarity Assumption has to verify that some validated product characteristics are indeed reliable predictors of safety and efficacy.

However, generic drugs and biosimilars are slightly different in drug interchangeability. As indicated in the subsection (k)(4) of the *Public Health Act subsection 351(k)(4)*, of the *BPCI Act*, the term interchangeable or interchangeability in reference to a biological product, means that the biological product may be biosimilar to the reference product and expected to produce the same clinical result in any given patient. In addition, the biological product may also be substituted for the reference product without the intervention of the health care provider who prescribed the reference product. However, unlike the interchangeability of small-molecule drug products, which could be described as prescribability and switchability, the FDA considers interchangeability includes the concepts of switching and alternating between an innovative biologic product (R) and its biosimilars (T). The concept of switching involves the switch from not only “R to T” or “T to R” but also “T to T” and “R to R.” As a result, in order to assess switching, biosimilarity for “R to T,” “T to R,” “T to T,” and “R to R” needs to be assessed based on some biosimilarity criteria under a valid study design.

On the other hand, the concept of alternating is referred to as either the switch from T to R and then switch back to T (i.e., “T to R to T”) or the switch from R to T and then switch back to R (i.e., “R to T to R”). Thus, the difference between “the switch from T to R” or “the switch from R to T” and “the switch from R to T” or “the switch from T to R” needs to be assessed for addressing the concept of alternating.

CONCLUSION

The concept of drug interchangeability for generic (small molecule) drug products and biosimilars (generic versions of large molecule biological drug products) is similar but different, and so the interpretation should be carefully handled. Based on the 80/125 rule, it is not usual but it is possible to cause a dramatic change in the concentration of a drug in blood when changing from one drug to another. This change may cause uncertainty in

the safety or efficacy. Based on the margins proposed in this entry, M1 has better ability to control safety and efficacy impact, and so it can be considered as a useful tool for safety monitoring for interchangeability. However, it has a higher probability of the CI for a study below the lower margin (below L), or above the upper margin (above U) indicating that if we set this margin as the new criteria rule rather than 80% to 125%, there will be so many bioequivalence/biosimilar studies that do not meet this criterion. Therefore, drug interchangeability for bioequivalence/biosimilar is a big challenge and we can do a future work to interpret this by using mathematical formula.

ACKNOWLEDGMENT

Parts of this entry originally appeared in the study by Liu and Chow.^[1]

REFERENCES

1. Liu, W.W.; Chow, S.C. Meta-analysis for safety monitoring of drug interchangeability. *J. Bioequiv. Bioavailab.* **2015**, *7*, (5), 239–243.
2. Chow, S.C. Bioavailability and bioequivalence in drug development. *Wiley Interd. Rev.* **2014**, *6*, (4), 304–312.
3. Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*, 3rd Ed.; Chapman Hall/CRC Press: New York, 2008.
4. FDA. *Guidance on Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Considerations*; FDA: Rockville, MD, 2003.
5. Chow, S.C.; Shao, J.; Wang, H. Individual bioequivalence testing under 2×3 designs. *Stat. Med.* **2002**, *21* (5), 629–648.
6. Chow, S.C.; Shao, J.; Wang, H. Statistical tests for population bioequivalence. *Stat. Sinica.* **2003**, *13* (2), 539–554.
7. Hyslop, T.; Hsuan, F.; Holder, D.J. A small sample confidence interval approach to assess individual bioequivalence. *Stat. Med.* **2000**, *19* (20), 2885–2897.
8. Chow, S.C.; Liu, J.P. Meta-analysis for bioequivalence review. *J. Biopharm. Stat.* **1997**, *7* (1), 97–111.
9. Chow, S.C. *Biosimilars: Design and Analysis of Follow-on Biologics*, 1st Ed.; Chapman Hall/CRC Press: New York, 2013.
10. EMA. *Guideline on Similar Biological Medicinal Products*; Canary Wharf: London, 2014.
11. FDA. *Biosimilars: Questions and Answers Regarding Implementation of the Biologics Price Competition and Innovation Act of 2009*; FDA: Silver Spring, MD, 2012.
12. Chow, S.C.; Lu, Q.; Tse, S.K.; Chi, E. Statistical methods for assessment of biosimilarity using biomarker data. *J. Biopharm. Stat.* **2010**, *20* (1), 90–105.

Drug Interchangeability: Generic Products, Prescribability and Switchability

Laszlo Endrenyi

Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Laszlo Tothfalusi

Department of Pharmacodynamics, Semmelweis University, Budapest, Hungary

Abstract

When the bioequivalence (BE) of a generic product to a reference drug is stated, their interchangeability is generally assumed. This implies that the drug products can be readily switched and substituted. However, it is important to distinguish between the conditions of prescribability and switchability when, respectively, a drug is provided to a patient either for the first time in any of its forms or administered by replacing another formulation of the same drug product. These two conditions are statistically characterized by the regulatory models of population and individual BE. These models embrace important principles.

Keywords: Bioequivalence; Interchangeability; Prescribability, Switchability; Population bioequivalence; Individual bioequivalence.

INTRODUCTION

When the patent of an originally developed drug expires, marketing authorization is often sought for alternative generic drug products. Drugs having small molecules can usually be synthesized chemically. In order to market the alternative, generic drugs and their bioequivalence (BE) with the brand name, the drug must be demonstrated and approved. The generics are then viewed to be as safe and effective as the original drugs. Consequently, they are assumed to be mutually interchangeable. This is expressed in the preface of the *Orange Book* of the Food and Drug Administration (FDA)^[1] as well as in a guideline of the European Medicines Agency (EMA).^[2] The view is confirmed in assurances by FDA commissioners.^[3,4] However, the implementation and application of interchangeability are subject to rules and regulations of the states in the United States (US) and of member nations in the European Union (EU).

Thus, there is a widely applied assumption that the BE of two drug products enables their interchangeability. While the assumption is very often reasonable, it has limitations. Therefore, it is important to keep in mind the distinction between the *prescribability* and *switchability* of drug products.^[5] Two corresponding scenarios can be envisaged.

Under the first scenario, a patient had no prior exposure to a drug in any of its forms, i.e., she/he has been naive to the drug. In such situations, any of the approved products are *prescribable* provided that their safety and efficacy are satisfactory and, therefore, they have been approved by regulators. However, after the patient has started to receive one of the products, it may not be substituted by another formulation.

Under the second scenario, the patient is already taking the drug but opts for a cheaper alternative for economic reasons. In this case, the alternative products must be therapeutically equivalent, so that the products will be *switchable* within individuals. The switch usually occurs from the reference product to the test product (denoted by $R \rightarrow T$).

Switching between marketed drug products can become more general, and they can be entirely *interchangeable*. Directions of switches are possible like the $T \rightarrow R$ and $T \rightarrow T'$ switch; here, T' is a second generic, test product. The products R , T , and T' must be interchangeable in any possible sense to achieve the same therapeutic effect after a switch.

Depending on the condition, two test models were proposed. To describe prescribability quantitatively the so-called *population BE* (PBE) model was developed. This required the comparison of the population distributions for the relevant pharmacokinetic parameters. A simple parallel group design study is considered adequate to establish PBE.

In contrast, interchangeability could be demonstrated by satisfying the requirements of a model for *individual BE* (IBE). The implementation of IBE calls for a more complex study design: IBE requires investigations with a replicated crossover design in which both drug products are measured (at least) twice within individuals.^[6]

The development of population and IBE reflects the approach of the US FDA. These two regulatory models will be described after an introduction to the procedure of average BE. A few remarks will conclude this entry.

AVERAGE BE

BE studies are required for granting marketing authorization for generic medicinal products. The exact requirements slightly differ between main jurisdictions such as the US, the EU, and Japan, but the general approach is similar.

The approval is based on the results of a BE study. The plasma concentrations of the generic (called Test product, T) and the originator (Reference, R) formulations are compared in a limited number (somewhere between 18 and 60 depending upon the associated variability) healthy volunteers in a crossover fashion. Half of the volunteers are randomly assigned to get the Test drug and, after it is eliminated from the body, they receive the Reference product. For the other half of the group the order is reversed, receiving first the Reference and then the Test product. After drug ingestion, several blood samples are obtained and the concentrations of the active substance are determined.

To demonstrate BE, it should be shown that derived parameters from the concentrations, such as the areas under the curve and the maximum concentration of the Test and Reference products, are not really different, but they are similar to each other.

BE is usually evaluated by the two one-sided tests (TOST) procedure. The logarithmic means of the two formulations (μ_T and μ_R , respectively) are compared. The two null hypotheses assume that the deviation between the logarithmic means is either too low:

$$H_{01} : \mu_T - \mu_R \leq -\theta \quad (1)$$

or too high:

$$H_{02} : \mu_T - \mu_R \geq \theta \quad (2)$$

θ is a BE limit which is usually thought to be symmetrical in the two directions. BE is demonstrated if the alternative hypotheses are satisfied:

$$H_{a1} : \mu_T - \mu_R > -\theta \quad (3)$$

and

$$H_{a2} : \mu_T - \mu_R < \theta \quad (4)$$

Another, operationally equivalent procedure to the TOST approach expects that the confidence interval around the difference between the logarithmic means should be between the upper and lower confidence limits:^[7,8]

$$\theta \leq \mu_T - \mu_R \leq \theta \quad (5)$$

The 90% confidence limits are usually between $-\ln(1.25)$ and $+\ln(1.25)$. Consequently, the type I error for both null hypotheses is not higher than 5%.

PBE FOR PRESCRIBABILITY

Schall and Luus^[9] introduced a statistical model for PBE. It was implemented in a statistical guidance of FDA.^[6] It suggests that the criterion for PBE includes not only the (squared) difference between the two means, $\delta^2 = (\mu_T - \mu_R)^2$, but also the difference between the corresponding total variances ($\sigma_{TT}^2 - \sigma_{TR}^2$); the total variances are the sums of the between- and within-subject variances:

$$\theta_p = (\delta^2 + \sigma_{TT}^2 - \sigma_{TR}^2) / \max\{\sigma_{T0}^2, \sigma_{TR}^2\} \quad (6)$$

Here, σ_{T0}^2 is a constant, the magnitude of which is established by regulators. PBE can be claimed if the one-sided 95% upper confidence bound for θ_p is less than a prespecified BE limit.

The model emphasizes that not only the difference between the two means is important but also the difference between the total variances. According to the denominator, θ_p remains at a constant level as long as σ_{TR}^2 does not exceed σ_{T0}^2 . However, θ_p increases at high total variation of the reference formulation.

In view of the above PBE criterion, PBE can be claimed if the null hypothesis in:

$$H_0 : \lambda \geq 0 \text{ vs. } H_a : \lambda < 0 \quad (7)$$

is rejected at the 5% level of significance and the observed geometric means ratio (GMR) is within the limits of 80% and 125%, where:

$$\lambda = \delta^2 + \sigma_{TT}^2 - \sigma_{TR}^2 - \theta_{PBE} \max(\sigma_{TR}^2, \sigma_0^2) \quad (8)$$

and θ_{PBE} is a constant specified in the 2001 FDA draft guidance.

Under a 2×2 crossover design, the one-sided 95% upper confidence bound for θ_p can be obtained under the following model:

$$y_{ijk} = \mu + F_i + P_j + Q_k + S_{ikl} + \epsilon_{ijk} \quad (9)$$

where μ is the overall mean, P_j is the fixed effect of the j^{th} period, Q_k is the fixed effect of the k^{th} sequence, F_i is the fixed effect of the i^{th} drug product, S_{ikl} is the random

effect of the i^{th} subject in the k^{th} sequence under the l^{th} drug product, and ϵ_{ijk} s are independent random errors distributed as $N(0, \sigma_{w1}^2)$. It is assumed that S_{ikl} s and ϵ_{ijk} s are mutually independent. It can be verified that (S_{ikT}, S_{ikR}) , $i = 1, 2, \dots, n_k$; $k = 1, 2$ are independent and identically distributed bivariate normal random vectors with mean 0 and an unknown covariance matrix:

$$\begin{pmatrix} \sigma_{BT}^2 & \rho\sigma_{BT}\sigma_{BR} \\ \rho\sigma_{BT}\sigma_{BR} & \sigma_{BR}^2 \end{pmatrix} \quad (10)$$

where σ_{Bl}^2 denotes the between-subject variability for the l^{th} drug product. Thus, we have:

$$\sigma_{TT}^2 = \sigma_{BT}^2 + \sigma_{WT}^2 \text{ and } \sigma_{TR}^2 = \sigma_{BR}^2 + \sigma_{WR}^2 \quad (11)$$

The means and total variations of the two drug products can be estimated from BE studies with two (or more) parallel groups.

IBE FOR SWITCHABILITY

Schall and Luus^[9] and Sheiner^[10] presented similar statistical models for the determination of IBE. The model was implemented in the statistical guidance of FDA.^[6] The model for IBE contrasts not only the squared deviation between the two means, $\delta^2 = (\mu_T - \mu_R)^2$, but also the difference between the corresponding within-subject variances ($\sigma_{WT}^2 - \sigma_{WR}^2$). It contains also an additional term for the variance component of the subject-by-formulation interaction, σ_D^2 :

$$\theta_1 = (\delta^2 + \sigma_{WT}^2 - \sigma_{WR}^2 + \sigma_D^2) / \max\{\sigma_{w0}^2, \sigma_{WR}^2\} \quad (12)$$

Here, σ_{w0}^2 is a constant, the magnitude of which is set by regulators. IBE can be claimed if the one-sided 95% upper confidence bound for θ_1 is less than a prespecified BE limit.

The IBE model compares the differences between both means and within-subject variances of the two drug products. Furthermore, the interaction between subjects and formulations is considered to be important. However, the procedure proposed for its estimation was questioned.^[11]

The denominator of the IBE model indicates that the magnitude of the criterion θ_1 is constant when σ_{WR}^2 does not exceed σ_{w0}^2 . However, θ_1 increases at high within-subject variations of the reference product.

FDA suggested at the time a value of $\sigma_{w0} = 0.20$.^[6] It recommended in a simplified model of IBE, the so-called reference scaled average BE, $\sigma_{w0} = 0.25$.^[12] EMA suggested $\sigma_{w0} = 0.294$ for a closely related model.^[2]

In view of the above IBE criterion, IBE can be claimed if the null hypothesis in

$$H_0 : \gamma \geq 0 \text{ vs. } H_a : \gamma < 0 \quad (13)$$

is rejected at the 5% level of significance and the observed GMR is within the limits of 80% and 125%, where

$$\gamma = \delta^2 + \sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2 - \theta_{IBE} \max(\sigma_{WR}^2, \sigma_{w0}^2) \quad (14)$$

and θ_{IBE} is a constant specified in the 2001 FDA draft guidance.

For the assessment of IBE, FDA recommends that a replicated 2×2 crossover design, i.e., (TRTR, RTTR) or (RTRT, TRTR) be used. Under the 2×2 replicated crossover design, the one-sided 95% upper confidence bound for θ_1 can be obtained under the following statistical model:

$$y_{ijk} = \mu + F_l + W_{ljk} + S_{ikl} + \epsilon_{ijk} \quad (15)$$

where μ is the overall mean, F_l is the fixed effect of the l^{th} drug product, W_{ljk} s are fixed period, sequence, and interaction effects, S_{ikl} is the random effect of the i^{th} subject in the k^{th} sequence under the l^{th} drug product, and ϵ_{ijk} s are independent random errors distributed as $N(0, \sigma_{w1}^2)$. It is assumed that S_{ikl} s and ϵ_{ijk} s are mutually independent. Under model (9), σ_D^2 is given by:

$$\sigma_D^2 = \sigma_{BT}^2 + \sigma_{BR}^2 - 2\rho\sigma_{BT}\sigma_{BR} \quad (16)$$

which is the variance of $S_{ikT} - S_{ikR}$. Note that σ_D^2 is usually referred to as the variance component due to the subject-by-formulation interaction. It can be verified that when $\sigma_{WR}^2 \geq \sigma_{w0}^2$, the linearized criterion γ can be decomposed as follows:

$$\gamma = \delta^2 + \sigma_{0.5,0.5}^2 + 0.5\sigma_{WT}^2 - (1.5 + \theta_{IBE})\sigma_{WR}^2 \quad (17)$$

Individual BE and interchangeability for small-molecule drug products were widely discussed for about a decade starting in the 1990s and eloquent recommendations were made.^[13,14] However, the need for more complicated studies, three- or even more than three-period crossover studies, was questioned.^[15,16]

CONCLUSION

The important concepts of prescribability and switchability should be kept in mind when pharmacokinetic features of two drug products are compared. Regulatory approval means directly an agreement only that they are prescribable, i.e., that the drug products may be provided to a naive patient who has not taken the drug until this point in any form.

For small-molecule generics, it is generally assumed that regulatory approval of BE with a reference drug indicates (or, better, implies) that the two drug products are therapeutically equivalent. Consequently, their free interchangeability is generally also assumed. This also enables

ready substitution of the various products in pharmacies. Since small-molecule drugs are generally safe and effective, these assumptions and practices usually work. However, there are important exceptions.

First, if several generics are bioequivalent to the same drug, then it is not obvious that they are bioequivalent also with each other. Anderson and Hauck^[17] demonstrated that with two or three generics, the probability of such lack of BE is fairly low. However, with five, six, or more products in the market, the concern becomes substantial. Also, strong clinical concerns were repeatedly raised about the safe generic-to-generic substitution of, e.g., antiepileptics^[18,19] and immunosuppressant products.^[20]

Granting the status of “interchangeability” between two drugs will lead to the practice of switching and substitution. Before substitutions of small-molecule products can be considered, regulators must declare that sufficient similarity of the safety and efficacy of the drug products has been demonstrated. With small-molecule, chemical drugs, *sufficient similarity* is usually based on a clearly defined, rigorous statement of the BE between the pharmacokinetic profiles of the contrasted drug products. The comparison typically involves the generic and the reference formulation. Thus, clear, simple regulatory requirements and criteria have been declared for establishing the BE between a small-molecule, generic formulation and a reference product as set out in regulatory guidances.^[2,6,21,22]

The 2001 draft guideline released by the FDA^[6] met serious criticism and was later superseded by another guideline^[23] in which the FDA reverted back to average BE. However, components of the IBE model, including the comparison of within-subject variances and the estimation of the subject-by-formulation variance component, have emerged in some of the FDA guidances.^[24,25] The concept of IBE was acknowledged but was never accepted by the Committee for Medicinal Products for Human Use (CHMP), the European regulatory body.

REFERENCES

- Food and Drug Administration. Approved drug products with therapeutic equivalence evaluations (Orange Book), 2016. Available at <http://www.fda.gov/Drugs/DevelopmentApprovalProcess/ucm079068.htm> (accessed September 2, 2016).
- European Medicines Agency. Guideline in the investigation of bioequivalence London. January 20, 2010. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/01/WC500070039.pdf (accessed September 2, 2016).
- Nightingale, S. Therapeutic equivalence of generic drugs: Letter to health practitioners, 2001. Available at <http://www.fda.gov/cder/news/nightentlett.htm> (accessed September 2, 2016).
- Hamburg, M.A. Information regarding anti-epileptic drugs: U.S. Food and Drug Administration in response to requests in Senate Report No. 111-39 and House Agricultural Committee Report No. 111-279, 2010. Available at www.hpm.com/pdf/FDA%20Anti%20Epileptic%20Drug%20Report%20to%20Congress%20-%202011.pdf (accessed September 2, 2016).
- Anderson, S.; Hauck, W.W. Consideration of individual bioequivalence. *J. Pharmacokin. Biopharm* **1990**, *18* (3), 259–273.
- Food and Drug Administration. *Guidance on Statistical Approaches to Establishing Bioequivalence*. Center for Drug Evaluation and Research; U.S. Food and Drug Administration (CDER): Rockville, MD, 2001.
- Schuurmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokin. Biopharm.* **1987**, *15* (6), 657–680.
- Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*, 3rd Ed.; Chapman Hall/CRC Press, Taylor & Francis: New York, 2008.
- Schall, R.; Luus, H. On population and individual bioequivalence. *Stat. Med.* **1993**, *12*, 1109–1124.
- Sheiner, L.B. Bioequivalence revisited. *Stat. Med.* **1992**, *30*, 1777–1788.
- Endrenyi, L.; Taback, N.; Tothfalusi, L. Properties of the estimated variance component for subject-by-formulation interaction in studies of individual bioequivalence. *Stat. Med.* **2000**, *19* (20), 2867–2878.
- Haidar, S.H.; Davit, B.; Chen, M.L.; Conner, D.; Lee, L.; Li, Q.H.; Lionberger, R.; Makhlof, F.; Patel, D.; Schuurmann, D.J.; Yu, L.X. Bioequivalence approaches for highly variable drugs and drug products. *Pharm. Res.* **2008**, *25*, 237–241.
- Schall, R.; Williams, R.L. Towards a practical strategy for assessing individual bioequivalence. Food and Drug Administration Individual Bioequivalence Working Group. *J. Pharmacokin. Biopharm.* **1996**, *24* (1), 133–149.
- Patnaik, R.N.; Lesko, L.J.; Chen, M.L.; Williams, R.L. Individual bioequivalence: new concepts in the statistical assessment of bioequivalence metrics. *Clin. Pharmacokin.* **1997**, *33*, 1–6.
- Chow, S.C. Individual bioequivalence - a review of FDA draft guidance. *Drug Inf. J.* **1999**, *33*, 435–444.
- Endrenyi, L.; Amidon, G.L.; Midha, K.K.; Skelly, J.P. Individual bioequivalence: attractive in principle, difficult in practice. *Pharm. Res.* **1998**, *15*, 1321–1325.
- Anderson, S.; Hauck, W.W. The transitivity of bioequivalence testing: potential for drift. *Int. J. Clin. Pharmacol. Ther.* **1996**, *34*, (9), 369–374.
- Privitera, M. Generic antiepileptic drugs: current controversies and future directions. *Epilepsy Curr.* **2008**, *8*, 113–117.
- Bialer, M.; Midha, K.K. Generic products of antiepileptic drugs (AEDs): A perspective on bioequivalence and interchangeability. *Epilepsia* **2010**, *51*, 941–950.
- van Gelder, T.; Gabardi, S. Methods, strengths, weaknesses, and limitations of bioequivalence tests with special regard to immunosuppressive drugs. *Transplant. Int.* **2013**, *26* (8), 771–777.
- Food and Drug Administration. *Draft Guidance for Industry: Bioequivalence Studies with Pharmacokinetic Endpoints for Drugs Submitted Under an ANDA*; Center

- for Drug Evaluation and Research (CDER): Silver Spring, MD, 2013.
22. Health Canada. *Guidance Document: Conduct and analysis of comparative bioavailability studies*; Ottawa, Ontario, 2012. Available at http://hc-sc.gc.ca/dhp/mps/alt_formats/pdf/prodpharma/applicdemande/guide-ld/bio/gd_cbs_ebc_ld-eng.pdf (accessed September 2, 2016).
 23. Food and Drug Administration. *Guidance on Bioavailability and Bioequivalence Studies for Orally Administered Drug Products – General Consideration*; Center for Drug Evaluation and Research (CDER): Silver Spring, MD, 2013.
 24. Food and Drug Administration. *Draft Guidance on Warfarin Sodium*; Center for Drug Evaluation and Research (CDER): Silver Spring, MD, 2012.
 25. Food and Drug Administration. *Draft guidance on methylphenidate*; Revised November 2014; recommended September 2012. Available at <http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm320007.pdf> (accessed September 2, 2016).

Ecologic Inference

Sander Greenland

Department of Epidemiology and Department of Statistics, University of California,
Los Angeles, California, U.S.A.

Abstract

This entry emphasizes the nonseparability of contextual (group-level) effects from individual-level effects that arises from the multi-level causal structure of ecological (aggregate) data.

Keywords: Aggregate studies; Contextual effects; Cross-level bias; Ecological bias; Ecological fallacy; Ecological studies.

INTRODUCTION

Ecological data (e.g., health-surveillance data such as vital statistics and registry data) are often collected at great expense and therefore should be exploited fully. They are useful for detecting trends in behaviors, health, and disease; for initial detection and screening of associations; and for describing outcomes across a broader spectrum of exposures than is found in typical cohort or case-control studies. This breadth suggests that the power to detect effects in ecological data would be greater than in cohort or case-control studies, if biases could be controlled. However, ecological data on risk factors and outcomes are usually unlinked on the individual level and thus allow only group-level (ecological or aggregate) associations to be examined.

The limitations of inferences from such data are illustrated in the often-cited characteristic of Florida's population, where people are "born Hispanic and die Jewish." These average characteristics of the state's population tell little about the reality of individuals living in Florida because the birth and death rates are group averages and are not even from the same individuals. Linked birth and death data would show little if any change in ethnic identity over a lifetime. Unlinked data only allow ecological associations to be examined; attempts to draw individual-level inferences or causal inferences (whether individual or group level) may suffer from problems subsumed under the topics of *ecological bias* and *ecological fallacies*. Here I review the theory needed to understand these problems. I emphasize the nonseparability of contextual (group-level) effects from individual-level effects that arises from the multi-level causal structure of ecological (aggregate) data. Contrary to common misperceptions, this problem afflicts ecological studies in which the sole objective is to estimate contextual effects, as well as studies in which the objective is to estimate individual effects. Multi-level

effects also severely complicate causal interpretations of model coefficients.

OVERVIEW

The terms *ecological inference* and *ecological analysis* refer to any analysis based on data that are limited to characteristics of groups (aggregates) of individuals.^[1] Typically, these aggregates are geographically defined populations, such as counties, provinces, or countries. An enormous body of literature on ecological inference extends at least as far back as the early twentieth century. See, for example, Iversen^[2] and Achen and Shively^[3] for reviews of the social science literature, and Morgenstern^[4] for a review of the health sciences literature.

The objectives of ecological analyses may be roughly classified into two broad categories: *individual objectives*, which examine relations of group and individual characteristics to the outcomes of the individuals in the groups, and *purely contextual objectives*, to examine relations of group characteristics to group outcomes. A study may have both objectives. Most authors have focused on studies with only individual-level objectives and have emphasized the obstacles to *cross-level inference* from group-level (ecological) observations to individual-level relationships. This focus has led to a common misperception that individual-level relationships can be safely ignored if contextual objectives are the only interest. Several authors have also criticized the common practice of using strong (and sometimes implausible) modeling assumptions to generate cross-level inferences.^[2-5]

An overview of the literature surrounding these problems and the controversy they generate is given in "Ecological vs. Individual-Level Sources of Bias." The section "Multi-level Model Theory for Ecological Inference" provides some basic multi-level (hierarchical) structural-model

theory to define the problems (“structural” here means the underlying regression structure, apart from issues of random error or sampling variation). Further statistical theory, including fitting methods, can be found in Navidi et al.,^[6] Guthrie and Sheppard,^[7] Wakefield and Salway,^[8] Sheppard,^[9] and Wakefield.^[10] Multi-level studies that combine data from individual and ecological (aggregate) levels are discussed in “Multi-level Studies.” The section “Other Special Problems of Ecological Analysis” briefly describes other special problems of inference from ecological data and discusses some difficulties in causally interpreting the coefficients of contextual variables in multi-level models.

ECOLOGICAL VERSUS INDIVIDUAL-LEVEL SOURCES OF BIAS

Introduction

Studies limited to characteristics of aggregates are usually termed *ecological studies*.^[1,4,11–13] This term is perhaps unfortunate because the word *ecological* suggests that such studies are especially appropriate for examining the impact of environmental factors, including societal characteristics, and invites confusion between an ecological perspective (addressing relations at the environmental or social level) and an ecological study. As many authors have pointed out (Firebaugh,^[14] Von Korff et al.,^[15] Navidi et al.,^[6] Susser and Susser,^[16] Duncan et al.,^[17,18] Greenland^[5,19]), overcoming this confusion requires adopting a *multi-level* perspective, which allows theory and observations to be integrated on all levels: physiologic (exposures and responses of systems within individuals), individual (exposures and responses of individuals), and aggregate or contextual (exposures and responses of aggregates or clusters of individuals, such as locales or societies).

Defenders of ecological studies argue (correctly) that many critics have presumed individual-level relationships are the ultimate target of inference of all ecological studies, when this is not always so,^[16,20,21] and that contagious outcomes necessitate group-level considerations in modeling, regardless of the target level.^[22] They also point out that an ecological summary may have its own direct effects on individual risk, beyond that conferred by the contributing individual values. For example, the average economic status of an area can have effects on an individual over and above the effects of the individual's economic status.^[23,24] Unfortunately, some defenders make implicit assumptions to “prove” that entire classes of ecological studies are valid, or at least no less valid than individual-level analyses; see Refs.^[25–27] for critical commentaries against such arguments in the health sciences. Some ecological researchers are well aware of these problems and so explain their assumptions,^[28,29] but they still draw criticism because inferences are sensitive to those assumptions.^[30–32]

In the present article, I review some controversial assumptions in ecological analyses and then briefly discuss multi-level methods that represent individual-level and ecological data in a single model. Simple illustrations are used to make the points transparent to nonmathematical readers, problems of confounding and specification bias are highlighted, and an overview of the underlying mathematical theory is provided. Researchers have raised many other issues in the ongoing ecological study controversy; see the references for details, especially those in the “Other Special Problems of Ecological Analysis” section.

AGGREGATE-LEVEL CONFOUNDING DUE TO INDIVIDUAL-LEVEL EFFECTS

Two major types of measurements are used to describe aggregated data: summaries of distributions of individuals within the aggregates, such as mean age or percent female, and purely ecological (contextual) variables that are defined directly on the aggregate, such as whether there is a needle-exchange program in an area. The causal effects of the purely contextual variables are the focus of much social and epidemiological research.^[2,16,17,20,33] Nonetheless, most outcome variables of public health importance are summaries of individual-level distributions, such as prevalence, incidence, mortality, and life expectancy, all of which can be expressed in terms of average individual outcomes.^[34] Furthermore, many contextual variables are surrogates that summarize individual data; for example, neighborhood social class is often measured by average income and average education.

The use of summary measures in an ecological analysis introduces a major source of uncertainty in ecological inference. Summary measures depend on the joint individual-level distributions within aggregates, but distributional summaries do not fully determine (and sometimes do not seriously constrain) those joint distributions. This problem corresponds to the concept of “information lost to aggregation,” and is a key source of controversy about ecological studies.^[3]

The confounding of contextual and individual effects is illustrated in [Table 1](#). For simplicity, just two areas are used, but examples with many areas are available.^[25] Suppose we wish to assess a contextual effect, such as the impact of an ecological difference between areas A and B (such as a difference in laws or social programs) on the rate of a health outcome. We measure this effect by the amount RR_A that this difference multiplies the rate (the true effect of being in A vs. being in B). One potential risk factor, X , differs in distribution between the areas; an effect of X (measured by the rate ratio RR_X comparing $X = 1$ to $X = 0$ within areas) may be present, but we observe no difference in rates between the areas.

Do the ecological data in panel 1a of [Table 1](#) show no contextual effect? That is, do they correspond best with

Table 1 An Example Demonstrating the Confounding of Both Contextual and Individual Effects in Ecological Data

	Group A			Group B		
	X = 1	X = 0	Total	X = 1	X = 0	Total
<i>a. Ecological (marginal) data</i>						
Y = 1	?	?	560	?	?	560
N	60	40	100	40	60	100
Rate	?	?	5.6	?	?	5.6
<i>b. Possibility 1 ($RR_X = 2$, $RR_A = 7/8$)</i>						
Y = 1	420	140	560	320	240	560
N	60	40	100	40	60	100
Rate	7.0	3.5	5.6	8.0	4.0	5.6
<i>c. Possibility 2 ($RR_X = 1/2$, $RR_A = 8/7$)</i>						
Y = 1	240	320	560	140	420	560
N	60	40	100	40	60	100
Rate	4.0	8.0	5.6	3.5	7.0	5.6

The ecological data cannot identify the effect of group (A vs. B) on the rate of the outcome $Y = 1$ when only a marginal summary of the individual-level covariate X is available. Panels b and c show how within-group differences in outcome rates can affect inferences based on the ecological data in panel a. N = denominator (in thousands of person-years); RR_A and RR_X are the rate ratios for the true effects of A vs. B and of $X = 1$ vs. $X = 0$, respectively.

$RR_A = 1$? Unfortunately, the ecological (marginal) data on X distributions and the outcome rates are mathematically equally compatible with a broad spectrum of possibilities, two of which are given in panels 1b and 1c of Table 1. In the first possibility, area A has benefited from the contextual difference ($RR_A < 1$), but this fact has been obscured by area A’s higher prevalence of X , which is harmful ($RR_X > 1$). In the second possibility, area A has been harmed by the contextual difference ($RR_A > 1$), but this fact has been obscured by area A’s higher prevalence of X , which is beneficial ($RR_X < 1$). The correct answers in either possibility could be obtained by comparing the area rates after they have been standardized directly to a common X distribution; however, such standardization would require knowing the X -specific rates *within* the areas, which are not available in the example. Furthermore, an ecological regression could not solve the problem because it would only regress the crude area rates on the proportion with $X = 1$ in each area: Because both the crude area rates are 5.6, the ecological X -coefficient would be zero, hence the regression would produce no X -adjustment of the area rates.

Lacking within-area data on the joint distribution of X and the outcome, an ecological analysis must necessarily rely on external (prior) information to provide inferences about the contextual effect, although the margins may impose some bounds on the possibilities.^[3,28,35] If the X -specific rates in each area were assumed to be proportional to those in some external reference population with known X -specific rates, those external rates could be used to construct and compare standardized morbidity ratios for the areas (“indirect adjustment”). Unfortunately, such external rates are rarely available for all important covariates, and so other external (prior) information must be used to estimate the effect.

The results of such an analysis can be disappointing if the prior information is ambiguous. If X indicates, say, regular cigarette use, and the outcome is total mortality, we might be confident that RR_X is well above 1 and, hence, that the contextual effect (i.e., the A-B rate ratio) is protective. However, if X indicates regular alcohol consumption, we might feel justified in ruling out scenarios involving values for the rate ratio RR_X that are very far from 1. Nonetheless, because alcohol may be protective at moderate levels and causative at higher levels, we could not be sure of the direction of RR_X ; that would depend on the relative proportion of moderate and heavy drinkers in the areas. As a consequence, we could not be sure of the direction (let alone degree) of confounding in the ecological estimate of the contextual effect (i.e., the ecological A-B rate ratio of $5.6/5.6 = 1$).

The problem of cross-level confounding just illustrated has been discussed extensively since the early twentieth century (see Chapter 1 of Achen and Shively^[3]) and is a mathematically trivial consequence of the fact that marginals do not determine joint distributions. Yet this logical problem, which differentiates ecological from individual-level studies, continues to be misunderstood or ignored by many ecological researchers, so much that Achen and Shively^[3] remarked:

A cynic might conclude that social scientists tend to ignore logical problems and contradictions in their methods if they do not see anything to be done about them.

Their remark applies to the health sciences as well, as illustrated by this quote:

In practice, it may be that confounding usually poses a more intractable problem for ecological than for

individual-level studies. But this is due to the greater reliance on secondary data and proxy measures in ecological studies, *not to any problem inherent in ecological studies*.^[20][emphasis added]

Although ecological studies do suffer from greater reliance on secondary data and proxy measures, this passage is typical of defenses that overlook the aspects of confounding that *are* inherent in ecological studies. Other examples of such defenses are Cohen,^[36–38] Susser,^[39] and Pearce.^[21] A point illustrated in Table 1 is that, given confounding by a measured risk factor X , the individual-level data allow one to control the confounding in the most straightforward way possible: stratify on X . In contrast, confounding by X cannot be controlled by using only the ecological data, despite the fact that the effect under study is contextual (the effect of the social differences between areas A and B on outcome rates). Because contextual and individual effects are completely confounded in ecological data,^[5,14,19] the only solutions to this problem are to obtain individual data within the ecological units or to use assumptions that are untestable with the ecological data.^[3,25,26]

Another fallacy in some defenses of ecological studies is the claim that individual-level information is not needed if one is interested only in contextual (ecological) effects. Examples such as that in Table 1 show that such “holistic” arguments are incorrect, especially in health studies in which the outcome measure is an individual-level summary, because individual-level factors can confound the ecological results even if the study factor is contextual.^[5,14,19] Holistic arguments also tend to overlook the fact that ecological data usually refer to arbitrary administrative groupings, such as counties, that are often poor proxies for social context or environmental exposures,^[3] and that ecological relations can be extremely sensitive to the degree of grouping.^[40]

The problem illustrated in Table 1 also applies to ecological estimates of average individual-level effects (cross-level inferences from the ecological to the individual level).^[3,9,19] For example, if the illustration in Table 1 represented a contrast of two areas, A and B, with the goal of estimating the impact of differences in the X distribution, we would see very small contextual effects obscure substantial X effects in the ecological data.

To summarize, observational ecological data alone tell us little about either contextual or individual-level effects on summary measures of population health because they are compatible with a broad range of individual-level associations. Even if the overall goal is to estimate contextual effects, ecological manifestations of those effects (panel 1a of Table 1) can be confounded by individual-level relations and are not estimable without information about those relations. Thus methods that purport to adjust ecological results for the confounding problem just described either must employ external data about individual relations, or they must make assumptions about those relations. The

unobserved nature of the individual relations means that neither approach can be fully tested (validated) against the ecological data.

SOME ASSUMPTIONS AND FALLACIES IN ECOLOGICAL ANALYSES

All too often, identification is forced by making arbitrary modeling assumptions. Controversy then surrounds the credibility or strength of the assumptions used to derive effect estimates from ecological data, the sensitivity of those estimates to changes in assumptions, and failures of proposed methods in real or simulated data. For examples, compare Freedman et al.^[41] with their discussants; compare Greenland and Morgenstern,^[42,43] Richardson and Hemon,^[44] Greenland,^[45,46] Greenland and Robins,^[25,26] Piantadosi,^[47] Stidley and Samet,^[48] and Lagarde and Pershagen^[49,50] with Cohen;^[36–38,51] compare King^[28,29] with Rivers,^[30] Cho,^[52] Freedman et al.,^[31,32] and the example in Stoto;^[53] and compare Wen and Kramer^[54] with Naylor.^[27]

All causal inferences from observational epidemiological data must rely on restrictive assumptions about the distributions of errors and background causes. Most estimates also rely on parametric models for effects. Thus the validity of inferences depends on examining the sensitivity of the estimates to violations of the underlying assumptions and models.

RANDOMIZATION ASSUMPTIONS

Interpreting an association as a causal effect depends on some sort of no-confounding or ignorability assumption, which in statistical methodology becomes operationalized as a covariate-specific randomization assumption.^[55,56] Such causal inferences are usually sensitive to violations of those assumptions, and this sensitivity is a major source of controversy in most nonexperimental research.

Suppose we are to study K communities. The distinction between ecological and individual-level confounding may become clearer if we contrast two levels of randomized intervention to reduce sexually transmitted diseases (STDs):

- C) A community health program (e.g., establishment of free STD clinics) is provided at random to half of the communities (K is assumed to be even).
- W) Within community k , a treatment (e.g., a series of free STD-clinic visits) is randomly provided to a proportion p_k of individuals, and p_k varies across communities.

Trial C is a cluster-randomized design. In this trial, the ecological data would comprise a community treatment-status indicator plus the outcome measures

(e.g., subsequent STD rates). These data would support randomization-based inferences about what the average causal effect of the program would be for the K communities (the communities would be the analysis units, and the sample size for the inferences would be K). Nonetheless, analyzing the individual-level data from trial C as a K -stratum study (the usual individual-level analysis) would support no inferences at all about treatment effects because each stratum (community) would have a zero margin. Put another way, community and treatment effects would be completely confounded in the standard individual-level model. Analyzing the individual-level data would instead require methods for cluster-randomized trials.

In trial W, the ecological data consist of the proportion treated (p_k) in each community, along with outcome measures. Unless the p_k had been randomly assigned across communities (as in trial C, in which $p_k = 0$ or 1), the ecological data would *not* support randomization-based inferences about treatment effects. If the p_k were constant, there would be no data for such an analysis; if the p_k varied, the community and treatment effects would be confounded. (This observation is essentially a contrapositive version of Goodman's identification condition for ecological regression,^[57] translated to the present causal setting.) Nonetheless, the individual-level data from any of or all the communities with $0 < p_k < 1$ would support the usual randomization-based inferences (e.g., exact tests stratified on community). Taking X as the treatment indicator and $k = A, B$, the information in Table 1 can be used as an example of trial W with $p_A = 0.6$ and $p_B = 0.4$. The example then exhibits confounding of the ecological data with no confounding of the individual-level data within community.

To summarize, different randomization assumptions underlie causal inferences from individual-level and ecological studies; consequently, valid inferences from these two study types may require controlling for different (although possibly overlapping) sets of covariates.^[25]

OBSERVATIONAL STUDIES

Causal interpretations of associations in observational data have a fundamental weakness: The validity of such interpretations depends on the assumptions that natural or social circumstances have miraculously randomized the study exposure, for example, by carrying out trial C or W for us. The number of individuals in an individual-level study tells us nothing about the validity of this or other such "no-confounding" assumptions. Larger size only increases the precision of randomization-based inferences by reducing the chance that randomization left large imbalances of uncontrollable covariates across treatment groups. This benefit of size stems from, and depends on, an assumption of exposure randomization, as in trial W. Systematic (non-random) imbalances within groups by definition violate that assumption.

The same argument applies to ecological studies. The number of ecological groups involved tells us nothing about the validity of an assumption that the exposure distribution (p_k in the above binary-treatment trials) was randomized across the groups. The benefit of a large number of groups stems from and depends on the assumption that those distributions were randomized, as in trial C. Again, systematic imbalances across groups by definition violate that assumption. Despite these facts, some defenses of ecological studies have based on the circular argument that large numbers of areas would reduce the chance of ecological confounding;^[36] this circularity arises because the large-number effect assumes randomization across areas, which is precisely the assumption in doubt.

COVARIATE CONTROL

To achieve some plausibility in causal inferences from observational data, researchers attempt to "control" for covariates that affect the outcome but that are not affected by the exposure (potential confounders). In individual-level studies, the traditional means of control is to stratify the data on these covariates, because within such strata the exposed and unexposed units cannot have any imbalance on the covariates beyond the stratum boundaries; e.g., within a 65- to 74-year-old age stratum, the exposed and unexposed could not be more than 10 years apart in age. Causal inferences then proceed by assuming randomization *within* these strata. However implausible this assumption may be, in the face of observed imbalances it is always more plausible than the assumption of simple (unstratified) randomization.

Ecological data can be stratified, but with few exceptions, the ecological exposures and covariates in public-use databases are insufficient in detail and accuracy to create strata with assured balance on key covariates. For example, in ecological studies of radon levels and lung cancer rates across U.S. counties, the key covariates are the county-specific distributions of age, sex, race, and smoking habits. To address concerns about bias from possible relationships of these strong lung cancer risk factors to radon exposure, the county data would have to be stratified by age, sex, race, and smoking behavior (smoking behavior is multi-dimensional: it includes intensity, duration, and type of cigarette use). The relationship of radon distributions to lung cancer rates would then be examined across the stratum-specific county data. This stratified analysis requires the county-specific joint distributions of age, sex, race, smoking behavior and radon, and age, sex, race, smoking behavior, and lung cancer. Unfortunately, to date no such data have been available. Although data on the age-sex-race-lung cancer distributions in counties are published, their joint distributions with radon and smoking are unobserved. Only marginal distributions of radon are surveyed, and only crude summaries of cigarette sales are available.

The limitations of ecological stratification data may be better appreciated by considering an analogous problem in an individual-level study of residential radon and lung cancer. One might attempt to “control” for smoking by using cigarette sales in a subject’s county of residence as a surrogate for smoking behavior, as in Cohen.^[36] Presumably, few epidemiologists would regard this strategy as adequately controlling for smoking, especially considering that it would impute an identical “smoking” level to every subject in the county, regardless of age, sex, or lung cancer status. The shortcomings of this strategy arise precisely because smoking behavior varies greatly among individuals in a county, much more so than average smoking behavior varies across counties.^[25]

MODELING ASSUMPTIONS

To circumvent the inherent limitations of ecological studies, investigators have employed models in which ecological data (simple marginal summaries of crude exposure and covariate measures) are sufficient for causal inference. As mentioned earlier, these models are restrictive and are no more supported by ecological data than randomization assumptions. For example, a common assumption in ecological analyses is that effects follow a multiple-linear regression model. This assumption is both natural and somewhat misleading: A multiple-linear model for individual-level effects induces a multiple-linear ecological model, but this parallel relation between the individual and ecological regressions fails in the presence of non-additive or non-linear effects in the ecological groups.^[25,58–62]

Not even the functional form of individual-level effects (which can confound ecological associations, as in Table 1) is identified by the marginal data in ecological studies. For example, suppose individual risk R depends on the covariate vector X (which may contain contextual variables) via $R = f(X)$, and A indicates a context (such as geographic area). The ecological observations identify only relations of average risks $E_A(R)$ to average covariate levels $E_A(X)$ across contexts. These ecological relationships generally do not follow the same functional form as the individual relationships because $E_A(R) = E_A[f(X)] \neq f[E_A(X)]$, except in very special cases, chiefly those in which f is additive and linear in all the X components.

To prevent negative rate estimates, most analyses of individual-level epidemiological studies assume a multiplicative (log-linear) model for the regression of the outcome rate on exposure and analysis covariates, in which $f(X) = \exp(X\beta)$. Such non-linearities in individual regressions virtually guarantee that simple ecological models will be misspecified. These non-linearities further diminish the effectiveness of controlling ecological confounding, although the problem can be mitigated somewhat by expanding the ecological model to include more detailed

covariate summaries, if available,^[7,25] and by including higher-order covariate terms.^[44,59,62]

Unfortunately, some authors have denied the misspecification problem by claiming that a linear regression is justified by Taylor’s theorem.^[36] This justification is circular because approximate linearity of $f(X)$ over the within-context (area-specific) range of X is required for a first-order Taylor approximation of $f(X)$ to be accurate.^[25,60] Furthermore, in most applications, this requirement is known to be violated. For example, the dependence of risk on age is highly non-linear for nearly all cancers. Attempts to circumvent this non-linearity by using age-specific outcomes encounter a lack of age-context-specific measures of potential confounders, such as smoking. Using age-standardized rates also fails to solve the problem because it requires use of age-standardized measures of the covariates in the regression model (see “Other Special Problems of Ecological Analysis”), and such measures are rarely available.

MULTI-LEVEL MODEL THEORY FOR ECOLOGICAL INFERENCE

Most of the points discussed in this section are consequences of standard multi-level modeling theory.^[63] Suppose Y and X are an outcome (response) variate and a covariate vector defined on individuals, and R and Z are an outcome and covariate vector defined directly on aggregates (clusters); X may contain products among individual and aggregate covariates. For example, Y could be an indicator of being robbed within a given year; X could contain age, income, and other variables. With state as the aggregate, R could be the proportion robbed in the state over the year (which is just the average Y in the state), and Z could contain state average age and income and an indicator of whether hypodermic needles are available without prescription in the state.

Let μ_{ik} and x_{ik} be the expected outcome and X -value of individual i in aggregate k , with $\bar{\mu}_k$ and $\bar{x}_k$ their averages and z_k the Z -value for aggregate k , $k = 1, \dots, K$. One generalized-linear model (GLM) for individuals is

$$\mu_{ik} = f(\alpha + x_{ik}\beta + z_k\gamma) \quad (1)$$

If $R = \bar{Y}$, as when R is a rate, proportion, or average lifespan, model 1 induces the aggregate-outcome model

$$E_k(R) = \bar{\mu}_k = \frac{\sum_{i=1}^{N_k} f(\alpha + x_{ik}\beta + z_k\gamma)}{N_k} \quad (2)$$

where N_k is the size of aggregate k .^[61,64] If $\beta \neq 0$, $\bar{\mu}_k$ will depend on the within-aggregate X -distribution as well as on the individual-specific effects β and the contextual effects γ .

SPECIAL CASES

If the individual model is linear (so that $f(u) = u$), and if $Z = \bar{X}$, models 1 and 2 become

$$\mu_{ik} = \alpha + x_{ik}\beta + \bar{x}_k\gamma \quad (3)$$

And

$$\bar{\mu}_k = \alpha + \bar{x}_k(\beta + \gamma) \quad (4)$$

Model 4 exhibits the complete confounding of the individual effects β and the aggregate (contextual) effects γ in ecological data.^[14,23] Although the induced regression is linear and can be fit with ecological data, the identified coefficient $\beta + \gamma$ represents a blend of individual and contextual effects, rather than just contextual effects, as often supposed. To see this point, let Y again be a robbery indicator, let X be income (in thousands of dollars), and let k index states. Then β represents the difference in individual robbery risks associated with a \$1,000 increase in individual income X within a given state; γ represents the difference in individual robbery risks associated with a \$1000 increase in state-average income $\bar{X}$ among individuals of a given income X ; and $\beta + \gamma$ represents the unconditional difference in average state risk associated with a \$1,000 increase in average state income. In other words, β represents an *intrastate* association of personal income with personal robbery risk; γ represents an *interstate* association of average (contextual) state income with personal robbery risk within personal-income strata; and $\beta + \gamma$ represents an unconditional interstate association of average income with average risk, blending the conditional associations with robbery risk of personal and state-average income differences.

Linear models usually poorly represent risk and rate outcomes because of their failure to constrain predicted outcomes to nonnegative values. If we drop the linearity requirement but maintain $Z = \bar{X}$, model 1 can be rewritten

$$\mu_{ik} = f[\alpha + (x_{ik} - \bar{x}_k)\beta + \bar{x}_k(\beta + \gamma)] \quad (5)$$

showing that β may also be interpreted as the within-aggregate (intra-cluster) association of the transformed individual expected outcome $f^{-1}(\mu)$ with departures of X from the aggregate-specific average $\bar{X}$; $\beta + \gamma$ is then the complementary association of $\bar{X}$ with $\bar{\mu}$ across aggregates.^[9] The departure $X - \bar{X}$ is, however, a contrast of individual and contextual characteristics X and $\bar{X}$, rather than a purely individual characteristic X as in model 1, and $\beta + \gamma$ remains a mixture of individual and contextual effects.

Unfortunately, not even the mixed effect $\beta + \gamma$ can be estimated unbiasedly without further assumptions if the model is non-linear and if the only available covariates are aggregate means $\bar{X}$ and purely contextual variables. Richardson et al.^[59] and Dobson^[60] discussed assumptions

for unbiased ecological estimation of $\beta + \gamma$ under log-linear models, in which $f(u) = e^u$, although they implicitly assumed $\gamma = 0$ (no contextual effect) and so described them as assumptions for estimating β ; see also Sheppard,^[9] Prentice and Sheppard,^[61] Greenland,^[45] and Wakefield.^[65] From a scientific perspective these assumptions can appear highly artificial (e.g., X multivariate normal with constant covariance matrix across aggregates, which cannot be satisfied if X has discrete components that vary in distribution across aggregates), although the bias from their violation might not always be large.^[59,60,64] Nonetheless, even if the bias in estimating $\beta + \gamma$ is small, β and γ remain inseparable without further assumptions or individual-level data.

EXTENSIONS OF THE BASIC MODEL

Model 1 is a bi-level model, in that it incorporates covariates for individual-level characteristics (the X) and for one level of aggregation (the Z). The model can be extended to incorporate characteristics of multiple levels of aggregation.^[63] For example, the model could include a covariate vector Z_c of census-tract characteristics, such as the average income in the tract, and another vector Z_s of state characteristics, such as the law indicators and enforcement variables (e.g., conviction rates and average time served for various felonies). The above points can be roughly generalized by saying that the absence of data on one level should be expected to limit inferences about effects at other levels. The term *cross-level bias* denotes bias in coefficient estimates on one level that results from incorrect assumptions about effects on other levels (e.g., linearity or absence of effects); similarly, *cross-level confounding* denotes inseparability of different levels of effect, as illustrated above.

An important extension of multi-level models adds between-level covariate products (“interactions”) to model 1, thus allowing the individual effects β to vary across groups. Even with correct specification and multi-level data, the number of free parameters that results can be excessive for ordinary methods. One solution is to constrain the variation with hierarchical coefficient models, which induces a mixed model for individual outcomes.^[63] For example, adding products among X and Z to model 1 is a special case of replacing β by θ_k , with θ_k constrained by the second-stage model

$$\theta_k = \beta + A_k \delta \quad (6)$$

where A_k is a known fixed matrix function of z_k . The ensemble of estimates of the vectors $\theta_1, \dots, \theta_K$ can then be shrunk toward a common estimated mean $\bar{\beta}$ by specifying a mean-zero, variance τ^2 distribution for δ . This process is equivalent to (empirical) Bayesian averaging of model 1 with the expanded model:^[66]

$$\mu_{ik} = f(\alpha + x_{ik}\beta + z_k\gamma + x_{ik}A_k\delta) \quad (7)$$

Unfortunately, the nonidentifiability and confounding problems of ecological analysis only worsen in the presence of cross-level interactions.^[42]

ECOLOGICAL REGRESSION

Many, if not most, ecological researchers do not consider the individual or aggregated model 1 or 2 but instead directly model the aggregated outcome $\bar{Y}$ with an ecological regression model. If $Z = \bar{X}$, an ecological generalized linear model is

$$E_k(\bar{Y}) = \bar{\mu}_k = h(a + \bar{x}_k b) \quad (8)$$

With $h(u) = u$ and individual linearity (model 3), the results given above imply that b will be a linear combination of individual and contextual effects. Given individual non-linearities, model 8 will at best approximate a complex aggregate regression like model 2, and b will represent a correspondingly complex mixture of individual and contextual effects.^[9] These results also apply when purely contextual variables are included with $\bar{X}$ in Z and $\bar{x}_k$ is replaced by z_k in model 8.

The classic “ecological fallacy” (perhaps more accurately termed cross-level bias in estimating individual effects from ecological data) is to identify b in model 8 with β in model 1. This identification is fallacious because unless $\gamma = 0$ (contextual effects absent), even linearity does not guarantee $b = \beta$. With individual-level non-linearity, even $\gamma = 0$ does not guarantee $b = \beta$. Ecological researchers interested in estimating contextual effects should note the symmetrical results that apply when estimating contextual effects from ecological data. Even with linearity, $b = \gamma$ will require $\beta = 0$, which corresponds to no individual-level effects of X , even when the aggregate summaries of X in Z (such as $\bar{X}$) have effects. In other words, for the ecological coefficient b to correspond to the contextual effect γ , any within-group heterogeneity in individual outcomes must be unrelated to the individual analogues of the contextual covariates. For example, if X is income and Z is mean income $\bar{X}$, $b = \gamma \neq 0$ would require individual incomes to have no effect except as mediated through average area income. Such a condition is rarely, if ever, plausible.

MULTI-LEVEL STUDIES

Problems arising from within-group heterogeneity should diminish as the aggregates are made more homogeneous on relevant covariates. Heterogeneity of social factors may also decline as the aggregates become more restricted (e.g., census tracts tend to be more homogeneous on income and education than counties). Nonetheless, the benefits of such restriction may be nullified if important covariate measures (e.g., alcohol and tobacco) are available only for the

broader aggregates, because confounding by those covariates may remain uncontrollable.

Suppose, however, that within-region, individual-level survey data are available for those covariates. Individual-level risk models may then be applied to within-region survey data and the resulting individual risk estimates aggregated for comparison to observed ecological rates.^[60,67] For example, suppose we have a covariate vector X measured on N_k surveyed individuals in region k , a rate model $r(x; \beta)$ with a β estimate $\hat{\beta}$ from individual-level studies (e.g., a proportional-hazards model derived from cohort-study data), and the observed rate $\bar{r}_k$ in region k . We may compare $\bar{r}_k$ to the area rate predicted from the model applied to the survey data, $\Sigma_i r(x_i; \hat{\beta}) / N_k$, where the sum is over the surveyed individuals $i = 1, \dots, N_k$ and x_i is the value of X for survey individual i . This approach is a regression analogue of indirect adjustment: $\hat{\beta}$ is the external information and corresponds to the reference rates used to construct expectations in standardized morbidity ratios.

Unfortunately, fitted models that generalize to the regions of interest are rarely available. Thus, starting from the individual-level model $r_k(x; \beta) = r_{0k} \exp(x\beta)$, Prentice and Sheppard^[61] and Sheppard and Prentice^[64] proposed estimating the individual parameters β by directly regressing the $\bar{r}_k$ on the survey data using the induced aggregate model $r_k = r_{0k} E_k[\exp(x\beta)]$, where $E_k[\exp(x\beta)]$ is the average of $\exp(x\beta)$ over the individuals in region k . Sheppard and Prentice^[64] showed how the observed rates $\bar{r}_k$ and the survey data (the x_i) can be used to fit this model. As with earlier methods, their method estimates region-specific averages from sample averages, but in the absence of external data on β , it constrains the region-specific baseline rates r_{0k} (e.g., by treating them as random effects); see Cleave et al.^[12] and Wakefield^[10] for related approaches.

Prentice and Sheppard^[61] called their method an “aggregate-data study”; however, much of the social science literature has long used this term as a synonym for “ecological study,”^[14,33] and so I would instead call it an incomplete multi-level study (“incomplete” because, unlike a complete multi-level study, individual-level outcomes are not obtained). Prentice and Sheppard^[61] conceived their approach in the context of cancer research, in which few cases would be found in modestly sized survey samples. For studies of common acute outcomes, Navidi et al.^[6] proposed a complete multi-level strategy in which the within-region surveys obtain outcome as well as covariate data on individuals, which obviates the need for identifying constraints.

Multi-level studies can combine the advantages of both individual-level and ecological studies, including the confounder control achievable in individual-level studies and the exposure variation and rapid conduct achievable in ecological studies.^[7] These advantages are subject to many assumptions that must be carefully evaluated.^[61,65] Multi-level studies share several of these assumptions with ecological studies. For example, multi-level studies based on

recent individual surveys must assume stability of exposure and covariate distributions over time to ensure that the survey distributions are representative of the distributions that determined the observed ecological rates. This assumption will be suspect when individual behavioral trends or important degrees of migration emerge after the exposure period relevant to the observed rates.^[68,69] Multi-level studies can also suffer from the problem, mentioned earlier, that the aggregate-level (ecological) data usually concern arbitrary administrative or political units, and so can be poor contextual measures. Furthermore, multi-level studies face a major practical limitation in requiring data from representative samples of individuals within ecological units, which can be many times more expensive to obtain than the routinely collected data on which most ecological studies are based.

In the absence of valid individual-level data, multi-level analysis can still be applied to the ecological data via non-identified random-coefficient regression. As in applications to individual-level studies,^[70] this approach begins with specifying a hierarchical prior distribution for parameters not identified by available data. The distribution for the effect of interest is then updated by conditioning on the available data. This approach is a multi-level extension of random-effects ecological regression to allow constraints on β (including a distribution for β) in the aggregate model, in addition to constraints on the r_{0k} . It is essential to recognize that these constraints are what identify exposure effects in ecological data; hence, as with all ecological results, a precise estimate should be attributed to the constraints that produced the precision.

OTHER SPECIAL PROBLEMS OF ECOLOGICAL ANALYSIS

Thus far, I have focused on problems of confounding and cross-level effects in ecological studies. These are not the only problems of ecological studies. Several others are especially noteworthy for their divergence from individual-level study problems or their absence from typical individual-level studies.

NON-COMPARABILITY AMONG ECOLOGICAL ANALYSIS VARIABLES

The use of non-comparably standardized or restricted covariates in an ecological model such as model (8) can lead to considerable bias.^[45,71] Typically, this bias occurs when $\bar{Y}$ is age-standardized and specific to sex and race categories but $\bar{X}$ is not. For example, disease rates are routinely published in age-standardized, sex-race-specific form, but most summary covariates (e.g., alcohol and tobacco sales) are not. Such non-comparability can lead to bias in the ecological coefficient b in model 8 as a representation of

the mixed effect $\beta + \gamma$, because the coefficients of crude covariate summaries will not capture the rate variation associated with variation in alcohol and tobacco use by age, sex, and race.

Unfortunately, non-comparable restriction and standardization of variables remain common in ecological analyses. Typical examples involve restricted standardized rates regressed on crude ecological variables, such as sex-race-specific, age-standardized mortality rates regressed on contextual variables (e.g., air-pollution levels) and unrestricted unstandardized summaries (e.g., per capita cigarette sales). If, as is usual, only unrestricted, unstandardized regressor summaries are available, less bias might be incurred by using the *crude* (rather than the standardized) rates as the outcome and controlling for ecological demographic differences by entering multiple age-sex-race-distribution summaries in the regression^[71] (e.g., proportions in different age-sex-race categories). More work is needed to develop methods for coping with non-comparability.

MEASUREMENT ERROR

The effects of measurement error on ecological and individual-level analyses can be quite different. For example, Brenner et al.^[72] found that independent non-differential misclassification of a binary exposure could produce bias away from the null and even reverse estimates from a standard linear or log-linear ecological regression analysis, although the same error would produce only bias toward the null in a standard individual-level analysis. Analogous results were obtained by Carroll^[73] for ecological probit regression with a continuous measure of exposure. Results of Brenner et al.^[72] also indicate that ecological regression estimates can be extraordinarily sensitive to errors in exposure measurement. On the other hand, Brenner et al.^[74] found that independent non-differential misclassification of a single binary confounder produced no increase in bias in an ecological linear regression. Similarly, Prentice and Sheppard^[61] and Sheppard and Prentice^[64] found robustness of their incomplete multi-level analysis to purely random measurement errors.

Besides assuming very simple models for individual-level errors, the foregoing results assume that the ecological covariates in the analysis are direct summaries of the individual measures. Often, however, the ecological covariates are subject to errors beyond random survey error or individual-level measurement errors, for example, when per capita sales data are used as a proxy for average individual consumption, and when area measurements such as pollution levels are subject to errors. Some researchers have examined the impact of such *ecological* measurement error on cross-level inferences under simple error models,^[6,8,75] but more research is needed, especially for the situation in which the grouping variable is

an arbitrary administrative unit serving as a proxy for a contextual variable.

PROBLEMS IN CAUSAL INTERPRETATION

The present article, like most of the literature, has been rather loose in using the word *effect*. Strictly speaking, the models and results discussed here apply only to interpretations of coefficients as measures of conditional associations. Their interpretation as measures of causal effects in formal causal models such as potential-outcome (counterfactual) models^[76–79] raises some problems not ordinarily encountered in individual-level modeling.

To avoid technical details, consider the linear-model example. A causal interpretation of β in model 3 is that it represents the change in an individual's risk that would result from increasing that person's income by \$1,000. Apart from the ambiguity of this intervention (does the money come from a pay raise or direct payment?), one might note that it would also raise the average income value $\bar{X}_k$ by $\$1,000/N_k$ and produce an additional change of γ/N_k in the risk of the individual. This interpretation apparently contradicts the original interpretation of β as *the* change in risk, from whence we see that β must be interpreted carefully as the change in risk given that average income is (somehow) held constant, for example, by reducing everyone else's income by $\$1,000/(N_k - 1)$ each.

If the size of the aggregate N_k is large, this problem is trivial; if it is small, however (e.g., the aggregate is a census tract), it may be important. One solution with a scientific basis is to recognize $\bar{X}$ as a contextual surrogate for socio-economic environment, in which case the contextual value $\bar{X}_k$ in model 1 might be better replaced by an individual-level variable $\bar{X}_{(-i)k}$, the average income of persons in aggregate k apart from individual i . Note that the mean of $\bar{X}_{(-i)k}$ is $\bar{X}_k$, and so its effect could not be separated from that of personal income when only aggregate means are available; that is, β and γ remain completely confounded.

Causal interpretation of the contextual coefficient γ is symmetrically problematic. This coefficient represents the change in risk from a \$1,000 increase in average income $\bar{X}_k$, while keeping the person's income x_{ik} the same. This income can be kept constant in many ways, for example, by giving everyone but person i a $\$1,000N_k/(N_k - 1)$ raise, or by giving one person other than i a $\$1,000N_k$ raise. The difference between these two interventions is profound and might be resolved by clarifying the variable for which $\bar{X}$ is a surrogate. Similar comments apply to causal interpretation of b in the ecological model (model 8).

Problems of causal interpretation only add to the difficulties in the analysis of contextual causal effects, but, unlike confounding, they may be as challenging for individual-level and multi-level studies as they are for ecological studies.^[2] Addressing these problems demands that detailed subject-matter considerations be used to construct

explicit multi-level causal models in which key variables at any level may be latent (not directly measured).

PROBLEMS IN AGGREGATION LEVEL

Another serious and often overlooked conceptual problem involves the special artificiality of typical model specifications above the individual level. Ecological units are usually arbitrary administrative groupings, such as counties, which may have only weak relations to contextual variables of scientific interest, such as social environment.^[3] Compounding this problem is the potentially high sensitivity of aggregate-level relationships to the grouping definitions.^[40] In light of these problems, it may be more realistic to begin with at least a tri-level specification, in which one aggregate level is of direct scientific interest (e.g., neighborhood or community) and the other (or others) are the aggregates for which data are available (e.g., census tracts or counties). This larger initial specification would at least clarify the proxy-variable (measurement surrogate) issues inherent in most studies of contextual effects.

CONCLUSION

The validity of any inference can only benefit from critical scrutiny of the assumptions on which it is based. Users of surveillance data should be alert to the special assumptions needed to estimate an effect from ecological data alone, and to the profound sensitivity of inferences to those assumptions (even if the estimate is of a contextual effect). The fact that individual-level studies have complementary limitations does not excuse the failure to consider these assumptions.

Nonetheless, ecological data are worth examining. Their value has been demonstrated by careful ecological analyses^[59,80] and by methods that combine individual and ecological data.^[6,18,61,64,65,81–83] Furthermore, the *possibility* of bias does not prove the presence of bias, and a conflict between ecological and individual-level estimates does not by itself mean that the ecological estimates are the more biased.^[20,25,26] The two types of estimates share some biases but not others, making it possible for the individual-level estimates to be the more biased.

The effects measured by the two types of estimates may differ as well; ecological estimates incorporate a contextual component that is absent from the individual estimates derived from a narrow context. Indeed, the contextual component may be viewed as both a key strength and weakness of ecological studies, because it is often of greatest importance, although it is especially vulnerable to confounding. Thus, in the absence of good multi-level studies, ecological studies will no doubt continue to fuel controversy and so inspire the conduct of potentially more informative studies.

ACKNOWLEDGMENTS

Most of the material was adapted from Refs. [5] (Copyright 2001, Oxford University Press) and [19] (Copyright 2002, John Wiley and Sons).

REFERENCES

- Langbein, L.I.; Lichtman, A.J. *Ecological Inference*; Series/No. 07-010; Sage: Thousand Oaks, CA, 1978.
- Iversen, G.R. *Contextual Analysis*; Sage: Thousand Oaks, CA, 1991.
- Achen, C.H.; Shively, W.P. *Cross-Level Inference*; University of Chicago Press: Chicago, IL, 1995.
- Morgenstern, H. Ecologic Studies. *Modern Epidemiology*, 3rd Ed.; Lippincott: Philadelphia, 2008; 511–531.
- Greenland, S. Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *Int. J. Epidemiol.* **2001**, *30*, 1343–1350.
- Navidi, W.; Thomas, D.; Stram, D.; Peters, J. Design and analysis of multilevel analytic studies with applications to a study of air pollution. *Environ. Health Perspect.* **1994**, *102* (Suppl. 8), 25–32.
- Guthrie, K.A.; Sheppard, L. Overcoming biases and misconceptions in ecologic studies. *J. R. Stat. Soc. Ser. A* **2001**, *164*, 141–154.
- Wakefield, J.; Salway, R. A statistical framework for ecological and aggregate studies. *J. R. Stat. Soc. Ser. A* **2001**, *164*, 119–137.
- Sheppard, L. Insights on bias and information in group-level studies. *Biostatistics* **2003**, *4*, 265–278.
- Wakefield, J. Ecological inference for 2×2 tables. *J. R. Stat. Soc. Ser. A* **2004**, *167*, 385–445.
- Piantadosi, S.; Byar, D.P.; Green, S.B. The ecological fallacy. *Am. J. Epidemiol.* **1988**, *127*, 704–893.
- Cleave, N.; Brown, P.J.; Payne, C.D. Evaluation of methods for ecological inference. *J. R. Stat. Soc. Ser. A* **1995**, *158*, 55–72.
- Plummer, M.; Clayton, D. Estimation of population exposure in ecological studies. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 113–126.
- Firebaugh, G. Assessing Group Effects. *Aggregate Data: Analysis and Interpretation*; Sage: Beverly Hills, CA, 1980; 13–24.
- Von Korff, M.; Koepsell, T.; Curry, S.; Diehr, P. Multilevel analysis in epidemiologic research on health behaviors and outcomes. *Am. J. Epidemiol.* **1992**, *135*, 1077–1082.
- Susser, M.; Susser, E. Choosing a future for epidemiology: II. From black box to Chinese boxes and eco-epidemiology. *Am. J. Public Health* **1996**, *86*, 674–677.
- Duncan, C.; Jones, K.; Moon, G. Health-related behaviour in context: A multilevel modeling approach. *Soc. Sci. Med.* **1996**, *42*, 817–830.
- Duncan, C.; Jones, K.; Moon, G. Context, composition and heterogeneity: Using multilevel models in health research. *Soc. Sci. Med.* **1998**, *46*, 97–117.
- Greenland, S. A review of multilevel theory for ecologic analysis. *Stat. Med.* **2002**, *21*, 389–395.
- Schwartz, S. The fallacy of the ecological fallacy: The potential misuse of a concept and the consequences. *Am. J. Public Health* **1994**, *84*, 819–824.
- Pearce, N. Traditional epidemiology, modern epidemiology, and public health. *Am. J. Public Health* **1996**, *86*, 678–683.
- Koopman, J.S.; Longini, I.M. Jr. The ecological effects of individual exposures and nonlinear disease dynamics in populations. *Am. J. Public Health* **1994**, *84*, 836–842.
- Firebaugh, G. A rule for inferring individual-level relationships from aggregate data. *Am. Sociol. Rev.* **1978**, *43*, 557–572.
- Hakama, M.; Hakulinen, T.; Pukkala, E.; Saxen, F.; Teppo, L. Risk indicators of breast and cervical cancer on ecologic and individual levels. *Am. J. Epidemiol.* **1982**, *116*, 990–1000.
- Greenland, S.; Robins, J. Ecologic studies—Biases, misconceptions, and counterexamples. *Am. J. Epidemiol.* **1994**, *139*, 747–760.
- Greenland, S.; Robins, J.M. Accepting the limits of ecologic studies. *Am. J. Epidemiol.* **1994**, *139*, 769–771.
- Naylor, C.D. Ecological analysis of intended treatment effects: Caveat emptor. *J. Clin. Epidemiol.* **1999**, *52*, 1–5.
- King, G. *A Solution to the Ecological Inference Problem*; Princeton University Press: Princeton, NJ, 1997.
- King, G. The future of ecological inference (letter). *J. Am. Stat. Assoc.* **1999**, *94*, 352–354.
- Rivers, D. Review of “A solution to the ecological inference problem.” *Am. Polit. Sci. Rev.* **1998**, *92*, 442–443.
- Freedman, D.A.; Klein, S.P.; Ostland, M.; Roberts, M.R. Review of “A solution to the ecological inference problem.” *J. Am. Stat. Assoc.* **1998**, *93*, 1518–1522.
- Freedman, D.A.; Ostland, M.; Roberts, M.R.; Klein, S.P. Reply to King (letter). *J. Am. Stat. Assoc.* **1999**, *94*, 355–357.
- Borgatta, E.F.; Jackson, D.J.; Eds. *Aggregate Data: Analysis and Interpretation*; Sage: Beverly Hills, CA, 1980.
- Greenland, S.; Rothman, K.J. Measures of Occurrence. *Modern Epidemiology*, 3rd Ed.; Lippincott: Philadelphia, PA, 2008, 32–50.
- Duncan, O.D.; Davis, B. An alternative to ecological correlation. *Am. Sociol. Rev.* **1953**, *18*, 665–666.
- Cohen, B.L. Ecological versus case-control studies for testing a linear-no-threshold dose-response relationship. *Int. J. Epidemiol.* **1990**, *19*, 680–684.
- Cohen, B.L. In defense of ecological studies for testing a linear no-threshold theory. *Am. J. Epidemiol.* **1994**, *139*, 765–768.
- Cohen, B.L. Re: Parallel analyses of individual and ecologic data on residential radon, cofactors, and lung cancer in Sweden (letter). *Am. J. Epidemiol.* **2000**, *152*, 194–195.
- Susser, M. The logic in ecological. *Am. J. Public Health* **1994**, *84*, 825–835.
- Openshaw, S.; Taylor, P.H. The Modifiable Area Unit Problem. *Quantitative Geography*; Routledge: London, 1981. Chap. 9.
- Freedman, D.A.; Klein, S.P.; Sacks, J.; Smyth, C.A.; Everett, C.G. Ecological regression and voting rights (with discussion). *Eval. Rev.* **1998**, *15*, 673–816.
- Greenland, S.; Morgenstern, H. Ecological bias, confounding, and effect modification. *Int. J. Epidemiol.* **1989**, *18*, 269–274.

43. Greenland, S.; Morgenstern, H. Neither within-region nor cross-regional independence of covariates prevents ecological bias (letter). *Int. J. Epidemiol.* **1991**, *20*, 816–818.
44. Richardson, S.; Hémon, D. Ecological bias and confounding (letter). *Int. J. Epidemiol.* **1990**, *19*, 764–766.
45. Greenland, S. Divergent biases in ecologic and individual-level studies. *Stat. Med.* **1992**, *11*, 1209–1223.
46. Greenland, S. Author's reply. *Stat. Med.* **1995**, *14*, 328–330.
47. Piantadosi, S. Ecologic biases. *Am. J. Epidemiol.* **1994**, *139*, 761–764.
48. Stidley, C.; Samet, J.M. Assessment of ecologic regression in the study of lung cancer and indoor radon. *Am. J. Epidemiol.* **1994**, *139*, 312–322.
49. Lagarde, F.; Pershagen, G. Parallel analyses of individual and ecologic data on residential radon, cofactors, and lung cancer in Sweden. *Am. J. Epidemiol.* **1999**, *149*, 268–274.
50. Lagarde, F.; Pershagen, G. The authors reply (letter). *Am. J. Epidemiol.* **2000**, *152*, 195.
51. Cohen, B.L. Letter to the editor. *Stat. Med.* **1995**, *14*, 327–328.
52. Cho, W.T.K. If the assumption fits: A comment on the King ecologic inference solution. *Polit. Anal.* **1998**, *7*, 143–163.
53. Stoto, M.A. Review of "Ecological inference in public health." *Public Health Rep.* **1998**, *113*, 182–183.
54. Wen, S.W.; Kramer, M.S. Uses of ecologic studies in the assessment of intended treatment effects. *J. Clin. Epidemiol.* **1999**, *52*, 7–12.
55. Greenland, S. Randomization, statistics, and causal inference. *Epidemiology* **1990**, *1*, 421–429.
56. Greenland, S.; Robins, J.M.; Pearl, J. Confounding and collapsibility in causal inference. *Stat. Sci.* **1999**, *14*, 29–46.
57. Goodman, L.A. Some alternatives to ecological correlation. *Am. J. Sociol.* **1959**, *64*, 610–625.
58. Vaupel, J.W.; Manton, K.G.; Stallard, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **1979**, *16*, 439–454.
59. Richardson, S.; Stücker, I.; Hémon, D. Comparison of relative risks obtained in ecological and individual studies: Some methodological considerations. *Int. J. Epidemiol.* **1987**, *16*, 111–120.
60. Dobson, A.J. Proportional hazards models for average data for groups. *Stat. Med.* **1988**, *7*, 613–618.
61. Prentice, R.L.; Sheppard, L. Aggregate data studies of disease risk factors. *Biometrika* **1995**, *82*, 113–125.
62. Lasserre, V.; Guihenneuc-Jouyaux, C.; Richardson, S. Biases in ecological studies: Utility of including within-area distribution of confounders. *Stat. Med.* **2000**, *19*, 45–59.
63. Goldstein, H. *Multilevel Statistical Models*, 2nd Ed.; Edward Arnold: New York, 1995.
64. Sheppard, L.; Prentice, R.L. On the reliability and precision of within-and between-population estimates of relative rate parameters. *Biometrics* **1995**, *51*, 853–863.
65. Wakefield, J. Multi-level modelling, the ecologic fallacy, and hybrid study designs. *Int. J. Epidemiol.* **2009**, *38*, 330–336.
66. Greenland, S. Multilevel modeling and model averaging. *Scand. J. Work Environ. Health* **1999**, *25* (Suppl. 4), 43–48.
67. Kleinman, J.C.; DeGruttola, V.G.; Cohen, B.B.; Madans, J.H. Regional and urban-suburban differentials in coronary heart disease mortality and risk factor prevalence. *J. Chronic Dis.* **1981**, *34*, 11–19.
68. Stavray, K.M. The role of ecologic analysis in studies of the etiology of disease: A discussion with reference to large bowel cancer. *J. Chronic Dis.* **1976**, *29*, 435–444.
69. Polissar, L. The effect of migration on comparison of disease rates in geographic studies in the United States. *Am. J. Epidemiol.* **1980**, *111*, 175–182.
70. Greenland, S. When should epidemiologic regressions use random coefficients? *Biometrics* **2000**, *56*, 915–921.
71. Rosenbaum, P.R.; Rubin, D.B. Difficulties with regression analyses of age-adjusted rates. *Biometrics* **1984**, *40*, 437–443.
72. Brenner, H.; Savitz, D.A.; Jöckel, K.-H.; Greenland, S. Effects of nondifferential exposure misclassification in ecologic studies. *Am. J. Epidemiol.* **1992**, *135*, 85–95.
73. Carroll, R.J. Some surprising effects of measurement error in an aggregate data estimator. *Biometrika* **1997**, *84*, 231–234.
74. Brenner, H.; Greenland, S.; Savitz, D.A. The effects of nondifferential confounder misclassification in ecologic studies. *Epidemiology* **1992**, *3*, 456–459.
75. Wakefield, J.; Shaddick, G. Health-exposure modelling and the ecological fallacy. *Biostatistics* **2006**, *7*, 438–455.
76. Greenland, S.; Rothman, K.J.; Lash, T.L. Measures of Effect and Measures of Association. *Modern Epidemiology*, 3rd Ed.; Lippincott: Philadelphia, PA, 2008; 51–70.
77. Pearl, J. *Causality*, 2nd Ed.; Cambridge University Press: New York, 2009.
78. Little, R.J.A.; Rubin, D.B. Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annu. Rev. Public Health* **2000**, *21*, 121–145.
79. Greenland, S. Causal analysis in the health sciences. *J. Am. Stat. Assoc.* **2000**, *95*, 286–289.
80. Prentice, R.L.; Sheppard, L. Validity of international, time trend, and migrant studies of dietary factors and disease risk. *Prev. Med.* **1989**, *18*, 167–179.
81. Künzli, N.; Tager, J.B. The semi-individual study in air pollution epidemiology: A valid design as compared to ecologic studies. *Environ. Health Perspect.* **1997**, *105*, 1078–1083.
82. Haneuse, S.; Wakefield, J. The combination of ecological and case-control data. *J. R. Stat. Soc. Series B*, **2008**, *70*, 73–93.
83. Wakefield, J.; Haneuse, S. Overcoming ecological bias using the two-phase study design. *Am. J. Epidemiol.* **2008**, *167*, 908–916.

Yangxin Huang

University of Rochester, Rochester, New York, U.S.A.

Abstract

The investigator may be interested in the dose that results in a specified level of response, typically 50% or 90%. Such doses are referred to as effective doses. This entry addresses several research and statistical models to further break down ED₅₀ and ED₉₀ results.

INTRODUCTION

Data in which each observation assumes only one of two possible outcomes are called binary. This type of data arises in many situations. One area where binary data are frequently observed is biological assay. Much has been written about the toxicity of drugs and their relative effectiveness when compared with each other.^[1,2] Typically, test subjects are given specified doses of a drug and one observes whether they respond or not. Interest is often focused on understanding the relationship between the probability that the subject responds and the dose administered to the subject.

OVERVIEW

The investigator may be interested in the dose that results in a specified level of response, typically 50% or 90%. Such doses are referred to as effective doses. An important quantity of interest in biological assays is the value of x such that $p(x) = 0.5$; that is, experiments are frequently summarized by an estimate of the median effective dose (ED₅₀), namely that dose that causes 50% of the subjects to respond. (See Refs. [3,4] for a discussion of the pros and cons of the usage of the ED₅₀ and further literature about it.)

As an illustration, Table 1 is the data set supplied to Abdelbasit and Plackett^[5] by P. S. Hewlett, upon which Sitter and Wu^[6] based their configurations. The maximum likelihood (ML) fit of the logistic and probit models (see below for details) to the data and estimated ED₅₀ are displayed in Fig. 1.

In this paper, in addition to the ED₅₀, the estimation of a more extreme percentage point will be of interest. The 90% effective dose (ED₉₀) is specifically considered. The ED₅₀ has been a crucially important summary measure in bioassay for many years. The ED₉₀ is probably the typical “extreme” dose studied when one is interested in extreme doses rather than the median effective dose. A large

literature^[6–14] has considered the coverage probabilities of various methods of interval estimation of the ED₅₀, assuming a logistic dose–response curve. Engeman et al.^[15] and Robertson et al.^[16] investigated the properties of several traditional analytical procedures for estimating the 90% lethal dose (LD₉₀), including the percentage coverage of the LD₉₀ by nominal 95% confidence intervals. Two of the procedures considered were maximum likelihood estimation and minimum chi-squared estimation, under the assumption of a logistic model. However, little attention has been paid to what happens to the performance of these estimators if the assumption of a logistic dose–response curve is incorrect. It is important to consider this eventually, as the dose–response relationship is usually unknown; thus we are usually never certain that what we assume is actually correct.

We focus our interest on the performance of four interval estimation methods of the ED₅₀/ED₉₀, derived from incorrectly assuming a logistic dose–response curve when an alternative dose–response model is actually true, and address the question of their robustness. Three alternative models are considered, these being a probit dose–response curve, a cubic logistic dose–response curve,^[17] and an asymmetric Aranda–Ordaz dose–response curve.^[18] The four methods of interval estimation of the ED₅₀/ED₉₀ considered here are Fieller’s (F) method,^[1,6] the likelihood ratio (LR) method,^[7,20] the Bartlett adjusted likelihood ratio (BLR) method,^[10,19] and the score test (ST) method.^[10,20] In the third section of this paper, a statistical model for binary response data is introduced to compare the performance of interval estimation methods. The four methods of interval estimation of the ED₅₀/ED₉₀ are briefly outlined in the fourth section. The numerical simulation experiments are briefly discussed in the fifth section. The distribution functions used for the true underlying dose–response relationship are described and the results of these experiments are presented in the sixth section. Finally, we conclude the paper with some discussions with regard to, in particular, the robustness of the four methods under model misspecification.

STATISTICAL MODEL

An observation at dose level x_i is classified as a response or nonresponse with respective probabilities p_i and $1 - p_i$. We consider the design problem where the statistician must choose a set of levels $\mathbf{x} = (x_1, \dots, x_k)$ and the number of independent observations $\mathbf{n} = (n_1, \dots, n_k)$ to take at these levels, subject to a fixed total $n = \sum_{i=1}^k n_i$, to make some inference on the basis of the observed frequencies of a response $\mathbf{r} = (r_1, \dots, r_k)$. The appropriate assumption would be to view $r_1, \dots, r_k$ as being mutually independent binomial (n_i, p_i) random variables with

$$\text{logit}(p_i) = \alpha + \beta x_i, \quad i = 1, \dots, k \quad (1)$$

The ED _{ν} ($\nu = 50, 90$) is then given by

$$\mu_\nu = (c_\nu - \alpha)/\beta \quad (2)$$

where $c_\nu = \ln[\nu/(100 - \nu)]$. In particular, $\mu_{50} = -\alpha/\beta$ and $\mu_{90} = (\ln 9 - \alpha)/\beta$. Our log-likelihood function is

$$L = \sum_{i=1}^k [r_i \log p_i + (n_i - r_i) \log(1 - p_i)] \quad (3)$$

The maximum likelihood estimator (MLE) of μ_ν is $\hat{\mu}_\nu = (c_\nu - \hat{\alpha})/\hat{\beta}$ (i.e. $\hat{\mu}_{50} = -\hat{\alpha}/\hat{\beta}$) and $\hat{\mu}_{90} = (\ln 9 - \hat{\alpha})/\hat{\beta}$, with $\hat{\alpha}$ and $\hat{\beta}$ denoting the respective MLEs of α and β .

When a model such as the logistic model is fitted to binary assay data, the model may be used to summarize the data, through the model parameters, and in addition the model may form the basis for comparison between different sets of data. In many cases, a binary response experiment is used to estimate the dose level x_i that will cause a particular response in 100 p_i % of the subjects to which it is administered. A commonly used summary of a fitted model is the ED₅₀. By itself, a point estimate of the ED₅₀ is generally viewed as being insufficient;^[4] a better summary usually requires the calculation of an interval estimate. Assuming a parametric model for the dose-response

Table 1 Toxicity of Oil to *Sitophilus granarius*

Log (dose) (x_i)	Number of subjects (n_i)	Number affected (r_i)	Observed proportion (p_i)
0.2810	47	47	1.00
0.2304	50	50	1.00
0.1523	50	50	1.00
0.0864	50	46	0.92
-0.0362	50	25	0.50
-0.0809	50	0	0.00
-0.1487	50	2	0.04
-0.2147	50	1	0.02
-0.3098	50	0	0.00

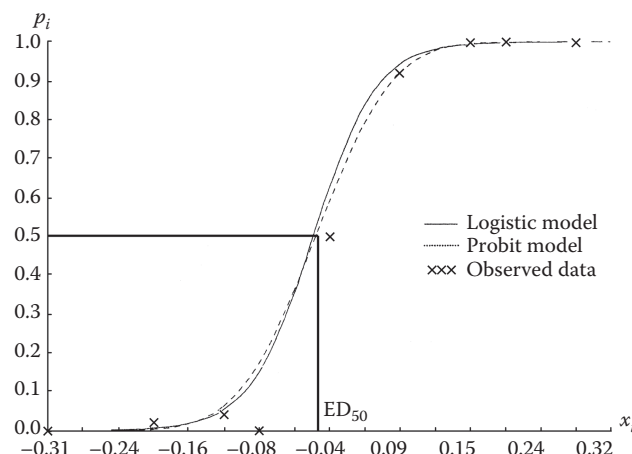


Fig. 1 The ML fit of the models to the data and estimated ED₅₀

relationship, several methods can be used to construct the confidence intervals for the ED ν ($\nu = 50, 90$).

THE FOUR METHODS OF INTERVAL ESTIMATION

The 95% Fieller interval for μ_ν is given by the set of $\mu_{\nu 0}$ satisfying

$$(\hat{\alpha} - c_\nu + \mu_{\nu 0} \hat{\beta})^2 (v_{11} + 2\mu_{\nu 0} v_{12} + \mu_{\nu 0}^2 v_{22})^{-1} < z_{0.05}^2 \quad (4)$$

where $V = (v_{ij})$ is the estimated asymptotic variance matrix of $(\hat{\alpha}, \hat{\beta})^T$, v_{11} and v_{22} are the estimated asymptotic variances, $\text{var}(\hat{\alpha})$ and $\text{var}(\hat{\beta})$, v_{12} is the estimated asymptotic covariance $\text{cov}(\hat{\alpha}, \hat{\beta})$, and $z_{0.05}^2$ is the upper 0.05 point of χ^2 with one degree of freedom (χ^2_1).

Suppose that under $H_0: \mu_\nu = \mu_{\nu 0}$, β is estimated, subject to $\mu_{\nu 0}$ by maximum likelihood and the value of the log-likelihood, $L_{H_0}(\mu_{\nu 0})$, is obtained. Under $H_1: \mu_\nu \neq \mu_{\nu 0}$, the overall maximum of the log-likelihood, maximized with respect to β and μ_ν , is L_{H_1} . Thus the 95% LR interval for μ_ν is given by the set of $\mu_{\nu 0}$ satisfying $\Gamma(\mu_{\nu 0}) < z_{0.05}^2$, where $\Gamma(\mu_{\nu 0}) = -2\{L_{H_0}(\mu_{\nu 0}) - L_{H_1}\} = D(\mu_{\nu 0}) - D(\hat{\mu}_\nu)$ is the deviance under the null hypothesis H_0 minus the deviance under the alternative hypothesis H_1 .

The Bartlett adjusted likelihood ratio test takes the form

$$\text{BLR} = \Gamma(\mu_{\nu 0})/\tilde{M} \quad (5)$$

Here M is the Bartlett factor, with the tilde denoting evaluation under $H_0: \mu_\nu = \mu_{\nu 0}$. The BLR statistic has an asymptotic χ^2_1 distribution under H_0 . As with the LR test itself, a 100(1- α)% BLR interval for μ_ν is the set of values $\mu_{\nu 0}$ not rejected by the test, namely those satisfying $\text{BLR} < z_{0.05}^2$. Explicitly, the Bartlett factor can be written as

$$M = 1 + (\varepsilon_2 - \varepsilon_1) \quad (6)$$

where

$$\varepsilon_2 = \text{tr}(\mathbf{H}\mathbf{Z}_d^2)/4 + \mathbf{1}_k^T \mathbf{F}(2\mathbf{Z}^{(3)} + 3\mathbf{Z}_d \mathbf{Z}_d) \mathbf{F} \mathbf{1}_k / 12 \quad (7)$$

and

$$\varepsilon_1 = \mathbf{1}_k^T \mathbf{H} \mathbf{x}^{(4)} / 4 (\mathbf{x}^T \mathbf{V} \mathbf{x})^2 + 5 (\mathbf{1}_k^T \mathbf{F} \mathbf{x}^{(3)})^2 / 12 (\mathbf{x}^T \mathbf{V} \mathbf{x})^3 \quad (8)$$

where the $k \times k$ matrices $\mathbf{F}$, $\mathbf{H}$, $\mathbf{V}$, and $\mathbf{Z}$ are given respectively by $\mathbf{F} = \mathbf{V}(\mathbf{I}_k - 2\mathbf{p})$, $\mathbf{H} = \mathbf{V}(6\mathbf{V}\mathbf{n}^{-1} - \mathbf{I}_k)$, $\mathbf{V} = \text{diag}\{n_i p_i (1 - p_i)\}$, and $\mathbf{Z} = \mathbf{X}(\mathbf{X}^T \mathbf{V} \mathbf{X})^{-1} \mathbf{X}^T$, $\mathbf{Z}_d$ is a diagonal matrix composed of the diagonal elements of $\mathbf{Z}$. Here $\mathbf{p} = \text{diag}(p_i)$, $\mathbf{n}^{-1} = \text{diag}(n_i^{-1})$, $\mathbf{x} = (x_1 - \mu_{v0}, \dots, x_k - \mu_{v0})^T$, $\mathbf{I}_k$ is a $k \times k$ identity matrix, $\mathbf{X}$ is the matrix with first column $\mathbf{1}_k$, $\mathbf{1}_k = (1, \dots, 1)^T$ of length k and second column the vector $(x_1, \dots, x_k)^T$. A superscript ($q = 3$ or 4) in brackets denotes the q th Hadamard product.^[10,19]

The score test under the null hypothesis $H_0: \mu_v = \mu_{v0}$ is given by

$$S = \frac{\sum_{i=1}^k \hat{w}_{i0} (x_i - \mu_{v0})^2 \left[\sum_{i=1}^k (r_i - n_i \hat{p}_{i0}) \right]^2}{\sum_{i=1}^k \hat{w}_{i0} \sum_{i=1}^k \hat{w}_{i0} (x_i - \bar{x}_{v0})^2} \quad (9)$$

where $\bar{x}_{v0} = \sum_{i=1}^k \hat{w}_{i0} x_i / \sum_{i=1}^k \hat{w}_{i0}$, $\hat{w}_{i0} = n_i \hat{p}_{i0} (1 - \hat{p}_{i0})$ with the circumflex denoting evaluation under $H_0: \mu_v = \mu_{v0}$. Under H_0 , S has, asymptotically, χ_1^2 distribution. Again, the confidence interval for μ_v is the set of values μ_{v0} not rejected by the test.

NUMERICAL EXPERIMENTS

The sets of experimental configurations used to compare the four methods of interval estimation of the ED $_{\nu}$ ($\nu = 50, 90$) are summarized in Table 2. Data sets (a) and (b) are based on configurations reported by William,^[7]

who referred to conducting experiments with a total of 20–30 subjects at four to six dose levels, and data set (c) is that reported by Sitter and Wu.^[6] data set (c) contains a number of extreme responses. These configurations are used to compare the four methods of interval estimation for the ED₅₀. The remaining three sets of configurations are used in the equivalent study for the ED₉₀. Data set (d) is based on the configurations with the six dose levels described by Engeman et al.,^[15] namely, equally spaced doses on a geometric scale with the theoretical levels of response for the first and last doses set at 0.05 and 0.95, respectively. Based on these response probabilities with $(\beta, \mu_{90}) = (1, 7.2)$, we calculated dose levels $x_1, \dots, x_6$. Data sets (e) and (f) reported by Robertson et al.^[16] corresponded to response probabilities $p_i = 0.2, 0.3, 0.4, 0.45, 0.55, 0.6, 0.7, 0.8$ and $p_i = 0.1, 0.15, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95$, respectively. Based on these response probabilities, we calculated dose levels presented in Table 2 corresponding to a logistic distribution with $\beta = 1$ and first dose level $x_1 = 0$. The choices of $(\beta_1, \mu_{90}) = (1, 3.5)$ and $(1, 4.4)$ in data sets (e) and (f), respectively, approximately correspond to the probabilities used by Robertson et al. All of the experiments were conducted with an equal number of observations per dose, so that $n_i = n/k$.

When the cubic logistic model was the true dose–response model, the values of γ chosen were $\gamma = 0, 0.05, 0.1, 0.15$, and 0.2 in data sets (a) and (b) and $\gamma = 0, 1, 2, 5$, and 10 in data set (c) for the ED₅₀, and $\gamma = 0, 0.001, 0.005, 0.01$, and 0.02 for the ED₉₀. When the asymmetric Aranda–Ordaz model was the true dose–response model, the values of λ used were $\lambda = 0.8, 0.9, 1, 1.1$, and 1.2 for both the ED₅₀ and the ED₉₀. The reason for using these values of γ or λ is to ensure that the data are generated from realistic dose–response curves.

The Genstat package^[21] was used to obtain estimated coverage probabilities (CP) corresponding to nominal 95% confidence intervals for each of the four types of intervals. For each of the experiments, 10,000 simulated data sets

Table 2 Sample Sizes (n_i), β , μ_v , and the Values of the Dose Levels (x_i) for the Six Sets of Experimental Configurations Investigated

1. Experimental configurations for the ED ₅₀				
Set	n_i	β	μ_{50}	Doses (x_i)
(a)	6	1, 2	3, 3.5, 4	1, 2, ..., 5
(b)	5	1, 2	3.5, 4, 4.5	1, 2, ..., 6
(c)	10, 20, 30, 50	7, 14	0.1, 0.2, 0.3	−0.3098, −0.2147, −0.1487, −0.0809, −0.0362, 0.0864, 0.1523, 0.2304, 0.2810
2. Experimental configurations for the ED ₉₀				
Set	n_i	β	μ_{90}	Doses (x_i)
(d)	30, 50, 70, 90	1, 2	7.1, 7.2, 7.3	2.056, 3.233, 4.411, 5.589, 6.767, 7.944
(e)	8, 15, 30, 60	1, 2	3.3, 3.5, 3.7	0.000, 0.539, 0.981, 1.186, 1.587, 1.792, 2.234, 2.773
(f)	8, 15, 30, 60	1, 2	4.2, 4.4, 4.6	0.000, 0.463, 3.045, 3.296, 3.584, 3.932, 4.394, 5.142

were generated. In the simulation study, our data are generated from the three true dose–response curves, but we (mistakenly) regard the data as having arisen from a logistic dose–response curve.

It is also useful to determine whether the extended model is actually a better fit to the data than the simpler logistic model. We could use a likelihood ratio test or score test to determine this. The score test is more convenient, as it only requires the logistic model to be fitted, whereas the likelihood ratio test requires fitting of the extended model as well, and this is somewhat more difficult. Thus the score test is used here. Data sets for which the score tests were significant at the 5% level of significance were excluded from the calculation of the coverage probabilities. This was performed to mimic actual data analysis in which one would expect the inappropriateness of the assumed model to be detected if the model fit was sufficiently bad. The results obtained for the cubic–logistic and asymmetric Aranda–Ordaz distributions are thus based on data sets for which a diagnostic test statistic has not found real evidence of the inadequacy of the assumed logistic model. The performance of these score tests is critical and may have a potential bearing on the final results. Here the effect of the extra parameter is taken into consideration, as it appears in the expression of μ_v for the true model and affects those data sets for which the score test is not rejected. It is this μ_v value for the extended model, not μ_v for the logistic model, for which the 95% confidence interval is appropriate.

In addition, excluded are intervals for data sets for which the null hypothesis $\beta = 0$ could not be rejected against the alternative hypothesis $\beta \neq 0$ at the 5% level of significance, using the Wald test.^[6] Also, to confine ourselves to cases in which the MLEs are calculable for all of the methods, data sets with either zero partial responses or one partial response were excluded from our analysis;^[6,22] in addition, the score tests for the extra parameter of the cubic logistic and asymmetric Aranda–Ordaz families cannot be determined if there are either zero partial responses or one partial response.

SIMULATION RESULTS

First, we describe the three models that were used to represent the true dose–response relationships in the simulation experiments outlined in “Numerical Experiments.”

The first distribution used to represent the true dose–response relationship was the normal distribution function

$$p(x_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\rho(x_i)} \exp(-z^2/2) dz \quad (10)$$

where $\rho(x_i) = (\sqrt{3}/\pi)(\alpha + \beta x_i)$ and our probit model has been parameterized to have the same variance as the logistic model^[3] for a more detailed discussion of this point.

The second distribution used was the cubic logistic distribution function.^[17]

$$p(x_i) = 1/\left\{1 + \exp\left[-\left(\beta(x_i - \mu_{50}) + \gamma(x_i - \mu_{50})^3\right)\right]\right\} \quad (11)$$

This is a symmetric distribution function that contains the logistic distribution function as a special case when $\gamma = 0$. Note that when $\gamma < 0$, $p(x)$ is not a monotonic function of x although it may be monotonic over the observed range. As a consequence, extreme doses either do not exist or are not unique if $\gamma < 0$. Therefore it is concluded that the model is most useful when $\gamma \geq 0$.^[17]

The third distribution used was the asymmetric Aranda–Ordaz distribution.^[18]

$$p(x_i) = \begin{cases} 1 - [1 + \lambda \exp(\alpha + \beta x_i)]^{-1/\lambda} & \lambda \exp(\alpha + \beta x_i) > -1 \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

This is an asymmetric distribution function that contains the logistic distribution function as a special case when $\lambda = 1$.

For the three true dose–response curves, we define a particular method of interval estimation as being “robust” if the CP is within the range $\omega = \{X: |X| \leq 0.5\}$. Here $X = \text{CP} - 95$ and CP is the estimated percentage coverage probability corresponding to a nominal 95% confidence interval, obtained via the fitting of a logistic model—whether or not this is the appropriate model. Values of X close to zero indicate that a method of interval estimation is achieving an observed level of coverage close to its nominal level.

From this definition, we consider that a method of interval estimation is liberal when $X < -0.5$ and conservative when $X > 0.5$, respectively. Here the range of $|X| \leq 0.5$ was chosen because of it being approximately equal to the range of values for which an approximate 95% confidence interval (for the CP) would cover 95%.

Robustness of Interval Estimators of the ED₅₀

The Probit Model

For the experimental configurations in data sets (a) and (b), the simulation results shown in Fig. 2 indicate that the Fieller method of interval estimation is always conservative. There are a number of liberal results that occur for the LR, BLR, and score test methods. The Fieller method displays no robust results; the LR method has only one robust result; and the BLR and score test methods each have four robust results from the 12 obtained.

The simulation results shown in Fig. 2, based on data set (c), indicate that the LR, BLR, and score test methods tend to be liberal. Fieller’s method is still generally conservative

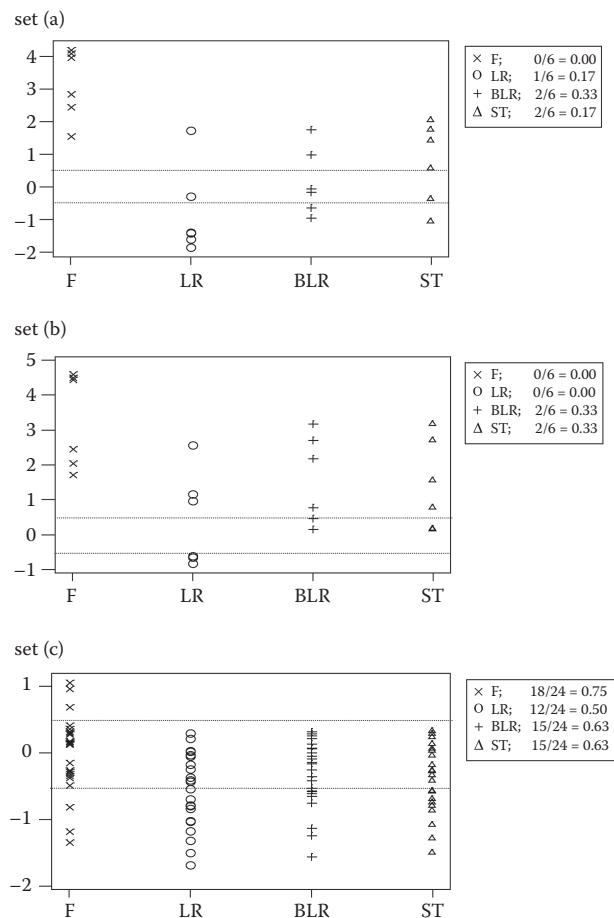


Fig. 2 A robustness comparison for the four methods of interval estimation of the ED₅₀, when the true dose–response is the probit model. Two dotted lines denote the left and right endpoints of the range $\omega = (-0.5, 0.5)$, respectively. The fractions in the key denote the proportion of points in ω

for small sample sizes, but as n_i becomes larger, it becomes more liberal. The performances of all four methods are similar. They are robust in more than 50% of the cases considered.

The Cubic Logistic Model

For the simulation results shown in Fig. 3 based on data sets (a) and (b), it can be seen, by noting the proportion of points inside ω , that the BLR and score test methods are more robust than the Fieller and LR methods, which have only 5 robust results out of the 60 considered for the LR method. In addition, the Fieller and score test methods are almost always conservative, while the LR and BLR methods are sometimes conservative and sometimes liberal.

The simulation results shown in Fig. 3 based on data set (c), indicate that 82, 96, 107, and 103 out of 120 results are within ω for the Fieller, LR, BLR, and score test methods, respectively. Thus all four methods of interval estimation are, to some degree, robust.

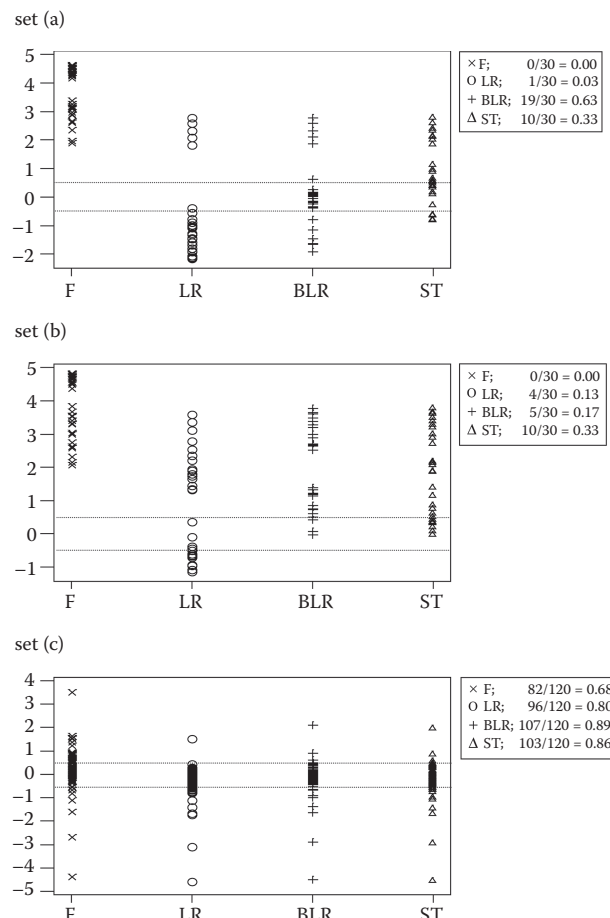


Fig. 3 A robustness comparison for the four methods of interval estimation of the ED₅₀, when the true dose–response is the cubic logistic model. Two dotted lines denote the left and right endpoints of the range $\omega = (-0.5, 0.5)$, respectively. The fractions in the key denote the proportion of points in ω

The Asymmetric Aranda–Ordaz Model

For the simulation results shown in Fig. 4 based on data sets (a) and (b), Fieller’s method tends to be conservative. The BLR and score test methods can also be conservative. The LR method can be quite liberal at times, when the other three methods are not. Overall, none of the four methods can really be classified as being robust.

For the simulation results shown in Fig. 4 based on data set (c), all four methods of interval estimation appear to show a good degree of robustness overall. It follows from the proportion of points inside ω that the performance of all four methods is reasonably comparable.

Summary

When the true dose–response curve is the probit model, none of the methods of interval estimation are robust for the configurations given by data sets (a) and (b). All four methods display a similar performance for the data set (c)

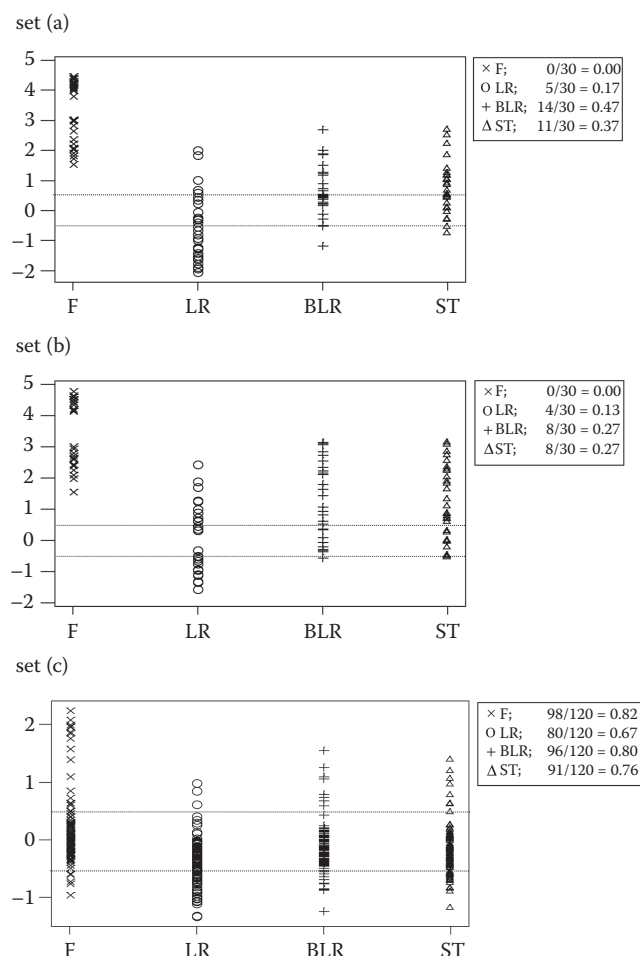


Fig. 4 A robustness comparison for the four methods of interval estimation of the ED₅₀, when the true dose-response is the asymmetric Aranda-Ordaz model. Two dotted lines denote the left and right endpoints of the range $\omega = (-0.5, 0.5)$, respectively. The fractions in the key denote the proportion of points in ω

configurations, and they are robust in more than one-half of the configurations considered.

For a true cubic logistic dose-response relationship, the BLR and score test methods perform best. They are at their most robust for the data set (c) configurations, where the other two methods also perform satisfactorily.

For departures from a logistic model given by the asymmetric Aranda-Ordaz model, the absolute performance of all four methods is very poor, specifically for the Fieller and LR methods, for data sets (a) and (b); on the other hand, the absolute performance of all four methods is very good for data set (c).

Robustness of Interval Estimators of the ED₉₀

The Probit Model

For the simulation results shown in Fig. 5, the performance of the LR, BLR, and score test methods is similar and

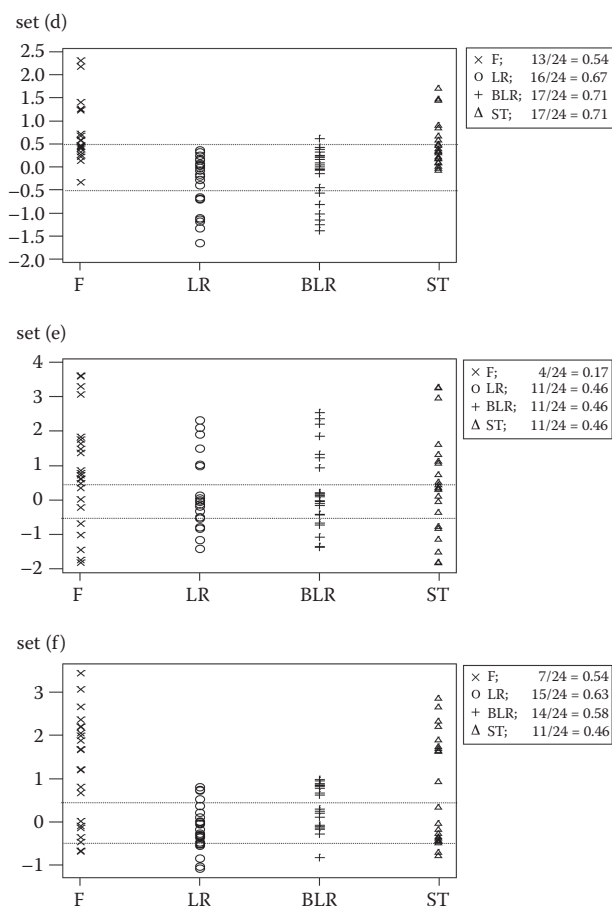


Fig. 5 A robustness comparison for the four methods of interval estimation of the ED₉₀, when the true dose-response is the probit model. Two dotted lines denote the left and right endpoints of the range $\omega = (-0.5, 0.5)$, respectively. The fractions in the key denote the proportion of points in ω

somewhat better than that of the Fieller method. However, none of the methods is robust in more than three quarters of the cases considered.

The Cubic Logistic Model

The results displayed in Fig. 6 show that in all three sets of experimental configurations, the performance of Fieller's method is poorer than those of the other three methods, which are similar in their performance, and are robust in at least two-thirds of the cases considered.

The Asymmetric Aranda-Ordaz Model

The simulation results shown in Fig. 7 suggest that the absolute performance of the Fieller method is always slightly inferior to those of the LR, BLR, and score test methods, which display a similar performance and achieve a result in ω in at least 50% of occasions.

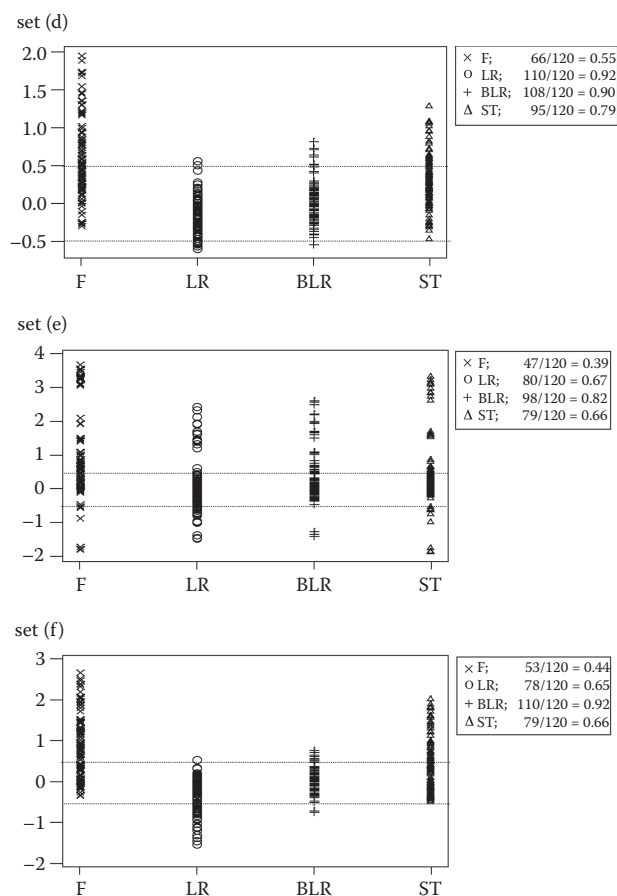


Fig. 6 A robustness comparison for the four methods of interval estimation of the ED₉₀, when the true dose–response is the cubic logistic model. Two dotted lines denote the left and right endpoints of the range $\omega = (-0.5, 0.5)$, respectively. The fractions in the key denote the proportion of points in ω

Summary

When the true dose–response models are either probit, cubic logistic, or asymmetric Aranda–Ordaz, the overall performances of the LR, BLR, and score test methods are similar, and noticeably better than that of Fieller’s method. The LR, BLR, and score test methods achieved a reasonable level of robustness.

CONCLUSION

Overall, using the definition of robustness given in “Simulation Results,” the BLR and score test methods are the most robust of the four methods of interval estimation of the ED₅₀ for the range of true models considered here. However, even their performance is poor in a number of the configurations studied. This is usually true when the sample size is much smaller, such as in data sets (a) and (b).

The results obtained under departures from the logistic model given by the two extended models are encouraging. Overall, all four methods of interval estimation of the ED₅₀

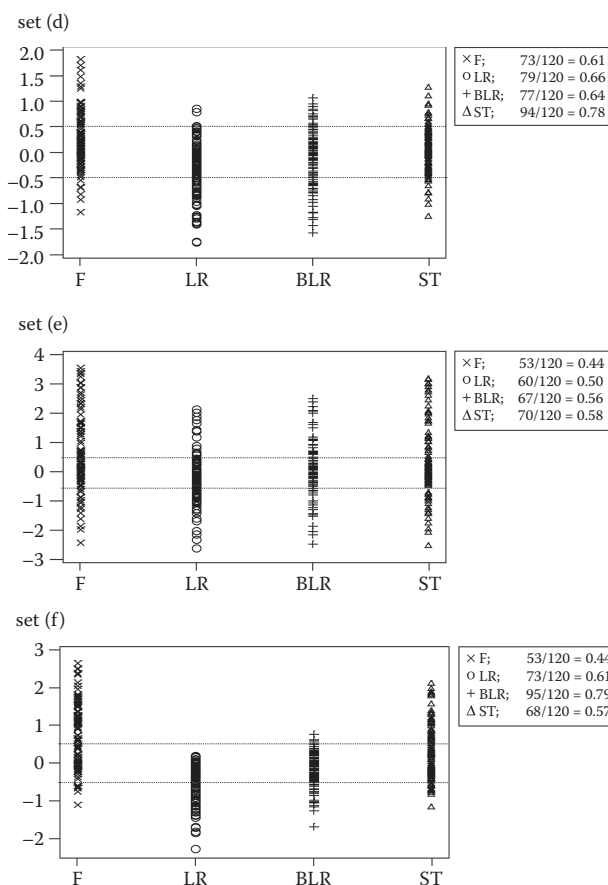


Fig. 7 A robustness comparison for the four methods of interval estimation of the ED₉₀, when the true dose–response is the asymmetric Aranda–Ordaz model. Two dotted lines denote the left and right endpoints of the range $\omega = (-0.5, 0.5)$, respectively. The fractions in the key denote the proportion of points in ω

are robust to some degree. However, as might be expected, there are differences according to the degree of departure from the logistic model. When the true dose–response relationship is cubic logistic, combining the results for all four methods for data set (c), when $\gamma = 1$, 88% of the results are robust; this figure drops to 65% for $\gamma = 10$. This is possibly not surprising, as $\gamma = 10$ appears to be a sizable departure in absolute terms. The asymmetric Aranda–Ordaz model provides a more interesting scenario. Departures from the logistic model, given by values of λ greater than 1, appear to give rise to fewer robust results than departures by an equivalent amount given by values of λ less than 1. For example, when $\lambda = 1.2$, 64% of results are robust, compared with 81% when $\lambda = 0.8$. Thus departures above $\lambda = 1$ may not have to be that great to render the methods nonrobust.

On the whole, for the ED₉₀, Fieller’s method did not perform usually as well as the LR, BLR, and score test methods, which achieve a reasonable level of robustness.

From the results presented here, one might say that the BLR and score test methods could be viewed as safer methods to use for interval estimation of effective doses in the presence of model misspecification.

ACKNOWLEDGMENTS

The author is grateful to Drs. Mark Hann, Peter Harris, and Simon Kirby for their thorough reading of and valuable comments on an earlier version of this paper.

REFERENCES

1. Finney, D.J. *Probit Analysis*, 3rd Ed.; Cambridge University Press: Cambridge, 1971.
2. Finney, D.J. *Statistical Methods in Biological Assay*; Griffin: London, 1978.
3. Collect, D. *Modelling Binary Data*; Chapman and Hall: London, 1991, 94.
4. Morgan, B.T.J. *Analysis of Quantal Response Data*; Chapman and Hall: London, 1992.
5. Abdelbasit, K.M.; Plackett, R.L. Experimental design for binary data. *J. Am. Stat. Assoc.* **1983**, *78*, 90–98.
6. Sitter, R.R.; Wu, C.F.J. On the accuracy of Fieller intervals for binary response data. *J. Am. Stat. Assoc.* **1993**, *88*, 1021–1025.
7. Williams, D.A. Interval estimation of the median lethal dose. *Biometrics* **1986**, *42*, 641–645.
8. Hoekstra, J.A. Estimation of the LC50, a review. *Environ-Metrics* **1991**, *2*, 139–152.
9. Alho, J.M.; Valtonen, E. Interval estimation of inverse dose–response. *Biometrics* **1995**, *51*, 491–501.
10. Harris, P.; Hann, M.; Kirby, S.P.J.; Dearden, J.C. Interval estimation of the median effective dose for a logistic dose–response curve. *J. Appl. Stat.* **1999**, *26*, 715–722.
11. Huang, Y. Interval estimation of the ED₅₀ when a logistic dose–response curve is incorrectly assumed. *Comput. Stat. Data Anal.* **2001**, *36* (4), 525–537.
12. Huang, Y. Various methods of interval estimation of the median effective dose. *Commun. Stat., Simul. Comput.* **2001**, *30*, 99–113.
13. Huang, Y.; Harris, P.; Kirby, S.P.J.; Dearden, J.C. Improved approaches to interval estimation of effective doses for a logistic dose–response curve. *J. Stat. Comput. Simul.* **2002**, *72*, 565–584.
14. Faraggi, D.; Izikson, P.; Reiser, B. Confidence intervals for the 50 percent response dose. *Stat. Med.* **2003**, *22*, 1977–1988.
15. Engeman, R.M.; Otis, D.L.; Dusenberry, W.E. Small sample properties of some parametric bioassay estimators of the LD90. *Comput. Biomed. Res.* **1986**, *19*, 588–595.
16. Robertson, J.L.; Smith, K.C.; Savin, N.E.; Lavigne, R.J. Effects of dose selection and sample size on the precision of lethal dose estimates in dose–mortality regression. *J. Econ. Entomol.* **1984**, *77*, 833–837.
17. Morgan, B.J.T. The cubic logistic model for quantal assay data. *Appl. Stat.* **1985**, *34*, 105–113.
18. Aranda-Ordaz, F.J. On two families of transformations to additivity for binary response data. *Biometrika* **1981**, *68*, 357–367.
19. Cordeiro, G.M. Improved likelihood ratio statistics for generalized linear models. *J. R. Stat. Soc.* **1983**, *45*, 404–413.
20. Alho, J.M. On the computation of likelihood ratio and score test based confidence intervals in generalized linear models. *Stat. Med.* **1992**, *11*, 923–930.
21. Genstat 5 Committee. *Genstat 5 Release 3 Reference Manual*; Clarendon Press: Oxford, 1993.
22. Hamilton, M.A. Robust estimates of the ED₅₀. *J. Am. Stat. Assoc.* **1979**, *74*, 344–354.

Enrichment Design

Jen-pei Liu

Division of Biometry, Department of Agronomy, and Institute of Epidemiology, National Taiwan University, Taipei, Taiwan; and Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan

Abstract

The main objective of an enrichment design is to screen for possible responders using the active treatments under investigation. This entry serves as a brief review of the phases of an enrichment design and the process by which they are conducted.

Keywords: Enrichment; Holter monitoring; Targeted clinical trials; Screening.

INTRODUCTION

In clinical trials, the target population is defined by the inclusion and exclusion criteria. As a result, a screening period is usually conducted for the purpose of identifying the eligible subjects before actually enrolling them into the study. On the other hand, a placebo run-in period is also usually performed after the screening period and before the double-blind, randomized, active treatment period to assess the patient's compliance and to estimate the potential placebo effect. Both the screening and placebo run-in periods can be viewed as sequential screening processes for the identification of the patients who meet the inclusion/exclusion and compliance criteria without giving the active treatment under investigation to the patients.

The objective of some trials is to evaluate the pharmaceuticals or treatments based on hard efficacy endpoints such as death, myocardial infarction, or bone fracture. However, these hard endpoints require a long follow-up period to observe. On the other hand, some short-term responses can be obtained much more quickly for some soft endpoints such as biological markers. In addition, some therapeutic agents are likely to be effective in a population of the patients who may have an underlying disorder that is responsive to the manipulation of dose levels of the agents or several agents. Therefore it is necessary to perform some additional screening processes using the active treatments under evaluation after either the screening period or the placebo run-in period to identify the candidates in whom test agents are likely beneficial in the early phase of the trial. This phase of an additional screening process of the same therapeutic agent or test of different agents for the identification of patients with drug efficacy is called the enrichment phase. The patients with drug efficacy identified in the enrichment phase are then randomized to receive either the efficacious dose (agent) or the matched

placebo or the active controls. This type of the design with additional screening processes with the active treatments evaluated in the study is called the enrichment design.^[1] According to this definition, the difference between the designs, with either screening or placebo run-in periods, and the enrichment designs is that the active treatments are used to identify the patients.

EXAMPLES

An enrichment design usually consists of at least two phases. The first phase is the enrichment phase, in which a titration design is usually conducted to classify patients into groups according to whether the pharmaceuticals under investigation present some benefits to the patients. The second usually is a randomized double-blind phase, possibly with a placebo concurrent control to investigate the effectiveness and the safety of the test agents in these patients formally and rigorously. The primary efficacy endpoints used for the evaluation during the two phases of an enrichment design, depending upon the objectives of the trial, can be the same or different. The concept of enrichment design is illustrated in three clinical trials in the areas of Alzheimer's disease and arrhythmia that have been described in detail in Chow and Liu.^[2]

The enrichment design was elected for a clinical trial conducted in the early development stage of tacrine, with doses of 40 and 80 mg four times a day in the treatment of patients with probable Alzheimer's disease. As indicated in Davis et al.,^[3] because of the clinical, biochemical, and pathological heterogeneity of the disease and clinical experience, not all patients with probable Alzheimer's disease would respond to any single treatment, and those who did respond might do so only within a limited dose range. Consequently, an enrichment design was elected to encounter

the above situations. This trial consisted of a total of four phases: a six-week double-blind dose-titration enrichment phase, a two-week placebo baseline phase, a six-week randomized, double-blind, placebo-controlled phase, and a six-week sustained active phase. After the patients met the inclusion and exclusion criteria, they were enrolled into the enrichment phase of the trial that consisted of three titration sequences. Each titration sequence consists of three two-week dosing periods. The dose in each titration sequence is always titrated up from 40–80 mg four times a day with a placebo in dosing periods 1, 2, and 3 for titration sequences 1, 2, and 3, respectively. The patients were randomized into one of the three titration sequences that were conducted in a double-blind fashion. The potentially therapeutic responses for each patient at each dose were then assessed at the end of each two-week dosing period. The “best dose” response for a patient was defined in advance in protocol as a reduction of at least four points from the screening value in a total score on the Alzheimer’s Disease Assessment Scale (ADAS)^[4] and without intolerable side effects. Patients with the identified “best dose” were then entered into a two-week placebo baseline period with the hope that this period would be sufficiently long for tacrine to wash out from the body and for patients to return to the screening pretreatment status with comparable efficacy outcomes. At the end of the two-week placebo baseline phase, the patients with a reduction of at least four points in ADAS during the enrichment phase entered the subsequent 6-week randomized, double-blind, parallel-group, placebo-controlled phase. They were randomized either to the active tacrine at their best dose or to the matched placebo. Patients who completed the six-week double-blind phase then entered into the sustained active treatment phase. As a result, the purpose of the selection of an enrichment design with three titration sequences for this study is to identify a group of patients who are likely to respond to tacrine at a certain dose; after a wash-out period of two weeks, these identified patients were randomized to either tacrine at their best dose or to placebo concurrent control in a double-blind phase.

The rationale for the selection of the enrichment design in one of the trials during the development of tacrine is to verify whether a short-term response to tacrine has predictability for the long-term efficacy in the prevention of progression of the patients with probable Alzheimer’s disease. The same primary efficacy endpoints such as ADAS or the Clinical Global Impression of Change were used in both the enrichment and double-blind phases for the evaluation of tacrine’s effectiveness. On the other hand, for other therapeutic agents, the real efficacy endpoint is mortality, which requires a long time to observe. Therefore the short-term efficacy of the agents is assessed by some other objective surrogate endpoints. It is then very important to know whether the short-term efficacy based on surrogate endpoint is predictive of the hard endpoint such as mortality. As a result, the enrichment design is

usually employed for the identification of the “short-term” responders at the initial stage followed by the main phase of the long-term study. Examples of this type of the trials can be found in the area of arrhythmia such as the Cardiac Arrhythmia Suppression Trial (CAST)^[5–7] and the Electrophysiologic Study versus Electrocardiographic Monitoring (ESVEM).^[8–12]

Cardiac Arrhythmia Suppression Trial is a multicenter, randomized, placebo-controlled study used to test the hypothesis whether the suppression of asymptomatic or mildly symptomatic ventricular arrhythmia after myocardial infarction would reduce the rate of death from arrhythmia. The drugs include two class-IC antiarrhythmic agents, encainide and flecainide, and moricizine with a placebo concurrent control. One of the objectives of the study is to test the predictability of adequate suppression of ventricular arrhythmia by the active drugs based on the ventricular premature contractions (VPC) as recorded by Holter monitor for mortality. As a result, an open-label enrichment design with two titration sequences involved with only active drugs was selected for this study. Patients with ejection fraction of at least 30% were randomly assigned to the two titration sequences: (encainide, moricizine, flecainide) or (flecainide, moricizine, encainide). The reason for the insertion of moricizine in the middle dosing period is its inferior efficacy in the suppression of VPC as compared with the other two active agents. Each drug is tested at two dose levels. The doses were 35 and 50 mg three times a day (tid) for encainide; 100 and 150 mg twice a day for flecainide; and 200 and 250 mg tid for moricizine. Because flecainide exhibits negative inotropic properties, it was not administered to the patients with an ejection fraction of less than 30%. The titration sequences for the patients with an ejection fraction of less than 30% were (encainide, moricizine) or (moricizine, encainide). The pre-specified criteria for an adequate suppression of ventricular arrhythmia were (i) a reduction of at least 80% in VPC and (ii) a reduction of at least 90% runs of unsustained ventricular tachycardia as measured by the 24-hour Holter recording four to ten days after each dose was begun. The titration process for a particular patient was stopped as soon as a drug and a dose were found to yield an adequate suppression. The patients whose arrhythmia was adequately suppressed were then randomized to either the “best drug” identified during the enrichment phase or to placebo for a three-year long-term follow-up.

The results of CAST indicate that the short-term efficacy measured as suppression of VPC based on Holter noninvasive ambulatory electrocardiographic monitoring might not be a good predictor for the long-term hard mortality endpoint. Others suggested that the failure to induce ventricular tachycardia or fibrillation by some drug assessed by the invasive electrophysiologic study might be a good independent predictor of the recurrence of arrhythmia. Consequently, the ESVEM trial was the first large

prospective, randomized trial conducted to compare the two methods in the predictability of long-term recurrence of arrhythmia by the short-term efficacy assessed by the two methods.^[8,10,11] Because it requires correlating the difference in recurrence rates of arrhythmia with the short-term efficacy by both methods, an inpatient enrichment phase was elected to identify a group of patients in whom a test drug exhibited a short-term efficacy assessed by either one of the two methods. After the patients had fulfilled entry criteria and 48-hour Holter monitoring and electrophysiologic study criteria, they were randomized to one of the two parallel groups for the two methods to assess the short-term drug efficacy. The first group employed non-invasive ambulatory electrocardiographic monitoring; the second group applied the invasive electrophysiologic study. Within each group, the patients received up to six arrhythmia agents in a random order until one drug was predicted to be efficacious or until all drugs have been tried that the patients were eligible to receive. A test drug was identified as effective as assessed by electrocardiographic monitoring during the inpatient enrichment phase if the following predefined criteria for short-term efficacy were met: (i) 70% reduction in mean VPC; (ii) 80% reduction in pairs of VPC; (iii) 90% reduction in mean ventricular tachycardia count; and (iv) absence of any runs of ventricular tachycardia longer than 15 seconds. Drug effectiveness evaluated by the electrophysiologic study during the enrichment phase is defined as the failure by the drug to induce a run of ventricular tachycardia longer than 15 seconds with V1V2V3 stimulation at the right ventricular apex. If a drug was proven to be efficacious for a patient during the enrichment phase, then that patient was discharged from the hospital for the long-term follow-up with the drug, and the accuracy of the prediction of efficacy is determined during the long-term follow-up. Patients in whom no drugs were proven to be effective during the enrichment phase were not randomized and were withdrawn from the study, but their vital signs and recurrence of arrhythmia were monitored. Enrichment design can also be used for the studies in non-responders or in the patients intolerant of another agent. Temple^[1] provided examples in the studies for severe hypertension and psychotropic agents.

The disease targets at the molecular level can be identified after completion of the Human Genome Project (HCP). Hence, treatment modalities specific for the patients with the identified molecular targets have been evaluated in the targeted clinical trials. Because only the patients tested positive for the molecular targets are enrolled and randomized into the targeted clinical trials, enrichment design is one of the most commonly employed design for the targeted clinical trials, examples are ALTTO trials,^[13] TAILORX trials,^[14] and MINDACT trial.^[15] Simon and Maitournam^[16] and Maitournam and Simon^[17] have discussed requirement of the sample sizes between the targeted clinical trials versus untargeted trials. In addition,

Liu and Lin^[18] and Liu et al.^[19] have suggested the use of EM algorithm for statistical inference of treatment effects for targeted clinical trials under enrichment.

CONCLUSION

The main objective of an enrichment design is to screen for possible responders using the active treatments under investigation. However, if a placebo is also included in the enrichment phase, then the possible placebo responders will also be identified in the enrichment phase. In addition, as demonstrated in the previous discussion, the design employed in the enrichment phase is usually a dose titration design with no wash-out periods between dosing interval. As a result, the treatment effects, carryover effects, and time effects all confound one another during the process of identification of responders. In CAST or ESVEM, the patients were randomized as soon as the first drug at the first dose was found to yield an adequate suppression of VPC. Another issue for the enrichment design therefore is whether or not the response observed is the best drug at its optimal dose for the responder. On the other hand, because of either a lack of randomization or different methods of randomization for the enrichment phase, statistical methods for analysis based on the data from both enrichment phases and double-blind phase of the trial are not fully developed. In summary, enrichment design further restricts the target population into a very small selective group. However, sometimes, it is not yet possible to distinguish this small group of patients from the rest with the same ailment in terms of demographic and other prognostic factors for grassroots clinical practice. Therefore, inference based on the statistical analysis for a trial using an enrichment design remains a challenge for all of us.

REFERENCES

1. Temple, R.J. Special study design, early escape, enrichment, studies in non-responders. *Commun. Stat. Theory Methods* **1994**, 23 (2), 499–531.
2. Chow, S.C.; Liu, J.P. *Design and Analysis of Clinical Trials: Concepts and Methodologies*; Wiley and Sons: New York, 1998.
3. Davis, K.L.; Thal, L.J.; Gamzu, E.R.; Davis, C.S.; Woolson, R.F.; Gracon, S.I.; Drachman, D.A.; Schneider, L.S.; Whitehouse, P.J.; Hoover, T.M.; Morris, J.C.; Kawas, G.H.; Knopman, D.S.; Earl, N.L.; Jumar, V.; Doody, R.S.; the Tacrine Collaborative Study Group. A double-blind, placebo-controlled multicenter study for Alzheimer's disease. *N. Engl. J. Med.* **1992**, 327, 1253–1259.
4. Folstein, M.F.; Folstein, S.E.; McHugh, P.R. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Rev.* **1975**, 12, 189–198.

5. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N. Engl. J. Med.* **1989**, *321*, 406–412.
6. Echt, D.S.; Liebson, P.R.; Mitchell, L.B.; Peters, R.W.; Obias-Manno, D.; Barker, A.H.; Arensberg, D.; Baker, A.; Friedman, L.; Greene, H.L.; Huther, M.L.; Richardson, D.W.; the CAST Investigators. Mortality and morbidity in patients receiving encainide and flecainide or placebo. *N. Engl. J. Med.* **1991**, *324*, 781–788.
7. Ruskin, J.N. The Cardiac Arrhythmia Suppression Trial (CAST). *N. Engl. J. Med.* **1989**, *321*, 386–388.
8. The ESVEM Investigators. The ESVEM trial: Electrophysiologic study versus electrocardiographic monitoring for selection of antiarrhythmic therapy of ventricular tachyarrhythmias. *Circulation* **1989**, *79*, 1354–1360.
9. The ESVEM Investigators. Determinants of predicted of efficacy antiarrhythmic drugs in the Electrophysiologic study versus electrocardiographic monitoring trials. *Circulation* **1993**, *87*, 323–329.
10. Mason, J.W.; for the ESVEM Investigators. A comparison of seven antiarrhythmic drugs in patients with ventricular tachyarrhythmias. *N. Engl. J. Med.* **1993**, *329*, 452–458.
11. Mason, J.W.; for the ESVEM Investigators. A comparison of Electrophysiologic testing versus Holter monitoring to predict antiarrhythmic-drug efficacy for ventricular tachyarrhythmias. *N. Engl. J. Med.* **1993**, *329*, 445–451.
12. Ward, D.E.; Camm, A.J. Dangerous ventricular arrhythmias can we predict drug efficacy. *N. Engl. J. Med.* **1993**, *329*, 498–499.
13. The ALTTO trial. <http://www.cancer.gov/> Accessed on June 1, 2009.
14. Sparano, J.; Hayes, D.; Dees, E.; et al. Phase III randomized study of adjuvant combination chemotherapy and hormonal therapy versus adjuvant hormonal therapy alone in women with previously resected axillary node-negative breast cancer with various levels of risk for recurrence (TAILORX Trial). <http://www.cancer.gov/clinicaltrials/ECOG-PACCT-1> (Accessed on June 1, 2009).
15. MINDACT Design and MINDACT trial overview. <http://www.breastinternationalgroup.org/transbig.html> (Accessed on June 20, 2009).
16. Simon, R.; Maitournam, A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin. Cancer Res.* **2004**, *10*, 6759–6763.
17. Maitournam, A.; Simon, R. On the efficiency of targeted clinical trials. *Stat. Med.* **2005**, *24*, 329–339.
18. Liu, J.P.; Lin, J.R. Statistical methods for targeted clinical trials under enrichment design. *J. Formos. Med. Assoc.* **2008**, *107* (12), S34–S41.
19. Liu, J.P.; Lin, J.R.; Chow, S.C. Inference on treatment effects for targeted clinical trials under enrichment design, published on line in *Pharmaceutical Statistics*, 2009. DOI:10.1002/pst.364.

Enrichment Design with Patient Population Augmentation

Yijie Zhou

Data and Statistical Science, AbbVie Inc., North Chicago, Illinois, U.S.A.

Bo Yang

Biometrics, Vertex Pharmaceuticals, Boston, Massachusetts, U.S.A.

Lanju Zhang and Lu Cui

Data and Statistical Science, AbbVie Inc., North Chicago, Illinois, U.S.A.

Abstract

Clinical trials can be enriched on subpopulations that may be more responsive to treatments to improve the chance of trial success. In 2012, Food and Drug Administration issued a draft guidance to facilitate enrichment design, where it pointed out the uncertainty on the subpopulation classification and on the treatment effect outside of the identified subpopulation. Related statistical literature is reviewed, and the developed enrichment design with patient population augmentation is described in this entry. In this newly developed design strategy, the identified subpopulation (biomarker positive) is augmented by some biomarker-negative patients, so that there is sufficient power to assess both the treatment effect in the biomarker-positive subpopulation and the overall treatment benefit. A weighted statistic is used to assess the overall treatment effect for this assessment, correcting for the disproportionality of biomarker-positive and biomarker-negative subpopulations under enriched trial setting. Screening information is utilized for weight determination.

Keywords: Biomarker; Enrichment design; Overall treatment effect; Screening; Two-step design; Weighted estimator.

INTRODUCTION

Enrichment designs have been widely discussed to establish treatment benefit in a selected (enriched) subpopulation. In this entry, literature and common practice of enrichment design is reviewed, and a newly proposed enrichment design with patient population augmentation is described.

BACKGROUND

In the past decades, enrichment designs have emerged and gained popularity in clinical development and subsequently in statistical research for trial design.^[1–7] They are closely related to the Food and Drug Administration (FDA)’s initiative on personalized medicine.^[8] One purpose to select such enriched population is to increase trial sensitivity of treatment response, e.g., the *Trastuzumab* benefit on HER2+ breast cancer patients.^[9–12] This type of enrichment is typically referred to as *predictive enrichment* and will be the focus of this entry. Specifically, predictive enrichment selects patients who are more likely to respond to a treatment of interest than other patients, so as to generate a larger treatment effect size and to permit use of a smaller study size. Typically, the predictive enrichment

selection maneuvers are well rationalized without controversy, and trial results are interpreted accordingly for the selected subpopulation. This is in contrary to the conventional “all-comer design” which refers to a clinical trial that enrolls general disease population, often referred as all-comer population, and makes inference on the (overall) treatment effect for the general disease population.

Several considerations arise and subsequently statistical methodology are proposed during the evolved thinking of predictive enrichment clinical trials, and the considerations can be divided into trial design perspective and data analysis perspective.

Design Aspect

When a predictive biomarker is adequately identified, a typical enrichment design enrolls only biomarker-positive patients. It is considered to be more efficient than an all-comer design as illustrated by Simon and Maitourman.^[2,3] The terminology “biomarker-positive” and “biomarker-negative” are used to denote the enriched subpopulation and its complementary (i.e., patients not satisfying the enrichment criteria), respectively. When such a biomarker is a post-treatment pharmacodynamics marker, a run-in design^[13,14] can be used to classify all-comers into biomarker-positive or biomarker-negative patients in the

first stage of the trial and patients are randomized in the second stage of the trial. In typical biomarker-based designs with or without a run-in stage, depending on the strength of prior evidence either only biomarker-positive patients are randomized or all patients are randomized stratified by the biomarker status (i.e., stratified all-comer design). A comprehensive review of biomarker-based design and testing strategies is provided by Freidlin and Korn.^[15]

In 2012, FDA issued a draft guidance on enrichment design.^[16] The guidance suggested inclusion of biomarker-negative patients to assess “the *uncertainty* about the marker cutoff and responsiveness of marker-negative patients.” Such uncertainty is not unreasonable considering multiple disease pathways, less-than-perfect accuracy of biomarker assays, and the uncertainty around the cutoff values to define biomarker-positive versus biomarker-negative patients, etc. One example is again the development of HER2-targeting agents and the *Trastuzumab* treatment for adjuvant breast cancer. In a retrospective analysis of a pivotal *Trastuzumab* adjuvant trial, 174 cases found actually lacked HER2 gene amplification, and they still benefited as much as HER2-positive patients.^[17] These findings were confirmed in a similar analysis from an independent *Trastuzumab* adjuvant study.^[18] Biological evidence suggested that efficacy of HER2-targeting agents may be through breast cancer stem cell, and the regulation of breast cancer stem cell by HER2 may extend beyond HER2 gene amplification.^[19]

In the presence of such uncertainty, although typical enrichment design (which only studies biomarker-positive patients) can achieve better trial efficiency in terms of a smaller sample size, it comes at the expense of lacking the ability to make inferences for an all-comer population. As the FDA guidance has explicitly requested to include biomarker-negative patients in enrichment design, or more importantly when there could be potential treatment effect in the biomarker-negative subpopulation and therefore the overall treatment effect in an all-comer population is also of interest, an enrichment design that randomizes biomarker-positive patients only may not be sufficient. On the other hand, if an all-comer population is randomized using the stratified all-comer design, the efficiency associated with the enrichment trial design feature is then lost. It would become a challenge, particularly when the biomarker-positive subpopulation is relatively small, that a large number of biomarker-negative patients have to be enrolled into such a trial in order to be able to recruit sufficient number of biomarker-positive patients. As an illustrative example, consider a 2-arm design where 468 patients are needed to provide 90% power to detect a treatment difference of 1.5 with σ of 5.0 at two-sided α of 0.05. Under a typical enrichment design to study biomarker-positive patients only, the study size would be 468. Under a stratified all-comer design, the study size would be 2,340 where biomarker-positive patients only account for 20% of the general disease population, which becomes a very large trial to enroll.

As an alternative, adaptive enrichment has been proposed in literature where biomarker-negative patients are also assessed as part of the trial design. Wang, Hung, and O'Neill^[20] proposed to start with an all-comer trial and use interim data to decide whether to continue with the all-comer trial or to enrich on one of the prespecified candidate biomarker-positive subpopulations. This design has the advantage of flexibility but suffers from the analysis and operational complexity of all adaptive trials. More importantly, depending on the strength of the interim data for enrichment decision-making, such designs could have the risk of making incorrect interim decision, resulting in wasting or losing a subpopulation. This is particularly a concern when randomization is not stratified for all candidate biomarker-positive subpopulations so bias can exist in the evaluation, or in situations where the candidate biomarker-positive subpopulations are small and therefore their interim data are limited and lack of power. A different type of adaptive enrichment was proposed by Simon and Simon,^[21] where the enrollment of each individual patient is determined via a probability calculation for potential treatment benefit after certain initial assessment period. Aside from the common adaptive design-associated features, Simon and Simon's approach lacks a systematic enrichment biomarker and therefore potentially lacks a clinically meaningful rationale or interpretation for such enrichment.

Analysis Aspect

In the enrichment design methods mentioned above, when the analysis (or final analysis in adaptive trials) is conducted on the biomarker-positive subpopulation only, standard analysis techniques can be directly applied. However, when a trial is enriched in a biomarker-positive group with some biomarker-negative patients included, statistical analysis needs to account for this design feature. There are two issues to be addressed—multiplicity adjustment for testing on multiple populations (biomarker-positive subpopulation and all-comer population) and inference of overall treatment effect for all-comer population. Commonly used approaches to handle multiplicity adjustment include sequential testing and splitting family-wise type I error rate.^[14,15] Several of the above mentioned design philosophies start the treatment assessment with all-comer population and proceed to biomarker-positive enrichment if all-comer population fails or is not likely to succeed. Therefore, assessment in the promising biomarker-positive subpopulation can be at risk because it is not the first step of statistical testing. Splitting type I error rate can mitigate the risk of mis-specifying the testing sequence as always, and the challenge shifts to how the type I error rate should be split, e.g., 0.01/0.04 or 0.04/0.01. In addition, it comes with the expense that the known-to-be-promising biomarker-positive subpopulation is not tested on the full type I error rate of 0.05, paying a multiplicity penalty for the less promising all-comer population. The multiplicity

penalty in the adaptive enrichment is similar but in a more complex manner.^[20,21]

How the overall treatment effect is estimated when the proportion of biomarker-positive patients in an enrichment design is typically larger (enriched) than that in an all-comer trial composition is another issue. Naive analysis via a simple pooling of biomarker-positive and biomarker-negative patients which did not take into account this disproportionality has been seen. One example is the TIMI (Thrombolysis in Myocardial Infarction) 38 study^[22] where two subpopulations were studied—it randomized 10,074 acute coronary syndromes (ACS) patients with non-ST elevation myocardial infarction (NSTEMI) which was the enriched subpopulation and 3534 ACS patients with STEMI. When treatment effect was established in the enriched NSTEMI subpopulation, a simple pooling of the two subpopulations was conducted by the sponsor to evaluate the overall treatment effect as if the trial was conducted in an all-comer trial composition. This analysis was improper because the proportion of the NSTEMI patients was inflated in the analysis. Therefore, new statistical techniques are needed in order to appropriately estimate the overall treatment effect under the enrichment trial setting.

In a newly proposed enrichment design with patient population augmentation,^[23] the biomarker-positive subpopulation is the primary focus, but this subpopulation is augmented with appropriate number of biomarker-negative patients. A weighted estimator is used to obtain proper statistical inference for the overall treatment effect in an all-comer population. If prior evidence suggests treatment benefit in biomarker-positive patients, a study such that the chance of trial success in biomarker-positive subpopulation is maximized, followed by the establishment of treatment benefit in an all-comer population with the addition of biomarker-negative patients could be designed. In contrast to randomizing all available subjects, the latter goal is achieved while maintaining the “enriched” design feature for trial efficiency. Comparison of this methodology with other enrichment design methods in literature is summarized in Table 1.

METHODS

Overall Treatment Effect in an All-Comer Population

Consider a typical randomized controlled two-arm clinical trial setting, where an experimental drug is compared to a control and the treatment effect is measured by treatment difference in response values. The trial enrolls both biomarker-positive patients and biomarker-negative patients.

Let Y_{+e} and Y_{-e} denote the response of a random biomarker-positive patient and a random biomarker-negative patient, respectively, in the experimental arm, both of which are normally distributed. Let W denote the biomarker-positive indicator, where $W \sim \text{Bernoulli}(p)$. Let Y_e denote the response of a patient randomly selected from the experimental arm. Then:

$$Y_e = WY_{+e} + (1 - W)Y_{-e} \quad (1)$$

Similarly for the control arm,

$$Y_c = WY_{+c} + (1 - W)Y_{-c} \quad (2)$$

Therefore,

$$Y_e - Y_c = W(Y_{+e} - Y_{+c}) + (1 - W)(Y_{-e} - Y_{-c}) \quad (3)$$

The expected treatment difference in a general patient population, i.e., in an all-comer population, is as follows:

$$\delta = E[Y_e - Y_c] = E[E[Y_e - Y_c | W]] = p\delta_+ + (1 - p)\delta_- \quad (4)$$

where δ_+ and δ_- are the expected treatment difference in the biomarker-positive subpopulation and the biomarker-negative subpopulation, respectively. When the overall treatment effect in an all-comer population is of interest, δ becomes the parameter of interest.

Table 1 Comparisons of different enrichment design methods

	Trial population	Treatment effect inference	Screening data used
Traditional enrichment design	Biomarker-positive only	Biomarker-positive only	No
Stratified all-comer design including “run-in” design	All comer	1. Biomarker positive then biomarker negative 2. Biomarker positive then all comer 3. All comer then biomarker positive 4. α splitting for biomarker positive and all comer	No
Adaptive enrichment	All comer then potentially biomarker positive	All comer then potentially biomarker positive	No
Enrichment with patient population augmentation	Biomarker positive and biomarker negative, enriched on biomarker positive	1. Biomarker positive then all comer 2. α splitting for biomarker positive and all comer	Yes

Weighted Estimator of Overall Treatment Effect

Under stratified randomization by the biomarker status, δ in Eq. 4 can be estimated by:

$$\hat{\delta} = \hat{p}\hat{\delta}_+ + (1 - \hat{p})\hat{\delta}_- \quad (5)$$

where $\hat{\delta}_+$ and $\hat{\delta}_-$ are the estimated treatment difference in the biomarker-positive and biomarker-negative subpopulations, respectively, and $\hat{p}$ is obtained at the screening stage to estimate the proportion of biomarker-positive patients under the real-world disease prevalence. Therefore, when a trial is enriched for the biomarker-positive subpopulations, $\hat{p}$ will adjust the contribution of biomarker-positive data and biomarker-negative data in estimating the overall treatment effect, so their contributions are aligned as in an all-comer population. It can be shown that $\hat{p}$ is an unbiased and consistent estimate of δ .

The variance of the proposed weighted estimator, again under stratified randomization, can be explicitly derived as:

$$\begin{aligned} \text{Var}[\hat{\delta}] = & \left(p^2 + \frac{p(1-p)}{m} \right) \frac{2\sigma_+^2}{n_+} \\ & + \left((1-p)^2 + \frac{p(1-p)}{m} \right) \frac{2\sigma_-^2}{n_-} \\ & + \frac{p(1-p)}{m} (\delta_+ - \delta_-)^2 \end{aligned} \quad (6)$$

where m is the number of screened patients, n_+ and n_- are the number of randomized patients per treatment arm in the biomarker-positive subpopulation and biomarker-negative subpopulation, respectively, and σ_+^2 and σ_-^2 are the common variance of response (common for the experimental arm and control arm) in the biomarker-positive subpopulation and the biomarker-negative subpopulation, respectively. $\hat{p}$, $\hat{\delta}_+$, $\hat{\delta}_-$ and sample variance for σ_+^2 and σ_-^2 can be plugged into $\text{Var}(\hat{\delta})$ to get a consistent variance estimator. Statistical inference follows by constructing a Z-test based on asymptotic normal distribution of $\hat{\delta}$.

Trial Design and Sample Size Considerations

In line with the enrichment philosophy, a two-step sample size determination strategy can be utilized. Specifically, the biomarker-positive subpopulation is sized first with power $(1 - \beta)$ (step I) at two-sided α to test for δ_+ . Knowing the sample size of the biomarker-positive subpopulation, the sample size of the biomarker-negative population is then calculated to achieve power $(1 - \beta')$ (step II) at two-sided α' when testing for δ . This calculation is carried out by solving:

$$\text{Var}(\hat{\delta}) = \frac{\delta^2}{(Z_{\alpha'/2} + Z_{\beta'})^2} \quad (7)$$

Table 2 Illustrative sample size and power calculations

p	Biomarker+	Overall n, (n ₊ /n)		
	n ₊	$\beta' = 90\%$	$\beta' = 80\%$	$\beta' = 70\%$
0.2	234	535 (0.44)	454 (0.52)	405 (0.58)
0.3	234	464 (0.50)	399 (0.59)	360 (0.65)
0.4	234	410 (0.57)	355 (0.66)	325 (0.72)

Note: Parameter settings: $\delta_+ = 1.5$; $\delta_- = 1.0$; $\sigma_+ = \sigma_- = 5$; $m = 2 \times n_+/p$; $\alpha = \alpha' = 0.05$, two-sided; $\beta = 90\%$; and $\beta' = 70\%$, 80% , and 90% .

with δ in Eq. 4 and $\text{Var}(\hat{\delta})$ in Eq. 6 and with given n_+ and $m = 2 \times n_+/p$. Note that it requires the value of p , and usually a good estimated value can be obtained from external sources prior to trial start.

When sequential testing is planned, α and α' would be the same, typically at 0.05. Under the enrichment philosophy, it is reasonable to test first for δ_+ on the biomarker-positive subpopulation and then for δ on the overall trial population. When type I error rate splitting is planned, α and α' can take different values. Again under the enrichment philosophy, the biomarker-positive subpopulation is recommended to be fully powered at 90% power, and therefore β is 10%. β' can be set as same as β if time and resource of drug development allows. It can also be set higher than β from a practical perspective. For example, β is 10% and β' is 30%, so we size the enriched subpopulation for 90% power and the total trial population for 70% power.

Table 2 presents illustrative calculations for power and sample size under this enrichment design. It can be seen that the real trial composition of biomarker-positive subpopulation, n_+/n , is larger than p which is the prevalence of biomarker-positive patients in general disease populations. It confirms that the weighted estimator down-weighs biomarker-positive data to its prevalence rate, so the overall treatment effect is not driven by the enriched biomarker-positive subpopulation. In addition, it can be seen that larger the p , the smaller the n , meaning that fewer biomarker-negative patients are to be enrolled because the biomarker-positive sample size n_+ remains at 234 to maintain 90% power in this subpopulation.

CONCLUSIONS AND DISCUSSIONS

Enrichment design has been around for a long time but is not widely used. One leading cause is that once we restrict the study population to biomarker-positive patients only, we are only able to establish treatment effect in this subpopulation and thus miss the opportunity to investigate the overall treatment effect in a broad patient population. In the newly proposed enrichment design method, biomarker-positive subpopulation is augmented by additional biomarker-negative patients and so the overall treatment effect can also be assessed. This design has the advantages of retaining a

design and analysis priority for the promising biomarker-positive subpopulation. At the same time, it allows the assessment of the overall treatment effect while maintaining the “enriched” feature of trial design for efficiency, compared to enrolling a large all-comer population. Not all these features can be achieved simultaneously via previous enrichment design approaches.

In order to make correct inference on the overall treatment effect when the biomarker-positive subpopulation is over-represented under the “enriched” trial setting, a weighted estimator is used to combine the treatment effect in the biomarker-positive and biomarker-negative subpopulations at appropriate level. The weighted estimator down-weights the enriched biomarker-positive data and thus will not let the treatment effect in biomarker-positive subpopulation to drive the overall treatment effect. An explicit variance formula for this weighted estimator is derived, which allows for explicitly powering the overall treatment effect, and thus facilitates simple sample size formulation.

Another important feature of the design is the use of screening data. Unlike usual clinical trials that screening data is ancillary for data analysis and definitely not contributing to trial design and sample size planning, screening data are involved in both trial design and data analysis under this new approach. This is *to our knowledge the first time* that screening information is incorporated into clinical trial design and analysis.

To summarize, clinical trials are expensive and time-consuming to conduct, and therefore a carefully selected design strategy can greatly help reduce the cost and expedite drug development. The new enrichment design with patient population augmentation is advocated when there is evidence of a promising treatment effect in a biomarker-positive subpopulation and there is uncertainty on the treatment effect in the complementary biomarker-negative subpopulation, especially when prevalence rate of the biomarker-positive subgroup is not high (e.g., ≤ 0.5).

ACKNOWLEDGMENTS

The research was sponsored by AbbVie. AbbVie contributed to writing, reviewing, and approving the publication.

REFERENCES

- Galera, B.S.; Rowbotham, M.C.; Perander, J.; Friedman, E. Topical lidocaine patch relieves postherpetic neuralgia more effectively than a vehicle topical patch: Results of an enriched enrollment study. *Pain* **1999**, *80* (3), 533–538.
- Simon, R.; Maitournam, A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin. Cancer Res.* **2004**, *10* (20), 6759–6763.
- Maitournam, A.; Simon, R. On the efficiency of targeted clinical trials. *Stat. Med.* **2005**, *24* (3), 329–339.
- Wang, S.J.; O'Neill, R.T.; Hung, H.M.J. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharm. Stat.* **2007**, *6* (3), 227–244. doi: 10.1002/pst.300.
- Simon, R. The use of genomics in clinical trial design. *Clin. Cancer Res.* **2008**, *14* (19), 5984–5993.
- Mackey, H.M.; Bengtsson, T. Sample size and threshold estimation for clinical trials with predictive biomarkers. *Contemp. Clin. Trials* **2013**, *36* (2), 664–672.
- Mandrekar, S.J.; Sargent, D.J. Drug designs fulfilling the requirements of clinical trials aiming at personalizing medicine. *Chinese Clin. Oncol.* **2014**, *3*, 2, 14.
- Paving the Way for Personalized Medicine: FDA's Role in a New Era of Medical Product Development*; U.S. Food and Drug Administration. 2013.
- Baselga, J. Herceptin alone or in combination with chemotherapy in the treatment of HER2-positive metastatic breast cancer: Pivotal trials. *Oncology* **2001**, *61*, Suppl. 2, 14–21.
- Joensuu, H.; Kellokumpu-Lehtinen, P.L.; Bono, P.; Alanko, T.; Kataja, V.; Asola, R.; Utriainen, T.; Kokko, R.; Hemminki, A.; Tarkkanen, M.; Turpeenniemi-Hujanen, T.; Jyrkkio, S.; Flander, M.; Helle, L.; Ingalsuo, S.; Johansson, K.; Jääskeläinen, A.S.; Pajunen, M.; Rauhala, M.; Kaleva-Kerola, J.; Salminen, T.; Leinonen, M.; Elomaa, I.; Isola, J. FinHer Study Investigators. Adjuvant docetaxel or vinorelbine with or without *Trastuzumab* for breast cancer. *New Engl. J. Med.* **2006**, *354*, 809–820.
- Piccari-Gebhart, M.J.; Procter, M.; Leyland-Jones, B.; Goldhirsch, A.; Untch, M.; Smith, I.; Gianni, L.; Baselga, J.; Bell, R.; Jackisch, C.; Cameron, D.; Dowsett, M.; Barrios, C.H.; Steger, G.; Huang, C.S.; Andersson, M.; Inbar, M.; Lichinitser, M.; Láng, I.; Nitz, U.; Iwata, H.; Thomssen, C.; Lohrisch, C.; Suter, T.M.; Rüschoff, J.; Suto, T.; Greatorex, V.; Ward, C.; Strahle, C.; McFadden, E.; Dolci, M.S.; Gelber, R.D. Herceptin Adjuvant (HERA) Trial Study Team. *Trastuzumab* after adjuvant chemotherapy in HER2-positive breast cancer. *New Engl. J. Med.* **2005**, *353* (16), 1659–1672.
- Smith, I.; Procter, M.; Gelber, R.D.; Guillaume, S.; Feyereislova, A.; Dowsett, M.; Goldhirsch, A.; Untch, M.; Mariani, G.; Baselga, J.; Kaufmann, M.; Cameron, D.; Bell, R.; Bergh, J.; Coleman, R.; Wardley, A.; Harbeck, N.; Lopez, R.I.; Mallmann, P.; Gelmon, K.; Wilcken, N.; Wist, E.; Sánchez Rovira, P.; Piccari-Gebhart, M.J. HERA study team. 2-Year follow-up of *Trastuzumab* after adjuvant chemotherapy in HER2-positive breast cancer: A randomised controlled trial. *Lancet* **2007**, *369* (9555), 29–36.
- Freidlin, B.; McShane, L.M.; Korn, E.L. Randomized clinical trials with biomarkers: Design issues. *J. Natl. Cancer I.* **2010**, *102* (3), 152–160.
- Hong, F.; Simon, R. Run-in phase III trial design with pharmacodynamics predictive biomarkers. *J. Natl. Cancer I.* **2013**, *105* (21), 1628–1633.
- Freidlin, B.; Korn, E.L. Biomarker enrichment strategies: matching trial design to biomarker credentials. *Nat. Rev. Clin. Oncol.* **2014**, *11* (2), 81–90.
- Guidance for Industry: Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products*; U.S. Food and Drug Administration, 2012.
- Paik, S.; Kim, C.; Wolmark, N. HER2 status and benefit from adjuvant *Trastuzumab* in breast cancer. *New Engl. J. Med.* **2008**, *358*, 1409–1411.

18. Perez, E.A.; Reinholz, M.M.; Hillman, D.W.; Tenner, K.S.; Schroeder, M.J.; Davidson, N.E.; Martino, S.; Sledge, G.W.; Harris, L.N.; Gralow, J.R.; Dueck, A.C.; Ketterling, R.P.; Ingle, J.N.; Lingle, W.L.; Kaufman, P.A.; Visscher, D.W.; Jenkins, R.B. HER2 and chromosome 17 effect on patient outcome in the N9831 adjuvant *Trastuzumab* trial. *J. Clin. Oncol.* **2010**, *28* (28), 4307–4315.
19. Korkaya, H.; Wicha, M.S. HER2 and breast cancer stem cells: More than meets the eye. *Cancer Res.* **2013**, *73* (12), 3489–3493.
20. Wang, S.J.; Hung, H.M.J.; O'Neill, R.T. Adaptive patient enrichment designs in therapeutic trials. *Biom. J.* **2009**, *51* (2), 358–374.
21. Simon, N.; Simon, R. Adaptive enrichment designs for clinical trials. *Biostatistics* **2013**, *14* (4), 613–625.
22. Wiviott, S.D.; Braunwald, E.; McCabe, C.H.; Montalescot, G.; Ruzyllo, W.; Gottlieb, S.; Neumann, F.J.; Ardissino, D.; De Servi, S.; Murphy, S.A.; Riesmeyer, J.; Weerakkody, G.; Gibson, C.M.; Antman, E.M. 38 Investigators TRITON-TIMI. Prasugrel versus clopidogrel in patients with acute coronary syndromes. *New Engl. J. Med.* **2007**, *357*, 2001–2015.
23. Yang, B.; Zhou, Y.; Zhang, L.; Cui, L. Enrichment design with patient population augmentation. *Contemp. Clin. Trials* **2015**, *42*, 60–67.

Equivalence Trials

Jen-pei Liu

Division of Biometry, Department of Agronomy, and Institute of Epidemiology, National Taiwan University, Taipei, Taiwan; Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan

Abstract

Equivalence trials attempt to compare a test treatment to an active control or the standard therapy. By giving statistical examples, this entry discusses benefits and methods of conducting equivalence trials.

Keywords: Average bioequivalence; Confidence intervals; Individual bioequivalence; Population bioequivalence.

INTRODUCTION

Most of the trials conducted during the developmental stage of investigational pharmaceuticals before filing the New Drug Application (NDA) or Product License Application (PLA) are for demonstration of better efficacy and safety of the test product over the concurrent placebo control. This type of trial is referred to as the superiority trial. On the other hand, some trials try to compare the test treatment to an active control such as some reference treatment or the standard therapy. The purpose of this type of trial is to verify whether the effectiveness or safety of the test treatment is similar to that of the active control. Depending on its objectives, this type of trial can be further classified into equivalence or non-inferiority trials.

After the patent of an innovative product expires, other pharmaceutical companies can manufacture generic copies under the Drug Price Competition and Patent Term Restoration Action passed in 1984 through the Abbreviated New Drug Application (ANDA) based on the pharmacokinetic (PK) responses derived from plasma or blood concentrations collected from bioequivalence studies, such as area under the plasma concentration–time curve (AUC) or the peak concentration (C_{max}). In order to reduce the variation, crossover designs with normal volunteers are usually employed for bioequivalence testing.^[1] Based on the “Fundamental Bioequivalence Assumption,”^[2] if the PK parameters of the generic and innovative products are close, within a pre-specified allowable limits, it is then assumed that they produce no clinically meaningful difference in effectiveness and safety and hence the generic drugs and the marketed innovative reference product can be used interchangeably. On the other hand, some drug products do not work through systemic absorption and have negligible plasma levels. These drugs include metered-dose inhalers (MDIs) indicated for the relief of

bronchospasm in patients with reversible obstructive airway disease, sucralfate for the treatment of acute duodenal ulcer and topical and vaginal antifungals. Because these drug products have negligible plasma concentration of the active ingredients, PK responses are no longer adequate for evaluation of bioequivalence between drug products. As a result, the U.S. Food and Drug Administration (FDA) suggested that clinical trials be conducted to demonstrate the therapeutic equivalence based on well-defined clinical endpoints.^[3,4] Both clinical equivalence or bioequivalence trials are required to demonstrate similarity based on either clinical endpoints or PK responses. They are referred to as “equivalence trials.”

For many diseases, there exists a standard drug that has been proved to be efficacious with a reasonably tolerable safety profile by more than one adequate well-controlled studies. But it is still worthwhile to develop new drugs for the same disease with similar efficacy if they can provide a better safety margin, are easy to administer, are less expensive, require a short duration of treatment, and can substitute for an invasive procedure. As a result, instead of a superior trial, it is sufficient for clinical trials to provide evidence that the new drug is therapeutically no worse than the standard reference treatments with respect to a pre-specified limit of no clinical significance.^[5–9] This type of trial is called the non-inferiority trial, which will be discussed in more detail in the entry *Therapeutic Equivalence*.

EQUIVALENCE TRIALS

The concept of equivalence asks for similarity in distributions of efficacy clinical endpoints or PK responses between the test and the reference drug. It is then referred to as “population equivalence.” Under the normality

assumption, the distributions of most PK responses or efficacy clinical endpoints or their transformation can be uniquely determined by the first two moments, the average and variability. Hence the equivalence between of the test and reference drug products can be established through the similarity of both the average and the variance. Average equivalence requires the similarity in averages of the distributions between the two drug products. On the other hand, population equivalence, under the normality assumption, asks for equivalence in both average and variability. However, in general, equivalence in the first two moments does not guarantee equivalence between distributions. But the assessment of equivalence between a test drug and a reference product, in general, focuses only on the comparison in averages. Only recently has some attention been paid to the equivalence in distributions of PK responses for generic products.^[10,11] For a complete description and recent development of bioequivalence, Chow and Liu^[2]. Chow and Liu^[9] provide further discussion on therapeutic equivalence. Failure to reject the null hypothesis of equality does not imply equivalence between the test and reference product.^[5,12–14] We hence use average as an example to demonstrate the concept of equivalence. Let μ_T and μ_R be the population averages of the test and reference products, respectively. The concept of the equivalence in average is that the difference in average is within the pre-specified allowable limits. In other words, the average of the test product is neither too low nor too high as compared with that of the reference product. As a result, the consumer's risk is the error for declaring average equivalence between the test and reference products when in fact they are not equivalent; the producer's risk is the error of declaring average inequivalence between the two products when in fact they are equivalent. The hypothesis of equivalence in terms of average can therefore be formulated as follows^[15]

$$\begin{aligned} H_0 : \mu_T - \mu_R \geq L \quad \text{or} \quad \mu_T - \mu_R \leq U \\ \text{vs.} \\ H_a : L < \mu_T - \mu_R < U \end{aligned} \quad (1)$$

where L and U are some pre-specified lower and upper equivalence limits, respectively.

Because the above alternative hypothesis is an interval, it is also called the “interval hypothesis of equivalence.” In addition, both lower and upper equivalence limits are required for interval hypothesis in Eq. (1). It is also referred to as “two-sided equivalence.” The interval hypothesis in Eq. (1) is the correct formulation of the hypothesis for the assessment of equivalence in such a manner that the type I error represents the consumer's risk and the type II error is the producer's risk. The selection of equivalence limits should require clinical or pharmacokinetic justification. For average bioequivalence, the FDA guidance^[16] requires a range of 80–125% for the ratio of average area under

curves or peak concentrations between the test product and the reference product.

The interval hypothesis in Eq. (1) can further be decomposed into two one-sided hypotheses as

$$\begin{aligned} H_{0L} : \mu_T - \mu_R \leq L \quad \text{vs.} \quad H_{aL} : \mu_T - \mu_R > L, \\ \text{and} \\ H_{0U} : \mu_T - \mu_R \geq U \quad \text{vs.} \quad H_{aU} : \mu_T - \mu_R < U \end{aligned} \quad (2)$$

A comparison of interval hypotheses in Eq. (1) and two one-sided test hypotheses in Eq. (2) reveals that the parameter space defined in the null hypothesis of the interval hypotheses is the union of the spaces defined in the two one-side null hypotheses. The parameter space defined by the alternative hypothesis in Eq. (2) is the intersection of the spaces of the two alternative hypotheses in Eq. (2). From the intersection–union principle (IUT,^[17]), for each of the two one-sided hypotheses, there is a level α test statistic to verify whether the average of the test product is not too low (high) as compared with that of the reference drug. Therefore if both null hypotheses of the two one-sided hypotheses are rejected at the pre-specified nominal significance level α , then equivalence in average between the test and reference products is also concluded at the α significance level. Because this procedure involves testing two one-sided hypotheses simultaneously, it is called the two one-sided tests procedure.^[2,18]

Because the equivalence between the two drugs is concluded only if the null hypothesis of both one-sided hypotheses is rejected at the pre-specified nominal level of significance, the two one-sided tests procedure can control the consumer's risk at the nominal level. However, if the variability is large or the sample size is too small, the two one-sided tests procedure may be very conservative in concluding equivalence. Hsu et al.^[19] and Brown et al.^[20] have proposed other procedures for the interval hypothesis of equivalence, which are claimed to be more powerful than the usual two one-sided tests procedure. However, their procedures are mathematically complicated and lack intuitive interpretation. The open non-convex shape of their rejection regions imply that any estimate of mean difference in average might conclude average equivalence as long as the standard error of the observed mean difference is sufficiently larger or if it increases to infinity. The hypothesis for evaluation of equivalence in variability can be similarly formulated^[2,21] as

$$\begin{aligned} H_0 : \sigma_T^2 / \sigma_R^2 \geq \delta_L \quad \text{or} \quad \sigma_T^2 / \sigma_R^2 \leq \delta_U \\ \text{vs.} \\ H_a : \delta_L < \sigma_T^2 / \sigma_R^2 < \delta_U \end{aligned} \quad (3)$$

where σ_T^2 and σ_R^2 are the variance of the test and reference product, respectively; $0 < \delta_L < 1 < \delta_U$ are the equivalence limits for evaluation of equivalence in variability. For

bioequivalence testing conducted under crossover design, they are referred to as the intrasubject variability. Under a standard two-sequence and two-period (2×2) crossover design, Liu and Chow^[21] proposed the Pitman–Morgan two one-sided tests procedure for Eq. hypothesis 3. Because each subject receives test or reference product only once under a 2×2 crossover design, estimators for σ_T^2 and σ_R^2 , and σ_T^2/σ_R^2 cannot be obtained without further assumptions. However, for PK responses obtained under replicated crossover designs, inference about equivalence based on σ_T^2/σ_R^2 including confidence interval is quite straightforward.^[2,21] This discussion about equivalence in average or variability can be applied to the responses obtained from either parallel designs or crossover designs. For bioequivalence trials using either the standard 2×2 or replicated crossover designs, see Chow and Liu^[2,15]. Chow and Liu^[2,22–24] have provided sample size determination for equivalence in averages.

CONFIDENCE INTERVAL APPROACHES

Although the discussion for assessment of equivalence is in the context of statistical hypothesis testing procedure, it also is an estimation problem. As a result, many have advocated the confidence interval approach to assessment of equivalence between the test and reference drug.^[2,6,13,14,25,26] The idea and procedure for the assessment of equivalence are quite straightforward. For two-sided average equivalence, if the nominal level is α , then the two drugs are concluded equivalent if the $100(1 - 2\alpha)\%$ confidence interval for the mean difference is completely contained within (L, U). The confidence interval approach is more appealing than the testing procedure because it can provide the magnitude and width of the average difference between the two drugs. In addition, the hypothesis testing and confidence interval approaches in some situations are in fact operationally equivalent.^[18] For many cases, the pivotal quantity for construction of confidence interval is the test statistic for the two one-sided tests procedure, which follows either a central t distribution or the standard normal distribution. As a result, for two-sided equivalence problem, the $100(1 - 2\alpha)\%$ confidence interval is totally contained within (L, U) if and only if both one-sided hypotheses in Eq. (2) are rejected at the α significance level.

EQUIVALENCE LIMITS

The decision for assessment of equivalence by either the confidence interval approach or the testing procedure depends upon the lower and upper equivalence limits and has nothing to do with whether the observed difference is zero or not. Table 1 illustrates the consistency between the two one-sided tests procedure and the confidence interval approach for the two-sided equivalence. As can be seen in Table 1, although the null hypothesis of equality is not rejected, the equivalence between the two drugs cannot be concluded because the 90% confidence interval may be too wide to be contained within the equivalence limits. Another situation is that the equivalence between the test and reference drugs is concluded because the confidence interval is very narrow and is totally contained within the limits although it does not contain 0. In other words, equivalence can be concluded even when the null hypothesis of equality is rejected; it is then concluded that there is a difference between the test and reference product. As a result, the length of equivalence interval, U–L for therapeutic equivalence, in general, should be smaller than the difference in averages between the test and concurrent placebo that superior trials try to detect.

Although the equivalence limits for average bioequivalence required by the FDA are well accepted and established, selection of the equivalence limits for assessing similarity is always controversial and remains a challenge in other areas.^[27,28] If the selected equivalence limits are too narrow, it is extremely difficult to conclude equivalence although the difference between them is of no clinical significance. On the other hand, if the chosen equivalence limits are very wide, then it is easy to declare the two drugs with a clinically meaningful difference to be equivalent.

In general, the equivalence limits depend on the response of the reference drug. For example, Huque and Dubey^[4] have proposed some equivalence limits for binary data such as eradication rate for antibiotics. They suggested that the equivalence limits be tighter for higher reference response rates. Suppose that the response rate was 75% from two adequate, well-controlled studies for approval of the reference drug. The equivalence limits suggested by Huque and Dubey^[4] is $\pm 20\%$. The sample size can also be determined based on $\pm 20\%$ equivalence limits. However, when the current equivalence trial yields a reference response rate of 85%, should the equivalence limits be

Table 1 Comparison of the Results by Different Procedures for Assessing Equivalence with a Known Symmetric Limit of 0.2

Observed difference	90% (C.I.)	H _{0E}	H _{0L}	H _{0U}	Two-sided equivalence
0.00	(−0.10, 0.10)	Fail to reject	Reject	Reject	Yes
0.00	(−0.25, 0.25)	Fail to reject	Fail to reject	Fail to reject	No
0.05	(0.02, 0.08)	Reject	Reject	Reject	Yes
0.10	(−0.02, 0.22)	Fail to reject	Reject	Fail to reject	No

H_{0E} is the null hypothesis for equality. (From Chow and Liu^[9].)

changed to $\pm 15\%$ as suggested by Huque and Dubey?^[4] Is the sample size calculated with the limits of $\pm 20\%$ large enough to provide sufficient power for the limits of $\pm 15\%$?

These questions reveal the issues that the equivalence limits based on the reference responses from any previous studies are not internally valid. Therefore the equivalence limits may be chosen based on the reference response obtained concurrently within the same equivalence trial for internal validity. In this case, the lower and upper equivalence limits in the interval hypothesis Eq. (2), however, are no longer fixed known quantities but random variables. The formulation for interval hypothesis should be reformulated as

$$H_a : L < \mu_T / \mu_R U \quad (4)$$

where $0 < L < 1 < U$.

The interval hypothesis in Eq. (4) implies that the two drugs are equivalent in average if the ratio of averages is between L and U . Logarithmic transformation of Eq. (4) reformulates the interval hypothesis from ratio to the difference as

$$H_a : \ln(L) < \ln(\mu_T) - \ln(\mu_R) < \ln(U)$$

where $\ln$ denotes the natural logarithm and $\ln(L) < 0 < \ln(U)$.

If the test and reference drug products are equivalent, both of them could be better than placebo or they could be worse than placebo. Evaluation of equivalence through the interval hypothesis in Eq. (4) therefore does not necessarily establish efficacy of the test drug product without a concurrent placebo control. Lack of assay sensitivity is one of the serious issues for active controlled trial. For more discussion, see Wang et al.^[29]

For assessment of average bioequivalence between a generic and a reference formulation, most regulatory agencies such as the FDA and EC request that PK responses such as AUC or $C_{\max}$ be analyzed on the logscale with $L = 80\%$ and $U = 125\%$, that is, $\ln(L) = -0.2231$ and $\ln(U) = 0.2231$.

On the other hand, if data are analyzed on the original scale, the equivalence limits in Eq. (2) depend on the observed reference mean for internal validity. Suppose that in Eq. (2) $L = f\mu_R$ and $U = g\mu_R$, then the two one-sided hypotheses corresponding to Eq. (2) can then be formulated as

$$\begin{aligned} H_{0L} : \mu_T - (1-f)\mu_R &\leq 0 \\ \text{vs.} \\ H_{aL} : \mu_T - (1-f)\mu_R &> 0, \\ \text{and} \\ H_{0U} : \mu_T - (1-g)\mu_R &\leq 0 \\ \text{vs.} \\ H_{aU} : \mu_T - (1-g)\mu_R &> 0. \end{aligned} \quad (5)$$

Consequently, the two one-sided tests procedure can be constructed based on the estimates for $\mu_T - (1-f)\mu_R$ and

$\mu_T - (1-g)\mu_R$ and their estimated standard errors. However, currently, the two one-sided tests procedure for average are performed with estimated equivalence limits based on the observed reference mean as the true unknown fixed quantities without taking into consideration the variability of estimated equivalence limits. Liu and Weng^[30] have shown that for the data obtained from the 2×2 crossover design, the traditional application of the two one-sided tests procedure cannot control the probability of type I error. When the correlation between the two responses from the same patient approaches 1, the consumer's risk goes to 50%. They proposed a modified two one-sided tests procedure based on the two one-sided hypotheses in Eq. (5), which adequately controls type I error rate at the nominal level.

POPULATION BIOEQUIVALENCE

As demonstrated above, average bioequivalence does not guarantee that the generic and reference product can be used interchangeably. In general, interchangeability between the generic product and the reference product can be classified as prescribability and switchability. "Prescribability" is referred to as the physician's choice, for prescribing a drug for new patients, between an innovative product and a number of its generic copies that have been shown to be bioequivalent to the innovative product.^[11] On the other hand, "switchability" is related to the switch from a drug such as an innovative product or a generic drug to an alternative drug such as a generic copy of a brand-name drug or another generic copy of the same innovative drug product within the same patient whose concentration of the drug has been titrated to a steady, efficacious, and safe level. To ensure switchability, the equivalence between drug products must be demonstrated within each individual. This concept is referred to as "individual bioequivalence."

For evaluation of population bioequivalence, as mentioned above, under normality assumption, equivalence in both average and variability is required. One approach is to conclude population bioequivalence at the α significance level if both the null hypothesis for average in Eq. (1) and that for variability in Eq. (4) are rejected at the α significance level.^[21] This disaggregate procedure is another application of the union-intersection principle. As a result, although it is a size α test, it can be very conservative. The FDA issued guidance on January 2001 on statistical approaches to establish bioequivalence. The guidance suggested using the following mixed-scaling aggregate criterion for assessment of population bioequivalence based on the log-scale of AUC or $C_{\max}$:

$$\theta = [(\mu_T - \mu_R)^2 + (\sigma_{TT}^2 - \sigma_{TR}^2)] / \max(\sigma_{TR}^2, \sigma_{TO}^2) \quad (6)$$

where σ_{TT}^2 is the total variance (sum of intrasubject and intersubject variability) of the test formulation, σ_{TR}^2 is the

total variance of the reference formulation, and σ_{TO}^2 is the specified constant total variance.

If $\sigma_{TR}^2 > \sigma_{TO}^2$, the aggregate criterion in Eq. (6) is scaled by the total reference variability. It is then referred to as the reference-scaled criterion. When σ_{TR}^2 is smaller than σ_{TO}^2 , the guidance proposes that σ_{TR}^2 in the denominator of Eq. 6 is replaced by σ_{TO}^2 . The resulting criterion is then referred to as constant-scaled criterion. Population bioequivalence is concluded if θ is smaller than some pre-specified upper equivalence limit θ_p . The corresponding hypothesis can be written as

$$H_0: \theta \geq \theta_p \quad \text{vs.} \quad H_a: \theta < \theta_p \quad (7)$$

The FDA guidance suggests that sponsors or applicants contact the FDA for further information about σ_{TO}^2 and θ_p if the sponsors or applicants wish to use the population bioequivalence criterion. The distribution of the estimator of θ is quite complicated. The confidence interval approach to test hypothesis Eq. (7) for evaluation of bioequivalence requires the use of the bootstrap procedure. As a result, the FDA guidance suggests using the following linearized criterion:

$$\eta = [(\mu_T - \mu_R)^2 + (\sigma_{TT}^2 - \sigma_{TR}^2)] - \theta_p \max(\sigma_{TR}^2, \sigma_{TO}^2) \quad (8)$$

The corresponding hypothesis is given as

$$H_0: \eta \geq 0 \quad \text{vs.} \quad H_a: \eta < 0 \quad (9)$$

The FDA guidance recommends the method proposed by Hyslop et al.^[31] to derive close-form upper approximate confidence limit for η . The population bioequivalence is concluded at the α significance level if the upper $(1 - \alpha)100\%$ confidence limit for η is smaller than 0 provided the point estimate of μ_T/μ_R based on the geometric mean is with 0.8 and 1.25. From Eq. 6, the aggregate criterion proposed in the FDA guidance is based on the difference in average bioavailability and difference in total variability scaled by the reference total variability. As a result, crossover design is no longer required for assessment of population bioequivalence. A two-group parallel design with a single PK response from each subject will suffice for evaluation of population bioequivalence.

Individual bioequivalence is one of the approaches used to assess bioequivalence between the generic and reference listed drug recommended by the FDA guidance.^[32] Aggregate referenced-scaled and constant-scaled criteria similar to those for population bioequivalence are also proposed for evaluation of individual bioequivalence. The criterion for individual bioequivalence is scaled by the reference intrasubject variability. The constant-scaled criterion is selected if the reference intrasubject variability is smaller than 0.04. There are many unresolved issues for the aggregate criterion such as masking effects, unknown bias and power function, inference for second moments, two-stage

procedure, interpretation of results, and others. These issues will be addressed in detail in the entry *Individual Bioequivalence*.

REFERENCES

1. Federal Register **1977**, 42 (5), 320.26(b) US Government Printing Office, Washington, DC.
2. Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*, 3rd Ed.; CRC/Chapman & Hall: New York, 2008.
3. Huque, M.F.; Dubey, S.; Fredd, S. Establishing Therapeutic Equivalence with Clinical Endpoints. In *The 1989 Proceedings of the Biopharmaceutical Section of the American Statistician Association*; 1989, 46–52. Alexandria, VA.
4. Huque, M.F.; Dubey, S. Design and Analysis of Therapeutic Equivalence Clinical with a Binary Clinical Endpoint. In *The 1990 Proceedings of the Biopharmaceutical Section of the American Statistician Association*; 1990, 46–52. Alexandria, VA.
5. Dunnett, C.W.; Gent, M. Significance testing to establish equivalence between treatments with special reference to data in the form of 2_2 table. *Biometrics* **1977**, 33, 593–602.
6. Makuch, R.W.; Johnson, M. Active Control equivalence Studies: Planning and Interpretation. In *Statistical Issues in Drug Research and Development*; Peace, K. E., Ed.; Marcel Dekker, Inc.: New York, 1990, 238–246.
7. Durrleman, S.; Simon, R. Planning and monitoring of equivalence studies. *Biometrics* **1990**, 46, 329–336.
8. Blackwelder, W.C. “Proving the null hypothesis” in clinical trials. *Control Clin. Trials* **1982**, 3, 345–353.
9. Chow, S.C.; Liu, J.P. *Design and Analysis of Clinical Trials: Concepts and Methodologies*, 2nd Ed.; John Wiley and Sons: New York, 2004.
10. Liu, J.P. Invited Discussion of “Individual Bioequivalence: a Problem for Switchability”. In *Biopharmaceutical Report*; Anderson, S., Ed.; The Biopharmaceutical Section of the American Statistical Association, 1994; Vol. 2, 7–9, Alexandria, VA.
11. Chow, S.C.; Liu, J.P. Current issues in bioequivalence trials. *Drug Inf. J.* **1995**, 29, 795–804.
12. Metzler, C.M. Bioavailability: A problem in equivalence. *Biometrics* **1974**, 30, 309–317.
13. Westlake, W.J. Use of confidence intervals in analysis of comparative bioavailability trials. *J. Pharm. Sci.* **1972**, 61, 1340–1341.
14. Westlake, W.J. Symmetrical confidence intervals for bioequivalence trials. *Biometrics* **1976**, 32, 741–744.
15. Chow, S.C.; Liu, J.P. On assessment of bioequivalence with high-order crossover designs. *J. Biopharm. Stat.* **1992**, 2, 239–256.
16. The FDA Draft Guidance. *Bioavailability and Bioequivalence Studies for Orally Administered Drug Products-General Considerations*; The Center for Drug Evaluation and Research, the Food and Drug Administration: Rockville, MD, 2002.
17. Berger, R.L. Multiparametric hypothesis testing and acceptance sampling. *Technometrics* **1982**, 24, 295–300.

18. Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioequivalence. *J. Pharmacokinet. Biopharm.* **1987**, *15*, 657–680.
19. Hsu, J.C.; Hwang, J.T.G.; Liu, H.K.; Ruberg, S.J. Confidence intervals associated with tests for bioequivalence. *Biometrika* **1994**, *81*, 103–114.
20. Brown, L.D.; Hwang, J.T.G.; Munk, A. An unbiased test for the bioequivalence problem. *Annual. Stat.* **1977**, *25*, 2345–2367.
21. Liu, J.P.; Chow, S.C. On the assessment of variability in bioavailability/bioequivalence studies. *Comm. Stat. Theory Methods* **1992**, *21*, 2591–2607.
22. Liu, J.P. Use of replicated cross-over designs in assessing bioequivalence. *Stat. Med.* **1995**, *14*, 1067–1078.
23. Liu, J.P.; Chow, S.C. Sample size determination for the two one-sided tests procedure in bioequivalence. *J. Pharmacokinet. Biopharm.* **1992**, *20*, 101–104.
24. Liu, J.P. Letter to the Editor. on “Sample size for therapeutic equivalence based on confidence interval” by S.C. Lin. *Drug Inf. J.* **1995**, *29*, 1063–1064.
25. Makuch, R.W.; Simon, R. Sample size requirements for evaluating a conservative therapy. *Cancer Treat. Rep.* **1978**, *6*, 1037–1040.
26. Jennison, C.; Turnbull, B.W. Sequential equivalence testing and repeated confidence intervals, with applications to normal and binary responses. *Biometrics* **1993**, *49*, 31–43.
27. Ebutt, A.F.; Frith, L. Practical issues in equivalence trials. *Stat. Med.* **1998**, *17*, 1691–1701.
28. Rohmel, J. Therapeutic equivalence investigations: Statistical considerations. *Stat. Med.* **1998**, *17*, 1703–1714.
29. Wang, S.J.; Hung, J.H.M.; Tsong, Y. Utility and pitfalls of some statistical methods in active controlled trials. *Control Clin. Trials* **2002**, *23*, 15–28.
30. Liu, J.P.; Weng, C.S. Bias of two one-sided tests procedures in assessment of bioequivalence. *Stat. Med.* **1995**, *14*, 853–862.
31. Hyslop, T.; Hsuan, F.; Holder, D.J. A small sample confidence interval approach to assess individual bioequivalence. *Stat. Med.* **2000**, *19*, 2885–2897.
32. The FDA Guidance. *Statistical Approaches to Establishing Bioequivalence*; The Center for Drug Evaluation and Research, the Food and Drug Administration: Rockville, MD, 2001.

Ethnic Factors

Melody H. Lin

Office for Human Research Protections, Department of Health and Human Services,
Rockville Maryland, U.S.A.

Helen McGough

Human Subjects Division, University of Washington, Seattle, Washington, U.S.A.

Abstract

Ethnicity may be used to refer to “racial” or “genetic” differences between groups of people. It may be used to refer to behavioral and social differences, such as language, religion, dress, mode of making a living, or kinship patterns. This entry addresses the factors associated with how the results of a clinical study translate across ethnic borders.

Keywords: Ethnicity; Extrinsic; Intrinsic.

INTRODUCTION

Can efficacy and safety data about an experimental drug or device from a trial conducted in one region be extrapolated to other regions? Will governmental regulatory bodies accept data collected in a region other than its own in support of drug or device registration? Will the health care environment in one region respond in the same manner to a new drug or device as in another region? Questions such as these often result from the issue of how important ethnic factors are in determining the safety, efficacy, dosage, and therapeutic regimen of a new drug or device.

DEFINING ETHNIC FACTORS

Ethnicity is a concept that is defined in many different ways and to meet different objectives. *Ethnicity* may be used to refer to “racial” or “genetic” differences between groups of people. It may be used to refer to behavioral and social differences, such as language, religion, dress, mode of making a living, or kinship patterns. It may also refer to differences in climate and environmental and occupational exposures. The problems inherent in the use of this term are reflected in the current dispute in the United States about how to categorize its citizens into ethnic groups. Should all Spanish-speaking people be labeled “Hispanic”? Such a category would lump together people from as disparate cultural origins as Spain, Argentina, and Cuba. Should the term “African American” apply to a sixth-generation descendant of West Africans brought to North America over 200 years ago as well as to a recent immigrant from Nigeria or South Africa?

In the conduct of clinical trials, the concept of *ethnicity* may also be used in a variety of ways. Ethnic factors may relate to differences in metabolism between groups of people, differences in health care practices, and differences in response to pain, for example. Ethnic factors may have an impact not only on the safety and efficacy of a new drug or device, but also on the acceptance the drug or device is likely to have—how well it will fit in to a particular population’s lifestyle, budget, and health care practices.

The International Conference on Harmonization (ICH) has proposed a distinction between what it describes as the physiologic aspects of ethnicity (*intrinsic* ethnic factors) and the cultural and environmental aspects of ethnicity (*extrinsic* ethnic factors). Under *intrinsic* factors, the ICH Guidelines include genetic makeup of a population, similarities in height, body mass and composition, and organ dysfunction. Under *extrinsic* factors, the ICH Guidelines include environmental factors, such as climate, pollution, and altitude, and cultural or behavioral factors, such as diet, religion, the use of drugs (including tobacco) and alcohol, and indigenous health care systems.

This is a useful way to categorize ethnic factors. However, such a categorization may gloss over the importance of considering the interrelationships between intrinsic and extrinsic ethnic factors. As in many disease syndromes, such as alcoholism, diabetes, heart disease, and mental illnesses, extrinsic and intrinsic factors may have to work together to potentiate the emergence of the disease itself. Genetic predisposition may not inevitably lead to disease, but genetic predisposition plus environmental exposure may make disease development certain.

WHY IS IT IMPORTANT TO INCLUDE ETHNIC FACTORS IN DESIGNING SAFE AND EFFECTIVE TRIALS?

There are at least two ways that ethnic factors influence the conduct of clinical trials. One is from the point of view of study design and implementation. The other is from the point of view of the utility of the data collected. Ethnic factors must be considered in the design and implementation of the trial in order to produce useful data for purposes of registration. But ethnic factors must also be considered before the decision to conduct a trial in a foreign region is made. What are some of the reasons for the importance of considering ethnic factors in the conduct of clinical trials?

1. Unless these factors are considered, we will not have an adequate understanding of the relationship among culture, race, environment, and disease in terms of pathobiology, etiology, diagnosis, progression, treatment, outcomes, risk reduction, and drug delivery. These issues will play an important role in the methodological design of a clinical trial.
2. Ethnicity is often a marker of other kinds of distinguishing features among people around the world. Ethnicity may mark disparities in levels of health (malnutrition, hopelessness) or in the transfer of information about health and disease (access to care, types of care available, attitudes toward risk reduction). Ethnic factors may determine whether a drug or device will have market acceptance once it is approved. These are issues to be considered before registration decisions are made.

IMPLICATIONS FOR THOSE CONDUCTING CLINICAL TRIALS IN FOREIGN REGIONS

Let us assume that the decision has been made to conduct a trial in a foreign region. This decision initiates special considerations, responsibilities, and opportunities for the researcher as well as for the region in which the study will be conducted.

1. *Involvement of local researchers and research staff, including social scientists such as anthropologists, sociologists, psychologists, and epidemiologists in the trial design as well as implementation.* Knowledgeable researchers can improve the way trials are conducted in many important ways, including issues of recruitment, retention, adherence, and loss to follow-up. Understanding regional attitudes toward medicine, for example, may improve the design of a trial of a new drug by recognizing potentially confounding variables, such as the use of traditional medicines,

diet, and eating and drinking patterns. Providing incentives that are culturally appropriate may improve retention and adherence. Recognizing mobility patterns and anticipating them may result in smaller losses to follow-up.

2. *Designing pharmacokinetic and pharmacodynamic studies tailored to what is known about the population in which the trial will be conducted.* It is important to conduct clinical trials in foreign regions that take into account such ethnic factors as differences in metabolism, dose response, and genetic differences. These factors may affect the levels at which unacceptable toxicity is reached or efficacy rates. Differences in the levels of clearance enzymes, lipids, fibrinogen, and glucose, for example, may have important implications for the extrapolation of data from one region to another. Without accurate information, the human subjects of such trials could be put at unnecessary risk, and the resulting data may be misunderstood and useless.
3. *Designing dose-response and efficacy studies to include well-defined endpoints and disease definitions that are recognized by the local researchers and by the population members.* Ethnic factors may play a role in how *health* and *wellness* are defined. For example, the degree to which the risk of impotence is acceptable may affect the success or failure of a trial of an experimental prostate cancer drug.
4. *Evaluating clinical trials conducted in the population of drugs in the same class as the experimental drug, or similar devices, or of different dosages in a determination of whether or not a new trial must be conducted.* Before a decision to conduct a trial in a foreign region is made, researchers must determine if ethnic factors affect a new drug's safety and efficacy. It may be that trials of drugs in the same class as the experimental drug have shown no outcome or toxicity differences between two regions. It may be that similar dosages and therapeutic regimens have been similar among the two populations. It may not be necessary to conduct the trial at all. On the other hand, it may be that no trials of similar drugs, dosages, or therapeutic regimens have been conducted, or that previous trials have produced dissimilar results. In this case, if the decision to seek registration in a foreign region has been made, it is important to conduct carefully controlled safety and efficacy trials among that region's population.

IMPLICATIONS FOR HUMAN SUBJECTS REVIEW

Clinical trials in foreign regions may pose special challenges for the ethical review boards charged with the

protection of human subjects, especially as global development of new drugs and devices expands and matures. The diversity and breadth of clinical trials will increase, and the expertise required of ethical review boards will also expand. What are some of the steps that ethical review boards can take to prepare for these new challenges?

1. *Learn the ICH's good clinical practices (GCP) with regard to Institutional Review Board (IRB) or Institutional Ethics Committee (IEC) guidance.* The guidance published in the Federal Register (June 10, 1998) outlines the important conditions and procedures that are required to conduct ethical reviews of clinical trials.
2. *Require local ethical review of the clinical trial and request documentation of its findings.* It will become even more important than it is today to make sure that clinical trials receive ethical review in the regions in which they are conducted. This may result in a multiplicity of boards reviewing the same trial from quite different perspectives. It will become incumbent upon each board to make sure it maintains open communication with the researcher as well as with the other ethical review boards considering the same trial. Each review board may request documentation of the local region's determination with regard to a particular trial before making its own decision. Not only is requesting documentation helpful to the boards reviewing a study, it also makes sure that the population of the region plays a part in determining whether or how the trial is conducted.
3. *Involve experts who can contribute information on ethnic factors in the consideration of a new clinical trial.* Traditionally, ethical review boards have been composed primarily of medical personnel who are experts in certain medical traditions. It will be important to expand the membership of these boards to incorporate less traditional members, including citizens of the region in which the trial will be conducted as well as experts in understanding and evaluating the impact of ethnic factors on the conduct of trials. The education involved in expanding membership is a two-way process. Conventional board members will need to learn more about ethnic factors, and the new members will have to learn about the ethics of research with human subjects.

BARRIERS TO THE ETHICAL CONDUCT OF TRIALS IN FOREIGN POPULATIONS

Time

In an increasingly competitive marketplace, speed is of the essence in bringing a drug to approval. Conducting trials in foreign populations will increase the time this process takes.

Budget

Conducting trials in foreign populations may be even more expensive than conducting trials in the sponsor's home country or in the country that is the primary marketing target.

Communication

Dealing with researchers and research subjects who are ethnically different from one another and from the sponsor may create problems leading to trials that are conducted poorly, are at variance with the protocol, and result in incomplete or inadequate data. Miscommunication may result not only from language differences but also from religious, political, or other "extrinsic" ethnic factors.

Stigmatization

Sponsors and researchers may need to work to overcome initial distrust and fear of "foreign" research and of researchers conducting trials in their regions. Education of the region's population and learning from the region may be a slow, and therefore costly, process.

CONCLUSION

Despite these barriers, the ethical conduct of clinical trials in all regions of the world is now an easier goal to accomplish than ever before. Increasing agreement among different regions on the ethical principles necessary to conduct high-quality trials while protecting human subjects will lead to heightened consistency and cooperation among ethical review boards. The basic ethical principles of respect, beneficence, and justice are grounded in the concept of regional review. Regional ethical review boards best evaluate ethnic factors. Cooperation between these boards and researchers will lead to safer, more efficient clinical drug trials and delivery of high-quality health care to humans everywhere.

Expiration Dating Period

Annpey Pong

Merck Research Laboratories, Summit, New Jersey, U.S.A.

Abstract

The expiration dating period (or shelf-life) of a drug product is defined as the time interval that the characteristics of the product will remain within the approved specification limits after manufactured. This entry focuses on the methods of determining the expiration dating period of a drug product and the regulatory requirements that influence this decision.

Keywords: Batch-to-batch; Direct method; Inverse method; Shelf-life.

INTRODUCTION

The Food and Drug Administration (FDA) requires that for every drug product on the market, an expiration dating period must be indicated on the immediate container label. The *expiration dating period* (or shelf-life) is defined as the time interval that the characteristics of the drug product will remain within the approved specification limits after manufactured.

For determination of the expiration dating period of a given batch of a drug product, as indicated in the FDA stability guideline in 1987,^[1] the expiration dating period can be estimated as the time at which the 95% one-sided lower (upper) confidence bound for the mean degradation curve intersects the approved lower (upper) specification limit if the drug characteristic is expected to decrease (increase) with time in a linear fashion (see Fig. 1). If the relationship between the drug characteristic of interest and time is not linear, it may be linearized using an appropriate transformation (e.g., a linear, quadratic, or cubic function on an arithmetic or logarithmic scale).

It should be noted that the term *shelf-life* and *expiration dating period* are often used interchangeably in the literature. In fact, the *expiration dating period* is the term used in the FDA's guideline for the labeled shelf-life, while the *shelf-life* is a broad name used in drug research and development. The true shelf-life is defined as the point of intersection between true degradation curve and the specification limit of a drug product.

REGULATORY REQUIREMENTS

To assure the integrity of commercial drugs, the U.S. Pharmacopeia (USP) included the first clause regarding drug shelf-life in 1975. In 1984, the FDA first required stability testing. To determine the expiration dating period,

a stability study is usually conducted to characterize the degradation of the drug product. The stability testing will provide evidence on how the quality of a drug product varies with time under the influence of a variety of environmental factors (e.g., temperature, humidity, and light) and to establish expiration dating period appropriate for the drug product. That is, stability studies will provide certain assurance or confidence that the drug product meets the standards for expected characteristics of identity, strength, quality, and purity prior to the expiration date.

The specific requirements on statistical design and analysis of stability studies for human drugs and biologics were not available until the FDA stability guideline issued in 1987. In 1993, the International Conference on Harmonization (ICH) issued guidelines^[2] on stability based on a strong industrial interest in international harmonization of regulatory requirements for marketing in the European Community (EC), Japan, and the United States. The six ICH member representatives are the European Commission, the European Federation of Pharmaceutical Industry Associations, the Japanese Ministry of Health and Welfare, the Japanese Pharmaceutical Manufacturers Association, the FDA, and the U.S. Pharmaceutical Manufacturers Association. At the same time, the World Health Organization (WHO) also published draft guidelines for stability. After ICH deliberations led to several stability guidelines in the period from 1994–1997, the draft *FDA Guideline on Stability Practice* was published in June 1998. This draft FDA guideline^[3] incorporates the texts of four ICH documents, which were revised and expanded from the 1987 stability guideline. There was no specific change in statistical principles in analyzing stability data in this draft guideline compared to the 1987 stability guideline. Note that the most recent issue refers to ICH Q1A(R2) *Stability Testing of New Drug Substances and Products* in 2003.^[4]

Generally, there are different criteria for acceptable levels of stability with respect to chemical, physical,

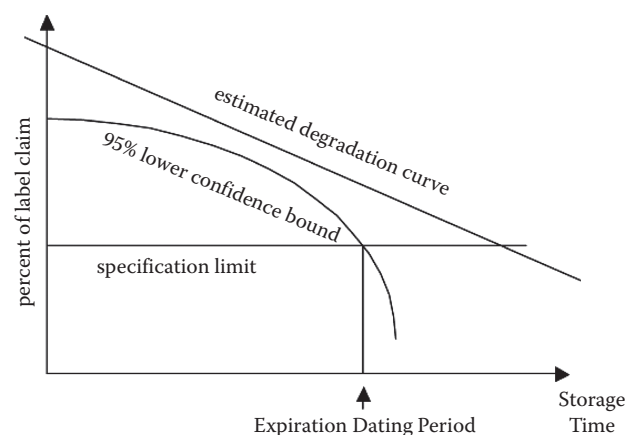


Fig. 1 Determination of drug expiration dating period

Table 1 Drug Characteristics for Different Dosage Forms

Dosage form	Drug characteristics
Tablets	appearance, friability, hardness, color, odor, moisture, strength, dissolution
Capsules	strength, moisture, color, appearance, shape, brittleness, dissolution
Emulsions	appearance, color, odor, pH, viscosity, strength
Oral solution and suspensions	appearance, strength, pH, color, odor, redispersibility, dissolution, clarity
Oral powder	appearance, pH, dispersibility, strength
Metered dose ophthalmic preparations	strength, delivered dose per actuation, number of metered doses, color, clarity, particle size, loss of propellant, pressure, valve corrosion, spray pattern
Topical and ophthalmic preparations	appearance, clarity, color, homogeneity, odor, pH, redispersibility, consistency, particle size distribution, strength, weight loss
Small-volume parenterals	strength, appearance, color, particulate matter, pH, sterility, pyrogenicity
Large-volume parenterals	strength, appearance, color, clarity, particulate matter, pH, volume, extractables, sterility, pyrogenicity
Suppositories	strength, softening, range, appearance, dissolution

microbiological, therapeutic, and toxicological characteristics of drug products.^[5] Table 1 gives a list of drug characteristics for different dosage forms that should be evaluated in a stability study. The strength of a drug product is defined as either (i) the concentration of the drug substance or (ii) the potency, i.e., the therapeutic activity of the drug product which can be determined by an appropriate laboratory test or by an adequately developed and controlled clinical data. For the analysis of stability data, the FDA requires that percent of label claim, not percent of initial average value, be used as the primary variable for strength. In practice, the determination of drug expiration

Table 2 Regulatory Differences

Regulatory factor	FDA	EC/Japan
Drug product commerce	Regulate drug products for U.S. and international commerce	Focused on intranational commerce
Specifications Strength (assay)	90–110% label claim	95–105% (release and shelf specifications) in many cases. Must provide justification for 90–110%
Labeling	Controlled room temperature (CRT)	Room temperature
Definition of temperature	15–30°C, USP-PF proposal	EC, defined as 15–25°C Japan, defined as 1–30°C
Expiration dating period	Based on statistical evaluation of three lots on stability studies-assay/impurities Physical attributes (e.g., hardness, moisture, dissolution, appearance)	Based on summary of at least two lots—no statistical requirement
Container/closure	Actual marketing package	No well-defined requirements

dating period is usually based on the primary drug characteristic of interest such as strength or potency rather than other drug characteristics in most stability studies.

The ICH guideline is similar to the FDA guideline in principle. However, the ICH guideline provides a general indication on the requirements for stability testing, but leaves sufficient flexibility to encompass the variety of different situations and characteristics of the materials being evaluated.^[6] To compare the regulatory difference among the United States, European Community, and Japan, Kumkumian^[7] summarized the results shown in Table 2.

The stability testing for post-approval changes is also addressed in the FDA draft guideline.^[4] In the past, all post-approval changes were considered together and required extensive stability documentation. With the publication of the Scale-Up and Post-Approval Change for Immediate Release/Modified Release (SUPAC-IR/MR) solid oral dosage forms guideline, this approach was modified. For example, the stability commitment to support post-approval changes could rely on the first one or three production batches and annual batches thereafter on long-term stability studies. If the data give no reason to believe that the proposed change will alter the stability of the drug product, the previously approved shelf-life can be used.

Statistical Methods

To determine the drug expiration dating period, three methods are illustrated. The first method is recommended in the FDA's guideline, and the other two methods, direct method and inverse method, were proposed by Shao and Chow.^[8] The completed information on statistical design and analysis of stability studies can be found in Chow, 2007.^[9]

FDA's Approach

In the guideline, assuming the drug characteristic decreases linearly with time, then the following linear regression model is used.

$$y_j = \alpha + \beta t_j + e_j, \quad j = 1, \dots, n \quad (1)$$

where y_j is the assay result (percent of labeled claim) at time t_j , α and β are the intercept and slope, respectively, t_j 's are the sampling time points selected in the stability study, e_j 's are random errors which are independent and identically distributed as $N(0, \sigma^2)$.

The commonly used method for the estimation of α , β , and σ^2 is the method of ordinary least squares (OLS), which yields $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{t}$, $\hat{\beta} = S_{TY}/S_{TT}$, $\hat{\sigma}^2 = S_{YY} - \hat{\beta}S_{TY}/n - 2$ where

$$\begin{aligned} \bar{t} &= \frac{1}{n} \sum_{j=1}^n t_j, \quad \bar{y} = \frac{1}{n} \sum_{j=1}^n y_j, \\ S_{TT} &= \sum_{j=1}^n (t_j - \bar{t})^2, \quad S_{TY} = \sum_{j=1}^n (t_j - \bar{t})(y_j - \bar{y}), \\ S_{YY} &= \sum_{j=1}^n (y_j - \bar{y})^2 \end{aligned}$$

At a specific time point, $t = t_0$, the least square estimate of mean degradation is given by $y(t_0) = \hat{\alpha} + \hat{\beta}t_0$ with the estimated variance $\hat{V}(y(t_0)) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(t_0 - \bar{t})^2}{S_{TT}} \right]$.

The 95% lower confidence limit of the mean degradation line is given by $L(t_0) = \hat{\alpha} + \hat{\beta}t_0 - t_{.95}(n-2) \left[\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(t_0 - \bar{t})^2}{S_{TT}} \right] \right]^{\frac{1}{2}}$ where $t_{.95}(n-2)$ is the 95th percentile of the central t distribution with $n-2$ degrees of freedom. Consequently, the expiration dating can be obtained as the minimum of the two solutions of t_0 from the following quadratic equation:

$$\left[\eta - (\hat{\alpha} + \hat{\beta}t_0) \right]^2 = t_{.95}(n-2) \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(t_0 - \bar{t})^2}{S_{TT}} \right] \quad (2)$$

where η is the specification limit. Therefore, the shelf-life estimator based on FDA's approach is the solution of Eq. (2), i.e.,

$$\hat{\theta}_F = \left[\hat{\beta}^2 - C_n \right]^{-1} \left[(\eta - \hat{\alpha})\hat{\beta} + B_n - \sqrt{\left[(\eta - \hat{\alpha})\hat{\beta} + B_n \right]^2 - (\hat{\beta}^2 - C_n) \left[(\eta - \hat{\alpha})^2 - A_n \right]} \right]$$

where

$$\begin{aligned} A_n &= \hat{\sigma}^2 t_{.95}^2 (n-2) \left(\frac{1}{n} + \frac{\bar{t}^2}{S_{TT}} \right), \\ B_n &= -\frac{\bar{t} \hat{\sigma}^2 t_{.95}^2 (n-2)}{S_{TT}}, \quad C_n = \frac{\hat{\sigma}^2 t_{.95}^2 (n-2)}{S_{TT}}. \end{aligned}$$

Direct Method

In the direct method, the shelf-life is determined by an approximate ($n \rightarrow \infty$ or $\sigma \rightarrow 0$) 95% lower confidence bound for the true shelf-life of $\theta = (\eta - \alpha)/\beta$, which is

$$\hat{\theta}_D = \frac{\eta - \hat{\alpha}}{\hat{\beta}} - \frac{\hat{\sigma} z_{.95}}{|\hat{\beta}|} \left(\frac{1}{n} + \frac{1}{S_{TT}} \left(\frac{\eta - \hat{\alpha}}{\hat{\beta}} - \bar{t} \right)^2 \right)^{\frac{1}{2}} \quad \text{where } z_{.95} \text{ is the 95th percentile of the standard normal distribution.}$$

This is based on the asymptotic theory that

$$\left(\frac{\eta - \hat{\alpha}}{\hat{\beta}} - \theta \right) \left(\frac{\hat{\sigma}}{|\hat{\beta}|} \sqrt{\frac{1}{n} + \frac{1}{S_{TT}} \left(\frac{\eta - \hat{\alpha}}{\hat{\beta}} - \bar{t} \right)^2} \right)^{-1} \rightarrow N(0,1) \text{ in law}$$

Thus, an asymptotic lower confidence bound for θ is defined as $\hat{\theta}_D = (\eta - \hat{\alpha}/\hat{\beta}) - 1.645s$ where $\hat{\theta}_D$ is the estimator of shelf-life from the direct method, and s is an estimated asymptotic standard deviation of $\eta - \hat{\alpha}/\hat{\beta}$.

Inverse Method

In this method, the shelf-life estimator is obtained based on the inverse regression model, which is defined as below:

$$t_j = \alpha^* + \beta^* y_j + e_j^*, \quad j = 1, \dots, n \quad (3)$$

The shelf-life estimator $\hat{\theta}_I$ based on the inverse method is the 95% lower confidence bound for $\alpha^* + \beta^*\eta$, because the true shelf-life is the t -value when the mean of $y = \eta$. By considering an ordinary linear regression model in Eq. (3), the following estimators are obtained $\hat{\alpha}^* = \bar{t} - (S_{TY}/S_{YY})\bar{y}$, $\hat{\beta}^* = S_{TY}/S_{YY}$.

Also, the estimator for the variance of $\hat{\alpha}^* + \hat{\beta}^*\eta$ is

$$\frac{1}{n-2} \left(S_{TT} - \frac{S_{TY}^2}{S_{YY}} \right) \left(\frac{1}{n} + \frac{(\eta - \bar{y})^2}{S_{YY}} \right) = \hat{\sigma}^2 \frac{S_{TT}}{S_{YY}} \left(\frac{1}{n} + \frac{(\eta - \bar{y})^2}{S_{YY}} \right)$$

Therefore, the shelf-life estimator based on the inverse method is

$$\hat{\theta}_l = \bar{t} + \frac{S_{TT}}{S_{rr}}(\eta - \bar{y}) - \hat{\sigma}_{t_{.95}}(n-2) \sqrt{\frac{S_{TT}}{S_{rr}} \left(\frac{1}{n} + \frac{(\eta - \bar{y})^2}{S_{rr}} \right)}.$$

These three methods were further compared based on their asymptotic biases and mean squared errors of the true shelf-life.^[8] It is concluded that the FDA's method is slightly better when σ is large, and the direct method is better when σ is small according to the bias, mean squared error of $\hat{\theta}_p$, $\hat{\theta}_l$, and $\hat{\theta}_r$, and coverage probability from the simulation result. The coverage probability is the proportion of times the confidence intervals cover the true shelf-life in a simulation study. In general, the inverse method is too conservative unless σ is very small.

Practical Issues

Batch-to-Batch Similarity

The FDA guideline requires that a stability study be conducted by testing at least three batches, and the batch similarity be examined by the preliminary test to verify the equality of slopes and intercepts of all batches. The guideline also indicates that the preliminary tests for the equality of slopes and equality of intercepts be performed at the 0.25 level of significance, which was suggested by Bancroft.^[10] If the preliminary tests for the equality of slopes and the equality of intercepts are not rejected at the 0.25 level of significance, then the common slope and intercept are estimated on the pooled stability data from all batches. Note that as pointed out by Ruberg and Stegeman,^[11] the use of the significance level of 0.25 for assessing batch-to-batch similarity for poolability (or combinability) might penalize good batches in drug shelf-life estimation. On the other hand, if preliminary tests are rejected at the 0.25 level of significance, the minimum of shelf lives obtained from individual batches is considered as the drug shelf-life.

The disadvantage of the minimum approach is too conservative and does not consider the batch-to-batch variation. Chow and Shao^[12,13] pointed out that the presence of batch-to-batch variation has an impact on the determination of shelf-life. Therefore, it is of interest to estimate the batch-to-batch variation. Chow and Wang^[14] indicated that the estimated batch variation can be used to determine the homogeneity of degradation rates for pooling data from a batch, and evaluate whether or not a desired labeled shelf-life can be reached. Because the primary goal of a stability study is to apply the established shelf-life to all future production batches, several approaches for the determination of shelf-life with random batches have been proposed.

Shao and Chow^[15] proposed a method that ensures that future batches have longer shelf lives than the labeled shelf-life with a given statistical assurance. According to Sun et al.,^[16] the individual expiration dating period proposed

in Shao and Chow^[15] is an asymptotically normally distributed random variable.

Two-Phase Shelf-Life Estimation

In practice, the drug products may be stored under different temperatures in stability testing. For example, the drug product may require (i) 24 months at -20°C followed by either 1 month at 5°C and 2 days at 25°C , or (ii) 24 months at -20°C followed by 2 weeks at 5°C and 1 day at 25°C .^[17] These types of drug products are called the frozen drug products. In stability study, the drug shelf-life of frozen drug products is determined by a two-phase stability study; that is, each phase has its own stability testing based on the different temperatures. Mellon^[18] suggested analyzing each phase separately to obtain a combined expiration dating and provided a number of approaches for the analysis of frozen drug products. However, this method does not account for the fact that the expiration dating at the second phase would depend on the first phase. Using the changing point which occurs at the time point of the 95% lower confidence limit of the mean degradation curve at the first phase that intersects the approved lower specification limit as the expiration dating was considered by Shaban^[19] and Hinkley.^[20] Furthermore, Shao and Chow^[21] proposed a method by using a two-phase regression analysis to determine the drug shelf-life of frozen drug products.

Shelf-Life Estimation with Discrete Responses

In the current stability guidelines, the expiration dating period is determined based on the mean degradation curve of continuous response (e.g., potency) over time. However, both the FDA and ICH stability guidelines also require that some other discrete characteristics such as color and hardness should also be considered when performing stability studies, but no written guideline to determine the shelf-life for discrete responses. Chow and Shao^[22] proposed two shelf-life estimators that are applicable to all future batches when the stability data are discrete. The first method is suitable for binary data without batch-to-batch variation, and the second method is useful for the random batches. Both methods are summarized below.

Binary Data Without Batch-to-Batch Variation

Let y_{ij} be the binary response of the i th batch at some time points n_i , assuming y_{ij} 's are independent. Then the following model in Eq. 4 is followed:

$$\begin{aligned} E(y_{ij}) &= \frac{e^{\beta^T x_{ij}}}{1 + e^{\beta^T x_{ij}}}, \quad i = 1, \dots, k, \\ \text{Var}(y_{ij}) &= \frac{e^{\beta^T x_{ij}}}{1 + e^{\beta^T x_{ij}}} \left(1 - \frac{e^{\beta^T x_{ij}}}{1 + e^{\beta^T x_{ij}}} \right), \quad j = 1, \dots, n_i \end{aligned} \quad (4)$$

where x_{ij} is a p -vector of covariates, such as time points or bottle size; β^T is the transpose of β , which is a p -vector of unknown parameters. The proposed shelf-life estimator is $\hat{t}^* = \inf\left(t : L(t) \leq \frac{1+e^\eta}{e^\eta}\right)$ which satisfies $P_Y(\hat{t}^* > t^*) \leq 5\%$ where $\hat{t}^*$ is an approximate 95% lower confidence bound for the true shelf-life t^* , P_Y is the approximate probability which is determined based on the distribution of $\hat{\beta}$, and $L(t) = \hat{\beta}^T x(t) - z_{0.95} \times \sqrt{x(t)^T \hat{V} x(t)}$ where $z_{0.95}$ is the 95th percentile of the standard normal distribution; $\hat{\beta}$ is obtained by solving the following equation $\sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij} \left(y_{ij} - \frac{e^{\beta^T x_{ij}}}{1 + e^{\beta^T x_{ij}}} \right) = 0$ and $\hat{V}$ is estimated by

$$\hat{V} = \left[\sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij} x_{ij}^T \frac{e^{\hat{\beta}^T x_{ij}}}{1 + e^{\hat{\beta}^T x_{ij}}} \left(1 - \frac{e^{\hat{\beta}^T x_{ij}}}{1 + e^{\hat{\beta}^T x_{ij}}} \right) \right]^{-1}$$

Random Batches

Assuming k batches be the random sample, then β_i 's in Eq. (4) are random effects. Therefore, a mixed-effect model with conditional expectation and variance of y_{ij} given the random effect β_i is considered. The true shelf-life is a random variable, which is given by $t_{\text{future}}^* = \inf(t : \beta_{\text{future}}^T x(t) \leq 1 + e^\eta / e^\eta)$. The proposed shelf-life estimator for t_{future}^* is $\hat{t}_{\text{future}}^* = \inf(t : L(t) \leq 1 + e^\eta / e^\eta)$ where $L(t) = \hat{\beta}^T x(t) - \rho_{0.95}(k) \sqrt{v(t)/k}$ with

$$v(t) = x(t)^T \left[\frac{1}{k-1} \sum_{i=1}^k (\hat{\beta}_i - \hat{\beta})(\hat{\beta}_i - \hat{\beta})^T \right] x(t)$$

$\hat{\beta}_i$ is the maximum likelihood estimator of β_i , and $\hat{\beta}$ is the average of $\hat{\beta}_i$ for i from 1 to k . Also, $\rho_{0.95}(k)$ satisfies $\int_0^1 P_r\{T_k(u) \leq \rho_{0.95}(k)\} du = .95$ where $T_k(u)$ is a random variable that follows the noncentral t -distribution with $k-1$ degrees of freedom and the noncentrality parameter of $\sqrt{k}\Phi^{-1}(1-u)$, and Φ is the standard normal distribution function.

Multiple Components

The statistical methods for the determination of a single shelf-life of drug products with a single active ingredient are well established. However, many drug products contain more than one active ingredient in practice. The examples include Levite® (oral contraceptive drug), which contains two active ingredients, levonorgestrel and ethinylestradiol, and Premarin® (conjugated estrogens), which contains more than three active ingredients, including estrone, equilin, and 17 α -dihydroequilin. Furthermore, the multiple ingredients are often applied to herbal medicines in preclinical and clinical development.^[23] It is estimated

that as many as a quarter of all drugs come from natural sources, mostly plants.^[24] Also, it is common that one dose of traditional Chinese herbal medicine may contain multiple active ingredients, such as Menopause®, which contains black cohosh, wheat berries, Jujube fruit, licorice root, and kava. Those multiple active ingredients in this formula work together to improve the balance of progesterin and estrogen to relieve menopause symptoms.

Nevertheless, there is no discussion regarding the statistical methods for the determination of the shelf-life of drug products with multiple active ingredients in both the ICH and the FDA stability guidelines. For a drug product with multiple ingredients (components), the usual practice for the estimation of the expiration dating period is to consider the drug product as a whole rather than as individual components in terms of the percentages of the label claim of the drug product. It is also recognized that the degradations of individual components may vary from component to component. To account for the practical issues that may be encountered in the estimation of the drug shelf-life of drug products with multiple components, the naive approach is to use the minimum of expiration dating periods obtained from individual components being considered as the labeled expiration dating. As discussed before, the minimum approach either is too conservative, which may lose the clinical interest in practice, or lacks statistical justification. In a similar manner, Chow and Shao^[25] proposed using simultaneous lower confidence bounds for estimation of drug products with multiple components. This method is also conservative. As an alternative, they also suggested using principal components in conjunction with factor analysis to identify the most active components and then performing stability analysis using multivariate analysis to account for the relationships among components. However, the relative merit and disadvantages need further investigation.

Additionally, Pong and Raghavarao^[26] proposed a statistical method for estimating the shelf-life of drug products with two active ingredients. The idea of the proposed method is to derive a shelf-life estimator that is the linear combination of individual shelf-life from two components. The optimal weights for each component are chosen based on the minimum mean squared error of the true shelf-life, which is sensitive to the variability and degradation rates from individual components. In this method, the minimum approach is a special case of the class of estimators under consideration. This method had been further extended for drug product with n (> 2) ingredients.^[27]

CONCLUSION

Stability studies are essential in pharmaceutical research and development, because they provide a certain assurance that the drug products will meet the standards of identity, strength, quality, and purity of drug characteristics before

the drug expiration date. However, there is no or little discussion regarding the determination of the drug shelf-life of drug products with multiple components in the FDA and ICH stability guidelines. Therefore, it would be of future interest to investigate the shelf-life of drug products with multiple components in the field of stability analysis.

REFERENCES

1. FDA. *Guideline for Submitting Documentation for the Stability of Human Drugs and Biologics*; Center for Drugs and Biologics, Office of Drug Research and Review, Food and Drug Administration: Rockville, MD, 1987.
2. ICH. *Stability Testing of New Drug Substances and Products. Tripartite International Conference on Harmonization Guideline*; 1993.
3. FDA. *Draft Guidance for Industry: Stability Testing of Drug Substances and Drug Products*; Center for Drugs and Biologics, Office of Drug Research and Review, Food and Drug Administration: Rockville, MD, 1998.
4. ICH Q1A. *Stability Testing of New Drug Substances and Products*. Tripartite International Conference on Harmonization Guideline; 2003.
5. Nf, Usp. *The United States Pharmacopeia XXIV and the National Formulary XIX*; United States Pharmacopeial Convention, Inc.: Rockville, MD, 2000.
6. Chow, S.C.; Pong, A. Current issues in regulatory requirements of drug stability. *J. Food Drug Anal* **1995**, *3* (2), 75–85.
7. Kumkumian, C. International council for harmonization stability guidelines: Food and drug administration regulatory perspective. *Drug Inf. J.* **1994**, *28*, 635–640.
8. Shao, J.; Chow, S.C. Drug shelf life estimation. *Stat. Sin.* **2001**, *11*, 737–745.
9. Chow, S.C. *Statistical Design and Analysis of Stability Studies*; Chapman & Hall/CRC, Taylor & Francis Group: Boca Raton, FL, 2007.
10. Bancroft, G.A. Analysis and inference for incompletely specified models involving the use of preliminary test(s) of significance. *Biometrics* **1964**, *20*, 427–442.
11. Ruberg, S.; Stegeman, J.W. Pooling data for stability studies: testing the equality of batch degradation slopes. *Biometrics* **1991**, *47*, 1059–1069.
12. Chow, S.C.; Shao, J. Test for batch-to-batch variation in stability analysis. *Stat. Med.* **1989**, *8*, 883–890.
13. Chow, S.C.; Shao, J. Estimating drug shelf-life with random batches. *Biometrics* **1991**, *47*, 1071–1079.
14. Chow, S.C.; Wang, S.-G. On the estimation of variance components in stability analysis. *Commun. Stat. Theory Methods* **1994**, *23* (1), 289–303.
15. Shao, J.; Chow, S.C. Statistical inference in stability analysis. *Biometrics* **1994**, *50*, 753–763.
16. Sun, Y.; Chow, S.C.; Li, G.; Chen, K.W. Assessing distribution of estimated drug shelf lives in stability analysis. *Biometrics* **1999**, *55*, 896–899.
17. Chow, S.C.; Liu, J.P. *Statistical Design and Analysis in Pharmaceutical Science*; Marcel Dekker, Inc: New York, 1995.
18. Mellon, J.I. Design and Analysis Aspects of Drug Stability Studies When the Product is Stored at Several Temperatures. *Presented at the 12th Annual Midwest Statistical Workshop*, Muncie, IN, 1991.
19. Shaban, S.A. Change point problem and two-phase regression: An annotated bibliography. *Int. Stat. Rev.* **1980**, *48*, 83–93.
20. Hinkley, D.V. Inference in two-phase regression. *J. Am. Stat. Assoc.* **1971**, *66*, 736–743.
21. Shao, J.; Chow, S.C. Two-phase shelf-life estimation. *Stat. Med.* **2001**, *20*, 1239–1248.
22. Chow, S.C.; Shao, J. Stability analysis with discrete response. *J. Biopharm. Stat.* **2003**, *13*, 451–462.
23. Habs, M. Herbal products: Advances in preclinical and clinical development. *Drug Inf. J.* **1999**, *33*, 993–1001.
24. Schwab, D. *The Article Entitled "The Odyssey" of the Star-Ledger Newspaper on May 2, 2001*; 2001—New Jersey.
25. Chow, S.C.; Shao, J. *Statistics in Drug Research—Methodology and Recent Development*; Marcel Dekker, Inc: New York, 2002.
26. Pong, A.; Raghavarao, D. Shelf Life Estimation for Drug Products with Two Components. *Proceedings of the Biopharmaceutical Section of the American Statistical Association*; 2001.
27. Pong, A. Comparing Designs for Stability Studies and Shelf Life Estimation for Drug Product with Multiple Components. Ph. D. Thesis. Temple University: Philadelphia, PA, 2001.

Exploratory Factor Analysis

Joseph C. Cappelleri and Robert A. Gerber

Pfizer Inc., New London, Connecticut, U.S.A.

Abstract

In exploratory factor analysis, there is initial uncertainty as to the number of factors being measured, and the results of the analysis are used to help solve for the number of factors. This entry introduces basic concepts of the common factor model and highlights a few major steps in conducting an exploratory factor analysis.

INTRODUCTION

Factor analysis is a multivariate method concerned with detecting and analyzing patterns based on the correlations among quantitative variables. In factor analysis, a set of measured variables is reduced into a smaller, more manageable set of unobservable factors that underlie the measured variables. The objective is to identify the number and the nature of the factors that are responsible for covariation in the data. For example, the Hospital Anxiety and Depression Scale (HADS) is a widely used questionnaire with 14 variables (or items), each measured on a four-point ordinal scale.^[1] A factor analysis revealed that, although the questionnaire contained 14 items, it was really just measuring two underlying factors (or constructs): anxiety and depression.^[2,3] Seven of the 14 questions measured anxiety, whereas the other seven measured depression.

Exploratory and confirmatory factor analyses are two major approaches to factor analysis. In exploratory factor analysis, there is initial uncertainty as to the number of factors being measured, and the results of the analysis are used to help solve for the number of factors. Exploratory factor analysis is suitable for generating hypotheses about the structure of the data. It can be used to lend structure and validity to the data. In addition, exploratory factor analysis can further refine an instrument by revealing what items may be dropped from the questionnaire because they contribute little to the presumed underlying factors.

Although exploratory factor analysis explores the patterns in the correlations of items (or variables), confirmatory factor analysis tests whether the correlations conform to the anticipated or expected scale structure. In confirmatory factor analysis, there must be not only a pre-specified number of unobserved factors (latent variables) being assessed, but also a theoretical framework of which observable (manifest) variables are related and not related to which factors. A more important distinction is that confirmatory factor analysis, unlike exploratory factor analysis, tests for a specified relationship between latent

factors and, in doing so, allows for testing hypotheses on what factors have effects on other factors.

In this entry, attention is restricted to exploratory factor analysis. Section “Basic Concepts” introduces basic concepts of the common factor model. Section “Comparison With Principal Component Analysis” compares exploratory factor analysis with principal component analysis, a related variable reduction technique. Section “Steps in Conducting an Exploratory Factor Analysis” highlights and illustrates a few major steps in conducting an exploratory factor analysis. Section “Further Reading” provides sources for further reading.

BASIC CONCEPTS

Common and Unique Factors

The matrix of correlations between observed variables (items) provides the basic building block for factor analysis. Consider a questionnaire with six items, each rated on a seven-point ordinal scale. The pattern of correlations between items may indicate, for example, that the first three questions happen to be strongly correlated with one another but essentially uncorrelated with the last three questions. Similarly, the last three questions may be strongly correlated with one another but essentially uncorrelated with the first three questions.

This grouping suggests two sets of variables. The three questions in each set are correlated with one another because they are all influenced by the same underlying factor. A factor, a hypothetical construct, is an unobservable (latent) variable that is believed to influence certain observable (manifest) variables that can be measured directly. A common factor is a factor that influences more than one observable variable and is “common” in the sense that more than one variable has it in common. In the current scenario, the first three observable variables are correlated because they have a factor in common; the

last three observable variables are correlated because they share a distinct, common factor.

The two common factors are not the only factors that influence the observed variables. A second type of factor—a unique factor—influences only one observed variable. A unique factor represents all of the independent factors, including the error component specific to a given variable, that are unique to that single variable. Each variable has its own unique factor.

Orthogonal versus Oblique Rotation

An orthogonal rotation is one in which the (common) factors are treated as uncorrelated and hence orthogonal. An oblique rotation, on the other hand, is one in which the factors are taken as correlated or associated. No implication is made, though, that one factor is expected to have a causal effect on the other.

Factor Loadings

Factor loadings are coefficients that indicate the importance of a variable to each factor. These coefficients are important because they signify the nature of the variables that most strongly relate to a factor; the nature of the variables helps to capture the nature and meaning of a factor.

Factor loadings appear in either a pattern matrix or a structure matrix. Factor loadings in a pattern matrix are essentially standardized regression coefficients. In this case the factors and observed variables are standardized to have unit variances, where the observable variables are dependent variables and the factors are explanatory variables. Pattern loadings may be therefore thought as optimal linear weights in predicting the variables from the factors. The factor pattern reflects the unique contribution that each factor makes to the variance of each variable.

Factor loadings in a structural matrix, on the other hand, represent the product-moment correlation between an observed variable and an underlying factor. Factor loadings, therefore, appear as standardized regression weights from the pattern matrix and correlation coefficients from the structural matrix. For an orthogonal factor analysis, but not for an oblique factor analysis, factor loadings from pattern and structural matrices are equivalent.

Model

In a factor analytic model, each observed variable can be viewed as a linear combination of the underlying factors, both common and unique. Suppose there are six observed variables and two retained common factors in a factor model. Each of the six observed variables can be viewed as a weighted sum of the two underlying factors, plus a variable-specific unique factor. For example, let Y_i be a subject's score on the first observed variable, F_1 and F_2 be

a subject's score on the two common factors, and U_i be the unique factor associated with Y_i . Algebraically,

$$Y_i = b_1(F_1) + b_2(F_2) + d_i(U_i)$$

where b_1 and b_2 represent the regression coefficients (weights) for the underlying common factors F_1 and F_2 , respectively, and d_i represents the regression weight for the unique factor associated with Y_i . A subject's score on Y_i is therefore determined by multiplying her scores on the underlying factors by the appropriate regression weights and then summing the results. The regression weights of the common factors are standardized and come from the factor pattern matrix. Generally, a different set of factor weights, and thus a different equation, is expected for each observed variable. A set of q common factors will have q corresponding regression weights ($b_1, b_2, \dots, b_{q-1}, b_q$).

An estimated factor score is an estimate of a subject's standing on a particular factor. Factor scores are estimated by creating linear composites of the observed variables, that is, by adding and optimally weighting each subject's scores on the observed variables. A score can be assigned to a subject on a given factor to indicate where he stands on each factor. If a factor analysis of 14 items suggests two factors, for instance, a subject will have an estimated factor score on one factor and another estimated factor score on the second factor. These two factor scores (instead of the larger set of 14 items) could then be used as predictor variables or dependent variables in subsequent analyses. Alternatively, instead of an estimated factor score, a factor-based scale score for an individual can be created by simply taking, say, an unweighted average of an individual's responses to the observed items belonging to a particular scale.

Communality and Unique Components

A communality is pegged to a given observed variable and refers to the variance in an observed variable that is accounted for by the set of common factors. Suppose a set of two common factors is retained. A communality of 0.8 for a variable, say the first variable, means that 80% of the variance for that variable is accounted by the two common factors. A high communality for a variable means the variable is strongly influenced by at least one of the factors. The communality is computed by squaring that variable's regression weights, or its factor pattern loadings, of all the retained common factors and then summing these squares.

In contrast to the communality, the unique component refers to the proportion of variance in a given observed variable that is not accounted for by the common factors. Once a communality of a variable is computed, it is easy to calculate the unique component: simply subtract the communality from one. If the communality of the first variable (Y_1) is 0.8, then 20% of the variance in Y_1 is accounted for by its unique factor, U_1 . The square root of the unique

component (i.e., the square root of 0.20), which is 0.45, becomes the weight given to U_1 (i.e., $d_1 = 0.45$).

COMPARISON WITH PRINCIPAL COMPONENT ANALYSIS

Exploratory factor analysis and principal components analysis are similar in some ways and, as a result, they are often mistaken for one another. Both exploratory factor analysis and principal component analysis are variable reduction procedures, reducing a number of variables to a smaller number. But the two procedures are generally different and not conceptually identical.

In factor analysis the observed variables are linear combinations of the underlying factors. In principal components analysis the principal components are linear combinations of the observed variables.

Factor analysis is better suited than principal component analysis to identify the number and nature of the latent factors that are responsible for covariation in the data set. In factor analysis, factors are extracted to account only for the common variance (variance accounted for by the common factors), while the unique variance remains unanalyzed. This extraction is accomplished by analyzing an adjusted correlation matrix with correlations between the observed variables off the diagonal of the matrix and communality estimates on the diagonal.

In principal component analysis, on the other hand, components are extracted to account for the total variance in the data set, both common and unique, and not just the common variance. This extraction is accomplished by analyzing an unadjusted correlation matrix with a correlation matrix having ones on the diagonal. Because principal component analysis makes no attempt to separate the common component from the unique component of each variable's variance, and factor analysis does, factor analysis is more appropriate than principal component analysis in identifying the factor structure of the data.

STEPS IN CONDUCTING AN EXPLORATORY FACTOR ANALYSIS

Factor analysis is typically conducted in a sequence of steps. A sequence of major steps is illustrated with the Patient Satisfaction with Insulin Therapy (PSIT) Questionnaire (Table 1). The PSIT is based on 15 items with each item based on a five-point Likert scale ranging from strongly agree to strongly disagree. Responses to each item on the questionnaire were analyzed so that a higher item score indicated a favorable attitude.

The questionnaire was self-administered at baseline and 12-week follow-up to subjects with type 1 diabetes recruited into a 12-week parallel study of an inhaled insulin

Table 1 The Patient Satisfaction with Insulin Therapy Questionnaire: Baseline Descriptive Statistics in a Trial ($n = 69$)^a

Items ^b	Standard	
	Mean	Deviation
1. I find it easy to take insulin the way I take it now.	3.58	1.28
2. I have no discomfort taking insulin the way I take it now.	3.07	1.38
3. I find it convenient to take insulin the way I take it now.	3.09	1.36
4. I am self-conscious about taking insulin away from home.	3.20	1.37
5. I find it easy to take all the doses of insulin my doctor recommends.	3.90	1.18
6. I find the time it takes for each dosing acceptable.	3.87	1.21
7. I find that my eating schedule can be flexible with few problems.	3.28	1.38
8. I prefer to stay home rather than take insulin away from home.	3.71	1.45
9. I do not mind measuring my blood glucose before each meal.	3.74	1.23
10. I feel good on my current insulin treatment schedule.	3.72	1.04
11. I find it difficult to take every dose of insulin my doctor recommends.	3.90	1.19
12. I find it difficult to take insulin away from home.	3.71	1.41
13. I would find it difficult to take insulin four times a day.	2.86	1.41
14. I find it easy to travel for a few days and take all my doses of insulin.	3.74	1.18
15. Overall, I am satisfied with my current way of taking insulin.	3.40	1.21

^aA higher item score (range 1–5) indicated a more favorable attitude. Hence, items 1, 2, 3, 5, 6, 7, 9, 10, 14, and 15 were reversed scored and analyzed so that 1 = strongly disagree, 2 = slightly disagree, 3 = neither agree nor disagree, 4 = slightly agree, 5 = strongly agree. Items 4, 8, 11, 12, and 13 were analyzed so that 1 = strongly agree, 2 = slightly agree, 3 = neither agree nor disagree, 4 = slightly disagree, 5 = strongly disagree.

^bCopyright © 1996 Pfizer Inc. All rights reserved (Patient Satisfaction with Insulin Therapy Questionnaire).

regimen versus a conventional subcutaneously injected insulin regimen.^[4] In this entry, three major steps are illustrated for the factor analysis of the baseline responses of 69 subjects (Table 1).^[5] These steps are intended to help guide the conduct of an exploratory factor analysis.

Step 1: Determine the Number of Retained Factors

Three recommendations are offered in determining the number of factors to retain. One recommendation

is to use the scree test.^[6] The scree test is a rule-of-thumb criterion and involves the plot of the eigenvalue (i.e., the amount of variance that is accounted for by a given factor) associated with each factor (Fig. 1). The objective of a scree plot is to look for a break between factors with relatively large eigenvalues and those with smaller eigenvalues. Factors that appear before the break are taken to be meaningful and retained for rotation.^[6] For the data on the PSIT questionnaire, the scree plot (Fig. 1) depicted an abrupt break or discontinuity before factor 3, suggesting that only the first two factors were meaningful to be retained. A straight line can be drawn from factor 3 to factor 15 but not from factor 2 or factor 1 to factor 15.

A second recommendation in deciding on the number of factors involves retaining a factor if it accounts for a certain percentage of the variance in the data set. Although no fixed percentage is standard, we prespecified and required that at least 10% of the variance be explained by a retained factor in the PSIT data. The first factor accounted for 66% of the variance, and the second factor accounted for an additional 20%. The third factor accounted only for an additional 7% of the variance, and subsequent factors independently accounted for progressively lower percentages of variance.

The final recommendation when solving for the number of factors embraces criteria for interpretability: interpreting the substantive meaning of the retained factors and verifying that the interpretation makes sense about the subject matter. Four criteria for interpretability are given in Step 3: Interpret the Factors (below) and concurred with a two-factor solution for the example data.

Therefore, a two-factor solution was chosen based on the scree test, the percentage of variance explained by each factor, and interpretability criteria.

Step 2: Rotate the Chosen Factors

In common factor analysis, the observed items are viewed as linear combinations of the factors, and the elements of the rotated factor pattern are regression coefficients (or regression weights) associated with each factor in the prediction of an item score. When all of the items and factors are standardized to have a mean of 0 and standard deviation of 1, the regression weights become standardized regression coefficients. Relative to the initial and unrotated solution, which tends to be ambiguous, rotating (transforming) factors fosters a simpler structure of the data that leads to a more interpretable pattern.

For the PSIT data, the rotated factor matrix of pattern loadings with a promax rotation of two factors was used (Table 2). A promax rotation allows the retained factors to be oblique (correlated). The pattern loadings are essentially standardized regression coefficients comparable with those obtained in multiple regression. Their absolute values reflect the unique contribution that each factor makes to the variance of the observed item. The rotated factor pattern matrix was used to determine which groups of variables are measuring a given factor and to determine the meaning of each factor.

Step 3: Interpret the Factors

Four criteria were applied to judge interpretability and overall results, including the number of factors retained. First, at least three items should load on each retained factor. Second, the items that loaded on a given factor should share a common conceptual meaning. Third, the items that loaded on different factors should measure different conceptual meanings. Finally, the rotated factor pattern should demonstrate simple structure in that (i) most of the items should have high pattern loadings (say, $\geq |0.40|$) on one

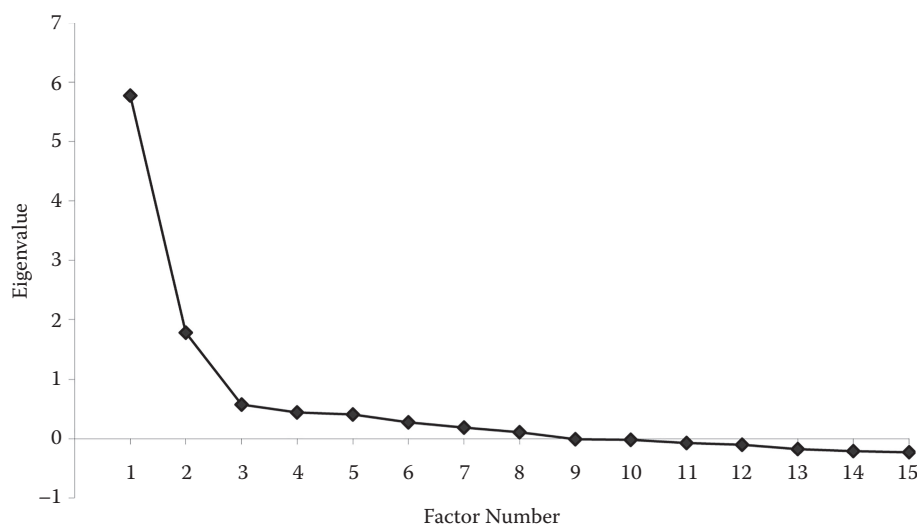


Fig. 1 Scree plot showing the amount of variance accounted for by each factor. Note: Because the cumulative percentage of variance exceeds 100% at some points in the analysis, it was necessary mathematically that some trivial factors have negative eigenvalues

Table 2 Rotated Factor Pattern: Standardized Regression Coefficients^a

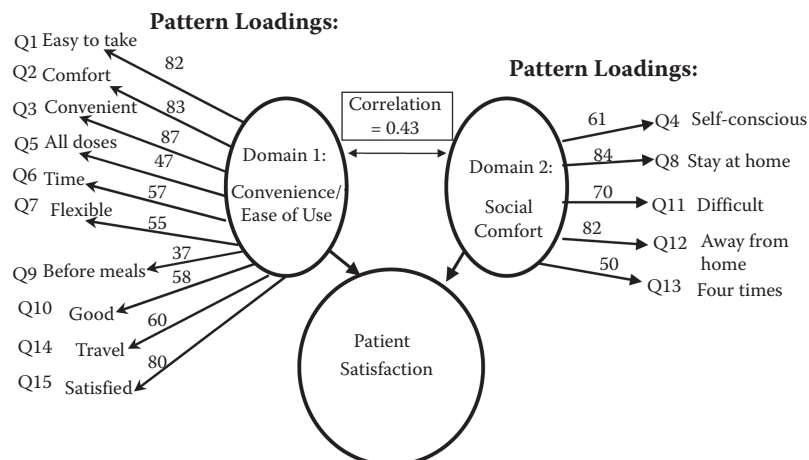
Items ^b	Factor 1	Factor 2
1. I find it easy to take insulin the way I take it now.	82 ^c	9
2. I have no discomfort taking insulin the way I take it now.	83 ^c	-15
3. I find it convenient to take insulin the way I take it now.	87 ^c	-20
4. I am self-conscious about taking insulin away from home.	3	61 ^d
5. I find it easy to take all the doses of insulin my doctor recommends.	47 ^c	34
6. I find the time it takes for each dosing acceptable.	57 ^c	25
7. I find that my eating schedule can be flexible with few problems.	55 ^c	16
8. I prefer to stay home rather than take insulin away from home.	-19	84 ^d
9. I do not mind measuring my blood glucose before each meal.	37 ^c	-7
10. I feel good on my current insulin treatment schedule.	58 ^c	-4
11. I find it difficult to take every dose of insulin my doctor recommends.	-1	70 ^d
12. I find it difficult to take insulin away from home.	9	82 ^d
13. I would find it difficult to take insulin four times a day.	5	50 ^d
14. I find it easy to travel for a few days and take all my doses of insulin.	60 ^c	21
15. Overall, I am satisfied with my current way of taking insulin.	80 ^c	10

^aMultiplied by 100 and rounded.^bCopyright © 1996 Pfizer Inc. All rights reserved (Patient Satisfaction with Insulin Therapy Questionnaire).^cThese items constitute Factor 1.^dThese items constitute Factor 2.

factor and low loadings on the other factor and (ii) each factor should have high loadings for some items and low loadings for the others. All four criteria were met in this factor analysis with two factors (Table 2).

Factor 1 made a large and unique contribution to the variance of the following 10 items: easy to take, comfort, convenient, all doses, time, flexible, before meals, feel good, easy to travel, and satisfied—items 1, 2, 3, 5, 6, 7, 9, 10, 14, and 15 in Table 2. Because these items relate to convenience and ease of use of insulin therapy, Factor 1 was labeled “convenience/ease of use.” Factor 2 made a unique and noticeable contribution to the variance of the following five items: self-conscious, stay at home, difficult to take dose, away from home, and insulin intake four times—items 4, 8, 11, 12, and 13 in Table 2. Because these items relate to social stigma and use of insulin delivery away from home, factor 2 was labeled “social comfort.” The inter-factor correlation of 0.43 revealed a moderate positive correlation between the two factors. Higher levels of convenience and ease of use were associated with higher levels of social comfort. Fig. 2 summarizes and depicts the results of the factor analysis.

Further support for a two-factor solution and its interpretation came from the simple structure of the semipartial correlations from the reference structure matrix and the simple bivariate correlations from the factor structure matrix (data not shown). A semi-partial correlation represents a correlation between a variable and a common factor, after removing the effects of other common factors. A simple bivariate correlation represents a correlation between an item and a common factor. The three-factor solution, on the other hand, gave a vague interpretation on how certain items related to particular factors. In addition, its semipartial correlations were more ambiguous and less compelling in demonstrating simple structure compared with the semipartial correlations from the two-factor solution.

**Fig. 2** Depiction of factor analysis results

FURTHER READING

Detailed descriptions of exploratory factor analysis, as well as related techniques, abound elsewhere.^[2,3,7–14] In particular, Gorsuch^[10] has been a leading text that covers factor analysis in depth while avoiding detailed mathematics. Ki and Chow^[12] considered factor analysis in conjunction with principal component analysis for identifying factors in a quality-of-life questionnaire. Confirmatory factor analysis, not covered in this entry, is a special case of a broad class of models called structural equation models. Many sources are available on confirmatory factor analysis and structural equation models.^[11,15–24] In particular, Bollen^[15] has been a leading text on structural equations with latent variables. Exploratory factor analysis and confirmatory factor analysis on different data sets have appeared within the same article.^[25–27] Standard statistical software [e.g., SAS^[11]] is available for exploratory and confirmatory factor analyses. Specialized software for structural equation models include AMOS,^[28] EQS,^[29] LISREL,^[30] and MPLUS.^[31]

CONCLUSIONS

Exploratory factor analysis is a multivariate, data-reduction method that identifies the number and the nature of the factors that are responsible for covariation in the data. The objective is to arrive at a manageable and meaningful set of unobservable factors that underlie the measured variables. In exploratory factor analysis, there is initial uncertainty as to the number of factors being measured. The results of exploratory factor analysis are used to help solve for the number of factors, generating hypotheses about the structure of the data. Exploratory factor analysis can be used to lend structure and validity to the data. In addition, it can further refine an instrument by revealing what items may be dropped from the questionnaire because they contribute little to the presumed underlying factors.

One recommendation on how many factors to retain is to use a scree test. In addition, other criteria can be used to evaluate the number of factors and their content for suitability. Among these criteria are items that load on a given factor should have a shared conceptual meaning and items that have high standardized pattern loadings (≥ 0.40 in absolute value) on one factor should also have low loadings on the other factors.

This entry introduced the basic concepts of the common factor model. Exploratory factor analysis was compared and contrasted with principal component analysis, a related variable reduction technique. Major steps of exploratory factor analysis were illustrated with an example on patient satisfaction for individuals with diabetes. Sources for further reading were provided.

REFERENCES

1. Zigmond, A.S.; Snaith, R.P. The hospital anxiety and depression scale. *Acta. Psychiatrica. Scand* **1983**, *67*, 361–370.
2. Fayers, P.M.; Hays, R.D., Eds.; *Assessing Quality of Life in Clinical Trials: Methods and Practice*, 2nd Ed.; Oxford University Press: New York, 2005.
3. Fayers, P.M.; Machin, D. *Quality of Life: Assessment, Analysis and Interpretation of Patient-Reported Outcomes*, 3rd Ed. John Wiley & Sons, Ltd.: Chichester, 2007.
4. Gerber, R.A.; Cappelleri, J.C.; Kourides, I.A.; Gelfand, R.A. Treatment satisfaction with inhaled insulin in patients with type 1 diabetes: A randomized controlled trial. *Diabetes Care* **2001**, *24*, 1556–1559.
5. Cappelleri, J.C.; Gerber, R.A.; Kourides, I.A.; Gelfand, R.A. Development and factor analysis of a questionnaire to measure patient satisfaction with injected and inhaled insulin for type 1 diabetes. *Diabetes Care* **2000**, *23*, 1799–1803.
6. Cattell, R.B. The scree test for the number of factors. *Multivariate Behav. Res.* **1966**, *1*, 245–276.
7. Rummel, R.J. *Applied Factor Analysis*; Northwestern University Press: Evanston, IL, 1970.
8. Kim, J.O.; Mueller, C.W. *Introduction to Factor Analysis: What It Is and How to Fix It*; Sage Publications, Inc.: Beverly Hills, CA, 1978.
9. Kim, J.O.; Mueller, C.W. *Factor Analysis: Statistical Methods and Practical Issues*; Sage Publications, Inc.: Beverly Hills, CA, 1978.
10. Gorsuch, R.L. *Factor Analysis*, 2nd Ed.; Lawrence Erlbaum Associates: Hillsdale, NJ, 1983.
11. Hatcher, L. *A Step-by-Step Approach to Using the SAS® System for Factor Analysis and Structural Equation Modeling*; SAS Institute Inc.: Cary, NC, 1994.
12. Ki, F.; Chow, S.C. Statistical justification for the use of composite scores in quality of life assessment. *Drug Inf. J.* **1995**, *29*, 715–727.
13. Pett, M.A.; Lackey, N.R.; Sullivan, J.L. *Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research*; Sage Publications: Thousand Oaks, CA, 2003.
14. Thompson, B. *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*; American Psychological Association: Washington, DC, 2004.
15. Bollen, K.A. *Structural Equations with Latent Variables*; John Wiley & Sons: New York, 1989.
16. Bentler, P.M.; Stein, J.A. Structural equation models in medical research. *Stat. Methods Med. Res* **1992**, *1*, 159–181.
17. Long, J.S. *Confirmatory Factor Analysis: A Preface to LISREL*; Sage Publications, Inc.: Beverly Hills, CA, 1983.
18. Long, J.S. *Covariance Structure Models: An Introduction to LISREL*; Sage Publications, Inc.: Beverly Hills, CA, 1983.
19. Hoyle, R.H., Ed. *Structural Equation Modeling: Concepts, Issues and Applications*; Sage Publications, Inc.: Thousand Oaks, CA, 1995.
20. Kaplan, D. *Structural Equation Modeling: Foundations and Extensions*; Sage Publications: Thousand Oaks, CA, 2000.
21. Kline, R. *Principles and Practice of Structural Equation Modeling*, 2nd Ed.; The Guilford Press: New York, 2005.
22. Brown, T.A. *Confirmatory Factor Analysis for Applied Research*; The Guilford Press: New York, 2006.

23. Cappelleri, J.C.; Bushmakina, A.G.; Baker, C.L.; Merikle, E.; Olufade, A.O.; Gilbert, D.G. Confirmatory factor analyses and reliability of the Modified Cigarette Evaluation Questionnaire. *Addict. Behav.* **2007**, *32*, 912–923.
24. Cappelleri, J.C.; Bushmakina, A.G.; Baker, C.L.; Merikle, E.; Olufade, A.O.; Gilbert, D.G. Multivariate framework of the Brief Questionnaire of Smoking Urges. *Drug Alcohol Depend.* **2007**, *90*, 234–242.
25. Cappelleri, J.C.; Bushmakina, A.G.; Baker, C.L.; Merikle, E.; Olufade, A.O.; Gilbert, D.G. Revealing the multidimensional framework of the Minnesota Nicotine Withdrawal Scale. *Curr. Med. Res. Opin.* **2005**, *21*, 749–760.
26. Cappelleri, J.C.; Bushmakina, A.G.; Gerber, R.A.; Kline-Leidy, N.; Sexton, C.C.; Lowe, M.R.; Karlsson, J. Psychometric analysis of the Three-Factor Eating Questionnaire-R21: results from a large diverse sample of obese and non-obese subjects. *Int. J. Obes.* **2009**, *33*, 611–620.
27. Cappelleri, J.C.; Bushmakina, A.G.; Gerber, R.A.; Kline-Leidy, N.; Sexton, C.C.; Karlsson, J.; Lowe, M.R. Evaluating the Power of Food Scale in obese subjects and a general sample of individuals: development and measurement properties. *Int. J. Obes.* **2009**, *33*, 913–922.
28. Arbuckle, J.L. *Amos 4.0 Users' Guide*; SmallWaters Corporation: Chicago, IL, 1999.
29. Bentler, P.M. *EQS Structural Equations Program Manual*; Multivariate Software, Inc.: Encino, CA, 1997.
30. Jöreskog, K.G.; Sörbom, D. *LISREL 8: User's Reference Guide*; Scientific Software International: Chicago, IL, 1996.
31. Muthén, L.K.; Muthén, B.O. *Mplus User's Guide*, 5th Ed.; Muthén & Muthén: Los Angeles, CA, 1998–2007.

Extra Variation Models

Jorge G. Morel

The Procter & Gamble Company, Winton Hill Business Center, Cincinnati, Ohio, U.S.A.

Nagaraj K. Neerchal

Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, Maryland, U.S.A.

Abstract

Presented in this entry is a description of the phenomenon known as extra variation or overdispersion, along with a review of the literature. Several testing models are detailed in an effort to give examples of this phenomenon.

INTRODUCTION

Consider a study where experimental units are clusters and each elemental unit within the cluster is classified into two mutually exclusive categories. If we assume elemental units are independent, the binomial model is appropriate for analyzing the data. But clustering introduces a lack of independence, and as a consequence the data usually exhibit larger variances than the variance permitted by the binomial model. This phenomenon is known as *extra variation* or *overdispersion*; in practice, seems to be the norm and not the exception.^[1] Extra variation is also observed for count data analyzed under the Poisson model when the residual variation obtained after a Poisson model may be greater than that can be attributed to the sampling variation assumed by the model. The phenomenon of *under dispersion* is less common, but it can be induced, for example, by stratified sampling. Under dispersion can also occur when there is competition for a positive response. As discussed in Stigler,^[2] examples of extra variation arising from a clustered population, can be traced in the literature as far back as the year 1876. Early remarks related with extra variation under the Poisson model are found in Student.^[3]

Examples of binomial (multinomial) models with extra variation may arise in the following: (i) teratological study of a genetic trait that is passed on with a certain probability to offspring of the same mother; (ii) reproductive toxicity experiments in which chemicals are administered to litters and responses are measured on individual fetuses; (iii) experiments in which individuals receiving treatments are allowed to interact with each other via support groups, as in stop-smoking programs or stop-drinking programs; (iv) household surveys in which respondents within a household (or a neighborhood) may be strongly influenced by one or two individuals; (v) in studies in which summary measures are reported after repeated application of a treatment or repeated use of a product. The phenomenon

of extra variation is widely recognized by researchers and practitioners alike. A large number of papers in the literature discuss examples, with real data illustrating the extra variation phenomenon and its consequences on erroneous inferences due to inflated type I error rates. Some relevant examples in the areas of teratology and toxicology are Williams,^[4–6] Kupper and Haseman,^[7] Haseman and Kupper,^[8] Kupper et al.,^[9] Shirley and Hickling,^[10] Paul,^[11] Pack,^[12] Lefkopoulou et al.,^[13] Carr and Portier^[14] Donner et al.,^[15] Morel and Neerchal.^[16] Other important examples of extra variation are found in Rosner^[17] and Donner^[18] in ophthalmology; Altham^[19] and Cohen^[20] in studies of hospitalized sibling pairs; Crowder^[21] in the analysis of seed germination; Heckman and Willis^[22] in connection with a panel study; Efron^[23] in the analysis of toxoplasmosis in 34 cities of El Salvador; Moore^[24] in the modeling of chromosome aberration in survivors of the atomic bomb of Hiroshima. This list of examples is by no means exhaustive. Real examples of overdispersion under the Poisson model are found in the analysis of number of tumors in rats, Gail et al.^[25] and Lawless,^[26] in a longitudinal clinical trial aimed to study the number of epileptic seizures, Breslow and Clayton,^[27] on the frequency of recurrence of bladder cancer, Dean and Balshaw.^[28] An assessment of efficacy of vitamin C in reducing child mortality in Indonesia is found in Zeger and Edelstein.^[29] In general, Poisson models with extra variation arise when any of the assumptions of the Poisson distribution is violated such as in the analysis of counts and rates of longitudinal studies.

OVERVIEW

The phenomenon of binomial outcomes with extra variation is usually characterized in terms of the first two moments. Suppose that T is a binomial random variable representing the number of successes out of m trials. Then,

$$E(T) = m\pi \quad (1)$$

where π represents the probability of success of each Bernoulli trial. Assume now that the Bernoulli trials $Y_1, Y_2, \dots, Y_m$ that encompass T are positively correlated, so the usual assumption of independence in the binomial model is violated. The case “positively correlated” corresponds to extra variation, which is the focus of this chapter. The occurrence of under dispersion is less common and corresponds to a negative correlation among the Bernoulli outcomes that comprise T . Thus, suppose that for two distinct Bernoulli outcomes the correlation is ρ^2 , $0 < \rho^2 < 1$. We use ρ^2 to indicate the correlation is non-negative. Then, it can be shown

$$\text{Var}(T) = m\pi(1 - \pi)\{1 + \rho^2(m - 1)\} \quad (2)$$

A random variable T with first two moments as in Eqs. (1) and (2) is said to have extra variation or overdispersion relative to the binomial model. When $\rho = 0$ Eq. (2) becomes the variance of a binomial distribution. Among sampling practitioners, the parameter ρ^2 represents the *intracluster correlation* and the factor $\{1 + \rho^2(m - 1)\}$ is known as *design effect*. The definition of extra variation can be extended to the case of multinomial distribution in a similar way. When Y is a Poisson-type of random variable, Y is said to have overdispersion if its first two moments are of the form

$$E(Y) = \mu, \quad \mu > 0 \quad (3)$$

and

$$\text{Var}(Y) = \mu + \kappa\mu^2 = \mu(1 + \kappa\mu), \kappa > 0 \quad (4)$$

The parameter κ is known as the dispersion parameter. Observe that when $\kappa = 0$, variance in Eq. (4) becomes that of the Poisson model. Dean^[30] provides an additional form of overdispersion by considering $\text{Var}(Y) = \mu(1 + \kappa)$.

There are several approaches for modeling data exhibiting extra variation. A widely used method is the one known as quasi-likelihood, where the binomial variance is inflated by a suitable constant. McCullagh and Nelder^[1] give details of the method. We discuss this method in “Quasi-likelihood Models.” Other methods include the use of likelihood models, as we will explain in “Likelihood Models.” We present a goodness-of-fit statistic to test the adequacy of likelihood models in “A Goodness-of-fit Test for Extra Variation Models.” The generalized estimating equations methodology proposed by Liang and Zeger^[31] and by Zeger and Liang,^[32] which is closely related to the quasi-likelihood method and generalized linear models, is reviewed in “Generalized Estimating Equations Models.” More recently, generalized models that include random effects to model the heterogeneity between individuals within a cluster have been proposed. Some references for

this method are McGilchrist,^[33] Drum and McCullagh,^[34] Breslow and Clayton,^[35] Karim and Zeger,^[36] Zeger and Karim,^[37] Goldstein,^[38] Schall,^[39] and Zeger et al.^[40] Random effect models are reviewed in “Random Effect Models.” We will present several examples related to preclinical and clinical experiments from biopharmaceutical studies throughout the remainder of this chapter.

QUASI-LIKELIHOOD MODELS

In 1974 Wedderburn^[41] proposed the use of quasi-likelihood models. A quasi-likelihood function is defined without the knowledge of the likelihood function, but through its relation between the mean and the variance of each observation for which the quasi-likelihood function is being constructed. The structure of the first derivative of the quasi-likelihood function (without knowing the function from which this derivative is obtained from) resembles the structure of the score function, that is, the derivative of the logarithm of the likelihood function. Parameter estimates are obtained by solving the score function, often iteratively by methods such as Gauss–Newton. In order to understand quasi-likelihood models it is necessary to have some familiarity with *generalized linear models* (GLM). A quick introduction to GLM can be presented via the popular logistic and Poisson regression models. Let Y_j , $j = 1, 2, \dots, n$, represent n independent random variables belonging to a distribution in the exponential family such that $E(Y_j) = \mu_j$. Because the Y_j 's belong to the exponential family, it is possible to prove

$$\text{Var}(Y_j) = \phi v(\mu_j),$$

where the multiplier ϕ is known as the *dispersion* (or *scale*) parameter. In the case of the normal distribution, the parameter ϕ corresponds to the variance σ^2 . For some members of the exponential family the dispersion parameter is not present in the sense that $\phi \equiv 1$, as in the case of the Binomial (Multinomial) and Poisson distributions. Overdispersion under these distributions arises when $\phi > 1$.

In GLM, the mean μ_j , $j = 1, 2, \dots, n$, depends on a k -dimensional row vector of known covariates x_j and a k -dimensional column vector of unknown parameters β , through a smooth function $g(\cdot)$ so that

$$g(\mu_j) = x_j \beta = \sum_{i=1}^k x_{ji} \beta_i \quad (5)$$

The function $g(\cdot)$ is known as the *link function* and $\eta_j = x_j \beta$ is known as the *linear predictor*. Note that the inverse of the link function yields $\mu_j = g^{-1}(x_j \beta)$, the mean function. Let y represent the vector of observations $y = (y_1, y_2, \dots, y_n)^t$, and let $L(\beta, \phi; y) = \sum_{j=1}^n \ln \{f(y_j; \beta, \phi)\}$ represent the

log-likelihood function. Here $f(y_j; \beta, \phi)$ represents the density (or probability function) of Y_j , $j = 1, 2, \dots, n$, and the Y_j 's belong to the exponential family. The k -dimensional column vector of first derivatives of the log-likelihood function with respect to β , is the *gradient* or *score vector* that is,

$$u(\beta) = \frac{\partial L(\beta, \phi; y)}{\partial \beta}$$

The $k \times k$ matrix of second derivatives is the *Hessian*, given by

$$H(\beta) = \frac{\partial^2 L(\beta, \phi; y)}{\partial \beta \partial \beta^t}$$

The *Fisher Information Matrix (FIM)* is negative of the expected value of the Hessian, that is

$$I = -E \left[\frac{\partial^2 L(\beta, \phi; y)}{\partial \beta \partial \beta^t} \right] = E \left[\frac{\partial L(\beta, \phi; y)}{\partial \beta} \frac{\partial L(\beta, \phi; y)}{\partial \beta^t} \right]$$

In GLM, in order to obtain $\hat{\beta}$, the *maximum likelihood estimator (MLE)* of β , one needs to solve the estimating equations

$$\frac{\partial L(\beta, \phi; y)}{\partial \beta} = 0 \quad (6)$$

Parameter estimates are obtained by solving Eq. (6), usually by iterative methods such as Newton–Raphson or Gauss–Newton procedure. Nelder and Wedderburn,^[41] and Wedderburn,^[42] showed that the solution of the maximum likelihood equations is equivalent to an iterative weighted least-squares procedure. If the current estimate of β is denoted by $\hat{\beta}_{(s)}$, the improved estimate $\hat{\beta}_{(s+1)}$ is

$$\hat{\beta}_{(s+1)} = \hat{\beta}_{(s)} + (\hat{D}_{(s)}^t \hat{V}_{(s)}^{-1} \hat{D}_{(s)})^{-1} \hat{D}_{(s)}^t \hat{V}_{(s)}^{-1} (y - \hat{\mu}_{(s)}) \quad (7)$$

where $\hat{\mu}_{(s)}$, $\hat{V}_{(s)}$ and $\hat{D}_{(s)}$ correspondent respectively to the n -dimensional column vector of means $\mu = (\mu_1, \mu_2, \dots, \mu_n)^t$, the $n \times n$ diagonal matrix of variance functions $V = \text{Diag}\{v(\mu_1), v(\mu_2), \dots, v(\mu_n)\}$ and the $n \times k$ matrix of first derivatives

$$D = \left(\frac{\partial \mu_1}{\partial \beta}, \frac{\partial \mu_2}{\partial \beta}, \dots, \frac{\partial \mu_n}{\partial \beta} \right)^t, \text{ all terms evaluated at } \beta = \hat{\beta}_{(s)}.$$

If the canonical link is assumed (logit link for the binomial distribution and logarithm link for the Poisson), it can be shown that iterative procedure in Eq. (7) further reduces to

$$\hat{\beta}_{(s+1)} = \hat{\beta}_{(s)} + (X^t \hat{V}_{(s)} X)^{-1} X^t (y - \hat{\mu}_{(s)}) \quad (8)$$

As in GLMs, in quasi-likelihood models the mean μ_j depends on a k -dimensional row vector of known covariates x_j and a k -dimensional column vector of unknown parameters β , through a smooth function $g(\cdot)$ as in Eq. (5). As pointed out earlier, in quasi-likelihood models, the structure of the first derivative of the quasi-likelihood function (without knowing the function from which this derivative is obtained from) resembles the structure of the *score function*, i.e., the derivative of the logarithm of the likelihood function. For the one-parameter exponential family, the log likelihood is the same as the quasi-likelihood. For a single observation, the quasi-likelihood function is defined as

$$\frac{\partial Q(Y_j, \mu_j)}{\partial \mu_j} = \frac{Y_j - \mu_j}{\phi v(\mu_j)} \quad (9)$$

Note that in Eq. (9), the function $Q(y, \mu)$ is such that its first derivative resembles the log-likelihood function of distributions in the exponential family, $Q(y, \mu) = \int_{-\infty}^{\mu} \frac{y-t}{\phi v(t)} dt + C$. The quantity

$$Q(y, \mu) = \sum_{j=1}^n Q(y_j, \mu_j) \quad (10)$$

is known as the *quasi-likelihood* (or more precisely as the *log-quasi-likelihood*) based on the data $y = (y_1, y_2, \dots, y_n)^t$ and the mean vector $\mu = (\mu_1, \mu_2, \dots, \mu_n)^t$. The quasi-likelihood estimate of β , say $\hat{\beta}$, is obtained by solving Eq. (10) with respect to β , that is

$$U(\beta) = \frac{\partial Q(y, \mu)}{\partial \beta} = 0 \quad (11)$$

These equations can be written as

$$\sum_{j=1}^n \frac{y_j - \mu_j}{\phi v(\mu_j)} \frac{\partial \mu_j}{\partial \beta} = 0 \quad (12)$$

whose solution in β does not require knowledge of ϕ . Therefore, they can be solved using the iterative weighted least-squares procedure as in Eq. (7) or as in Eq. (8). Once the iterative procedure has converged, the estimated covariance matrix is inflated by a consistent estimator of ϕ . Two common estimators of ϕ are based on the *deviance* and on *Pearson's Chi-square*, each statistic divided by its degrees of freedom. Assuming t_j is a realization of $T_j \sim \text{Binomial}(\pi_j, m_j)$, and y_j a realization of $Y_j \sim \text{Poisson}(\lambda_j)$, T_j 's and Y_j 's independent, $j = 1, 2, \dots, n$, the deviance and Pearson's chi-square for the binomial distribution are

$$D = 2 \sum_{j=1}^n \left\{ t_j \log \left(\frac{t_j}{m_j \hat{\pi}_j} \right) + (m_j - t_j) \log \left(\frac{m_j - t_j}{m_j - m_j \hat{\pi}_j} \right) \right\} \quad (13)$$

$$X^2 = \sum_{j=1}^n \frac{(t_j - m_j \hat{\pi}_j)^2}{m_j \hat{\pi}_j (1 - \hat{\pi}_j)} \quad (14)$$

Similarly, for the Poisson distribution we have

$$D = 2 \sum_{j=1}^n \left\{ y_j \log \left(\frac{y_j}{\hat{\lambda}_j} \right) - (y_j - \hat{\lambda}_j) \right\} \quad (15)$$

$$X^2 = \sum_{j=1}^n \frac{(y_j - \hat{\lambda}_j)^2}{\hat{\lambda}_j} \quad (16)$$

We now see some examples.

We first consider the simple case where $T_1, T_2, \dots, T_n$ are random variables with extra variation relative to the binomial model so that the mean and the variance of each random variable are as in Eqs. (1) and (2). Using the notation for quasi-likelihood models, we have $\mu_j = mp$, $v(\mu_j) = mp(1-p)$, and $\phi = \{1 + \rho^2(m-1)\}$ for each j , $j = 1, 2, \dots, n$. Here, the link function is the identity. Also, there are not covariates associated with the responses. To illustrate the estimation procedure, consider the data in Table 1, taken from a clinical trial. The study consisted of $n = 17$ subjects who brushed their teeth for six consecutive weeks with an experimental dental paste. The pH was measured at the end of the study at four sites of each subject's mouth ($m = 4$). It is believed that if the $\text{pH} \leq 6.5$, the subject is at higher risk of oral disease. For each site of each subject, we created a binary outcome variable, taking the value of "1" if the $\text{pH} \leq 6.5$ and "0" otherwise. Table 1 provides the distribution of the T_j 's (total number of mouth sites with $\text{pH} \leq 6.5$). The quasi-likelihood estimate of π is

$$\hat{\pi} = \frac{\sum_{j=1}^n T_j}{nm} \quad (17)$$

The estimated variance under the assumption that T_j 's are truly Binomial random variables, the so called naïve variance estimate, would be

$$\hat{V}(\hat{\pi}) = \frac{\hat{\pi}(1-\hat{\pi})}{nm} \quad (18)$$

The Pearson's chi-square statistic in Eq. (14) divided by its degrees of freedom ($n-1$) is a consistent estimator of ϕ . In this example

Table 1 pH Data

Number of mouth sites with $\text{pH} \leq 6.5$	0	1	2	3	4
Number of subjects	4	4	6	2	1

$$\hat{\phi} = \frac{\sum_{j=1}^n (t_j - m\hat{\pi})^2}{m\hat{\pi}(1-\hat{\pi})(n-1)} \quad (19)$$

The quasi-likelihood estimated variance of $\hat{\pi}$ is the naïve variance in Eq. (18) times the estimate of the scale parameter in Eq. (19). Thus, the quasi likelihood results are: $\hat{\pi} = 0.3824$, $\hat{\phi} = 1.4712$ and $se(\hat{\pi}) = 0.0715$. An approximated 95% confidence interval for π is given by (0.24, 0.52). Because the disease rate (proportion of sites with $\text{pH} \leq 6.5$) was approximately 71% at baseline, it is believed that the experimental toothpaste was efficacious in reducing the proportion of mouth sites at higher risk of oral disease.

For a discussion on quasi-likelihoods models we refer the reader to Cox and Snell.^[44] The use of quasi-likelihood models becomes more complex in the situation we are going to consider next.

Suppose that $T_1, T_2, \dots, T_n$ are independent random variables with extra variation relative to the binomial model such that $E(T_j) = m_j \pi_j$, $\text{Var}(T_j) = m_j \pi_j (1 - \pi_j) \{1 + \rho^2(m_j - 1)\}$, and $\ln(\pi_j / (1 - \pi_j)) = \mathbf{x}_j \boldsymbol{\beta}^0$. In order to estimate $\boldsymbol{\beta}$ and ρ simultaneously, Williams^[5] modified the iterative weighted least-squares procedure in Eq. (8) to allow the weights (inflation factors) $w_j^{-1} = \{1 + (m_j - 1)\rho^2\}$ to be used in conjunction with the variance function $v_j = m_j \pi_j (1 - \pi_j)$. In Williams's paper the intra-cluster correlation is denoted by ϕ . Note that the variance of T_j is $v_j w_j^{-1}$. For known ρ^2 , the quasi-likelihood estimator of $\boldsymbol{\beta}$ can be solved iteratively as

$$\hat{\boldsymbol{\beta}}_{(s+1)} = \hat{\boldsymbol{\beta}}_{(s)} + (\mathbf{X}' \mathbf{W} \mathbf{V} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W} (\mathbf{y} - \boldsymbol{\mu}) \quad (20)$$

where $\mathbf{V} = \text{Diag}\{m_1 \pi_1 (1 - \pi_1), m_2 \pi_2 (1 - \pi_2), \dots, m_n \pi_n (1 - \pi_n)\} \equiv \text{Diag}(v_1, v_2, \dots, v_n)$ and

$$\mathbf{W} = \text{Diag}\left\{\frac{1}{1 + (m_1 - 1)\rho^2}, \frac{1}{1 + (m_2 - 1)\rho^2}, \dots, \frac{1}{1 + (m_n - 1)\rho^2}\right\} \equiv \text{Diag}(w_1, w_2, \dots, w_n).$$

Williams performed Taylor's series expansion on $E(X^2)$ in order to derive the moment estimator ρ^2 with X^2 defined as in Eq. (14). At each step of the iterative procedure the updated $\hat{\boldsymbol{\beta}}$ is used to obtain an updated $\hat{\rho}^2$. The latter is used to improve $\hat{\boldsymbol{\beta}}$ in the next iteration. We refer the reader to Williams^[5] for more details.

To illustrate the single variance inflation procedure and Williams^[5] approach, consider the ossification data of Table 2, which has been previously analyzed by Morel and Neerchal.^[16] The data come from a two-way factorial design with 81 pregnant C57BL/6J mice. The purpose of the experiment was to investigate the synergistic effect of phenytoin and trichloropropene oxide. The presence or absence of ossification at the phalanges is considered a measure of the teratogenic effect of the substances under study. At some

Table 2 Ossification Data^a

Group	Observations
Control	8/8, 9/9, 7/9, 0/5, 3/3, 5/8, 9/10, 5/8, 5/8, 1/6, 0/5, 8/8, 9/10, 5/5, 4/7, 9/10, 6/6, 3/5
Sham	8/9, 7/10, 10/10, 1/6, 6/6, 1/9, 8/9, 6/7, 5/5, 7/9, 2/5, 5/6, 2/8, 1/8, 0/2, 7/8, 5/7
PHT	1/9, 4/9, 3/7, 4/7, 0/7, 0/4, 1/8, 1/7, 2/7, 2/8, 1/7, 0/2, 3/10, 3/7, 2/7, 0/8, 0/8, 1/10, 1/1
TCPO	0/5, 7/10, 4/4, 8/11, 6/10, 6/9, 3/4, 2/8, 0/6, 0/9, 3/6, 2/9, 7/9, 1/10, 8/8, 6/9
PHT + TCPO	2/2, 0/7, 1/8, 7/8, 0/10, 0/4, 0/6, 0/7, 6/6, 1/6, 1/7

^aNumber of fetuses showing ossification / litter size. PHT: phenytoin; TCPO: trichloropropene oxide.

phalanges the ossification did not occur at all; in some others the ossification was almost complete. The middle third phalanges were selected because they provided about 50% chance of ossification in the control groups. For illustration purposes, we only analyzed the response of the left middle third phalanx. In the experiment each pregnant mouse was randomly allocated to an untreated control group (Control), a sham control group (Sham) receiving gavages of water, and three treated groups receiving daily, by gastric gavages, 60 mg/kg of phenytoin (PHT), 100 mg/kg of trichloropropene oxide (TCPO), and 60 mg/kg phenytoin and 100 mg/kg of trichloropropene oxide (PHT + TCPO). For illustrative purposes, the two controls were combined and will refer to as combined controls. The experiment thus can be seen as a 2×2 factorial, with TCPO and PHT as the two factors. The levels of TCPO are 0 mg/kg and 100 mg/kg, and the levels of PHT are 0 mg/kg and 60 mg/kg. The data on the left middle third phalanx are shown in Table 2.

Table 3 depicts the beta estimates and their standard errors after fitting the full factorial logistic regression model under no extra variation (fetuses within the litters are wrongly assumed to be independent) and under the two extra variation quasi-likelihood models previously discussed. The single variance inflation approach is based

on multiplying the “naïve” covariance matrix of the beta estimates by a consistent estimator of the scale parameter. In this example we picked the deviance in Eq. (13) divided by its degrees of freedom. The no-extra variation model and the single-inflation-factor quasi-likelihood model differ in that the variance-covariance matrix of the latter is 3.77 times the variance-covariance matrix of the former. The estimation procedure proposed by Williams^[5] produced results very consistent with those obtained under the single-inflation-factor quasi-likelihood model. Williams’s model differs from the “naïve” model in that Williams uses the weights $w_j = 1/\{1 + 0.410875(m_j - 1)\}$ which were obtained iteratively.

We now turn to extra variation Poisson models. Similar to extra variation binomial models, extra variation Poisson models are used to accommodate higher variability in the data than permitted by the usual Poisson model. One way to accomplish this is by letting the mean of the Poisson distribution vary according to a prior distribution, with possibly some unknown parameters. That is, we consider $m_1, m_2, \dots, m_n$ are random counts such that m_j for $j = 1, 2, \dots, n$ is distributed as a Poisson random variable with mean λ_j . Furthermore, assume that λ_j is a random variable with density function $f(\lambda)$ such that $E(\lambda_j) = \mu$ and $V(\lambda_j) = \psi\mu^2$. We are assuming $\psi > 0$; otherwise the distribution of the λ_j ’s would be degenerate at the value of μ . Then the unconditional mean and variance of m_j are $E(m_j) = \mu$ and $\text{Var}(m_j) = \mu(1 + \psi\mu)$, and the counts m_j ’s have extra variation relative to the Poisson model. When the prior distribution on the λ_j ’s is gamma, then the marginal distribution (unconditional distribution) of the m_j ’s is negative binomial (see Johnson et al.^[44]) The assumption of common mean μ can be relaxed and, as in the binomial case, the variance of the Poisson model can be inflated with a constant ϕ ($\phi > 1$) in order to account for extra variation. Thus we can have the counts m_j ’s such that $E(m_j) = \mu_j$ and $\text{Var}(m_j) = \phi\mu_j$, whereas before μ_j depends on a k -dimensional row vector of known covariates x_j and an k -dimensional column vector of unknown parameters β^0 , throughout a smooth function $g(\cdot)$, $g(\mu_j) = x_j\beta^0$. The natural choice for $g(\cdot)$ is

Table 3 Beta Estimates and Standard Errors of Ossification Data for Three Models

Parameter	Model					
	Naïve		Quasi-likelihood (single-inflation factor)		Quasi-likelihood (Williams ^[5] approach)	
	Estimate	Standard error	Estimate	Standard error	Estimate	Standard error
Intercept	0.83	0.14	0.83	0.27	0.72	0.25
TCPO	−0.85	0.22	−0.85	0.43	−0.74	0.43
PHT	−2.11	0.25	−2.11	0.49	−1.95	0.47
TCPO + PHT	1.05	0.41	1.05	0.80	1.04	0.76
Scale Parameter	—		3.77		—	
ρ^2	—		—		0.41	

PHT: phenytoin; TCPO: trichloropropene oxide.

Table 4 Vertebral Fracture Data^a

Group	Observations
Active (Bisphosphonate)	(1, 2.9897), (2, 2.9843), (0, 2.9897), (1, 2.9843), (2, 2.9459), (0, 3.0445), (1, 3.0582), (0, 2.9897), (1, 1.2567), (0, 2.9678), (3, 2.9596), (1, 2.9843), (0, 2.9925), (1, 1.4346), (1, 3.0116), (0, 2.9733), (2, 3.0637), (1, 2.9706), (0, 1.1882), (0, 2.9733), (1, 1.2676), (2, 2.9541), (0, 2.3874), (0, 1.2868), (0, 1.2813), (2, 2.7625), (2, 2.4504), (4, 2.8747)
Placebo	(0, 1.3771), (11, 3.0418), (0, 2.4312), (3, 3.1650), (1, 3.0034), (2, 2.9843), (0, 2.9651), (2, 1.3087), (5, 3.0363), (8, 3.0691), (2, 2.9706), (2, 3.2553), (4, 2.9624), (0, 3.2060), (1, 2.9596), (7, 3.0144), (0, 2.9569), (2, 2.9843), (0, 1.1937), (3, 3.0062), (1, 2.9870), (4, 3.0089), (1, 1.2621), (1, 2.9788), (5, 2.8556)

^aNumber of incident vertebral fractures, person-years of observation.

the logarithm function, which is the “canonical link” if the m_j 's were really distributed as Poisson ($\phi = 1$). Let us consider an example of Poisson counts with extra variation.

In a clinical study of osteoporotic patients at high risk of vertebral fractures, subjects were randomly assigned either to a group receiving active therapy (bisphosphonate) ($n = 28$) or to a placebo group ($n = 25$). The number of incident vertebral fractures (new and worsening) was measured during a period of time that lasted between approximately 14 and 39 months. The data are presented in Table 4.

Let m_{ij} and Y_{ij} represent the number of vertebral fractures and the years of observation of the j -th subject in the i -th treatment group. Conditioning on Y_{ij} , we fitted the extra variation Poisson model $E(m_{ij}) = \mu_{ij}$, $\text{Var}(m_{ij}) = \phi\mu_{ij}$, $\phi > 1$, and link function $\ln(\mu_{ij}/Y_{ij}) = \alpha + \tau_i$, where α is an overall mean and τ_i the effect of the i -th treatment group. The estimates are obtained by using iterated re-weighted least squares as previously described under the assumption that $\phi = 1$. Then, the variance-covariance matrix of the parameter estimates is inflated with a consistent estimate of ϕ . In this example we have chosen the Pearson's chi-square statistic in Eq. (16) divided by its degrees of freedom. The results of this estimation procedure for the vertebral fracture data are presented in Table 5. The estimated fractures-per-year rates are $e^{-0.0449} = 0.96$ for the placebo group and $e^{-0.0449-0.9000} = 0.39$ for the bisphosphonate group, indicating a highly significant (p -value = 0.0025) reduction in the number of vertebral fractures in the bisphosphonate group, compared to placebo. The estimate of the inflation factor (scale parameter) ϕ is 1.7311. Another common estimator of ϕ is the deviance in Eq. (15) divided

by its degrees of freedom, which in this example turned out to be 1.8472.

LIKELIHOOD MODELS

In this section we will discuss estimating extra variation models using proper likelihood functions. A likelihood with first two moments as in Eqs. (1) and (2) can be obtained as the marginal distribution of the total number of successes, where its conditional distribution is binomial and the probability of success varies from cluster to cluster according to a certain distribution. Let $T_1, T_2, \dots, T_n$, represent independent random variables such that given the j -th random variable, $j = 1, 2, \dots, n$, T_j is distributed binomial (P_j, m). Furthermore, assume that the P_j 's are themselves random variables distributed according to a distribution in the interval (0,1) with density function $f(P)$ (or probability function if the distribution of P is discrete) such that $E_f(P_j) = \pi$ and $V_f(P_j) = \rho^2\pi(1 - \pi)$. Then it can be shown that the unconditional mean and variance of each T_j are as in Eqs. (1) and (2). Under a prior $f(P)$, for any $t = 0, 1, 2, \dots, m$,

$$\Pr(T = t) = E_f\{\Pr(T = t)\} = \int_0^1 \binom{m}{t} P^t (1-P)^{m-t} f(P) dP \quad (21)$$

A distribution of the form in Eq. (21) is said to be a mixture, with the prior $f(P)$ being the mixing distribution. If $f(P)$ is continuous, the unconditional distribution of T is an infinite mixture. If the prior $f(P)$ is discrete and finite, the integral symbol in Eq. (21) changes to a finite summation, and the resulting distribution is a finite mixture. Two examples of mixture distributions used to model extra variation are the beta-binomial distribution originally used by Skellam^[45] and the finite-mixture distribution proposed by Morel and Nagaraj.^[46] We first consider the beta-binomial distribution.

Let $f(P)$ be the density function of a beta distribution with parameters $a = C\pi$ and $b = C(1 - \pi)$, where $C = (1 - \rho^2)/\rho^2$. Note that $E_f(P) = a/(a + b) = \pi$ and

Table 5 Estimates, Standard Errors, Chi Squares, and p -values for Poisson Extra Variation model of Vertebral Fracture Data

Parameter	Estimate	Standard Error	Wald Chi-square	p -value
Intercept	-0.0449	0.1632	0.08	0.7833
Bisphosphonate	-0.9000	0.2974	9.16	0.0025
Scale (ϕ)	1.7311			

$V_f(P) = ab/(a+b+1)(a+b)^2 = \rho^2\pi(1-\pi)$. Thus, as noted earlier, the unconditional distribution of T exhibits extra variation. It can be shown that, for any t , $t = 0, 1, 2, \dots, m$,

$$\Pr(T = t) = \binom{m}{t} \frac{\prod_{k=1}^t (C\pi + k - 1) \prod_{k=1}^{m-t} \{C(1-\pi) + k - 1\}}{\prod_{k=1}^m (C + k - 1)} \quad (22)$$

where $C = (1 - \rho^2)/\rho^2$ and a product from $k = 1$ to $k = 0$ is defined as 1. The probability function of Eq. (22) defines the beta-binomial distribution with parameters π , ρ , and m and denoted as beta-binomial $(\pi, \rho; m)$.

Morel and Nagaraj^[46] proposed a new distribution for modeling extra variation where the unconditional distribution of the total number of successes T can be obtained from placing on P a two-point prior distribution on the points $(1 - \rho)\pi + \rho$ and $(1 - \rho)\pi$ with probabilities π and $(1 - \pi)$, respectively. In this case, the unconditional probability function of T turns out to be, for $t = 0, 1, \dots, m$,

$$\Pr(T = t) = \pi \Pr(X_1 = t) + (1 - \pi) \Pr(X_2 = t) \quad (23)$$

where $X_1 \sim \text{binomial}((1 - \rho)\pi + \rho; m)$ and $X_2 \sim \text{binomial}((1 - \rho)\pi; m)$. Note that this is a mixture of two binomial distributions. If T has the distribution given by Eq. (23), it is denoted as $T \sim \text{finite-mixture}(\pi, \rho; m)$. The finite-mixture distribution results from an effort to model meaningfully the physical mechanism behind the extra variation. It can be derived in the following way.

Let $Y, Y_1^0, \dots, Y_m^0$ be independent and identically distributed Bernoulli (π) random variables. For each i , $i = 1, \dots, m$, define Y_i as Y with probability ρ , or as Y_i^0 with probability $(1 - \rho)$, $0 \leq \rho \leq 1$. Each Y_i can be represented as $Y_i = YI(U_i \leq \rho) + Y_i^0I(U_i > \rho)$, where the U_i 's are independent uniform $(0, 1)$ random variables, and $I(\cdot)$ denote an indicator function. In this model for clumped data, the response of any member of the cluster is the same as Y , with probability ρ . Also, for $i \neq j$ it can be shown that $\Pr(Y_i = 1 | Y_j = 1) = \rho^2$. That is, the conditional probability that a member of the cluster responds positively, given that one of the others in the cluster has already responded positively, is ρ^2 . If we let T be the sum of the Y_i 's, then $T = Y \sum_{i=1}^m I(U_i \leq \rho) + \sum_{i=1}^m Y_i^0 I(U_i > \rho)$. This leads to the following representation of T :

$$T = (Y'N) + (X | N) \quad (24)$$

where $N \sim \text{binomial}(\rho; m)$, $Y \sim \text{Bernoulli}(\pi)$, N and Y independent, and $(X | N) \sim \text{binomial}(\pi; m - N)$ if $N < m$. It can be shown [see Morel^[47]] that the probability function of T in Eq. (24) is given by Eq. (23). Under representation in Eq. (24), the outcome given by Y is duplicated a random number of times N , $N = 0, 1, \dots, m$. This is the number

of units in the cluster exhibiting the same behavior. The remaining $m - N$ units within the cluster provide independent Bernoulli responses. Contribution of these $m - N$ units to the total count T is represented by $(X | N)$. Fig. 1 illustrates the representation of a cluster of size $m = 7$. The cluster is thought of as having two subgroups. One is of size 4 (N ; in general $0 \leq N \leq 7$), and it consists of those elemental units that greatly influence each other and provide identical responses.

The three elemental units forming the second subgroup respond independently of each other and of the first subgroup. Their responses have been shown by a “?” to indicate that they may be either a “0” or a “1.” Clearly, the binomial count T in Eq. (24) exhibits extra variation, because the individual respondents (either by consulting each other before responding or by a natural biological inheritance) provide more similar responses than one would expect from independent Bernoulli trials. As far as the authors know, the finite-mixture is the only distribution that provides a clear interpretation of the underlying cluster mechanism. To illustrate the fitting of the beta-binomial and finite-mixture distribution, we now present an example on chromosome association.

Skellam^[45] analyzed data on the secondary association of chromosomes in *Brassica*, a group of Old World temperate zone herbs of the mustard family with beaked cylindrical pods (similar to cabbage). This is the first example in the literature where the beta-binomial distribution was formally used to model extra variation. At meiosis, the special process of cell division, there are three pairs of bivalents. A *bivalent* is a structure consisting of two paired homologous chromosomes. The paired chromosomes in each bivalent may or may not show association. For a nucleus i , $i = 1, 2, \dots, n$, let T_i represent the total number of pairs of bivalents showing association, T_i , $i = 0, 1, 2, 3$. If the probability of association is constant for all nuclei and the same for all pairs of bivalents, the binomial distribution would be appropriate for this problem. As we will see next, extra variation is present in this example, and both

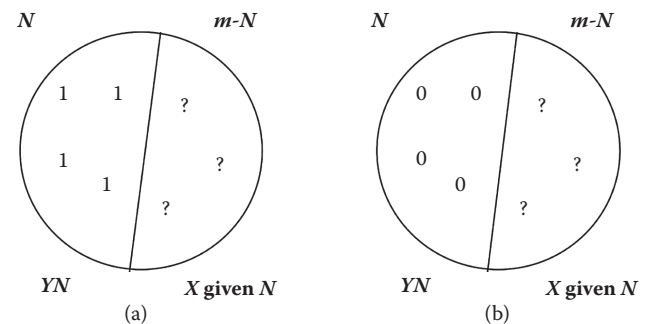


Fig. 1 Two possible outcomes for a cluster of size $m = 7$ under the finite-mixture distribution. Some elemental units provide identical responses, either “1,” as in (a), or “0,” as in (b). The “?” symbol represents elemental units responding independently, with either a “0” or a “1”

Table 6 Chromosome Data

Number of Associations	0	1	2	3
Frequency	32	103	122	80

the beta-binomial and finite-mixture distributions provide reasonable fits to this problem.

Table 6 shows the distribution of the number of chromosome associations for 337 nuclei. We fit three distributions to these data: the binomial, the beta-binomial, and the finite-mixture. The maximum likelihood results of fitting these distributions are given in **Table 7**.

As we can see in this example, the two extra variation distributions produce almost identical results. The variance estimates under the beta-binomial and the finite-mixture distributions are 1.17 times the estimate obtained by using the binomial distribution. This value is an estimate of the multiplier $\{1 + \rho^2(m - 1)\}$ that appears in Eq. (2). To test for the presence of extra variation under a specific overdispersion model, we need to test the hypothesis $H_0: \rho = 0$ versus the alternative hypothesis $H_1: \rho > 0$. Properties of the likelihood ratio test cannot be obtained by the usual asymptotic theory since the parameter being tested is on the boundary of the parameter space, $0 < \rho < 1$. Self and Liang^[48] have shown that in this case (see case 5 in their paper), the asymptotic distribution of the likelihood ratio test is a 50:50 mixture of a point mass at zero and at a chi square distribution with 1 degree of freedom. The likelihood ratio test statistic is 7.2 under the beta-binomial and 7.0 under the finite-mixture, with p -values approximately 0.004. These statistics are almost the same, and they strongly support the hypothesis of extra variation under either model.

To conclude this example, we compute the traditional Pearson's goodness-of-fit chi square test $X^2 = \sum_{s=0}^3 (O_s - E_s)^2 / E_s$ to see if either distribution is appropriate for the data. **Table 8** shows the observed and expected frequencies of the number of associations under the three distributions considered in this example. The binomial distribution does not fit the data well, given that $X^2 = 8.11$, with an associated p -value of 0.017. The beta-binomial and the finite-mixture distributions fit the distribution of the number of associations fairly well, with P -values of the chi square goodness-of-fit tests being 0.383 and 0.348, respectively.

Neerchal and Morel^[49] investigated the “relative asymptotic efficiency” of the quasi-likelihood estimator given in Eqs. (17)–(19) with respect to the maximum likelihood estimate of π of the beta-binomial and finite-mixture distributions. The relative efficiency of the quasi-likelihood estimator relative to the maximum likelihood estimator of π under the beta-binomial is

$$\frac{\pi(1-\pi)\phi}{\frac{nm}{V(\hat{\pi}_{bb})}}$$

where $\phi = \{1 + \rho^2(m - 1)\}$, $\hat{\pi}_{bb}$ is the maximum likelihood estimator of π under the beta-binomial distribution and $V(\hat{\pi}_{bb})$ is the asymptotic covariance matrix when the cluster size m is large. The relative asymptotic efficiency with respect to the finite-mixture distribution is defined in a similar way. **Table 9** shows the asymptotic relative efficiencies for the combinations $\pi = 0.1, 0.3, 0.5$; $\rho^2 = 0.09, 0.25, 0.49$; and $m = 10, 40$. The results from **Table 9** indicate

Table 7 Results of Fitting Binomial, Beta-Binomial, and Finite-Mixture Distributions to Chromosome Data

Parameter	Distribution					
	Binomial		Beta binomial		Finite mixture	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
π	0.5806	0.0155	0.5807	0.0168	0.5805	0.0168
ρ			0.2944	0.0577	0.2930	0.0580
$-2 \times \text{Log-likelihood}$	880.8		873.6		873.8	

Table 8 Chi-Square Goodness-of-Fit Tests for Chromosome Data

Number of associations	Observed frequency	Expected frequencies		
		Binomial	Beta-binomial	Finite-mixture
0	32	24.86	33.97	34.08
1	103	103.24	97.16	96.82
2	122	142.94	127.67	128.24
3	80	65.96	78.20	77.85
X^2		8.11	0.76	0.88
Degrees of Freedom		2	1	1
p -value		0.017	0.383	0.348

Table 9 Asymptotic Relative Efficiency of the Quasi-Likelihood Estimator Relative to Beta-Binomial and Finite-Mixture Distributions

ρ^2	m	Distribution	$\pi = 0.1$	$\pi = 0.3$	$\pi = 0.5$
0.09	10	Beta binomial	1.00	1.00	1.00
		Finite mixture	1.03	1.01	1.01
	40	Beta binomial	1.00	1.00	1.00
		Finite mixture	2.33	2.03	1.95
0.25	10	Beta binomial	1.01	1.01	1.01
		Finite mixture	1.25	1.19	1.18
	40	Beta binomial	1.03	1.03	1.03
		Finite mixture	3.85	3.82	3.81
0.49	10	Beta binomial	1.05	1.05	1.05
		Finite mixture	1.46	1.43	1.42
	40	Beta binomial	1.10	1.11	1.12
		Finite mixture	4.05	4.05	4.05

that there is little gain in efficiency in using the beta-binomial distribution over the quasi-likelihood model. The two distributions for modeling extra variation have essentially the same efficiency of the quasi-likelihood estimator when $\rho^2 = 0.09$ and $m = 10$. But as ρ^2 increases and/or the cluster size m gets larger, the gain in efficiency of the finite-mixture distribution over the quasi-likelihood model becomes greater.

We now revisit the teratology example of Table 2 and fit the beta-binomial and finite-mixture models, assuming a common intra-litter correlation ρ^2 and the logistic link with a full factorial design. The results of the fittings are giving in Table 10. As we can see from Table 10, the results of the beta-binomial and finite-mixture models are very consistent with each other. The negative coefficients indicate an increased risk of no ossification. Thus, the toxic effect of the chemicals is to impede ossification. If $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ represent, respectively, the beta estimates for Intercept, TCPO, PHT, and TCPO + PHT, then an approximate 95%

confidence interval for the odds ratio of PHT versus the combined controls when TCPO = 0 mg/kg is given by $e^{\hat{\beta}_2 \pm 1.96 \sqrt{\hat{v}(\hat{\beta}_2)}}$, where $\hat{v}(\hat{\beta}_2)$ represents the estimated asymptotic variance of $\hat{\beta}_2$. Similarly, an approximate 95% confidence interval for the odds ratio of PHT versus the combined controls when TCPO = 100 mg/kg is given by $e^{(\hat{\beta}_2 + \hat{\beta}_3) \pm 1.96 \sqrt{\hat{v}(\hat{\beta}_2 + \hat{\beta}_3)}}$. Table 11 shows the approximate 95% confidence intervals for the two levels of TCPO based on the beta estimates of the binomial, single-inflation-factor and Williams's^[5] approach in Table 3 and the beta estimates based on the beta-binomial and finite-mixture models in Table 10. Recall single-inflation factor was estimated by computing the deviance over its degrees of freedom. The confidence intervals under the binomial model are much shorter than for all the extra-variation models. This suggests that the level of confidence obtained under the binomial model is below the nominal one. The remaining models, which incorporate the litter effect, suggest that in the presence of TCPO, the chances to impede ossification may be the same for both PHT and the combined control.

A GOODNESS-OF-FIT TEST FOR EXTRA VARIATION MODELS

Tests for goodness-of-fit with extra variation models have typically focused on testing for the presence of extra variation in the data rather than model adequacy. Paul and Plackett,^[50] and Collings and Margolin,^[51] describe a variety of statistics for testing the null hypothesis that there is no overdispersion against specific alternatives such as the beta-binomial model. They consider likelihood functions that reduce to “no extra variation” when the extra variation parameter is zero. An illustration of such a test for overdispersion was given in “Likelihood Models,” when we tested the hypothesis $H_0 : \rho = 0$ versus the alternative hypothesis $H_1 : \rho > 0$ under either the beta-binomial and finite-mixture model, for the teratology data of Table 2. Note that the construction of goodness-of-fit tests for testing the adequacy of a specific extra variation model is straightforward if the

Table 10 Beta Estimates and Standard Errors of the Ossification Data for the Full Factorial Logistic Regression Model Under Beta-Binomial and Finite-Mixture Models

Parameter	Model			
	Beta binomial		Finite mixture	
	Estimate	Standard error	Estimate	Standard error
Intercept	0.70	0.23	0.64	0.23
TCPO	−0.78	0.40	−0.95	0.37
PHT	−1.69	0.40	−1.53	0.40
TCPO + PHT	0.68	0.69	0.62	0.67
ρ^2	0.34	0.05	0.34	0.04
−2*Log-likelihood	306.6		305.1	

PHT: phenytoin; TCPO: trichloropropene oxide.

Table 11 Approximate 95% Confidence Intervals for Odds Ratio of PHT Under Binomial, Single-Inflation-Factor, Williams,^[5] Beta-Binomial, and Finite-Mixture Models

Model	TCPO = 0 mg/kg	TCPO = 100 mg/kg
Binomial	0.07, 0.20	0.18, 0.65
Single Inflation Factor	0.05, 0.31	0.10, 1.19
Williams ^[5]	0.06, 0.36	0.12, 1.31
Beta-binomial	0.08, 0.41	0.12, 1.13
Finite-mixture	0.10, 0.48	0.13, 1.23

PHT: phenytoin; TCPO: trichloropropene oxide.

clusters sizes are the same, that is, $m_j \equiv m$, for $j = 1, 2, \dots, n$. This is the case of the chromosome example in “Likelihood Models” (Table 8), where the traditional goodness-of-fit test given in Rao (^[52], pages 391–398) and Cramer (^[53], pages 477–506) applies. When the m_j ’s are different, the construction of a Pearson’s goodness-of-fit statistic is not straightforward because the observed and expected frequencies are not associated with a unique value of m . Neerchal and Morel^[49] proposed an extension of the traditional Pearson’s chi-square statistic $X^2 = \sum_{s=0}^3 (O_s - E_s)^2 / E_s$ when the clusters sizes m_j ’s are allowed to be different. Asymptotic properties of this test, as well as extensions to the case where covariates are available, have been investigated by Sutradhar^[54] and Sutradhar et al.^[55]

Let $t_1, t_2, \dots, t_n$ represent a realization of n -independent random variables $T_1, T_2, \dots, T_n$, these variables assumed to have extra variation relative to the binomial distribution, that is, the first two moments are as in Eqs. (1) and (2). It is hypothesized that each T_j has a certain probability function depending on the parameters π, ρ , and the cluster size m_j , where the m_j ’s are allowed to be distinct. One way of extending the traditional Pearson’s statistic is by pooling the data across all clusters and comparing the observed distribution of t_j/m_j to the corresponding expected frequencies. To this end, first divide the unit interval into k mutually exclusive interval $A_s, s = 1, 2, \dots, k$. Then, compute the observed and expected frequencies as

$$O_s = \sum_{j=1}^n I\left(\frac{t_j}{m_j} \in A_s\right)$$

and

$$E_s = \sum_{j=1}^n \sum_{t=0}^{m_j} \Pr(T_j = t) I\left(\frac{t}{m_j} \in A_s\right)$$

where $I(\cdot)$ represents an indicator function and $\Pr(T_j = t)$ is based on the maximum likelihood estimates of π and ρ . An example of this test with varying m_j ’s can be found in Neerchal and Morel.^[49] They fit the beta-binomial and finite-mixture distributions to three data sets from a study to induce mutagenic damage previously analyzed

by Haseman and Soares.^[56] In the study, a group of male mice were first treated with a suspected mutagen and later paired with one or more untreated females. The female mice became pregnant and after a certain time they were sacrificed and their uteri examined for the presence of living and dead implantations. (An implantation is an early stage of a fetus.) Because implantations or fetuses from the same mother tend to respond similarly, it is reasonable to treat them as a natural cluster. Table 12 depicts the results of the goodness-of-fit test for one of the three data sets ($n = 554$). For this data set the average clusters size was 13.1, with cluster sizes varying between 1 and 18. The p -values of the goodness-of-fit test under the beta-binomial and finite-mixture distributions were 0.0014 and 0.2205 respectively. These tests are based on six degrees of freedom. It is clear that for this example, the finite-mixture distribution fits the data well while the beta-binomial distribution does not. The p -values were also estimated by bootstrapping the data parametrically with 20,000 replications. They turned to be 0.0012 and 0.2124, close to the ones obtained by using the chi-square distribution.

GENERALIZED ESTIMATING EQUATIONS

“Generalized Estimating Equations,” better known as GEE or GEEs, is a methodology adapted by Liang and Zeger,^[31] and Zeger and Liang^[32] for the analysis of longitudinal data when regression is the main focus. The methodology is related with quasi-likelihood models (Wedderburn^[42] and McCullagh^[57]) and generalized linear models (McCullagh and Nelder^[58]). Although the GEE methodology was originally proposed for longitudinal data analysis, it has been applied to almost any type of clustered data. In the forthcoming discussion we restrict ourselves to binary responses or counts. However the methodology can be extended to

Table 12 Goodness-of-Fit Results for Haseman and Soares^[54] Data ($n = 554$)

Interval A_s	Observed (O_s)	Expected (E_s)	
		Beta-binomial	Finite-mixture
[0, 1/19]	286	298.8457	283.0477
(1/19, 2/19]	141	113.9472	147.4433
(2/19, 3/19]	51	45.8545	44.7637
(3/19, 4/19]	28	34.5916	26.9008
(4/19, 5/19]	20	24.0023	13.4963
(5/19, 6/19]	7	14.5022	8.0440
(6/19, 7/19]	4	9.1030	7.5816
(7/19, 8/19]	3	5.1611	6.2479
(8/19, 1]	14	7.9922	16.4746
Total	554	554.0000	554.0000
X^2		21.6384	8.2477

multinomial outcomes. For each subject j , $j = 1, 2, \dots, n$, let $\mathbf{Y}_j = (y_{j1}, y_{j2}, \dots, y_{jm_j})^t$ represent a m_j -dimensional column vector of responses, and let $\mathbf{X}_j = (x_{j1}, x_{j2}, \dots, x_{jm_j})^t$ denote an associated $m_j \times k$ matrix of observed covariates, where x_{jl} , $l = 1, 2, \dots, m_j$, represents a k -dimensional row vector. Let μ_{jl} and $v(\mu_{jl})$ represent, respectively, the mean and the “natural variance” of y_{jl} , the latter being the variance of the y_{jl} ’s when they are independent and belong to the exponential family. If the y_{jl} ’s were either binomial or Poisson distributed, their corresponding natural variances would be $\pi_{jl}(1 - \pi_{jl})$ or λ_{jl} , respectively. It is further assumed that the μ_{jl} ’s are related to the x_{jl} ’s throughout a link function $g(\cdot)$ such that $g(\mu_{jl}) = x_{jl}\beta^0$, where β^0 is a k -dimensional column vector of unknown parameters. The canonical links for the binomial and Poisson distributions are the logit and logarithmic functions, respectively. Other links for the binomial distribution are the “identity,” the “probit” and the “complementary log-log” functions. The identity function may be used as a link function for the Poisson case.

Let $\mathbf{A}_j$ represent the $m_j \times m_j$ diagonal matrix of natural variances of the vector $\mathbf{Y}_j$, $\mathbf{A}_j = \text{diag}\{v(\mu_{j1}), v(\mu_{j2}), \dots, v(\mu_{jm_j})\}$. In order to apply the GEE methodology one needs first to define the correlation structure of $\mathbf{Y}_j = (y_{j1}, y_{j2}, \dots, y_{jm_j})^t$ through an $m_j \times m_j$ “working” correlation matrix $\mathbf{R}_j(\alpha)$ such that

$$\mathbf{V}_j = \mathbf{A}_j^{1/2} \mathbf{R}_j(\alpha) \mathbf{A}_j^{1/2} \phi \quad (25)$$

would be equal to the variance-covariance matrix of $\mathbf{Y}_j$, if $\mathbf{R}_j(\alpha)$ were the true correlation matrix for the $\mathbf{Y}_j$ ’s. Then define the estimating equations as

$$\sum_{j=1}^n \mathbf{D}_j^t \mathbf{V}_j^{-1} (\mathbf{Y}_j - \mu_j) = \mathbf{0} \quad (26)$$

where $\mu_j = (\mu_{j1}, \mu_{j2}, \dots, \mu_{jm_j})^t$, $\mathbf{D}_j^t$ represents the $k \times m_j$ matrix of first derivatives of μ_j with respect to β^0 , and $\mathbf{V}_j$ is defined as in Eq. (25).

In order to solve the estimating Eq. (26) for β^0 , consistent estimators of ϕ and α (say, $\hat{\phi}$ and $\hat{\alpha}$) are needed. These estimators are functions of the data and of β^0 . So they can be acquired by first obtaining $\hat{\beta}$, the solution of estimating Eq. (26) assuming that $\mathbf{R}_j(\alpha)$ is an identity matrix and that ϕ is equal to 1. Then $\hat{\phi}$ and $\hat{\alpha}$ are plugged into the estimating Eq. (26), which is then solved to obtain an improved version of $\hat{\beta}$. The improved version of $\hat{\beta}$ incorporates the hypothesized covariance matrix structure given in Eq. (25) and can be used to obtain new estimators $\hat{\phi}$ and $\hat{\alpha}$. The procedure continues until no changes are observed on $\hat{\beta}$, $\hat{\phi}$, and $\hat{\alpha}$. Liang and Zeger^[31] have shown the consistency of $\hat{\beta}$ under mild regularity conditions. They also showed that a consistent estimator of the variance-covariance matrix of $\hat{\beta}$

$$\tilde{\mathbf{A}} = \left\{ \mathbf{H}(\hat{\beta}, \hat{\alpha}, \hat{\phi}) \right\}^{-1} \mathbf{G} \left\{ \mathbf{H}(\hat{\beta}, \hat{\alpha}, \hat{\phi}) \right\}^{-1} \quad (27)$$

where $\mathbf{H}(\hat{\beta}, \hat{\alpha}, \hat{\phi}) = \sum_{j=1}^n \mathbf{D}_j^t(\hat{\beta}) \mathbf{V}_j^{-1}(\hat{\beta}, \hat{\alpha}, \hat{\phi}) \mathbf{D}_j(\hat{\beta})$, $\mathbf{G} = \frac{(n^* - 1)}{(n^* - k)} \frac{n}{(n - 1)}$

$$\sum_{j=1}^n (d_j - \bar{d})(d_j - \bar{d})^t, \quad n^* = \sum_{j=1}^n m_j, \quad d_j = \mathbf{D}_j^t(\hat{\beta}) \mathbf{V}_j^{-1}(\hat{\beta}, \hat{\alpha}, \hat{\phi})$$

$\mathbf{Y}_j - \mu_j(\hat{\beta})$, and $\bar{d} = \sum_{j=1}^n d_j / n$. Usually the matrix $\mathbf{G}$ does not

contain the multiplier $(n^* - 1)/(n^* - k)$ and some authors treat $\bar{d}$ as a null vector. If each cluster contains exactly one elemental unit, the factor $\frac{(n^* - 1)}{(n^* - k)} \frac{n}{(n - 1)}$ reduces to $n/(n - k)$, which is the degrees of freedom correction applied to the residual mean square for ordinary least squares in which k parameters are estimated. The quantity $(n^* - 1)/(n^* - k)$ reduces the small sample bias associated with using the estimated link function to calculate deviations. For large samples, the variance-covariance $\tilde{\mathbf{A}}$ in Eq. (27) provides correct Type I error rates for arbitrary choice of the working correlation matrix $\mathbf{R}_j(\alpha)$. This is one of the most interesting characteristics of GEE: the inferences based on $\hat{\beta}$ are correct as far as one uses the “robust” (also known as “sandwich” or “empirical”) variance-covariance matrix given in Eq. (27).

This is under the assumption that the mean of the model has been correctly estimated. Liang and Zeger^[31] provide results of a simulation study where they show gains in efficiency when the working correlation matrix coincides with the true and unknown one. Some examples of working correlation matrices are: independence, exchangeable, m -dependent, first-order autoregressive, and unspecified. For each of these cases, Liang and Zeger^[31] provide indications on how to estimate ϕ and α .

It is well known that use of the estimated variance-covariance matrix given in Eq. (27) in the construction test of Wald’s statistics leads to inflated Type I errors in small samples. Morel,^[59] Morel, Bokossa and Neerchal,^[60] and Morel and Neerchal,^[61] have proposed a modified version of in Eq. (27) to reduce the small sample bias. The modified variance-covariance matrix is asymptotically equivalent to the one defined in Eq. (27). The small sample correction is given by

$$\hat{\mathbf{A}} = \hat{\mathbf{A}} + \hat{\delta}_n \hat{\phi} \left\{ \mathbf{H}(\hat{\beta}, \hat{\alpha}, \hat{\phi}) \right\}^{-1} \quad (28)$$

where $\tilde{\mathbf{A}}$ is defined as in Eq. (27), $\hat{\phi} = \max\{1, \text{trace}\{(\mathbf{H}(\hat{\beta}, \hat{\alpha}, \hat{\phi}))^{-1} \mathbf{G}\} / v\}$, and $\hat{\delta}_n = \min\{0.5, v/(n - v)\}$. The quantity $\hat{\phi}$ is an estimate of the design effect, which has been bounded below by 1. The factor $\hat{\delta}_n$ converges to zero as the sample size increases, and has been bounded above by 0.5. The upper bound on $\hat{\delta}_n$ is arbitrary and usually does not come into place when sample sizes are moderate. It has been demonstrated in Monte Carlo studies that the variance-covariance matrix in Eq. (28) reduces the inflated Type error rates due to small samples. We recommend the use of the modified estimator in practice. This adjusted

variance estimator falls into the category of “additive correction” and it is always positive definite provided $\tilde{A}$ in Eq. (28) is. Other small sample bias corrections have been proposed by Fay and Graubard^[62] and Mancl and DeRouen.^[63]

Let us now consider the ossification data of Table 2. These data were used to illustrate the fitting of two quasi-likelihood models in “Quasi-likelihood Models,” and the fitting of two likelihood models in “Likelihood Models.” We fit these data using the GEE methodology with an “independence” working correlation matrix. These results, which are reported in Table 13, illustrate the use of the Morel-Bokossa-Neerchal small sample bias correction. The adjusted standard errors are slightly larger to correct for the small sample bias correction embedded in the “sandwich.” As the number of clusters n increases, this correction vanishes.

Liu and Liang^[64] present a method for computing sample size and power for studies with correlated observations. The term “sample size” refers to the number of clusters in which the cluster is formed by individuals in longitudinal studies and by families in family studies. The sample size determination is based on the GEE methodology and a quasi-score test statistic. We will next consider two important special cases namely the two-sample problem with correlated binary responses, and the case of a cross-over design with binary responses.

In the two-sample problem two independent treatment groups are considered. Let P_1 and P_2 represent their hypothesized proportions. Let ρ represent the intra-cluster correlation and m the number of repeated (correlated) measurements. At a α level and a $1 - \beta$ power, an estimate of the sample size n (number of clusters) is given by

$$n = \frac{(Z_{\beta} + Z_{\alpha/2})^2 \{P_1(1 - P_1) + P_2(1 - P_2)\} \{1 + \rho(m - 1)\}}{(P_1 - P_2)^2 m} \quad (29)$$

Here Z_p represents the upper p -th percentile of a standard normal distribution.

As an example, consider a dental care study for the detection of caries. Children used a tooth paste containing as

active ingredient either stannous fluoride or sodium fluoride for three years. The estimated proportions of observing caries for stannous fluoride and sodium fluoride were respectively 9.8% and 8.6%. The corresponding odds ratio is 1.15 (relative risk is 1.14). The average number of teeth surfaces investigated were approximately $m = 110$ surfaces per subject. Furthermore, an estimate of the “design effect” $\{1 + \rho(m - 1)\}$ turned out to be 5.50. A new study is being planned to establish the superiority of sodium fluoride over the active ingredient stannous fluoride. Assuming a power of 82% and a Type I error rate of 10% (5% for a one-sided test), then using Eq. (29) the estimated sample size is $n = 380$ subjects per treatment group.

Consider now a 2×2 cross-over design, that is, two treatments (denoted by A and B, say) and two periods. Subjects are randomly allocated to one of the sequences AB and BA. That is, half of the subjects receive treatment A during the first period, and then treatment B during the second period. The order of application for subjects in the BA sequence is reversed. Let P_1 and P_2 represent the success rates of treatments A and B, respectively. Let ρ represent the correlation of responses between the two periods. At a α level and a $1 - \beta$ power, an estimate of the sample size n is

$$n = \frac{(Z_{\alpha/2} + Z_{\beta})^2 \{P_1(1 - P_1) + P_2(1 - P_2) - 2\rho\sqrt{P_1(1 - P_1)P_2(1 - P_2)}\}}{(P_1 - P_2)^2} \quad (30)$$

Here n stands for the total number of subjects in the study, thus half of the sample size n should be allocated to the sequence AB and the other half to BA. As an example, consider a clinical 2×2 crossover study where a test compound to treat sleep disorders will be compared against a placebo. The trial consists of two nights with a proper wash out in between nights. The primary endpoint is if the subject experienced weakness after the onset of sleep. The proportion of subjects experiencing weakness without any medication is believed to be $P_1 = 0.80$. Researchers expect to reduce this rate to $P_1 = 0.50$ when subjects are on active. From previous studies, researchers expect the correlation

Table 13 Beta Estimates and Standard Errors of Ossification Data for Full Factorial Logistic Regression Model Using GEE with Independence Working Correlation Matrix and with the Morel-Bokossa-Neerchal Correction

Parameter	Model		
	GEE with independence working correlation matrix		GEE with Morel-Bokossa-Neerchal Correction
	Estimate	Standard error	Standard Error
Intercept	0.83	0.25	0.26
TCPO	−0.85	0.41	0.43
PHT	−2.11	0.34	0.37
TCPO+PHT	1.05	0.76	0.80

PHT: phenytoin; TCPO: trichloropropene oxide.

ρ is approximately $\rho = 0.05$. Using Eq. (30), for a two-sided test with 5% Type I error rate and 90% power the estimated number of subjects for the 2×2 cross-over is about $n = 46$.

RANDOM EFFECT MODELS

The Generalized Linear Models (GLM) framework provided by McCullagh and Nelder^[1] unifies the analysis of continuous and discrete response variables when the responses can be assumed to be independent. As previously discussed, in GLM, moments of the response y_{ji} , from the i -th elemental unit from the j -th cluster, $j = 1, 2, \dots, n$, $i = 1, 2, \dots, m_j$, are modeled such that $E(y_{ji}) = \mu_{ji}$ and $V(y_{ji}) = \phi v(\mu_{ji})$, where μ_{ji} depends on a k -dimensional row vector of known covariates $\mathbf{x}_{ji}$ and an unknown k -dimensional column vector of unknown parameters $\boldsymbol{\beta}^0$ through a link function $g(\cdot)$ such that $g(\mu_{ji}) = \mathbf{x}_{ji}\boldsymbol{\beta}^0$. As before, $v(\mu_{ji})$ is a known function that gives the variance of the observation in terms of the mean. As described in “Quasi-likelihood Models,” GLM formulation is closely related to the quasi-likelihood method of obtaining the estimates. However, the quasi-likelihood method does not facilitate the simultaneous estimation of the mean (μ) and dispersion (ρ) parameters in the extra variation models. The GLM framework has been extended by Zeger, Liang, and Albert^[40] to include random effects. The resulting models, *generalized linear mixed models* (GLMM), can accommodate the analysis of longitudinal data, where subjects are monitored over a period of time, and the analysis of observations obtained on different subjects belonging to some kind of cluster, like those usually obtained in cluster sampling. As indicated earlier, overdispersed binomial (or multinomial) is usually a result of natural association among observations from individuals belonging to the same cluster. Thus, we include a brief review of GLMM in this chapter.

Extension of GLM to obtain GLMM is straightforward and is analogous to the extension of linear models to linear mixed models. A random effect is introduced corresponding to each subject, which describes the deviation of the subject's response from the group average. We consider only the simplified case of subjects treated as random. Real life examples might be more complex. For instance, in a multi-center study where subjects within centers provide repeated measurements across time, it would be reasonable to treat “centers” and “subjects within centers” as random. Following the previous notations, under GLMM, the response y_{ji} for the i -th elemental unit in the j -th cluster is assumed to satisfy

$$g(\mu_{ji}) = \mathbf{x}_{ji}\boldsymbol{\beta}^0 + z_{ji}b_j \quad (31)$$

and

$$\text{Var}(y_{ji}|b_j) = \phi v(\mu_{ji}) \quad (32)$$

where $\mu_{ji} = E(y_{ji} | b_j)$ is an independent observation from a mixture distribution. Note that conditional on b_j , Eqs. (31) and (32) describe a GLM with link function $g(\cdot)$ and variance function $v(\cdot)$. Unconditional moments may be obtained by taking expectations of the conditional moments given by Eqs. (31) and (32) with respect to the distribution of b_j 's. Thus under GLMM,

$$\mu_{ji} = E(y_{ji}) = E\{E(\mu_{ji} | b_j)\} = \int g^{-1}(\mathbf{x}_{ji}\boldsymbol{\beta}^0 + z_{ji}b) \partial F(b_j) \quad (33)$$

and

$$\begin{aligned} V(y_{ji}) &= E\{V(y_{ji}|b_j)\} + V\{E(y_{ji}|b_j)\} \\ &= \int \phi V\{g^{-1}(\mathbf{x}_{ji}\boldsymbol{\beta}^0 + z_{ji}b_j)\} \partial F(b_j) + z_{ji}z_{ji}'\sigma_b^2 \end{aligned} \quad (34)$$

Models described by Eqs. (33) and (34) are referred to as *population average* (PA) models by Zeger, Liang, and Albert.^[40] They refer to the models given by Eqs. (31) and (32) as *subject specific* (SS) models. Such models are most suitable when response for an individual (rather than description of the population as a whole) is the focus of the study, as in studies of growth curves. Population-average models are most useful in epidemiological studies. Here the focus is in studying the difference between the average responses of groups. A salient example given by Zeger, Liang, and Albert^[40] is as follows: Suppose x_{ji} indicates whether or not the subject j at time i smokes and y_{ji} is the presence or absence of a respiratory infection. Then the PA model describes the difference in infection rates between smokers and non-smokers, whereas the SS model estimates the expected change in the probability of infection of a subject given a change in smoking behavior.

Estimation of a GLMM can be done, in principle, by specifying a distribution for the random effects b_j 's, obtaining (by integration) the marginal (unconditional) likelihood of the data and then proceeding with likelihood estimation. Even in the case of linear mixed models, closed form for likelihood estimates may not be available. Harville^[65] implements this approach to the linear mixed models, assuming a normal distribution for the random effects and a normal distribution for the data conditional on the random effects. Anderson and Aitkin^[66] and Im and Gianola^[67] use maximum-likelihood estimation in logistic and probit models, where random effects are assumed to follow a normal distribution and the conditional distribution response is binomial. In general, being able to obtain likelihood estimates for GLMM, depends on whether or not the marginal (unconditional) distribution of the y_{ji} 's can be obtained by integrating with respect to a justifiable distribution of the random effects. Zeger and Karim^[37] use the connection between GLMM and the Bayesian framework to give a method of estimation that overcomes the computational limitations using a Gibbs sampling

approach. One way of avoiding numerical computation of these integrals is to approximate the integrals by simple expansions or a method such as Laplace's method of approximating integrals. This approach has been explored by Stiratelli et al.^[68] and Breslow and Clayton^[35]. Zeger, Liang, and Albert^[40] use the generalized estimating equations approach to estimate GLMM. Schall^[39] gives an algorithm for estimation in GLMM which is easy to program. This algorithm also accommodates the specification of just the first two moments of the distribution of the random effects. Drum and McCullagh^[34] adapt the residual maximum-likelihood (REML), a method that has been successfully used in linear mixed models, to GLMM with logistic links. Their [Table 1](#) gives a succinct summary of methods used in estimating GLMM. We now provide an example.

Beitler and Landis^[69] analyzed data from a multicenter randomized clinical trial aimed to investigate the efficacy of two topical cream preparations (active drug and control) in curing an infection. The data are provided in [Table 14](#).

We consider the following logistic model with random effects (clinics as random).

$$t_{ij} \text{ given } (x_{ij}; b_j) \sim \text{Binomial}(\pi_{ij}; m_{ij})$$

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} + b_j$$

$$x_{ij} = 1 \text{ if } i = 1 \text{ (subjects receiving the drug), zero otherwise}$$

$$b_j \text{'s independent and identically distributed Normal } (0, \sigma^2)$$

Table 14 Favorable Responses to Active Drug and Control Treatment from a Multicenter Randomized Clinical Trial

Clinic	Treatment	Favorable responses	Total
1	Drug	11	36
1	Control	10	37
2	Drug	16	20
2	Control	22	32
3	Drug	14	19
3	Control	7	19
4	Drug	2	16
4	Control	1	17
5	Drug	6	17
5	Control	0	12
6	Drug	1	11
6	Control	0	10
7	Drug	1	5
7	Control	1	9
8	Drug	4	6
8	Control	6	7

Here the subscript i , $i = 1, 2$ represents the cream preparation ($i = 1$ active drug, $i = 2$ control). The centers are represented with the subscript j , $j = 1, 2, \dots, 8$. We fit the previous GLMM to the multicenter data using the Laplace method. The maximum likelihood results are provided in [Table 15](#).

The active drug effect is significant at the customary 5%. The estimated odds-ratio turned out to be $e^{0.7382} = 2.09$. An approximate 95% for the odds-ratio is (0.03, 1.45).

We conclude the discussion of GLMM with a few general remarks. GLMM methodology provides useful ways of modeling clustered responses, which occur routinely in biopharmaceutical studies. Responses from subjects who belong to the same neighborhood and responses collected from the same subject over a period of time in a longitudinal research are examples of such data. As seen in "Likelihood Models," the extra variation models presented in this chapter, namely, the beta-binomial models and finite-mixture models, can, in fact, be considered as random effect models. However, it should be noted that the models described in this chapter focus on situations where responses from elemental units within the cluster are not available.

SUMMARY

In this chapter, we presented a description of the phenomenon known as extra variation or overdispersion, along with a review of the literature. There are four different sections, each section describing a different approach for modeling extra variation. These sections are "Quasi-Likelihood Models," "Likelihood Models," "Generalized Estimating Equations Models," and "Random Effect Models." Another section was devoted to a goodness-of-fit statistic for testing the adequacy of overdispersion models. Several examples discussing a variety of real data sets are cited. Quasi-likelihood models are illustrated with data from two clinical studies (pH data of [Table 1](#) and vertebral fracture data of [Table 4](#)) and one preclinical experiment (ossification data of [Table 2](#)). In the section of likelihood models, two distributions for modeling extra variation are presented, namely, the beta-binomial and the finite-mixture. An insightful representation of the finite-mixture model is provided ([Fig. 1](#)). An example on chromosome association is thoroughly discussed ([Table 6](#)), and the example on ossification data of [Table 2](#) is revisited.

Table 15 Maximum Likelihood Results Using the Laplace Method from a Multicenter Randomized Clinical Trial

Parameter	Estimate	Standard error	p-value
Intercept (β_0)	-1.1965	0.55291	0.0672
Drug Effect (β_1)	0.7382	0.3004	0.0436
Variance due to clinics (σ^2)	1.9313	1.16873	
-2* Log Likelihood	74.20		

The sections on generalized estimating equation models and random effect models are clearly presented. The ossification data of Table 2 is used one more time to illustrate the use of the generalized estimating equations approach as well as a small sample bias correction usually embedded into the sandwich estimator. An epidemiological example is described in the section of random-effect models to clarify the difference between population-average models and subject-specific models. An example (Table 14) of a GLMM application using the Laplace estimation method is presented. In practice, overdispersion may be far too common than one would anticipate and therefore we recommend that one should always check for the presence of overdispersion before proceeding with the analysis under the assumption of independence.

REFERENCES

- McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd Ed.; Chapman and Hall: London, 1989, 124–128.
- Stigler, S.M. *The History of Statistics. The Measurements of Uncertainty before 1990*; The Belknap Press of Harvard University: Cambridge, MA, 1986, 229–238.
- “Student.” An explanation of deviations from Poisson’s law in practice. *Biometrika* **1919**, *12*, 211–215.
- Williams, D.A. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **1975**, *31*, 949–952.
- Williams, D.A. Extra-binomial variation in logistic linear models. *Appl. Stat* **1982**, *31*, 144–148.
- Williams, D.A. Extra-binomial variation in logistic linear models. *Biometrics* **1988**, *44*, 305–308.
- Kupper, L.L.; Haseman, J.K. The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics* **1978**, *34*, 69–76.
- Haseman, J.K.; Kupper, L.L. Analysis of dichotomous response data from certain toxicological experiments. *Biometrics* **1979**, *35*, 281–293.
- Kupper, L.L.; Portier, C.; Hogan, M.D.; Yamamoto, E. The impact of litter effects on dose-response modelling in teratology. *Biometrics* **1986**, *42*, 85–98.
- Shirley, E.A.C.; Hickling, R. An evaluation of some statistical methods for analysing numbers of abnormalities found amongst litters in teratology studies. *Biometrics* **1981**, *37*, 819–829.
- Paul, S.R. Analysis of proportions of affected fetuses in teratological experiments. *Biometrics* **1982**, *38*, 361–370.
- Pack, S.E. Hypothesis testing for proportions with overdispersion. *Biometrics* **1986**, *42*, 967–972.
- Lefkopoulou, M.; Moore, D.; Ryan, L. The analysis of multiple correlated binary outcomes: Application to rodent teratology experiments. *J. Am. Stat. Assoc* **1989**, *84*, 810–815.
- Carr, G.; Portier, C.J. An evaluation of some methods for fitting dose-response models to quantal-response developmental toxicology data. *Biometrics* **1993**, *49*, 779–791.
- Donner, A.; Eliasziw, M.; Klar, N. A comparison of methods for testing homogeneity of proportions in teratologic studies. *Stat. Med* **1994**, *13*, 1253–1264.
- Morel, J.G.; Neerchal, N.K. Clustered binary logistic regression in teratology data using a finite mixture distribution. *Stat. Med* **1997**, *16*, 2843–2853.
- Rosner, B. Multivariate methods in ophthalmology with application to other paired-data situations. *Biometrics* **1984**, *40*, 1025–1035.
- Donner, A. Statistical methods in ophthalmology: An adjusted chi-square approach. *Biometrics* **1989**, *45*, 605–611.
- Altham, P.M.E. Discrete variable analysis for individuals grouped into families. *Biometrika* **1976**, *63*, 263–269.
- Cohen, J.E. The distribution of the chi-squared statistic under clustered sampling from contingency tables. *J. Am. Stat. Assoc* **1976**, *71*, 665–670.
- Crowder, M.J. Beta-binomial Anova for proportions. *Appl. Stat* **1978**, *27*, 34–37.
- Heckman, J.J.; Willis, R.J. A beta-logistic model for the analysis of sequential labor force participation by married women. *J. Polit. Econ* **1977**, *85*, 27–58.
- Efron, B.E. Double exponential families and their use in generalized linear regression. *J. Am. Stat. Assoc* **1986**, *81*, 709–721.
- Moore, D.F. Modelling the extraneous variance in the presence of extra-binomial variation. *Appl. Stat* **1987**, *36*, 8–14.
- Gail, M.H.; Santner, T.J.; Brown, C.C. An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics* **1980**, *36*, 255–266.
- Lawless, J.F. Regression methods for Poisson process data. *J. Am. Stat. Assoc* **1987**, *82*, 808–815.
- Breslow, N.E.; Clayton, D.G. Approximate inferences in generalized linear models. *J. Am. Stat. Assoc* **1993**, *88*, 9–25.
- Dean, C.B.; Balshaw, R. Efficiency lost by analyzing counts rather than event times in Poisson and overdispersed Poisson regression models. *J. Am. Stat. Assoc* **1997**, *92*, 1387–1398.
- Zeger, S.L.; Edelstein, S.L. Poisson regression with a surrogate X; an analysis of vitamin A and Indonesian children’s mortality. *Appl. Stat* **1989**, *38*, 309–331.
- Dean, D.B. Testing for overdispersion in Poisson and binomial regression models. *J. Am. Stat. Assoc* **1992**, *87*, 451–457.
- Liang, K.Y.; Zeger, S.L. Longitudinal data analysis using generalized linear models. *Biometrika* **1986**, *73*, 13–22.
- Zeger, S.L.; Liang, K.Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **1986**, *42*, 121–130.
- McGilchrist, C.A. Estimation in generalized mixed models. *J. R. Stat. Soc. Ser. B* **1994**, *56*, 61–69.
- Drum, M.L.; McCullagh, P. REML estimation with exact covariance in the logistic mixed model. *Biometrics* **1993**, *49*, 677–689.
- Breslow, N.E.; Clayton, D.G. Approximate inferences in generalized linear models. *J. Am. Stat. Assoc* **1993**, *88*, 9–25.
- Karim, M.R.; Zeger, S.L. Generalized linear models with random effects; Salamander mating revisited. *Biometrics* **1992**, *48*, 631–644.
- Zeger, S.L.; Karim, M.R. Generalized linear models with random effects: A Gibbs sampling approach. *J. Am. Stat. Assoc* **1991**, *86*, 79–86.

38. Goldstein, H. Nonlinear multilevel models, with an application to discrete response data. *Biometrika* **1991**, 78, 45–51.
39. Schall, R. Estimation in generalized linear models with random effects. *Biometrika* **1991**, 78, 719–727.
40. Zeger, S.L.; Liang, K.Y.; Albert, P.S. Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **1988**, 44, 1049–1060.
41. Nelder, J.A.; Wedderburn, R.W.M. Generalized linear models. *J.R. Stat. Soc. Ser. A* **1972**, 135, 370–384.
42. Wedderburn, R.W.M. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **1974**, 61, 439–447.
43. Cox, D.R.; Snell, E.J. *Analysis of Binary Data*, 2nd Ed.; Chapman and Hall: New York, 1989.
44. Johnson, N.L.; Kotz, S.; Kemp, A.W. *Univariate Discrete Distributions*, 2nd Ed; John Wiley & Sons: New York, 1992, 204.
45. Skellam, J.G. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *J. R. Stat. Soc. Ser. B* **1948**, 10, 257–261.
46. Morel, J.G.; Nagaraj, N.K. A finite mixture distribution for modelling multinomial extra variation. *Biometrika* **1993**, 80, 363–371.
47. Morel, J.G. A simple algorithm for generating multinomial random vectors with extravariation. *Commun. Stat. Simul* **1992**, 21, 1255–1268.
48. Self, S.G.; Liang, K.Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *J. Am. Stat. Assoc* **1987**, 82, 605–610.
49. Neerchal, N.K.; Morel, J.G. Large cluster results for two parametric multinomial extra variation models. *J. Am. Stat. Assoc* **1998**, 93, 1078–1087.
50. Paul, S.R.; Plackett, R.L. Inference sensitivity for Poisson mixtures. *Biometrika* **1978**, 65, 591–602.
51. Collings, B.J.; Margolin, B.H. Testing goodness of fit for the Poisson assumption when observations are not identically distributed. *J. Am. Stat. Assoc* **1985**, 80, 411–418.
52. Rao, C.R. *Linear Statistical Inferences and Its Applications*; John Wiley and Sons: New York, 1965.
53. Cramer, H. *Métodos Matemáticos de Estadística*, 4th Ed.; Ediciones Aguilar: Madrid, 1968.
54. Sutradhar, S.C. *Testing the Goodness-of-Fit of Overdispersion Models for Categorical Data*; Unpublished Dissertation. University of Maryland Baltimore County, 2001.
55. Sutradhar, S.C.; Neerchal, N.K.; Morel, J.G. A goodness-of-fit test for overdispersed binomial or multinomial models. *J. Stat. Plann. Infer* **2008**, 138, 1459–1471.
56. Haseman, J.K.; Soares, E.R. The distribution of fetal death in control mice and its implications on statistical tests for dominant lethal effects. *Mutat. Res* **1976**, 41, 277–288.
57. McCullagh, P. Quasi-likelihood functions. *Ann. Stat* **1983**, 11, 59–67.
58. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*; Chapman and Hall: London, **1983**.
59. Morel, J.G. Logistic regression under complex survey designs. *Surv. Methodol.* **1989**, 15, 203–223.
60. Morel, J.G.; Bokossa, M.C.; Neerchal, N.K. Small sample correction for the variance of GEE estimators. *Biom. J* **2003**, 5, 395–409.
61. Morel, J.G.; Neerchal, N.K. Comparison of four small sample bias corrections of the empirical covariance estimators; University of Maryland Baltimore County, Technical Report, 2006-02.
62. Fay, M.P.; Graubard, B.I. Small-sample adjustments for Walt-type tests using sandwich estimators. *Biometrics* **2001**, 57, 1198–1206.
63. Mancl, L.A.; DeRouen, T.A. A covariance estimator for GEE with improved small-sample properties. *Biometrics* **2001**, 57, 126–134.
64. Liu, G.; Liang, K.Y. Sample size calculations for studies with correlated observations. *Biometrics* **1997**, 53, 937–947.
65. Harville, D.A. Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* **1977**, 72, 320–340.
66. Anderson, D.A.; Aitkin, M. Variance component models with binary response: Interviewer variability. *J. R. Stat. Soc. Ser. B* **1985**, 47, 203–210.
67. Im, S.; Gianola, D. Mixed models for binomial data with an application to lamb mortality. *Appl. Stat.* **1988**, 37, 196–204.
68. Stiratelli, R.; Laird, N.; Ware, J.H. Random-effects models for serial observations with binary response. *Biometrics* **1984**, 40, 961–971.
69. Beittler, P.J.; Landis, J.R. A mixed-effects model for categorical data. *Biometrics* **1985**, 41, 991–1000.

Factor Analysis

Mary D. Sammel and Sarah J. Ratcliffe

Department of Biostatistics and Epidemiology, Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, U.S.A.

Benjamin E. Leiby

Department of Pharmacology and Experimental Therapeutics, Thomas Jefferson University, Philadelphia, Pennsylvania, U.S.A.

Abstract

The goal of factor analysis is data reduction of multiple measurements into a smaller number of latent outcomes or factors that are not directly observable. This entry provides some general background information on factor analysis and discusses the standard methods of in practice for continuous data.

Keywords: Correlation; Data reduction; Latent trait; Measurement model; Multivariate analysis.

INTRODUCTION

Factor analytic methods are among a rich body of literature on statistical methods for multivariate continuous outcomes. The goal of factor analysis is data reduction of multiple measurements into a smaller number of latent outcomes or factors that are not directly observable. The factors are derived via a decomposition of the covariance or correlation matrix of the observed, often called manifest, data. We begin with a brief overview of the controversial origin of this method; describe the standard methods in practice for continuous data; discuss extensions of these methods to data with other distributions assumed; and present an example. We conclude with some discussion and description of further methodological extensions.

BACKGROUND

The development of factor analysis was first attributed to Spearman, who sought to measure “intelligence.”^[1] Although it is often taught as purely deductive mathematics,^[2] this method was derived with the above goal in mind. A student of Spearman’s, Cyril Burt, is credited with the reification of intelligence, that is, utilizing this method to justify his theory about the heritability of intelligence and to derive a unilinear ranking system. Unable to resist human nature, these rankings have been used to discriminate against various subgroups of society. In addition, controversy arose when Thurstone published his work entitled “Vectors of Mind,”^[3] which illustrated the non-uniqueness of Spearman’s principal factors (principal components solution), and proposed rotation of the

factors to identify “logical” vectors. This approach challenged Spearman’s theory that the first principal factor was “general intelligence.” As the two solutions explained an equivalent proportion of the sample variability, both models are equally plausible.

Despite its checkered past, factor analysis has maintained its popularity in the social sciences.^[4] In health research, measurement of abstract concepts such as “quality of life” or “depression” are of interest. While these concepts are generally measured via questionnaires, others have used them to discern structure from biological parameters, such as quantifying the severity of birth defects,^[5] to identify a metabolic syndrome,^[6] and to evaluate the impact of different types of pollution on mortality.^[7] Even when it is hard to justify the theoretical existence of a “latent variable,” these summary scores can be considered logical constructs of the observed data.^[8] There are two approaches to the implementation of factor analysis. The first, termed exploratory, has the goal of identifying the latent structure that is data driven. Confirmatory factor analysis (CFA) begins with a theoretic framework or proposed structure, which is then confirmed by evaluating goodness-of-fit to the observed data. Both frameworks are described below.

EXPLORATORY FACTOR ANALYSIS

Theoretic Model

Exploratory factor analysis (EFA) aims to partition the covariance matrix of the data into two components, a piece that is common to all the measures (or variables), and a piece which is specific to each measurement. In truth,

the underlying structure is unknown. Let $\mathbf{X} = (x_1, \dots, x_p)'$ be a p -variate random (column) vector with mean $\mu = (\mu_1, \dots, \mu_p)'$ and covariance matrix $\Sigma = [\sigma_{ij}]_{p \times p}$. The vector $\mathbf{X}$ is assumed to be linearly dependent on a set of $m < p$ random, unobserved common factors $\mathbf{f} = (f_1, \dots, f_m)$, plus residual or specific factors $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_p)$. Further, assume $E(\mathbf{f}) = \mathbf{0}$, $\text{cov}(\mathbf{f}) = \mathbf{I}$, $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\text{cov}(\boldsymbol{\varepsilon}) = \Psi = \text{diag}(\Psi_1, \dots, \Psi_p)$, and $\text{cov}(\mathbf{f}, \boldsymbol{\varepsilon}) = \mathbf{0}$. Then $\mathbf{X} - \mu = \mathbf{L}\mathbf{f} + \boldsymbol{\varepsilon}$ where $\mathbf{L} = [l_{ij}]_{p \times m}$ is a matrix of factor “loadings” or regression weights. The resulting covariance matrix is

$$\begin{aligned}\Sigma &= \mathbf{L}\mathbf{L}' + \Psi \\ \Rightarrow \sigma_{ii}^2 &= \sum_{k=1}^m l_{ik}^2 + \Psi_i = h_i^2 + \Psi_i\end{aligned}$$

The variance of the i th variable explained by the common factors is called the i th communality, h_i^2 , while Ψ_i is called the unique or specific variance. The above model can easily be generalized so that the common factors are intercorrelated by assuming $\text{cov}(\mathbf{f}) = \Phi \neq \mathbf{I}$. In this case the factors are said to be oblique. When $\text{cov}(\mathbf{f}) = \mathbf{I}$, the factors are said to be orthogonal.

A basic property of EFA is that it is scale independent. That is, if $\mathbf{A}$ is an orthogonal matrix, then $\mathbf{A}\mathbf{X}$ will have loadings $\mathbf{A}\mathbf{L}$ and factors $\mathbf{f}$. Another key property is that $\mathbf{f}$ and $\mathbf{L}$ are not unique except when the dimension of $\mathbf{f}$ is one. So, $\mathbf{L}\mathbf{f} = \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{f}$ for any $m \times m$ orthogonal matrix $\mathbf{T}$. This non-uniqueness property leads to the idea of factor rotation.

Parameter Estimation

Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be a sample of size n observations of random vector $\mathbf{X}$. The factors and loadings are derived from the sample covariance matrix $\mathbf{S}$, an estimator of the true variance/covariance matrix Σ . That is, the estimated EFA model is found by solving $\mathbf{S} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Psi}$. As the estimates for $\mathbf{L}$ and Ψ ($\hat{\mathbf{L}}$ and $\hat{\Psi}$ respectively) are not unique, initial solutions are obtained by imposing further constraints. The most popular methods for obtaining the initial solution are the principal component, principal factor, and maximum likelihood. The solution from any method is then rotated for ease of interpretation. Ideally, each variable will have a high loading on only one factor. One of the most common rotations is the varimax rotation,^[9] which selects the orthogonal matrix $\mathbf{T}$ that maximizes the sum of variances of the squared loadings of each factor. That is, it attempts to have only large or negligible coefficients in any column of the rotated loadings matrix $\hat{\mathbf{L}}$. Other orthogonal rotations include the quarimax and equamax rotations. Oblique rotations, while not orthogonal, can often be easier to interpret but result in correlated factors. One example is the PROMAX rotation,^[10] which takes orthogonally rotated factors (usually the varimax rotation) and rotates them again to oblique factors that approximate a target matrix $\mathbf{Q}$. $\mathbf{Q} = [q_{ij}]_{p \times m}$ is structured such that $q_{ij} = |l_{ij}^{-1}| l_{ij}$ where $r > 1$ is the power of the rotation. Oblique

factors have been demonstrated to replicate better than orthogonal solutions.^[11]

Number of Factors

An important issue is how many factors to include in the model, as different m 's will lead to different solutions and interpretations. Parameter estimation is accomplished conditional on a pre-specified number of factors. When the population is normally distributed, a likelihood ratio test^[12] can be used to test the adequacy of the choice of the number of factors. However, this test can lead to overfitting, that is, selecting an m such that the final factor(s) may be “significant” but provide no additional insight into the data. A number of ad hoc methods have also been proposed. These include using factors with unrotated eigenvalues greater than one, factors that account for approximately 70%–80% of the original variation in the data, and factors with “large” values on a scree plot of the sample eigenvalues of $\mathbf{S} - \Psi$. The term “scree” refers to the debris that falls off the mountain and collects at its base. Therefore, we desire to have our statistical model describing the mountain while discarding the debris. There is no one method giving the “correct” answer. Trying a few values of m is recommended, with the final choice being driven by the subject matter and the interpretability of the solution.

Computation

Most standard statistical software packages have functions or procedures for performing factor analysis. Both the SAS and STATA procedure are called “factor,” while the S-PLUS function is called “factanal.” SPSS can also perform factor analysis via a pull down menu. All packages have the most popular methods for initial estimation and rotations. Each package also has a variety of other techniques available. Estimation of the unobservable factors is often a goal, and can be produced by these software packages.^[13] Once estimated, these factors are often utilized as if they were observed data. However, given that the estimation error is ignored, the results underestimate the true associations.^[14]

CONFIRMATORY FACTOR ANALYSIS

In EFA no prior knowledge about the underlying structure is assumed. That is, the number of factors, m , is unknown. In CFA, the researcher has a preconceived notion of the structure, which is based upon a theoretic framework. This method can also be employed to validate structures previously identified by EFA. Elements of the factor loading matrix are restricted so that factors are associated with specific variables. This approach falls under a broader class of structural equation models.^[15] CFA is often performed using a dedicated software package such as AMOS, EQS,

or LISREL. Alternatively, the factors can be modeled using “proc calis” in SAS.

Once the proposed model is estimated, the goodness-of-fit of the model must be assessed. The first step is generally to check that the model is locally identified, that is, the estimated parameters are of the correct sign and magnitude, variance estimates are positive, and the entries in the residual matrix $S - \hat{\Sigma}$ are uniformly small. Secondly, tests or indices of fit are used to evaluate the overall goodness-of-fit. Over 30 such tests have been proposed.^[16] Owing to their similarity with the R^2 and adjusted R^2 tests in multiple regression, the goodness-of-fit index (GFI) and adjusted GFI are two of the more commonly employed indices. Further, a reasonably fitting model will have a Bentler comparative fit index and a Bentler and Bonett index over 0.9. In practice, more than one test is needed to evaluate the overall fit. Usually, the model does not fit the data. In such cases, the complexity of the proposed model should be reduced, rather than expanded. Likelihood ratio, Lagrange Multiplier, and Wald tests exist to compare nested models.

Extensions

Adoption of factor analytic methods by the medical community has been slow^[17] because they were developed largely for continuous responses. It has been pointed out that ignoring the true distribution of the data and forcing dichotomous or ordinal responses into continuous data models can be problematic.^[18] Extensions of classic factor analytic methods that utilize estimated or polychoric correlations among sets of dichotomous responses or groups of ordered responses have been proposed.^[19] One class of models, Item Response Theory (IRT), was originally developed within the context of educational testing to combine dichotomous measures into a single latent construct. The original form of these IRT models (the Rasch model) utilized a joint logit model to estimate separate difficulty parameters for each item, but assumed a constant association (discrimination parameter) with the latent trait across all items within a domain, allowing for ranking and scaling of items along the range of the latent trait. Extensions to sets of nominal, ordered categorical, or rating scale responses also have been proposed.^[20] When each item is allowed to have a differential association with the latent trait, the discrimination parameters have a similar interpretation to the factor loading parameters in their continuous data counterparts.^[21] Just as with factor analysis, the IRT model evaluates sets of observed outcomes that are all of the same type. Recently, software has been developed allowing the modeling of mixed outcome types.^[22] The MPLUS package can fit EFA models with continuous and ordered categorical data, while the CFA module can handle any combination of continuous, ordered categorical, and count data. This program utilizes a probit distribution for the categorical measurements and, thus, is different from the IRT models.

Although it is easy to conceptualize models for mixed outcome types, estimation via maximum likelihood quickly becomes challenging as this method of estimation requires numerical integration. Bayesian methods provide an alternative to dealing with complex estimation. Bayesian approaches to factor analysis involve specifying a prior distribution for model parameters and then using the data to estimate the posterior distribution.^[8] One advantage of the Bayesian approach is that the choice of appropriate prior distributions will produce valid estimates for all model parameters in contrast to methods such as maximum likelihood.^[23] Given that the posterior density is often not of closed form, numerical methods are required for estimation. Advances in computing power and development of Markov Chain Monte Carlo algorithms have made Bayesian factor analysis models easier to implement in practice and have led to renewed interest in these models. Bayesian methods have been applied to non-linear factor analysis models as well as mixed outcome types and longitudinal data.^[24] The Bayesian framework also allows for a unified procedure of model fitting and selection.^[25] Bayesian factor analysis functions (MCMCfactanal, MCMCordfactanal, MCMCmixfactanal) are available as part of the MCMCpack in R.^[26] The example below compares standard software EFA assuming normality of the observed data to an analysis where the observed form of the data has been properly accounted for.

Example

Our example is an EFA of data from the Interstitial Cystitis Database (ICDB) study where the measured variables are of mixed types (normal, binary, and ordinal).

Interstitial Cystitis (IC) is a chronic syndrome characterized by urinary frequency, urgency, and/or pain in the absence of any identifiable cause. The ICDB cohort study was created to follow the natural history of IC in treated patients.^[27] A secondary goal of the study was to investigate the association of IC symptoms with pathologic bladder features. One-third of the subjects agreed to undergo cystoscopy, hydrodistension, and bladder biopsy at the beginning of the study. Data on bladder biopsies were recorded using a 39-item questionnaire designed to capture a wide range of potential processes affecting the development of IC-related symptoms.^[28] By analyzing the correlation among biopsy features, it is hoped that it will be possible to identify some of the biologic processes contributing to IC and the key features affected by these processes.

For our analysis, we selected a subset of 24 items that had a low percentage of missing data and at least some variability among responses. One continuous variable, 16 binary variables, and 7 ordinal variables constitute these items. Of the 211 subjects who underwent bladder biopsy, 203 had complete data for the selected items and were included in these analyses.

Two factor analyses were performed in MPLUS.^[22] The first assumes that all of the variables are normally

distributed while the second accounts for the fact that the distributions are of different types. Estimation of model parameters was done using the method of unweighted least squares. To determine the number of factors for our model, we created scree plots of the eigenvalues in descending order. In both analyses, a three-factor model was suggested because the slope of the plot was steepest through the third factor (Figs. 1 and 2). Consultation with a pathologist

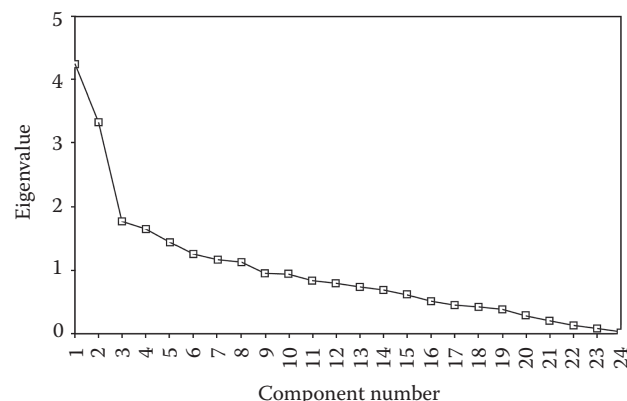


Fig. 1 Scree plot of eigenvalues for analysis 1, assuming all observed variables are normally distributed

confirmed our choice of the three-factor models as they yielded the most biologically plausible interpretation.

Factor loadings using a PROMAX rotation and groupings of the variables based on their factor loadings are presented in Table 1 (for analysis 1) and Table 2 (for analysis 2). The largest factor loading in absolute value for each variable has been highlighted, indicating the factor to which that variable is most strongly related. Factor 2

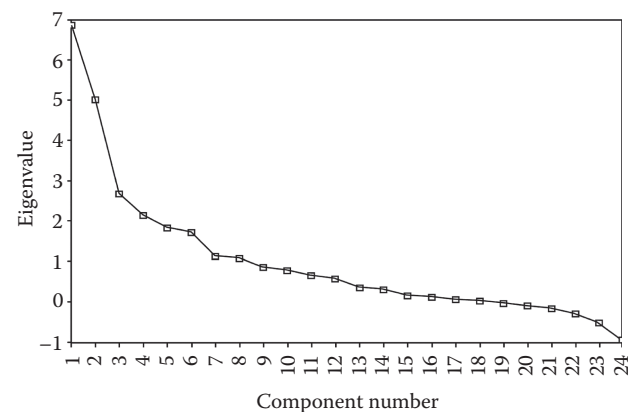


Fig. 2 Scree plot of eigenvalues for analysis 2, assuming mixed outcome types (continuous, binary, and ordered categorical)

Table 1 Factor Loadings for Analysis 1, Assuming All Observed Variables Are Normally Distributed

Variable (codes)	Factor 1	Factor 2	Factor 3
5: Mast cell counts	0.15	−0.07	0.24
9a: Urothelium completely denuded	0.10	0.98	0.11
9b: Urothelial discontinuities	−0.10	−0.93	−0.11
10: Granulation tissue lamina propria	0.76	0.07	−0.07
11: Percent of mucosa denuded of urothelium	0.22	0.68	−0.07
12: Submucosal hemorrhage	−0.09	0.16	0.92
13a: Submucosal hemorrhage aggregated	−0.07	0.01	0.70
13d: Submucosal hemorrhage diffuse	−0.09	0.07	0.60
14: Percentage of submucosa with granulation tissue	0.86	0.15	−0.01
16b: Eosinophilia in lamina propria	0.15	0.06	0.14
17: Mononuclear endothelitis	0.55	0.07	−0.21
18: Transmural mononuclear vasculitis	0.47	−0.13	−0.21
20: Mucosal edema	0.12	−0.09	0.12
22: Reactive urothelial change	0.18	−0.35	−0.03
23: Urothelial hyperplasia	−0.10	−0.08	0.04
25: Urothelial hematopoietic cells	0.10	−0.32	0.24
26: Lamina propria hematopoietic cells	0.56	−0.08	0.27
27a: Lamina propria infiltrate aggregated	0.64	0.08	−0.02
27d: Lamina propria infiltrate diffuse	0.23	−0.17	0.28
30: Lamina propria collage dense	−0.06	0.03	0.11
32: Percentage of lamina propria with nerves	0.41	−0.02	0.15
34: Percentage of lamina propria with S100 positive mononuclear cells	0.12	−0.07	0.00
36: Percentage of lamina propria with vessels	0.25	0.06	0.26
37: Vessels in lamina propria clustered	0.13	−0.06	0.26

Table 2 Factor Loadings for Analysis 2, Assuming Mixed Outcome Types (Continuous, Binary, and Ordered Categorical)

Variable	Factor 1	Factor 2	Factor 3
5: Mast cell counts (continuous)	0.28	−0.16	−0.03
9a: Urothelium completely denuded (0,1)	0.33	1.04	−0.10
9b: Urothelial discontinuities (0,1)	−0.30	−0.90	0.12
10: Granulation tissue lamina propria (0,1)	0.84	0.30	0.05
11: Percent of mucosa denuded of urothelium (0–4)	0.38	0.88	0.07
12: Submucosal hemorrhage (0–3)	0.47	−0.29	0.17
13a: Submucosal hemorrhage aggregated (0,1)	0.42	−0.34	0.26
13d: Submucosal hemorrhage diffuse (0,1)	0.38	−0.28	0.17
14: Percentage of submucosa with granulation tissue (0–3)	1.04	0.39	0.20
16b: Eosinophilia in lamina propria (0,1)	0.55	−0.01	−0.41
17: Mononuclear endothelitis (0,1)	0.77	0.44	−0.58
18: Transmural mononuclear vasculitis (0,1)	0.62	0.11	−0.74
20: Mucosal edema (0,1)	0.48	−0.04	0.75
22: Reactive urothelial change (0,1)	0.14	−0.47	−0.08
23: Urothelial hyperplasia (0,1)	−0.09	−0.18	−0.37
25: Urothelial hematopoietic cells (0–3)	0.13	−0.83	−0.12
26: Lamina propria hematopoietic cells (0–3)	0.76	−0.23	−0.32
27a: Lamina propria infiltrate aggregated (0,1)	0.77	0.20	−0.22
27d: Lamina propria infiltrate diffuse (0,1)	0.39	−0.44	0.01
30: Lamina propria collage dense (0,1)	0.04	0.03	0.43
32: Percentage of lamina propria with nerves (0–3)	0.57	−0.03	−0.09
34: Percentage of lamina propria with S100 positive mononuclear cells (0,1)	0.27	−0.15	−0.65
36: Percentage of lamina propria with vessels (0–2)	0.56	0.05	0.05
37: Vessels in lamina propria clustered (0,1)	0.33	−0.19	0.03

Table 3 Correlations among the Factors for Analyses 1 and 2

	Analysis 1		Analysis 2	
	Factor 1	Factor 2	Factor 1	Factor 2
Factor 2	0.021		−0.211	
Factor 3	0.383	−0.331	−0.030	−0.142

is clearly defined in both analyses and is related to items 9a, 9b, 11, 22, and 25, which record the amount of urothelium lost and the capacity for urothelial regeneration. The remaining items are related most strongly to factors 1 and 3, although the grouping of the variables differs somewhat between the two analyses. These factors represent the role of mastocytosis and cellular inflammation in the development of IC. The difference in item grouping also results in a difference in correlation among the factors (Table 3).

As this example illustrates, assuming normality when it is not appropriate can have a substantial effect on the results of the analysis. This is especially true when the categorical items have asymmetric distributions, as is the case here. In general, the factor loadings for analysis 1 are smaller

than those for analysis 2, while the correlations among the factors are larger for analysis 1. This leads to less certainty about variable groupings and less clearly-defined factors.

CONCLUSIONS

Given today's computing capabilities, Spearman's goal of summarizing multiple measures to derive information about latent constructs has become a reality. Even when this is not a primary goal, his method and various extensions are valuable tools for multivariate data reduction. As with any form of research, there is a need to use the principles of sound statistical practice. Issues of validity, reliability, and potential for bias need to be addressed for all proposed constructs and measurement instruments; Streiner and Norman^[29] describe these issues. Factor analytic methods have their foundation in the social sciences, but this broad class of models can be utilized in many different settings. Muthén^[30] discusses links between these models, random effects, and causal models, as well as extensions including latent class and growth curve models.

REFERENCES

1. Spearman, C. General intelligence objectively determined and measured. *Am. J. Psychol.* **1904**, *15*, 201–293.
2. Gould, S.J. The real error of Cybil Burt: factor analysis and the reification of intelligence. *The Mismeasure of Man*; W.W. Norton and Company Inc.: New York, 1981, 234–320.
3. Thurstone, L.L. *The Vectors of Mind*; University of Chicago Press: Chicago, IL, 1935.
4. Everitt, B.S. *An Introduction to Latent Variable Models*; Chapman and Hall: New York, 1984.
5. Sammel, M.D.; Ryan, L.M. Latent Variable Models with Fixed Effects. *Biometrics* **1996**, *52*, 650–663.
6. Henley, A.J.; Festa, A.; D'Agostino, R.B.; Wagenknecht, L.E.; Savage, P.J.; Tracy, R.P.; Saad, M.F.; Haffner, S.M. Metabolic and inflammation variable clusters and prediction of type 2 diabetes: factor analysis using directly measured insulin sensitivity. *Diabetes* **2004**, *53* (7), 1773–1781.
7. Laden, F.; Neas, L.M.; Dockery, D.W.; Schwartz, J. Association of fine particulate matter from different sources with daily mortality in six US cities. *Environ. Health Perspect.* **2000**, *108* (10), 941–947.
8. Bartholomew, D.J. Posterior analysis of the factor model. *Br. J. Math. Stat. Psychol.* **1981**, *34*, 93–99.
9. Kaiser, H. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **1958**, *23*, 187–200.
10. Hendrickson, A.E.; White, P.O. Promax: a quick method for rotation to orthogonal oblique structure. *Br. J. Stat. Psychol.* **1964**, *17*, 65–70.
11. Coste, J.; Fermanian, J.; Venot, A. Methodological and statistical problems in the construction of composite measurement scales: a survey of six medical and epidemiological journals. *Stat. Med.* **1995**, *14*, 331–345.
12. Bartlett, M.S. A note on the multiplying factors for various χ^2 approximations. *J. Royal Stat. Soc., Ser. B* **1954**, *16*, 296–298.
13. Johnson, R.A.; Wichern, D.W. *Applied Multivariate Statistical Analysis*, 5th Ed.; Prentice Hall: Upper Saddle River, NJ, 1992.
14. Sammel, M.D.; Ryan, L.M. Effects of covariance misspecification in a latent variable model for multiple outcomes. *Stat. Sinica* **2002**, *12* (4), 1207–1222.
15. Bollen, K.A. *Structural Equations with Latent Variables*; John Wiley: New York, 1989.
16. March, H.W.; Balla, J.R.; McDonald, R.P. Goodness-of-fit and indices in confirmatory factor analysis. *Psychol. Bull.* **1988**, *103*, 391–410.
17. Bentler, P.M.; Stein, J.A. Structural equation models in medical research. *Stat. Methods Med. Res.* **1992**, *1*, 159–181.
18. Glonek, G.F.V.; McCullagh, P. Multivariate logistic models. *J. Royal Stat. Soc., Ser. B Method.* **1995**, *57*, 533–546.
19. Olsson, U.L.F. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **1979**, *44*, 443–460.
20. *Handbook of Modern Item Response Theory*; Springer: New York, 1996.
21. Douglas, J.A. Item response models for longitudinal quality of life data in clinical trials. *Stat. Med.* **1999**, *18*, 2917–2931.
22. Muthén, L.K.; Muthén, B.O. *MPLUS User's Guide*, 3rd Ed.; Muthén and Muthén: Los Angeles, CA, 1998–2004.
23. Martin, J.K.; McDonald, R.P. Bayesian estimation in unrestricted factor analysis: a treatment for Heywood Cases. *Psychometrika* **1975**, *40*, 505–517.
24. Dunson, D.B. Bayesian methods for latent trait modelling of longitudinal data. *Stat. Methods Med. Res.* **2007**, *16*, 399–415.
25. Lopes, H.F.; West, M. Bayesian Model assessment in factor analysis. *Statistica Sinica* **2004**, *14*, 41–67.
26. Martin, A.D.; Quinn, K.M.; Park, J.H. MCMCpack: Markov chain Monte Carlo (MCMC) Package. R package version 1.0–3, 2009. Available at: <http://CRAN.R-project.org/package=MCMCpack>.
27. Simon, L.J.; Landis, J.R.; Tomaszewski, J.E.; Nyberg, L.M. The ICDB Study Group. The Interstitial Cystitis Data Base (ICDB) Study. *Interstitial Cystitis*; Lippincott-Raven Press: Philadelphia, PA, 1997; 17–24.
28. Tomaszewski, J.E.; Landis, J.R.; Russack, V.; Williams, T.; Wang, L.P.; Hardy, C.; Brensinger, C.; Matthews, Y.L.; Abele, S.T.; Kusek, J.W.; Nyberg, L.M. The Interstitial Cystitis Database Group. Biopsy features are associated with primary symptoms in interstitial cystitis: results from the Interstitial Cystitis Database study. *Urology* **2001**, *57* (Suppl. 6A), 67–81.
29. Streiner, D.L.; Norman, G.R. *Health Measurement Scales: A Practical Guide to Their Development and Use*, 2nd Ed.; Oxford University Press: New York, 1995.
30. Muthén, B.O. Beyond SEM: general latent variable modeling. *Behaviormetrika* **2002**, *29* (1), 81–117.

Factorial Designs

Junfang Li

Aventis Pharma, Ltd., Tokyo, Japan

Abstract

Factorial design theory is emerging as a fast, effective way of accumulating knowledge in most major sciences. This entry gives a brief review of the history of factorial design and some relevant theories developed recently and also touches on applications and clinical trials using the method.

INTRODUCTION

Factorial design theory is emerging as a fast, effective way of accumulating knowledge in most major sciences. It has been widely used in agricultural research, quality improvement, and cost reduction. Recently, we have seen its application growing fast in biology, medical research, and clinical trials. In biological processes, for example, the polymerase chain reaction (PCR) of DNA diagnosis and identification is affected by many factors. In a clinical study, when patients need multiple therapies, such as combinations of several medications, investigators may have to study the toxicity and benefit of different drug combinations.^[1] A factorial design can be a cost-effective way to study multiple factors in a single study.

During the last two decades, new theory and methodology, such as search design, sequential probing design, and orthogonal and optimal parallel flats type designs, have been developed in the field of factorial designs. The advancement in factorial design provides a powerful instrument for researchers to generate, interpret and apply scientific experiments much more effectively. We realize that there is a great need for understanding new theory and implementing new methodology in practical problems. However, many of the new techniques can only be found in statistical literature, and not organized for the comprehension of applied statisticians.

In this entry, the brief review of the history of factorial design and some relevant theories developed recently are given. A series of orthogonal designs, which can only be generated by advanced design theories, are presented for some specified two-factor interactions. In addition, applications of mixed orthogonal designs in the PCR process of DNA identification and drug interactions in clinical trials are discussed in the section of “Orthogonal Designs for 3^m Factorial Experiments.” This work may guide applied statisticians to implement optimal designs for practical problems without going through the lengthy and complex mathematics involved in generating orthogonal designs.

HISTORY

Factorial design theory was first introduced by Fisher^[2] in agricultural applications. Bose^[3] established the mathematical foundation of experimental design theory by applying Galois fields and the associated finite geometries. For a symmetric s^n factorial experiment, Rao^[4] showed how to construct an orthogonal array of strength $2d$ [OA($2d$)] from a single flat design for estimating all effects up to d -factor interactions.

A design is said to be orthogonal if and only if the corresponding information matrix is diagonal. Orthogonal designs are optimal designs with respect to many optimality criteria, such as D -, E -, and A -optimality. Using an orthogonal design, the estimators of effects are uncorrelated and the variances of estimates are minimized. Orthogonal arrays play an important role in optimal designs. If a design is an orthogonal array of strength $2d$, it is an optimal design with respect to general criteria^[5,6] from which all factorial effects up to d -factor interactions can be estimated, with the assumption that other higher-order interactions are zero.

In practice, it is difficult to use orthogonal arrays of strength $2d$ for $d > 2$. The number of treatments in orthogonal arrays of the strength $2d$ increases exponentially when the number of factors increases. For an experiment with a large number of factors, the number of treatments in an orthogonal array is too large. Often it is not practical to collect all data corresponding to the treatments. In addition, it may be unnecessary to estimate all effects up to d -factor interactions. Prior knowledge may indicate that some higher-order interactions can be ignored. Therefore it is desirable to develop orthogonal designs with a small number of treatments in order to estimate a specified set of higher-order interactions (not necessarily all effects up to d -factor interactions).

One effective way to construct factorial designs is to use “parallel flats.” Connor and Young^[7] introduced the class of parallel flats designs (PFD). Srivastava and Throop^[8] showed how to generate an OA($2d$) of PFD. The general

theory of parallel flats designs is further studied in Refs. [9] and [10]. In her dissertation, Li^[11] gave a necessary and sufficient condition to generate an orthogonal parallel-flats type of design for estimating a specified set of s^n factorial effects. Series of orthogonal designs for 2^n experiments were also presented in Ref. [11]. These results were also published in Ref. [12]. Using representations of complex theory in factorial effects, Liao^[13] extended the foregoing work to mixed $2^n \times 3^m$ factorial experiments.

ORTHOGONAL DESIGNS FOR S^m FACTORIAL EXPERIMENTS

Consider an s^m factorial experiment. Let $\underline{Y} = (y_1, y_2, \dots, y_N)'$ be an $(N \times 1)$ vector of observations corresponding to N treatment combinations in a design T of size $(m \times N)$. Consider a linear model

$$E(\underline{Y}) = X\underline{F}, \quad \text{var}(\underline{Y}) = \sigma^2 I_N \quad (1)$$

where $\underline{F}$ is a $(\nu \times 1)$ vector of unknown effect parameters, X is the $(N \times \nu)$ design matrix of rank ν , and I_N is the $(N \times N)$ identity matrix. The elements of $\underline{F}$ form a subset of all factorial effects, assuming other effects not in $\underline{F}$ are zero. The estimate of $\underline{F}$ is given by $\hat{\underline{F}} = (X'X)^{-1}X'\underline{Y}$, and the variance matrix of $\hat{\underline{F}}$ is $\text{var}(\hat{\underline{F}}) = \sigma^2(X'X)^{-1}$. The corresponding information matrix is $M = X'X$. The design T is said to be an orthogonal design if and only if M is diagonal.

Consider the defining vector for any effect in $\underline{F}$. Suppose that $F_1, F_2, \dots, F_m$ are m factors of interest, and each factor takes three levels. A general factorial effect may be denoted as

$$e = F_1^{\varepsilon_1} \dots F_m^{\varepsilon_m} \quad (2)$$

where $\varepsilon_i \in \text{GF}(3)$ ($i = 1, \dots, m$). The vector

$$\underline{e} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m)' \quad (3)$$

of size $(m \times 1)$ is called the defining vector of the factorial effect e . The general mean μ will have the defining vector $\underline{0}$.

Consider the parallel-flats type of designs for s^m factorial experiments. Suppose that T_i is an $(m \times s^{m-r})$ matrix, whose columns $\underline{t}$ are the solutions to the linear system

$$A\underline{t} = \underline{c}_i, \quad (i = 1, \dots, f) \quad (4)$$

over $\text{GF}(3)$, where A is an $(r \times m)$ matrix of rank r and $\underline{t}$ and $\underline{c}_i$ are vectors of the sizes $(m \times 1)$ and $(r \times 1)$, respectively. Let T be the collection of all the T_i ; i.e.,

$$T = [T_1 : T_2 : \dots : T_i : \dots : T_f] \quad (5)$$

Notice that T is a matrix of size $(m \times N)$, where $N = fs^{m-r}$. Let

$$A = [B : I_{n-r}] \quad \text{and} \quad C = (\underline{c}_1, \dots, \underline{c}_f) \quad (6)$$

The design T is called a parallel-flats type of design determined by the pair of matrices (A, C) , or given by Eq. 4.

In Ref. [12], a simple necessary and sufficient condition on A and C such that T is an orthogonal design is given for a general s^m factorial experiment, where s is a prime or a power of a prime. Theorem 1 provides a way to construct an orthogonal design with a possibly smaller number of treatments.

Theorem 1. Consider the linear model of Eq. 1.^[12] Let T be determined by (A, C) as in Eqs. 4–6. Then T is an orthogonal design for estimating $\underline{F}$ if and only if the following holds.

1. Let e be any effect in $\underline{F}$ with the defining vector $\underline{e}$, if $\underline{e} = \underline{u}'A$, where the vector $\underline{u}(r \times 1)$ is not zero, then $\underline{u}'C(1 \times f)$ must be an OA(1).
2. Let e_1 and e_2 be any two effects in $\underline{F}$ with the defining vectors $\underline{e}_1$ and $\underline{e}_2$, respectively. If $\underline{e}_1 \neq \beta \underline{e}_2$, and $\underline{e}_1 + \gamma \underline{e}_2 = \underline{u}'A$ over $\text{GF}(3)$, where β and γ are nonzero in $\text{GF}(s)$ and $\underline{u}$ is a nonzero vector of size $(r \times 1)$, then $\underline{u}'C$ must be an OA(1).

Now we apply Theorem 1 to generate orthogonal designs for intermediate two-factor interactions, i.e., a subset of all two-factor interactions. For simplicity, we denote any vector of parameters as follows:

$$\begin{aligned} \underline{F}' &= (\mu, F_1, F_1^2, \dots, F_i, F_i^2, F_i^{\varepsilon_i} F_j^{\varepsilon_j}, \dots) \\ &= (0, 1, 1^2, \dots, i, i^2, i^{\varepsilon_i} j^{\varepsilon_j}, \dots) \end{aligned}$$

For example,

$$\begin{aligned} \underline{F}' &= (\mu, F_1, F_1^2, F_2, F_2^2, F_1 F_2, F_1^2 F_2, F_1 F_2^2, F_1^2 F_2^2) \\ &= (0, 1, 1^2, 2, 2^2, 12, 1^2 2, 12^2, 1^2 2^2) \end{aligned}$$

In “Orthogonal Designs for 2^n Factorial Experiments” and “Orthogonal Designs for 3^m Factorial Experiments,” N is the total number of treatments in a design, n (or m) is the number of factors of interest, and ν is the number of parameters in $\underline{F}$. We assume that n (or m) factors are divided into p subgroups, and each subgroup has m_k ($k = 1, \dots, p$) factors in it. The two-factor interactions are either within or between some of the subgroups. All the designs of the following parallel-flats type satisfy the conditions in Theorem 1.

ORTHOGONAL DESIGNS FOR 2^n FACTORIAL EXPERIMENTS

In this section, we provide a class of examples (from 2^n factorial experiments) to illustrate the foregoing theory on orthogonal designs. Given the theory of the last section for any given situation, it is generally relatively simple to obtain A and C matrices by trial and error. Our illustrations include cases where we wish to estimate the general

mean, the main effects, and some two-factor interactions, assuming the remaining factorial effects are negligible. Our designs are such that they enable us to estimate the required parameters under the given conditions. However, in many cases, they may turn out to be better than what we claim; that is, they may enable us to estimate more parameters or to estimate parameters under less-restrictive conditions. Illustrations for other cases of the 2^n series and for general s series or mixed cases will be given elsewhere.

In practice, we often do not have to estimate all the two-factor interactions. We may be interested only in estimating some specific type of two-factor interactions plus the general mean and all the main effects because the remaining effects may be known to be negligible. We shall refer to this as the case of “intermediate resolution.” For example, it is very common to see that elderly patients and AIDS patients take combination treatments. Suppose that a study patient population takes four active combination treatments and that each active treatment has two dose levels. To study the effect of interactions within the first three active treatments, a 2^4 factorial experiment given in Case I, Example 1, may be considered.

For a 2^n factorial experiment, orthogonal fractional designs can be used to estimate the factorial effects by applying the classical theory of experimental design. However, the number of treatments is a power of 2. When the number of factors n is large, this generally gives a large number of treatments. In practice, an individual experimental treatment could be very expensive. Therefore one tries to minimize the number of treatments required to estimate effects. Theorem 1 provides a way to construct an orthogonal design with a possibly smaller number of treatments. By applying it, a number of orthogonal designs are presented in this section for 2^n factorial series of intermediate resolution, where n is in the practical range. In the following, we discuss the intermediate situation in detail.

Suppose that n factors are divided into p subgroups, where the i th ($i = 1, \dots, p$) subgroup has n_i factors in it. Let p' be the number of subgroups of interest, where $p' = p$ or $p' = p - 1$, according to whether the p th subgroup is ignored or not. In other words, if $p' = p - 1$, the factors in the p th subgroup are not involved in the higher-order-factor interactions. Consider the following three cases:

- The interactions within p' subgroups.
- The interactions that involve factors from different subgroups.
- The interactions that arise within some of the subgroups or between some of the subgroups.

We now present the designs for different cases. For convenience, we list the parameters and symbols:

n = number of factors under consideration
 p = number of subgroups

p' = number of subgroups actually involved in the two-factor interactions

n_i = number of the factors in the i th ($i = 1, \dots, p$) subgroup

ν = number of nonnegligible effects

f = number of flats

N = total number of treatments

Let $\tilde{B} = (\underline{b}'_1, \dots, \underline{b}'_p)$ and $\tilde{C} = (\underline{c}'_1, \dots, \underline{c}'_{p'})$, where $\underline{b}'_k$ ($k = 1, \dots, p$) is the k th row of B in Eq. 6 and $\underline{c}'_k$ ($k = 1, \dots, p'$) is the k th row of C . Here, the $\tilde{B}$ is of size $[r(n-r) \times 1]$ and is a “rolled out” form of B ; similarly, $\tilde{C}$ is $(rf \times 1)$ and is a “rolled out” form of C . We also denote the parameter vector $\underline{F} = (\mu, F_1, \dots, F_p, \dots, F_n, F_1F_2, \dots, F_iF_j)'$ as $\underline{F} = F(0, 1, \dots, i, \dots, n, 1:2, \dots, i:j)$. For example, $\underline{F} = (\mu, F_1, F_2, F_3, F_1F_2, F_1F_3, F_2F_3)'$ as $\underline{F} = F(0, 1, 2, 3, 1:2, 1:3, 2:3)$. Hereafter all the designs presented are determined by pairs (A, C) , where (A, C) satisfy the condition in Theorem 1 and where A and C are defined by the corresponding $\tilde{B}$ and $\tilde{C}$.

Case I: Interactions Within Subgroups

Here the n factors are partitioned into p subgroups, the i th subgroup has n_i factors in it, and the two-factor interactions are within p' subgroups, where $p' = p$ or $p' = p - 1$. In Table 1, we list the parameters n, ν, N, f, p, p' , and $n_1, \dots, n_p$ for each of 21 designs, where the number of factors n is from 4 to 10. The designs in Table 1 utilize a single flat

Table 1 Designs for Interactions Within Subgroups

id	n	ν	N	f	p	p'	$(n_1, \dots, n_p)$
1	4	8	8	1	2	1	(3, 1)
2	5	7	8	1	2	1	(2, 3)
3	6	8	8	1	2	1	(2, 4)
4	6	10	16	1	3	3	(2, 2, 2)
5	6	13	16	1	1	1	(4, 2)
6	6	11	16	1	3	2	(3, 2, 1)
7	7	14	16	1	2	1	(4, 3)
8	7	11	16	1	4	3	(2, 2, 2, 1)
9	7	23	32	1	2	1	(6, 1)
10	8	13	16	1	4	4	(2, 2, 2, 2)
11	8	15	16	1	2	1	(4, 4)
12	8	24	32	1	2	1	(6, 2)
13	8	20	32	1	3	2	(2, 5, 1)
14	9	16	16	1	2	1	(4, 5)
15	9	14	16	1	5	4	(2, 2, 2, 2, 1)
16	9	25	32	1	2	1	(6, 3)
17	9	19	32	1	3	1	(3, 3, 3)
18	10	14	16	1	2	1	(3, 7)
19	10	16	16	1	5	5	(2, 2, 2, 2, 2)
20	10	26	32	1	2	1	(6, 4)
21	10	20	32	1	4	3	(3, 3, 3, 1)

(i.e., C is a vector, say, $\underline{c}$). Without loss of generality, we just take $\underline{c} = \underline{0}$. So in the following, we present only the vector of effects $\underline{F}$, the parameters n, N, ν , and the matrices $\tilde{B}$. The set T of the solutions to the linear system $A\underline{t} = \underline{0}$ is then an orthogonal design for estimating the corresponding vector of effects $\underline{F}$. Next, the 21 designs are presented as Cases 1–21.

1. Consider a 2^4 factorial experimental design (Case 1 in Table 1). We are interested in estimating the general mean, all the main effects, and the two-factor interactions within the subgroup $\{F_1, F_2, F_3\}$. The factor F_4 is not involved in the two-factor interactions. Hence we have $n = 4, \nu = 8, p = 2, p' = 1, (n_1, n_2) = (3, 1)$, which are listed in Table 1. The vector of effects is $\underline{F} = (\mu, F_1, F_2, F_3, F_4, F_1F_2, F_1F_3, F_2F_3)' = F(0, 1, 2, 3, 4, 1\cdot2, 1\cdot3, 2\cdot3)$. The orthogonal design T is the set of solutions to the linear equation $A\underline{t} = \underline{0}$ over $\text{GF}(2)$, with total number of treatments $N = 8$, where $A = (1, 1, 1, 1)$ and $\tilde{B} = (111)$.

Cases 2–21 in Table 1 are similar to the preceding. We list only the vector of effects $\underline{F}$, parameters n, N, ν , and the $\tilde{B}$ vector for each case.

2. $\underline{F} = F(0, 1, \dots, 5, 1\cdot2), n = 5, N = 8, \nu = 7, \tilde{B} = (101, 011)$.
3. $\underline{F} = F(0, 1, \dots, 6, 1\cdot2), n = 6, N = 8, \nu = 8, \tilde{B} = (110, 011, 111)$.
4. $\underline{F} = F(0, 1, \dots, 6, 1\cdot2, 3\cdot4, 5\cdot6), n = 6, N = 16, \nu = 10, \tilde{B} = (0101, 1010)$.
5. $\underline{F} = F(0, 1, \dots, 6, 1\cdot2, 1\cdot3, 1\cdot4, 2\cdot3, 2\cdot4, 3\cdot4), n = 6, N = 16, \nu = 13, \tilde{B} = (1110, 0111)$.
6. $\underline{F} = F(0, 1, \dots, 6, 1\cdot2, 1\cdot3, 2\cdot3, 4\cdot5), n = 6, N = 16, \nu = 11, \tilde{B} = (0011, 1101)$.
7. $\underline{F} = F(0, 1, \dots, 7, 1\cdot2, 1\cdot3, 1\cdot4, 2\cdot3, 2\cdot4, 3\cdot4), n = 7, N = 16, \nu = 14, \tilde{B} = (0110, 0101, 1111)$.
8. $\underline{F} = F(0, 1, \dots, 7, 1\cdot2, 3\cdot4, 5\cdot6), n = 7, N = 16, \nu = 11, \tilde{B} = (1100, 0011, 0101)$.
9. $\underline{F} = F(0, 1, \dots, 7, 1\cdot2, \dots, 1\cdot6, 2\cdot3, \dots, 2\cdot6, \dots, 5\cdot6), n = 7, N = 32, \nu = 23, \tilde{B} = (11001, 00111)$.
10. $\underline{F} = F(0, 1, \dots, 8, 1\cdot2, 3\cdot4, 5\cdot6, 7\cdot8), n = 8, N = 16, \nu = 13, \tilde{B} = (1100, 0011, 0101, 1011)$.
11. $\underline{F} = F(0, 1, \dots, 8, 1\cdot2, 1\cdot3, 1\cdot4, 2\cdot3, 2\cdot4, 3\cdot4), n = 8, N = 16, \nu = 15, \tilde{B} = (1100, 1010, 1001, 0111)$.
12. $\underline{F} = F(0, 1, \dots, 8, 1\cdot2, \dots, 1\cdot6, 2\cdot3, \dots, 2\cdot6, \dots, 5\cdot6), n = 8, N = 32, \nu = 24, \tilde{B} = (11010, 10101, 00011)$.
13. $\underline{F} = F(0, 1, \dots, 8, 1\cdot2, 3\cdot4, \dots, 3\cdot7, 4\cdot5, \dots, 4\cdot7, \dots, 6\cdot7), n = 8, N = 32, \nu = 20, \tilde{B} = (10001, 11100, 00111)$.
14. $\underline{F} = F(0, 1, \dots, 9, 1\cdot2, 1\cdot3, 1\cdot4, 2\cdot3, 2\cdot4, 3\cdot4), n = 9, N = 16, \nu = 16, \tilde{B} = (1100, 1010, 1001, 0111, 1111)$.
15. $\underline{F} = F(0, 1, \dots, 9, 1\cdot2, 3\cdot4, 5\cdot6, 7\cdot8), n = 9, N = 16, \nu = 14, \tilde{B} = (1100, 1001, 1101, 0011, 1111)$.
16. $\underline{F} = F(0, 1, \dots, 9, 1\cdot2, \dots, 1\cdot6, 2\cdot3, \dots, 2\cdot6, \dots, 5\cdot6), n = 9, N = 32, \nu = 25, \tilde{B} = (10100, 10010, 10001, 00111)$.
17. $\underline{F} = F(0, 1, \dots, 9, 1\cdot2, 1\cdot3, 2\cdot3, 4\cdot5, 4\cdot6, 5\cdot6, 7\cdot8, 7\cdot9, 8\cdot9), n = 9, N = 32, \nu = 19, \tilde{B} = (10100, 01010, 11001, 10111)$.

18. $\underline{F} = F(0, 1, \dots, 10, 1\cdot2, 1\cdot3, 2\cdot3), n = 10, N = 16, \nu = 14, \tilde{B} = (1100, 1010, 1001, 0111, 1111, 0110)$.
19. $\underline{F} = F(0, 1, \dots, 10, 1\cdot2, 3\cdot4, 5\cdot6, 7\cdot8, 9\cdot10), n = 10, N = 16, \nu = 16, \tilde{B} = (1010, 0101, 0110, 1101, 1001, 1110)$.
20. $\underline{F} = F(0, 1, \dots, 10, 1\cdot2, \dots, 1\cdot6, 2\cdot3, \dots, 2\cdot6, \dots, 5\cdot6), n = 10, N = 32, \nu = 26, \tilde{B} = (01100, 01010, 01001, 11110, 10011)$.
21. $\underline{F} = F(0, 1, \dots, 10, 1\cdot2, 1\cdot3, 2\cdot3, 4\cdot5, 4\cdot6, 5\cdot6, 7\cdot8, 7\cdot9, 8\cdot9), n = 10, N = 32, \nu = 20, \tilde{B} = (11000, 01001, 00111, 00101, 10011)$.

Case II: Interactions Between Subgroups

In this section, we consider designs of intermediate resolution, where the two-factor interactions arise between the subgroups. In Table 2, 35 such cases are listed; the parameters n, ν, N, f, p, p' , and $n_1, \dots, n_p$ are also given for each case.

The first 12 designs are given by a single flat. We select $\underline{c} = \underline{0}$. Next, we present the vector of effects $\underline{F}$, the parameters n, N, ν , and the $\tilde{B}$ matrix for each case.

1. $\underline{F} = F(0, 1, 2, 3, 4, 1\cdot2, 1\cdot3), n = 4, N = 8, \nu = 7, \tilde{B} = (111)$.
2. $\underline{F} = F(0, 1, \dots, 5, 1\cdot2, 1\cdot3), n = 5, N = 8, \nu = 8, \tilde{B} = (110, 011)$.
3. $\underline{F} = F(0, 1, \dots, 6, 1\cdot4, 1\cdot5, 1\cdot6, 2\cdot4, 2\cdot5, 2\cdot6, 3\cdot4, 3\cdot5, 3\cdot6), n = 6, N = 16, \nu = 16, \tilde{B} = (1111, 0111)$.
4. $\underline{F} = F(0, 1, \dots, 6, 1\cdot2, \dots, 1\cdot6), n = 6, N = 16, \nu = 12, \tilde{B} = (1110, 0011)$.
5. $\underline{F} = F(0, 1, \dots, 7, 1\cdot2, \dots, 1\cdot7), n = 7, N = 16, \nu = 14, \tilde{B} = (1110, 0101, 0011)$.
6. $\underline{F} = F(0, 1, \dots, 7, 1\cdot4, 1\cdot5, 1\cdot6, 1\cdot7, 2\cdot4, 2\cdot5, 2\cdot6, 2\cdot7, 3\cdot4, 3\cdot5, 3\cdot6, 3\cdot7), n = 7, N = 32, \nu = 23, \tilde{B} = (01111, 11111)$.
7. $\underline{F} = F(0, 1, \dots, 8, 1\cdot2, 1\cdot3, \dots, 1\cdot8), n = 8, N = 16, \nu = 16, \tilde{B} = (0111, 1010, 1001, 1011)$.
8. $\underline{F} = F(0, 1, \dots, 8, 1\cdot6, 1\cdot7, 1\cdot8, 2\cdot6, 2\cdot7, 2\cdot8, 3\cdot6, 3\cdot7, 3\cdot8, 4\cdot6, 4\cdot7, 4\cdot8, 5\cdot6, 5\cdot7, 5\cdot8), n = 8, N = 16, \nu = 16, \tilde{B} = (11000, 01111, 1111)$.

The following orthogonal designs of parallel-flats type consist of more than one flat.

9. Consider a 2^6 factorial experiment. We are interested in estimating the general mean, all the main effects, and the two-factor interactions between the subgroup $\{F_1, F_2, F_3\}$ and the subgroup $\{F_4, F_5\}$. The factor F_6 is not involved into the two-factor interactions. So we have $n = 6, p = 3, p' = 2, \nu = 13, (n_1, n_2, n_3) = (3, 2, 1)$, which are listed in Table 2, Case 9. The vector of effects is $\underline{F} = F(0, 1, \dots, 6, 1\cdot4, 1\cdot5, 2\cdot4, 2\cdot5, 3\cdot4, 3\cdot5)$. The orthogonal design with total number of treatments $N = 16$ is determined by (A, C) , where

Table 2 Designs for Interactions Between Subgroups

id	n	ν	N	f	p	p'	$(n_1, \dots, n_p)$
1	4	8	8	1	2	2	(1, 3)
2	5	7	8	1	3	2	(1, 2, 2)
3	6	1	16	1	2	2	(3, 3)
4	6	12	16	1	2	2	(1, 5)
5	7	14	16	1	2	2	(1, 6)
6	7	23	32	1	2	2	(3, 4)
7	8	16	16	1	2	2	(1, 7)
8	8	24	32	1	2	2	(5, 3)
9	6	13	16	4	3	2	(3, 2, 1)
10	9	18	24	12	2	2	(1, 8)
11	10	20	24	12	2	2	(1, 9)
12	11	22	24	12	2	4	(1, 10)
13	11	31	48	12	3	3	(1, 1, 9)
14	12	24	24	12	2	2	(1, 11)
15	12	34	48	12	3	3	(1, 1, 10)
16	13	37	48	12	3	3	(1, 1, 11)
17	14	38	48	12	4	3	(1, 1, 11, 1)
18	17	34	40	20	2	2	(1, 16)
19	18	36	40	20	2	2	(1, 17)
20	18	52	80	20	3	3	(1, 1, 16)
21	19	38	40	20	2	2	(1, 18)
22	19	55	80	20	3	3	(1, 1, 17)
23	20	40	40	20	2	2	(1, 19)
24	20	58	80	20	3	3	(1, 1, 18)
25	21	61	80	20	3	3	(1, 1, 19)
26	22	62	80	20	3	2	(1, 1, 19, 1)
27	21	42	48	24	2	2	(1, 20)
28	22	44	48	24	2	2	(1, 21)
29	22	64	96	24	2	2	(1, 1, 20)
30	23	46	48	24	2	2	(1, 22)
31	23	67	96	24	3	3	(1, 1, 21)
32	24	48	48	24	2	2	(1, 23)
33	24	70	96	24	3	3	(1, 1, 22)
34	25	73	96	24	3	3	(1, 1, 23)
35	26	74	96	24	4	3	(1, 1, 23, 1)

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

and

$$C = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

For simplicity, we say that A and C are given by $\tilde{B} = (00, 00, 00, 11)$ and $\tilde{C} = (1010, 1100, 1001, 0111)$, respectively.

In the following, the C matrices are obtained from the (12×12) Hadamard matrix

$$H_{12 \times 12} = \begin{bmatrix} + & + & + & + & + & + & + & + & + & + & + & + \\ + & - & - & + & - & - & - & + & + & + & - & + \\ + & - & + & - & - & - & + & + & + & - & + & - \\ + & + & - & - & - & + & + & + & - & + & - & - \\ + & - & - & - & + & + & + & - & + & - & - & + \\ + & - & - & + & + & + & - & + & - & - & + & - \\ + & - & + & + & + & - & + & - & - & + & - & - \\ + & + & + & + & - & + & - & - & + & - & - & - \\ + & + & + & - & + & - & - & + & - & - & - & + \\ + & + & - & + & - & - & + & - & - & + & + & + \\ + & - & + & - & - & + & - & - & - & + & + & + \\ + & + & - & - & + & - & - & - & + & + & + & - \end{bmatrix}$$

Here $+$ and $-$ represent $+1$ and -1 , respectively. Let $\tilde{H}_{12 \times 12}$ be the matrix obtained from $H_{12 \times 12}$ by changing all -1 's into 0 's, and $\tilde{H}_{11 \times 12}$ be the matrix obtained by deleting the first row of $\tilde{H}_{12 \times 12}$. It is known that $\tilde{H}_{11 \times 12}$ and $\tilde{H}_{k \times 12}$ are $OA(2)$, where $\tilde{H}_{k \times 12}$ is the matrix with any k rows ($k \leq 11$) of $\tilde{H}_{11 \times 12}$. In the following eight cases, we select C as the matrices $\tilde{H}_{k \times 12}$.

- $\underline{F} = F(0, \dots, 9, 1-2, 1-3, \dots, 1-9)$, $n = 9$, $N = 24$, $\nu = 18$, $\tilde{B} = (1, 0, 0, 0, 0, 0, 0, 0, 0)$, $C = \tilde{H}_{8 \times 12}$.
- $\underline{F} = F(0, \dots, 10, 1-2, \dots, 1-10)$, $n = 10$, $N = 24$, $\nu = 20$, $\tilde{B} = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$, and $C = \tilde{H}_{9 \times 12}$.
- $\underline{F} = F(0, 1, \dots, 11, 1-2, 1-3, \dots, 1-11)$, $n = 10$, $N = 24$, $\nu = 22$, $\tilde{B} = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$, and $C = \tilde{H}_{10 \times 12}$.
- $\underline{F} = F(0, 1, \dots, 11, 1-2, 1-3, \dots, 1-11, 2-3, \dots, 2-11)$, $n = 11$, $N = 48$, $\nu = 31$, $\tilde{B} = (10, 01, 00, 00, 00, 00, 00, 00, 00, 00)$, and $C = \tilde{H}_{9 \times 12}$.
- $\underline{F} = F(0, 1, \dots, 12, 1-2, 1-3, \dots, 1-12)$, $n = 12$, $N = 24$, $\nu = 24$, $\tilde{B} = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$, and $C = \tilde{H}_{11 \times 12}$.
- $\underline{F} = F(0, 1, \dots, 12, 1-2, 1-3, \dots, 1-12, 2-3, \dots, 2-12)$, $n = 12$, $N = 48$, $\nu = 34$, $\tilde{B} = (10, 01, 00, 00, 00, 00, 00, 00, 00, 00, 00)$, and $C = \tilde{H}_{10 \times 12}$.
- $\underline{F} = F(0, 1, \dots, 13, 1-2, 1-3, \dots, 1-13, 2-3, \dots, 2-13)$, $n = 13$, $N = 48$, $\nu = 37$, $\tilde{B} = (10, 01, 00, 00, 00, 00, 00, 00, 00, 00, 00, 00)$, and $C = \tilde{H}_{11 \times 12}$.
- $\underline{F} = F(0, 1, \dots, 14, 1-2, 1-3, \dots, 1-13, 2-3, \dots, 2-13)$, $n = 14$, $N = 48$, $\nu = 38$, $\tilde{B} = (10, 01, 00, \dots, 00, 11)$. Here C is derived from $\tilde{H}_{12 \times 12}$ by exchanging the positions of the first row and the last row of $\tilde{H}_{12 \times 12}$.

In the following, we use the (20×20) Hadamard matrix (see Equation on next page) to select the C matrices. Let $\tilde{H}_{20 \times 20}$ be the matrix obtained from $\tilde{H}_{19 \times 20}$ by changing -1 's into 0 's, and $\tilde{H}_{19 \times 20}$ be the matrix obtained by deleting the first row of $\tilde{H}_{20 \times 20}$. It is known that $\tilde{H}_{19 \times 20}$ and $\tilde{H}_{k \times 20}$ are $OA(2)$, where $\tilde{H}_{k \times 20}$ is the matrix with any k rows ($k \leq 19$) of $\tilde{H}_{19 \times 20}$. In the following cases, we select C as the matrices $\tilde{H}_{k \times 20}$.

30. $\underline{F} = F(0, 1, \dots, 23, 1\cdot2, 1\cdot3, \dots, 1\cdot23), n = 23, N = 48, \nu = 46, \tilde{B} = (1, 0, \dots, 0),$ and $C = \tilde{H}_{22 \times 24}.$
31. $\underline{F} = F(0, 1, \dots, 23, 1\cdot2, 1\cdot3, \dots, 1\cdot23, 2\cdot3, \dots, 2\cdot23), n = 23, N = 96, \nu = 67, \tilde{B} = (10, 01, \dots, 00),$ and $C = \tilde{H}_{21 \times 24}.$
32. $\underline{F} = F(0, 1, \dots, 24, 1\cdot2, 1\cdot3, \dots, 1\cdot24), n = 24, N = 48, \nu = 48, \tilde{B} = (1, 0, \dots, 0),$ and $C = \tilde{H}_{23 \times 24}.$
33. $\underline{F} = F(0, 1, \dots, 24, 1\cdot2, 1\cdot3, \dots, 1\cdot24, 2\cdot3, \dots, 2\cdot24), n = 24, N = 96, \nu = 70, \tilde{B} = (10, 01, \dots, 00),$ and $C = \tilde{H}_{22 \times 24}.$
34. $\underline{F} = F(0, 1, \dots, 25, 1\cdot2, 1\cdot3, \dots, 1\cdot25, 2\cdot3, \dots, 2\cdot25), n = 25, N = 96, \nu = 73, \tilde{B} = (10, 01, 00, \dots, 00),$ and $C = \tilde{H}_{23 \times 24}.$
35. $\underline{F} = F(0, 1, \dots, 26, 1\cdot2, 1\cdot3, \dots, 1\cdot25, 2\cdot3, \dots, 2\cdot25), n = 26, N = 96, \nu = 74, \tilde{B} = (10, 01, 00, \dots, 00, 11),$ and $C(24 \times 24)$ is obtained by exchanging the first row and the last row of $\tilde{H}_{24 \times 24}.$

Case III: Interactions Between and Within Subgroups

Consider the situations where the two-factor interactions occur either within subgroups or between subgroups. In Table 3, we list eight such examples. Here all the factors are divided into three or four subgroups. The two-factor interactions are between the first two subgroups and within the third subgroup. The fourth subgroup (if it exists) is not involved in the two-factor interactions. Since the designs in Table 3 consist of a single flat, i.e., $C = c = 0$, we give the vector of effects $\underline{F}$, parameters n, N, ν , and the corresponding $\tilde{B}$ matrix for each case.

1. $\underline{F} = F(0, 1, \dots, 6, 1\cdot3, 1\cdot4, 2\cdot3, 2\cdot4, 5\cdot6), n = 6, N = 16, \nu = 12, \tilde{B} = (1110, 1101).$
2. $\underline{F} = F(0, 1, \dots, 7, 1\cdot2, 1\cdot3, 1\cdot4, 5\cdot6, 5\cdot7, 6\cdot7), n = 7, N = 16, \nu = 14, \tilde{B} = (0111, 1100, 1001).$
3. $\underline{F} = F(0, 1, \dots, 7, 1\cdot3, 1\cdot4, 2\cdot3, 2\cdot4, 5\cdot6), n = 7, N = 16, \nu = 13, \tilde{B} = (0101, 0011, 1001).$
4. $\underline{F} = F(0, 1, \dots, 8, 1\cdot3, 1\cdot4, 2\cdot3, 2\cdot4, 5\cdot6), n = 8, N = 16, \nu = 14, \tilde{B} = (0101, 0011, 1010, 1110).$
5. $\underline{F} = F(0, 1, \dots, 8, 1\cdot4, 1\cdot5, 1\cdot6, 2\cdot4, 2\cdot5, 2\cdot6, 3\cdot4, 3\cdot5, 3\cdot6, 7\cdot8), n = 8, N = 32, \nu = 19, \tilde{B} = (11100, 01110, 11101).$
6. $\underline{F} = F(0, 1, \dots, 9, 1\cdot3, 1\cdot4, 2\cdot3, 2\cdot4, 5\cdot6), n = 9, N = 16, \nu = 15, \tilde{B} = (11100, 01110, 10001, 00111).$
7. $\underline{F} = F(0, 1, \dots, 9, 1\cdot4, 1\cdot5, 1\cdot6, 2\cdot4, 2\cdot5, 2\cdot6, 3\cdot4, 3\cdot5, 3\cdot6, 7\cdot8, 7\cdot9, 8\cdot9), n = 9, N = 32, \nu = 22, \tilde{B} = (11100, 00111, 11010, 10001).$

Table 3 Designs for Interactions Within and Between Subgroups

id	n	ν	N	f	p	p'	$(n_1, \dots, n_p)$
1	6	12	16	1	3	3	(2, 2, 2)
2	7	14	16	1	3	3	(1, 3, 3)
3	7	13	16	1	4	3	(2, 2, 2, 1)
4	8	14	16	1	4	3	(2, 2, 2, 2)
5	8	19	32	1	3	3	(3, 3, 2)
6	9	15	16	1	4	3	(2, 2, 2, 3)
7	9	22	32	1	3	3	(3, 3, 3)
8	10	16	16	1	4	3	(2, 2, 2, 4)

8. $\underline{F} = F(0, 1, \dots, 10, 1\cdot3, 1\cdot4, 2\cdot3, 2\cdot4, 5\cdot6), n = 10, N = 16, \nu = 16, \tilde{B} = (11000, 10100, 10010, 10001, 00111).$

ORTHOGONAL DESIGNS FOR 3^m FACTORIAL EXPERIMENTS

In the following examples, N is the total number of treatments in a design, m is the number of factors of interest, and ν is the number of parameters in F . We assume that m factors are divided into p subgroups, and each subgroup has m_k ($k = 1, \dots, p$) factors in it. The two-factor interactions are either within or between some of the subgroups. All of the following designs of parallel-flats type satisfy the conditions in Theorem 1.

Case 1. Consider a 3^4 factorial experiment. The factorial effects given in the vector

$$\begin{aligned} \underline{F} = & (\mu, F_1, F_2, F_3, F_4, F_1^2, F_2^2, F_3^2, F_4^2, F_1F_2, F_1F_3, F_2F_3, F_1F_2^2, F_1F_3^2, \\ & F_2F_3^2, F_1^2F_2, F_1^2F_3, F_2^2F_3, F_1^2F_2^2, F_1^2F_3^2, F_2^2F_3^2) \\ = & (0, 1, 2, 3, 4, 1^2, 2^2, 3^2, 4^2, 12, 13, 23, 12^2, 13^2, 23^2, 1^22, 1^23, \\ & 2^23, 1^22^2, 1^23^2, 2^23^2) \end{aligned}$$

have the property that the two-factor interactions occur within the subgroup $\{1, 2, 3\}$. Consider the design T given by the linear equation

$$A\underline{t} = t_1 + t_2 + t_3 + t_4 = C$$

over $\text{GF}(3)$, where $A = (1 \ 1 \ 1 \ 1)$, and $C = 0$. By Theorem 1, the design (see Equation below) is an orthogonal design for estimating $\underline{F}$. Here $m = 4, \nu = 21, p = 2, m_1 = 3, m_2 = 1, N = 27$.

In the following examples of orthogonal designs, we only give matrices A, C , and $\underline{F}$.

$$T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 & 0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 \\ 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 2 & 1 & 2 & 1 & 0 & 1 & 0 & 2 & 2 & 1 & 0 & 1 & 0 & 2 & 0 & 2 & 1 & 1 & 0 & 2 & 0 & 2 & 1 & 2 & 1 & 0 \end{pmatrix}$$

Case 2. Consider a 3^5 factorial experiment. The factorial effects given in the vector

$$\underline{F}' = (0, 1, \dots, 5, 1^2, \dots, 5^2, 12, 13, 23, 12^2, 13^2, 23^2, 1^2 2, 1^2 3, 2^2 3, 1^2 2^2, 1^2 3^2, 2^2 3^2)$$

are such that the two-factor interactions occur within the subgroup $\{1, 2, 3\}$. The design T given by $A_t = C$, where

$$A = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix}, \quad C = \underline{0}$$

is an orthogonal design for estimating F . Here $m = 5$, $\nu = 23$, $p = 2$, $m_1 = 3$, $m_2 = 2$, $N = 27$.

Case 3. Consider a 3^6 factorial experiment. The factorial effects given in the vector

$$\underline{F}' = (0, 1, \dots, 6, 1^2, \dots, 6^2, 12, 13, 23, 12^2, 13^2, 23^2, 1^2 2, 1^2 3, 2^2 3, 1^2 2^2, 1^2 3^2, 2^2 3^2)$$

are such that the two-factor interactions occur within the subgroup $\{1, 2, 3\}$. The orthogonal design T given by $A_t = \underline{0}$, where

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}, \quad C = \underline{0}$$

is an orthogonal design for estimating F . Here $m = 6$, $\nu = 25$, $p = 2$, $m_1 = 3$, $m_2 = 3$, $N = 27$.

In the following examples, the two-factor interactions occur between subgroups.

Case 4. Consider a 3^4 factorial experiment. The factorial effects given in the vector

$$\underline{F}' = (0, 1, 2, 3, 4, 1^2, 2^2, 3^2, 4^2, 14, 24, 34, 14^2, 24^2, 34^2, 1^2 4, 2^2 4, 3^2 4, 1^2 4^2, 2^2 4^2, 3^2 4^2)$$

are such that the two-factor interactions are between the subgroups $\{1, 2, 3\}$ and $\{4\}$. By Theorem 1, the design T given by the solution to the linear equation $A_t = t_1 + t_2 + t_3 + t_4 = C$, where $A = (1 \ 1 \ 1 \ 1)$ and $C = \underline{0}$, is an orthogonal design for estimating $\underline{F}$. Here $m = 4$, $\nu = 21$, $p = 2$, $m_1 = 3$, $m_2 = 1$, $N = 27$.

Case 5. Consider a 3^5 factorial experiment. The factorial effects given in the vector

$$\underline{F}' = (0, 1, \dots, 5, 1^2, \dots, 5^2, 15, 25, 35, 45, 15^2, 25^2, 35^2, 45^2, 1^2 5, 2^2 5, 3^2 5, 4^2 5, 1^2 5^2, 2^2 5^2, 3^2 5^2, 4^2 5^2)$$

are such that the two-factor interactions are between the subgroups $\{1, 2, 3, 4\}$ and $\{5\}$. The design T given by $A_t = C$ is an orthogonal design for estimating $\underline{F}$, where

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \end{pmatrix}, \quad C = \underline{0}$$

Here $m = 5$, $\nu = 27$, $p = 2$, $m_1 = 4$, $m_2 = 1$, $N = 27$. The number of the parameters in $\underline{F}$ is equal to the number of treatments in T . The orthogonal design T is a saturated orthogonal design.

Case 6. Consider a 3^6 factorial experiment. The factorial effects given in the vector

$$\underline{F}' = (0, 1, \dots, 6, 1^2, \dots, 6^2, 14, 24, 34, 14^2, 24^2, 34^2, 1^2 4, 2^2 4, 3^2 4, 1^2 4^2, 2^2 4^2, 3^2 4^2)$$

are such that the two-factor interactions are between the subgroups $\{1, 2, 3\}$ and $\{4\}$. The subgroup $\{5, 6\}$ is not involved in the two-factor interactions. By Theorem 1, the design T given by $A_t = C$ is an orthogonal design for estimating $\underline{F}$, where

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}, \quad C = \underline{0}$$

Here $m = 6$, $\nu = 25$, $p = 3$, $m_1 = 3$, $m_2 = 1$, $m_3 = 2$, $N = 27$.

Case 7. Consider a 3^7 factorial experiment. The factorial effects given in the vector

$$\underline{F}' = (0, 1, \dots, 7, 1^2, \dots, 7^2, 14, 24, 34, 14^2, 24^2, 34^2, 1^2 4, 2^2 4, 3^2 4, 1^2 4^2, 2^2 4^2, 3^2 4^2)$$

are such that the two-factor interactions are between the subgroups $\{1, 2, 3\}$ and $\{4\}$. The subgroup $\{5, 6, 7\}$ is not involved in the two-factor interactions. The design T given by $A_t = C$ is an orthogonal design for estimating $\underline{F}$, where

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}, \quad C = \underline{0}$$

Here $m = 7$, $\nu = 25$, $p = 3$, $m_1 = 3$, $m_2 = 1$, $m_3 = 3$, $N = 27$.

Bose and Bush^[14] gave an orthogonal array $H(7 \times 18)$ of strength 2 as follows:

$$H = \begin{pmatrix} 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 1 & 2 & 0 & 1 & 2 & 1 & 2 & 0 & 2 & 0 & 1 & 1 & 2 & 0 & 2 & 0 & 1 \\ 0 & 1 & 2 & 1 & 2 & 0 & 0 & 1 & 2 & 2 & 0 & 1 & 2 & 0 & 1 & 1 & 2 & 0 \\ 0 & 1 & 2 & 2 & 0 & 1 & 2 & 0 & 1 & 0 & 1 & 2 & 1 & 2 & 0 & 1 & 2 & 0 \\ 0 & 1 & 2 & 1 & 2 & 0 & 2 & 0 & 1 & 1 & 2 & 0 & 0 & 1 & 2 & 2 & 0 & 1 \\ 0 & 1 & 2 & 2 & 0 & 1 & 1 & 2 & 0 & 1 & 2 & 0 & 2 & 0 & 1 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 1 & 2 \end{pmatrix} \quad (7)$$

Notice that any submatrix $\hat{H}(2 \times 18)$ of H is OA(2). Therefore a linear combination of any two rows of H over GF(3) is an OA(1). Now we use the matrix H to construct orthogonal designs.

Case 8. Consider a 3^4 factorial experiment. The factorial effects given in the vector

$$\underline{F}' = (0, 1, 2, 3, 4, 1^2, 2^2, 3^2, 4^2, 12, 13, 23, 12^2, 13^2, 23^2, 1^2 2, 1^2 3, 2^2 3, 1^2 2^2, 1^2 3^2, 2^2 3^2)$$

are such that the two-factor interactions occur within the subgroup $\{1, 2, 3\}$. An orthogonal design for estimating $\underline{F}$ is given by $A_t = C_{3 \times 18}$, where

$$A = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

and $C_{3 \times 18}$ consists of the first three rows of H in Eq. 7. Here $m = 4$, $\nu = 21$, $p = 2$, $m_1 = 3$, $m_2 = 1$, $N = 3 \times 18 = 54$.

Case 9. Consider a 3^5 factorial experiment. The factorial effects given in the vector

$$\underline{F}' = (0, 1, \dots, 5, 1^2, \dots, 5^2, 15, \dots, 45, 15^2, \dots, 45^2, 1^2 5, \dots, 4^2 5, 1^2 5^2, \dots, 4^2 5^2)$$

are such that the two-factor interactions occur between the subgroups $\{1, 2, 3, 4\}$ and $\{5\}$. An orthogonal design for estimating $\underline{F}$ is given by $A_t = C_{4 \times 18}$, where

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

and $C_{4 \times 18}$ consists of the first four rows of H in Eq. 7. Here $m = 5$, $\nu = 27$, $p = 2$, $m_1 = 4$, $m_2 = 1$, $N = 3 \times 18 = 54$.

Case 10. Consider a 3^6 factorial experiment. The factorial effects given in the vector

$$\underline{F}' = (0, 1, \dots, 6, 1^2, \dots, 6^2, 16, \dots, 56, 16^2, \dots, 56^2, 1^2 6, \dots, 5^2 6, 1^2 6^2, \dots, 5^2 6^2)$$

are such that the two-factor interactions occur between the subgroups $\{1, 2, 3, 4, 5\}$ and $\{6\}$. An orthogonal design for estimating $\underline{F}$ is given by $A_t = C_{5 \times 18}$, where

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

and $C_{5 \times 18}$ consists of the first five rows of H in Eq. 7. Here $m = 6$, $\nu = 33$, $p = 2$, $m_1 = 5$, $m_2 = 1$, $N = 3 \times 18 = 54$.

Case 11. Consider a 3^7 factorial experiment. The factorial effects given in the vector

$$\underline{F}' = (0, 1, \dots, 7, 1^2, \dots, 7^2, 17, \dots, 67, 17^2, \dots, 67^2, 1^2 7, \dots, 6^2 7, 1^2 7^2, \dots, 6^2 7^2)$$

have the property that the two-factor interactions occur between the subgroups $\{1, \dots, 6\}$ and $\{7\}$. An orthogonal design for estimating $\underline{F}$ is given by $A_t = C_{6 \times 18}$, where

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

and $C_{6 \times 18}$ consists of the first six rows of H in Eq. 7. Here $m = 7$, $\nu = 39$, $p = 2$, $m_1 = 6$, $m_2 = 1$, $N = 3 \times 18 = 54$.

Case 12. Consider a 3^8 factorial experiment. The factorial effects given in the vector

$$\underline{F}' = (0, 1, \dots, 8, 1^2, \dots, 8^2, 18, \dots, 78, 18^2, \dots, 78^2, 1^2 8, \dots, 7^2 8, 1^2 8^2, \dots, 7^2 8^2)$$

have the property that the two-factor interactions occur between the subgroups $\{1, \dots, 7\}$ and $\{8\}$. An orthogonal design for estimating $\underline{F}$ is given by $A_t = C_{7 \times 18}$, where

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

and $C_{7 \times 18} = H$ given by Eq. 7. Here $m = 8$, $\nu = 45$, $p = 2$, $m_1 = 7$, $m_2 = 1$, $N = 3 \times 18 = 54$.

ORTHOGONAL DESIGNS FOR MIXED $2^n \times 3^m$ FACTORIAL EXPERIMENTS

The mixed $2^n \times 3^m$ factorial experiments are very common in practice. It is useful to be able to generate an orthogonal design to estimate any specified set of mixed factorial effects. Consider the parallel-flats type of designs for the mixed factorial experiment. Let T_{li} be the set of solutions to the linear system

$$A_1 t_1 = c_{1i}, \quad (i = 1, 2, \dots, f) \quad (8)$$

over GF(2). Let T_{2i} be the set of solutions to the linear system

$$A_2 t_2 = c_{2i}, \quad (i = 1, 2, \dots, f) \quad (9)$$

over GF(3). The PFD design T for $2^n \times 3^m$ with $N = f \times 2^{n-r_1} \times 3^{m-r_2}$ treatments is defined as follows:

$$T = T_{11} \otimes_f T_{21} + \dots + T_{1f} \otimes_f T_{2f} \quad (10)$$

where $\otimes_f$ is the product defined in Ref. [7], and shown as in Cases 1–4 of this section.

The extension of Theorem 1 to the mixed case was given in Ref. [13]. Now we review the results. First consider the following definition.

Definition 1. Let Q be a $2 \times f$ array whose first row has the elements from the set $\{0, 1\}$, and whose second row has the elements from the set $\{0, 1, 2\}$. Let k_{ij} be the number of column vectors $(i, j)'$ of Q ($i = 0, 1; j = 0, 1, 2$).

1. Q is said to have the property O_1 if $k_{00} - k_{10} = k_{01} - k_{11} = k_{02} - k_{12}$.
2. Q is said to have property Q_2 if $k_{00} = k_{10}, k_{01} = k_{11}$ and $k_{02} = k_{12}$.
3. Q is said to have property O_3 if $k_{00} = k_{01} = k_{02}$ and $k_{10} = k_{11} = k_{12}$.
4. Q is said to have property O_4 if $k_{00} = k_{01} = k_{02} = k_{10} = k_{11} = k_{12}$, i.e., Q is an $OA(2)$.

Notice that the model of Eq. 1 and the defining vectors can easily be extended to the mixed $2^n \times 3^m$ factorial experiment. We use the same notation as in “Orthogonal Designs for s^m Factorial Experiments” for the mixed experiments.

Let $\underline{F}$ be a set of factorial effects in a $2^n \times 3^m$ experiment. Suppose that e_1 and e_2 are any effects in $\underline{F}$, and $\underline{e}_1$ and $\underline{e}_2$ are the corresponding defining vectors. Let $\underline{e}'_1 = (\underline{e}'_{11}, \underline{e}'_{21})$ and $\underline{e}'_2 = (\underline{e}'_{12}, \underline{e}'_{22})$, where $\underline{e}'_{11}(1 \times n)$ and $\underline{e}_{12}(1 \times n)$ are over $GF(2)$, $\underline{e}'_{21}(1 \times m)$ and $\underline{e}_{22}(1 \times m)$ are over $GF(3)$. Then we define the operation $\oplus$ as

$$\underline{e} \oplus \gamma \underline{e}_2 = (\underline{e}_{11} + \gamma \underline{e}_{12}, \underline{e}_{21} + \gamma \underline{e}_{22})$$

Define

$$\underline{F}^2 = \{\underline{e}_1 \oplus \gamma \underline{e}_2 \mid \text{for } e_1, e_2 \in \underline{F}, \gamma = 1, 2\}$$

Theorem 2. In a $2^n \times 3^m$ experiment, let design T be described as in Eq. 10, and let M be the information matrix for estimating $\underline{F}$ by using $T^{[13]}$. Then T is orthogonal if and only if the following conditions hold.

1. For $(\underline{e}_1, \underline{0}) \in \underline{F}^2$, if there exists a nonzero vector $\underline{u}'_1(1 \times r_1)$ over $GF(2)$ such that $\underline{e}_1 = \underline{u}'_1 A_1$, then $\underline{u}'_1 C_1$ must be an $OA(1)$, where $C_1 = (\underline{c}_{11}, \underline{c}_{12}, \dots, \underline{c}_{1f})$.
2. For $(\underline{0}, \underline{e}_2) \in \underline{F}^2$, if there exists a nonzero vector $\underline{u}'_2(1 \times r_2)$ over $GF(3)$ such that $\underline{e}_2 = \underline{u}'_2 A_2$, then $\underline{u}'_2 C_2$ must be an $OA(1)$, where $C_2 = (\underline{c}_{21}, \underline{c}_{22}, \dots, \underline{c}_{2f})$.
3. Suppose $(\underline{e}_1, \underline{e}_2) \in \underline{F}^2$, but $(\underline{e}_1, \underline{0}), (\underline{0}, \underline{e}_2) \notin \underline{F}^2$. If there exist a nonzero vector $\underline{u}'_1(1 \times r_1)$ over $GF(2)$ and a nonzero vector $\underline{u}'_2(1 \times r_2)$ over $GF(3)$ such that $\underline{e}_1 = \underline{u}'_1 A_1$ and $\underline{e}_2 = \underline{u}'_2 A_2$, then the matrix

$$\begin{bmatrix} \underline{u}'_1 & C_1 \\ \underline{u}'_2 & C_2 \end{bmatrix}$$

must have property O_1 .

4. Suppose $(\underline{e}'_1, \underline{e}_2), (\underline{e}_1, \underline{0}) \in \underline{F}^2$, but $(\underline{0}, \underline{e}_2) \notin \underline{F}^2$. If there exist a nonzero vector $\underline{u}'_1(1 \times r_1)$ over $GF(2)$ and a nonzero vector $\underline{u}'_2(1 \times r_2)$ over $GF(3)$ such that $\underline{e}_1 = \underline{u}'_1 A_1$ and $\underline{e}_2 = \underline{u}'_2 A_2$, then

$$\begin{bmatrix} \underline{u}'_1 & C_1 \\ \underline{u}'_2 & C_2 \end{bmatrix}$$

must have property O_2 .

5. Suppose $(\underline{e}_1, \underline{e}_2), (\underline{0}, \underline{e}_2) \in \underline{F}^2$, but $(\underline{e}_1, \underline{0}) \notin \underline{F}^2$. If there exist a nonzero vector $\underline{u}'_1(1 \times r_1)$ over $GF(2)$ and a nonzero vector $\underline{u}'_2(1 \times r_2)$ over $GF(3)$ such that $\underline{e}_1 = \underline{u}'_1 A_1$ and $\underline{e}_2 = \underline{u}'_2 A_2$, then

$$\begin{bmatrix} \underline{u}'_1 & C_1 \\ \underline{u}'_2 & C_2 \end{bmatrix}$$

must have property O_3 .

6. Suppose $(\underline{e}_1, \underline{e}_2), (\underline{e}_1, \underline{0}), (\underline{0}, \underline{e}_2) \in \underline{F}^2$. If there exist a nonzero vector $\underline{u}'_1(1 \times r_1)$ over $GF(2)$ and a nonzero vector $\underline{u}'_2(1 \times r_2)$ over $GF(3)$ such that $\underline{e}_1 = \underline{u}'_1 A_1$ and $\underline{e}_2 = \underline{u}'_2 A_2$, then

$$\begin{bmatrix} \underline{u}'_1 & C_1 \\ \underline{u}'_2 & C_2 \end{bmatrix}$$

must have property O_4 .

Now we apply Theorem 2 to generate the mixed factorial experiments.

Case 1. Consider a $2^3 \times 3$ factorial experiment. There are four factors F_1, F_2, F_3 , and F_4 , where F_1, F_2 , and F_3 take two levels $\{0, 1\}$ and F_4 takes three levels $\{0, 1, 2\}$. The factorial effects of interest are given in the vector

$$\begin{aligned} \underline{F}' &= (\mu, F_1, F_2, F_3, F_4, F_1 F_2, F_1 F_3, F_1 F_4, F_2 F_3, F_2 F_4, F_3 F_4, F_1 F_2 F_3, F_1 F_2 F_4, F_1 F_3 F_4, F_2 F_3 F_4) \\ &= (0, 1, 2, 3, 4, 4^2, 14, 24, 34, 14^2, 24^2, 34^2) \end{aligned}$$

The two-factor interactions in $\underline{F}$ are between the subgroups $\{1, 2, 3\}$ and $\{4\}$. Let $T_1 = [T_{11}, T_{12}, T_{13}]$, and T_{1i} ($i = 1, 2, 3$) with size (3×4) be the solutions to the linear $A_1 t_1 = \underline{c}_{1i}$, where $t_1' = (t_1, t_2, t_3)$, $A_1 = (111)$, and $\underline{c}_{1i} = 0$. Let $T_2 = [T_{21}, T_{22}, T_{23}]$ and let T_{2i} ($i = 1, 2, 3$) with size (1×3) be the solutions to the linear $A_2 t_2 = \underline{c}_{2i}$, where $t_2' = (t_4)$, $A_2 = (1)$, and $\underline{c}_{2i} = 0$. Here $n = 3, m = 1, f = 3$. Clearly T_1 and T_2 are orthogonal designs for estimating the main effects of a 2^3 and a 3^1 factorial, respectively. The linear system

$$\begin{aligned} A_1 t_1 &= \underline{c}_{1i} \\ A_2 t_2 &= \underline{c}_{2i}, \quad \text{for } (i = 1, 2, 3) \end{aligned}$$

satisfies the conditions in Theorem 2. Therefore the design

$$T = T_{11} \otimes_f T_{21} + T_{12} \otimes_f T_{22} + T_{13} \otimes_f T_{23}$$

$$= \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \end{pmatrix}$$

$$T = T_{11} \otimes_f T_{21} + T_{12} \otimes_f T_{22} + T_{13} \otimes_f T_{23}$$

$$= \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 \end{pmatrix}$$

obtained as in Eq. 10 is an orthogonal design for estimating $\underline{F}$. Here number of treatments is $N = 3 \times 4 = 12$; the number of parameters in $\underline{F}$ is $\nu = 12 = N$. The design T is a saturated orthogonal design.

Case 2. Consider a $2^4 \times 3$ factorial experiment. The factorial effects of interest given in the vector $\underline{F}$ as

$$\underline{F}' = (0, 1, 2, 3, 4, 5, 5^2, 12, 13, 23, 15, 25, 35, 45, 15^2, 25^2, 35^2, 45^2)$$

are such that the two-factor interactions are either within the subgroup $\{1, 2, 3\}$ or are between the subgroups $\{1, 2, 3, 4\}$ and $\{5\}$. Let $T_1 = [T_{11}:T_{12}:T_{13}]$ and let T_{1i} ($i = 1, 2, 3$) with size (4×8) be the solutions to the linear $A_1 t_{1i} = \underline{c}_{1i}$, where $A_1 = (1111)$ and $\underline{c}_{1i} = 0$. Let $T_2 = [T_{21}:T_{22}:T_{23}]$, where T_{2i} ($i = 1, 2, 3$) with size (1×3) are the solutions to the linear $A_2 t_{2i} = \underline{c}_{2i}$, where $\underline{t}_2' = (t_4)$, $A_2 = (1)$, and $\underline{c}_{2i} = 0$. Here $n = 4$, $m = 1$, $f = 3$. The linear system

$$A_1 t_{1i} = \underline{c}_{1i}$$

$$A_2 t_{2i} = \underline{c}_{2i}, \quad \text{for } (i = 1, 2, 3)$$

satisfies the conditions in Theorem 2. Therefore the design (see Equation below) obtained as in Eq. 10 is an orthogonal design for estimating $\underline{F}$. Here number of

treatments is $N = 3 \times 8 = 24$, and the number of parameters in $\underline{F}$ is $\nu = 18$.

Now consider orthogonal arrays for mixed designs. An $(n + m) \times N$ matrix T is said to be an $OA[(n + m), N, 2^n \times 3^m, d]$, if:

1. Its first n rows have elements from the set $\{0, 1\}$ and the last m rows have elements from the set $\{0, 1, 2\}$.
2. Any $d \times N$ submatrix T_0 of T has the property that all possible columns vectors of T_0 occur in T_0 with the same frequency.

It is easy to see that the following matrix H is an $OA(5, 12, 2^4 \times 3, 2)$:

$$H = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 \end{pmatrix}$$

Now we use the matrix H to generate the following orthogonal designs.

Case 3. Consider a $2^5 \times 3$ factorial experiment. The factorial effects given in the vector

$$\underline{F}' = (0, 1, \dots, 5, 6, 6^2, 16, 26, 36, 46, 56, 16^2, 26^2, 36^2, 46^2, 56^2)$$

have the property that the two-factor interactions are between the subgroups $\{1, 2, 3, 4, 5\}$ and $\{6\}$. Let C_1 (4×12) be the first four rows of H , and let $T_1 = [T_{1,1}:\dots:T_{1,12}]$, $T_{1,i}$ ($i = 1, \dots, 12$) with size (5×2) be the solutions to the linear $A_1 t_{1i} = \underline{c}_{1,i}$, where $\underline{c}_{1,i}$ is the i th column of C_1 , and

$$T = T_{1,1} \otimes_f T_{2,1} + \dots + T_{1,12} \otimes_f T_{2,12}$$

$$= \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \end{pmatrix}$$

$$A_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Let $C_2(1 \times 12)$ be the last row of H . Let $T_2 = [T_{2,1}, \dots, T_{2,12}]$, and let $T_{1,i} (i = 1, \dots, 12)$ with size (1×1) be the solutions to the linear $A_2 t_2 = c_{2,i}$ over $GF(3)$, where $A_2 = (1)$, $t_2 = t_6$, $c_{2,i}$ is the i th column of C_2 . The linear system

$$\begin{aligned} A_1 t_1 &= c_{1i} \\ A_2 t_2 &= c_{2i}, \quad \text{for } (i = 1, 2, 3) \end{aligned}$$

satisfies the conditions in Theorem 2. Therefore the design (see Equation above) obtained as in Eq. 10 is an orthogonal design for estimating $\underline{F}$. Here the number of treatments is $N = 12 \times 2^{5-4} \times 3^{1-1} = 24$, and the number of parameters in $\underline{F}$ is $\nu = 18$.

Case 4. Consider a $2^4 \times 3^2$ factorial experiment. The factorial effects of interest given in the vector

$$\underline{F}' = (0, 1, \dots, 5, 6, 5^2, 6^2, 16, 26, 36, 46, 56, 16^2, 26^2, 36^2, 46^2, 56^2, 5^2 6, 5^2 6^2)$$

are such that the two-factor interactions are between the subgroups $\{1, 2, 3, 4, 5\}$ and $\{6\}$. Take $T_1 = H$. Therefore the design (see Equation below) is an orthogonal design for estimating $\underline{F}$. Here the number of treatments is $N = 12 \times 3 = 36$, and the number of parameters in $\underline{F}$ is $\nu = 21$.

APPLICATION OF MIXED FACTORIAL DESIGN IN BIOMEDICAL RESEARCH

In this section, we consider two examples in the biomedical research area, in which factorial experiments are widely used for screening the important factors or studying the higher-factor interactions.

Example 1: Allele-Specific Polymerase Chain Reaction for Screening Point Mutation

Polymerase chain reaction is a powerful biological technique to facilitate molecular analysis, including the DNA

diagnosis and identification of a human being's genetic disease. Many components affect a PCR process. In addition, there may be complex interactions among the components of PCR. Factorial designs are considered as cost-effective techniques to achieve optimization of PCR.

In an experiment we consider, the PCR yield is determined by two procedures, DNA extraction and PCR conditions. The five most important factors (or components) are identified in the PCR array as follows:

1. F_1 = autoclave, the degree of autoclave blood spots prior to DNA extraction.
2. F_2 = DNA extraction, the methods of extracting human genetic DNA from dried blood spots.
3. F_3 = number of blood spots in each DNA extraction tube.
4. F_4 = template DNA, the volume of template DNA used in the PCR.
5. F_5 = magnesium concentration.

The first four factors, F_1 to F_4 , take on two levels, coded as $\{0, 1\}$. The last factor, F_5 (magnesium), takes on three levels, coded as $\{0, 1, 2\}$. Prior experience indicates that there may be interactions among autoclave, DNA extraction, and number of blood spots in each of the DNA levels. Also, magnesium is very active, and it may interact with other factors. Therefore we need to estimate the two-factor interactions within the subgroup $\{F_1, F_2, F_3\}$ and the two-factor interactions between the subgroups $\{F_1, F_2, F_3, F_4\}$ and $\{F_5\}$, besides the general mean and all the main effects. The parameter vector of interest can be written as

$$\underline{F}' = (\mu, F_1, F_2, F_3, F_4, F_5, F_5^2, F_1 F_2, F_1 F_3, F_2 F_3, F_1 F_5, F_2 F_5, F_3 F_5, F_4 F_5, F_1 F_5^2, F_2 F_5^2, 2 F_3 F_5^2, F_4 F_5^2)$$

By Case 2 in "Orthogonal Designs for Mixed $2^n \times 3^m$ Factorial Experiments," there exists an orthogonal design with 24 treatment combinations to estimate the parameter vector $\underline{F}$. The orthogonal design is presented above. Here the 24 columns correspond to the 24 treatments of different factorial combinations, which can give optimal estimates of 18 parameters in $\underline{F}$.

$$\begin{aligned} T &= T_1 \otimes_f (0, 1, 2)' \\ &= H \otimes_f (9, 1, 2)' \end{aligned}$$

$$= \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \end{pmatrix}$$

Treatment index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Autoclave	0	1	1	1	0	0	0	1	0	1	1	1	0	0	0	1	0	1	1	1	0	0	0	1
DNA extraction	0	1	0	0	1	1	0	1	0	1	0	0	1	1	0	1	0	1	0	0	1	1	0	1
No. blood spots	0	0	1	0	1	0	1	0	1	1	0	0	1	0	1	1	0	0	1	0	1	0	1	1
Template DNA	0	0	0	1	0	1	1	1	0	0	0	1	0	1	1	1	0	0	0	1	0	1	1	1
Magnesium	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
Gender	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
Drug A	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
Drug B	0	0	1	1	1	0	0	1	0	0	1	0	0	1	1	1	0	0	1	0	0	1	0	0	1	1	1	0	0	1	0	0	1	0	0	1
Drug C	0	1	1	0	1	0	1	0	0	1	0	1	0	1	1	0	1	0	0	1	0	1	0	1	1	0	1	0	1	0	0	1	0	1	0	1
Severity	0	0	0	0	1	1	1	1	2	2	2	0	0	0	0	1	1	1	1	2	2	2	0	0	0	0	1	1	1	1	2	2	2	2	2	2
Drug D	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2

Example 2: Drug Interaction in a Clinical Study

It is very common for patients with life-threatening diseases, such as AIDS and cancer, to undergo multiple therapies. Suppose, in a clinical study, we want to study the drug interactions among patients in different demographic groups. The following factors affect the outcome of therapy:

1. F_1 = gender of patients, with two levels (male = 0, female = 1).
2. F_2 = drug A, with two doses, denoted as 0 and 1.
3. F_3 = drug B, with two doses, denoted as 0 and 1.
4. F_4 = drug C, with two doses, denoted as 0 and 1.
5. F_5 = severity of the disease, with three levels (mild = 0, moderate = 1, severe = 2) (see 1st equation above).
6. F_6 = newly approved drug D, with three doses, denoted as (0, 1, 2).

The objective is to study the synergistic or antagonistic effects between newly approved treatment D and existing treatments A, B, and C in male and female patients with different levels of severity of the disease. Based on previous experiments, it is believed that interactions occur between D and A, B, C. The three-factor interactions can be ignored. We are interested in estimating all the main effects for each factor and two-factor interactions between drug D and the rest of the factors. Therefore the 21 factor effects of interest are given in the vector

$$\underline{F}' = (\mu, F_1, \dots, F_6, F_1^2, F_2^2, F_3^2, F_4^2, F_5^2, F_6^2, F_1F_2, F_1F_3, F_1F_4, F_1F_5, F_1F_6, F_2F_3, F_2F_4, F_2F_5, F_2F_6, F_3F_4, F_3F_5, F_3F_6, F_4F_5, F_4F_6, F_5F_6, F_1^2F_2^2, F_1^2F_3^2, F_1^2F_4^2, F_1^2F_5^2, F_1^2F_6^2, F_2^2F_3^2, F_2^2F_4^2, F_2^2F_5^2, F_2^2F_6^2, F_3^2F_4^2, F_3^2F_5^2, F_3^2F_6^2, F_4^2F_5^2, F_4^2F_6^2, F_5^2F_6^2)$$

By Case 4 in "Orthogonal Designs for Mixed $2^n \times 3^m$ Factorial Experiments," the following design is an orthogonal design for estimating $\underline{F}$: (See Equation 2nd equation above) where the 36 columns of factor level combinations are the treatments in the orthogonal design.

CONCLUSION

Notice that the numbers of treatment combinations required by a full or fracture factorial design increases geometrically as the numbers of factors increases. This leads to the major interest of factorial design to be efficient. One may generate design structure through assistance of a modern computer. However, without knowledge of advanced mathematics in finite geometries, number theories, group theories etc., it may become an exercise of trial and error to obtain an experimental design. In practice, the higher factors and high factor interactions do exist even though we wishfully think they shall be negligible. In order to give the estimates of higher factor effects which are uncorrelated with each other, one can use orthogonal arrays with higher strength. From the above discussions and examples, factorial design is an effective approach for quantitatively studying cause and effect relationship.

REFERENCES

1. Kesler, D.A.; Feiden, K.L. Faster evaluation of vital drugs. *Sci. Am.* March **1995**, 48–54.
2. Fisher, R.A. *The Design of Experiments*, 9th Ed.; MacMillan Pub. Co.: June, 1971.
3. Bose, R.C. Mathematical theory of the symmetrical factorial design. *Sankhya* **1947**, 8, 107–166.
4. Rao, C.R. The theory of fractional replications in factorial experiments. *Sankhya* **1950**, 10, 81–86.
5. Kiefer, J. Optimal Design in regression problems, II. *Ann. Math. Stat.* **1961**, 32, 298–325.
6. Cheng, C.S. Optimality of some weighing and 2^n fractional factorial designs. *Ann. Stat.* **1980**, 8, 437–446.
7. Connor, W.S.; Young, S. Fractional Factorial Designs for Experiments with Factors at Two and Three Levels. In *Applied Mathematics Series-National Bureau of Standards*; U.S. Government Printing Office: Washington, DC, 1961.

8. Srivastava, J.N.; Throop, D. Orthogonal Arrays Obtainable as Solutions to Linear Equations over Finite Fields. In *Linear Algebra and Its Applications*; 1990, Vol. 127.
9. Srivastava, J.N.; Anderson, D.A.; Mardekian, J. Theory of factorial designs of parallel flat type. I. The coefficient matrix. *J. Stat. Plan. Inference*. **1984**, 9, 22–252.
10. Srivastava, J.N.; Advances in the general theory of factorial designs on parallel pencils in Euclidean n -space. *Util. Math.* **1987**, 32, 75–94.
11. Li, J. Sequential and Optimal Single Stage Factorial Designs, with Industrial Applications. Ph.D. Thesis. Department of Statistics, Colorado State University: Fort Collins, CO, 1991.
12. Srivastava, J.N.; Li, J. Orthogonal designs of parallel flats type. *J. Stat. Plan. Inference* **1996**, 53, 261–283.
13. Liao, C.T. Fractional Factorial Designs for Estimating Location Effects and Screening Dispersion Effects; Ph.D. Dissertation. Colorado State University: Fort Collins, CO, 1994.
14. Bose, R.C.; Bush, K.A. Orthogonal arrays of strength two and three. *Ann. Math. Stat.* **1952**, 23, 502–524.

False Discovery Rate (FDR)

Avner Bar-Hen

INA-PG/INRA Biometrie, Paris, France

Philippe Broet and Cyril Dalmasso

INSERM U472, Villejuif Cedex, France

Jean-Jacques Daudin and Stephane Robin

Applied Mathematics and Computer Sciences, INA-PG/INRA, Paris, France

Abstract

The false discovery rate is defined as the expected proportion of false positives among the total number of positives. This entry provided statistical models for determining the FDR.

INTRODUCTION

Multiple testing is a classical problem for many high-dimensional data sets. The breakthrough of technology for image analysis or genomics has given a new interest for this question. A classical application is microarrays. This technology is part of a new class of biotechnologies that allow the monitoring of the expression level of thousands of genes simultaneously. Among the applications of microarrays, the identification of differentially expressed genes, i.e., genes whose expression is associated with the status of patients (treatment/control, for example) is an important one. The biological question of identifying differentially expressed genes can be restated as a two-sample hypothesis testing procedure: Is the gene differentially expressed between the two situations? However, when thousands of genes in a microarray data set are evaluated simultaneously by fold changes and significance tests approach, multiple testing problems immediately arise and lead to many false positive genes. In this “one-by-one gene” strategy the probability of detecting false positives rises sharply.

In this article we focus on differentially expressed genes; but the proposed results are applicable for all kinds of null hypothesis. In the framework of test theory, most of the desirable properties are defined for a single test: type I or II error, sensitivity, power, specificity, etc. In the case of multiple testing, extensions are far from obvious.

A small example illustrates the problem of multiple testing. Let us consider that m_0 genes among m are not differentially expressed, and that all the tests are performed with level α . Let us assume independent tests. Then the number of false positives (V) is a random variable, with binomial probability distribution

$$V \sim B(m_0, \alpha)$$

This simple modeling leads to the conclusion that the expected number of false positives when m_0 true null hypothesis are tested simultaneously is $E(V) = m_0 \alpha$. Regarding the high number of tests performed in microarray experiments (for instance $m = 10,000$), even if no gene is differentially expressed (in reality), a level $\alpha = 0.05$ leads to declare 500 genes as differentially expressed. The purpose of multiple testing is then to control the global risk of the procedure.

ERROR CRITERIA FOR MULTIPLE COMPARISONS: FALSE DISCOVERY RATE AND RELATED QUANTITIES

Definition

Consider a multiple testing situation in which m tests are being performed. The m genes can be classified according to the following 2×2 table:

	H_0 not rejected	H_0 rejected	Total
H_0 true	U	V	m_0
H_0 false	T	S	m_1
Total	W	R	m

To date, the aim of the main approaches taking into account the multiple testing problem has been to control the familywise error rate (FWER) that is defined as the probability of falsely reject at least one null hypothesis

$$\text{FWER} = \Pr[V \geq 1]$$

As an example, the Bonferroni procedure, which is the simplest and the most known procedure controlling the

FWER, consists in performing each test at level $\alpha^* = \alpha/m$. Then

$$\begin{aligned} \text{FWER} &= \Pr\{V > 0\} = \Pr\left(\min_{i \in I} (P_i) \leq \alpha^*\right) \\ &= \Pr\left(\bigcup_{i \in I} \{P_i \leq \alpha^*\}\right) \leq \sum_{i \in I} \Pr(P_i \leq \alpha^*) \\ &= m_0 \alpha^* \leq m \alpha^* = \alpha \end{aligned}$$

where $I = \{i \text{ such that } H_{0i} \text{ true}\}$.

However, as argued by Benjamini and Hochberg,^[1] controlling the FWER in multiple testing settings may not always be appropriate. As an alternative and less stringent concept of error control they introduced the false discovery rate (FDR). The main interest of controlling the FDR is that it is an appealing quantity that leads to more powerful procedures than those relying on the FWER. The FDR is defined as the expected proportion of false positives among the total number of positives

$$\text{FDR} = E(Q), \quad \text{where } Q = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases}$$

Broadly speaking, the FDR corresponds to the proportion of rejections that is incorrect.

It is worth noting that a dual quantity of the FDR is the false non-discovery rate (FNR) as defined by Genovese and Wasserman.^[2]

If FWER is the oldest and easiest extension to the classical testing procedure that is relevant for decisional studies (e.g., clinical trials), the FDR seems particularly well suited for exploratory analysis. In practice, the FDR leads to more powerful procedures than those relying on the FWER and gives an idea of the proportion of false positive hypothesis that a practitioner can expect if the experiment is done an infinite number of time. However, being an expectation, it gives very little information about the proportion of false discovery hypothesis in a given experiment.

In their seminal paper, Benjamini and Hochberg^[1] discussed another criterion, later called the positive false discovery rate (pFDR) by Storey, Taylor, and Siegmund^[3] and defined as $\text{pFDR} = E(V/R | R > 0)$. However, Benjamini and Hochberg^[1] did not consider this criterion as it cannot be controlled since under the complete null hypothesis (all null hypothesis tested are true), all significant results (if there are significant ones) are necessary false discoveries. So, it is impossible to insure that $\text{pFDR} < \alpha$ if $\alpha < 1$.

In 2001, Storey^[4] demonstrated that if the test statistics T is independent and identically distributed for a fixed rejection region Γ , which is the same for every test, the pFDR is defined as

$$\begin{aligned} \text{pFDR}(\Gamma) &= \Pr(\mathcal{H}_0 \text{ true} | T \in \Gamma) \\ &= \frac{\Pr(\mathcal{H}_0 \text{ true}) \Pr(T \in \Gamma | \mathcal{H}_0 \text{ true})}{\Pr(T \in \Gamma)} \end{aligned}$$

However, from its definition, the pFDR is obviously related to the FDR through $\text{pFDR} = \text{FDR} / \Pr(R > 0)$. As $\Pr(R > 0)$ tends to 1 when the number of tested hypotheses tends to infinity, these two criteria are asymptotically equivalent.

Two other quantities related to the FDR were also introduced: the q -value and the local FDR. The q -value has been introduced by Storey^[4] and measures the minimum FDR that is incurred when considering that a test is significant. The FDR that has been introduced by Efron et al.^[5] is defined as the FDR in an infinitesimal interval. It may be interpreted as an estimate of the FDR attached to each gene.

CONTROL AND ESTIMATION OF THE FDR

Control of the FDR

In their seminal paper, Benjamini and Hochberg^[1] developed a step-up procedure (here denoted BH) under the hypothesis of independency that controls FDR at a prespecified level α .

The BH procedure works as follows: Suppose $P_1, \dots, P_m$ are p -values for the m tests. Let $P_{(1)} < \dots < P_{(m)}$ denote the ordered p -values. Calculate $k = \max_i \{P_{(i)} \leq \alpha i/m\}$. The procedure rejects all null hypothesis for which $P_{(i)} \leq P_{(k)}$. If the tests are independent, this procedure ensures that

$$\text{FDR} \leq \frac{m_0}{m} \alpha \leq \alpha$$

Benjamini and Yekutieli^[6] extended the procedure to the case of dependent tests. Genovese and Wasserman^[2] and Storey, Taylor, and Siegmund^[3] provided additional results on control of FDR based on martingales.

Estimation of the FDR

Note that the number of false positives and the total number of positive genes depend on a threshold t fixed by the practitioner thus the FDR is also a function of this threshold. In our context, the threshold will be given by ordered p -values

$$V(t) = \#\{\mathcal{H}_0 \text{ true and } p_i \leq t, i = 1, \dots, m\}$$

$$R(t) = \#\{p_i \leq t, i = 1, \dots, m\}$$

If the procedure stops at threshold t , the observed number of positives is g

$$R(t) = g$$

The problem is rather in the calculus of the expected number of false positives.

As it is known that under $\mathcal{H}_0$ p -values are uniformly distributed over $[0,1]$ we have

$$\Pr\{P_i \leq t\} = \mathcal{H}_0 t$$

The expectation of the number of false positives is then $E[V(t)] = m_0 t$, m_0 being the number of true non-differentially expressed genes, which is unknown. Thus a natural estimate of FDR is

$$\widehat{\text{FDR}}(t) = \frac{m_0 t}{g}$$

As the number of true non-differentially expressed genes m_0 is unknown, a classical strategy^[1] is to replace it with m that is known but larger than m_0 . If the aim of the procedure is to control the FDR at level α , then the stopping rule will be

$$P_{(g)} < \frac{g\alpha}{m}$$

Benjamini and Hochberg^[1] and Storey, Taylor, and Siegmund^[3] used convexity argument to prove that $E(\widehat{\text{FDR}}(t)) \geq \text{FDR}(t)$.

As seen previously, $\text{pFDR}(\Gamma) = \Pr(\mathcal{H}_0 \text{ true} | T \in \Gamma)$ that can be reexpressed from the Bayes formula

$$\begin{aligned} \text{pFDR}(\Gamma) &= \Pr(\mathcal{H}_0 \text{ true} | T \in \Gamma) \\ &= \frac{\Pr(T \in \Gamma | \mathcal{H}_0 \text{ true}) \Pr(\mathcal{H}_0 \text{ true})}{\Pr(T \in \Gamma)} \end{aligned}$$

The method proposed by Storey^[7] and Storey and Tibshirani^[8] for estimating the pFDR is based on the separate estimation of the three terms: $\Pr(T \in \Gamma) | \mathcal{H}_0 \text{ true}$, $\Pr(T \in \Gamma)$, and $\pi_0 = \Pr(\mathcal{H}_0 \text{ true})$, which is the probability for a gene to not be differentially expressed. Different methods based on the distribution of the test statistics or the corresponding p -values with or without making any assumption on the distribution related to the modified genes have been proposed. It is worth noting that without any assumption on the distribution of modified genes, only an upper bound estimator of π_0 can be obtained, leading to a conservatively biased estimator for the pFDR.

Estimation Based on the Marginal Density of the Statistics Without Parametric Model on the Modified Gene Distribution

Distribution of the Gene Statistics. In their seminal paper, Tusher, Tibshirani, and Chu^[9] introduced significance analysis of microarrays (SAM) as a statistical technique for finding significant genes in microarrays. This technique aims to estimate the FDR based on the distribution of particular test statistics (T).^[9] The principle is to estimate the quantities $\Pr(T \in \Gamma | \mathcal{H}_0 \text{ true})$, $\Pr(T \in \Gamma)$, and π_0 using resampling techniques^[10] such as:

- The empirical proportion of rejected hypothesis is used to estimate $\Pr(T \in \Gamma)$.
- $\Pr(T \in \Gamma | \mathcal{H}_0 \text{ true})$ is estimated with resampling techniques.^[10]
- π_0 is estimated by $\min(\{\#T \in [\hat{q}_{25} - \hat{q}_{75}]\} / (0.5m); 1)$.

where $\hat{q}_{25}$ and $\hat{q}_{75}$ are the first and third quartiles of the distribution of the test statistics estimated under the null hypothesis with resampling techniques.

More details on the method can be found at the SAM Website: <http://www.stat.stanford.edu/~tibs/SAM/>

Distribution of the p -values. From the distribution of the p -values, other procedures have been proposed that rely on the hypothesis that the p -values are uniformly distributed under the null hypothesis, so that $\Pr(P_i \leq t | \mathcal{H}_0 \text{ true})$ is equal to t . Thus:

$$\text{pFDR}(t) = \frac{\pi_0 t}{\Pr(P_i \leq t)}$$

where the main problem is to estimate the quantity π_0 .

Let f be the marginal probability density function (pdf) of P , f_0 the conditional pdf of P under the null hypothesis, and f_1 be the conditional pdf of P under the alternative hypothesis

$$f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_1(p)$$

Under the null hypothesis, the P -values are supposed to be uniformly distributed on $[0, 1]$ so that $f_0(p) = 1_{[0,1]}(p)$. As $(1 - \pi_0) f_1(p)$ is non-negative, $f(p)$ is an upper bound of π_0 whatever p . Assuming that f_1 (which could itself be a mixture of distribution) is a non-increasing function, the smallest upper bound for π_0 based on previous equation is $f(1)$. Thus, assuming that $f_1(1) = 0$, an unbiased estimator of $f(1)$ provides a conservatively biased estimator of π_0 .

Considering

$$\hat{\pi}_0(\lambda) = \frac{\#\{P_i > \lambda; i = 1, \dots, m\}}{m(1 - \lambda)}$$

as a function of λ , Storey and Tibshirani^[8] proposed the usage of a cubic spline based method to estimate the quantity $\lim_{\lambda \rightarrow 1} \hat{\pi}_0(\lambda)$. Let F be the marginal cumulative distribution function of P ; this estimator can be viewed as $\hat{\pi}_0(\lambda) = [1 - \hat{F}(\lambda)] / (1 - \lambda)$. Then, the estimated quantity is:

$$\lim_{\lambda \rightarrow 1} \hat{\pi}_0(\lambda) = \lim_{\lambda \rightarrow 1} \frac{1 - \hat{F}(\lambda)}{1 - \lambda} = \frac{d\hat{F}}{d\lambda}(1) = \hat{f}(1)$$

Relying on the same framework, two procedures named BUM^[11] and SPLOSH^[12] have been recently proposed. The first one is based on a Beta-Uniform modeling approach, in which the parameters are estimated using the maximum likelihood method, and $\hat{\pi}_0 = \hat{f}(1)$. The second one is based on a local regression method^[13] to obtain a smooth estimate of f in a transformed space (for more details on the transformation used, see Ref. [12]).

Another class of estimators named LBE (Location Based Estimator) has been recently proposed by Dalmaso, Broet, and Moreau^[14] that is based on the expectation of the p -values. At first, let us note that

$$\frac{E(P)}{E_0(P)} = \pi_0 + (1 - \pi_0) \frac{E_1(P)}{E_0(P)}$$

where $E_0(P)$ and $E_1(P)$ are the expectations of the conditional distribution of P under the null and the alternative hypothesis, respectively.

Since the p -values are uniformly distributed under the null hypothesis, $E_0(P) = 1/2$. Then, $\hat{\pi}_0 = 2 \frac{1}{m} \sum_{i=1}^m P_i$ an unbiased estimator of $E(P)/E_0(P)$, is a conservatively biased estimator of π_0 . Under suitable conditions on a function φ the authors show that a transformation of the p -values can lead to a less biased estimator of π_0 . In the set of the suitable functions, they consider the functions $\varphi(P) = -\ln(1-x)^n, n \in \mathbb{N}$, and they show that these functions lead to a family of estimators for which the bias for π_0 is decreasing with n . As under the null hypothesis, $-\ln(1-P)$ follows an exponential distribution with parameter 1, $E_0([-\ln(1-P)]^n) = n!$, and the proposed family of estimators is:

$$\hat{\pi}_{0(n)} = \frac{\frac{1}{m} \sum_{i=1}^m [-\text{Log}(1-p_i)]^n}{n!}, \quad n \in \mathbb{N}$$

For this family of estimators, an upper bound of the asymptotic variance of the estimator is $[(\frac{n}{2n}) - 1]/m$. As there is a balance between bias (decreasing as n increases) and variance (increasing as n increases), for a specified number m of tested hypotheses, the authors intended in practice to choose n according to a threshold l for the variance's upper bound such as

$$n = \max \left(1, \max \left(n \in \mathbb{N} \mid \frac{(\frac{n}{2n}) - 1}{m} \leq l \right) \right)$$

Then, the estimator $\hat{\pi}_0$ is plugged in the FDR formula.

The estimators LBE, QVALUE, BUM, and SPLOSH were compared through simulations. This simulation study has shown that BUM and SPLOSH procedures underestimate π_0 in most of cases, leading to an underestimation of the estimator of the FDR. The simulation study has also shown that, as compared to the procedure QVALUE, the procedure LBE has the smallest mean square error in most cases.

Estimation Based on the Marginal Density of the Statistics with Parametric Model on the Modified Gene Distribution

In the framework of the mixture model with hypothesis on the modified gene distribution, frequentist and Bayesian approaches have been proposed.

A key point is to note that the distribution of the p -values is uniform on the interval $[0,1]$ regardless of the statistical test used (as long as that test is valid) and sample size. In contrast, under the alternative hypothesis, the distribution of p -values is concentrated near 0 and null near 1. Allison et al.^[15] proposed the use of finite mixture models of beta laws (because a beta distribution with parameter 1 and 1 is a uniform distribution). They

estimated the number of component as well as the parameters. BUM^[12] is a restricted version of this work. This work has been extended to the mixture of a uniform and a beta distribution by Liao et al.^[16] Other approaches based on mixture models in a Bayesian framework have also been proposed (for a few: 1) detecting differential gene expression with a semiparametric hierarchical mixture method^[17] or 2) a mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments^[18]).

THE LOCAL FDR

The value of 1%, 5% or 10% for the FDR, which determines the threshold t , is arbitrary. Storey^[7] has stressed the importance of assessing the own measure of significance to each feature. They intended to use the q -value

$$\frac{\hat{m}_0 P_i}{R_i}$$

where P_i is the p -value of the ordered gene i , and R_i is the total number of rejected genes with the threshold $t = P_i$.

The q -value is appealing because it gives a measure of significance that can be attached to each gene; but it must be stressed that it is not an estimate of the probability for the gene to be a false positive. The q -value is generally lower than the latter because it is computed using all the genes that are more significant than gene i . Obviously a gene whose p -value is near to the threshold t does not have the same probability to be differentially expressed than a gene whose p -value is close to 0. Therefore the q -value gives a too optimistic view of the probability for the gene to be a false positive. Therefore it is interesting to obtain an estimate of the FDR attached to each gene from an inferential point of view and without any assumption about the distribution of the p -values under H_1 . It is the idea of the local FDR introduced by Efron et al.^[5,19]

The general model of mixture can be written as $f(t) = p_0 f_0(t) + (1-p_0) f_1(t)$, where p_0 is the fraction of differentially expressed genes, and f_0 and f_1 are the densities of the statistics for non-differentially and differentially expressed genes, respectively. Bayes' theorem generates, under this model, for each gene the probability of no differential expression, given that the value of the statistic is the local FDR and is defined as

$$\frac{p_0 f_0(t)}{f(t)}$$

The local FDR is the a posteriori probability of effect for individual genes. In Ref. [5], the ratio $f_0/f(t)$ is estimated by bootstrap.

We now present an estimation method developed in Aubert et al.^[20]

Let us consider a particular experiment. We observe the differential expression of the genes and compute the

p -values P_i . For simplicity, we consider the ordered p -value and we drop the bracket in the subscript of P_i . At first we assume that m_0 the number of genes under $\mathcal{H}_0(i)$ is known. This assumption will be relaxed later.

The P_i are not random, but the status of the gene associated with the P_i is an unobserved value. It is the same framework that points process on a lattice.^[21] In fact we observe $R(t)$, the sum of two marked point process. The first one $V(t)$ is a counting process associated with non-differentially expressed gene. As the p -value under $\mathcal{H}_0$ is uniformly distributed, $V(t)$ is distributed as a binomial law with parameter m_0 and t . $S(t)$ is the counting process associated with gene under H_1 and very few can be said about its distribution. One may expect that intensity of $S(t)$ is higher around 0 than around 1.

The true parameter of interest for a given experiment is $V(t)$. Using basic properties of binomial law, we obtain

$$P(V(t) \geq v) = \sum_{i \geq v} C_{m_0}^i (m_0 t)^i (1 - m_0 t)^{m_0 - i}$$

From a practical point of view it gives bound to the number of false positive genes selected with a threshold value of t for a given experiment.

It is more useful to get results bounds on a given experiment than on expectation on all possible experiments. Following the same idea, it would be more useful to have results gene by gene rather than globally among all genes. Obviously a gene whose p -value is near the threshold t does not have the same probability to be differentially expressed than a gene whose p -value is closed to 0. The question can be restated as estimation of the ratio of the intensities around a given P_i . Let FDR_i be the local ratio of intensity around P_i . The naive estimate $\hat{E}(V_i)$ is:

$$\hat{E}(V_i) = \frac{m_0 (P_{i+1} - P_{i-1})}{2} \quad (1)$$

It corresponds to an interval around P_i . This estimate is unbiased if the points are uniformly distributed on $[0,1]$. However it is a very variable estimator. This fact is very well known in point process literature.^[21] A more stable estimate given in Ref. [20] is very close to the moving average estimate and corresponds to a generalization of Eq. (1).

Using the definition of FDR_i , it is direct to obtain

$$\sum_{i \leq j} \widehat{\text{FDR}}_i = \widehat{\text{FDR}} \left(\frac{P_j + P_{j+1} - P_1}{2} \right)$$

It is not a strong assumption to assume $p_1 \approx 0$.

An approximation to $\mathbb{P}(V_i = 1)$ can be obtained by noting that $\mathbb{E}(V_i) = \sum_k k \mathbb{P}(V_i = k) \approx \mathbb{P}(V_i = 1)$ for a small enough interval. As we omit positive terms, $\mathbb{E}(V_i) \geq \mathbb{P}(V_i = 1)$. This quantity can give a guideline for the choice of a threshold. Using cost of the experiment, cost of false positive gene

validation, and the profit of discovering a differentially expressed gene, it is direct to compute the optimal strategy for choosing the threshold.

As an interesting application it is possible to compute the FDR associated with a class of gene (for example, genes known to be involved in a given regulation network). These genes do not need to be consecutive for their p -values. It gives the opportunity to compute the FDR of any given group of clones or genes. The following sections demonstrate how the local FDR can be used on several classical datasets.

Local FDR on Golub Data Set

Golub et al.^[22] were interested in identifying genes that are differentially expressed between patients with two types of leukemias (ALL, and AML). Gene expression levels were measured using Affymetrix high-density chips containing 6817 human genes. The learning set comprises 27 ALL cases and 11 AML cases.

Data are available in the R multtest package. We used the preprocessing proposed by the authors and the p -values based on random permutations of the ALL/AML labels on Welch t —statistics for each gene^[23] on the 3051 remaining genes. m_0 is estimated with bootstrap method as suggested by Storey and Tibshirani and implemented in the library GeneTS of software R.

Fig. 1(A) presents the $\widehat{\text{FDR}}(i)$ for ordered genes, and 1(B) presents the smooth curves obtained using lowess with a span of 0.2 and an adaptative moving average method.

We can see that there is an abrupt change of the local FDR around $t = 0.15$ in the smooth local FDR. This may be an indication about the threshold. Fig. 1(C) presents a zoom of Fig. 1(B) for the first 600 p -values. We can see in Fig. 1(C) that if we select the 438 (14%) top genes, we obtain a q -value equal to 0.0078 while the 438th gene has a local FDR equal to 0.027. It must be noticed that there is a big difference between the two measures of FDR because the numerous regulated genes with very small p -values have an important weight in the q -value, which is not the case of the local FDR [Fig. 1(C)].

The p -values have been obtained using random permutations. Therefore the p -values are discrete with several genes possessing the same p -value. Therefore the values of $\widehat{\text{FDR}}(i, \lambda)$ may be equal to 0 because the difference between two successive p -values is 0. The discrete structure of the p -values implies a departure from the theoretical continuous uniform distribution. This explains why the moving average smoothing creates discrete jumps that appear in Fig. 1(C).

If the distribution of the statistics under $\mathcal{H}_0$ is correct, the p -values are distributed as a uniform distribution over $[0,1]$. The empirical distribution is far from the uniform distribution. This makes the hypothesis of the t -statistics very questionable.

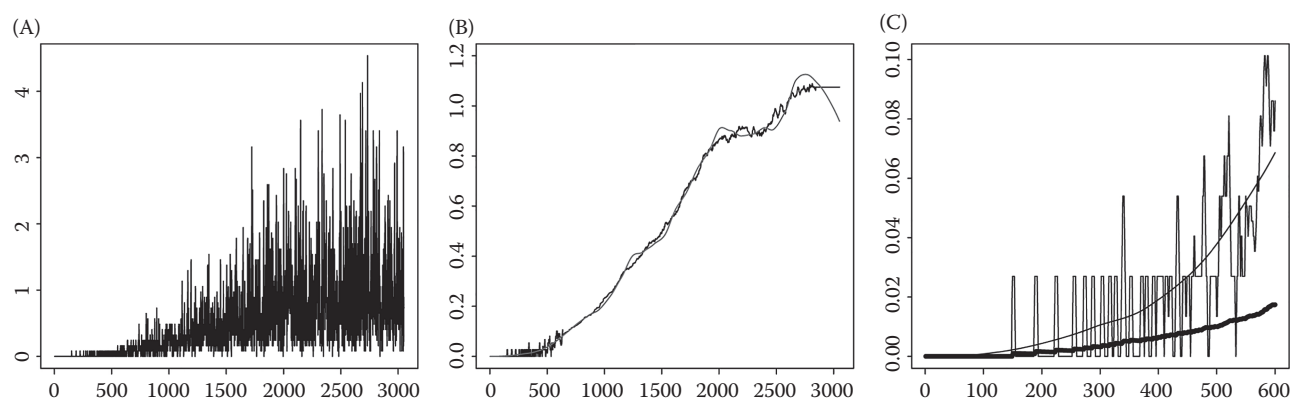


Fig. 1 Plots of the local FDR estimate for Golub data. *x*-axis: index of genes ordered along their *p*-values, *y*-axis: local FDR estimate. (A) raw values, (B) smooth estimates: moving average (discrete jumps), lowess (smooth curve); and (C) zoom on the first 600 genes of (B) moving average (discrete jumps), lowess (upper smooth curve), and *q*-value (lower thick smooth curve). (View this art in color at www.informaworld.com.)

Local FDR on Breast Cancer Data Set

Storey and Tibshirani^[8] have analyzed in detail data from Hedenfalk et al.^[24] on 15 microarrays on breast cancer. Using the same *p*-values, we have computed local FDR estimates. The three genes that have been analyzed in detail by Storey and Tibshirani^[8] are presented in Table 1.

One can see that the smooth local FDR estimate is greater than the *q*-value and gives a better idea of the probability that a gene is false positive. The difference of the two measures is sufficiently high to lead us to reconsider the scientific conclusion.

Fig. 2(A) presents the $\widehat{FDR}(i)$ for ordered genes, and 2(B) presents the smooth curves obtained using lowess with a span of 0.2 and moving average methods. The two smoothing methods give similar results. Setting $\lambda = 0.5$, Storey and Tibshirani^[8] estimate that 67% of the 3170 genes in the data are not differentially expressed. The asymptote near 1 of the smooth curve supports this estimation.

Local FDR on ApoAI Data

The goal of the study is to identify genes with altered expression in the livers of two lines of mice with very low high density lipoproteins (HDL) cholesterol levels compared to inbred control mice. The mouse model is the apolipoprotein AI (ApoAI) knock-out mice. ApoAI is a gene known to play a pivotal role in HDL metabolism. The statistical analysis is described in Dudoit, Shaffer, and Boldrick.^[23] Height clones are expected to be regulated between the control and the knock-out mice because they are clones of the ApoAI gene or of genes coregulated with ApoAI. The height clones are actually the eight top clones detected by the statistical tests. However there are other

Table 1 *p*-Value, *q*-Value and Local FDR Estimates for Three Genes in Hedenfalk data

Gene	<i>p</i> -value	<i>q</i> -value	Local FDR
MSH2	0.00005	0.0013	0.010
PDCD5	0.00048	0.022	0.033
CTGF	0.0036	0.049	0.098

following clones that seem statistically significant if we consider the *q*-value. We can see in Fig. 3(C) that the local FDR values are much higher than the *q*-values.

Fig. 3(A) presents the $\widehat{FDR}(i)$ for ordered clones and Fig. 3(B) presents the smooth curves obtained using lowess with a span of 0.2 and moving average methods. The two smoothing methods give different results at the two ends of the $[0,1]$ interval. The moving average method that uses a special adaptative algorithm for the ends gives a better smoothing. This is particularly important for the clones with a small *p*-value for which it is crucial to obtain good estimates of the probability of being false positives. The lowess smoothing does not work well for the 50 first clones. In this particular case the default smoothing parameter $f = 0.2$ is not well suited and should be lower. However if it is chosen too low, the smoothing will not be well suited for the rest of the curve.

There are two clones on the gene ApoAI. If we want to estimate the FDR of these two clones taken in a whole, we compute the sum of the smoothed local FDR of the two clones (the first and the height top clones) and obtain a local FDR for the gene Apo AI, which is equal to $(0 + 0.00048)/2 = 0.00024$. This example shows that it is possible to estimate the local FDR of any group of clones. This opportunity provided by the local FDR is certainly one of its major advantage with many potential applications.

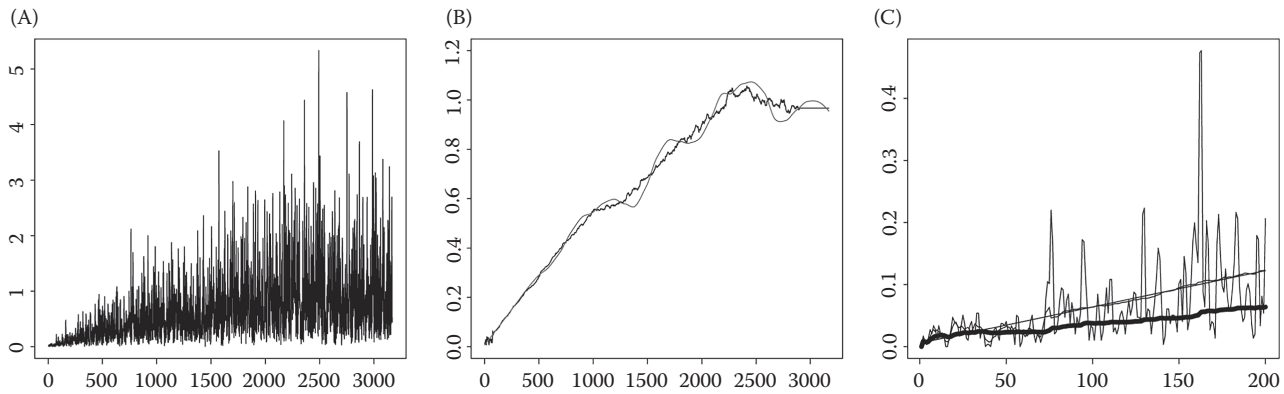


Fig. 2 Plots of the local FDR estimate for Hedenfalk data. *x*-axis: index of genes ordered along their *p*-values, *y*-axis: local FDR estimate, (A) raw values; (B) smooth estimates: moving average (discrete jumps), lowess (smooth curve); and (C) zoom on the first 200 genes of (B) raw values (discrete jumps), moving average and lowess (smooth curves); and *q*-value (lower thick smooth curve). (View this art in color at www.informaworld.com.)

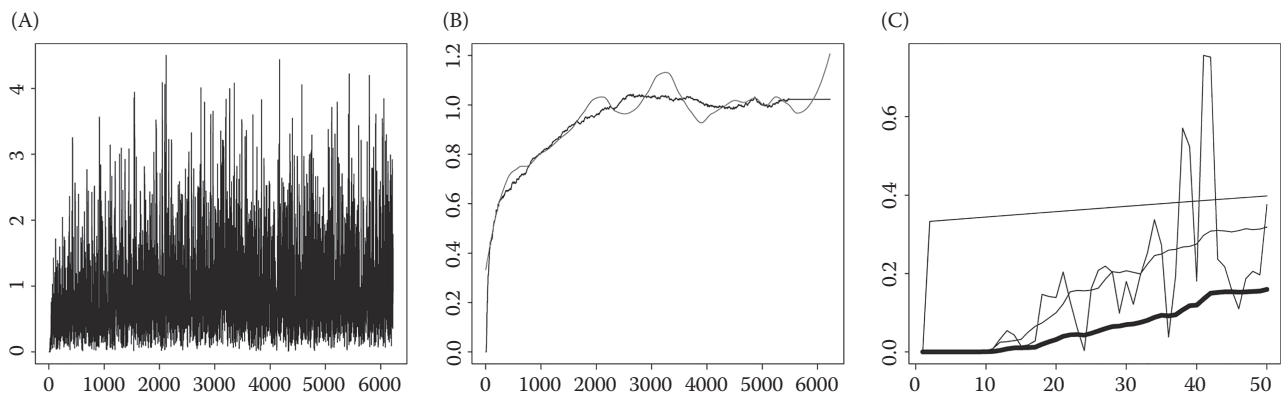


Fig. 3 Plots of the local FDR estimate for Apo-AI data. *x*-axis: index of clones ordered along their *p*-values, *y*-axis: local FDR estimate, (A) raw values; (B) smooth estimates: moving average (small discrete jumps), lowess (smooth curve); and (C) zoom on the 50 first genes of (B) raw values (discrete jumps), moving average (smooth curve) lowess (upper rectangular curve), and *q*-value (lower thick smooth curve). (View this art in color at www.informaworld.com.)

CONCLUSIONS

To date, statistical procedures devoted to the multiple testing problem mostly focused on controlling the FWER, which is the easiest extension to the classical test theory. However, controlling the FWER usually leads to a severe loss of power, particularly when analyzing high-dimensional data sets. In this context, the expected proportion of false discoveries among the total number of discoveries introduced in 1995 by Benjamini and Hochberg seems more appealing. Then, the FDR (and its related quantities) is nowadays widely used, and more and more procedures based on this criterion are developed.

While some procedures^[1,6] are based on a strict control of the FDR, many new procedures are focused on the estimation of this criterion through the relation between the *p*FDR and the FDR.^[7] In the framework of estimating procedures that do not make any assumption on the distribution

of modified genes, it is worth noting that all the proposed estimators necessarily overestimate the FDR. This problem is related to the fact that only an upper bound estimate can be obtained for the key quantity π_0 , which is the probability for a gene to be unmodified. Procedures based on distributional hypothesis for modified genes provide more accurate estimators of the FDR when the distributional hypotheses are true. However, these procedures are more complex to use in practice.

Among the extensions of the FDR, one of the most promising criterion is the local FDR introduced by Efron in 2001^[5] with an estimating procedure, which may be interpreted as an estimate of the FDR attached to each gene.

In the method proposed by Aubert et al.,^[20] the curve of the smoothed local FDR is an efficient tool to summarize the information about the number and the statistical significance of regulated genes, and may also be used to give an

indication about the validity of the statistical assumptions. Moreover it is a valuable tool to choose the threshold for separating the regulated genes from the unregulated one: One can choose a value of t maximizing the second derivative. Alternatively one can use a cost function and choose the threshold that minimizes the mean cost for a given cost function: Using cost of the experiment, cost of false positive gene validation, and the profit of discovering a differentially expressed gene, it is direct to compute the optimal strategy for choosing the threshold. One of the major interest of the local FDR is that it gives the opportunity to compute the FDR of any given group of clones (of the same gene) or genes pertaining to the same regulation network or the same chromosome.

REFERENCES

- Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. Ser. B.* **1995**, *57* (1), 289–300.
- Genovese, C.; Wasserman, L. Operating characteristics and extensions of the FDR procedure. *J. Royal Stat. Soc. Ser. B.* **2002**, *64* (1), 499–518.
- Storey, J.D.; Taylor, J.E.; Siegmund, D. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Royal Stat. Soc. Ser. B.* **2004**, *66*, 187–205.
- Storey, J.D. A direct approach to false discovery rates. *J. Royal Stat. Soc. Ser. B.* **2001**, *64*, 479–498.
- Efron, B.; Tibshirani, R.; Storey, J.D.; Tusher, V. Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **2001**, *96*, 1151–1160.
- Benjamini, Y.; Yekutieli, D. The control of the false discovery rate under dependency. *Ann. Stat.* **2001**, *29*, 1165–1188.
- Storey, J.D. The positive false discovery rate: a Bayesian interpretation and the q -value. *Ann. Stat.* **2003**, *31*, 2013–2035.
- Storey, J.D.; Tibshirani, R. Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci.* **2003**, *100*, 9440–9445.
- Tusher, V.G.; Tibshirani, R.; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA.* **2001**, *98*, 5116–5121.
- Ge, Y.; Dudoit, S.; Speed, T.P. Resampling-based multiple testing for microarray data hypothesis. *Test* **2002**, *12*(1), 1–44.
- Pounds, S.; Cheng, C. Improving false discovery rate estimation. *Bioinformatics.* **2004**, Advance Access published on February 26.
- Pounds, S.; Morris, S.W. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p -values. *Bioinformatics.* **2003**, *19*, 1236–1242.
- Loader, C. *Local Regression and Likelihood*; Springer-Verlag: New York, 1999.
- Dalmasso, C.; Broët, P.; Moreau, T. A simple procedure for estimating the false discovery rate. *Bioinformatics.* **2004**, Advance Access published on 12th Oct.
- Allison, D.B.; Gadbury, G.; Heo, M.; Fernandez, J.; Lee, C.-K.; Prolla, T.A.; Weindrich, R. A mixture model approach for the analysis of microarray gene expression data. *Comput. Stat. Data Anal.* **2002**, *39*, 1–20.
- Liao, J.G.; Lin, Y.; Selvanayagam, Z.E.; Weichung, J.S. A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics* **2004**, *20* (16), 2694–2701.
- Newton, M.A.; Noueiry, A.; Sarkar, D.; Ahlquist, P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **2004**, *5*, 155–176.
- Broët, P.; Lewin, A.; Richardson, S.; Dalmasso, C.; Magdelenat, H. A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics.* **2004**, *20*, 2562–2571.
- Efron, B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Stat. Assoc.* **2004**, *99*, 96–104.
- Aubert, J.; Bar-Hen, A.; Daudin, J.J.; Robin, S. Determination of the differentially expressed genes in microarray experiments using local FDR. *BMC Bioinform.* **2004**, *5* (1), 1–125.
- Cressie, N. *Statistics for Spatial Data*; revised edition, Wiley: New York, 1993.
- Golub, T.R.; Slonim, D.K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, M.; Coller, J.P.; Loh, M.; Downing, J.R.; Caligiuri, M.A.; Bloomfield, C.D.; Lander, E.S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.* **1999**, *286*, 531–537.
- Dudoit, S.; Shaffer, J.P.; Boldrick, J.C. Multiple hypothesis testing in microarray experiments. *Stat. Sci.* **2003**, *18* (1), 71–103.
- Hedenfalk, I.; Duggan, D.; Chen, Y.; Radmacher, M.; Bittner, M.; Simon, R.; Meltzer, P.; Gusterson, B.; Esteller, M.; Kallioniemi, O.P. Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* **2001**, *344*, 539–548.

GEE in Cancer Research

Sin-Ho Jung

Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, U.S.A.

Abstract

Cancer clinical trials often use a repeated measurement design in which individuals are randomly assigned to treatment arms and followed for measurements at intervals across a treatment period of fixed duration. In such studies, one of popular primary interests is the between-arm comparison of change rate in the expected value of a response variable over time. Because of the robustness to possible misspecification of the correlation structure among the repeated measurements, generalized estimating equation (GEE) method has been one of the most popular methods to fit the regression models and to test on the change rates. In this article, we review GEE method for analysis and sample size estimation methods for design of such trials. We consider the cases where the response variable is continuous or dichotomous.

Keywords: AR(1); Compound symmetry; Independent working correlation; Independent missing; Logit link; Missing completely at random; Monotone missing.

1 INTRODUCTION

Fisch et al.^[1] randomized advanced cancer outpatients between fluoxetine and placebo arms, and longitudinal assessments of quality of life (QOL) and depression were performed at baseline and every 3 to 6 weeks thereafter. The visit interval varied among patients and often depended on the schedule for anticancer therapy. Patients were assessed for 12 weeks, and complete assessment involved three to five sessions of data collection depending on the individual patients visit intervals. After the 12-week study period, the patients were given the option to continue the study drug for up to 9 more months. There were missing assessments too. The primary objective of this study was to compare the change rate of QOL and depression in advanced cancer patients over time. In this study, the outcome variables are continuous.

GENISOS (Genetics vs. Environment In Scleroderma Outcome Study) is an observational study designed as collaboration of the University of Texas-Houston Health Science Center with the University of Texas Medical Branch at Galveston and the University of Texas-San Antonio Health Science Center (Reveille et al.^[2]). Scleroderma, or systemic sclerosis, is a multisystem disease of unknown etiology characterized by cutaneous and visceral fibrosis, small blood vessel damage and autoimmune features (Medsger^[3]). The study subjects are regularly followed to check the occurrence pulmonary fibrosis. In this case, the outcome is a binary variable.

Typically repeated measurements from each subject are correlated. Furthermore, the subjects may skip some

visits so that the resulting data are subject to missing. GEE method (Liang and Zeger^[4]) has been one of the most popular methods to relate repeatedly measured outcome variable with covariates including the measurement time. If the missing is completely at random (Rubin^[5]), GEE method provides consistent regression estimator whether the working correlation structure specified for the repeated measurements is the true one.

In this article, we consider comparing the change rate of repeated measurements over time between two treatment groups. We propose a closed-form sample size formula for between-group comparison that can be used when designing a randomized longitudinal study with either continuous or binary outcome variable. Missing data attenuate the power of tests on the regression coefficients. Our formula account for the missing probabilities under independent or monotone missing pattern. We assume that the working independent correlation model is used in the final data analysis, but the sample size formula is derived to incorporate the influence of the true correlations, especially under AR(1) or compound symmetry.

The methods presented in this chapter can be used for randomized trials and observational studies for any diseases including cancer.

2 GENERALIZED ESTIMATING EQUATIONS

Suppose that n_k subjects are allocated to treatment group k ($k = 1, 2$), $n_1 + n_2 = n$. Let, for group k , $N_k = \sum_{i=1}^{n_k} m_{ki}$ denote

the total number of observations and $r_k = n_k/n$ the allocation proportion ($r_1 + r_2 = 1$).

For subject $i(i = 1, \dots, n_k)$ in group k , let y_{kij} denote the outcome variable at measurement time $t_{kij}(j = 1, \dots, m_{ki})$ with $\mu_{kij} = E(y_{kij})$ that is expressed as

$$g(\mu_{kij}) = a_k + b_k t_{kij},$$

where $g(\cdot)$ is a link function. To simplify the discussions, we use the identity link $g(\mu) = \mu$ for continuous outcome variables and the logit link $g(\mu) = \log\{\mu/(1-\mu)\}$ for binary outcome variables, but the generalization to the use of other links is simple.

The coefficient b_k represents the change rate per unit time in mean response if y is continuous and that in log-odds if y is binary. The measurement times may vary subject by subject due to missing measurements, patients' visits for measurements at unscheduled times, loss to follow-up, or other causes. In this article, we assume missing completely at random by Rubin (1976).

The repeated measurements ($y_{kij}, 1 \leq j \leq m_{ki}$) within each subject tend to be correlated. However, the true correlation structure is usually unknown or of secondary interest. Liang and Zeger^[4] proposed a consistent estimator, called the GEE estimator, based on a working correlation structure. Using either the identity or logit link, the GEE estimator $(\hat{a}_k, \hat{b}_k)$ based on the working independent structure solves $U_k(a, b) = 0$, where

$$U_k(a, b) = \frac{1}{\sqrt{n_k}} \sum_{i=1}^{n_k} \sum_{j=1}^{m_{ki}} \{y_{kij} - \mu_{kij}(a, b)\} \begin{pmatrix} 1 \\ t_{kij} \end{pmatrix}$$

and $\mu_{kij}(a, b) = g^{-1}(a + b t_{kij})$.

2.1 Continuous Outcome Variable Case

In the continuous outcome variable case, we have a closed form solution to $U_k(a, b) = 0$

$$\hat{b}_k = \frac{\sum_{i=1}^{n_k} \sum_{j=1}^{m_{ki}} (t_{kij} - \bar{t}_k) y_{kij}}{\sum_{i=1}^{n_k} \sum_{j=1}^{m_{ki}} (t_{kij} - \bar{t}_k)^2}$$

and $\hat{a}_k = \bar{y}_k - \hat{b}_k \bar{t}_k$, where $\bar{t}_k = N_k^{-1} \sum_{i=1}^{n_k} \sum_{j=1}^{m_{ki}} t_{kij}$ and $\bar{y}_k = N_k^{-1} \sum_{i=1}^{n_k} \sum_{j=1}^{m_{ki}} y_{kij}$.

By Liang and Zeger^[4], as $n \rightarrow \infty$, $\sqrt{n_k}(\hat{b}_k - b_k) \rightarrow N(0, v_k)$ in distribution. Here v_k is consistently estimated by

$$\hat{v}_k = \frac{\sum_{i=1}^{n_k} \left\{ \sum_{j=1}^{m_{ki}} (t_{kij} - \bar{t}_k) \hat{\epsilon}_{kij} \right\}^2}{\left\{ \sum_{i=1}^{n_k} \sum_{j=1}^{m_{ki}} (t_{kij} - \bar{t}_k)^2 \right\}^2}$$

where $\hat{\epsilon}_{kij} = y_{kij} - \hat{a}_k - \hat{b}_k t_{kij}$.

2.2 Binary Outcome Variable Case

Let $p_{kij} = \mu_{kij}$ in the binary outcome variable case. We have to solve the equation using a numerical method, such as the Newton-Raphson algorithm: at the l -th iteration,

$$\begin{pmatrix} \hat{a}_k^{(l)} \\ \hat{b}_k^{(l)} \end{pmatrix} = \begin{pmatrix} \hat{a}_k^{(l-1)} \\ \hat{b}_k^{(l-1)} \end{pmatrix} + n_k^{-1/2} A_k^{-1}(\hat{a}_k^{(l-1)}, \hat{b}_k^{(l-1)}) U_k(\hat{a}_k^{(l-1)}, \hat{b}_k^{(l-1)}),$$

where

$$A_k(a, b) = -n_k^{-1/2} \frac{\partial U_k(a, b)}{\partial(a, b)} = \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{m_{ki}} p_{kij}(a, b) q_{kij}(a, b) \begin{pmatrix} 1 & t_{kij} \\ t_{kij} & t_{kij}^2 \end{pmatrix}$$

$$p_{kij}(a, b) = g^{-1}(a + b t_{kij}) = \frac{e^{a + b t_{kij}}}{1 + e^{a + b t_{kij}}} \quad (1)$$

and $q_{kij}(a, b) = 1 - p_{kij}(a, b)$.

By Liang and Zeger^[4], for large n , $\sqrt{n_k}(\hat{a}_k - a_k, \hat{b}_k - b_k)^T$ is asymptotically normal with mean 0 and variance V_k that can be consistently estimated by

$$\hat{V}_k = A_k^{-1}(\hat{a}_k, \hat{b}_k) \hat{\Sigma}_k A_k^{-1}(\hat{a}_k, \hat{b}_k)$$

Where

$$\hat{\Sigma}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \left\{ \sum_{j=1}^{m_{ki}} \hat{\epsilon}_{kij} \begin{pmatrix} 1 \\ t_{kij} \end{pmatrix} \right\}^{\otimes 2},$$

$\hat{\epsilon}_{kij} = y_{kij} - p_{kij}(\hat{a}_k, \hat{b}_k)$ and $c^{\otimes 2} = c c^T$ for a vector c . Let $\hat{v}_k$ be the $(2, 2)$ -component of $\hat{V}_k$.

3 SAMPLE SIZE CALCULATION

Suppose that we want to test on the rate of change between two groups, i.e. $H_0: b_1 = b_2$. Based on the asymptotic results from the previous section, we can reject $H_0: b_1 = b_2$, in favor of $H_1: b_1 \neq b_2$ when

$$\left| \frac{\hat{b}_1 - \hat{b}_2}{\sqrt{\hat{v}_1 / n_1 + \hat{v}_2 / n_2}} \right| > z_{1-\alpha/2}, \quad (2)$$

where $z_{1-\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of a standard normal distribution.

In this section, we derive a sample size formula for the two-sided α test (2) to detect $H_1: |b_2 - b_1| = d(>0)$ with power $1 - \beta$. When designing a study, we usually

schedule fixed visit times $t_1 < \dots < t_m$ for m repeated measurements from each subject. We often set $t_1 = 0$ for the baseline measurement time. When the study is conducted, however, the subjects may skip some visit times due to various causes, which results in missing values, or may not follow the visit schedule correctly so that the observed visit times may be variable. Jung and Ahn^[6] show through simulations in continuous outcome variable case that the sample size formula based on fixed measurement times is very accurate even when the observed measurement times are widely distributed around the scheduled times.

By simple algebra, we can derive a sample size formula to detect the specified difference $|b_2 - b_1| = d$ with power $1 - \beta$,

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 (v_1/r_1 + v_2/r_2)}{d^2}, \quad (3)$$

where $v_k = \lim_{n \rightarrow \infty} \hat{v}_k$. The expression of v_k is slightly different between continuous and binary outcome variable cases as shown in the following subsections.

In order to allow for missing, let δ_j denote the proportion of patients with observation at t_j and $\delta_{jj'}$ the proportion of patients with observations at both t_j and $t_{j'}$. Also, let $\rho_{jj'} = \text{corr}(y_{kij}, y_{kij'})$. We assume that the missing pattern and correlation structure are common in two treatment groups. We will discuss specific models for missing pattern and correlation structure in Section 4.

3.1 Continuous Outcome Variable Case

We assume that the continuous outcome variable has a constant variance $\text{var}(y_{kij}) = \sigma^2$ over time which is common between two treatment groups. Jung and Ahn^[6] show that $v_1 = v_2 = v$ is expressed as

$$v = \frac{\sigma^2(s^2 + c)}{s^4},$$

Where

$$s^2 = \sum_{j=1}^m \delta_j (t_j - \tau)^2$$

$$c_k = \sum_{j \neq j'} \sum \delta_{jj'} \rho_{jj'} (t_j - \tau)(t_{j'} - \tau)$$

and

$$\tau = \frac{\sum_{j=1}^m \delta_j t_j}{\sum_{j=1}^m \delta_j}.$$

Hence, (3) is simplified to

$$n = \frac{v(z_{1-\alpha/2} + z_{1-\beta})^2}{d^2 r_1 r_2}. \quad (4)$$

Note that we do not have to specify the true values for $\{(a_k, b_k), k=1,2\}$, but only the difference $|b_2 - b_1| = d$ in sample size calculation. The sample size is proportional to the variance of measurement error σ^2 , and decreases as r_1 approaches $1/2$.

Chow et al.^[7] propose a sample size formula for ANOVA with repeated measures. Their formula is similar to the formula for K-sample by Jung and Ahn^[8] under CS and no missing.

3.2 Binary Outcome Variable Case

Let $p_{kj} = a_k + b_k t_j$ denote the success probabilities under H_1 . Jung and Ahn^[9] show that

$$v_k = \frac{(s_k^2 + c_k)}{s_k^4},$$

Where

$$s_k^2 = \sum_{j=1}^m \delta_j p_{kj} q_{kj} (t_j - \tau_k)^2$$

$$c_k = \sum_{j \neq j'} \sum \delta_{jj'} \rho_{jj'} \sqrt{p_{kj} q_{kj} p_{kj'} q_{kj'}} (t_j - \tau_k)(t_{j'} - \tau_k),$$

and

$$\tau_k = \frac{\sum_{j=1}^m \delta_j p_{kj} q_{kj} t_j}{\sum_{j=1}^m \delta_j p_{kj} q_{kj}}.$$

As shown here, the sample size formula under binary outcome variable case depends on the probabilities $(p_{kj}, j=1, \dots, m)$ under H_1 , so that we need to specify all regression parameters $\{(a_k, b_k), k=1,2\}$. Let $d = b_2 - b_1$ denote the difference in slope between two groups under H_1 . If there exist pilot data for the control group (group 1), then we may use the estimates as the parameter values, a_1 and b_1 . If $t_1 = 0$ is the time of randomization, the intercepts in the two groups are set the same, i.e. $a_1 = a_2$. By setting $b_2 = b_1 + d$, we can specify all regression coefficients under H_1 .

If there are no pilot data available, we may specify the binary probabilities at the baseline, p_{11} , and at the end of follow-up, p_{1m} , for the control group. Then we obtain (a_1, b_1) by

$$b_1 = \frac{g(p_{1m}) - g(p_{11})}{t_m - t_1}. \quad (5)$$

$$a_1 = g(p_{11}) - b_1 t_1 = g(p_{11}).$$

And we set $a_2 = a_1$ (since $p_{11} = p_{21}$ at the baseline) and $b_2 = b_1 + d$.

4 MODELING MISSING PATTERN AND CORRELATION STRUCTURE

Calculation of n (or v_k) requires projection of the missing probabilities and the true correlation structure.

4.1 Missing Pattern

In order to specify $\delta_{jj'}$, we need to estimate the missing pattern. If missing at time t_j is independent of missing at time $t_{j'}$ for each patient, then we have $\delta_{jj'} = \delta_j \delta_{j'}$ and call this type *independent missing*.

In some studies, subjects missing at a measurement time may be missing at all following measurement times as in the labor pain study discussed in Introduction. This type of missing is called *monotone missing*. In this case, we have $\delta_{jj'} = \delta_{j'}$ for $j < j'$ ($\delta_1 \geq \dots \geq \delta_m$). In monotone missing case, one may want to specify the proportion of patients who will contribute exactly the first j observations, say η_j . Then, noting that $\eta_j = \delta_j - \delta_{j+1}$ for $j = 1, \dots, m-1$ and $\eta_m = \delta_m$, we can obtain δ_j recursively starting from δ_m . Note also that $\sum_j \eta_j = \delta_1$, which equals 1 if all patients have measurements at the first measurement time t_1 .

In summary, we consider two missing patterns as candidates to the true one: (a) independent missing, where $\delta_{jj'} = \delta_j \delta_{j'}$; (b) monotone missing, where $\delta_{jj'} = \delta_{j'}$ for $j < j'$ ($\delta_1 \geq \dots \geq \delta_m$).

4.2 Correlation Structure

Now, we specify the ‘true’ correlation structure $\rho_{jj'}$. A reasonable model for $\epsilon_{ij} = y_{ij} - g^{-1}(a + bt_{ij})$ may be

$$\epsilon_{ij} = u_i + e_{ij},$$

where u_i are subject-specific error terms with variance σ_u^2 and e_{ij} are serially correlated within-subject error terms with variance σ_e^2 and correlation coefficients $\tilde{\rho}_{jj'}$. Assuming that u_i and e_{ij} are independent, we have

$$\text{cov}(\epsilon_{ij}, \epsilon_{ij'}) = \sigma_u^2 + \sigma_e^2 \tilde{\rho}_{jj'}.$$

Often, the variation between subjects (σ_u^2) is much larger than that within subject (σ_e^2). In this case, we have

$$\text{cov}(\epsilon_{ij}, \epsilon_{ij'}) \approx \sigma_u^2$$

and a compounding symmetry (CS) correlation structure, $\rho_{jj'} = \rho$ for $j \neq j'$, may be a reasonable approximation to the true one.

On the other hand, if the variation within subject (σ_e^2) dominates over the variation between subjects (σ_u^2), then we will have

$$\text{cov}(\epsilon_{ij}, \epsilon_{ij'}) \approx \sigma_e^2 \tilde{\rho}_{jj'},$$

so that a serial correlation structure may be a reasonable approximation to the true one. One of the most popular serial correlation structure, especially when measurement times are not equidistant, is a continuous autocorrelation model with order 1, AR(1), for which $\rho_{jj'} = \rho^{|t_j - t_{j'}|}$.

We consider two correlation structures as candidate approximations to the true one: (i) CS, where $\rho_{jj'} = \rho$; (ii) AR(1), where $\rho_{jj'} = \rho^{|t_j - t_{j'}|}$.

5 EXAMPLES

For sample size calculation, following parameters need to be specified commonly in both continuous and discrete outcome variable cases:

- Type I and II errors α and β , respectively.
- Clinically significant difference, $|b_2 - b_1| = d$.
- Allocation proportion for group k , r_k .
- Correlation structure and the associated correlation parameter ρ , i.e. (i) $\rho_{jj'} = \rho$ for CS or (ii) $\rho_{jj'} = \rho^{|t_j - t_{j'}|}$ for AR(1).
- Proportion of patients with an observation at t_j , δ_j .
- Missing pattern: (a) independent missing ($\delta_{jj'} = \delta_j \delta_{j'}$) or (b) monotone missing ($\delta_{jj'} = \delta_{j'}$ for $j < j'$).

In addition, we need to specify $\sigma^2 = \text{var}(\epsilon_{ij})$ in the continuous outcome variable case, and the regression coefficients $\{(a_k, b_k), k = 1, 2\}$ under H_1 in the binary outcome variable case.

We demonstrate our sample size formula with real longitudinal studies.

5.1 Labor Pain Study (Continuous Outcome Case)

Davis^[10] introduced a study on labor pain where women in labor were randomly assigned to pain medication group and placebo arms. At 30 minutes intervals, the self-reported amount of pain was marked on a 100 mm line, where 0 = no pain and 100 = extreme pain. The maximum number of measurements for each woman was $m = 6$, but there were numerous missing values at later measurement times with monotone missing pattern. A simple approach to such study objective may be to estimate and compare the slopes of pain scores over time in two treatment groups. In this study, the outcome variable is continuous. Suppose that we want to design a new study on labor pain based on the data reported by Davis^[10]. As in the original study, we assume monotone missing. In this study, the measurement times

were equispaced, so that we set $t_j = j - 1$ ($j = 1, \dots, 6$) for convenience. From the data, we obtain $\sigma^2 = 815.84$. Suppose we want to detect a difference of σ in mean pain score between two groups at t_6 . So, we project $d = \sigma/(t_6 - t_1) = 5.71$ in a new study. We consider a balanced randomization, i.e. $r_1 = r_2 = 1/2$. Also, from the data, the proportion of observed measurements are

$$(\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6) = (1, .90, .78, .67, .54, .41).$$

From these results, we obtain $\tau = 2.02$ and $s^2 = 11.42$.

Suppose that we want to detect a difference of $d = 5.71$ with 80% of power ($z_{1-\beta} = 0.84$) using a two-sided $\alpha = .05$ ($z_{1-\alpha/2} = 1.96$) test. Under CS, we obtain $\rho = .64$ and $c = -3.13$ from the data, so that we have $v = 815.84 \times (11.42 - 3.13)/11.42^2 = 0.0635$. Hence, from (4), the required sample size is calculated as

$$n = \left\lceil \frac{0.0635 \times (1.96 + 0.84)^2}{5.71^2} \right\rceil + 1 = 50.$$

where $[x]$ is the largest integer not exceeding x .

Under AR(1), we obtain from the data $\rho = .80$ and $c = 2.31$, so that the required sample size for detecting $d = 5.71$ with 80% of power using a two-sided $\alpha = .05$ test is given as $n = 83$. The sample size under AR(1) is larger than that under CS, by about 66%, in this example.

5.2 GENISOS (Binary Outcome Case)

As stated in Introduction, GENISOS is an observational study. Suppose that we want to develop a new clinical trial to examine the effect of a new drug in preventing the occurrence of pulmonary fibrosis in subjects with scleroderma based on the observations from the observational study. The parameter values specified below for sample size calculation are approximated by the estimates from the current data set of GENISOS.

We want to estimate the sample size for the new clinical trial using $\alpha = .05$ and $1 - \beta = .8$. As in GENISOS, presence or absence of pulmonary fibrosis will be assessed at baseline, and at months 6, 12, 18, 24 and 30. Since the measurement times are equidistant, we set $t_j = j - 1$ ($j = 1, \dots, 6$) for the $m = 6$ time points. The within-group correlation structure of the repeated measurements conforms to AR(1) with the adjacent correlation equal to $\rho = .8$, that is, $\rho_{jj'} = .8^{|j-j'|}$. We consider assigning equal number of scleroderma patients in each of two groups, i.e. $r_1 = r_2 = 1/2$.

Approximately 75% of scleroderma patients do not have pulmonary fibrosis at the baseline in the ongoing GENISOS. We project that the proportion of subjects without pulmonary fibrosis is 75% at baseline, i.e. $p_{1,1} = .75$, and 50% at 30 months, i.e. $p_{1,6} = .50$, in a placebo group. We assume that a new therapy will prevent or delay further

occurrence of pulmonary fibrosis. That is, the proportion of subjects without pulmonary fibrosis will remain 75% during 30-month study in a new therapy group, i.e. $p_{2,1} = p_{2,6} = .75$. From these values and (5), we obtain

$$b_1 = \frac{g(.5) - g(.75)}{5 - 0} = -0.220,$$

and $a_1 = g(.75) = 1.099$. Similarly, we obtain $(a_2, b_2) = (1.099, 0)$, so that $d = 0 - (-0.220) = 0.220$. By (1), the probabilities of no pulmonary fibrosis at the 6 time points are obtained as $(.750, .707, .659, .608, .555, .500)$ for the placebo group, and $(.750, .750, .750, .750, .750, .750)$ for the treatment group.

The proportions of observed measurements are expected to be

$$(\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6) = (1, .95, .90, .85, .80, .75).$$

Suppose that we expect independent missing. Then, we obtain $\delta_{jj'}$ from the specified δ_j values using $\delta_{jj'} = \delta_j \delta_{j'}$. Now we have all the parameter values required, and obtain $v_1 = 0.305$ and $v_2 = 0.353$. Finally, from (3), we obtain

$$n = \left\lceil \frac{(1.96 + 0.84)^2 (0.305 / 0.5 + 0.353 / 0.5)}{0.220^2} \right\rceil + 1 = 215.$$

If we assume monotone missing ($\delta_{jj'} = \delta_{j \vee j'}$), we obtain $v_1 = 0.324$, $v_2 = 0.308$ and $n = 229$.

ARTICLES OF RELATED INTEREST

Analysis of Repeated Measures Data with Missing Values: An Overview of Method; Clinical Trials; Logistic Regression; Randomization.

REFERENCES

1. Fisch, M.J.; Loehrer, P.J.; Kristeller, J.; Passik, S.; Jung, S.H.; Shen, J.; Brames, M.J.; Einhorn, L.H. Fluoxetine versus placebo in advanced cancer outpatients: A double-blinded trial of the Hoosier Oncology Group. *Journal of Clinical Oncology* **2003**, *21* (10), 1937–1943.
2. Reveille, J.; Fischbach, M.; McNearney, T.; Friedman, A.W.; Aguilar, M.B.; Lisse, J.; Fritzler, M.J.; Ahn, C.; Amett, F.C. Systemic sclerosis in 3 US ethnic groups: A comparison of clinical, sociodemographic, serologic and immunologic determinants. *Seminars in Arthritis and Rheumatism* **2001**, *30* (5), 332–346.
3. Medsger, T.A.J. Systemic sclerosis (scleroderma): clinical aspects. In *Arthritis and Allied Conditions*, 13th Ed.; Koopman W.J., Ed.; Williams and Wilkins: Baltimore, MD, 1997, 1433–1464.

4. Liang, K.Y.; Zeger, S.L. Longitudinal data analysis using generalized linear models. *Biometrika* **1986**, *73* (1), 13–22.
5. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63* (3), 581–592.
6. Jung, S.H.; Ahn, C. Sample size estimation for GEE method for comparing slopes in repeated measurements data. *Statistics in Medicine* **2003**, *22* (8), 1305–1315.
7. Chow, S.C.; Shao, J.; Wang, H. *Sample Size Calculations in Clinical Research*. Taylor & Francis: New York, 2003.
8. Jung, S.H.; Ahn, C. K-sample test and sample size calculation for comparing slopes in data with repeated measurements. *Biometrical Journal* **2004**, *46* (5), 554–564.
9. Jung, S.H.; Ahn, C. Sample size for a two-group comparison of repeated binary measurements using GEE. *Statistics in Medicine* **2005**, *24* (17), 2583–2596.
10. Davis, C.S. Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials. *Statistics in Medicine* **1991**, *10* (12), 1959–1980.

Generalizability Probability in Clinical Research

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

This entry evaluates the generalizability of clinical results observed from a clinical trial by means of a sensitivity analysis with respect to changes in mean and standard deviation of the primary clinical endpoints of the study.

INTRODUCTION

For marketing approval of a new drug product, the U.S. Food and Drug Administration^[1] (FDA) requires that at least two adequate and well-controlled clinical trials be conducted to provide substantial evidence regarding the effectiveness and safety of the drug product under investigation. The purpose of requiring at least two clinical studies is not only to assure the reproducibility, but also to provide valuable information regarding *generalizability*. Generalizability refers to whether the clinical results can be generalized to other similar patient populations within the same region (e.g., the United States, Europe, or Asia-Pacific countries) or from region to region. When the sponsor of a newly developed or approved drug product is interested in getting the drug product into the marketplace from one region (e.g., where the drug product is developed and approved) to another region, differences in ethnic factors could alter the efficacy and safety of the drug product in the new region. The International Conference on Harmonization (ICH) recommends that a *bridging study* be conducted to generate a limited amount of clinical data in the new region in order to extrapolate the clinical data between the two regions.^[2]

In practice, it is important to determine whether a clinical trial that produced positive clinical results provides substantial evidence to assure generalizability of the clinical results from one region to another. This entry evaluates the generalizability of clinical results observed from a clinical trial by means of a sensitivity analysis with respect to changes in mean and standard deviation of the primary clinical endpoints of the study.

The next section introduces the concept of generalizability probability of a positive clinical result is introduced. Statistical methods for the evaluation of generalizability probability are discussed in “The Frequentist Approach” and “The Bayesian Approach.” Some applications in clinical research are provided in “Applications.”

GENERALIZABILITY

Consider the situation where the target patient population of the second clinical trial is similar but slightly different (e.g., because of the difference in ethnic factors) from the population of the first clinical trial. The difference in population may be reflected by the difference in mean, variance, or both mean and variance of the primary study endpoints. For example, the population means may be different when the patient populations in two trials are of different race. When a pediatric population is used for the first trial and adult patients or elderly patients are considered in the second trial, the population variance is likely to be different.

Shao and Chow^[3] define the reproducibility probability of the first trial as an estimated power of the second trial using the data from the first trial. The reproducibility probability, considered in Shao and Chow,^[3] with the population of the second trial slightly deviated from the population of the first trial is referred to as the generalizability probability. In the next section, we illustrate how to obtain generalizability probabilities in the case where the study design is a two-group parallel design with a common variance. Other cases can be similarly treated. More details can be found in Chow and Shao.^[4]

THE FREQUENTIST APPROACH

Consider the two-group parallel design. Let $\mu_1 - \mu_2$ be the population mean difference and σ^2 be the common population variance of the first trial. Suppose that in the second trial, the population mean difference is changed to $\mu_1 - \mu_2 + \varepsilon$ and the population variance is changed to $C^2\sigma^2$, where $C > 0$. If $|\mu_1 - \mu_2|/\sigma$ is the signal-to-noise ratio for the population difference in the first trial, then the signal-to-noise ratio for the population difference in the second trial is

$$\frac{|\mu_1 - \mu_2 + \varepsilon|}{C\sigma} = \frac{|\Delta(\mu_1 - \mu_2)|}{\sigma}$$

where

$$\Delta = \frac{1 + \varepsilon/(\mu_1 - \mu_2)}{C} \quad (1)$$

is a measure of change in the signal-to-noise ratio for the population difference. For most practical problems, $|\varepsilon| < |\mu_1 - \mu_2|$ and thus $\Delta > 0$. Suppose a total of $n = n_1 + n_2$ patients are randomly assigned to a treatment group and a control group, then, the power for the second trial is

$$1 - T_{n-2}(t_{0.975; n-2} | \Delta\theta) + T_{n-2}(-t_{0.975; n-2} | \Delta\theta)$$

where $T_{n-2}(\cdot | \Delta\theta)$ is the non-central t -distribution with $n - 2$ degrees of freedom and the non-centrality parameter $\Delta\theta$ and

$$\theta = \frac{\mu_1 - \mu_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (2)$$

Note that the power for the second trial is the same as the power for the first trial if and only if Δ in Eq. 1 equals 1. Define

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (3)$$

where $\bar{x}_i$'s and s_i^2 's are the sample means and variances, respectively, based on the data from the first trial. Then, the generalizability probability on the estimated power approach, as described in the entry *Reproducibility Probability in Clinical Research*, can be obtained with T replaced by ΔT , assuming that Δ is known, that is,

$$\hat{P}_\Delta = 1 - T_{n-2}(t_{0.975; n-2} | \Delta T) + T_{n-2}(-t_{0.975; n-2} | \Delta T)$$

The generalizability probability based on the confidence bound approach as outlined in the entry *Reproducibility Probability in Clinical Research*, is given by

$$\hat{P}_\Delta = 1 - T_{n-2}(t_{0.975; n-2} | \Delta\theta) + T_{n-2}(-t_{0.975; n-2} | \Delta\theta)$$

if $\theta > 0$ and $\hat{P}_{\Delta-} = 0$ if $\theta = 0$.

Values of $\hat{P}_\Delta$ and $\hat{P}_{\Delta-}$ can be obtained using Tables 2 and 3 in the entry of *Reproducibility Probability in Clinical Research* once n , Δ , and $T(x)$ are known. In practice, the value of Δ may be unknown. We may either consider a maximum possible value of $|\Delta|$ or a set of Δ values to carry out a sensitivity analysis.

THE BAYESIAN APPROACH

When σ^2 is unknown, consider the commonly used non-informative prior for σ^2 , i.e., $\pi(\sigma^2) = \sigma^{-2}$. Assume that the priors for μ_1, μ_2 , and σ^2 are independent. The posterior density for (δ, u^2) is

$$\pi(\delta | u^2, x) \pi(u^2 | x)$$

where

$$\begin{aligned} \delta &= \frac{\mu_1 - \mu_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \\ u^2 &= \frac{(n - 2)\sigma^2}{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2} \\ \pi(\delta | u^2, x) &= \frac{1}{u} \phi\left(\frac{\delta - T(x)}{u}\right) \end{aligned}$$

ϕ is the density function of the standard normal distribution, T is given by (Eq. 3), and $\pi(u^2 | x) = f(u)$ with

$$f(u) = \left[\Gamma\left(\frac{n-2}{2}\right) \right]^{-1} \left(\frac{n-2}{2}\right)^{(n-2)/2} u^{-n} e^{-(n-2)/(2u^2)}.$$

Because θ in (Eq. 2) is equal to δu , the posterior mean of $p(\theta)$ is given by

$$\hat{P} = \int_0^\infty \left[\int_{-\infty}^\infty p\left(\frac{\delta}{u}\right) \phi\left(\frac{\delta - T(x)}{u}\right) d\delta \right] 2f(u) du \quad (4)$$

which is the so-called reproducibility probability under the Bayesian approach.

For the Bayesian approach with a known Δ given by Eq.1, the generalizability probability is given by

$$\hat{P}_\Delta = \int_0^\infty \left[\int_{-\infty}^\infty p\left(\frac{\Delta\delta}{u}\right) \phi\left(\frac{\delta - T(x)}{u}\right) d\delta \right] 2f(u) du \quad (5)$$

Table 5 in the entry of *Reproducibility Probability in Clinical Research* can be used to obtain the values of $\hat{P}_\Delta$ once n , Δ , and $T(x)$ are known. When Δ is unknown, the method considered above can be applied. Alternatively, we may consider the average of $\hat{P}_\Delta$ over a noninformative prior density $\pi(\Delta)$, that is,

$$\hat{P} = \int \hat{P}_\Delta \pi(\Delta) d\Delta$$

APPLICATIONS

In clinical development, after the investigational drug product has been shown to be effective and safe with respect

Table 1 Generalizability Probabilities and Sample Sizes for Clinical Trials in a New Population

Δ	The estimated power approach			The Bayesian approach		
	New sample size n^*			New sample size n^*		
	$\hat{P}_\Delta$	70% power	80% power	$\hat{P}_\Delta$	70% power	80% power
1.20	0.929	52	66	0.821	64	90
1.10	0.879	62	80	0.792	74	102
1.00	0.814	74	96	0.742	86	118
0.95	0.774	84	106	0.711	98	128
0.90	0.728	92	118	0.680	104	140
0.85	0.680	104	132	0.645	114	154
0.80	0.625	116	150	0.610	128	170
0.75	0.571	132	170	0.562	144	190

to a target patient population (e.g., adults), it is often of interest to study how likely the clinical result is reproducible in a similar but slightly different patient population (e.g., elderly patient population with the same disease under study or a patient population with different ethnic factors). This information is useful in regulatory submission or supplement new drug application (e.g., when generalizing the clinical results from adults to elderly patient population) and regulatory evaluation for bridging studies (e.g., when generalizing clinical results from Caucasian to Asian patient population). Detailed information regarding bridging studies can be found in ICH.^[2]

An Example

A double-blind randomized trial was conducted in patients with schizophrenia for comparing the efficacy of a test drug with a standard therapy. A total of 104 chronic schizophrenic patients participated in this study. Patients were randomly assigned to receive the treatment of the test drug or the standard therapy for at least one year, where the test drug group has 56 patients and the standard therapy group has 48 patients. The primary clinical endpoint of this trial was the total score of Positive and Negative Symptom Scales (PANSS). No significant differences in demographics and baseline characteristics were observed for baseline comparability.

Mean changes from baseline in total PANSS for the test drug and the standard therapy are $\bar{x}_1 = -3.51$ and $\bar{x}_2 = 1.41$, respectively, with $s_1^2 = 76.1$ and $s_2^2 = 74.9$. The difference $\mu_1 - \mu_2$ is estimated by $\bar{x}_1 - \bar{x}_2 = -4.92$ and is considered to be statistically significant with $T = -2.88$, a p -value of 0.004, and a reproducibility probability of 0.814 under the estimated power approach or 0.742 under the Bayesian approach.

The sponsor of this trial would like to evaluate the probability for reproducing the clinical result in a different population where Δ , the change in the signal-to-noise ratio, ranges from 0.75 to 1.2. The generalizability probabilities are given in Table 1 (because $|T| < 4$ the confidence bound

approach is not considered). In this example, $|T|$ is not very large and thus a clinical trial is necessary to evaluate in the new population. The generalizability probability can be used to determine the sample size n^* for such a clinical trial. The results are given in Table 1. For example, if $\Delta = 0.9$ and the desired power (reproducibility probability) is 80%, then $n^* = 118$ under the estimated power approach and 140 under the Bayesian approach; if the desired power (reproducibility probability) is 70%, then $n^* = 92$ under the estimated power approach and 104 under the Bayesian approach. A sample size smaller than that of the original trial is allowed if $\Delta \geq 1$; that is, the new population is less variable.

The sample sizes n^* in Table 1 are obtained as follows. Under the estimated power approach,

$$n^* = \left(\frac{T^*}{\Delta T} \right)^2 \left/ \left(\frac{1}{4n_1} + \frac{1}{4n_2} \right) \right.$$

where T^* is the value obtained from Table 2 of the entry *Reproducibility Probability in Clinical Research* for which the reproducibility probability has the desired level (e.g., 70% or 80%). Under the Bayesian approach, for each given Δ , we first compute the value T_Δ^* , at which the reproducibility probability has the desired level, and then use

$$n^* = \left(\frac{T_\Delta^*}{T} \right)^2 \left/ \left(\frac{1}{4n_1} + \frac{1}{4n_2} \right) \right.$$

CONCLUSION

In clinical research, it is important to determine whether the observed clinical results can be generalized from a patient population to another similar but slightly different patient population within the same region or from region to region. The assessment of generalizability probability using the frequentist approach or Bayesian approach is

helpful. The sensitivity analysis of generalizability probability is useful for determining whether a bridging clinical study is necessary from one region to another for regulatory review/approval of the drug product under investigation.

REFERENCES

1. FDA. *Guideline for Format and Content of the Clinical and Statistical Sections of New Drug Applications*; Center for Drug Evaluation and Research. The United States Food and Drug Administration: Rockville, MD, 1988.
2. ICH. *Ethnic Factors in the Acceptability of Foreign Clinical Data*; Tripartite International Conference on Harmonization Guideline, 1998, E5.
3. Shao, J.; Chow, S.C. Reproducibility probability in clinical trials. *Stat. Med.* **2002**, *21*, 1727–1742.
4. Chow, S.C.; Shao, J. *Statistics in Drug Research—Methodologies and Recent Development*; Marcel Dekker, Inc.: New York, 2002.

Generalized Estimating Equation

Myunghee Cho Paik

Mailman School of Public Health, Columbia University, New York City, New York, U.S.A.

Hye-Seung Lee

Pediatrics Epidemiology Center, Department of Pediatrics, University of South Florida, Tampa, Florida, U.S.A.

Abstract

The Generalized Estimating Equation method is an inferential procedure concerning the marginal mean of a multivariate outcome through regression models. This entry presents various test statistics to discuss the application of the GEE.

INTRODUCTION

The Generalized Estimating Equation (GEE,^[1,2]) method is an inferential procedure concerning the marginal mean of a multivariate outcome through regression models. The GEE method is based on assumptions regarding the first two moments, rather than the full distribution of an outcome vector. Since its introduction, GEE has been very popular, receiving more than 3,000 citations in the literature. Before the GEE era, regression models for multivariate outcomes, especially for non-normal outcomes, had no unifying approach and had been in limited use. While the classical linear model was extended naturally by generalized linear models (e.g., Ref. [3]) to outcomes arising from an exponential family distributions using various non-linear link functions, an analogous extension for linear models of multivariate normal outcomes to broader types of outcomes was not available mainly due to a lack of a rich class of multivariate distributions. Liang and Zeger^[1] cleverly circumvented this dilemma by borrowing the idea of quasi-likelihood,^[4] and presented GEE as a multivariate version of generalized linear models. Instead of making a full distributional assumption, GEE makes assumptions only on the mean and variance of an outcome as in quasi-likelihood models. In this sense, GEE can be viewed as a multivariate extension of the quasi-likelihood method. A technical distinction between the quasi-likelihood model and GEE is that the quasi-likelihood model does not allow nuisance parameters, except for the constant scale factor for the variance, whereas GEE does.

be zero is called the *estimating function*. A simple example of an estimating equation for the mean μ of random variable Y with a sample of n observations is:

$$\sum_{i=1}^n (Y_i - \mu) = 0.$$

This estimating equation gives the sample mean as the solution. Another well-known estimating equation is the normal equation that gives the Ordinary Least Squares Estimate as the solution. The score equation that yields maximum likelihood estimate as a solution is another example of an estimating equation. All the estimating functions mentioned above share the common characteristic of having zero expectation. For example,

$$E \left\{ \sum_{i=1}^n (Y_i - \mu) \right\} = n\mu - n\mu = 0.$$

Also, all of the estimators mentioned above are consistent. We show later that having zero expectation is key to proving consistency of the resulting estimator. We say that an estimating equation is *unbiased* if the corresponding estimating function has zero expectation. The Generalized Estimating Equation (GEE) is an unbiased *estimating equation* which has the form of a *generalized*, or modified, likelihood equation of a GLM.

Generalized Estimating Equations

Let $Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, \dots, Y_{in_i})^T$ ($i = 1, 2, \dots, K$) be a vector of outcomes for subject i , and $X_i = (X_{i1}^T, X_{i2}^T, \dots, X_{in_i}^T)$ be a $p \times n_i$ covariate matrix where X_{ij} is a $1 \times p$ vector of covariates. Conditioning on X_i , the mean of Y is denoted by $E(Y_i | X_i) = \mu_i$, where $\mu_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{in_i})^T$. Y_i could represent measurements taken at n_i different occasions, or

NOTATION AND DEFINITION

Estimating Equations

An equation which yields an estimator as a solution is called an *estimating equation*, and the function equated to

time points, for subject i . While elements in Y_i may be correlated, $Y_1, Y_2, \dots, Y_K$ are independent vectors. The mean of Y_{ij} is

$$E(Y_{ij} | X_{ij}) = \mu_{ij} = g^{-1}(X_{ij}\beta),$$

where $g(\cdot)$ is a link function which relates the linear predictor $X_{ij}\beta$ to the mean μ_{ij} , and β is an unknown parameter of interest. For example, for binary outcomes using the logit link function,

$$\log \frac{\mu_{ij}}{1 - \mu_{ij}} = X_{ij}\beta,$$

the link function is $g(a) = \log \frac{a}{1-a}$. The parameter of interest β has interpretation of log odds ratio. We denote the variance of Y_{ij} by $v_{ij}(\mu_{ij})/\phi$, where ϕ is a scalar dispersion parameter. For example, if Y_{ij} is binary, $E(Y_{ij}) = \mu_{ij}$, $v_{ij}(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$, and $\phi = 1$. If Y_{ij} is distributed univariate normal with mean μ_{ij} and variance (σ^2), then $v_{ij}(\mu_{ij}) = 1$, and $\phi = 1/\sigma^2$.

We assume the variance-covariance matrix of a random vector Y_i has the form:

$$V_i = A_i(\mu_i)^{1/2} R_i(\alpha) A_i(\mu_i)^{1/2} / \phi,$$

where A_i is a $n_i \times n_i$ diagonal matrix with the j^{th} diagonal element $v_{ij}(\mu_{ij})$, and $R_i(\alpha)$ is a correlation matrix. The GEE is defined as

$$U(\beta, \alpha) = \sum_{i=1}^K \frac{\partial \mu_i^T}{\partial \beta} V_i^{-1} (Y_i - \mu_i) = 0, \quad (1)$$

If α is known, the estimate of β , say $\hat{\beta}$, can be obtained by solving Eq. (1). The estimating function $U(\beta, \alpha)$ has a comparable form to the score function of a Generalized Linear Model (GLM). In fact, if $R_i = I_i$, where I_i is $n_i \times n_i$ identity matrix, the estimating function is identical to the score function of a GLM. However, it should be emphasized that specifying $R_i = I_i$ is not equivalent to assuming independence among $(Y_{i1}, Y_{i2}, Y_{i3}, \dots, Y_{in_i})$. This point will be revisited in the Working Correlation Section. The fundamental difference is that the estimating function $U(\beta, \alpha)$ is constructed, and is not derived from any target function to be maximized, whereas the score function is the derivative of the log-likelihood.

ASYMPTOTIC PROPERTIES

For any estimating equation to yield a consistent estimate it should be constructed to have zero expectation. This property is key to proving consistency and asymptotic normality of $\hat{\beta}$. To understand the proof heuristically, first consider the case in which α is known. Observe that

$$U(\hat{\beta}, \alpha) = 0.$$

Then, we can write

$$U(\hat{\beta}, \alpha) = 0 = U(\beta, \alpha) + \frac{\partial U(\beta, \alpha)}{\partial \beta^T} \bigg|_{\beta=\beta^*} (\hat{\beta} - \beta_0),$$

where $|\hat{\beta} - \beta^*| < |\hat{\beta} - \beta_0|$. This leads to

$$(\hat{\beta} - \beta_0) = \frac{\partial U(\beta, \alpha)}{\partial \beta^T} \bigg|_{\beta=\beta^*}^{-1} U(\beta_0, \alpha).$$

Denoting $E\left(\frac{\partial U(\beta, \alpha)}{\partial \beta^T}\right)$ by $W_0(\beta, \alpha)$, we have

$$W_0(\beta, \alpha) = \sum_{i=1}^K \frac{\partial \mu_i^T}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta^T}.$$

It can be shown that $\frac{\partial U(\beta, \alpha)}{\partial \beta^T} \bigg|_{\beta=\beta^*}$ converges to $W_0(\beta_0, \alpha)$. Then we have

$$(\hat{\beta} - \beta_0) \approx W_0(\beta_0, \alpha)^{-1} \{U(\beta_0, \alpha)\}. \quad (2)$$

Since $U(\beta_0, \alpha)$ is a sum of random vectors with mean zero and finite variance, by the central limit theorem, $U(\beta_0, \alpha)$ is distributed as multivariate normal. Then, Eq. (2) shows that $(\hat{\beta} - \beta_0)$ is a linear combination of random vectors from a multivariate normal distribution. Therefore $(\hat{\beta} - \beta_0)$ itself is distributed as multivariate normal, and we may write

$$\sqrt{K}(\hat{\beta} - \beta_0) \sim N_p(0, KV_{\hat{\beta}}),$$

where

$$V_{\hat{\beta}} = W_0(\beta_0, \alpha)^{-1} \text{Var}\{U(\beta_0, \alpha)\} W_0(\beta_0, \alpha)^{-1}$$

WORKING CORRELATION

Before we further discuss the variance of $U(\beta_0, \alpha)$, we need to revisit the correlation $R_i(\alpha)$. Zeger and Liang called it the “working correlation” to reflect the fact that R does not have to be the true correlation. If $R_i(\alpha)$ is not the true correlation but a specified “working” correlation, V_i is not the true variance of Y_i but only a specified variance, i.e. $E(Y_i - \mu_i)(Y_i - \mu_i)^T \neq V_i$. Let the variance of $U(\beta_0, \alpha)$ be $W_1(\beta_0, \alpha)$. Then

$$\begin{aligned} W_1(\beta_0, \alpha) &= \text{Var} \left(\sum_{i=1}^K \frac{\partial \mu_i^T}{\partial \beta} V_i^{-1} (Y_i - \mu_i) \right) \\ &= E \left(\sum_{i=1}^K \frac{\partial \mu_i^T}{\partial \beta} V_i^{-1} (Y_i - \mu_i) (Y_i - \mu_i)^T V_i^{-1} \frac{\partial \mu_i}{\partial \beta^T} \right). \end{aligned}$$

If $R(\alpha)$ is correctly specified, V_i is the correctly specified variance, i.e., $E(Y_i - \mu_i)(Y_i - \mu_i)^T = V_i$. Then the right-hand side becomes

$$\sum_{i=1}^K \frac{\partial \mu_i^T}{\partial \beta} V_i^{-1} E(Y_i - \mu_i)(Y_i - \mu_i)^T V_i^{-1} \frac{\partial \mu_i}{\partial \beta^T} = \sum_{i=1}^K \frac{\partial \mu_i^T}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta^T},$$

resulting in $W_i = W_0$. Even when R is misspecified, the variance $U(\beta_0, \alpha), W_1(\beta_0, \alpha)$, can be consistently estimated empirically by $\hat{W}_1$, that is

$$\sum_{i=1}^K \frac{\partial \mu_i^T}{\partial \beta} V_i^{-1} (Y_i - \mu_i)(Y_i - \mu_i)^T V_i^{-1} \frac{\partial \mu_i}{\partial \beta^T},$$

evaluated at $\beta = \hat{\beta}$. Then the variance of $\hat{\beta}$ is of ‘sandwich’ form, $W_0^{-1} W_1 W_0^{-1}$, and can be consistently estimated by

$$\hat{V}_{\hat{\beta}} = \hat{W}_0^{-1} \hat{W}_1 \hat{W}_0^{-1}.$$

Even when the correlation structure R is misspecified, this sandwich variance estimate is consistent for $V_{\hat{\beta}}$, and thus enables us to draw correct inferences. Misspecified correlation structure does not affect the consistency of $\hat{\beta}$ and $\hat{V}_{\hat{\beta}}$.

Since we can draw the correct inference even when we misspecify the working correlation structure R_i , we can intentionally choose $R_i = I_i$, although we disbelieve that Y_{ij} ’s are independent. In the case of a binary outcome with logit link, solving a GEE with $R_i = I_i$ is equivalent to fitting a logistic regression model for correlated Y_{ij} ’s as if all Y_{ij} ’s are independent. The GEE estimate in this case maximizes a product of Bernoulli distributions, which is not the true likelihood function, because Y_{ij} ’s are correlated. The above results show that the ordinary logistic regression parameter estimate is consistent even when Y_{ij} ’s are correlated, but that the correct variance estimate is $\hat{W}_0^{-1} \hat{W}_1 \hat{W}_0^{-1}$, not the usual variance estimate ($\hat{W}_0^{-1}$) for independent binary outcomes.

Alternatively the sandwich variance can be estimated by the Jackknife method. Let $\hat{\beta}_{-i}$ be the GEE estimate obtained after deleting the i^{th} independent unit or cluster. The Jackknife variance V_J ,

$$V_J = \sum_{i=1}^K (\hat{\beta}_{-i} - \hat{\beta})(\hat{\beta}_{-i} - \hat{\beta})^T,$$

can consistently estimate the sandwich variance $W_0^{-1} W_1 W_0^{-1}$. Use of the Jackknife variance to estimate a sandwich variance is discussed in Ziegler,^[5] and Lipsitz, Dear and Zhao.^[6]

Efficiency

If the misspecified working correlation yields consistent estimates, what is the advantage of correctly specifying the correlation structure? The answer is efficiency. Let the correct correlation matrix be R_i^* and corresponding W_0 be W_0^* . The variance of $\hat{\beta}$ is smallest when R is correctly

specified, and has the form W_0^{*-1} , whereas the variance of $\hat{\beta}$ is $W_0^{-1} W_1 W_0^{-1}$ when $R_i \neq R_i^*$. If the dimension of β is one, the asymptotic relative efficiency (ARE) of the estimator with misspecified correlation structure is

$$\text{ARE} = W_0^{-1} W_1 W_0^{-1} W_0^*.$$

By applying the maximization and minimization theorem of quadratic forms, we can see that the ARE of a linear combination of $\hat{\beta}$ lies between the largest and smallest eigenvalues of $W_0^{-1} W_1 W_0^{-1} W_0^*$. Rotnitzky and Jewell^[7] provide the upper and lower bounds for efficiency under some special correlation structures. Efficiency of GEE when $R(\alpha)$ is misspecified has been examined analytically^[8] and by simulation.^[9,10]

Estimation of Correlation α

When $R(\alpha)$ is a function of unknown parameter α , we need to estimate α . Note that α is assumed common to all i . The correlation matrix could be different depending on the unit, but should be completely determined by α , a vector of a relatively small dimension, for the asymptotic results of $\hat{\beta}$ to hold.

Estimators of α sometimes involve another unknown parameter ϕ . Then unknown ϕ can be replaced by a consistent estimate of ϕ , which can be obtained by

$$\hat{\phi}^{-1} = \sum_{i=1}^K \sum_{j=1}^{n_i} \frac{(Y_{ij} - \mu_{ij})}{\sqrt{v(\mu_{ij})}} \bigg/ \sum_{i=1}^K n_i.$$

Liang and Zeger^[1] have provided several examples of correlation structures. One simple example is the intraclass correlation matrix.

$$R_i(\alpha) = \begin{pmatrix} 1, \alpha, \alpha, \dots, \alpha \\ \alpha, 1, \alpha, \dots, \alpha \\ \alpha, \alpha, 1, \dots, \alpha \\ \dots, \dots, \dots, \dots, \dots \\ \alpha, \alpha, \alpha, \dots, 1 \end{pmatrix}.$$

In this case, the estimate of α can be obtained as

$$\hat{\alpha} = \left\{ \sum_{i=1}^K \sum_{j>k} \frac{(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})}{\sqrt{v(\mu_{ij})} \sqrt{v(\mu_{ik})}} \right\} \hat{\phi} \bigg/ \sum_{i=1}^K n_i(n_i - 1)/2.$$

Another example is auto-regressive of order 1 (AR-1), where the $(j, k)^{\text{th}}$ element of matrix $R_i(\alpha)$ is $\alpha^{|j-k|}$:

$$R_i(\alpha) = \begin{pmatrix} 1, \alpha, \alpha^2, \dots, \alpha^{n_i-1} \\ \alpha, 1, \alpha, \dots, \alpha^{n_i-2} \\ \alpha^2, \alpha, 1, \dots, \alpha^{n_i-3} \\ \dots, \dots, \dots, \dots, \dots \\ \alpha^{n_i-1}, \alpha^{n_i-2}, \alpha^{n_i-3}, \dots, 1 \end{pmatrix}.$$

In this case, α can be estimated as the regression coefficient in a regression model with zero intercept, $(Y_{ij} - \mu_{ij})/\sqrt{v(\mu_{ij})}$ as the dependent variable, and $(Y_{ij-1} - \mu_{ij-1})/\sqrt{v(\mu_{ij-1})}$ as the independent variable.

Muñoz et al.^[11] proposed a flexible family of correlation structures in which the $(j, k)^{\text{th}}$ element of matrix $R_i(\alpha)$ is $\alpha^{|j-k|^\delta}$. With an additional parameter δ , this family can express various types of correlation structures including AR-1 ($\delta = 1$) and intraclass correlation ($\delta = 0$).

When the number of sub-units is the same for all units ($n_i = n$ for all i), we can assume that $R_i(\alpha) = R(\alpha)$, and can estimate all $n(n-1)/2$ unknown parameters without specifying a particular structure:

$$\hat{R}(\alpha) = K^{-1} \sum_{i=1}^K \hat{A}_i^{-\frac{1}{2}} (Y_i - \hat{\mu}_i) (Y_i - \hat{\mu}_i)^T \hat{A}_i^{-\frac{1}{2}} / \hat{\phi}.$$

Note that the estimation of α depends on β , and in turn, the estimation of β depends on α . Each step should be alternated until both parameters converge. It turns out that estimating unknown α does not affect inference on β given some assumptions.^[11] A crucial assumption is that the estimate of α is $\sqrt{K}$ -consistent. To see why heuristically, first we write the estimating function as follows:

$$U(\beta, \hat{\alpha}(\beta)) = 0.$$

Taylor's expansion gives

$$U(\beta, \hat{\alpha}(\beta)) \approx U(\beta, \alpha_0(\beta)) + (\hat{\alpha} - \alpha_0)^T \frac{\partial U}{\partial \alpha}.$$

Denoting the j^{th} element of α by α_j , we have

$$\begin{aligned} \frac{\partial U}{\partial \alpha_j} &= \sum_{i=1}^K \frac{\partial \mu_i^T}{\partial \beta} \frac{\partial V_i^{-1}}{\partial \alpha_j} (Y_i - \mu_i) \\ &= \sum_{i=1}^K \frac{\partial \mu_i^T}{\partial \beta} V_i^{-1} \frac{\partial V_i^{-1}}{\partial \alpha_j} V_i^{-1} (Y_i - \mu_i) = O_p(K^{\frac{1}{2}}), \end{aligned}$$

because it is a sum of random vector with mean zero. With the condition $\sqrt{K}(\hat{\alpha} - \alpha_0) = O_p(1)$, we find

$$(\hat{\alpha} - \alpha_0)^T \frac{\partial U}{\partial \alpha} = O_p(1),$$

which is of smaller order than $U(\beta, \alpha_0(\beta))$. Since $U(\beta, \hat{\alpha}(\beta))$ is asymptotically equivalent to $U(\beta, \alpha_0(\beta))$, the inference about β can be drawn by ignoring the fact that α is estimated. Uncertainty added by $\hat{\alpha}$ replacing α is of small order and does not affect the asymptotic results of $\hat{\beta}$.

TEST STATISTICS

In likelihood theory, we can draw inference using three different, but related test statistics: Wald tests, Score tests,

and Likelihood ratio tests. Wald tests are based on the asymptotic normality of the maximum likelihood estimator. Score tests are based on the asymptotic normality of the score function evaluated at the null value. Likelihood ratio tests are based on the asymptotic chi square distribution of the difference of the log likelihoods evaluated for two nested models. Since the GEE is not derived from a likelihood function, the aforementioned three test statistics do not exist. However, we can construct similar test statistics. Consider a partition, $\beta^T = (\beta_1^T, \beta_2^T)$, where β_1 is the parameter of interest of length q and β_2 is a nuisance parameter with length $p - q$. We are interested in testing hypothesis $H_0: \beta_1 = \beta_{10}$.

Wald-Type Statistics

The Wald-type statistics can be formed by

$$T_w = (\hat{\beta}_1 - \beta_{10})^T \text{var}(\hat{\beta}_1)^{-1} (\hat{\beta}_1 - \beta_{10}),$$

where $\hat{\beta}_1$ is a $q \times 1$ sub-vector of the GEE estimate $\hat{\beta}$, and $\text{var}(\hat{\beta}_1)$ is a corresponding $q \times q$ sub-matrix of the sandwich variance $\hat{W}_0^{-1} \hat{W}_1 \hat{W}_0^{-1}$. The test statistic is asymptotically chi square distributed with q degrees of freedom.

Score-Type Statistics

For correlated binary outcomes, Lepkopolou, Moore and Ryan^[12] examined the score-type statistics for inference of GEE under the independent working correlation structure, i.e., $R = I$. The score-type statistics can be considered in the same way for general structure of $R(\alpha)$ and for other types of outcomes. Consider the partition of estimating function $U(\beta, \alpha) = \begin{pmatrix} U_1(\beta_1, \beta_2, \alpha) \\ U_2(\beta_1, \beta_2, \alpha) \end{pmatrix}$, where $U_1(\beta_{10}, \tilde{\beta}_2, \alpha)$ is a q -variate sub-vector of estimating function $U(\beta, \alpha)$ corresponding to β_1 . The Score-type statistics also can be formed by

$$T_s = U_1(\beta_{10}, \tilde{\beta}_2, \alpha)^T \text{var}\{U_1(\beta_{10}, \tilde{\beta}_2, \alpha)\}^{-1} U_1(\beta_{10}, \tilde{\beta}_2, \alpha),$$

where $\tilde{\beta}_2$ is the solution of $U(\beta_{10}, \beta_2, \alpha) = 0$, i.e., the GEE estimate of β_2 under the restriction $\beta_1 = \beta_{10}$. Consider a similar partition for matrices W_0 and W_1 :

$$\begin{aligned} W_0 &= E \begin{pmatrix} \frac{\partial U_1}{\partial \beta_1^T}, \frac{\partial U_1}{\partial \beta_2^T} \\ \frac{\partial U_2}{\partial \beta_1^T}, \frac{\partial U_2}{\partial \beta_2^T} \end{pmatrix} = \begin{pmatrix} D_{11}, D_{12} \\ D_{21}, D_{22} \end{pmatrix}, \text{ and} \\ W_1 &= \begin{pmatrix} M_{11}, M_{12} \\ M_{21}, M_{22} \end{pmatrix}. \end{aligned}$$

It can be easily verified that $D_{21} = D_{12}^T$ and $M_{21} = M_{12}^T$. To obtain the variance of $U(\beta_{10}, \tilde{\beta}_2, \alpha)$ we can consider the following expression:

$$U(\beta_{10}, \tilde{\beta}_2, \alpha) \approx U_1(\beta_{10}, \beta_{20}, \alpha) - D_{12} D_{22}^{-1} U_2(\beta_1, \beta_2, \alpha).$$

Then

$$\text{var}\{U(\beta_{10}, \tilde{\beta}_2, \alpha)\} = M_{11} + D_{12} D_{22}^{-1} M_{22} D_{22}^{-1} D_{21} - 2 D_{12} D_{22}^{-1} M_{21}.$$

The test statistic T_s is distributed as chi square with q degrees of freedom. Replacing α with $\tilde{\alpha}$ (the estimate of α under the null) does not affect the asymptotic distribution of the test statistic. One disadvantage of using T_s is that the computation of the score statistic requires a specialized program.

Likelihood Ratio-Type

When $R = I$, the GEE estimating function is the same as the score function of GLM, and the GEE estimate maximizes the function that is the likelihood function if the Y_{ij} 's are independent. Since the Y_{ij} 's are *not* independent, the maximized function by the GEE estimator is not the likelihood function. For example, when Y_{ij} is binary, the maximized function is

$$\sum_{i=1}^K \left(\sum_{j=1}^{n_i} Y_{ij} \log \left\{ \mu_{ij} / (1 - \mu_{ij}) \right\} - (1 - Y_{ij}) \log(1 - \mu_{ij}) \right).$$

In general, suppose that marginally Y_{ij} at the j^{th} occasion or time point arises from an exponential family distribution. Let the product of the densities of Y_{ij} 's be $L(\beta; Y)$. Although $L(\beta; Y)$ is not the likelihood function, we can form a likelihood ratio-like statistic for testing $\beta_1 = \beta_{10}$ in the presence of the nuisance parameter β_2 , and consider its asymptotic properties when the Y_{ij} 's are in fact correlated. Let the likelihood ratio-like statistic be T_L and

$$T_L = 2 \left\{ L(\hat{\beta}_1, \hat{\beta}_2) - L(\beta_{10}, \tilde{\beta}_2) \right\},$$

where $\tilde{\beta}_2$ is the maximizer of L under the restriction $\beta_1 = \beta_{10}$.

Rotnitzky and Jewell^[7] considered the distribution of T_L , when Y_{ij} 's are correlated. They showed that T_L can be expressed as a weighted sum of independent chi square variables with one degree of freedom. Denoting $\chi_1, \dots, \chi_q$ as q independent chi square random variables with one degree of freedom, they showed

$$T_L \approx \sum_{j=1}^q d_j \chi_j^2,$$

where $d_1 \geq d_2 \geq \dots \geq d_q$ are eigenvalues of $P \approx P_0^{-1} P_1$, where

$$P_0 = D_{11} - D_{12} D_{22}^{-1} D_{21}, \quad \text{and} \\ P_1 = M_{11} + D_{12} D_{22}^{-1} M_{22} D_{22}^{-1} D_{21} - 2 D_{12} D_{22}^{-1} M_{21},$$

where all sub-matrices D and M in P_0 and P_1 are evaluated at $\alpha = 0$. When there are no nuisance parameters, $P_0 = W_0$ and $P_1 = W_1$.

Rotnitzky and Jewell^[7] also derived the distributions of the "aïve" Wald and Score tests where statistics are formed as if all observations were independent. Those statistics are found to be asymptotically equivalent to T_L .

USING GEE WITH CAUTION

The GEE draws inference on the mean of Y given X . Although it is conditional on X , in the literature it is referred to as the marginal mean in the sense that it is the mean of an outcome at a single time point without conditioning on the other elements of the outcome vector. The GEE is the method to use when the interest is on the marginal mean. The interpretation of the regression parameters is also in terms of the marginal mean. Specifically, the regression coefficient β in GEE features *population average effects* that compare the mean outcome for a group of subjects in the population who have a specific value X , say x_1 versus another group who have $X = x_2$. On the other hand, one may be interested in *subject specific effects* that compare the mean outcome when a subject has a specific value $X = x_1$ vs. when the same subject has $X = x_2$.^[13] The GEE does not address subject-specific effects. Such inference can be drawn using mixed effects models. Several ways of extending GLM to mixed effects models are suggested.^[13–15]

Given that the main interest is in population average effects, one should note that the asymptotic properties of GEE depend on large K and fixed n_i . For studies with large n_i or small K , one may consider alternative methods. Even when K is large with fixed n_i , problems associated with specification of the working correlation structure may arise. For some working correlation structures such as m -dependence, there is a range of α that yields non-positive definite matrices. Users should make sure that the estimated correlation matrix is positive definite. A more fundamental problem is the lack of definition of the working correlation coefficient α .^[16] One of the assumptions used to derive the asymptotic properties of $\hat{\beta}$ is $\sqrt{K}$ -consistency of $\hat{\alpha}$. The problem is that α is not a true correlation, but merely a working correlation, and sometimes it is not clear what $\hat{\alpha}$ is estimating. Crowder^[16] examined this issue and cautioned that in some cases, α is subject to uncertainty of definition, leading to a breakdown of the asymptotic properties of $\hat{\beta}$. In the case of independent working correlation models, this problem can be avoided. Especially in the case of

canonical link function (e.g., logit link for binary; log link for Poisson; identity link for normal), the GEE estimates are guaranteed to exist. It is a good practice to fit an independent working correlation model first before fitting more complicated working correlation structures.

EXTENSIONS OF GEE

Various extensions of GEE have been proposed by many authors. Ziegler et al.^[17] gives an extensive annotated bibliography on GEE.

Joint Modeling

Prentice^[18] considered modeling not only the mean as a function of covariates, but also correlation as a function of covariates. The estimating equation for the correlation is stacked with the original GEE for β to form a joint estimating equation. Liang, Zeger and Qaqish^[19] modeled jointly the mean and the odds ratio between the repeated measurements, instead of correlation, and called it GEE2. Paik^[20] considered joint modeling of mean and scale parameter, ϕ . A joint modeling of mean, scale, and correlation parameter by separate link functions was proposed by Yan and Fine^[21] and later improved by Lee, Paik and Lee^[22] by employing the covariance structure of multivariate normal distribution. For multi-level multivariate data, for example, data with multiple measurements per each family member, correlation structure becomes complicated. In this case Lee, Paik and Lee^[23] proposed joint modeling of mean, scale, and canonical correlation. Instead of modeling pairwise correlation, they proposed to model canonical correlation which is maximal correlation between linear combinations of multiple measurements for two family members. Modelling aspects allow the canonical correlation to be expressed as a function of pair specific covariates.

A class of multivariate exponential family called partial exponential family is proposed by Zhao, Prentice, and Self,^[24] who have shown that the score function of the partial exponential family has the same form as GEE if the GEE variance matrix is correctly specified. However, partial exponential family distributions cannot be expressed in terms of (β, α, ϕ) , and iterative computation is required to translate (β, α, ϕ) , to the parameters in a partial exponential family. Nonetheless, mere existence of such family of distributions provides a basis of data generating mechanisms useful for research in GEE. Prentice and Zhao^[25] considered joint modeling using multivariate binary distribution that belongs to the partial exponential family.

Nonparametric GEE

Lin and Carroll^[26] developed nonparametric GEE to handle clustered data employing polynomial kernel smoothing.

Lin and Carroll^[27,28] extended their approach for the partial linear model, where the linear predictor is parametric for some covariates and non-parametric for others. They noted that their method requires either independent working correlation or an artificial under smoothing to guarantee the $\sqrt{K}$ -consistency of the slope. Although the original GEE is efficient with correctly specified within-subject correlation, their kernel GEE obtained higher efficiency with independent working correlation than with true correlation. Wang^[29] and Wang, Carroll, and Lin^[30] modified the Lin and Carroll's method by defining neighbors accounting for the correlation structure, and improved the efficiency. Chen and Jin^[31] proposed an alternative method having $\sqrt{K}$ -consistency for the regression estimator using "local" working correlation matrix which performs better than working independent correlation structure.

Missing Data

Another important issue is missing data. Note that GEE naturally handles varying numbers of sub-units in the independent unit. For example, varying numbers of repeated measurements across subjects possibly resulting from dropout is not an obstacle to fitting GEE models. However, it is valid only when the missingness probability does not depend on the outcome Y (called missing completely at random). If the probability of missingness depends on observed or on unobserved Y , the GEE estimate is not consistent. To fix this problem, Robins, Rotnitzky, and Zhao^[32] have proposed a weighted GEE in which each observation is weighted by the inverse of probability of observation. The authors also proposed a modified weighted GEE to improve the efficiency of the initially proposed weighted GEE. Paik^[33] has proposed imputed GEE where missing outcome is replaced with the estimates of the conditional expectation given observed data. Fitzmaurice et al.^[34] have examined the bias when the missingness mechanism is informative. Xie and Paik^[35,36] have proposed an imputation approach when covariates are missing. Lipsitz et al.^[37] have considered imputation approach for multivariate binary data using multivariate Gaussian model.

GENETIC APPLICATION OF GEE

One of the most popular applications of GEE has been in the field of quantitative genetics. In genetics, it is often of interest to investigate if sharing genes through familial relationship increases the risk of developing a disease. When the trait of interest was continuously measured, the interest would be to see if sharing genes through familial relationship produces a similar level for two quantitative traits from a relative pair. A multivariate outcome from family members is often to be analyzed, so the GEE has been used in the analysis along with likelihood-based modeling approach.

To examine whether or not there is any genetic effect on a trait, the correlation of the trait among family members can be compared with that among genetically unrelated individuals, after adjusting for other confounders. The joint modeling of mean, scale, and correlation parameters has been found to be useful in attacking this problem. Yan and Fine^[21] have illustrated their methods in the application of familial correlation analysis for a single trait, and Lee, Paik, and Lee,^[23] have done the same for multiple traits.

One of the most popular approaches is to implement a subject specific random genetic and environmental effect model. In this case, heritability, which is the ratio of the variance of random genetic effects given the expected proportions of sharing identity by descent (IBD) genes to the total variance of a trait, measures the genetic effect on the trait. Based on the random effect model, various approaches have been developed to locate genes by inferring a function of the variance of random genetic effects given the *locus specific* estimates of proportions of sharing IBD genes. One approach is to directly infer the variance component. Many authors applied joint modeling for GEE by Prentice and Zhao^[25] to estimate the locus specific genetic variance and compared the results with maximum likelihood estimates.^[38–40] Another approach is the Haseman-Elston regression^[41] where outcome is the squared differences in the trait values of relative pairs and independent variable is the estimates of the proportions of sharing IBD genes. Variance estimation and the Haseman-Elston regression approaches are shown to be framed in the context of GEE by Chen, Broman, and Liang.^[42,43]

COMPUTATION

The GEE regression estimates under the independent working correlation structure can be obtained by using any software that fits GLM. However, the variance reported by this software would be the diagonal elements of $\hat{W}_0^{-1}$. To compute the correct variance, we need to calculate $\hat{W}_1$ as well. Some software packages have routines to fit GEE and report the sandwich variance estimates. SAS (PROC GENMOD), Stata (XTGEE), and SUDAAN (PROC MULTILOG, PROC LOGISTIC and PROC REGRESS) are among the commercially available programs. SUDAAN also supports a jackknife variance estimator using independent working correlation. For review and comparison of those software, see Horton and Lipsitz.^[44] GEE is also available in S-PLUS(GEE or YAGS) or in R(GEE or GEEPACK).

REFERENCES

- Liang, K.-Y.; Zeger, S.L. Longitudinal data analysis using generalized linear models. *Biometrika* **1986**, *73*, 13–22.
- Zeger, S.L.; Liang, K.-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **1986**, *42*, 121–130.
- McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd Ed.; Chapman and Hall: London, 1989.
- Wedderburn, R.W.M. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika* **1974**, *61*, 439–447.
- Ziegler, A. Practical considerations of the jackknife estimator of variance for generalized estimating equations. *Stat. Papers* **1997**, *38*, 363–369.
- Lipsitz, S.R.; Dear, K.B.G.; Zhao, L. Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics* **1994**, *50*, 842–846.
- Rotnitzky, A.; Jewell, N.P. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **1990**, *77*, 485–497.
- Mancini, L.A.; Leroux, B.G. Efficiency of regression estimates for clustered data. *Biometrics* **1996**, *52*, 500–511.
- Paik, M.C. Repeated measurement analysis for nonnormal data in small samples. *Commun. Stat.* **1988**, *B17*, 1155–1171.
- Park, T. A comparison of the Generalized Estimating Equation approach With the maximum likelihood approach for repeated measurements. *Stat. Med.* **1993**, *12*, 1723–1732.
- Muñoz, A.; Carey, V.; Schouten, J.P.; Segal, M.; Rosner, B. A Parametric family of correlation structures for the analysis of longitudinal data. *Biometrics* **1992**, *48*, 733–742.
- Lefkopoulou, M.; Moore, D.; Ryan, L. The analysis of multiple correlated binary outcomes: application to rodent teratology experiments. *J. Am. Stat. Assoc.* **1989**, *84*, 810–815.
- Zeger, S.L.; Liang, K.-Y.; Albert, P.S. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **1988**, *44*, 1049–1060.
- Lee, Y.; Nelder, J.A. Hierarchical generalized linear models (Disc: P656–678). *J. R. Stat. Soc. Ser. B* **1996**, *58*, 619–656.
- Breslow, N.E.; Clayton, D.G. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* **1993**, *88*, 9–25.
- Crowder, M. On the use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika* **1995**, *82*, 407–410.
- Ziegler, A.; Kastner, C.; Blettner, M. The generalised estimating equations: an annotated bibliography. *Biom. J.* **1988**, *40*, 115–139.
- Prentice, R.L. Correlated binary regression with covariates specific to each binary observation. *Biometrics* **1988**, *44*, 1033–1048.
- Liang, K.-Y.; Zeger, S.L.; Qaqish, B. Multivariate regression analyses for categorical data. *J. R. Stat. Soc. Ser. B.* **1992**, *54* (1), 3–40.
- Paik, M.C. Parametric variance function estimation for nonnormal repeated measurement data. *Biometrics* **1992**, *48*, 19–30.
- Yan, J.; Fine, J. Estimating equations for association structures. *Stat. Med.* **2004**, *23*, 859–874.
- Lee, H.-S.; Paik, M.C.; Lee, J.H. Genotype adjusted familial correlation analysis using three generalized estimating equations. *Stat. Med.* **2008**, *27* (26), 5471–5483.

23. Lee, H.-S.; Paik, M.C.; Lee, J.H. Estimating a multivariate familial correlation using Joint models for canonical correlations: application to memory score analysis from familial Hispanic Alzheimer's disease study. *Biometrics* **2009**, *65*, 463–469.
24. Zhao, L.P.; Prentice, R.L.; Self, S.G. Multivariate mean parameter estimation by using a partly exponential model. *J. R. Stat. Soc. Ser. B-Methodol.* **1992**, *54* (3), 805–811.
25. Prentice, R.L.; Zhao, L.P. Estimating equations for parameters in mean and covariances of multivariate discrete and continuous response. *Biometrics* **1991**, *47* (3), 825–839.
26. Lin, X.; Carroll, R.J. Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Am. Stat. Assoc.* **2000**, *95*, 520–534.
27. Lin, X.; Carroll, R.J. Semiparametric regression for clustered data using generalized estimating equations. *J. Am. Stat. Assoc.* **2001**, *96*, 1045–1056.
28. Lin, X.; Carroll, R.J. Semiparametric regression for clustered data. *Biometrika* **2001**, *88*, 1179–1185.
29. Wang, N. Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* **2003**, *90*, 43–52.
30. Wang, N.; Carroll, R.J.; Lin, X. Efficient semiparametric marginal estimation for longitudinal/clustered data. *J. Am. Stat. Assoc.* **2005**, *100*, 147–157.
31. Chen, K.; Jin, Z. Partial linear regression models for clustered data. *J. Am. Stat. Assoc.* **2006**, *101*, 195–204.
32. Robins, J.M.; Rotnitzky, A.; Zhao, L.P. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Stat. Assoc.* **1995**, *90*, 106–121.
33. Paik, M.C. The generalized estimating equation approach when data are not missing completely at random. *J. Am. Stat. Assoc.* **1997**, *92*, 1320–1329.
34. Fitzmaurice, G. M.; Molenberghs, G.; Lipsitz, S.R. Regression models for longitudinal binary responses with informative drop-outs. *J. Royal Stat. Soc. Ser. B* **1995**, *57*, 691–704.
35. Xie, F.; Paik, M.C. Generalized estimating equation model for binary outcomes with missing covariates. *Biometrics* **1997**, *53*, 1458–1466.
36. Xie, F.; Paik, M.C. Multiple imputation methods for missing covariates problem in generalized estimating equation. *Biometrics* **1997**, *53*, 1538–1546.
37. Lipsitz, S. R.; Molenberghs, G.; Fitzmaurice, G.M.; Ibrahim, J. GEE with Gaussian estimation of the correlations when data are incomplete. *Biometrics* **2000**, *56*, 528–536.
38. Stram, D.O.; Lee, H.; Thomas, D.C. Use of generalized estimating equations in segregation analysis of continuous outcomes. *Genet. Epidemiol.* **1993**, *10*, 575–579.
39. Bull, S.B.; Chapman, N.H.; Greenwood, C.M.T.; Darlington, G.A. Evaluation of genetic and environmental effects using GEE and APM methods. *Genet. Epidemiol.* **1995**, *12*, 729–734.
40. Amos, C.I.; Zhu, D.K.; Boerwinkle, E. Assessing genetic linkage and association with robust components of variance approaches. *Ann. Hum. Genet.* **1996**, *60*, 143–160.
41. Haseman, J.K.; Elston, R.C. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **1972**, *2*, 3–19.
42. Chen, W.-M.; Broman, K.W.; Liang, K.-Y. Quantitative trait linkage analysis by generalized estimating equations: unification of variance components and Haseman-Elston regression. *Genet. Epidemiol.* **2004**, *26*, 265–272.
43. Chen, W.-M.; Broman, K.W.; Liang, K.-Y. Power and robustness of linkage test for quantitative traits in general pedigree. *Genet. Epidemiol.* **2005**, *28*, 11–23.
44. Horton, N.; Lipsitz, S.R. Review of software to fit generalized estimating equation regression models. *Am. Stat.* **1999**, *53*, 160–169.

Generalized Inference

Chen-Tuo Liao

National Taiwan University, Taipei, Taiwan, R.O. China

Chi-Rong Li

Chung Shan Medical University, Taichung, Taiwan, R.O. China

Abstract

This entry provides a brief description of generalized test variables and generalized pivotal quantities and supplies application examples.

INTRODUCTION

Generalized test variables (GTVs) and generalized pivotal quantities (GPQs) were proposed by Tsui and Weerahandi,^[1] and Weerahandi,^[2] respectively. When the conventional frequenting methods fail to provide satisfactory solutions for hypothesis test or interval estimation about some complex parametric functions, GTVs and GPQs may serve as an easy tractable means of computing generalized p -values and generalized confidence intervals (GCI) for them. According to many simulation studies, the generalized inference often performs satisfactorily in the repeated sampling argument.

Generalized p -values and GCIs have been applied successfully to many biological studies such as occupational exposure limit estimation,^[3,4] population and individual bioequivalence,^[5] accuracy comparison with ROC curves,^[6,7] confidence limits about effect measures,^[8] tolerance intervals for linear models,^[9,10] growth curve estimation,^[11,12] common mean estimation for several normal populations,^[13,14] the means and variances of a bivariate log-normal distribution,^[15] and others. An in-depth review of these procedures and their applications was presented in the two textbooks by Weerahandi.^[16,17]

We introduce the concepts of generalized p -values and GCIs in the next section. We demonstrate the applications with ROC curve and tolerance interval inference in the section “Examples,” and we discuss some possible drawbacks of generalized inference in the section “Limitations.” Finally, a brief summary is given in the conclusion.

GENERALIZED p -VALUE AND GENERALIZED CONFIDENCE INTERVAL

Suppose that Y is a random sample whose distribution depends on a vector of unknown parameters $\zeta = (\theta, \eta)$,

where θ is a parameter of interest, and η is a vector of nuisance parameters. Let y be the observed value of Y . A random quantity $T = T(Y; y, \zeta)$ is said to be a GTV for θ if it satisfies the following three conditions:

- The distribution of T is free of nuisance parameters η .
- The observed value of T , say $t_{obs} = T(y; y, \zeta)$, does not depend on unknown parameters.
- For fixed y and η , $P[T \leq t | \theta]$ is monotonic in θ for all t .

A generalized extreme region for testing null hypothesis $H_0 : \theta \in \Theta_0$ versus alternative hypothesis $H_1 : \theta \in \Theta_1$ is defined as $C = \{Y : T(Y; y, \zeta) \geq t_{obs}\}$ such that small probabilities of C indicate evidence against the null hypothesis, where Θ_0 is some subset of the parameter space and Θ_1 is its complement. Then, a generalized p -value is defined as

$$p = \sup_{\theta \in \Theta_0} P[C | \theta].$$

For the test with a nominal significance level α , one typically rejects null hypothesis if $p < \alpha$. The exact distribution function of T is usually too complicated to derive; thus, the generalized p -value is often computed by Monte Carlo algorithm.

Similarly, a random quantity $R = R(Y; y, \zeta)$ is said to be a GPQ for θ if it satisfies the following two conditions:

- The distribution of R is free of unknown parameters.
- The observed value of R , say $r_{obs} = R(y; y, \zeta)$, does not depend on nuisance parameters η .

A GTV and a GPQ for θ can be easily transformed to be each other, because both of them must fulfill two identical conditions. A GCI for θ is usually obtained from appropriate percentiles of a sampling distribution of R based on Monte-Carlo algorithms.

EXAMPLES

The Area Under the ROC Curve

Suppose that Y and X are the two variables from a diagnostic device for a diseased patient and a non-diseased patient, respectively. Let θ be the area under the ROC curve (AUC), measuring the accuracy of the diagnostic device. Then, it can be verified that $\theta = P(Y > X)$.^[18] Under the normality assumption that $Y \sim N(\mu_Y, \sigma_Y^2)$ and $X \sim N(\mu_X, \sigma_X^2)$,

$$\theta = \Phi \left(\frac{\mu_Y - \mu_X}{\sqrt{\sigma_Y^2 + \sigma_X^2}} \right).$$

We now show how to find a GTV for θ . Suppose $\bar{Y}$ and S_Y^2 are the sample mean and sample variance for the diseased with n samples, and $\bar{X}$ and S_X^2 are those for the non-diseased with m samples. Thus, $Z = [(\bar{Y} - \bar{X}) - (\mu_Y - \mu_X)] / \sqrt{\sigma_Y^2/n + \sigma_X^2/m} \sim N(0,1)$, $U = (n-1)S_Y^2/\sigma_Y^2 \sim \chi_{n-1}^2$, and $V = (m-1)S_X^2/\sigma_X^2 \sim \chi_{m-1}^2$. We have GPQs for σ_Y^2 and σ_X^2 as $R_{\sigma_Y^2} = (n-1)S_Y^2/U$ and $R_{\sigma_X^2} = (m-1)S_X^2/V$, respectively. Also, a GPQ for $\mu_Y - \mu_X$ and is $R_{\mu_Y - \mu_X} = \bar{y} - \bar{x} - Z \sqrt{R_{\sigma_Y^2}/n + R_{\sigma_X^2}/m}$, where $s_Y^2, s_X^2, \bar{y}$ and $\bar{x}$ are the observed values of $S_Y^2, S_X^2, \bar{Y}$ and $\bar{X}$, respectively. Consequently, we obtain a GTV for θ given by

$$T = T(\bar{Y}, \bar{X}, S_Y^2, S_X^2; \bar{y}, \bar{x}, s_Y^2, s_X^2, \theta) = R_\theta - \theta,$$

where

$$R_\theta = \Phi \left(\frac{R_{\mu_Y - \mu_X}}{\sqrt{R_{\sigma_Y^2} + R_{\sigma_X^2}}} \right).$$

It is easy to verify R_θ is a GPQ for θ , and the distribution function of T is a monotonic function in θ , thus T is a GTV for θ . For testing $H_0: \theta \leq \theta_0$ versus $H_1: \theta > \theta_0$, where θ_0 is a pre-specified value, one can generate a sampling distribution of R_θ by mutually independent $R_{\sigma_Y^2}, R_{\sigma_X^2}$ and $R_{\mu_Y - \mu_X}$. The required generalized p -value consequently is estimated by the proportion of the realized values of R_θ that are greater than or equal to θ_0 .

Tolerance Intervals for Balanced Linear Models

Let F denote the cumulative distribution for a random variable W . An interval $[L(Y), U(Y)]$, based on a random sample Y , is said to be a two-sided β -content, γ -confidence tolerance interval, denoted by (β, γ) -tolerance interval, for F if

$$Pr\{[F(U(Y)) - F(L(Y))] \geq \beta\} = \gamma.$$

Thus, one can state with confidence coefficient γ that at least a proportion β of the population modeled by F is contained in the interval $[L, U]$. Similarly, one-sided tolerance limits can be defined.

Suppose that $W \sim N(\theta, T^2)$, where $T^2 = \sum_{i=1}^q h_i \sigma_i^2$. Also, h_i are known constants and each σ_i^2 denotes a linear combination of the variance components in a balanced linear model. Thus, mutually independent observable statistics $G, S_1^2, S_2^2, \dots, S_q^2$ are available such that

- $G \sim N(\theta, \sigma^2)$, where $\sigma^2 = \sum_{i=1}^q c_i \sigma_i^2$ and c_i are known constants.
- $f_i S_i^2 / \sigma_i^2 \sim \chi_{f_i}^2$, for $i = 1, 2, \dots, q$.

It is easy to show that an upper (β, γ) -tolerance limit for W is in fact a γ -confidence upper bound for $\delta = \theta + z_\beta \tau$. Because an exact γ -confidence upper bound for δ is generally unavailable, an upper generalized γ -confidence bound for δ , denoted by $R_{\delta, \gamma}$, is thus a useful surrogate. Let g denote the observed value of G and $s_1^2, s_2^2, \dots, s_q^2$ denote the observed values of $S_1^2, S_2^2, \dots, S_q^2$. We thus have $Z = (G - \theta)/\sigma \sim N(0,1)$ and $U_i = f_i S_i^2 / \sigma_i^2 \sim \chi_{f_i}^2$. Define

$$\begin{aligned} R_\theta &= g - Z \sqrt{\max \left\{ 0, \sum_{i=1}^q \frac{c_i f_i s_i^2}{U_i} \right\}} \\ &= g - \left(\frac{G - \theta}{\sigma} \right) \sqrt{\max \left\{ 0, \sum_{i=1}^q \frac{c_i \sigma_i^2 s_i^2}{S_i^2} \right\}} \end{aligned}$$

and

$$R_T = \sqrt{\max \left\{ 0, \sum_{i=1}^q \frac{h_i f_i s_i^2}{U_i} \right\}} = \sqrt{\max \left\{ 0, \sum_{i=1}^q \frac{h_i \sigma_i^2 s_i^2}{S_i^2} \right\}}.$$

It can easily be verified that R_θ and R_T are GPQs for θ and T respectively. So a GPQ for δ is given by

$$R_\delta = R_\theta + z_\beta R_T.$$

Similarly, the desired $R_{\delta, \gamma}$ for an upper (β, γ) -tolerance limit can be simply obtained from a Monte-Carlo algorithm.

Limitations

Stimulation studies have shown that generalized inference can be successfully applicable to many practical situations. More interestingly, Hannig^[19] provided the connection between generalized inference and fiducial inference. Also, Hannig et al.^[20] proposed that GCIs asymptotically maintain the target coverage level for a large class of parametric problems. These imply that the GCIs are no worse asymptotically than the intervals based on asymptotic normality of the estimators of the parametric functions of interest.

However, the generalized inference still suffers from some possible limitations. First, the choice of a GTV or GPQ for a parametric function can be non-unique. In addition, there is no guarantee that all of them enjoy good performance in the repeated sampling argument.^[19,20] Second, generalized inference is usually made on a limited number of continuous distributions like normal and log-normal distributions. There is no much work has been studied about it for a parametric function of a discrete distribution. Third, it still needs further theoretical investigation on the sufficient conditions that generalized inference can be a workable tool in the frequenting argument, particularly in the small sample situations. For the most cases according to generalized inference, one still needs an intensive simulation study to verify their performance. Certainly, a simulation study cannot cover all parameter spaces.

CONCLUSION

Even though generalized inference has been successfully applicable to a wide range of statistical problems and many problems of practical importance. A useful procedure of finding a GTV or GPQ for a parametric function if it exists can be found in.^[16] However, the user should be careful to make inference based on it.

REFERENCES

1. Tsui, K.W.; Weerahandi, S. Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *J Am. Stat. Assoc.* **1989**, *84*, 602–607.
2. Weerahandi, S. Generalized confidence intervals. *J Am. Stat. Assoc.* **1993**, *88*, 899–905.
3. Krishnamoorthy, K.; Mathew, T. Assessing occupational exposure via the one-way random effects model with balanced data. *J.Agric. Biol. Environ. Stat.* **2002**, *7*, 440–451.
4. Krishnamoorthy, K.; Guo, H.Z. Assessing occupational exposure via the one-way random effects model with unbalanced data. *J. Stat. Plan. Inference* **2005**, *128*, 219–229.
5. McNally, R.J.; Iyer, H.K.; Mathew, T. Tests for individual and population bioequivalence based on generalized p-values. *Stat. Med.* **2003**, *22*, 31–53.
6. Li, C.R.; Liao, C.T.; Liu, J.P. On the exact interval estimation for the difference in paired are as under the ROC curves. *Stat. Med.* **2008**, *27*, 224–242.
7. Li, C.R.; Liao, C.T.; Liu, J.P. A non-inferiority test for diagnostic accuracy based on the paired partial areas under ROC curves. *Stat. Med.* **2008**, *27*, 1762–1776.
8. Zou, G.Y.; Donner, A. Construction of confidence limits about effect measures: a general approach. *Stat. Med.* **2008**, *27*, 1693–1702.
9. Liao, C.T.; Lin, T.Y.; Iyer, H.K. One-and two-sided tolerance intervals for general balanced mixed linear models and unbalanced one-way random models. *Technometrics* **2005**, *47*, 323–335.
10. Lin, T.Y.; Liao, C.T.; Iyer, H.K. Tolerance intervals for unbalanced one-way random effects models with covariates and heterogeneous variances. *J. Agric. Biol. Environ. Stat.* **2008**, *13*, 221–241.
11. Weerahandi, S.; Berger, V.W. Exact inference for growth curves with intraclass correlation structure. *Biometrics* **1999**, *55*, 921–924.
12. Lin, S.H.; Lee, J.C. Exact tests in simple growth curve models and one-way ANOVA with equicorrelation error structure. *J. Multivar. Anal.* **2003**, *84*, 351–368.
13. Krishnamoorthy, K.; Lu, Y. Inferences on the common mean of several normal populations based on the generalized variable method. *Biometrics* **2003**, *59*, 237–247.
14. Lin, S.H.; Lee, J.C. Generalized inferences on the common mean of several normal populations. *J. Stat. Plan. Inference* **2005**, *134*, 568–582.
15. Bebu, I.; Mathew, T. Comparing the means and variances of a bivariate log-normal distribution. *Stat. Med.* **2008**, *27*, 2684–2696.
16. Weerahandi, S. *Exact Statistical Methods for Data Analysis*; Springer: New York, 1995.
17. Weerahandi, S. *Generalized Inferences in Repeated Measures: Exact Methods in MANOVA and Mixed Models*; Wiley: New York, 2004.
18. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36.
19. Hannig, J. On fiducial inference, the good, the bad, and the ugly. Technical Report 2006/3, Colorado State University, Department of Statistics. Available at http://www.stat.colostate.edu/research_reports.html.
20. Hanning, J.; Iyer, H.K.; Patterson, P. Fiducial generalized confidence intervals. *J. Am. Stat. Assoc.* **2006**, *101*, 254–269.

Genetic Linkage and Linkage Disequilibrium Analysis

Kongming Wang

Department of Statistics, Pfizer Inc., New York, New York, U.S.A.

Bernice Porjesz and Henri Begleiter

Department of Psychiatry, HSCB, State University of New York, Downstate Medical Center, Brooklyn, New York, U.S.A.

Kevin Jones

Department of Psychiatry, HSCB, State University of New York, Brooklyn, New York, U.S.A.

Abstract

This entry describes the basic concepts of and commonly used methods in genetic mapping [linkage and linkage disequilibrium (LD) analysis] of quantitative trait loci (QTL).

INTRODUCTION

Genetic mapping of human traits (phenotypes) aims to identify chromosomal regions that contain genes affecting traits of interest and especially genes that affect human susceptibility to particular diseases. There are many single-gene diseases (e.g., Huntington's disease) that are caused by a change of a single gene. There are also many polygenic diseases resulting from the combined action of alleles of more than one gene (e.g., heart disease and diabetes). A great deal of attention has been focused on identifying disease genes in order that they may be used for disease diagnoses, personalized treatment, and the prediction of treatment outcomes.

Pharmacogenetics can aid biopharmaceutical research and development in many ways. First, it is applied in diagnosis of disease. An important example is prenatal genetic screening, such as amniocentesis for the Down syndrome and newborn genetic screening for inherited metabolic disorders. Second, it is applied in drug efficacy prediction. A typical clinical trial would have a third or more patients who do not respond to the test drug. By identifying the genes that are associated with drug response, it is possible to improve drug response rates in clinical trials by selecting targeted patient populations. See Ref. [1] for a list of genes that have been significantly associated with drug response. Third, similar to predicting drug efficacy, it is applied in side effects prediction of a drug (common side effects in clinical trials and/or rare side effects of marketed drug). Fourth, it would be applied in market expansion in the sense of identifying patients who are currently unsuited to the drug but potentially responsive to the drug when dosage or formula is selected based on their genotypes. Finally, clinical, genetic, and molecular phenotypes can be integrated to identify drug targets.

In this article we will describe the basic concepts of and commonly used methods in genetic mapping [linkage and linkage disequilibrium (LD) analysis] of quantitative trait loci (QTL). Interested readers can go through the references cited here or elsewhere for more information on mapping QTL. Classical linkage analysis is described in section "Classical Linkage Analysis," and it can be used to construct a genetic map according to the distances between marker pairs. The idea of linkage and LD analysis is to scan the genome to detect the correlation between phenotype and the markers. Two commonly used methods used for linkage and LD analysis, regression and variance-component methods, are described in section "Mapping QTL." Section "Genetics of Complex Diseases: The Endophenotype Approach" describes some of our work on mapping QTL of neurological and psychiatric disorders. The last section is devoted to discussion.

CLASSICAL LINKAGE ANALYSIS

Chromosomes are long strands of deoxyribonucleic acid (DNA) made up of four bases or nucleotides: A (adenine), C (cytosine), T (thymine), and G (guanine). Each individual has 23 pairs of chromosomes. Each chromosome pair has a paternal homologue (a copy carried by the sperm) and a maternal homologue (a copy carried by the egg). Genes are basic units of inheritance (short segments of DNA) and act as if linearly arranged at field places (loci) on a chromosome transmitted in the gametes from parent to offspring. It is now believed that the human genome has around 20,000–25,000 genes. A genetic marker is a segment of DNA with an identifiable physical location on a chromosome and whose inheritance can be followed. A marker can be a gene or a group of genes, or it can be

some section of DNA with no known function. One type of marker is a Short Tandem Repeat, which is a sequence of DNA nucleotides (dinucleotides, trinucleotides, etc.) that varies among individuals in terms of how many times the sequence is repeated (AT, ATAT, ATATAT, etc.). A more recently developed marker is the single nucleotide polymorphism (SNP) in which individuals differ in a single nucleotide at that location (e.g., some individuals have the sequence ATAGCTA, while others have the sequence ATAGGTA). Markers are selected that vary among individuals, i.e., the selected markers are informative. Because DNA segments that lie near each other on a chromosome tend to be inherited together, markers are often used as indirect ways of tracking the inheritance pattern of a gene that has not yet been identified but whose approximate location is known. An allele is one of several alternative forms of a marker occupying the same locus on a particular chromosome. Genotype data at a marker refer to the maternal and paternal alleles at the marker.

The process of forming germ cells is called meiosis. During meiosis, the chromosome pairs are split and the maternal and paternal homologues recombine (crossover), forming two new homologues. This random shuffling of genetic material is called recombination. Every germ cell (sperm or egg) contains one homologue from each of the 23 chromosome pairs. During fertilization, a sperm and an egg combine, and the paternal and maternal homologues form a full chromosomal set. As a result of crossover, each offspring will carry genes from all four grandparents. If no crossover had occurred, the offspring would have only inherited genes from two grandparents.

The genetic distance between two markers is defined as the expected number of crossovers between them. Note that genetic distance is not a measure of the physical distance between the two markers on a chromosome. Haldane's model of recombination is that crossover occurs according to a Poisson process, leading to the mapping function

$$\theta = (1 - e^{-2x})/2$$

where θ is the recombination probability between two markers and x is the genetic distance in Morgans (expected number of crossovers) between the two markers. See Ref. [2] for other mapping functions. With all distances between marker pairs being computed, a genetic map can be constructed to show the relative locations of the markers along the chromosome. Genetic maps are used in linkage analysis to identify markers for disease traits. New techniques have produced dense maps with more markers (e.g., SNPs), and hence enable fine mapping of disease genes.

MAPPING QTL

A qualitative trait is expressed qualitatively, which means that the phenotype (what you see) falls into different

categories (e.g., blood types) and is a clear representation of its genotype. A quantitative trait shows continuous variation (e.g., height or blood pressure) because the trait is under the influence of many genes. Most of these genes have a small effect on the total phenotype value. In addition to the genetic effects associated with the quantitative trait, the phenotype is directly influenced by environmental factors. This results in the observed phenotypic variation appearing continuous in nature. This is further complicated by gene–gene interactions and gene–environment interactions. The observed phenotype for a quantitative trait is the sum of all genetic effects across all involved genes and environmental effects.

Mapping of qualitative and QTL (genes) is used to study the relation between a phenotypic value and one or more genetic markers. A large correlation between a phenotype value and a marker genotype indicates that the marker is close to a gene that influences the phenotype. Various statistical models and methods have been developed to study this correlation. Some methods are based on pedigree data (linkage analysis) while other methods are based on population and pedigree data (association methods). Combined linkage and association methods have improved statistical power. We will focus on statistical methods for mapping QTL, which are more complicated than mapping qualitative trait loci. These methods are also applicable or can be adapted to map qualitative trait loci.

Heritability

The concept of heritability was introduced to quantify the level of predictability of passage of a biologically interesting phenotype from parent to offspring.^[3] The phenotype is seen as a function of environment and genotype. Assuming additive and independent effects of genes and environment (a simplified model), phenotypic variance (V_p) of a trait in a population may be expressed by one component of genetic variance (V_G) and one of environmental variance (V_E), so that $V_p = V_G + V_E$.

Heritability measures how the genetic contribution to a trait might vary in a population, and the heritability coefficient, H^2 , is given by the ratio of the total genetic variance to the phenotypic variance: $H^2 = V_G/V_p$. This is the broad heritability. Another measure, the narrow heritability, is only based on the additive genetic variation. We will discuss additive genetic variance and dominant genetic variance in the next section when we describe the variance-component model. The heritability can be used to predict the effects of searching for genes. It is easier to detect genes determining a trait if heritability of the trait is stronger. But heritability analysis does not provide any information about the locations of the contributing genes. This is the purpose of linkage/LD analysis.

The phenotypic variance (V_p) is observed, but the genetic variance (V_G) has to be estimated. If genotype data are available, marker-based methods can be employed to

estimate heritability.^[4,5] Note that heritability can be estimated without knowing genotypes. To estimate heritability from phenotypes without genotype data, one would need to make additional assumptions or focus on specific pairs of relatives so that phenotype variance is not completely described by environmental effects. A popular approach is to use twins^[6] and assume equal environmental effects.

Linkage Analysis

As QTL are unknown, linkage analysis studies the coinheritance of phenotypes and marker genotypes within families. One expects relatives who have similar phenotypes to have similar genotypes at marker loci close to genetic loci that influence the trait, while markers at distant loci behave stochastically according to the rules of Mendelian inheritance. We will review two of the most commonly used methods and their extensions: Haseman–Elston regression^[7] and variance-component model.^[4,8]

Two genes are identical by descent (IBD) if one is a copy of the other or if both are physical copies of the same ancestral gene. Thus a parent and a child always share one allele IBD at each locus. Two siblings can share 0, 1, or 2 alleles IBD at each locus with probabilities 1/4, 1/2, and 1/4, respectively. For a distant pair of relatives in a complex pedigree, it is impossible to enumerate all the possibilities of allele sharing and all the breeding types, to calculate the allele sharing IBD probabilities between the pair. It is even more difficult for multipoint analysis. Statistical methods have been developed to estimate the allele sharing IBD probabilities between any pair of a pedigree.^[4,9,10] Some methods are deterministic while most are stochastic (Monte Carlo techniques).

If linkage exists between a trait locus and a marker, then alleles sharing IBD at the two loci are correlated. The Haseman and Elston^[7] method (H–E method) is derived for sib pairs and for regressing the squared sib-pair trait phenotype difference on the proportion of alleles the sibs are estimated to share IBD at a marker locus. Let y_{1j} and y_{2j} be the normalized phenotype values of the j th sib pair such that the population mean is zero. Let π_j be the estimated proportion of alleles the j th sib pair shared IBD at a marker locus. The slope β from the regression line

$$(y_{1j} - y_{2j})^2 = \beta_0 + \pi_j \beta + \varepsilon_j$$

indicates whether a linkage exists between the trait locus and the marker locus. The intercept β_0 is of no particular interest, and ε_j is the residual. The slope β is negatively proportional to the genetic part of the phenotype variance. If β is negative then there exists linkage between a trait locus and the marker locus, because it correlates similarity at a trait locus with similarity at the marker locus (sharing more alleles IBD makes the phenotype difference between the sib pair smaller). For each marker on a genetic map,

one can carry out the regression and find out which markers are statistically significantly linked to the trait loci. The trait loci are close to these chromosome locations. With more dense maps available because of advanced technology, one can do two sweeps to save computational time: The first sweep is on a relatively sparse map to find the rough locations of the trait loci, and the second sweep is on all the available markers in a neighborhood of those rough locations. Note that this is a multitest problem, but the p -values are not adjusted.

Extensions of the H–E method have been developed to increase statistical power. Wright^[11] pointed out that sib-pair QTL linkage information is not fully utilized by considering only phenotype differences of sib pairs. Further information can be obtained from phenotype sums of sib pairs. Sham and Purcell^[12] proposed a combined H–E method. This method regresses a linear combination of $(y_{1j} + y_{2j})^2$ and $(y_{1j} - y_{2j})^2$ on π_j , the estimated proportion of alleles shared IBD at a marker locus by sib pairs. Elston et al.^[13] introduced an improved H–E method which regresses the products of the sib phenotypes, $y_{1j}y_{2j}$, on the estimated proportion of alleles shared IBD. Olson and Wijsman^[14] extended the H–E method for use with general pedigrees, considering $(y_{1j} - y_{2j})^2$ for all relative pairs. These H–E methods typically use the Wald statistic based on the SE from ordinary least squares.

For the variance-component model,^[4,8] the phenotype values of a pedigree are assumed to follow a multivariate normal distribution. Let $\mathbf{Y}_i = (y_{i1}, y_{i2}, \dots)'$ be the phenotypes of the i th pedigree. If there are n trait loci influencing the trait, then the variance-component model can be written as

$$y_{ik} = \mu + x_{ik}\beta + \sum_{j=1}^n \gamma_{ikj} + \varepsilon_{ik} \quad (1)$$

where μ is the grand mean, x_{ik} is a vector of covariates (e.g., age, gender), β represents fixed effects corresponding to the covariates, γ_{ikj} is the effect of the j th QTL, and ε_{ik} represents a random environmental deviation. Assume γ_{ikj} and ε_{ik} are uncorrelated random variables with mean 0 and variance σ_j^2 and σ_ε^2 , respectively. With both additive and dominant effects, $\sigma_j^2 = \sigma_{aj}^2 + \sigma_{dj}^2$, where σ_{aj}^2 is the additive genetic variance and σ_{dj}^2 is the dominant genetic variance. The variance of y_{ik} is

$$\sigma_y^2 = \sum_{j=1}^n \sigma_j^2 + \sigma_\varepsilon^2$$

and the phenotype covariance between a pair of relatives indexed by k and m is

$$\begin{aligned} \text{COV}(y_{ik}, y_{im}) &= E[(y_{ik} - \mu)(y_{im} - \mu)] \\ &= \sum_{j=1}^n [\pi_{kmj} \sigma_{aj}^2 + \kappa_{kmj} \sigma_{dj}^2] \end{aligned}$$

where π_{kmj} is the proportion of alleles shared IBD at the j th QTL by the relative pair and κ_{kmj} is the j th QTL-specific probability of the pair of relatives sharing two alleles IBD. Both π_{kmj} and κ_{kmj} can be estimated from genetic marker data.

To search for the QTL that influence the phenotype values, one would focus on one QTL at a time. If we focus on the j th QTL, then model (1) can be written as

$$y_{ik} = \mu + x_{ik}\beta + \gamma_{ikj} + g_{ik} + \varepsilon_{ik}$$

where g_{ik} represents the random genetic effect by all QTL other than the j th QTL. The covariance between a pair of relatives indexed by k and m is then

$$\text{COV}(y_{ik}, y_{im}) = \pi_{kmj}\sigma_{aj}^2 + \kappa_{kmj}\sigma_{dj}^2 + 2\phi_{km}\sigma_g^2 + \delta_{km}\sigma_d^2 \quad (2)$$

where $\phi_{km} = E[\pi_{kmj}]/2$ is the expected kinship coefficient over the genome and $\delta_{km} = E[\kappa_{kmj}]$ is the expected probability of sharing two alleles IBD.

The variance-component model uses likelihood ratios to test significance of each marker. Let $\Delta_i = Y_i - (\mu, \mu, \dots)' - [(x_{i1}, x_{i2}, \dots)\beta]'$ be the centered phenotype values of the i th pedigree and Ω_i be its covariance matrix at the j th QTL with elements given by Eq. (2). Assume Y_i follows a multivariate normal distribution. The log-likelihood of the i th pedigree with t individuals at the j th QTL is

$$\ln L_i(\mu, \beta, \sigma_{aj}^2, \sigma_{dj}^2, \sigma_g^2, \sigma_d^2 | Y_i, (x_{i1}, x_{i2}, \dots)) = -\frac{t}{2} \ln(2\pi) - \frac{1}{2} \ln(|\Omega_i|) - \frac{1}{2} \Delta_i' \Omega_i^{-1} \Delta_i \quad (3)$$

If there are N pedigrees, then the log-likelihood of all pedigrees is simply the sum of the log-likelihood for each pedigree. Likelihood estimation results in consistent parameter estimates even when multivariate normality assumption is violated.^[8] Simulations confirm the consistency of variance-component estimates of genetic effect size.^[15]

LOD Score

If the j th QTL in formula (2) does not contribute to the trait under study, then Eq. (2) becomes a reduced model

$$\text{COV}(y_{ik}, y_{im}) = 2\phi_{km}\sigma_g^2 + \delta_{km}\sigma_d^2 \quad (4)$$

and the log-likelihood (3) is reduced to

$$\ln L_i(\mu, \beta, \sigma_g^2, \sigma_d^2 | Y_i, (x_{i1}, x_{i2}, \dots)) = -\frac{t}{2} \ln(2\pi) - \frac{1}{2} \ln(|\Omega_i^r|) - \frac{1}{2} \Delta_i' (\Omega_i^r)^{-1} \Delta_i \quad (5)$$

where Ω_i^r is a covariance matrix at the j th QTL with elements given by Eq. (4). If there are N pedigrees, then the log-likelihood of all pedigrees for the reduced model is the sum of the log-likelihood for each pedigree.

The hypothesis of no effect of j th QTL can be tested by likelihood ratio test. To this end, we calculate the logarithm of the odds (to the base 10) (LOD) score

$$\text{LOD} = -\log_{10} \frac{\hat{L}_r}{\hat{L}}$$

Here $\hat{L}$ is the likelihood of full model (3) with parameters estimated by maximum-likelihood method, and $\hat{L}_r$ is the likelihood of reduced model (5) with parameters estimated by maximum-likelihood method.

It is typical to interpret a LOD score above 3 as evidence of linkage. Roughly speaking, a LOD score of 3 corresponds to a p -value of 0.0001. This is much more stringent than the usual significance level of 0.05 and is for controlling the false positive rate when performing multiple tests. To find a chromosome location (marker) close to a QTL, we need to scan the genome at many markers.

IBD Estimation

Both H-E and variance-component methods rely on IBD sharing probabilities between relatives within a pedigree, but it is impossible to obtain IBD sharing probabilities directly at QTL. Many statistical methods have been developed to estimate IBD allele sharing at QTL from genetic marker data. The idea is that if a QTL is close to a marker, then recombination between the QTL and the marker is rare. Consequently, IBD allele sharing is almost identical at the QTL and the marker. More accurate estimate of IBD allele sharing at a QTL can be obtained by using multiple markers in the vicinity of the QTL.

Davis et al.^[16] proposed an algorithm to calculate IBD probabilities for complicated pedigrees when there is no missing genetic marker information. The method of Fulker, Cherny, and Garton^[17] is a multipoint estimation of IBD allele sharing for sib pairs, and Almasy and Blangero^[4] extend this multipoint approach to arbitrary pairs in a pedigree. Jung, Fan, and Jin^[18] estimate IBD allele sharing at a QTL by a linear combination of IBD allele sharing at multiple markers in the vicinity of the QTL.

LD and Fine Mapping

Linkage analysis is a pedigree-based method that relies on the phenomena of recombination (each chromosome that an offspring receives from a parent consists of segments derived from both grandparents). In practice, one can only obtain DNA samples from a few generations of a pedigree. It is unlikely to observe recombination events between two closely linked markers from pedigree data. Therefore linkage analysis has limited resolution and is usually not able to localize genes within 1.0 cM.^[19]

As a complement to linkage analysis, association mapping or LD mapping is a population-based analysis and offers an approach for obtaining more precise location of

disease genes—the so-called fine mapping. LD refers to the statistical dependence of alleles at different loci, i.e., $P(AB) \neq P(A)P(B)$. A frequently used measure is the correlation coefficient Δ^2 .^[20] Usually fine mapping uses dense markers of SNP. A SNP gives rise to two alleles. Consider two biallelic loci with alleles A and a at one locus and alleles B and b at the other, where labeling is arbitrary. Let π_A , π_a , π_B , and π_b be the allele frequencies, and let π_{AB} , π_{aB} , π_{Ab} , and π_{ab} be the four haplotype frequencies. Then

$$\Delta^2 = (\pi_{AB} - \pi_A \pi_B)^2 / (\pi_A \pi_a \pi_B \pi_b)$$

A Chi-square test can be constructed to test the hypothesis of independence between marker pairs.^[19] Note that linkage always implies association, but association (correlation) between two markers does not necessarily imply linkage, because the correlation could be because of population stratification. So association mapping has high resolution but with a high false positive rate. One can use linkage analysis to scan the genome for markers affecting a trait, and then use association mapping in the neighborhoods of those markers for precise locations of the QTL.

For mapping QTL one has to detect correlation between markers and quantitative traits. Luo, Tao, and Zeng^[21] proposed a likelihood analysis method to estimate LD between a marker and a QTL using unrelated individuals in a natural population. With this maximum-likelihood analysis, the QTL genotypes are considered as missing data and an expectation-maximization algorithm is applied to infer the genotype information. Then QTL allele frequencies, LD between the QTL and the marker under consideration, and QTL effects are estimated simultaneously. Thus the association analysis does not require pedigree data, though pedigree information can be used to avoid bias because of population stratification and to have additional power.

Association studies can also be regarded as identity by state (IBS) analysis. Two alleles are IBS if they are of the same type but are not necessarily pure copies. Two individuals sharing alleles IBS at a marker are not necessarily sharing alleles IBD. The regression and variance-component methods can be applied to both pedigree data and population data using allele sharing IBS. LD between an SNP and a QTL can be achieved by incorporation of the mean effects model as a measured covariate^[22–25] with likelihood approach. Consider a QTL with alleles A and a that contribute to the phenotype y . There are three genotypes, $G = AA, Aa$, and aa , at the SNP. Because the contribution of the SNP to trait y depends on the genotype G , the mean of the trait $E(y)$ will depend on the genotype G as well. Assuming additive effects of SNP alleles, the mean of the trait can be modeled as

$$E(y|G) = \mu + \beta N_A$$

where N_A is the number of A alleles corresponding to genotype G , and β is the regression coefficient describing the

strength of the influence of the genotype on the trait.^[25] It is possible to test whether the phenotypic mean differs between individuals with different genotypes by comparing the likelihood of a model with genotype-specific trait means to that of a model with equal means for all genotypes.

Combined Linkage and LD Analysis

Family pedigree data can be used for both linkage and LD analysis, and population data can be used in LD analysis. Combined linkage and LD mapping would allow the use of both pedigree and population data, reduce the false positive linkage because of the use of pedigree allele sharing IBD information, and at the same time allow the use of more dense maps (e.g., SNPs) for fine mapping because of LD analysis. Combined linkage and LD analysis will have more power in detecting QTL.

Almasy et al.^[26] proposed an approach to use allele sharing IBS information between unrelated individuals to augment the linkage information obtained from allele sharing IBD among relatives. This is an extension of the variance-component method for linkage analysis of Almasy and Blangero.^[4] Fulker et al.^[23] proposed a combined linkage and LD analysis using sib pairs. Their likelihood approach models linkage parameters in the covariance structure while they model LD parameters on the means of the phenotype. The idea is that population stratification will influence sib-pair means of the phenotype but not sib-pair differences. So modeling the gene effect differently for pair means and pair differences will reduce spurious LD. Fan and Xiong^[24] proposed a two-point method based on variance-component models. In their model, linkage information is contained in the variance-covariance matrix based on family data, and the LD information is contained in the mean parameter based on both family pedigree and population data.

Software

The literature on linkage and LD analysis is very rich, and software has been developed for QTL mapping in abundance. The webpage <http://www.nslj-genetics.org/soft/> has a comprehensive list of software for genetic linkage analysis, marker mapping, LD mapping, and pedigree drawing. Selecting an analysis method and software depends on the experimental design (twin study, sib-pair data, general pedigree, natural population, etc.).

Regression and variance-component methods are commonly used for QTL mapping. The H-E regression method using sib-pair data is available in GENEHUNTER, and the regression method is extended to general pedigree data in Multipoint Engine for Rapid Likelihood Inference (MERLIN). The variance-component method is available in GENEHUNTER, MERLIN, and Sequential Oligogenic Linkage Analysis Routines (SOLAR). We have been using

SOLAR^[4] for combined linkage and LD mapping of QTL of psychiatric disorders. Variance-component methods are more powerful than regression methods but with more restricted assumptions such as normality of the phenotypes. SOLAR is robust in phenotype distribution and it has a *t*-distribution option.

GENETICS OF COMPLEX DISEASES: THE ENDPHENOTYPE APPROACH

Linkage analysis techniques have been successfully utilized to locate genes of rare Mendelian diseases; however, these methods have had only marginal success in locating genes associated with more common complex (non-Mendelian) disorders such as neurological and psychiatric diseases. Some of the reasons for these difficulties include the disorders resulting from small effects of many genes (polygenic disorders), incomplete or low penetrance, clinical and genetic heterogeneity, the presence of phenocopies, and diagnosis uncertainty. However, some recent successes have resulted from the use of methods, which efficiently utilize full pedigree structures in extended pedigrees, and which employ QTL underlying disease phenotypes (endophenotypes or intermediate phenotypes).^[27] These quantitative biological markers serve as covariates that correlate with the main trait of interest (diagnosis) and serve to better define that trait or its underlying genetic mechanism.^[28] Endophenotypes therefore represent the genetic liability of the disorder among non-affected relatives of affected individuals.

To qualify as an endophenotype of a disorder, the trait in question must meet a number of criteria:

1. The trait must be present in affected individuals and correlate with diagnosis and severity of disease or age of onset.
2. The trait must reflect susceptibility and not be the consequence of transient states.
3. The trait must be present in unaffected relatives of affected controls with levels significantly higher than in random controls.
4. The trait must be heritable.

Here we provide a brief overview of work being undertaken as part of the Collaborative Study on the Genetics of Alcoholism (COGA) project. The overarching goal of COGA is to identify genes that affect the susceptibility to develop alcohol dependence and related phenotypes. The gene mapping strategy employed by COGA integrates quantitative endophenotypes derived from human electroencephalogram (EEG) recording along with extensive psychiatric diagnostic tools. This methodology has resulted in the identification of several susceptibility genes, two of which are summarized here.

The COGA

Alcoholism is a complex disease influenced by underlying biological susceptibility factors, by environmental factors, and by complex interactions among genes and between genes and the environment. Since its inception COGA has carefully ascertained and assessed a large sample of densely affected alcoholic families, a sample of less densely affected families, and a sample of comparison families unselected for psychiatric status. This task was conducted to obtain highly informative families and to obtain reasonable power for genetic analyses. A unique approach taken by COGA is the adoption of neuroelectric measures as quantitative endophenotypes to the study of genetic risk for the development of alcoholism and related disorders. Inherent in this approach is the ability to study neuroelectric phenomenon as phenotypes of cognition, which has the potential to elucidate the genetic underpinnings of human brain oscillations.

COGA has reported several important linkage findings for the major clinical phenotype, alcohol dependence, as well as for related clinical features of alcohol dependence such as smoking, depression, suicidal behavior, and conduct disorder.^[29,30] A number of the significant COGA linkage findings for alcoholism have been consequent to highly significant linkage findings using neuroelectric endophenotypes.^[31,32] These significant linkage findings have been followed up with LD studies to identify specific genes within these regions of linkage.

The Utility of EEG as a Quantitative Endophenotype

Recording brain electrical activity using scalp electrodes provides a non-invasive, sensitive measure of brain function in humans. These neuroelectric phenomena may be recorded during the continuous EEG when the subject is at rest, or one may record the time-specific event-related potentials (ERPs) and event-related oscillations (EROs) during specific cognitive tasks. These neuroelectric phenomena represent important measurable correlates of human information processing and cognition. They represent traits less complex and more proximal to gene function than either diagnostic labels or traditional cognitive measures do.^[33] The human brain response, as measured by recording and measuring EEG features, may be utilized as phenotypes of cognition, and as valuable tools for the understanding of some complex genetic disorders.^[34]

Two examples of the use of neuroelectric endophenotypes from the COGA project are summarized below. These neuroelectric measures were obtained using two types of EEG experiment: eyes closed resting EEG and the visual oddball ERP. The neuroelectric endophenotypes used in the examples have been shown to adhere

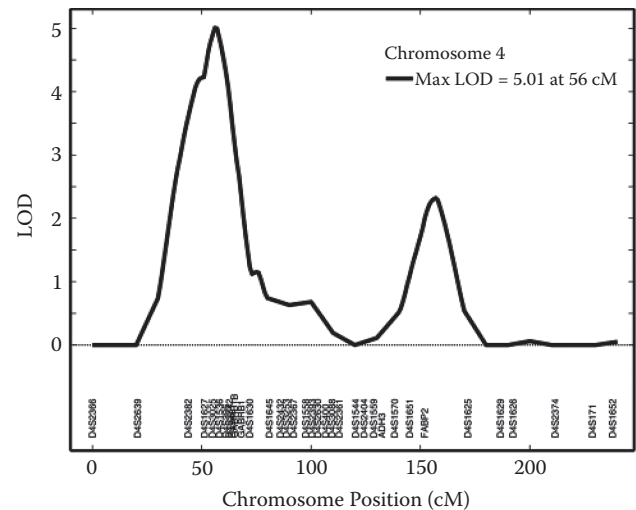
to the endophenotype qualification criteria outlined above. The first example uses resting EEG β band power which has been shown in several studies to have increased power in alcoholics, offspring of alcoholics, and high-risk individuals, particularly during the resting condition.^[34] Using twin studies, resting EEG β power has been found to be highly heritable with estimates between 80% and 90%.^[35] The second example uses measures derived from the P300 wave evoked during the visual oddball event-related experiment; these measures were designed to quantify the EROs underlying the P300 potential. Several studies have demonstrated that reduced P300 amplitude is found in alcoholics and their offspring, and is associated with the risk for alcoholism^[34,36,37] suggesting that the P300 is a useful endophenotype indexing familial risk for alcoholism. A P300 amplitude heritability estimate of 60% has been established through a meta-analysis of five P300 twin studies.^[38]

Example 1: Spontaneous β Band (16–20 Hz) EEG Oscillations During the Eyes Closed Resting State

EEG recordings were obtained with non-invasive scalp electrodes in the individuals who were awake but eyes closed. The filtered, artifact-free data were transformed into horizontal bipolar derivations. EEG absolute power between 3 and 28 Hz were subdivided into θ (3.0–7.0 Hz), $\alpha 1$ (7.5–9.0 Hz), $\alpha 2$ (9.5–12.0 Hz), $\beta 1$ (12.5–16.0 Hz), $\beta 2$ (16.5–20.0 Hz), and $\beta 3$ (20.5–28.0 Hz). A singular value decomposition procedure^[39] was utilized to obtain phenotypic data for each of the six EEG power bands.

Using the EEG- $\beta 2$ phenotypic data, a variance-component linkage analysis^[4] revealed significant linkage on a region of chromosome 4p, with a LOD score of 5 in alcoholic families (Fig. 1). Furthermore, a microsatellite marker in *GABRB1* in the center of this region showed LD with EEG- $\beta 2$. A combined linkage/LD analysis^[26] using this marker resulted in an increased LOD score of 6.5, which equates to an association p -value of 0.004.^[31] The estimated disequilibrium parameter ($\rho_d = 0.57$) indicated LD between the *GABRB1* microsatellite marker and the functional QTL. In addition, a novel non-parametric multipoint linkage technique also gave strong evidence for linkage in this region ($P < 1.0 \times 10^{-6}$) using the same EEG- $\beta 2$ phenotype.^[40]

These results implicated the GABA_A gene cluster as containing the causative functional genetic variants influencing the phenotype. This cluster includes the *GABRA2*, *GABRA4*, *GABRG1*, and *GABRB1* genes. The actions of GABA_A receptors are believed to be a fundamental requirement for the generation of high-frequency EEG oscillations (β and γ band), and blockade of these receptors has been observed to result in the loss of synchronization of these oscillations.^[41] Also, a number of important effects of alcohol are mediated by GABA transmission,^[42] making



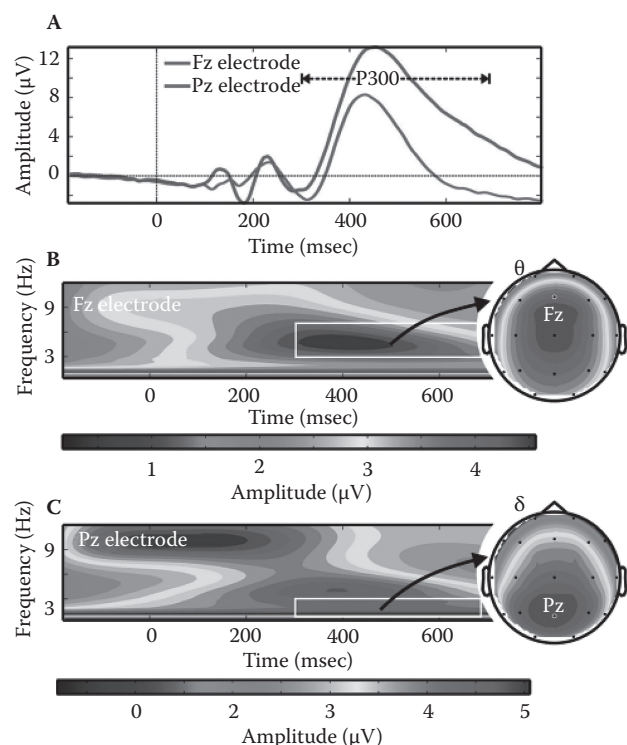


Fig. 2 Illustration of the ERO phenotypes used in the genetic analysis of visual oddball data. Traditional grand averaged evoked response potentials are depicted in (A) corresponding to target condition data for a frontal (Fz) and posterior (Pz) electrode. The P300 event is observed as a positive-going deflection starting approximately 300 msec after the target stimulus presentation. Time-frequency representations (TFR) of the target condition data (calculated using the S-transform) are depicted in (B) for the Fz electrode and (C) for the Pz electrode. These representations are obtained by averaging the instantaneous amplitude of the individual trial TFR data, and hence incorporate both stimulus phase locked and nonphase locked energy. Mean values calculated from individual subjects within time-frequency windows of interest were used as phenotypes in the genetic analysis. The inset head plots in (B) and (C) illustrate the distribution of these phenotypic measures across the head averaged within the θ and δ frequency bands and during the “P300” poststimulus time window (300–700 msec). (View this art in color at <http://www.informaworld.com>)

observed in the head plot insets of Fig. 2B and C that the θ band ERO component of the P300 response has a fronto-central distribution on the head, whereas the δ component shows a more posterior distribution.

A genome-wide variance-component linkage analysis of frontal channel θ band ERO data revealed significant linkage (LOD = 3.5) on chromosome 7q at 171 cM (Fig. 3). A cholinergic muscarinic receptor gene, *CHRM2*, resides near this locus and is a likely candidate to account for these linkage findings. The results of subsequent association analysis of the phenotype using SNPs genotyped within and flanking the *CHRM2* gene are provided in Table 1.

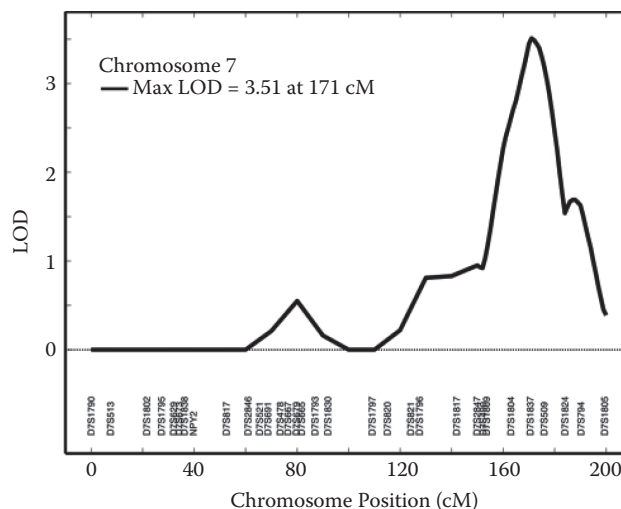


Table 1 Genetic Association *p*-Values (Uncorrected) for the Frontal Electrode Visual Oddball θ Phenotype with *CHRM2* SNPs

<i>CHRM2</i> SNPs (allele frequencies)	Measured genotype association		Gamete competition
	Additive model— <i>p</i> -value	Dominant model— <i>p</i> -value	<i>p</i> -Value
rs1824024 (G: 0.33/T: 0.67)	0.0008***	0.09	0.007**
rs2350786 (A: 0.29/G: 0.71)	0.015*	0.11	0.02*
rs8191992 (A: 0.41/T: 0.59)	0.53	0.48	0.14
rs1378650 (C: 0.58/T: 0.42)	0.2	0.68	0.015*

These results were obtained using a Caucasian-only subset of the original data to curtail possible stratification effects. The additive model measured genotype results (implemented in SOLAR) were obtained using a population-based measured genotype SNP model in which AA SNP genotypes are coded as −1, Aa genotypes are coded as 0, and aa genotypes are coded as 1. The dominant model measured genotype results were obtained using a population-based measured genotype SNP model in which AA and aa SNP genotypes are coded as 1 and Aa SNP genotypes are coded as 0. Age, gender, age-squared, and age by gender were included in the measured genotype models as fixed effects on the phenotype trait mean. The gamete competition association method (implemented in MENDEL) tests whether transmission of an allele is correlated with higher (or lower) trait values in the recipient of the allele. Age and gender effects on the quantitative traits were removed by linear regression prior to the gamete competition analysis.

**p* < 0.05.

***p* < 0.01.

****p* < 0.001.

CONCLUSIONS

A few of the current relevant issues regarding genetic mapping of complex traits will now be discussed. One important question is how large a sample size is required to have a reasonable power to detect linkage/association if it exists. Pfeiffer and Gail^[51] calculated sample size for population- and family-based case-control association studies using additive scores for various choices of marker allele frequencies (Table 2 in Ref. [51]). Olson and Wijsman^[52] considered several designs and sample size requirements in detecting LD. The simulation of Ekström^[53] shows that the LOD score with 300 sib pairs is much smaller than the LOD score with 100 four-sib nuclear families without information on parental genotypes using variance-component linkage models. This confirms that a smaller number of larger pedigrees are more informative than a larger number of smaller pedigrees for linkage analysis. Multipoint analysis gives increased power over single-point analysis to map QTL. Similarly, multitrait analysis provides accurate estimates of QTL positions.^[54] Multitrait methods use multivariate statistical models (multitrait single-point and multitrait multipoint). The methods described in section “Mapping QTL” can be extended to multitrait cases.

Sample size depends on required level of statistical significance, required power, design, test statistic to be used, and factors such as magnitude of effect. For linkage and LD analysis, such factors include the proportion of variance explained by QTL, gene action, marker heterozygosity, and density. The closed-form formula for sample size calculations is available in the literature. For the variance-component model, the generalized likelihood ratio test has a central Chi-square distribution under null hypothesis (no genetic component in phenotype variance) and has a non-central Chi-square distribution under alternative hypothesis. For sibship data and the variance-component model, Sham et al.^[55] derived closed-form

expressions of the non-centrality parameter, λ_s , per sib pair. A critical value is obtained based on the required significance level and the central Chi-square distribution. Then the required total non-centrality, λ_T , can be calculated based on the required power, the critical value, and the non-central Chi-square distribution under alternative hypothesis. The number of required sib pairs is given by

$$N = \lambda_T / \lambda_s$$

The non-centrality parameter per sib pair is:

$$\begin{aligned} \lambda_s \approx & V_A^2/8 + 3V_D^2/16 + V_A V_D/4 \\ & + 7V_A^4/64 + 63V_D^4/512 + 45V_A^2 V_D^2/63 \\ & + 7V_A^3 V_D/16 + 31V_A V_D^3 \\ & + V_S[3V_A^3/8 + 15V_D^3/32 + 9V_A^2 V_D/8 + 21V_A V_D^2/8] \\ & + V_S^2[3V_A^2/8 + 9V_D^2/16 + 3V_A V_D/4] \end{aligned}$$

for linkage analysis

$$\lambda_s = (3V_A/2 + 5V_D/4)/(V_N + 2V_S)$$

for association test based on difference between-sibship means, and

$$\lambda_s = (V_A/2 + 3V_D/4)/V_N$$

for association test based on difference within sibship, where V_A , V_D , V_S , and V_N are the additive, dominant, residual shared, and residual nonshared proportions of variance. If the association test is based on individual differences, then the non-centrality per sib pair is the sum of the non-centrality parameters per sib pair of the between-sibships and within-sibship tests. The critical value is 13.8 for a required significance of LOD = 3. With this critical value, the total non-centrality to have 80% power is 20.8. So given V_A , V_D , V_S , and V_N , one can calculate N , the required number

of sib pairs. Purcell, Cherny, and Sham^[56] developed software for sample size calculation, which is available online (<http://pngu.mgh.harvard.edu/~purcell/gpc/>). The closed-form formula of sample size calculations for a popular association test, the transmission disequilibrium test that is not discussed in this article, can be found in Iles.^[57]

Finally, it is expected that the future of genetic mapping will be influenced by the emerging microarray technologies. Genotypes and gene expressions of thousands of genes can be obtained simultaneously using microarrays. This provides huge amounts of data. Some statistical issues in microarray data analysis are discussed in Smyth, Yang, and Speed.^[58] One of the biggest issues with such data is that of multiple statistical testing. A possible solution is to use linkage analysis to locate the approximate locations of QTL and then use microarray to study genes around those locations. Linkage/LD analysis methods can also be extended to microarray data.^[59]

ACKNOWLEDGMENTS

The Collaborative Study on the Genetics of Alcoholism (COGA) (Principal Investigator: H. Begleiter; Co-Principal Investigators: L. Bierut, H. Edenberg, V. Hesselbrock, B. Porjesz) includes nine different centers where data collection, analysis, and storage take place. The nine sites and Principal Investigators and Co-Investigators are: University of Connecticut (V. Hesselbrock); Indiana University (H. Edenberg, J. Nurnberger Jr., P.M. Conneally, T. Foroud); University of Iowa (S. Kuperman, R. Crowe); SUNY Downstate Medical Center (B. Porjesz, H. Begleiter); Washington University in St. Louis (L. Bierut, A. Goate, J. Rice); University of California at San Diego (M. Schuckit); Howard University (R. Taylor); Rutgers University (J. Tischfield); Southwest Foundation (L. Almasy). Zhaoxia Ren serves as the NIAAA Staff Collaborator. This national collaborative study is supported by the NIH Grant U10AA008401 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and the National Institute on Drug Abuse (NIDA).

In memory of Theodore Reich, M.D., Co-Principal Investigator of COGA since its inception and one of the founders of modern psychiatric genetics, we acknowledge his immeasurable and fundamental scientific contributions to COGA and the field.

REFERENCES

- Goldstein, D.; Tate, S.; Sisodiya, S. Pharmacogenetics does genomic. *Nat. Rev. Genet.* **2003**, *4*, 937–947.
- Ott, J. *Analysis of Human Genetic Linkage*; Johns Hopkins University Press: Baltimore, MD and London, 1991.
- Feldman, M.W. Heritability: some theoretical ambiguities. In *Keywords in Evolutionary Biology*; Lloyd, E.A.; Keller, E.F., Eds.; Harvard University Press: Cambridge, MA, 1992, 151–157.
- Almasy, L.; Blangero, J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **1998**, *62*, 1198–1211.
- Ritland, K. Marker-inferred relatedness as a tool for detecting heritability in nature. *Mol. Ecol.* **2000**, *9*, 1195–1204.
- Neale, M.; Cardon, L. *Methodology for Genetic Studies of Twins and Families; NATO ASCI Series D: Behavioural and Social Sciences*; Kluwer Academic Publishers, 1992.
- Haseman, J.K.; Elston, R.C. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **1972**, *2*, 3–19.
- Amos, C.I. Robust variance-components approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.* **1994**, *54*, 535–543.
- Amos, C.I.; Dawson, D.V.; Elston, R.C. The probabilistic determination of identity-by-descent sharing for pairs of relatives from pedigrees. *Am. J. Hum. Genet.* **1990**, *47*, 842–853.
- Pong-Wong, R.; George, A.W.; Woolliams, J.A.; Haley, C.S. A simple and rapid method for calculating identity-by-descent matrices using multiple markers. *Genet. Sel. Evol.* **2001**, *33*(5), 453–471.
- Wright, F.A. The phenotypic difference discards sib-pair QTL linkage information. *Am. J. Hum. Genet.* **1997**, *60*, 740–742.
- Sham, P.; Purcell, S. Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. *Am. J. Hum. Genet.* **2001**, *68*, 1527–1532.
- Elston, R.C.; Buxbaum, S.; Jacobs, K.B.; Olson, J.M. Haseman and Elston revisited. *Genet. Epidemiol.* **2000**, *19*, 1–17.
- Olson, J.M.; Wijsman, E.M. Linkage between quantitative trait and marker loci: methods using all relative pairs. *Genet. Epidemiol.* **1993**, *10*, 87–102.
- Blangero, J.; Almasy, L. Multipoint oligogenic linkage analysis of quantitative traits. *Genet. Epidemiol.* **1997**, *14*, 959–964.
- Davis, S.; Schroeder, M.; Goldin, L.; Weeks, D. Nonparametric simulation-based statistics for detecting linkage in general pedigrees. *Am. J. Hum. Genet.* **1996**, *58*, 867–880.
- Fulker, D.; Cherny, S.; Gardon, L. Multipoint interval mapping of quantitative trait loci, using sib pairs. *Am. J. Hum. Genet.* **1995**, *56*, 1224–1233.
- Jung, J.; Fan, R.; Jin, L. Combined linkage and association mapping of quantitative trait loci by multiple markers. *Genetics* **2005**, *170*, 881–898.
- Weir, B. *Genetic Data Analysis II*; Sinauer Associates: Sunderland, MA, 1996.
- Pritchard, J.; Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **2001**, *69*, 1–14.
- Luo, Z.L.; Tao, S.H.; Zeng, Z.-B. Inferring linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. *Genetics* **2000**, *156*, 457–467.
- Boerwinkle, E.; Chakraborty, R.; Sing, C.F. The use of measured genotype information in the analysis of quantitative phenotypes in Man. I. Models and analytical methods. *Ann. Hum. Genet.* **1986**, *50*, 181–194.

23. Fulker, D.; Cherny, S.; Sham, P.; Hewitt, J. Combined linkage and association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.* **1999**, *64*, 259–267.
24. Fan, R.; Xiong, M. Combined high resolution linkage and association mapping of quantitative trait loci. *Eur. J. Hum. Genet.* **2002**, *11*, 125–137.
25. Iachine, I. Master of Applied Statistics, 2004. Available at <http://statmaster.sdu.dk/courses/st115>.
26. Almasy, L.; Williams, J.T.; Dyer, T.D.; Blangero, J. Quantitative trait locus detection using combined linkage/disequilibrium analysis. *Genet. Epidemiol.* **1999**, *17*, S31–S36.
27. Almasy, L.; Blangero, J. Endophenotypes as quantitative risk factors for psychiatric disease: rationale and study design. *Am. J. Med. Genet.* **2001**, *105*, 42–44.
28. Gottesman, I.; Gould, T.D. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am. J. Psychiatr.* **2003**, *160*, 636–645.
29. Edenberg, H.J.; Dick, D.M.; Xuei, X.; Tian, H.; Almasy, L.; Bauer, L.O.; Crowe, R.R.; Goate, A.; Hesselbrock, V.; Jones, K.; Kwon, J.; Li, T.K.; Nurnberger, J.I.; O'Connor, S.J.; Reich, T.; Rice, J.; Schuckit, M.A.; Porjesz, B.; Foroud, T.; Begleiter, H. Variations in GABRA2, encoding the $\alpha 2$ subunit of the GABA(A) receptor, are associated with alcohol dependence and with brain oscillations. *Am. J. Hum. Genet.* **2004**, *74*, 705–714.
30. Wang, J.C.; Hinrichs, A.L.; Stock, H.; Budde, J.; Allen, R.; Bertelsen, S.; Kwon, J.M.; Wu, W.; Dick, D.M.; Rice, J.; Jones, K.; Nurnberger, J., Jr.; Tischfield, J.; Porjesz, B.; Edenberg, H.J.; Hesselbrock, V.; Crowe, R.; Schuckit, M.; Begleiter, H.; Reich, T.; Goate, A.M.; Bierut, L.J. Evidence of common and specific genetic effects: association of the muscarinic acetylcholine receptor *M2* (*CHRM2*) gene with alcohol dependence and major depressive syndrome. *Hum. Mol. Genet.* **2004**, *13*, 1903–1911.
31. Porjesz, B.; Almasy, L.; Edenberg, H.; Wang, K.; Chorlian, D.B.; Foroud, T.; Goate, A.; Rice, J.P.; O'Connor, S.; Rohrbach, J.; Kuperman, S.; Bauer, L.; Crowe, R.R.; Schuckit, M.A.; Hesselbrock, V.; Conneally, P.M.; Tischfield, J.A.; Li, T.K.; Reich, T.; Begleiter, H. Linkage disequilibrium between the beta frequency of the human EEG and a GABA(A) receptor gene locus. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 3729–3733.
32. Jones, K.A.; Porjesz, B.; Almasy, L.; Bierut, L.; Goate, A.; Wang, J.C.; Wu, W.; Bauer, L.O.; Chorlian, D.B.; Dick, D.M.; Edenberg, H.J.; Foroud, T.; Hesselbrock, V.; Kuperman, S.; Nurnberger, J. Jr.; O'Connor, S.J.; Schuckit, M.A.; Stimus, A.T.; Reich, T.; Begleiter, H. Linkage and linkage disequilibrium of evoked EEG oscillations with *CHRM2* receptor gene polymorphisms: implications for human brain dynamics and cognition. *Int. J. Psychophysiol.* **2004**, *53*, 75–90.
33. Tsuang, M.T.; Faraone, S.V. The future of psychiatric genetics. *Curr. Psychiatr. Rep.* **2000**, *2*, 133–136.
34. Porjesz, B.; Rangaswamy, M.; Kamarajan, C.; Jones, K.A.; Padmanabhapillai, A.; Begleiter, H. The utility of neurophysiological markers in the study of alcoholism. *Clin. Neurophysiol.* **2005**, *116*, 993–1018.
35. Van Beijsterveldt, C.E.M.; Boomsma, D.I. Genetics of the human electroencephalogram (EEG) and event-related brain potentials (ERPs): a review. *Hum. Genet.* **1994**, *94*, 319–330.
36. Begleiter, H.; Porjesz, B.; Bihari, B.; Kissin, B. Event-related potentials in boys at risk for alcoholism. *Science* **1984**, *225*, 1493–1496.
37. Polich, J.; Pollock, V.E.; Bloom, F.E. Meta-analysis of P300 amplitude from males at risk for alcoholism. *Psychol. Bull.* **1994**, *115*, 55–73.
38. Van Beijsterveldt, C.E.M.; Van Baal, G.C.M. Twin and family studies of the human electroencephalogram: a review and a meta-analysis. *Biol. Psychol.* **2002**, *61*, 111–138.
39. Wang, K.; Begleiter, H.; Porjesz, B. Trilinear modeling of event-related potentials. *Brain Topogr.* **2000**, *12*, 263–271.
40. Ghosh, S.; Begleiter, H.; Porjesz, B.; Edenberg, H.; Foroud, T.; Reich, T. Linkage mapping of beta 2 EEG waves via non-parametric regression. *Am. J. Med. Genet.* **2003**, *118B*, 66–71.
41. Haenschel, C.; Baldeweg, T.; Croft, R.J.; Whittington, M.; Gruzelier, J. Gamma and beta frequency oscillations in response to novel auditory stimuli: a comparison of human electroencephalogram (EEG) data with in vitro models. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 7645–7650.
42. Harris, R.A.; Mihic, S.J.; Valenzuela, C.F. Alcohol and benzodiazepines: recent mechanistic studies. *Drug Alc. Dep.* **1998**, *51*, 155–164.
43. Covault, J.; Gelernter, J.; Hesselbrock, V.; Nellissery, M.; Kranzler, H.R. Allelic and haplotypic association of GABRA2 with alcohol dependence. *Am. J. Med. Genet.* **2004**, *129B*, 104–109.
44. Xu, K.; Westly, E.; Taubman, J.; Astor, W.; Lipsky, R.H.; Goldman, D. Linkage disequilibrium relationships among GABRA cluster genes located on chromosome 4 with Alcohol dependence in two populations. *Alcohol Clin. Exp. Res.* **2004**, *28*, 48A.
45. Stockwell, R.G.; Mansinha, L.; Lowe, R.P. Localization of the complex spectrum: the S-transform. *IEEE Trans. Sig. Proc.* **1996**, *44*, 998–1001.
46. Frodl-Bauch, T.; Bottlender, R.; Hegerl, U. Neurochemical substrates and neuroanatomical generators of the event-related P300. *Neuropsychobiology* **1999**, *40*, 86–94.
47. Comings, D.E.; Wu, S.; Rostamkhani, M.; McGue, M.; Lacono, W.G.; Cheng, L.S.-C.; MacMurray, J.P. Role of cholinergic muscarinic 2 receptor (*CHRM2*) gene in cognition. *Mol. Psychiatr.* **2003**, *8*, 10–13.
48. Perry, E. Cholinergic mechanisms and cognitive decline. *Eur. J. Anaesthesiol.* **1998**, *15*, 768–773.
49. Mitrofanis, J.; Guillery, R.W. New views of the thalamic reticular nucleus in the adult and in the developing brain. *Trends Neurosci.* **1993**, *16*, 240–245.
50. Callaway, E. The pharmacology of human information processing. *Psychopharmacologia* **1983**, *20*, 359–370.
51. Pfeiffer, R.; Gail, M. Sample size calculations for population- and family-based case-control association studies on marker genotypes. *Genet. Epidemiol.* **2003**, *25*, 136–148.
52. Olson, J.M.; Wijsman, E.M. Design and sample-size considerations in the detection of linkage disequilibrium with a disease locus. *Am. J. Hum. Genet.* **1994**, *55*(3), 574–580.
53. Ekström, C.T. Power of multipoint identity-by-descent methods to detect linkage using variance component models. *Genet. Epidemiol.* **2001**, *21*, 285–298.

54. Lund, M.; Sørensen, P.; Guldbrandtsen, B.; Sørensen, D. Multitrait fine mapping of quantitative trait loci using combined linkage disequilibria and linkage analysis. *Genetics* **2003**, *163*, 405–410.
55. Sham, P.; Cherny, S.; Purcell, S.; Hewitz, J. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am. J. Hum. Genet.* **2000**, *66*, 1616–1630.
56. Purcell, S.; Cherny, S.; Sham, P. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **2003**, *19*(1), 149–150.
57. Iles, M. On calculating the power of a TDT study—comparison of methods. *Ann. Hum. Genet.* **2002**, *66*, 323–328.
58. Smyth, G.; Yang, Y.; Speed, T. Statistical issues in cDNA microarray data analysis. In *Functional Genomics: Methods and Protocols*; Brownstein, M.J., Khodursky, A.B., Eds.; Humana Press: Totowa, NJ, 2003; Vol. 224 — *Methods in Molecular Biology*, 111–136.
59. Ott, J.; Hoh, J. Set association analysis of SNP case-control and microarray data. *J. Comput. Biol.* **2003**, *10*, 569–574.

Global Database and System

Alice T. M. Hsuan

Aventis Pharmaceuticals, Inc., Bridgewater, New Jersey, U.S.A.

Patrick Genyn

Johnson & Johnson PRD, Raritan, New Jersey, U.S.A.

Abstract

This entry briefly touches upon the system architecture and the methodology applied to support the development, deployment, and maintenance of such a validated system. It also devotes time to the data components of such a standardized database that are based upon a simple implementation of data-warehouse techniques.

INTRODUCTION

Standardization is a generic solution applied in business reengineering. Analysis of the clinical data management, statistical analysis, and reporting processes has identified several elements where standardization would result in time savings. Analysis has identified the database as the true center of the clinical development process. In the database, the “upstream” of clinical research raw data obtained in clinical trials, transformed in analyzable variables, meets the “downstream”—statistical analysis and reporting and clinical reporting. Standardization of this element will yield large time-saving benefits for the development process and higher-quality databases and reports; in addition, it will make more effective use of available resources. At the same time, a standardized database will allow for different modes of data entry, e.g., centralized as well as geographically distributed. It will also take advantage of technological changes such as electronic data capture versus manual data entry methods and provide customized analyses to meet different regulatory preferences in different countries and to efficiently answer additional regulatory requests during the review process. Standardized databases also provide the opportunity to harmonize and streamline the systems for clinical data management and statistical analysis. Applying such standardized systems will definitely result in consistent and comprehensive results and high-quality reporting, independent of where the system is applied in a decentralized approach of processing clinical trials.

This entry will give an overview of the requirements leading to such a standardized environment. It will briefly touch upon the system architecture and the methodology applied to support the development, deployment, and maintenance of such a validated system. A section is devoted to the data components of such a standardized database that

are based upon a simple implementation of data-warehouse techniques. The last sections will briefly lay out the transition into a standardized approach to processing clinical data and the critical success factors involved.

REQUIREMENTS ANALYSIS

With the guidelines of the International Conference on Harmonization (ICH) of Technical Requirements for Registration of Pharmaceuticals for Human Use^[1] in place, regulatory submissions to different authorities, such as the Department of Health and Human Services, the Food and Drug Administration (FDA), and the European Agency for the Evaluation of Medicinal Products (EMA), can be based on studies conducted all over the world. Consequently, studies conducted in the United States are used for submissions to EMA, and studies conducted in the rest of the world can be used in a new drug application (NDA) submitted to the FDA.

Managing, analyzing, and reporting individual trials can be done independently. However, consolidating the data and the results, for instance, into integrated summaries for safety and efficacy, can become a very challenging task when the databases of the individual trials were designed and constructed without having this final global objective of integration in mind. Therefore it is necessary to define and construct similar databases for each trial in terms of structure and content that are based on similar protocols and case report forms (CRFs).

In a decentralized environment for processing clinical data, the likelihood of duplicating work is always present. Standardized databases and systems, however, provide the opportunity to leverage resources globally by reusing database designs and programs from one trial to another. Appropriate and to-the-point communication is

an important and necessary proactive tool in leveraging efforts around the globe.

Standardized approaches in clinical databases and systems, leveraging resources by reusing designs and programs, and standardizing the processes of data management and statistical analysis are solutions for shortening the time to the market and for producing higher-quality databases and reports. However, these solutions do not guarantee these important results when the approach is not global and applied in all departments involved in clinical trials. Proactive management and careful planning are key factors in achieving those quality and time objectives.

Many regulatory authorities, such as the FDA and the EMEA, and other organizations, such as ICH and the International Organization on Standardization (ISO), are also imposing more strict regulations in the area of computerized systems.^[2] These regulations require that a pharmaceutical company be able to show at any time that the applied computer systems produce reliable, accurate, and complete data and results. All requirements mentioned in this paragraph will help in validating the programs and systems used for clinical data processing; however, the challenge of building and, especially, of maintaining a validated system is still there and must be a goal for all levels and functions in the different departments involved.

SYSTEM ARCHITECTURE

The architecture of a clinical system needs to be considered from the perspectives of data management, statistics, and system development. The foundation of the architecture is the physical computing environment. It is essential to separate the computing environment into three distinct environments—development, testing, and production—so that each function can be performed without adversely affecting other functions.

The development environment needs to be a very open one in which the developers have few restrictions. The developers should be able to try out as many alternatives as possible for designs and programs, some of which may even affect other developers in a negative way, such as bad performance or disk congestion. Ultimately, one developer could even cause the computer system to crash. However, by applying a split environment, such a tryout should not affect the production environment or the user acceptance environment because this environment will be separate.

User acceptance is a formal stage during development and testing of a computer system, and the results of the acceptance test are part of the validation package. Therefore the environment for this function needs to be more controlled and restricted as compared to the development environment and should be set up according to formal installation guidelines. At the same time, a user, being very familiar with real data, expects copies of real data or simulated data in the test environment. The user

can then use and manipulate these data to cover as many test scenarios as possible.

The actual processing of the clinical trial data is performed in a secured, validated production environment that is separated from the development and testing functions. The entire physical computing environment as part of the system architecture is depicted in Fig. 1.

The next building block in this computer environment is the data. The database containing the data refers in a broad sense to a “knowledge database” of a clinical project for developing a compound of a certain indication and formulation. The limits are:

Upstream: the construction of a normalized, relational, “raw” database, the design for which is based on the layout of the case record forms. This database contains the actual data collected using the CRF, diaries, quality-of-life questionnaires, etc.

Downstream: datasets containing the data for the tables and figures that format the output of the statistical analysis, and which are used for the clinical trial report and its addenda.

Another level in the hierarchy of the system architecture consists of the basic software for storing, retrieving, and analyzing data. A relational database management system (RDBMS), such as Oracle®, is most often used for storing data in a normalized and structured way. Clinical data management systems installed upon such as RDBMS are used for entering, cleaning, and reviewing the data. Examples are Clintrial™, Recorder™, and Oracle Clinical™. Statistical analysis of the clinical data is supported mostly by a separate package, such as SAS® or SPlus™. However, this package is closely integrated with the RDBMS containing the raw data.

A modular approach to developing systems will result in a very maintainable system. Therefore the system can be split into a number of tools and a number of subsystems. These tools and subsystems form the last level of the infrastructure hierarchy.

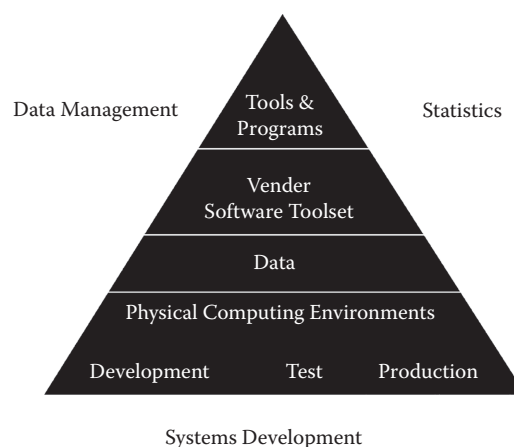


Fig. 1 The system architecture supports system- and trial-specific development in a validated environment

A subsystem will have a very specific function supporting a specific task in the process. An example of such a system is the data display subsystem that provides computer-based CRF printouts. These are used for evaluating the trial database; in addition, they are included in submissions to regulatory agencies. Another example of a subsystem is the efficacy system that analyzes all types of efficacy data: discrete, ordinal, and continuous.

Tools, on the other hand, do not support a specific task as a subsystem does, but is a piece of code with specific input and output that can be used for building the subsystems. In other words, the tools are the building blocks for subsystems. Tools can have different objectives, such as data handling, graphic representation, reporting, and applied statistics. Examples are tools to merge two or more datasets, to create summary reports, to output to a word processor, and to calculate the *p* value according to a specific statistical test.

DATA COMPONENTS

The global database contains many types of data and is used for many purposes. It will be much easier to manage the complex data by separating them into logical components. One such way is to separate them into the following five components (see Fig. 2).

- 1. The first component contains all data collected in a clinical trial—whether they be paper CRFs, ancillary data on tape, or other paper-based or computer-based data. These assorted data are named A database.
- 2. The second component is the base physical CRF database, the B database. This database is usually a relational database that contains raw CRF data and other ancillary data and that is organized in a normalized structure most suited for data entry.
- 3. The third component, the C database, is intended for clinical usage. The C database structure provides a view into the raw CRF data in the B database, but it

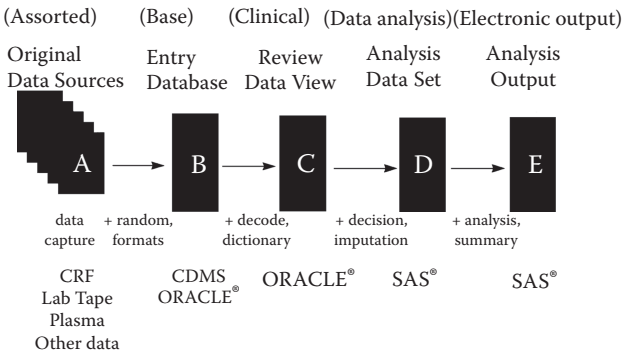


Fig. 2 The different database components enable an efficient workflow by mirroring the processes in data management and biostatistics

is organized in a denormalized structure most suited for retrieval and data review. Dictionaries are linked to the C views.

- 4. The fourth component, the D datasets, is designed for data analysis. The D database is constructed by adding calculated values, derived variables, and analysis decisions to the B database. Statisticians provide the requirements and specifications of the D database. The D datasets are usually in SAS format. They are organized on a subject level and include separate datasets for demographics, adverse events, efficacy, etc.
- 5. The fifth component contains the electronic outputs of the statistical summary and analysis, the E datasets. The statistical subsystem generates statistical summary and analysis outputs, i.e., summary tables, listings, and graphics.

SYSTEM DEVELOPMENT

Methodology

One of the most commonly applied development and qualification methodologies is the V model.^[3] This model is very simple and straightforward and is schematically described in Fig. 3.

- 1. The first step is to specify the system concept, including the scope of the project, a preliminary project schedule, and a first assessment of costs and resources needed.
- 2. The next step is to define and formalize the user and functional requirements of the system.^[4] The system requirement specification (SRS) describes each of the essential requirements of the software and the external interfaces. One or more reviews will be performed to verify its compliance with stated user requirements.
- 3. After approval of the SRS, the design specifications are formalized in the system design description (SDD) for the system architecture, software components, interfaces, and data. One or more reviews will be performed to ensure that the SDD for a given module achieves the objectives provided in the applicable SRS.
- 4. The system is then built during step 4, according to programming guidelines and naming conventions. Every developer performs construction testing that is freeform testing; however, some documentation is maintained. The objective is to ensure that the programming was carried out properly and that individual modules meet their specifications.
- 5. During step 5, technical testers will perform verification, in which formal testing is performed, independent of the developer, in a separate test environment.

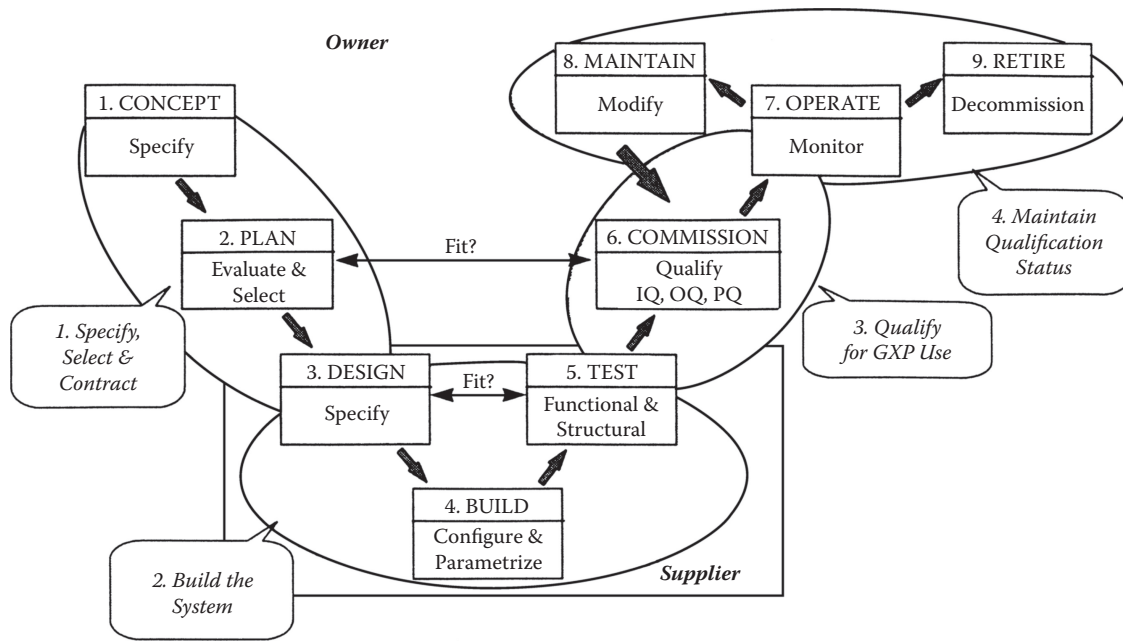


Fig. 3 The V model is an accepted development methodology with built-in validation components

The objective is to ensure that the programming of modules and interfaces was done properly and that the resulting product conforms to design specifications.

6. System qualification (“user acceptance testing”) during step 6 is a formal testing performed in an operating environment that is equivalent to the one used for normal operations. The objective is to ensure that the integrated system performs as expected under normal working conditions. System qualification consists of installation qualification (IQ), operational qualification (OQ), and performance qualification (PQ).

Validation

In the computer business, many definitions and approaches for the validation of computer systems are available. The definition applied for the qualification of the system is the following:

Validation of a system is to establish documented evidence that provides a high degree of assurance that a specific process will consistently produce results meeting its predetermined specifications and quality attributes.

The implementation of this validation approach resulted in three formal stages of validation: installation qualification, operational qualification, and performance qualification. Every qualification phase consists of three steps: defining the plan for qualifying the system, defining and executing the test scripts, and writing a summary report of the qualification. These steps are formally approved and

signed off by the system owners. This process is repeated for every qualification.

The IQ will demonstrate that the computerized system installed is the system designed to meet the requirements. The IQ will contain the evidence that the computerized system has been installed in an appropriate environment and that all system documentation is in place to start the use and maintenance of the system.

The OQ will produce the evidence that the system meets all its requirements by means of functional testing of the system under test conditions. The OQ will demonstrate that the system performs as specified according to predefined test scripts that are executed in a separate test environment but within the boundaries of normal usage of the system.

The PQ provides evidence that the system functions as intended by the user, using the defined procedures. The PQ will guarantee that the necessary standard operating procedures are in place, that users are properly trained, that errors are tracked and managed, and that changes to the system are controlled, etc.

MAINTENANCE

Change Management and Training

Systems to support clinical data management and biostatistics are very complex and not straightforward to use. Change management of people and customized training sessions are key solutions to a successful implementation and delivery of such a system.

A formal definition of change management is:

The process of aligning an organization's people and culture with changes in strategy, structure, and systems.

A formal change management approach will clearly define the customers and all other people that are directly or indirectly involved in the project and how they will be affected by the implementation of the system. By laying-out this structure, it becomes obvious who can help in implementing the system and who might be obstructive. The change management approach must be supported by a good communication plan. Having all these parts of a change management procedure defined, the actual change management can start during the development of the system, way before the actual implementation.

Adequate training sessions are part of the implementation of a good change management procedure and of performance qualification. The audience for the training for this system is very heterogeneous, e.g., technical and nontechnical people, or people from different functional areas, such as data management, biostatistics, and medical writing. Also, before a new hire can start working with the system, a more personalized training is required.

The training therefore is divided into three parts. The first part is an introduction to the new system. The second part explains and clearly defines the rationale behind the data management databases and systems and the biostatistics databases and systems. Newly introduced definitions, structures, techniques, and tests, among other items, are explained in detail, along with why a certain solution is chosen. The third part of the training package consists of technical usage of the system and is focused primarily on technical people such as database administrators, analysts, and programmers.

Change Control

Research in general and clinical research in particular are very dynamic and change very frequently. Clinical trials within the same development plan will vary; clinical trials can have a standard part, e.g., safety, and will have a very specific part as well, e.g., efficacy. Many changes and enhancements will be requested of a system supporting the process of analyzing clinical data. A change request is a request to modify one or more existing deliverables, such as documents or system components, for reasons other than nonconformance, and which have a direct impact on the configuration of the system.

A good clinical practices (GCP) system must remain in a validated state during its operational use. All changes to the system have to be recorded and evaluated for their impact on the validation status. If necessary, the system owner will perform a repeat of validation activities or have them performed. Therefore it is crucial to have a good change control approach in place to track, resolve, and

implement changes to the system. The main objectives of the change control procedure that is applied is to collect all bugs, anomalies, and enhancements, to track these requests from originator to implementation of the change, and to document changes in code, documentation, procedures, etc. The impact of such a system change on cleaned and closed databases or on reported trials needs to be carefully assessed because clinical development programs generally span many years.

TRANSITION

The introduction of a new clinical database and system is always complicated by the question of how to handle the ongoing trials, existing databases, and legacy systems. It is not possible to leave all current databases and systems to run their own course because clinical programs usually last many years. It is also not possible to move all current databases and systems into the new one, due to incompatibility and high cost. Therefore decisions regarding the transition level for the current database and system need to be made. The transition level may depend on the layout of the CRFs, the programming status, and the submission status. One other factor determining transition level is the balancing of the benefits with the cost in resources and budget. After a detailed inventory, the transition level may range from no conversion to full conversion. For example, we may opt not to convert projects without a plan for further development, and convert only integrated safety-summary-related safety data for projects that are at the submission stage. A detailed transition plan, including time schedule, resource needs, cost, and transition level, will provide a guide as to when, what, and how to transition all existing databases and systems.

CONCLUSION

The successful development and implementation of a new global system require both good system development methodology and good process. In previous sections, the importance of training and transition when implementing a new system was discussed. Throughout the life of the system, all involved parties need to co-own the system by being actively involved in the design, planning, development, implementation, and maintenance of the system. The involved parties include users and customers from all relevant sites of the company. Good project management in planning, resource, and communication are critical in ensuring the success of system development.

The main objective of a global clinical system is to shorten the drug development time while improving the quality of clinical data, analysis, summary, and reports. A good system alone will not reach this goal. It is necessary to review, modify, or even reengineer the analysis process,

from finalized protocol and CRFs to the generation of individual and integrated reports. The analysis process conducive to up-front planning, good communication and teamwork, and global collaboration will make optimal use of the global system. Coupling the new process with the new global clinical database and system will achieve the benefits of improved submission quality, synergy among sites, shared resources among sites, shortened submission preparation time, shortened response time for regulatory agency inquiries, and an enabling of the electronic data review internally and by the regulatory agencies.

REFERENCES

1. International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use. E6: Good Clinical Practice: Consolidated Guidelines; E8: General Considerations for Clinical Trials; E9: Statistical Considerations in the Design of Clinical Trials; E3: Clinical Trial Reports, Structure and Content.
2. Department of Health and Human Services, Food and Drug Administration. 21 CFR Part 11, Electronic records, Electronic signatures; Guidance for industry, computerized systems used in clinical trials, draft guidance.
3. Stokes, T.; Branning, R.C.; Chapman, K.G.; Hambloch, H.; Trill, A.J. *Good Computer Validation Practices, Common Sense Implementation*; Interpharm Press.
4. Institute of Electrical and Electronic Engineers. ANSI/IEEE Std 1058.1—1987, IEEE Standard for Software Project Management Plans; ANSI/IEEE Std 730.1—1989, IEEE Standard for Software Quality Assurance Plans; ANSI/IEEE Std 830—1984, IEEE Guide to Software Requirements Specifications; ANSI/IEEE Std 1016—1987, IEEE Recommended Practice for Software Design Description; ANSI/IEEE Std 1012—1986, IEEE Standard for Software Verification and Validation Plans.

Good Clinical Practice (GCP)

Mamoru Narukawa

Division of Pharmaceutical Medicine, Kitasato University Graduate School, Tokyo, Japan

Masahiro Takeuchi

Kitasato University Graduate School, Tokyo, Japan

Abstract

Clinical research is essential for the progress of medical science. This entry reviews the ethical and research guidelines for good clinical practice meant to be understood as international standards.

Keywords: Declaration of Helsinki; Good clinical practice; Harmonization.

INTRODUCTION

Clinical research is essential for the progress of medical science. Because it involves human subjects, their rights, safety, and well being must be protected, and the research should bring about meaningful results. Good Clinical Practice (GCP) is an ethical and scientific quality principle for designing, conducting, and reporting clinical research.

Originating from the Declaration of Helsinki, guidelines and rules named GCP have been developed, adopted, and implemented in several countries/regions, and these GCP guidelines have been in the process of international harmonization. The internationalized GCP guideline is expected to enable and to promote multi-regional clinical trials and a mutual acceptance of clinical trial data worldwide.

Because statistical justification of clinical research is a necessary condition for the fulfillment of scientific and ethical quality, statisticians' good understanding of GCP guidelines and their active participation in the research process are expected.

CLINICAL RESEARCH AND GOOD CLINICAL PRACTICE

Clinical research is an experimentation involving human subjects with a view to judging the effect of medical intervention, learning the natural course of a disease, and/or investigating factors that cause a disease. Although in vitro studies and nonclinical (animal) studies are essential and important experimental processes in medical science, information obtained from such studies is not always sufficient for clinical use. And even when the results of nonclinical studies suggest the medical intervention's efficacy and safety, we must reaffirm its result on human beings before applying it generally in clinical practice.^[1] The Declaration

of Helsinki says, "Medical progress is based on research that ultimately must include studies involving human subjects."^[2] This is the basis for necessitating the conducting of clinical research.

Considering such nature of clinical research, we must take care that the rights, safety, and well being of subjects who participate in the research are protected throughout the process. Good Clinical Practice is an ethical and scientific quality principle that is applied in designing, conducting, and reporting medical research that involves the participation of human subjects.^[3]

ETHICAL AND SCIENTIFIC QUALITY OF CLINICAL RESEARCH

Declaration of Helsinki

The Declaration of Helsinki, which was adopted by the 18th World Medical Association (WMA) General Assembly in Helsinki in June 1964, has been acknowledged as the cornerstone of research ethics to provide guidance to physicians and other participants in medical research that involves human subjects.^[2] It is a code that was declared based on the following, seemingly opposing, two principles:

- a. Physicians should give priority to protecting their patient's health over everything else.
- b. Experimentation involving human subjects is essential for the progress of science as well as for the relief of human beings who are suffering from diseases.

The declaration aims primarily at protecting the personal integrity and the welfare of subjects who participate in clinical research. Basic principles described in the declaration are shown in [Table 1](#).

Table 1 Basic Principles of the Declaration of Helsinki

- The well-being of the individual research subject must take precedence over all other interests.
- Medical research must conform to generally accepted scientific principles.
- A protocol, which describes the design and the performance of the research study, must be submitted for consideration, comment, guidance, and approval to a research ethics committee before the study begins.
- Every medical research study must be preceded by careful assessment of predictable risks and burdens in comparison with foreseeable benefits.
- Potential subjects must be adequately informed of the aims, methods, any possible conflicts of interest, the anticipated benefits and potential risks of the study, and others, and the subject's freely given informed consent must be obtained.

Among them, articles related to the subjects' informed consent and the peer review of protocols by the Institutional Review Board (IRB) are two key practical elements. The declaration has been amended several times so far, and the article related to the IRB (called "Research Ethics Committee" in the declaration), which was not included in the declaration when it was initially adopted in 1964, was added at a Tokyo amendment in 1975.

Now many countries and medical associations endorse the Declaration of Helsinki as their ethical code for clinical research.

Quality Assurance and Quality Control

The aim of quality assurance and quality control is to ensure the reliability of the data obtained in clinical research as well as to ensure the protection of human subjects from research risk. In other words, it is to guarantee that clinical research is conducted according to GCP guidelines.

In such activities, standard operating procedures (SOPs) and quality assurance units (QAUs) play key roles to achieve its aim. The SOPs are written instructions describing the general operations to be performed in proceeding with clinical research. Sponsors, investigators, and other medical professionals who are involved in clinical research should be fully acquainted with these instructions, and should fulfill their responsibilities. The QAU is responsible for surveying the operations related to clinical research in accordance with SOPs, and ensuring that the generated data will satisfy the requirements for quality.

Throughout the research, monitoring and audit are particularly important to assure quality. As for clinical trials sponsored by pharmaceutical companies, monitoring and audit by sponsors throughout the trials are essential, while the regulatory authorities also audit the trial independently. The monitoring and audit serve not only to assure data reliability but also to familiarize investigators and other medical professionals with the regulations and rules of clinical

trials. This will also lead to the improvement of the quality of clinical trials.

Thus the implementation of GCP guidelines will be reinforced and ensured by the monitoring and audit by the regulatory authorities as well as by the sponsors.

HISTORY OF GOOD CLINICAL PRACTICE IMPLEMENTATION BY THE UNITED STATES, EUROPEAN UNION, JAPAN, AND WORLD HEALTH ORGANIZATION

United States

The requirements for GCP were established as a matter of urgency by the United States in the early 1970s, responding to gross misconduct and, in many instances, fraud in clinical trials.^[4] A variety of proposed rules and regulations have been published between the latter half of the 1970s and the first half of the 1980s under the pretext of protection of human subjects and rules of new drug application (NDA).

It is well known that there exist no comprehensive regulations named GCP in the United States. The U.S. GCP comprises several federal regulations and other guidelines. The main components of the U.S. GCP are shown in Table 2.

The Code of Federal Regulations, Parts 50 and 56, describe the issues concerning ethics such as the rules and procedures for informed consent, and the role and responsibilities of IRBs, respectively. In Part 312, the responsibilities of the investigator and the sponsor/contract research organization (CRO) in conducting clinical trials are prescribed under the rules of investigational new drug application (IND). Part 314 sets forth procedures and requirements for the submission to, and the review by, the U.S. Food and Drug Administration (FDA) of applications to market a new drug.

The Compliance Program Guidance Manual is a manual for on-site inspection of IRBs, sponsors/CROs/monitors,

Table 2 Main Components of the U.S. GCP

Title 21 Code of Federal Regulations
Part 50: Protection of Human Subjects
Part 56: Institutional Review Boards
Part 312: Investigational New Drug Application
Part 314: Applications for FDA Approval to Market a New Drug
Food and Drug Administration Compliance Program Guidance Manual
– Bioresearch Monitoring
Institutional Review Board
Sponsors, Contract Research Organizations, and Monitors
Clinical Investigators

and investigators by the FDA staff, and explains the procedures, points for inspection, etc.

Over the years, these are updated and modified to include, for example, the 21 CFR Part 54 financial disclosure by clinical investigators. This is a regulation that requires an applicant to disclose certain financial arrangements between sponsors of the covered clinical trials and the investigators, to avoid the situation in which those arrangements could be a potential source of bias in the trials.

The U.S. FDA started audits of the work of clinical investigators in 1962, and audits on routine basis have been conducted since 1977.^[5] Common deficiencies following FDA site inspections for 2,711 investigators in North America between 1997 and 2008 are shown in Table 3. Common deficiencies include a failure to follow investigational plan as well as inadequate and inaccurate records (Table 3).^[6]

When deficiencies are found as a result of such audits, which indicate that the investigator has repeatedly or deliberately violated relevant regulations or has submitted false information to the sponsor, the FDA will initiate official actions including the measure of so-called “disqualification.” This means that the investigator may not use investigational drugs in the future.

European Union

The concept of GCP in European Union (EU) countries dates back to the 1960s, with the Nordic countries pioneering the concept around the ethical responsibility of the clinical investigator. Since those days, regional and local Independent Ethics Committees (IECs) have been established in EU member countries, applying the principles of research ethics based on the Declaration of Helsinki. The IEC is a committee that assesses the project’s reliability and validity, whether it entails the risks of harm or discomfort for research subjects, and estimates the anticipated value of the knowledge to be gained. The IEC examines whether the risks of harm or discomfort are balanced against the benefit, and whether the information and consent procedures proposed are appropriate in light of the ethical principles.^[7]

Table 3 Common Deficiencies Found in the FDA Audit for Clinical Investigators in North America (2,711 Investigations Between 1997 and 2008)

Most common deficiencies	Percentage
Failure to follow investigational plan	34.0
Inadequate and inaccurate records	23.7
Inadequate drug accountability	9.4
Inadequate informed consent form	9.1
Failure to report adverse drug reactions	8.2

From Ref. [6].

In July 1991, the Enforcement of the EEC Note for Guidance: “Good Clinical Practice for Trials on Medicinal Products in the European Community” was issued. This enforcement was setting into operation the GCP guidelines, but was not yet legally binding at that time. Then, in September 1991, the European Commission (EC) Directive 91/507/EEC, “Analytical, pharmacotoxicological and clinical standards and protocols in respect of the testing of medicinal products,” was published. By this enforcement, EC member countries were obliged to bring into force their laws, regulations, and administrative provisions necessary to comply with the directive, requesting all clinical trials to be designed, implemented, and recorded in accordance with the GCP guideline by January 1992.

In May 2001, the Directive 2001/20/EC of the European Parliament and of the Council on the approximation of the laws, regulations, and administrative provisions of the member states relating to the implementation of GCP in the conduct of clinical trials on medicinal products for human use was published. It finally came into effect in 2004 with the intention of streamlining the process of clinical trials throughout Europe. However, it was reported to have had negative impact on clinical trials, especially investigator-initiated academic research, because of the increased responsibility and bureaucratic burden.^[8] A system of routine GCP inspection has been in place since 2006.

Japan

The history of GCP in Japan is relatively new. The first GCP in Japan was published in 1989, as a notification of the Director General of the Pharmaceutical Affairs Bureau, Ministry of Health and Welfare (MHW). Its main principles were: (i) proper contract between the sponsor and the medical institution; (ii) establishment of IRB in the medical institution; (iii) protection of subjects’ rights; and (iv) record keeping. This old GCP had some peculiar aspects such as: (i) influential professors (called “principal investigators”) had clout with clinical trials and were in charge of crucial parts of the trial, including preparation of protocols and study reports; (ii) sponsors were not able to conduct on-site monitoring and source data verification; and (iii) informed consent obtained orally was accepted.

Then, in April 1997, a new GCP guideline, which was based on the ICH GCP (see “International Harmonization of Good Clinical Practice”), was introduced. By the adoption of this new GCP, Japanese GCP-related regulations had been replaced by international norms.

The MHW/PMDA (Pharmaceuticals and Medical Devices Evaluation Agency) has been conducting audit for clinical trials for regulatory registration on medical institutions including investigators as well as sponsors since 1993. Table 4 summarizes common deficiencies found in the audit for hospitals in fiscal year 2004, responding to 50 NDAs in the period.^[10] While Japanese GCP audits are

Table 4 Common Deficiencies Found in Japan MHW/PMDA Audit for Hospitals (from April 2004–March 2005)

Deficiencies	Number of findings in 114 audits/189 trials
Qualification of hospitals not met	1
Lack of appropriate SOPs for hospitals and IRBs	18
IRB not appropriately worked	76
Inappropriate contract between sponsors and hospitals	7
Protocol deviation	374
Incorrect/insufficient CRFs	222
Inappropriate informed consent	24
Unsatisfactory investigational drug accountability	8
Inappropriate archives	89

From Ref. [9].

targeted on clinical data and submitted documents, audits by the U.S. FDA are mainly focused on the investigators.

If critical deficiencies in clinical trials are found, official actions would be taken in the form of rejection of the submitted study results.

The implementation of the new GCP guideline has improved the ethical and scientific quality of trials conducted in Japan. It also leads to the harmonization of regulatory practice concerning clinical trials that contributes to the mutual acceptance of foreign clinical data.

World Health Organization

The World Health Organization (WHO) had been developing a GCP guideline named “Guidelines for Good Clinical Practice (GCP) for Trials on Pharmaceutical Products” in consultation with national drug regulatory authorities within WHO’s member states, and it was finally published in 1995.^[11] A number of developing countries have no national regulations for clinical trials or the regulations require supplementation. The establishment of internationally accepted GCP guidelines would be a great value in order to establish globally applicable standards for the conduct of biomedical research on human subjects, considering the future situation where more clinical trials would be conducted worldwide, including developing countries.

INTERNATIONAL HARMONIZATION OF GOOD CLINICAL PRACTICE

The activities for international harmonization of technical requirements for the registration of pharmaceuticals for human use, called ICH, have been carried out in the past decades by the EU, Japan, and the United States. The ICH is an international conference with the aim of harmonizing

technical requirements for NDA in the three regions and of promoting mutual acceptance of NDA data and, as a result, of providing good new medicines to patients sooner. More than 60 harmonized guidelines have been developed in the areas such as quality, safety (nonclinical studies), and efficacy (clinical studies).

In the area of efficacy, the mutual acceptance of clinical trial data (i.e., utilization of clinical data collected in foreign countries for NDA without conducting additional domestic trials) is a key issue. Clinical trial data with ethical and scientific quality are the prerequisite for considering mutual acceptance.

Good Clinical Practice is one of the topics selected as an issue for harmonization in the efficacy area. The objective of the ICH GCP guideline is to provide a unified standard for the EU, Japan, and the United States to facilitate a mutual acceptance of clinical data by the regulatory authorities in these regions.^[13]

Table 5 shows a brief history of the development and implementation of the ICH GCP guideline. The guideline was developed with consideration of the GCPs of the EU, Japan, and the United States, as well as those of Australia, Canada, the Nordic countries, and the WHO.

Members agreed on the draft guideline in May 1995, and the draft was circulated for comments in each region. After discussing comments submitted by relevant parties in the regions, the members reached an agreement on the final document. Regulatory authorities of the EU, Japan, and the United States have adopted and published the final guideline according to their domestic procedure.

The principles of ICH GCP guideline are described in Table 6.

The purposes of the ICH GCP guideline are to protect the rights of human subjects participating in clinical trials and to ensure the scientific validity and credibility of the data collected in the trials. The main principle in the guideline is that the rights, safety, and well being of the trial subjects are the most important considerations and should prevail over the interests of science and society.

Table 5 Brief History of ICH GCP

<i>Development of ICH-GCP</i>	
November 1991	Agreed to start cooperative work toward harmonized GCP
May 1995	Reached an agreement on the draft guideline
May 1996	Reached an agreement on the final guideline
<i>Implementation in each region</i>	
July 1996	EU: CPMP/ICH/135/95/Step 5
March 1997	Japan: PAB Notification No. 430, MHLW Ordinance No. 28
May 1997	United States: Federal Register, Vol. 62, No. 90

Table 6 Principles of the ICH GCP Guideline

1. Clinical trials should be conducted in accordance with the ethical principles that have their origin in the Declaration of Helsinki, and that are consistent with GCP and the applicable regulatory requirement(s).
2. Before a trial is initiated, foreseeable risks and inconveniences should be weighted against the anticipated benefit for the individual trial subject and society. A trial should be initiated and continued only if the anticipated benefits justify the risks.
3. The rights, safety, and well being of the trial subjects are the most important considerations and should prevail over the interests of science and society.
4. The available nonclinical and clinical information on an investigational product should be adequate to support the proposed clinical trial.
5. Clinical trials should be scientifically sound, and described in a clear, detailed protocol.
6. A trial should be conducted in compliance with the protocol that has received prior IRB/IEC approval/favorable opinion.
7. The medical care given to, and medical decisions made on behalf of, subjects should always be the responsibility of a qualified physician or, when appropriate, of a qualified dentist.
8. Each individual involved in conducting a trial should be qualified by education, training, and experience to perform one's respective task(s).
9. Freely given informed consent should be obtained from every subject prior to clinical trial participation.
10. All clinical trial information should be recorded, handled, and stored in a way that allows its accurate reporting, interpretation, and verification.
11. The confidentiality of records that could identify subjects should be protected, respecting the privacy and confidentiality rules in accordance with the applicable regulatory requirement(s).
12. Investigational products should be manufactured, handled, and stored in accordance with applicable good manufacturing practice (GMP). They should be used in accordance with the approved protocol.
13. Systems with procedures that assure the quality of every aspect of the trial should be implemented.

From Ref. [3].

The guideline has an important and beneficial impact on the clinical trials conducted in the three participating regions (Europe, Japan, and the United States) as well as in many other regions throughout the world. And it should fulfill its intended purpose of providing for a more economical use of resources and the elimination of unnecessary delays in the global development and availability of new drugs, and at the same time maintaining safeguards on quality, safety, and efficacy and regulatory obligations to protect public health.^[12]

FUTURE PERSPECTIVES

Biostatistics and Good Clinical Practice

The assurance of quality and credibility of clinical trial data is a prerequisite for the further analysis and evaluation of the trial results. In this context, clinical trials should be conducted to fulfill this requirement according to GCP guidelines. Biostatisticians play one of the key roles in such situations, and they are requested to participate in all stages of the trial process, from designing the protocol, planning the analyses, to preparing clinical study reports in order to assure the scientific quality as well as the ethical quality.^[3]

Because they involve human subjects, clinical trials that result in useless outcome from the viewpoint of their plans and procedures will create ethical problems, and should be avoided by all means. It is the task of biostatisticians to ensure that statistical principles are applied appropriately in clinical trials to demonstrate the drug's safety and efficacy, and make the results meaningful.^[13]

There seem some situations in which biostatistics directly relate to the interpretation and implementation of ethical principles or GCP guidelines in the planning and conduct of clinical trials. One of the examples is the issue related to the use of placebo in controlled clinical trials.

When the Declaration of Helsinki was revised in October 2000, the wording of a paragraph that referred to a placebo control caused a lot of discussion on the circumstances where the use of placebo is allowed.

A strict interpretation of the sentence seemed to rule out the use of placebo in circumstances where proven prophylactic, diagnostic, or therapeutic methods exist. This interpretation brought about concerns such as:

- a. Results of clinical trials, which are placebo-controlled, conducted in developing countries might be used for the benefits of patients in developed countries, as a result of the difference in the availability of a certain proven method between developed and developing countries, which would be a potential control treatment.
- b. Placebo-controlled trials in the areas where proven prophylactic, diagnostic, or therapeutic methods exist are ruled out, although such studies are necessary in some cases in order to obtain sufficient proof of the efficacy of a new treatment.

For this reason, the WMA Council decided to publish a note of clarification on the interpretation of this issue, in that, although the WMA opposes the notion that the nonavailability of drugs should be used as a justification

Table 7 Paragraph 32, the Declaration of Helsinki

The benefits, risks, burdens, and effectiveness of a new intervention must be tested against those of the best current proven intervention, except in the following circumstances:

- The use of placebo, or no treatment, is acceptable in studies where no current proven intervention exists; or
- Where for compelling and scientifically sound methodological reasons the use of placebo is necessary to determine the efficacy or safety of an intervention and the patients who receive placebo or no treatment will not be subject to any risk of serious or irreversible harm. Extreme care must be taken to avoid abuse of this option.

From Ref. [2].

to conduct placebo-controlled trials, a placebo-controlled trial may be ethically acceptable under certain situations, even if proven therapy is available.^[14] A trial of add-on treatment to a standard therapy, or a trial for a minor condition where there is no additional risk or irreversible harm for the subjects in placebo group, may be considered in such a case. Anyway, all other provisions of the declaration, including the subjects' informed consent and the protocol review by IRB, must be observed. The note was finally incorporated into the text, Paragraph 32, in 2008, as shown in [Table 7](#).

The ICH Guideline of Topic E10, Choice of Control Group and Related Issues in Clinical Trials, describes general principles concerning the choice of a control group for clinical trials intended to demonstrate the efficacy of a drug.^[15] Because it would not be feasible to establish an absolute standard concerning this issue, statisticians, together with medical professionals and other experts, should fully participate in the process of protocol development, and discuss the best choice of a control group deeply based on each condition in light of the Declaration of Helsinki, GCP guidelines, and relevant ICH guidelines, so that clinical trials are conducted in a scientific and ethical manner.

Globalization of Clinical Development

If pharmaceutical companies can make an NDA for approval by utilizing clinical data collected in foreign countries without conducting additional domestic trials, it gives a great impact on new drug development from the viewpoint of saving resources as well as providing new drugs to patients in a timely fashion. The ICH Guideline of Topic E5, Ethnic Factors in the Acceptability of Foreign Clinical Data, was established in 1998, and now the acceptability of foreign clinical data for domestic NDA is evaluated according to the guideline from the viewpoint of scientific quality as well as ethical quality. Because the guideline clearly states that clinical trials establishing dose-response, efficacy, and safety should be conducted

according to GCP guidelines, the observance of GCP guidelines is a prerequisite for examining the usability of the result as NDA data in foreign countries.^[16] In this sense, the harmonized ICH GCP guideline plays an important role in the new drug development, which is being internationalized continuously, because the observance of the ICH GCP guideline means the observance of international standard.

These guidelines and others at ICH opened the door for globalization process of clinical development of new drugs. By adopting the principles of the ICH GCP guideline, pharmaceutical companies accelerate their activity of conducting clinical trials worldwide, rather than only in conventional regions such as North America and Western Europe, from the viewpoints of cost, time, and regulatory environment. Over the past years, the share of Phase II/III trial sites for East Europe, Asia, and Latin America has been rapidly increasing. In this situation, there is growing concern among regulators and in public debate about how well these trials are conducted from an ethical and scientific standpoint. In order to address this concern, we need to further promote and establish good clinical practice standards worldwide as well as to strengthen the regulatory review and inspection process and to increase its transparency to assure the rights, safety, and well-being of trial subjects and the credibility of trial data.

CONCLUSION

In the past decades, GCP guidelines have been developed, adopted, and implemented in several countries/regions, and these GCP guidelines have been in the process of international harmonization. It is a natural course as well as a necessary course of action considering the recent internationalization of new drug development, which would enable and promote a mutual acceptance of clinical trial data worldwide. The role of internationalized GCP guideline will become more significant with the advances of medical science.

The adoption of GCP guidelines itself, however, is not a goal, nor does it necessarily assure a good quality of clinical trials. The quality assurance and quality control activities, including monitoring and audit by regulatory authorities as well as sponsors, are expected to ensure and to enhance the level of ethical and scientific quality of clinical research. Also it calls for the active participation of various experts in clinical research including statisticians. The statistical justification of clinical research is a necessary condition for the fulfillment of scientific and ethical quality, and therefore statisticians should be familiarized with GCP guidelines and take part in every stage of the process of clinical research.

REFERENCES

1. Sunahara, S. Introduction to clinical research (in Japanese). Igakushoin (Tokyo) **1988**, 14–15.
2. The World Medical Association. *Declaration of Helsinki*;—<http://www.wma.net/e/policy/pdf/17c.pdf> (accessed August 2009).
3. ICH Secretariat. *ICH Harmonized Tripartite Guideline: Guideline for Good Clinical Practice*; 1996—Geneva.
4. Dent, N.J. European good laboratory and clinical practices: their relevance to clinical pathology laboratories. *Qual. Assur.* **1991**, 1 (1), 82–88.
5. Lisook, A.B. FDA audits of clinical studies: policy and procedure. *J. Clin. Pharmacol.* **1990**, 30, 296–302.
6. Karlberg, J.P.E. US FDA Site inspection findings, 1997–2008, fail to justify globalization concerns. *Clinical Trial Magnifier* **2009**, 2, 194–212.
7. Strandberg, K. Independent Ethics Committees (IEC) in the European Union (EU). *Proceedings of the Third International Symposium on Clinical Research of New Drugs*, Osaka, March 2001; Mix: Tokyo; 2001.
8. Hemminki, A.; Kellokumpu-Lehtinen, P.L. Harmful impact of EU clinical trials directive. *BMJ* **2006**, 332, 501–502.
9. Saito, K.; Kodama, Y.; Ono, S.; Maida, C.; Fujimura, A.; Miyamoto, E. Reliability of Japanese clinical trials estimated from GCP audit findings. *Int. J. Clin. Pharmacol. Ther.* **2008**, 46, 415–420.
10. Ono, S.; Kodama, Y.; Nagao, T.; Toyoshima, S. The quality of conduct in Japanese clinical trials deficiencies found in GCP inspections. *Control Clin. Trials* **2002**, 23, 29–41.
11. World Health Organization. *Guidelines for Good Clinical Practice (GCP) for Trials on Pharmaceutical Products*;—<http://apps.who.int/medicinedocs/en/d/Jwhozip13e/> (accessed August 2009).
12. Dixon Jr., J.R. The International Conference on Harmonization Good Clinical Practice guideline. *Qual. Assur.* **1998**, 6 (2), 65–74.
13. ICH Secretariat. *ICH Harmonized Tripartite Guideline: Statistical Principles for Clinical Trials*; 1998—Geneva.
14. The World Medical Association. Press Release WMA clarifies its ethical guidance on the use of placebo-controlled trials;—http://www.wma.net/e/press/2001_5.htm (accessed August 2009).
15. ICH Secretariat. *ICH Harmonized Tripartite Guideline: Choice of Control Group and Related Issues in Clinical Trials*; 2000—Geneva.
16. ICH Secretariat. *ICH Harmonized Tripartite Guideline: Ethnic Factors in the Acceptability of Foreign Clinical Data*; 1998—Geneva.

Good Laboratory Practice (GLP)

Ying Geng, Baoming Ning, and Dejiang Tan

National Institutes for Food and Drug Control, Beijing, China

Abstract

Good laboratory practice (GLP) is a quality system in which nonclinical health and environmental safety studies are planned, performed, monitored, recorded, archived, and reported. The purpose of GLP is to ensure the adequate environment and the satisfied quality, integrity, traceability, and reliability of safety data through the implementation of well-established quality management system for safety assessment for human and environment. GLP was first introduced officially in 1972 New Zealand Testing Laboratory Act; however, the formal regulatory concept of GLP was created by U.S. Food and Drug Administration in 1978. Other two representative GLP regulations are GLP issued by the U.S. Environmental Protection Agency and GLP harmonized by the Organization for Economic Co-operation and Development. These three GLP regulations are different in scope as well as the definition of nonclinical laboratory study. GLP system shares common principles with laboratory accreditation quality system in a large sense, which differs significantly from accreditation system with respect to the quality assurance system and regulatory purpose. In order to promote the market approval procedure for new products, more and more national regulatory authorities have signed the Memorandum of Understanding or Mutual Recognition Agreement on GLP. Good statistic practice plays the key role in many aspects such as scientific and reliable study design, sampling plan, data analysis, instrument calibration, and decision rules.

Keywords: Good laboratory practice; Laboratory accreditation; Mutual recognition; Organization for Economic Co-operation and Development; Quality system; U.S. Environmental Protection Agency; U.S. Food and Drug Administration.

INTRODUCTION

Good laboratory practice (GLP) is a quality system in which nonclinical health and environmental safety studies are planned, performed, monitored, recorded, archived, and reported. Operated by government, laboratory studies under GLP regulation are conducted to obtain data on properties and/or its safety, intended for submission to appropriate regulatory authorities. After more than 40 years of practice and application in pharmaceutical industry, in a narrow sense, GLP is widely recognized as a common quality standard for organizations that perform the non-clinical safety studies on potential drug or medical devices. However, in a broad sense, the purpose of GLP is to ensure the adequate environment and the satisfied quality, integrity, traceability, and reliability of safety data through the implementation of a well-established quality management system for safety assessment for human and environment.

GLP quality system is different from laboratory accreditation quality systems like ISO17025, WHO/GPCL (Good Practice for Pharmaceutical Quality Control Laboratories). However, the essential principles of the two quality systems have much in common and are comparable to each

other to some extent. With respect to the quality control section of GMP (Good Manufacture Practice), GMP regulations is applied to products for human use, while GLP pertains to quality control system for safety assessment for human and environment. Both GMP and GLP ensure high quality of products and laboratory procedures. With the ever increasing globalization of industries and products, in spite of the fact that there are so many national or regional GLPs, in order to speed up of the market approval procedure for new products, internationally by harmonizing regulation and avoiding the duplicate many time-consuming and expensive test procedures, more and more national regulatory authorities has signed the Memorandum of Understanding (MOU) or Mutual Recognition Agreement (MRA) on GLP.

Since GLP principles ensure that the study falls under a rigorous quality assurance system, neither the scientific validity of method nor the ability to generate accurate and precise measurements, is the main purpose of GLP. In this sense, good statistic practice played the key role in the safety assessment studies in many aspects such as scientific and reliable study design, sampling plan, data analysis, instruments calibration, and decision rules.

NONCLINICAL LABORATORY STUDY AND SCOPE OF GLP

GLP was first introduced officially in 1972 New Zealand Testing Laboratory Act to provide for the development and maintenance of good conformity assessment practice that is consistent with international practice and facilitates trade.^[1] However, the formal regulatory concept of GLP was created by the U.S. Food and Drug Administration (FDA) in 1978.^[2] Other two representative GLP regulations^[3,4] are GLP issued by the U.S. Environmental Protection Agency (EPA) and GLP harmonized by the Organization for Economic Co-operation and Development (OECD). These three GLP regulations are different in scope as well as the definition of nonclinical laboratory study.

GLP scopes described in FDA, EPA and OECD were different to the extent necessary to reflect the agencies' different statutory responsibilities. As reflected by the definition of "nonclinical laboratory study," the FDA regulates studies conducted in the scope of health effects, while the EPA regulates studies conducted in the scope of health effects, chemical characterization, and environmental fate studies. As harmonized by OECD, nonclinical study

referred to an experiment or a set of experiments in which a test item is examined under laboratory conditions or in the environment to obtain data on its properties and/or its safety. Currently, GLP standards have been established across the world, making these standards a universal quality assurance system. [Table 1](#) illustrates the comparison of the definition of nonclinical laboratory study and scope from GLP regulations of FDA, EPA, and OECD.

GLP PRINCIPLES

General GLP principles include (1) organization, personnel, and responsibility; (2) the quality assurance program; (3) facilities; (4) equipment, reagents, and materials; (5) test systems; (6) test and reference items; (7) standard operating procedures; (8) study performance and reporting; and (9) the archives. With the more detailed requirement and development of electronic technology, more principles were implemented such as the principles to the organization and management of multi-site studies,^[5] to in vitro studies,^[6] to computerized systems,^[7] and for peer review of histopathology.^[8]

Table 1 Comparison of the Definition of Nonclinical Laboratory Study and Scope from GLP Regulations of FDA, EPA, and OECD

	FDA	EPA	OECD
Nonclinical laboratory study	Nonclinical laboratory study means in vivo or in vitro experiments in which test articles are studied prospectively in test systems under laboratory conditions to determine their safety.	Study means any experiment at one or more test sites, in which a test substance is studied in a test system under laboratory conditions or in the environment to determine or help predict its effects, metabolism, product performance (efficacy studies only as required by 40 CFR 158.640), environmental and chemical fate, persistence and residue, or other characteristics in humans, other living organisms, or media.	Nonclinical health and environmental safety study, henceforth referred to simply as "study," means an experiment or a set of experiments in which a test item is examined under laboratory conditions or in the environment to obtain data on its properties and/or its safety, intended for submission to appropriate regulatory authorities.
Scope	... for conducting nonclinical laboratory studies that support or are intended to support applications for research or marketing permits for products regulated by the Food and Drug Administration, including food and color additives, animal food additives, human and animal drugs, medical devices for human use, biological products, and electronic products.	(1) ... for conducting studies relating to health effects, environmental effects, and chemical fate testing. This part is intended to ensure the quality and integrity of data ... (2) ... applies to any study described by paragraph "a." of this section which ... (3) It is EPA's policy that all data developed under section 5 of TSCA be in accordance with provisions of this part.	... should be applied to the non-clinical safety testing of test items contained in pharmaceutical products, pesticide products, cosmetic products, veterinary drugs as well as food additives, feed additives, and industrial chemicals. ... include work conducted in the laboratory, in greenhouses, and in the field. ... apply to all nonclinical health and environmental safety studies required by regulations for the purpose of registering or licensing pharmaceuticals, pesticides, food and feed additives, cosmetic products, veterinary drug products and similar products, and for the regulation of industrial chemicals.

HISTORY OF GOOD CLINICAL PRACTICE REGULATION BY THE UNITED STATES, THE EUROPEAN UNION, JAPAN, CHINA, AND THE WORLD HEALTH ORGANIZATION

US Food and Drug Administration

The Federal Food, Drug, and Cosmetic Act, passed by Congress in 1938, gave authority to the FDA to oversee the safety of food, drugs, and cosmetics. The Public Health Service Act requires that a sponsor establish the safety and efficacy of biological products. These laws place on the FDA the responsibility for reviewing the sponsor's test results and determining whether the results establish the safety and efficacy of the product. If the agency accepts that safety and efficacy are adequately established, the sponsor is permitted to market the product.

At the end of the 1960s and the beginning of the 1970s, with a rapidly increasing use of chemical substances and the emergence of "contract research organizations" (CROs), tightened regulatory requirements especially for safety testing of pharmaceuticals became imperative. The "thalidomide disaster" and other safety cases are notable for their severe influence. Under the discrepancies between the data and conclusions submitted to the FDA, a series of hearings on the "Preclinical and Clinical Testing by the Pharmaceutical Industry" were started by US senate. Proposed regulations for GLP in nonclinical laboratory studies were then published by FDA on November 1976. After that, the FDA finalized GLP regulations and published them in the Federal Register on December 22, 1978 (21 CFR 58), and these GLP regulations came into effect on June 20, 1979.^[9]

The FDA performs two types of GLP inspections: routine inspections and "for-cause" inspections. The former are periodic evaluation of a laboratory's compliance with the GLP regulations on an at least once every 2 years basis, and the latter are invoked for one or more serious reasons as follows: deficiencies found in a routine inspection, concerns raised in the study reports submitted, validation of studies submitted, a third-party validation, verification, and investigation of questionable circumstances.^[10]

When failure of the facility to comply with one or more of the GLP regulations or other regulations published in Chapter 21 CFR, adverse effects on the validity of studies, or failure to achieve compliance with regard to lesser regulatory actions was inspected, the FDA would enforce the strategy of disqualification of a testing facility, in which case the facility fails to comply with the GLP regulations.

US Environmental Protection Agency

Confronting with the similar problems as FDA concerns, the EPA published its own GLP standards under the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA), which regulates pesticides and their safety, and

the Toxic Substances Control Act (TSCA), concerned with chemical substances. The EPA's FIFRA and TSCA GLP regulations were finalized on November 29, 1983, with the FIFRA GLP regulations codified as 40 CFR 160^[11] and the TSCA GLP regulations as 40 CFR 792.^[12] The Laboratory Data Integrity Branch is in charge of the direction of GLP programs, with the primary mission to assure the quality and integrity of studies submitted to the EPA and to assure compliance with the GLP regulations by conduction laboratory inspection and data audits.

The European Union

The first council directive issued in Europe is the 79/831/EEC of 18 September 1979 amending for the sixth time Directive, which regulated that the tests carried out shall comply with the principles of good current laboratory practice. Then in 1987 and 1988, a series of directives were published, the 87/18/EEC, 87/19/EEC, 87/20/EEC, and 88/320/EEC, which regulated that the safety evaluation of chemical substances, medicinal products, and veterinary products in compliance with GLP principle and under inspection and verification of GLP. After that, the two Directives 2004/9/EC and 2004/10/EC were issued, replacing Directives 87/18/EEC and 88/320/EEC, by the European Parliament and the Council of the European Union. These directives (as shown in Table 2) had adopted the principles and the revised OECD Guides for Compliance Monitoring Procedures for GLP.

The OECD took the task of the international harmonization of GLP standards, aiming to avoid nontariff barriers to trade in chemicals, promote mutual acceptance of data, and eliminate unnecessary duplication of labor. As stated by OECD, "to avoid different schemes of implementation that could impede international trade in chemicals, OECD member countries have pursued international harmonization of test methods and good laboratory practice."^[4] In 1979 and 1980, an international group of experts established under the Special Program on the Control of Chemicals developed the OECD Principles of Good Laboratory Practice, utilizing common managerial and scientific practices, and experiences from various national and international sources.

These principles of GLP were adopted by the OECD Council in 1981, as an Annex to the Council Decision on the Mutual Acceptance of Data in the Assessment of Chemicals [C(81)30(Final)].^[13] The OECD GLP regulations were revised and finally adopted on November 26, 1997 [C(97)186/Final].^[4] They were widely introduced into OECD countries. Moreover, many countries have developed GLP standard of their own based on OECD principles. The definition of GLP was first explicated to "nonclinical and environmental safety studies," and the scope was eliminated as "all non-clinical health and environmental safety studies required by regulations for the purpose of registering or licensing pharmaceuticals,

Table 2 Directives Adopted the Principles and the Revised OECD Guides for Compliance Monitoring Procedures for GLP

Council Directive	Contents
79/831/EEC	Council Directive 79/831/EEC of 18 September 1979 amending for the sixth time Directive 67/548/EEC on the approximation of the laws, regulations, and administrative provisions relating to the classification, packaging, and labeling of dangerous substances
87/18/EEC	Council Directive 87/18/EEC of 18 December 1986 on the harmonization of laws, regulations, and administrative provisions relating to the application of the principles of GLP and the verification of their applications for tests on chemical substances
87/19/EEC	Council Directive 87/19/EEC of 22 December 1986 amending Directive 75/318/EEC on the approximation of the laws of the Member States relating to analytical, pharmacotoxicological, and clinical standards and protocols in respect of the testing of proprietary medicinal products
87/20/EEC	Council Directive 87/20/EEC of 22 December 1986 amending Directive 81/852/EEC on the approximation of the laws of the Member States relating to analytical, pharmacotoxicological, and clinical standards and protocols in respect of the testing of veterinary medicinal products
88/320/EEC	Council Directive 88/320/EEC of 9 June 1988 on the inspection and verification of GLP
2004/9/EC	Directive 2004/9/EC of the European Parliament and of the Council of 11 February 2004 on the inspection and verification of GLP
2004/10/EC	Directive 2004/10/EC of the European Parliament and of the Council of 11 February 2004 on the harmonization of laws, regulations, and administrative provisions relating to the application of the principles of GLP and the verification of their applications for tests on chemical substances

pesticides, food and feed additives, cosmetic products, veterinary drug products and similar products, and for the regulation of industrial chemicals.”^[4]

Japan

The GLP regulations were issued by multiple agencies: Ministry of Health, Labour and Welfare,^[14,15] relating to pharmaceuticals, medical devices, and workplace chemicals; Ministry of Economy, Trade and Industry and Ministry of Environment, both relating to industrial chemicals; and Ministry of Agriculture, Forestry and Fisheries relating to pesticides, veterinary drugs, and feed additives,^[16] which were mainly based on Revised Guidance for the Exchange of Information concerning National Programs for Monitoring of Compliance with the Principles of Good Laboratory Practices.

China

The first GLP tentative regulation, “Guidelines of Drug Non-clinical Study Good Laboratory Practice (draft),” was promulgated by the State Science and Technological Commission (the former Ministry of Science and Technology of the People’s Republic of China). In 1993, the national drug administration system reformation occurred, which resulted in the alternation of law enforcement subject. The newly revised tentative guideline of GLP^[17] was then published by China Food and Drug Administration in 1999; final regulations such as “Guideline of Drug Non-clinical Study Good Laboratory Practice”^[18] were published in 2003, to ensure the validity and reliability of nonclinical laboratory studies in order to support the safety of CFDA regulation. The updated regulations were revised in 2017

recently.^[19] Other GLP regulations were issued by the Ministry of Agriculture of People’s Republic of China and the former Ministry of Environmental Protection of People’s Republic of China.

The World Health Organization

The World Health Organization (WHO)/Research and Training in Tropical Diseases (TDR) developed a GLP series in 2001, comprising a GLP Handbook as well as GLP training manuals, with the second edition of GLP Handbook issued in 2009 as well as that of training manuals issued in 2008. WHO/TDR GLP series were developed on the basis of OECD principles, in order to assist countries in conducting nonclinical research and drug development, especially aiming to empower disease endemic countries to develop and lead research activities at internationally recognized standards and quality, and anticipating that the use of GLP resources will help promote cost-effective and efficient preclinical research with a long-term positive effect on the development of products for the improvement of human health.^[20,21]

MUTUAL RECOGNITION

The objective of mutual recognition of GLP signed by two or more countries is to enable the reciprocal recognition of each country’s standard, protocol, test data, and conformity assessments of GLP program for the evaluation of safety, and inspectional procedures are mutually acceptable.

The divergence of technical requirements from country to country was such that industries (such as pharmaceutical industry, food industry, and cosmetic industry)

found it necessary to harmonize the regulation by achieving science-based mutual acceptance of research and development data. FDA has signed MOU or MRA with counterparties of Germany, Sweden, Switzerland, the Netherlands, Italy, Canada, and Japan.

As an international organization with more than 30 member countries, OECD adopted Mutual Acceptance of data (MAD),^[22] which states that data generated in any member country in accordance with OECD GLP principles shall be accepted in other member countries. The MAD system substantially saves governments and chemical producers every year from the expensive and labor-intensive testing.^[23] OECD also implemented nonmember adherents to the MAD system, for the OECD full adherent countries such as Argentina, Brazil, India, Malaysia, South Africa, and Singapore. For the emerging economies such as China, India, and Brazil, international recognized OECD GLP is a suitable template to achieve a global harmonized GLP standard by incorporation with the consideration and concerns of all stake holders.

LABORATORY ACCREDITATION QUALITY SYSTEM AND GLP

The standards of quality system are implemented to quality management of laboratory, aiming at improving the ability to consistently produce valid results. Such accreditation system as ISO/IEC 17025 specifies the general requirements for the competence to carry out tests and/or calibrations. These standards are widely used as a quality system in environmental, food, chemical, and clinical testing laboratories. Testing laboratories seek accreditation by such quality systems to demonstrate their competence.

The most distinctive difference between GLP and laboratory quality system lies in the regulatory control mechanism. GLP compliance is written into law, regulating the studies that nonclinical health and environmental safety studies intended for regulatory submission. To comply with GLP regulations is to assure that work is being performed by qualified personnel, under appropriate direction, in adequate facilities, with calibrated and maintained equipment, using standard operating procedures, documenting raw data, having in-process work inspected, and providing reports that are reviewed by a QA professional.^[9] To this extent, GLP is a quality system aiming at guaranteeing neither scientific validity of method nor the ability to generate accurate and precise measurements.

GOOD STATISTICS PRACTICE AND GLP

GLP principles provide a quality system that ensures the studies are planned, performed, monitored, recorded, archived, and reported under quality assurance. Nevertheless, it is the good statistics practice (GSP) that plays

a key role in many aspects such as scientific and reliable study design, sampling plan, data analysis, instrument calibration, and decision rules. Statistical considerations were important for conducting safety studies, such as sampling plan, conventional statistical power, significance, critical study endpoint, and validation of the study. With the tendency of ever-increasing collaboration with multiple site or laterals, and explosive growing data capacity, statistics was driven to take more and more important part in the design of study plan, the analysis of data, risk assessment, and scientific decision rules.

CONCLUSION

The principles of GLP promote the quality and validity of data generated in the testing of chemicals and prevented fraudulent practices. As a common quality standard for nonclinical safety studies on human and environment safety assessment, GLP has been globally accepted and played an important role over decades of practice. The good statistics practice was implemented in GLP providing a valid and fair assessment of the study conducted.

GLP, ISO 17025, WHO/GPCL and GMP share some essential universal principles: Say what you do and do what you say. With the continuous deepening of economic globalization and fast speed of new technology development, there are increasing demands and challenges in the international harmonization, scientific regulatory technical requirements, and modern technology innovations. In order to promote the market approval procedure for new products, more and more national regulatory authorities have signed the MOU or MRA on GLP. International recognized GLP shall play a key role in global harmonization process by incorporation with the consideration and concerns of all stake holders.

REFERENCES

1. Testing Laboratory Registration Council. *Testing Laboratory Registration Act 1972, Public Act 1972 No 36*; Wellington, New Zealand, 1972.
2. FDA. *Code of Federal Regulation 21, PART 58—Good Laboratory Practice for Nonclinical Laboratory Studie*; FDA: Rockville, MD. <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?cfrpart=58> (accessed Aug 2017).
3. EPA. *Good Laboratory Practices Standards Compliance Monitoring Program*; United States Environmental Protection Agency. <https://www.epa.gov/compliance/good-laboratory-practices-standards-compliance-monitoring-program> (accessed Aug 2017).
4. OECD Principles on Good Laboratory Practice. *OECD Series on Principles of Good Laboratory Practice and Compliance Monitoring—Number 1, Environment Directorate Organization for Economic Co-operation and Development*; Paris, 1998.

5. OECD Principles on Good Laboratory Practice. *OECD Series on Principles of Good Laboratory Practice and Compliance Monitoring—Number 13, The Application of the OECD Principles of GLP to the Organisation and Management of Multi-Site Studies*; Paris, 2002.
6. OECD Principles on Good Laboratory Practice. *OECD Series on Principles of Good Laboratory Practice and Compliance Monitoring—Number 14, The Application of the Principles of GLP to In Vitro Studies*; Paris, 2004.
7. OECD Principles on Good Laboratory Practice. *OECD Series on Principles of Good Laboratory Practice and Compliance Monitoring—Number 17, Advisory Document of the Working Group on Good Laboratory Practice. Application of GLP Principles to Computerised Systems*; Paris, 2016.
8. OECD Principles on Good Laboratory Practice. *OECD Series on Principles of Good Laboratory Practice and Compliance Monitoring—Number 16, Advisory Document of the Working Group on Good Laboratory Practice, Guidance on the GLP Requirements for Peer Review of Histopathology*; Paris, 2014.
9. Jury, P. S. *Good Laboratory Practice the Why and the How*; Springer-Verlag: Berlin, Heidelberg, New York, 2001.
10. Weinberg, S. *Good Laboratory Practice Regulations*, 4th Ed.; Tikvah Therapeutics, Inc.: Atlanta, 2007.
11. Protection of Environment. *Code of Federal Regulations. Title 40. Chapter I—Environmental Protection Agency, Part 160—Good Laboratory Practice Standards*; 1999.
12. Protection of Environment. *Code of Federal Regulations. Title 40. Chapter I—Environmental Protection Agency (continued), Part 792—Good Laboratory Practice Standards*; 1999.
13. OECD Council. *Council Decision Amending Annex II to the Council Decision Concerning the Mutual Acceptance of Data in the Assessment of Chemicals [C(81)-30(FINAL)]*; Organisation for Economic Co-operation and Development: Paris, 1997.
14. MHLW. *Ministerial Ordinance on Good Laboratory Practice for Nonclinical Safety Studies of Drugs*; the Minister of Health, Labour and Welfare. Ordinance of the Ministry of Health and Welfare No.21 of March 26, 1997 (as last amended by the Ordinance of the Ministry of Health, Labour and Welfare No.114 of June 13, 2008).
15. MHLW. *Ministerial Ordinance on Good Laboratory Practice for Nonclinical Safety Studies of Medical Devices*; the Minister of Health, Labour and Welfare. Ordinance of the Ministry of Health, Labour and Welfare No.37 of March 23, 2005 (as last amended by the Ordinance of the Ministry of Health, Labour and Welfare No.115 of June 13, 2008).
16. MAFF. *Notification on the Good Laboratory Practice—The notification from the Director-General of Agricultural Production Bureau, the Ministry of Agriculture, Forestry and Fisheries*; Ministry of Agriculture, Forestry and Fisheries, October, 1999.
17. CFDA. *The Guidelines of Drug Non-Clinical Study Good Laboratory Practice (Tentative Guideline)*; China Food and Drug Administration. <http://www.sda.gov.cn/WS01/CL0053/24492.html> (accessed Aug 2017).
18. CFDA. *The Guidelines of Drug Non-Clinical Study Good Laboratory Practice (CFDA No.2)*; China Food and Drug Administration. <http://www.sda.gov.cn/WS01/CL0053/24472.html> (accessed Aug 2017).
19. CFDA. *The Guidelines of Drug Non-Clinical Study Good Laboratory Practice (CFDA No.34)*; China Food and Drug Administration. <http://www.sfda.gov.cn/WS01/CL0053/175598.html> (accessed Aug 2017).
20. WHO. *Handbook: Good Laboratory Practice (GLP): Quality Practices for Regulated Non-clinical Research and Development*, 2nd Ed.; The Special Program for Research and Training in Tropical Diseases (TDR)/World Health Organization, Geneva, 2009.
21. WHO. *Good Laboratory Practice Training Manual for the Trainee: A Tool for Training and Promoting Good Laboratory Practice (GLP) Concepts in Disease Endemic Countries*, 2nd Ed.; The Special Program for Research and Training in Tropical Diseases (TDR)/World Health Organization, Geneva, 2008.
22. Organization for Economic Co-operation and Development (OECD). *Mutual Acceptance of Data (MAD)*. <http://www.oecd.org/env/ehs/mutualacceptanceofdatamad.htm> (accessed Aug 2017).
23. Organization for Economic Co-operation and Development (OECD). *Non-member Adherents to the OECD System for Mutual Acceptance of Chemical Safety Data*. <http://www.oecd.org/env/ehs/non-member-adherens-to-oecd-system-for-mutual-acceptance-of-chemical-safety-data.htm> (accessed Aug 2017).

Good Programming Practice (GPP)

Aileen L. Yam

Independent Consultant

Abstract

This entry presents ideas on the fitness, robustness, remodeling friendliness, extendibility, and economy of program construction. The importance of documenting, organizing, validating, and archiving programs, as well as the need to keep updated and to have a cooperative spirit in a shared project, are highlighted.

Keywords: Coding; Designing; Good practice; Programming; Quality; Validation.

INTRODUCTION

The operation of a biostatistics group in the drug development process is analogous to the process of building a house. Data coordinators prepare building blocks and lay a solid foundation. Statisticians are the architects with overall responsibility for the conception and realization of the project. Programmers are the construction engineers who execute the statisticians' plan and transform building materials into a house. In other words, a synergy of efforts among data coordinators, statisticians, and programmers is necessary. Data coordinators ensure the integrity of the data. Statisticians do detailed analysis of the drug, choose appropriate statistical methods, write specifications, and analyze results. Programmers implement statistical rules to convert clinical data into useful information for assessing the safety and efficacy of a drug.

This entry presents ideas on the fitness, robustness, remodeling friendliness, extendibility, and economy of program construction. The importance of documenting, organizing, validating, and archiving programs, as well as the need to keep updated and to have a cooperative spirit in a shared project, are highlighted.

requirements. Quality needs to be defined within business concerns, corporate and industry culture, on the basis of the expected life span of the programming code, and, sometimes, on a project-by-project basis.

In general, quality for programming in the pharmaceutical industry refers to *fitness* (whether the programs meet functional requirements and are suitable for the purpose for which they are intended), *robustness* (accurate and reliable under different programming conditions and data scenarios), *remodeling friendliness* (easy to read, understand, and maintain), *extendibility* (re-usable and automatable to handle multiple studies of the same drug), and *economy* (on time and within budget).

With the stringent requirements, the emphasis here is on programming techniques and processes that build in quality to do things right the first time.

The SAS® System is the most commonly used software in clinical trials so “SASese” is inevitably spoken. However, the general principles discussed within the context of programming in SAS can be applied regardless of the programming tool or language selected. Analogies and simple programming examples are given to make the concepts behind the principles more concrete.

QUALITY IS IN THE EYE OF BEHOLDER?

The goal of good programming practice is quality work. Quality is an elusive term. Quality as practiced may sometimes be different from quality as idealized in theoretical discussions. There are constraints to balance among available resources, budgets, schedules, and programming

DESIGNING

Just as building a house needs blueprints, designing is an important prerequisite for good programming practice before actual coding starts. Winograd et al.^[1] stressed the importance of bringing design to programming. Their book is based on a workshop on software design. Two software developers at the workshop listed the components in the development of designs as understanding, abstracting, structuring, representing, and detailing. The crucial role of design as a counterpart of programming in clinical trials is very similar.

Note: The SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Understanding involves learning the nature of the drug and the protocol as well as comprehending the analysis specification. What are the rules? Are the rules appropriate, feasible, sufficient, or contradictory? Understanding also means having a good grasp of the data structure, formats, and categories. Are the data standardized across studies? Do the data match the case report form or the protocol requirements?

Abstracting means sketching out the main elements for programming. Are all the requirements covered? Can a consistent vision of all the elements in the whole project be built in, from Investigational New Drug (IND) to New Drug Application (NDA) or Biologic License Application (BLA)?

Structuring entails the ordering of relationships among the elements. Is there logical, functional, sequential, or procedural coherence? Is the layout elegantly clear?

Representing refers to representing the elements in computer language. What are the steps or procedures selected? Are the selections relevant? Will they work in various situations? Do the methodologies conform to the Clinical Data Interchange Standards Consortium (CDISC) standards^[2] and Food and Drug Administration (FDA) requirements?^[3]

Detailing is the careful attention to the fine points of programming. Are the approaches lean, simple, and straightforward? How can programs be written so that they are easy to debug, modify, maintain, and enhance? What are the goals of a specific program? Are the goals to build applications or systems where continuous fine-tuning and addition of features are needed? Or are the goals to write good enough programs for basic operation? What is seen as indispensable in some situations can be unnecessary overhead in other situations.

The process of designing fosters communication among programmers, statisticians, clinicians, and data coordinators, and encourages different perspectives of solutions to problems. Good designs clarify a program and convey the content. Good programmers are designer-programmers. The design approach selected has to be tailored to project needs. Design deals with the issue of whether the programs are *fitting* and appropriate to address the statistical questions of the drug under study.

CODING

In addition to good designs, quality engineering is needed. Good programming practice as defined and illustrated herein is organized around the goals of robustness, remodeling friendliness, extendibility, and economy. Not all of these goals are completely compatible. For example, making a program more robust may add complexity to a program. The pragmatic challenge lies in prioritizing and finding a happy medium among these quality goals.

ROBUSTNESS

Robustness is the most essential characteristic in the construction of a quality program. Robustness encompasses accuracy, reliability, thoroughness, and reproducibility. As a general rule, making a program robust means writing solid code and having a good understanding of how to effectively prevent bugs and how to easily or automatically detect bugs.^[4]

Here is an example of built-in robustness. If you need to create a response flag that indicates an average increase of 1% from baseline over 3 years at each of the scheduled 6-month intervals, you can code the program as follows:

```
if month = 6 and pchange = .50 then response = 1;
else if month = 12 and pchange = 1.0 then response = 1;
else if month = 18 and pchange = 1.5 then response = 1;
else if month = 24 and pchange = 2.0 then response = 1;
else if month = 30 and pchange = 2.5 then response = 1;
else if month = 36 and pchange = 3.0 then response = 1;
else response = 0;
```

If subjects do not follow the study protocol and come in at unscheduled time points or after 3 years, or if there are missing data, the preceding code will not work properly. A more robust way to write the code is:

```
if pchange = month / 12 then response = 1;
else if pchange > 3.0 then response = 0;
```

Not all the programming code can be made robust and elegant simultaneously as effortlessly as this example. It sometimes depends on the level of over-engineering expended to check in every conceivable way and at every conceivable place.

What follows is a discussion of the various aspects of strengthening robustness.

Syntax errors, uninitialized variables, automatic data type conversions, division by zero, unknown formats, undeclared array elements, and merging with repeats of key variables are some of the ordinary types of bugs that usually exist in the early stage of programming development. They are the easiest to locate and fix. Compilers usually give error or warning messages on these kinds of errors.

Logic errors are a less obvious kind of bugs that inhibit the proper function of a program. They are less obvious because the program is syntactically correct but may not give the correct result. An example of logic error is referring to the first or the last observation that has just been deleted. It is very important to test programming code in small chunks. It is also important to know the assumptions behind the design of the program, and to build in

statements to output warning messages if assumptions are violated. Bugs multiply quickly. Any bug found needs to be immediately fixed, especially at a step that supplies data to another step.

A common type of bug is introduced by data. A program may not have accounted for all the data scenarios or for unexpected data errors. For example, the studies of a drug are conducted in both the United States and Europe, and if a programmer expects only inches and centimeters in the conversion of height values, the program will miss the height values that are in meters or have missing units. One way to avoid this kind of bug is to do a frequency check on the unit categories and to write the program based on the frequency results. But the studies may not have been completed when the program is written and new data may create bugs in the program. A more reliable way is to build routines into the program to flag unexpected data values and to output warning messages, or to capture those values and print them out. It also helps to draw a Venn diagram or to sketch out a list to see if all possible combinations of conditions are covered.

Missing values are a type of bug most often neglected by programmers. A simple example is to write a condition for test results of less than 5, to indicate significant improvement; subjects with missing test results end up in the “significant improvement” category:

```
if score < 5 then improve = 'SIG';
elseif score >= 5 then improve = 'NOT SIG';
```

This code is syntactically correct, but there are two other problems besides assigning missing values to the wrong category. First, the length of the variable *improve* is not defined. Its length is defined by the first IF statement. Thus the value of *improve* in the second IF statement is truncated to 3. Second, invalid scores, such as negative scores or zeroes, are not accounted for. One way to prevent the bugs is to change the code to the following:

```
length improve $7;
if score = . then improve = '';
else if 1 <= score < 5 then improve = 'SIG';
else if score >= 5 then improve = 'NOT SIG';
else put 'DATA ERROR: ' SUBJECT = 'subjid'
      'SCORE = ' score;
```

One other kind of data-related bug comes from merging or concatenating data. It is necessary to examine database structures before combining the data, to sort the data and join them by key variables, to keep only the variables in each data set that are necessary in order to prevent variables of the same name from being overwritten in merging, to add debugging code to output warning if the data cannot be combined properly, and to verify the number of observations before and after the data are combined.

In clinical trials, because NDA or BLA submissions usually require the integration of information from several studies of a potential new drug, maintaining internal consistency is a very important aspect of ensuring robustness. Consistency is needed across all studies of the same drug in naming variables, in choosing data type, in formatting and labeling variables, in assigning the values of categorical variables, and in structuring derived data. Here are some reasons why consistency is important:

1. If a character variable is assigned different lengths in different studies, the values of the character variable are truncated when data from the studies are concatenated.
2. If the same variable is given different names in different studies, proper integration of data is not possible until the variable is renamed to the same name in all studies.
3. The effects of inconsistency are subtle; usually there is no error message. It is necessary to adopt a uniform approach for all studies to rule out unexpected errors.

Numeric representation errors, also known as rounding errors, are one type of bug that can crawl into programming code without notice. Fractional values and values that do not have enough storage space have the potential for numeric representation problems. Numeric representation errors propagate. The more computation the values go through, the larger the errors will become. An example of the occurrence of numeric representation errors in programming laboratory data is the comparison of laboratory values to normal ranges, where both the laboratory values and the normal ranges are fractional values. Here are some major ways to minimize numeric representation problems in SAS software:

1. Only use the LENGTH statement or the ATTRIB statement to truncate values when disk space is limited.
2. Use integers if possible.
3. Use the ROUND function or the FUZZ function on fractional values before doing a comparison on the values.

Related to numeric representation errors is mixed variable types that result in automatic data type conversions. Avoid comparing numeric variables with character variables because the comparison figures out the conversion type, converts one type into another, does rounding, and determines the answer.

Hard-coding is a cardinal sin. Hard-coding is usually an attempt to programmatically fix data errors. Hard-coding disrupts the integrity of data and the replication of results. It also undermines the robustness of a program. Data errors are best handled at the source, i.e., in the database.

Using available formats, functions, and routines from software is a way to prevent bugs. The software is usually thoroughly tested before it is released. In SAS software, for example, directly outputting p -values from a procedure into a data set is easier and less error-prone than writing a program to extract the p -values. There is also less chance for error using formats such as PVALUE and PERCENT than building the formats with FORMAT statements. The PVALUE format automatically transforms values of less than zero to the standard SAS output p -value format. The PERCENT format can automatically take care of rounding.

REMODELING FRIENDLINESS

Change is a fact of life in programming for clinical trials. Therefore making a program easy to understand, easy to use, easy to modify, and easy to debug is vital. Some requests for changes are made for cosmetic reasons. Most cosmetic changes are like repainting a house: they are simple and they rarely affect original program design. Requirement or analysis changes are usually structural changes; the sections that are intricately interwoven may need major rework. Changes also arise from error correction.

No matter how big or small changes are, each change requires review of the changed code, testing of the impact of the changes on the whole program, and revalidation and sometimes rewriting of all the results. It is better to funnel all the change requests through lead statisticians or programming managers who are familiar with the intent and the budget of the studies than to make changes on demand. The leads will coordinate the changes where ideas for changes contradict among clinicians, data coordinators, statisticians, medical writers, and the regulatory group. The leads will also separate the wheat from the chaff change requests and will estimate and make sure everyone knows the cost of each change.

To insulate a program from the effects of remodeling, the following are some guidelines to write or revise programs by anticipating and documenting changes:

1. Write simple and straightforward code in a clear layout, and have sufficient comments so that any programmer can understand and modify the code with little chance of introducing bugs.
2. Comply with your company's programming conventions; they are tools to help communicate the nature of a program or a variable. For example, all program names that start with the letter m are macro programs, or all variables that start with an underscore are array variables.
3. Keep the initialization and the reference of variables close together so that they are easy to locate.
4. Do not nest too deep. Cut down interrelated parts to minimize the impact of changes.

5. Global macro variables are convenient to access, but the values of the global macro variables may be inadvertently changed in a program, whereas local macro variables that are confined to a small section of code are easier to maintain.
6. Modularize parts that are likely to change.
7. When there are extensive programming structural changes, it is sometimes easier to rewrite a program than to modify an old one.
8. Avoid tricks; the time spent writing and documenting tricky code is better spent in writing plain code that is easy to maintain.
9. Make changes that are compatible with the existing program design.
10. Adhere to your company's system for documenting and tracking changes.

EXTENDIBILITY

In clinical trials, usually multiple studies are performed on the same drug. A major goal in programming is to attain extendibility. Extendibility here means that the programs are flexible in accommodating varying number of treatments, visits, hours, etc., and are re-usable for many studies. A basic example of extendibility is to make study number a macro variable so that programs can be used for any study with the same data structure and analysis requirements by reassigning the value of the macro variable.

The advantages of extendibility are that reusing the code makes the output identical and that the value of the code increases if the code is re-usable for multiple studies of one or more drugs.

Some of the techniques for building extendible programs are: look for patterns in data and analysis; design easily shared code; make use of macros and parameter-passing mechanism; avoid data-driven coding; modularize parts that are different and make it easy to isolate and replace a section of code; use local macro variables when possible so that it is easy to pull out some parts of the code and plug them into another program.

There is a tradeoff between generality and specificity. The more general and flexible the design is, the harder it is to write the program and to detect bugs, and sometimes the harder it is to use or to modify.

It is usually easier to write general programs for safety analyses than for efficacy analyses because safety analyses tend to be similar, except for the number of treatment groups or visit numbers, while efficacy analyses tend to be substantively different.

ECONOMY

Conventional wisdom in programming defines economy as leanness of programming code. Each and every piece

of code is important and cannot be taken away. In a more practical sense from the viewpoint of research on drugs, economy means completing the project and being able to submit NDAs or BLAs on time and within budget.

Economy is often discussed in terms of computer efficiency and human efficiency. Computer efficiency means minimal use of system resources, such as execution time, input-output time, memory, and data storage. Human efficiency means less programming time spent in designing, coding, validating, documenting, and maintaining.

Some techniques for ensuring computer efficiency are: order IF conditions by decreasing frequency of occurrence; eliminate unnecessary input and output operations; minimize macro reference; compile frequently used macros; use random access (WHERE condition) instead of sequential access (IF condition); use the fewest array dimensions possible; and automate repetitive tasks.

Here is an example of human efficiency in creating a variable called *block* where subjects 1–10 are in block 1, subjects 11–20 are in block 2, and subjects 21–30 are in block 3:

$$\text{block} = 1 * (1 \leq \text{subject} \leq 10) + 2 * (11 \leq \text{subject} \leq 20) + 3 * (21 \leq \text{subject} \leq 30);$$

There is only one statement to type. The tradeoffs are that it is hard to maintain and may lack robustness. The code is fine in this case, because randomization schemes for assigning subjects to blocks are usually established at the beginning of a study and seldom change.

In clinical trials, good programming design, careful algorithm selection, the prevention of errors, and the writing of clean code that is easy to maintain contribute to economy. Having a few dedicated programmers on a project, and not rotating programmers in and out of projects, saves time and increases quality because each additional person added to a project has a learning curve. In addition, economy means joint efforts among data coordinators, statisticians, programmers, and medical writers to increase productivity. Clean and complete data and less rework from rule changes are important. Locate data problems and promptly inform data coordinators to rule out programming errors. Work with statisticians in early stages to reduce the amount of time spent in requirement changes or debugging. If each and every person makes an effort to do the job right the first time, economy is built into the process. It also spares the medical writers from having to rewrite results because of changes or errors.

Economy is integral to the overall program quality but it is not the most critical condition. Priorities need to be established if economy conflicts with other goals, especially robustness.

DOCUMENTATION AND ORGANIZATION

Documentation and organization are controversial topics because they are products of programming style. Rigid

standards cannot be imposed without contention; the best way is to work out a team style. Style is not just a matter of aesthetics; good style communicates content. The essence of good style is discussed here.

Documentation in programming is generally known as comments. Effective comments are not the comments that simply repeat the code; they should explain the objectives at a higher level of abstraction than the code.^[5] Code-level comments can be replaced by good programming style, which includes meaningful variable names, descriptive labels, easily understood logic, well-thought-out design, clear layout, and naming conventions that distinguish among constants, temporary variables, local or global macro variables, loop index variables, and array variables. Good style documents code without extensive comments. The best documentation is self-documenting code. Comments are expensive, not only to write but also to maintain. Therefore comments need to be accurate, brief, and to the point.

Organization refers to the formatting of the code. Good organization provides clues to the logic of a program. Some of the techniques of good organization include: group related topics together; use alignment, white space, blank lines, indentation, blocks; use parentheses to clarify how expressions are evaluated; use a minimum of nesting; be consistent throughout the whole project; make the code easy to read and maintain. The purpose of organization is to make the visual structure match the logical structure of the code.

VALIDATION

The FDA's definition of process validation^[6] is: "Establishing documented evidence that provides a high degree of assurance that a specific process will consistently produce a product meeting its predetermined specifications and quality attributes."

Every company needs to establish a checklist of validation procedures that reflect the principles of the FDA's requirements. Validation is making sure that the code does what it purports to do and that it is stable and reliable. The emphasis is on documentation. If there is no written document, there is no proof of validation. The validation procedures have to be understood and followed.

Validation is costly and is potentially endless. Good programming practice at every development step is essential to ensure high quality, regardless of the amount of final testing and inspection. It is hard to define what constitutes meaningful validation, but here are some ways to produce bug-free code:

1. Design programs to include debugging aids during the development phase and to exclude debugging aids during the production phase.
2. Test every step of programming logic.

3. Try out various data scenarios.
4. Create test data with errors to see if the code breaks down.
5. Devise a group review and change control process.
6. Thoroughly review initial program development and changes.
7. Conduct independent testing.

Most of the testing procedures can prove only the presence, but not the absence, of bugs. Doing a combination of various validation approaches increases the chances of finding bugs. If budget and resources are limited, fixing problems is more important than fixing symptoms. All the test results, the comments from table and listing review by statisticians, data coordinators, clinicians, medical writers, and the regulatory group need to be kept as proofs of validation.

SECURITY PLAN

Many factors can contaminate a program. A computer virus can alter the behavior of a program. A person can accidentally delete or modify the wrong program. Power failure may cause loss of information. A disk may crash. A hard drive may be accidentally reformatted.

Most companies have standard operating procedures (SOPs) for storage, archival, and retrieval. The keys are to have appropriate antivirus software and procedures in place, to have password authentication or audit trail controls, to periodically backup, to have off-site storage, to test the backup recovery procedures, and to have contingency plans for conducting the studies in the event of disasters.

KEEPING UPDATED

Programming is a fast-paced occupation. The SAS software and other application software frequently come out with updated versions with new and improved features. Keeping updated with the technology enhancements not only helps to streamline programming but also allows greater access to certain statistics, better control of output, more appealing presentation of results, and easier data access and information sharing across environment.

It is also necessary to keep up with the FDA's ongoing efforts to improve the review process. Some of the major efforts are *standardized* electronic submissions, the adoption of universal file formats such as using Adobe Acrobat software for transferring documents across platforms, or using Extensible Markup Language (XML) for sharing information and moving data among diverse computer types and operating systems. CDISC (Clinical Data Interchange Standards Consortium) standards have

been recommended for FDA regulatory submissions since 2004 and continue to be developed and improved over time. CDISC provides global consensus-based standards from data collection (CDASH) to submissions with Study Data Tabulation Model (SDTM) or Analysis Data Model (ADaM) as well as metadata for preparing define.xml (CRT-DDS). Thanks to CDSIC standards, programming code can be made more re-usable and validation by double-programming becomes easier.

The Windows point-and-click or drag-and-drop environment makes programming easier and provides multiple options to perform the same task. At the same time, programming can be more error-prone in such environment. A single movement of the mouse button can delete a whole directory. It is crucial to select the correct variables and the correct directory and to define standards common to all programmers. For dynamic reports or graphs, make sure they always reflect the most current data available. Here are the important steps for building an application in an interactive Windows environment:

1. Determine the variables and valid ranges for variable values.
2. Determine the necessary sequences.
3. Group the sequences according to function.
4. Use descriptive labels, icons, pushbuttons, or other graphical elements.
5. Maintain internal consistency in the interface design, work with users to make the interface user-friendly, and work with other applications programmers to jointly set up design standards.
6. Make sure numeric and logical operations are correct.
7. Test the individual sequences.
8. Link the sequences and test the entire application.
9. Add code to capture errors and make a graceful exit instead of letting the application crash.
10. Document the application.

PROGRAMMING ETIQUETTE ON A SHARED PROJECT

Beyond good coding disciplines, the remaining programming virtue is attitude. Programmers usually share different aspects or different studies of programming a clinical trial. A cooperative and understanding attitude facilitates the sound and speedy completion of a project.

Standards and conventions are communication tools. Abide by the CDISC standards and company conventions so the whole team speaks a common language. Write readable, easily shared, and re-usable code whenever possible, keeping in mind that others may have to maintain, modify, or validate the code. Clean up test programs, or

stack the most recent test on top of the previous ones in the same program, and comment out the previous tests. Putting all test programs in one file not only makes the programming area cleaner; it also makes it easy to locate previous tests.

Make the project a good learning experience for everyone. Be tolerant of individual approaches, keep an open mind, and learn from one another. To err is human. We learn not only from good programming tips and techniques but also from making mistakes.

CONCLUSION

In programming, the chain is only as strong as its weakest link, so make good programming practice a habit.

REFERENCES

1. Winograd, T.; Bennett, J.; DeYoung, L.; Hartfield, B. *Bringing Design to Software*; ACM Press: New York, 1996, xiii–xxv, 46–47.
2. <http://www.cdisc.org> for publications on global clinical data and submission standards.
3. <http://www.fda.gov> for guidance documents on submission.
4. Maguire, S.A. *Debugging the Development Process*; Microsoft Press: Redmond, WA, 1994, 31–32.
5. McConnell, S.C. *Code Complete: A Practical Handbook of Software Construction*; Microsoft Press: Redmond, WA, 1993, 453–491.
6. Stokes, T.; Branning, R.C.; Chapman, K.G.; Hambloch, H.; Trill, A.J. *Good Computer Validation Practices: Common Sense Implementation*; Interpharm Press: Buffalo Grove, IL, 1994, 76.

Good Statistics Practice (GSP)

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

The purpose of good statistics practice is not only to minimize bias but also to minimize variability that may occur before, during, and after the conduct of the studies. This entry discusses the role of statistics in drug development and their importance in ensuring the effectiveness and safety of a new drug.

INTRODUCTION

Good statistics practice (GSP) in pharmaceutical research and development is defined as a set of statistical principles and/or standard operating procedures for the best biopharmaceutical practices in design, conduct, analysis, evaluation, reporting, and interpretation of studies at various stages of pharmaceutical research and development.^[1–3] The purpose of GSP is not only to minimize bias but also to minimize variability that may occur before, during, and after the conduct of the studies. More importantly, GSP provides a valid and fair assessment of the drug product under study. The concept of GSP in drug research and development can be seen in many regulatory requirements, standards/specifications, and guidelines/guidances set by most health authorities, such as the U.S. Food and Drug Administration (FDA) and the Committee for Proprietary Medicinal Products (CPMP) in the European Community.^[4] For example, the U.S. regulatory requirements for drug research and development are codified in the U.S. Code of Federal Regulations (CFR), while the U.S. Pharmacopeia and National Formulary (USP/NF) includes standard procedures, test and sampling plans, and acceptance criteria and specifications of many pharmaceutical compounds. In addition, the FDA also develops a number of guidelines and guidances to assist the sponsors in drug research and development. These guidelines and guidances are considered gold standards for achieving good laboratory practice (GLP), good clinical practice (GCP), current good manufacturing practice (cGMP), and good regulatory practice (GRP). The concept of GSP is well outlined in the guideline on *Statistical Principles for Clinical Trials* recently issued by the International Conference on Harmonization.^[5] As a result, GSP not only provides accuracy and reliability of the results derived from the studies, but also ensures the validity and integrity of the studies. Note that in February 2005, the statistical Program Committee (CPE) of European Community adopted the European Statistics Code of good practice and undertook to observe the fifteen principles for GSP in Europe.

ROLE OF STATISTICS IN DRUG DEVELOPMENT

In the research and development of a drug product, statistics are necessarily applied at various critical stages to meet regulatory requirements for the effectiveness, safety, identity, strength, quality, purity, and stability of the drug product under investigation. These critical stages include pre-IND (investigational new drug application), IND, NDA (new drug application), and post-NDA. The roles of statistics at these critical stages are briefly described later.

At the very early stages of pre-IND, pharmaceutical scientists may have to screen thousands of potential compounds in order to identify a few promising compounds. An appropriate use of statistics with efficient screening and/or optimal designs will assist pharmaceutical scientists to identify the promising compounds within a relatively short time frame and cost effectively.

As indicated by the FDA, an IND should contain information regarding chemistry, manufacturing, and controls (CMC) of the drug substance and drug product to ensure the drug identity, strength, quality, and purity of the investigational drug. In addition, the sponsors are required to provide adequate information about pharmacological studies for absorption, distribution, metabolism, and excretion (ADME) and acute, sub-acute, and chronic toxicological studies and reproductive tests in various animal species to show that the investigational drug is reasonably safe to be evaluated in clinical trials in humans. At this stage, statistics are usually applied to (i) validate a developed analytical method, (ii) establish a drug expiration dating period through stability studies, and (iii) assess toxicity through animal studies. Statistics are necessarily applied to meet standards of accuracy and reliability.

Before the drug can be approved, the FDA requires that substantial evidence of the effectiveness and safety of the drug be provided in the Technical Section of Statistics of an NDA submission. Because the validity of statistical inference regarding the effectiveness and safety of the drug is always a concern, it is suggested that a careful review be

performed to ensure an accurate and reliable assessment of the drug product. In addition, to have a fair assessment, the FDA also establishes advisory committees, each consisting of clinical, pharmacological, and statistical experts and one advocate (not employed by the FDA) in designated drug classes and subspecialties to provide a second but independent review of the submission. The responsibility of the statistical expert is not only to ensure that a valid design is used but also to evaluate whether statistical methods used are appropriate for addressing the scientific and medical questions regarding the effectiveness and safety of the drug.

After the drug is approved, the FDA also requires that the drug product be tested for its identity, strength, quality, and purity before it can be released for use. For this purpose, the current good manufacturing practice is necessarily implemented to (i) validate the manufacturing process, (ii) monitor the performance of the manufacturing process, and (iii) provide quality assurance of the final product. At each stage of the manufacturing process, the FDA requires that sampling plans, acceptance criteria, and valid statistical analyses be performed for the intended tests, such as potency, content uniformity, and dissolution.^[6] For each test, sampling plan, acceptance criteria, and valid statistical analysis are crucial for determining whether the drug product passes the test based on the results from a representative sample.

STATISTICAL PRINCIPLES

In this section we will discuss some key statistical principles in the design and analysis of studies that may be encountered at various stages of drug development.

Bias and Variability

For approval of a drug product, regulatory agencies usually require that the results of the studies conducted at various stages of drug research and development are accurate and reliable to provide a valid and fair assessment of the treatment effect. The accuracy and reliability are usually referred to as the *closeness* and the *degree of closeness* of the results to the true value (i.e., true treatment effect). Any deviation from the true value is considered a bias, which may be due to selection, observation, or statistical procedures. Pharmaceutical scientists would make any attempt to avoid bias, whenever possible, to ensure that the collected results are accurate.

The reliability of a study is an assessment of the precision of the study, which measures the degree of the closeness of the results to the true value. The reliability reflects the ability to repeat or reproduce similar outcomes in the targeted population. The higher precision a study is, the more likely it is that the results would be reproducible. The precision of a study can be characterized by the variability incurred during the conduct of the study.

In practice, because studies are usually planned, designed, executed, analyzed, and reported by a team that consists of pharmaceutical scientists from different disciplines, bias and variability inevitably occur. Possible sources of bias and variability should be identified at the planning stage of the study not only to reduce the bias but also to minimize the variability.

Confounding and Interaction

In drug research and development, there are many sources of variation that have an impact on the evaluation of the treatment. If these variations are not identified and properly controlled, then they may be mixed up with the treatment effect that the studies are intended to demonstrate. In this case, the treatment is said to be confounded with the effects due to these variations. For a better understanding, consider the following example. Last winter, Dr. Smith noticed that the temperature at the emergency room was relatively low, which had caused some discomfort among medical personnel and patients. Dr. Smith suspected that the heating system might not function properly and decided to improve it. As a result, the temperature of the emergency room has been raised to a comfortable level this winter. However, this winter is not cold as last winter. Therefore, it is not clear whether the improvement of emergency room temperature was due to the improvement of the heating system or the effect of a warmer winter.

The statistical interaction is to investigate whether the joint contribution of two or more factors is the same as the sum of the contributions from each factor when considered alone. If an interaction between factors exists, an overall assessment cannot be made. For example, suppose that a placebo-controlled clinical trial was conducted at two study centers to assess the effectiveness and safety of a newly developed drug product. Suppose that the results turned out that the drug is efficacious (better than placebo) at one study center and inefficacious (worse than placebo) at the other study center. As a result, a significant interaction between treatment and study center occurred. In this case, an overall assessment of the effectiveness of the drug product can be made.

In practice, possible confounding factors should be identified and properly controlled at the planning stage of the studies. When significant interactions among factors are observed, subgroup analyses may be necessary for a careful evaluation of the treatment effect.

Type I Error, Significance Level, and Power

In statistical analysis, two different kinds of mistakes are commonly encountered when performing hypotheses testing. For example, suppose that a physician is to determine whether or not one of his or her patients is still alive. If the patient is dead, then the physician may remove his or her life support equipment for other patients who need it.

Therefore, the null hypothesis of interest is that the patient is still alive, whereas the alternative hypothesis is that the patient is dead. Under these hypotheses, the physician may make two mistakes, which are (i) he or she may conclude that the patient is dead when in fact the patient is still alive, and (ii) he or she may claim that the patient is still alive when in fact the patient is dead. The first kind of mistake is usually referred as a type I error; the latter is the so-called type II error. Since a type I error is usually considered more important or serious, we would like to limit the probability of committing this kind of error to an acceptable level. This acceptable level of probability of committing a type I error is known as the *significance level*. As a result, if the probability of observing a type I error based on the data is less than the significance level, we conclude that a statistically significant result is observed. A statistically significant result suggests that the null hypothesis be rejected in favor of the alternative hypothesis. The probability of observing a type I error is usually referred to as the *p*-value of the test. On the other hand, the probability of committing a type II error subtracted from 1 is called the *power* of the test. In our example, the power of the test is the probability of correctly concluding the death of the patient when the patient is dead.

For the pharmaceutical application, suppose that a pharmaceutical company is interested in demonstrating that newly developed drug is efficacious. The null hypothesis is often chosen, as the drug is inefficacious, versus the alternative hypothesis that the drug is efficacious. The objective is to reject the null hypothesis in favor of the alternative hypothesis and consequently to conclude that the drug is efficacious. Under the null hypothesis, a type I error is made if we conclude that the drug is efficacious when in fact it is not. This error is also known as *consumer's risk*. Similarly, a type II error is committed if we conclude that the drug is inefficacious. This error is sometimes called *producer's risk*. The power is then considered to be the probability of correctly concluding that the drug is efficacious, when in fact it is. For assessment of drug effectiveness and safety, a sufficient sample size is often selected to have a desired power with a pre-specified significance level. The purpose is to control both type I (significance level) and type II (power) errors.

Randomization

Statistical inference on a parameter of interest of a population under study is usually derived under the probability structure of the parameter. The probability structure depends upon the randomization method employed in sampling. The failure of the randomization will have a negative impact on the validity of the probability structure. Consequently, the validity, accuracy, and reliability of the resulting statistical inference of the parameter is questionable. Therefore, the randomization should be performed using an appropriate randomization method under a valid

randomization model according to the study design to ensure the validity, accuracy, and reliability of the derived statistical inference.

Sample Size Determination/Justification

One of the major objectives of most studies during drug research and development is to determine whether the drug is effective and safe. During the planning stage of a study, the following questions are of particular interest to the pharmaceutical scientists: (i) How many subjects are needed in order to have a desired power for detecting a meaningful difference? (ii) What is the tradeoff if only a small number of subjects are available for the study due to a limited budget and/or some scientific considerations? To address these questions, a statistical evaluation for sample size determination/justification is often employed. Sample size determination usually involves the calculation of sample size for some desired statistical properties, such as power or precision; sample size justification is to provide statistical justification for a selected sample size, which is often a small number.

For a given study, sample size can be determined/justified based on some criteria of a type I error (a desired precision) or a type II error (a desired power). The disadvantage for sample size determination/justification based on the criteria of precision is that it may have a small chance of detecting a true difference. As a result, sample size determination/justification based on the criteria of power becomes the most commonly used method. Sample size is selected to have a desired power for detection of a meaningful difference at a pre-specified level of significance.

In practice, however, it is not uncommon to observe discrepancies among study objective (hypotheses), study design, statistical analysis (test statistic), and sample size calculation. These inconsistencies often result in (i) the wrong test for the right hypotheses, (ii) the right test for the wrong hypotheses, (iii) the wrong test for the wrong hypotheses, or (iv) the right test for the right hypotheses with insufficient power. Therefore, before the sample size can be determined, it is suggested that the following be carefully considered: (i) the study objective or the hypotheses of interest should be clearly stated, (ii) a valid design with appropriate statistical tests should be used, and (iii) sample size should be determined based on the test for the hypotheses of interest.

Statistical Difference and Scientific Difference

A *statistical difference* is defined as a difference that is unlikely to occur by chance alone, while a *scientific difference* is a difference that is considered to be of scientific importance. A statistical difference is also referred to as a *statistically significant difference*. The difference between the concepts of *statistical difference* and *scientific*

difference is that statistical difference involves chance (probability) while scientific difference does not. When we claim there is a statistical difference, the difference is reproducible with a high probability.

When conducting a study, there are basically four possible outcomes. The result may show that (i) the difference is both statistically and scientifically significant, (ii) there is a statistically significant difference yet the difference is not scientifically significant, (iii) the difference is scientifically significant yet it is not statistically significant, and (iv) the difference is neither statistically significant nor scientifically significant.

If the difference is both statistically and scientifically significant or it is neither statistically or scientifically significant, then there is no confusion. However, in many cases, a statistically significant difference does not agree with the scientifically significant difference. This inconsistency has created confusion/arguments among pharmaceutical scientists and biostatisticians. The inconsistency may be due to large variability and/or insufficient sample size.

One-Sided Test versus Two-Sided Test

For the evaluation of a drug product, the null hypothesis of interest is often that there is no difference. The alternative hypothesis is usually that there is a difference. The statistical test for this setting is called a two-sided test. In some cases, the pharmaceutical scientist may test the null hypothesis of no difference against the alternative hypothesis that the drug is superior to the placebo. The statistical test for this setting is known as a *one-sided test*.

For a given study, if a two-sided test is employed at the significance level of 5%, then the level of proof required is one out of 40. In other words, at the 5% level of significance, there is 2.5% chance (or 1 out of 40) that we may reject the null hypothesis of no difference in the positive direction and conclude that the drug is effective at one side. On the other hand, if a one-sided test is used, the level of proof required is 1 out of 20. It turns out that a one-sided test allows more ineffective drugs to be approved because of chance as compared to the two-sided test. It should be noted that when testing at the 5% level of significance with 80% power, the sample size required increases by 27% for a two-sided test as compared to a one-sided test. As a result, there is a substantial cost saving if a one-sided test is used.

However, agreement is not universal among regulatory entities, academia, and the pharmaceutical industry as to whether a one-sided test or a two-sided test should be used. The U.S. Food and Drug Administration tends to oppose the use of a one-sided test, though this position has been challenged by several pharmaceutical companies on Drug Efficacy Study Implementation (DESI) drugs at the Administrative Hearing. Dubey^[7] has pointed out that several viewpoints that favor the use of one-sided tests were discussed in an administrative hearing. These points indicated that a one-sided test is appropriate in the following

situations: (i) where there is truly concern with outcomes in one tail only, and (ii) where it is completely inconceivable that the results could go in the opposite direction.

European Statistics Code of Good Practice

In February 2005, the Statistical Program Committee (CPE) adopted the European Statistics Code of Good Practice and undertook to observe the 15 principles established therein, as well as to periodically review their application using the good practice indicators corresponding to each of the 15 principles. This code has been embraced by *Instituto Nacional de Estadística* (INE) by way of a resolution by the Board of Directors, which thus undertakes to comply with the aforementioned when establishing the general principles regulating the generating of statistics for State purposes. In this way, INE endeavors to guarantee an improvement in the service it provides to society, which will undoubtedly reinforce its image as an institution. In May 2005, the CPE agreed a formula for monitoring implementation of the code for a duration of three years. During that period, the various countries must carry out quality self-assessment, taking as a reference the aforementioned good practice indicators, which in turn must be contrasted and checked via so-called peer reviews. The end result was submitted to the Board and to the European Parliament in 2008. The 15 principles are briefly described below:

- Principle 1: Professional Independence—The professional independence of statistical authorities from other policy, regulatory, or administrative departments and bodies, as well as from private sector operators, ensures the credibility of European statistics.
- Principle 2: Mandate for data collection—Statistical authorities must have a clear legal mandate to collect information for European statistical purposes. Administrations, enterprises, and households and the public at large may be compelled by law to allow access to or deliver data for European statistical purposes at the request of statistical authorities.
- Principle 3: Adequacy of resources—The resources available to statistical authorities must be sufficient to meet European statistics requirements.
- Principle 4: Quality commitment—All ESS members commit themselves to work and cooperate according to the principles fixed in the “Quality declaration of the European statistical system.”
- Principle 5: Statistical confidentiality—The privacy of data providers (households, enterprises, administrations, and other respondents),

- the confidentiality of the information they provide, and its use only for statistical purposes must be absolutely guaranteed.
- Principle 6: Impartiality and objectivity—Statistical authorities must produce and disseminate European statistics respecting scientific independence and in an objective, professional, and transparent manner in which all users are treated equitably.
- Principle 7: Sound methodology—Sound methodology must underpin quality statistics. This requires adequate tools, procedures, and expertise.
- Principle 8: Appropriate statistical procedures—Appropriate statistical procedures, implemented from data collection to data validation, must underpin quality statistics.
- Principle 9: Non-excessive burden on respondents—The reporting burden should be proportionate to the needs of the users and should not be excessive for respondents. The statistical authority monitors the response burden and sets targets for its reduction overtime.
- Principle 10: Cost-effectiveness—Resources must be effectively used.
- Principle 11: Relevance—European statistics must meet the needs of users.
- Principle 12: Accuracy and reliability—European statistics must accurately and reliably portray reality.
- Principle 13: Timeliness and punctuality—European statistics must be disseminated in a timely and punctual manner.
- Principle 14: Coherence and comparability—European statistics should be consistent internally and over time and comparable among regions and countries; it should be possible to combine and make joint use of related data from different sources.
- Principle 15: Accessibility and clarity—European statistics should be presented in a clear and understandable form, disseminated in a suitable and convenient manner, and available and accessible on an impartial basis with supporting metadata and guidance.

IMPLEMENTATION OF GSP

The implementation of GSP in drug research and development is a team project that requires mutual communication, confidence, respect, and cooperation among statisticians, pharmaceutical scientists in the related areas, and regulatory agents. The implementation of GSP

involves some key factors that have an impact on the success of GSP. These factors include (i) regulatory requirements for statistics, (ii) the dissemination of the concept of statistics, (iii) appropriate use of statistics, (iv) effective communication and flexibility, and (v) statistical training. These factors are briefly described next.

In the drug development and approval process, regulatory requirements for statistics are the key to the implementation of GSP. They not only enforce the use of statistics but also establish standards for statistical evaluation of the drug products under investigation. An unbiased statistical evaluation helps pharmaceutical scientists and regulatory agents in determining (i) whether the drug product has the claimed effectiveness and safety for the intended disease, and (ii) whether drug product possesses good drug characteristics, such as the proper identity, strength, quality, purity, and stability. A set of guideline standard operating procedures is often developed to fulfill with regulatory requirements for good statistics practice. For example, Spriet and Dupin-Spriet^[1] have proposed a set of procedures to fulfill quality requirements set by company policy according to regulatory requirements of good clinical practice. Wiles et al.^[2] have indicated that the Professional Standards Working Party of the Statisticians in the Pharmaceutical Industry (PSI) in the United Kingdom has developed a set of guideline standard operating procedures for good statistics practice. These guideline standard operating procedures cover clinical development plan, clinical trial protocol, statistical analysis plan, determination of evaluability of subjects for analysis, randomization and blinding procedure, data management, interim analysis plan, statistical report, archiving and documentation, data overview, and quality assurance and quality control.

In addition to regulatory requirements, it is always helpful to disseminate the concept of statistical principles described earlier whenever possible. It is important for pharmaceutical scientists and regulatory agents to recognize that (i) a valid statistical inference is necessary to provide a fair assessment with certain assurance regarding the uncertainty of the drug product under investigation, (ii) an invalid design and analysis may result in a misleading or wrong conclusion about the drug product, and (iii) a larger sample size is often required to increase the statistical power and precision of the studies. The dissemination of the concept of statistics is critical to establish the pharmaceutical scientists' and regulatory agents' brief in statistics for scientific excellence.

One of the commonly encountered problems in drug research and development is the misuse or abuse of statistics in some studies. The misuse or abuse of statistics is a critical issue that may result either in having the right question with the wrong answer or in having the right answer for the wrong question. For example, for a given study, suppose that a right set of hypotheses (the right question) is established to reflect the study objective. A misused statistical test may provide a misleading or wrong answer to the right question. On the other hand, in many clinical trials,

point hypotheses for equality (the wrong question) are often wrongly used for establishment of equivalency. In this case, we have the right answer (for equality) for the wrong question. As a result, it is recommended that appropriate statistical methods be chosen to reflect the design that should be able to address the scientific or medical questions regarding the intended study objectives for implementation of GSP.

Communication and flexibility are important factors to the success of GSP. Inefficient communication between statisticians and pharmaceutical scientists or regulatory agents may result in a misunderstanding of the intended study objectives and consequently in an invalid design and/or inappropriate statistical methods. Thus, effective communication among statisticians, pharmaceutical scientists, and regulatory agents is essential for the implementation of GSP. In addition, in many studies, the assumption of a statistical design or model may not be met, due to the nature of the drug product under investigation, the experimental environment, and/or other causes related/unrelated to the studies. In this case, the traditional approach of doing everything by the book does not help. In practice, since a concern from a pharmaceutical scientist or the regulatory agent may translate into a constraint for a valid statistical design and appropriate statistical analysis, it is suggested that a flexible yet innovative solution be developed under the constraints for the implementation of GSP.

Because regulatory requirements for the drug development and approval process vary from drug to drug and from country to country, various designs and/or statistical methods are often required for a valid assessment of a drug product. Therefore, statistical continued/advanced education and training programs should be routinely held for both statisticians and non-statisticians, including pharmaceutical scientists and regulatory agents. The purpose of such a continued/advanced education and/or training program is threefold. First, education and training enhance communications within the statistical community. Statisticians can certainly benefit from such a training and/or educational program by acquiring more practical experience and knowledge. In addition, education and training provide the opportunity to share/exchange information, ideas, and/or concepts regarding drug development between professional societies. Finally, education and training identify critical practical and/or regulatory issues that are commonly encountered in the drug development and regulatory approval process. A panel discussion from different disciplines may result in some consensus to resolve the issues, which helps in establishing standards of statistical principles for the implementation of GSP.

CONCLUSION

During the development and regulatory approval process, good pharmaceutical practices are necessarily implemented (i) to ensure the effectiveness and safety of the drug product under investigation before approval and (ii) to ensure that the drug product possesses good drug characteristics, such as the proper identity, strength, quality, purity, and stability, in compliance with the standards as specified in the USP/NF after regulatory approval. These good pharmaceutical practices include GLP for animal studies, GCP for clinical development, cGMP for chemistry, manufacturing, and controls (CMC), and GRP for regulatory review and approval process. In essence, GSP is the foundation of GLP, GCP, cGMP, and GRP. The implementation of GSP is a team project that involves statisticians, pharmaceutical scientists, and regulatory agents as well. The success of GSP depends upon mutual communication, confidence, respect, and cooperation among statisticians, pharmaceutical scientists, and regulatory agents.

REFERENCES

1. Spriet, A.; Dupin-Spriet, T. Good biometrics practice proposals for a set of procedures. *Drug Inf. J.* **1992**, *26*, 405–409.
2. Wiles, A.; Atkinson, G.; Huson, L.; Morse, P.; Struthers, L. Good statistical practices in clinical research: guideline standard operating procedures. *Drug Inf. J.* **1994**, *28*, 615–627.
3. Chow, S.C. Good statistics practice in the drug development and regulatory approval process. *Drug Inf. J.* **1997**, *31*, 1157–1166.
4. CPMP. *The Committee for Proprietary Medicinal Products Working Party on Efficacy of Medicinal Products. Note for Guidance; Good Clinical Practice for Trials on Medicinal Products in the European Community*, Commission of European Communities: Brussels, Belgium, July, 1990, 111/396/88-EN Final.
5. ICH. *Statistical Principles for Clinical Trials*; International Conference on Harmonization; 1997.
6. USP/NF. *The United States Pharmacopeia XXIV and the National Formulary XIX*; United States Pharmacopeial Convention, Inc.: Rockville, MD, 2000.
7. Dubey, S.D. Some thought on the one-sided and two-sided tests. *J. Biopharm. Stat.* **1991**, *1*, 139–150.

Good Validation Practice (GVP)

Ying Geng and Lan He

National Institutes for Food and Drug Control, Beijing, China

Abstract

Validation practice in the pharmaceutical industry is indispensable for quality assurance and quality control. It confirms that the facility, process, operating system, analytical method, or product meets its intended use. Therefore, it assures the quality of pharmaceutical manufacture and the justification of analytical results. This article first presents the definition of validation in pharmaceutical industry, and then summarizes good validation practice on two basic ranges, process validation and analytical method validation, from regulated requirement, quality-by-design concept, and technology transfer perspectives. The statistical justification of pharmaceutical validation plays an important role for its valid and fair assessment.

Keywords: Analytical life cycle; Analytical method validation; Pharmaceutical validation; Process validation; Quality by design; Technology transfer.

INTRODUCTION

Validation practice in the pharmaceutical industry is indispensable for quality assurance and quality control. It confirms that the facility, process, operating system, analytical method, or product meets its intended use. Therefore, it assures the quality of pharmaceutical manufacture and the justification of analytical results. Good validation practice regulates the pharmaceutical validation practice with scientific principles and/or standard operating procedures for the validation practices in design, conduct, monitoring, evaluation, reporting, and decision-making in pharmaceutical manufacturing and regulatory registration. Statistics plays an important role for the validation practice for it provides scientific tools for valid design and fair assessment.

This article first presents the definition of the validation in pharmaceutical industry, and then summarizes good validation practice on two basic ranges, process validation and analytical method validation, from regulated requirement, quality-by-design (QbD) concept, and technology transfer (TT) perspectives. Finally, the important role that statistics plays is highlighted.

DEFINITION OF VALIDATION IN THE PHARMACEUTICAL INDUSTRY

In pharmaceutical industry, the objective of validation is to ensure that the drug product meet specification for identity, strength, quality, purity, stability and reproducibility. The concept of validation was introduced firstly by

Ted Byers and Bud Loftus, in order to improve the quality of pharmaceutical products.^[1] Generally, the definition of validation was described by subpart H part 820 in CRF 21 for Chapter I, food and drug administration, as “Validation means confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use can be consistently fulfilled.”^[2] Two more specific definitions were described in CRF 21, one as “establishing documented evidence that provides a high degree of assurance that a system will consistently produce a product meeting its predetermined specifications and quality characteristics” in subpart B of part 106 for cGMP,^[3] and the other as “obtaining and evaluating scientific and technical evidence that a control measure, combination of control measures, or the food safety plan as a whole, when properly implemented, is capable of effectively controlling the identified hazards”.^[4]

Under the general definition of validation, validation practice can refer to people, equipment, facility, method, process, or other concerns covered in the Current Good Manufacturing Practice (cGMP); the FDA published a cluster of guidelines regulating the validation practices varying from analytical method validation (AMV)^[5,6] and process validation (PV)^[7] to the sterilization PV,^[8] software validation,^[9] computer system validation,^[10] as well as clinical validation.^[11] Basically, pharmaceutical validations include AMV for the purpose that the validation of an analytical method or a test procedure is to assure the accuracy and reliability of the test results, and PV, which is aiming to assure that the final product has a high probability of meeting the standards for the identity, strength, quality, purity, stability, and reproducibility of the drug product.^[12]

PV OF PHARMACEUTICAL INDUSTRY

Process validation began from the validation of sterile process in the 1970s. With several decades of development, the process validation evolved into a wide range of processes in the pharmaceutical industry, such as aseptic process for bulk pharmaceutical chemicals, container preparation processes, lyophilization process, biotechnology process, solid dosage forms process, and packaging process.

Regulated Requirement

The cGMP regulations in 21 Code of Federal Regulations (CFR) Parts 210 and 211 described general and specific requirements for validation of manufacturing process, which are mainly on designing, operating, and controlling a process, to assure the product quality. The FDA published an updated “Guidance for Industry Process Validation: General Principles and Practices” in January 2011,^[7] replacing “Guideline on General Principles of Process Validation” (the 1987 Guidance).^[13] In the 2011 Guidance, PV is defined as “the collection and evaluation of data, from the process design stage through commercial production, which establishes scientific evidence that a process is capable of consistently delivering quality product,” and described in three stages^[7]:

- Stage 1—Process design: the commercial manufacturing process is defined during this stage based on knowledge gained through development and scale-up activities.
- Stage 2—Process qualification: the process design is evaluated to determine if the process is capable of reproducible commercial manufacturing.
- Stage 3—Continued process verification: ongoing assurance is gained during routine production that the process remains in a state of control.

As stated in the guideline 2011,^[7] Effective PV contributes significantly to assuring drug quality. The basic principle of quality assurance is that a drug should be produced that is fit for its intended use. This principle incorporates the understanding that the following conditions exist:

- a. Quality, safety, and efficacy are designed or built into the product.
- b. Quality cannot be adequately assured merely by in-process and finished product inspection or testing.
- c. Each step of a manufacturing process is controlled to assure that the finished product meets all quality attributes including specifications.

QbD Quality-by-Design Concept and PV

Since the FDA began with the initiative entitled “Pharmaceutical Quality for the 21st Century” in 2002,^[14] new

technological advances, modern quality management techniques, and risk-based approaches have been encouraged into adoption in the pharmaceutical industry. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) developed a series of guidance, Q8 Guideline Pharmaceutical Development,^[15] Q9 guideline Quality Risk Management,^[16] and Q10 Pharmaceutical Quality System.^[17] The “Guidance for Industry Process Validation: General Principles and Practices (Guideline 2011)” aligns PV activities with a product life cycle concept and risk management. QbD uses a science and risk-based approach that emphasizes the importance of developing scientific knowledge and thorough understanding of both the product and the process.

TT of Pharmaceutical Process

TT of pharmaceutical process refers to the transfer of a manufacturing process from one site to another site. TT could be from an early-stage research laboratory to a manufacture site, between two clinical manufacture sites, from one clinical manufacture site to commercial site, or between commercial sites, and be with the same scale or a scale-up strategy. A successful TT requires a comprehensive knowledge of process characteristics to demonstrate that the transferred product meet in-process release and stability specifications, and be consistent and reproducible with the donor site. The prospective validation with QbD approach may be applied to identify critical process parameters that may have impacts on critical quality attributes. Moreover, gap analysis and risk assessment may be utilized for the purpose of quality risk management and understanding the operation differences between the donor and receiving sites.

VALIDATION OF ANALYTICAL PROCEDURES

Good Analytical Method Validation Practice ensures accurate and reliable test results, and therefore the safety and quality of the product.^[18]

Regulated Requirement

In the United States, the Federal Food, Drug, and Cosmetic Act regulated that assays and specifications in monographs of the United States Pharmacopeia and the National Formulary (USP–NF) constitute legal standards. The cGMP regulations 21 CFR 211.194(a) require that test methods, “which are used for assessing compliance of pharmaceutical articles with established specifications, must meet proper standards of accuracy and reliability.”^[19] Similar regulations for the regulation of analytical procedure validation as part of registration application and newly carried in pharmacopoeia were drafted in various compendia of

the United States,^[20] the European Union,^[21] Japan,^[22] and China.^[23] The ICH also published harmonized guideline in order to provide some guidance and recommendations for analytical procedure.^[24,25]

Harmonized Characteristics

As harmonized by ICH, during the validation of analytical procedures, appropriated validation characteristics can be considered simultaneously to provide a sound, overall knowledge of the capabilities of the analytical procedure.^[24] These validation characteristics are accuracy, precision (repeatability, intermediate precision, and reproducibility), specificity, detection limit, quantitation limit linearity, and robustness. There are four common types of analytical procedure: (1) identification tests, (2) quantitative tests for impurities' content, (3) limit tests for the control of impurities, and (4) quantitative tests for active moiety in samples of drug substance or drug product or other selected component(s) in the drug product.^[25]

Assay validation is the process of demonstrating and documenting that the performance characteristics of the procedure and its underlying method meet the requirements for the intended application and that the assay is therefore suitable for its intended use. Biological assays (bioassays) typically exhibit greater variability than chemically based tests, for the inherent characteristics of biological response to the product; a biochemical or physiological response at the cellular level; enzymatic reaction rates or biological responses induced by immunological interactions; or ligand and receptor binding.^[26,27] Therefore, unlike the straightforward validation strategy for well-characterized, chemically based small-molecule drug product, the validation of bioassays is more flexible and requires more profound statistical consideration.

Analytical Life Cycle and Method Transfer

In 2010, Nethercote et al. suggested that AMV can benefit from a life cycle approach and proposed that method validation be defined as “the collection and evaluation of data and knowledge from the method design stage through its life cycle of use which establishes scientific evidence that a method is capable of consistently delivering quality data.”^[28] The USP general chapter <1224> describes the transfer of analytical procedure (also referred to as method transfer) as “the documented process that qualifies a laboratory (the receiving unit) to use an analytical test procedure that originate in another laboratory (the transferring unit), thus ensuring that the receiving unit has the procedural knowledge and ability to perform the transferred analytical procedure as intended.”^[29] Under the concept of life cycle, method transfer may be conducted at any stage from the initializing site (originator) to either another internal site(s) or an external receiving laboratory (recipient). The principle goal of method transfer is to

demonstrate that the method is appropriately transferred and validated at the receiving laboratory.

STATISTICS AND GOOD VALIDATION PRACTICE

Statistics is deeply involved in pharmaceutical validation, from experimental/method design, analytical characteristic evaluation, sampling strategy, quality control, specification setting to process capability/risk assessment. It provides scientific tools for good validation practice, such as risk assessment, process capability control, design of experiment, multiple stages tests approach, and sampling strategy.

Statistics is directly regulated by cGMP guidelines: CFR 21 211.110(b) requires that in-process specifications be “derived from previous acceptable process average and process variability estimates where possible and determined by *the application of suitable statistical procedures* where appropriate,”^[30] and cGMP regulations specify that samples must meet specifications and *statistical quality control criteria* as condition of approval and release (§ 211.165(d)),^[31] and the report must *provide appropriate confidence interval* (§ 870.1415(c)).^[32] Moreover, in the 2011 Guidance, it is recommended that a statistician or a person with adequate training in statistical process control techniques develop the data collection plan and statistical methods and procedures used in measuring and evaluating process stability and process capability. As stated, “The 2011 Guidance references a number of acceptable industry standards but clarifies that manufacturers must make deliberate decisions about which statistical tools and analyses are appropriate for their products and processes.”^[7]

CONCLUSION

Validation is a scientific endeavor at various stages of life cycle of pharmaceutical product development. Since the issue of the Guideline on General Principles of Process Validation (the 1987 Guidance), good validation practices have been broadened to a wide range through pharmaceutical industry. Good validation practices ensure the analytical method is suitable for its intended use throughout the life cycle, the well-designed experiment is conducted, the variation due to process capability is under control, and the products have required quality and reproducibility from regulatory agencies. The adoption of GVP guidelines is inherently an important part of the QbD concept, modernizing the pharmaceutical industry progress and producing high-quality drug products, and is critical for the desired state of under-control pharmaceutical manufacturing and high-quality products without extensive regulation oversight. The statistical justification of pharmaceutical validation plays an important role for its valid and fair assessment.

REFERENCES

1. Agalloco, J. Validation: An unconventional review and reinvention. *PDA J. Pharm. Sci. Technol.* **1995**, 49 (4), 175–179.
2. General Provisions. Current Good Manufacturing Practice. Code of Federal Regulations Title 21. Chapter I—Food and Drug Administration Department of Health and Human Service. Subpart H—Medical Devices. Part 820. Subpart A.
3. Current Good Manufacturing Practice. Code of Federal Regulations Title 21. Chapter I—Food and Drug Administration Department of Health and Human Service. Subpart B—Food for Human Consumption. Part 106. Subpart B.
4. General Provisions. Code of Federal Regulations Title 21. Chapter I—Food and Drug Administration Department of Health and Human Service. Subpart B—Food for Human Consumption. Part 117. Subpart A.
5. FDA. *Guidance for Industry—Validation of Analytical Procedures: Definition and Terminology*; FDA: Rockville, MD, 1999.
6. FDA. *Guidance for Industry—Validation of Analytical Procedures: Methodology*; FDA: Rockville, MD, 1999.
7. FDA. *Guidance for Industry—Process Validation: General Principles and Practices*; FDA: Rockville, MD, 2011.
8. FDA. *Guidance for Industry and FDA Staff—Medical Device User Fee and Modernization Act of 2002, Validation Data in Premarket Notification Submissions (510(k)s) for Reprocessed Single Use Medical Devices*; FDA: Rockville, MD, September 2006.
9. FDA. *General Principles of Software Validation*; Final Guidance for Industry and FDA Staff; FDA: Rockville, MD, January 2002.
10. FDA. *Guidance for Industry—Blood Establishment Computer System Validation in the User's Facility*; FDA: Rockville, MD, April 2013.
11. FDA. *Guidance for Industry—In the Manufacture and Clinical Evaluation of In Vitro Tests to Detect Nucleic Acid Sequences of Human Immunodeficiency Viruses Types 1 and 2*; FDA: Rockville, MD, December 1999.
12. Chow, S.C. *Statistics in Drug Research*; Marcel Dekker: New York, 2002.
13. FDA. *Guideline on General Principles of Process Validation*; FDA: Rockville, MD, May 1987.
14. FDA. *Final Report: Pharmaceutical CGMPs for the 21st Century—A Risk-Based Approach*; FDA: Rockville, MD, September 2004.
15. ICH Secretariat. *ICH Harmonized Tripartite Guideline: Pharmaceutical Development Q8(R2)*; Geneva, Switzerland, 2009.
16. ICH Secretariat. *ICH Harmonized Tripartite Guideline: Quality Risk Management Q9*; Geneva, Switzerland, 2005.
17. ICH Secretariat. *ICH Harmonized Tripartite Guideline: Pharmaceutical Quality System Q10*; Geneva, Switzerland, 2008.
18. Krause, S.O. Good analytical method validation practice, part I: Setting-up for compliance and efficiency. *J. Validation Technol.* **2002**, 9 (1), 23–32.
19. Current Good Manufacturing Practice for finished pharmaceuticals. Code of Federal Regulations Title 21. Chapter I—Food and Drug Administration Department of Health and Human Service. Subpart C—Drugs: General. Part 211. Subpart J.
20. General Chapter <1225>. Validation of Compendial Procedures. United States Pharmacopeia and the National Formulary (USP40–NF35).
21. General Tests 5.3. Statistical Analysis of Results of biological Assays and Tests. European Pharmacopoeia 9.2.
22. General Information—Validation of Analytical Procedures. Japanese Pharmacopoeia 17th Ed., JP XIV, 2001.
23. General Chapter <9101> *Guidelines for Validation of Analytical Method Adopted in Pharmaceutical Quality Specification*. Pharmacopoeia of the People's Republic of China 2015 Vol. IV.
24. ICH Secretariat. *Guidance for Industry—Q2A Text on Validation of Analytical Procedures*; Geneva, Switzerland, 1995.
25. ICH Secretariat. *Guidance for Industry—Q2B Validation of Analytical Procedures: Methodology*; Geneva, Switzerland, 1996.
26. General Chapter <1033>. Biological Assay Validation. United States Pharmacopeia and the National Formulary (USP40–NF35).
27. ICH Secretariat. *ICH Harmonized Tripartite Guideline: Specifications: Test Procedures and Acceptance Criteria for Biotechnological/Biological Products Q6B*; Geneva, Switzerland, 1999.
28. Nethercote, P.; Borman, P.; Bennett, T.; GSK; Martin, G.; Complectors Consulting LLC; McGregor, P.; PMcG Consulting. QbD for Better Method Validation and Transfer. *Pharmaceutical Manufacturing* 2010; <http://www.pharmamanufacturing.com/articles/2010/060.html> (accessed in August 18, 2017).
29. General Chapter <1224>. Transfer of Analytical Procedures. United States Pharmacopeia and the National Formulary (USP40–NF35).
30. Current Good Manufacturing Practice for finished pharmaceuticals. Code of Federal Regulations Title 21. Chapter I—Food and Drug Administration Department of Health and Human Service. Subpart C—Drugs: General. Part 211. Subpart F.
31. Current Good Manufacturing Practice for finished pharmaceuticals. Code of Federal Regulations Title 21. Chapter I—Food and Drug Administration Department of Health and Human Service. Subpart C—Drugs: General. Part 211. Subpart I.
32. Cardiovascular devices. Code of Federal Regulations Title 21. Chapter I—Food and Drug Administration Department of Health and Human Service. Subpart H—Medical devices. Part 870. Subpart B.

Group Sequential Methods

Cheng-Tao Chang

Novo Nordisk Inc., Princeton, New Jersey, U.S.A.

Abstract

The group sequential procedure (design, monitoring, and analysis) is the most commonly applied form of sequential procedure to facilitate the conduct of interim analysis in the biopharmaceutical field. The focus of this entry is on the regulatory requirements, statistical methods, and design monitoring of group sequential procedures.

INTRODUCTION

The group sequential procedure (design, monitoring, and analysis) is the most commonly applied form of sequential procedure to facilitate the conduct of interim analysis in the biopharmaceutical field. A primary motivation of a group sequential analysis, in general, is the conservation of experimental materials and/or research effort. An important biopharmaceutical application in this aspect was in drug screening.^[1] Recently, active developments of the group sequential procedure have been in clinical trials, especially trials with mortality or irreversible morbidity as the primary endpoint. In these clinical trials, a further consideration is an ethical obligation to terminate the study at the earliest opportunity to prevent the administration of a less favorable therapy to additional patients. Conducting the clinical trial in stages, with group sizes predetermined by a certain schedule, provides logistical feasibility over fully sequential methods because these experiments or clinical trials generally will involve a lengthy time to respond. Yet the advantage of reducing the average sample number is largely retained.^[2] Many different methods of group sequential procedure are available: different stopping boundaries, different Type I error spending/use functions, conditional power or predictive power approaches, Bayesian methods, etc. The logistics of a group sequential trial are, however, complicated and require careful planning and conducting. A review article is given in Ref. [3], and a rather complete book has been published recently as well.^[4]

REGULATORY REQUIREMENTS

The International Conference on Harmonization (ICH) Guidelines (February 1998)^[5] discuss the regulatory requirements on interim analyses in general (E3, Section 11.4.2.3) and the group sequential procedures in particular regarding data monitoring and early stopping

(E9, Sections 3.4 and 4.1–4.6). The following is a summary:

- All interim analyses should be carefully planned in advance and should be described in the protocol. Any unplanned interim analysis should be avoided if at all possible. In the event that an unplanned interim analysis was conducted, a protocol amendment describing the interim analysis should be completed prior to an unblinded access to treatment comparison data, and the study report should explain why it was necessary and the degree to which the blindness had to be broken, and should provide an assessment of the potential magnitude of bias introduced and of the impact on the interpretation of the results.
- The schedule of analyses, or at least the considerations that will govern its generation, should be stated in the protocol or amendment before the time of the first interim analysis; the stopping guidelines and their properties should also be clearly stated there.
- The procedure plan should be written or approved by the Data Monitoring Committee (DMC), when the trial has one.
- Any changes to the trial and any consequent changes to the statistical procedures should be specified in an amendment to the protocol at the earliest opportunity.
- The procedures selected should always ensure that the overall probability of Type I error is controlled.
- The execution of an interim analysis must be a completely confidential process when unblinded data and results are potentially involved. All staff involved in the conduct of the trial should remain blind to the results of such analyses. Investigators should only be informed about the decision to continue or discontinue the trial, or to implement modifications to trial procedures.
- Any interim analysis planned for administrative purposes only should also be described in the protocol and subsequently reported. The specific purposes

should be clearly stated and should specifically exclude any possibility of early stopping. The blind should not be broken.

STATISTICAL METHODS

Fundamentals

We start by considering the design of a clinical trial with a plan of monitoring the data at calendar time $ct_1, ct_2, \dots, ct_K$ for some fixed $K > 1$. For example, a trial monitoring committee might plan to meet every March 31 and September 30 for the first 2 years and then every June 15 for the rest of the trial duration. (The DMC will be discussed later.) Corresponding to the ct_i , $i = 1, \dots, K$ are the information times (or *information fractions*) $t_1, t_2, \dots, t_K = 1$. (The *information time/fraction* and the *maximum duration trial* vs. *maximum information trial* will be discussed later.) At information time t_i , $i = 1, \dots, K$, standardized Z -statistics $Z_1, Z_2, \dots, Z_K$ are calculated based on the accumulated data for testing hypotheses about the treatment effect represented by a parameter θ . A focal statistical problem of the group sequential procedures is to find critical values to satisfy certain desirable operational characteristics. These critical values are the so-called *group sequential boundaries*. For example, for a one-sided hypothesis with the decision options being “either reject $H_0: \theta = 0$ or continue at the interim stages” and “reject or accept H_0 at the final stage,” the problem is to find values $b_1, b_2, \dots, b_K$ such that the overall Type I error α is maintained at a prespecified level:

$$P_{H_0}(\text{Reject } H_0) = P_{H_0}(Z_1 \geq b_1, \text{ or } Z_2 \geq b_2, \text{ or } \dots Z_K \geq b_K) = \alpha \quad (1)$$

(Different forms of one-side hypotheses are given in Ref. [3].) For a corresponding two-sided hypothesis test, a symmetric boundary would replace α by $\alpha/2$ in Eq. 1 and then apply the one-sided boundary symmetrically. (Asymmetric boundaries and interim analyses with a chance to accept H_0 at an early stage will be discussed later.) In the following, we first review several commonly cited group sequential boundaries.

Equally Spaced Group Sequential Boundaries

Suppose that the (information) times are equally spaced: $t_1 = 1/K$, $t_2 = 2/K, \dots, t_K = K/K = 1$.

- Pocock:^[6] Find $b_i = c$ for $i = 1, \dots, K$ such that:

$$P_{H_0}(Z_1 \geq c, \text{ or } Z_2 \geq c, \text{ or } \dots Z_K \geq c) = \alpha \quad (2)$$

For example, if set $\alpha = 0.025$ and $K = 5$, then $c = 2.413$. The boundaries are (2.413, 2.413, 2.413, 2.413, 2.413).

- O’Brien and Fleming:^[7] Find c such that:

$$P_{H_0}(Z_1 \sqrt{t_1} \geq c, \text{ or } Z_2 \sqrt{t_2} \geq c, \text{ or } \dots Z_K \sqrt{t_K} \geq c) = \alpha \quad (3)$$

For $\alpha = 0.025$ and $K = 5$, $c = 2.04$. Hence $b_i = 2.04/\sqrt{t_i} = 2.04/\sqrt{i/5}$. The boundaries are (4.562, 3.226, 2.634, 2.281, 2.040).

- Haybittle^[8] and Peto et al.^[9] Find $b_i = c$ for $i = 1, \dots, K-1$ such that:

$$P_{H_0}(Z_1 \geq c, \text{ or } Z_2 \geq c, \text{ or } \dots Z_K \geq z_{1-\alpha}) = \alpha \quad (4)$$

For example, if set $\alpha = 0.025$ and $K = 5$, then $c = 3.291$. The boundaries are (3.291, 3.291, 3.291, 3.291, 1.960).

The preceding boundaries, for the two-sided case, are not “closed” at the final analysis. Whitehead and Stratton^[10] and Whitehead and Brunier^[11] proposed “triangular” closed boundaries, which plot the unstandardized test statistic vs. its variance.

The calculations of the boundaries require iterative numerical integrations.^[12,13] Commercial computer packages available for obtaining the boundaries for a variety of K and α include EaSt^[14] and S + ^[15] for the Pocock and O’Brien–Fleming, and PEST^[16] for the Whitehead boundaries. Useful programs and documents are also publicly available on the World Wide Web sites (see Refs. [17,18]).

The choice of the boundaries in the trial design depends on the intention of early stopping of a trial. O’Brien–Fleming boundaries are the most stringent among the three at early times, while Pocock boundaries maintain a constant throughout. Haybittle–Peto boundaries are similar to Pocock boundaries at interim stages but set the final analysis resembling the fixed sample size design.

At the design stage, it may be reasonable to plan for fixed K analyses at equally spaced information times. But when the trial is ongoing (i.e., when monitoring the study), changes often occur. Interim analyses may well not be on the preplanned schedule, and the frequency of the analyses K may also change. This requires a more flexible procedure. The Slud–Wei^[19] and Lan–DeMets^[20] Type I error spending/use function approach meets this need.

Type I Error Spending/Use Function Approach

The idea of the Type I error spending/use function approach can be illustrated as follows. Untie the fixed K in Eq. expression 1 and decompose the rejection region:

$$R = \{Z_1 \geq b_1, \text{ or } Z_2 \geq b_2, \text{ or } Z_3 \geq b_3, \dots\}$$

into disjoint regions:

$$R_1 = \{Z_1 \geq b_1\}$$

$$\begin{aligned}
& \{Z_{t_1} \geq b_1, \text{ or } Z_{t_2} \geq b_2\} \\
&= \{Z_{t_1} \geq b_1\} \cup \{Z_{t_1} < b_1, Z_{t_2} \geq b_2\} = R_1 \cup R_2 \\
& \{Z_{t_1} \geq b_1, \text{ or } Z_{t_2} \geq b_2, \text{ or } Z_{t_3} \geq b_3\} \\
&= (R_1 \cup R_2) \cup \{Z_{t_1} < b_1, Z_{t_2} < b_2, Z_{t_3} \geq b_3\} \\
&= (R_1 \cup R_2) \cup R_3
\end{aligned}$$

and so on. Hence:

$$\begin{aligned}
P_{H_0}(R) &= P_{H_0}(R_1 \cup R_2 \cup R_3 \cup \dots) \\
&= P_{H_0}(R_1) + P_{H_0}(R_2) + P_{H_0}(R_3) + \dots = \alpha
\end{aligned}$$

At t_1 , we specify a value of $\alpha(t_1) = P_{H_0}(R_1)$ and solve for b_1 . At t_2 , specify:

$$\begin{aligned}
\alpha(t_2) &= P_{H_0}(R_1 \cup R_2) = P_{H_0}(R_1) + P_{H_0}(R_2) \\
&= \alpha(t_1) + P_{H_0}(R_2)
\end{aligned}$$

This leads to $P_{H_0}(R_2) = \alpha(t_2) - \alpha(t_1)$, and we can solve for b_2 because b_1 is already known from the first stage. Similarly at t_3 , specify:

$$\begin{aligned}
\alpha(t_3) &= P_{H_0}(R_1 \cup R_2 \cup R_3) = (P_{H_0}(R_1) + P_{H_0}(R_2)) + P_{H_0}(R_3) \\
&= \alpha(t_2) + P_{H_0}(R_3)
\end{aligned}$$

Thus $P_{H_0}(R_3) = \alpha(t_3) - \alpha(t_2)$. We then solve for b_3 because b_1 and b_2 have already been solved in the previous steps. The process continues.

Note that, when solving for b_2 , we only need the joint distribution of (Z_{t_1}, Z_{t_2}) . For a partial sum process with independent increments (see later discussion), where we can write

$$Z_i = \frac{S(t_i)}{\sqrt{I(t_i)}}$$

with the numerator being the cumulative sum of the observations (i.i.d. variates with mean θ) and the denominator being the square root of the cumulative statistical information in estimating θ at time t_i , the correlation is:

$$\begin{aligned}
\rho(t_2, t_2) &= \text{Corr} = \text{Cov}(Z_{t_1}, Z_{t_2}) \\
&= \text{Cov}\left(\frac{S(t_1)}{\sqrt{I(t_1)}}, \frac{S(t_2)}{\sqrt{I(t_2)}}\right) \\
&= \text{Cov}\left(\frac{S(t_1)}{\sqrt{I(t_1)}}, \frac{S(t_1) + (S(t_2) - S(t_1))}{\sqrt{I(t_2)}}\right) \\
&= \frac{1}{\sqrt{I(t_1)I(t_2)}} \text{Var}(S(t_1)) = \sqrt{\frac{I(t_1)}{I(t_2)}}
\end{aligned}$$

Other time point boundaries follow similarly. [When information is proportional to the sample size, then $\rho(t_1, t_2) = (n_1/n_2)^{1/2} = (t_1/t_2)^{1/2}$.] Therefore the joint distribution only involves the ratio of information at the current stage and before, not the later time points. This indicated the flexibility of this approach for monitoring trials because there is no requirement for a given K , the total number of analyses, or for the “equally spaced t_i ” restriction. All that is needed is $\alpha(t_1) < \alpha(t_2) < \dots < \alpha(1) = \alpha$, a strictly increasing function $\alpha(t)$, known as the “Type I error (α) spending/use function” prespecified in the study design protocol.

Example: Two-stage design (one interim and one final analyses) symmetrical boundaries.

Specify the overall Type I error $\alpha = 0.05$ (two-sided test) and the spending function as $\alpha(t) = \alpha t$. Suppose we conduct an interim analysis when $t_1 = 1/2$, then $\alpha(t_1 = 1/2) = 0.025$. The corresponding boundary point at $t_1 = 1/2$ is $b_1 = 2.244$. This leaves $\alpha - \alpha(t_1 = 1/2) = 0.025$ to be spent at the final stage $t_2 = 1$. Solving $P\{|Z_{t_1}| < 2.244, |Z_{t_2}| \geq b_2\} = 0.025$ given that the correlation of the bivariate normal distribution is $t_1/t_2 = 1/2$, we obtain $b_2 = 2.126$. Notice that at the final stage, $\alpha_2 \equiv P(|Z| > 2.126) = 0.034$. That is, with this linear/uniform spending function, if an interim testing is to be conducted in the middle of the trial, the P value at the final stage has to be less than 0.034—instead of the 0.05 level—to be significant.

A natural question is “What kind of α spending function describes the Pocock boundary and the O’Brien–Fleming boundary?”

Example: $K = 5$, $\alpha = 0.025$, Pocock boundary: (2.413, 2.413, 2.413, 2.413, 2.413).

- i. $P_{H_0}(Z_{t_1} \geq 2.413) = 0.0079 \Rightarrow \alpha(t_1 = 1/5) = 0.0079$
- ii. $P_{H_0}(Z_{t_1} \geq 2.413 \text{ or } Z_{t_2} \geq 2.413)$
 $= P_{H_0}(Z_{t_1} \geq 2.413) + P_{H_0}(Z_{t_1} < 2.413, Z_{t_2} \geq 2.413)$
 $= 0.0079 + 0.0059 = 0.0138$
 $\Rightarrow \alpha(t_2 = 2/5) = 0.0138$
- iii. $P_{H_0}(Z_{t_1} \geq 2.413, \text{ or } Z_{t_2} \geq 2.413, \text{ or } Z_{t_3} \geq 2.413) = P_{H_0}(Z_{t_1} \geq 2.413 \text{ or } Z_{t_2} \geq 2.413) + P_{H_0}((Z_{t_1} < 2.413, Z_{t_2} < 2.413), (Z_{t_3} \geq 2.413)) = 0.0138 + 0.0045 = 0.0183$
 $\Rightarrow \alpha(t_3 = 3/5) = 0.0183$
- iv. $P_{H_0}(Z_{t_1} \geq 2.413, \text{ or } Z_{t_2} \geq 2.413, \text{ or } Z_{t_3} \geq 2.413, \text{ or } Z_{t_4} \geq 2.413)$
 $= \dots (\text{similar steps}) = 0.0183 + 0.0036$
 $= 0.0219 \Rightarrow \alpha(t_4 = 4/5) = 0.0219$
- v. $P_{H_0}(Z_{t_1} \geq 2.413, \text{ or } Z_{t_2} \geq 2.413, \text{ or } Z_{t_3} \geq 2.413, \text{ or } Z_{t_4} \geq 2.413, \text{ or } Z_{t_5} \geq 2.413) = \dots (\text{similar steps})$
 $= 0.0219 + 0.0031 = 0.0250 \Rightarrow \alpha(t_5 = 1) = 0.0250$

Example: $K = 5$, $\alpha = 0.025$, O'Brien–Fleming boundary: (4.56, 3.23, 2.63, 2.28, 2.04).

Following similar steps as described previously, we can obtain the (discrete) Type I error spending function $\alpha(t_1 = 1/5) = 0.0000$, $\alpha(t_2 = 2/5) = 0.0006$, $\alpha(t_3 = 3/5) = 0.0045$, $\alpha(t_4 = 4/5) = 0.0128$, and $\alpha(t_5 = 1) = 0.0250$.

Removing the restriction of a fixed K and equally spaced $t_i = i/K$, Lan and DeMets^[20] gave the continuous α spending function for the general Pocock boundary as $\alpha_{\text{Pocock}}(t) = \alpha \ln[1 + (e - 1)t]$, and for the general O'Brien–Fleming boundary as:

$$\alpha_{\text{OB-F}}(t) = 2 \left[1 - \Phi \left(\frac{z_{\alpha/2}}{\sqrt{t}} \right) \right]$$

where $\Phi(\cdot)$ is the c.d.f. of the standard normal variate.

Other Type I error rate spending functions have also been proposed in the literature. Lan and DeMets^[20] and Kim and DeMets^[21] studied a certain member of the “power family” $\alpha(t) = \alpha t^c$, $c > 0$. The landmark clinical trial “Simvastatin Scandinavian Survival Study” or the “4S”^[22] used a member of the “truncated exponential distribution family” of α spending functions proposed by Hwang et al.^[23]

$$\begin{aligned} \alpha(\gamma, t) &= \alpha \left[\frac{1 - e^{-\gamma t}}{1 - e^{-\gamma}} \right], \quad \gamma \neq 0 \quad \text{for } 0 \leq t \leq 1 \\ &= \alpha t, \gamma = 0 \end{aligned} \quad (5)$$

As noted in Hwang et al., the continuous Pocock boundary is included in Eq. 5, with $\gamma = 1$; the continuous O'Brien–Fleming boundary is also a member of Eq. 5, with approximately $\gamma = -4$ or -5 ; the “power family” member $\alpha(t) = \alpha t^{3/2}$ is closely approximated by $\gamma = -1$, and $\alpha(t) = \alpha t^2$ is approximated by $\gamma = -2$.

Group sequential methods provide a chance for early rejection of H_0 and for stopping a trial in such a way that the overall Type I error rate is maintained. The tradeoff is some loss of power, because the tests at early stages involve smaller sample sizes, and at later stages, the rejection regions are tighter (both implying lower power.) To have the same power as the fixed size design, a group sequential trial would require a larger *maximum* sample size, but the *average* sample size would still be smaller (under H_A). The maximum sample size factor is a function of the Type I and Type II errors, the number of analyses, and the way of spending α .

Some authors^[21,24] tried to search for “optimal boundaries” that minimize *average* sample size or the expected first boundary crossing time under a given H_A and power. However, in practice, the choice of boundaries is usually based more on clinical decisions and the nature of the trial than the statistical properties. For example, with either the “power family” or the “truncated exponential family,” one can choose a member (e.g., $\gamma = 0$ in Eq. 5) for a uniform spending of the Type I error. Choosing a member

(e.g., $\gamma > 0$, usually $1 \leq \gamma \leq 4$) for a concave spending of the Type I error is suitable for short-term trials with immediate response, such as single-dose analgesic studies using time-to-pain-relief as the endpoint, or when accelerated early stopping is desirable in the early phase of drug development. A negative γ in Eq. 5 (usually $-5 \leq \gamma \leq -1$, including the O'Brien–Fleming boundary) is for convex spending, which is suitable for a large mortality trial where the patient recruitment is slow and staggered and the follow-up period is long term so that very early stopping with limited samples is not much encouraged. As was pointed out by Kim and DeMets,^[21] the more convex the Type I error spending function is, the more powerful the group sequential test is.

Partial Sum Process with Independent Increments

Most of the group sequential methods were developed earlier using immediate response, either continuous or binary, where cumulative sums of the independent random variables from time to time are easily seen to have independent increments. The usefulness of this “partial sum with independent increments” property was seen in the previous section on “Type I Error Spending/Use Function Approach.”

An important development was provided by Tsiatis,^[25,26] who showed that the log-rank statistic computed over time behaves much like a partial sum of independent normal random variables. This result extended the use of group sequential methods to clinical trials with survival data. (See the entry on *Survival Analysis*.) Recently, Jennison and Turnbull^[27] provided a unified theory that explains the “independent increments” structure commonly seen in group sequential test statistics. Scharfstein et al.^[28] demonstrated that all sequentially computed Wald statistics based on efficient estimators [e.g., maximum likelihood estimators (MLEs)] of the parameter of interest will, under mild regularity conditions, have the asymptotic multivariate normal distribution similar to the earlier setup. Hence the group sequential procedure extends to more complicated situations such as the proportional hazards model^[29] and correlated observations including longitudinal data with random effects model^[30,31] or with distribution-free analyses.^[32]

Information Time/Fraction and Maximum Duration Trial Vs. Maximum Information Trial

Taking statistical information as the inverse of the variance of the parameter estimate, Lan and Zucker^[33] defined the *information time/fraction* as the amount of information accrued by calendar time divided by the total information at the scheduled end of the trial. We have seen that the information time played an essential role in the Type I error spending function approach. Depending on the statistics used (see Lan et al.^[34] for a general discussion), a complete unit information can be approximated by either a

patient (for comparing means) or an event (for comparing survival distributions). In either case, the total information must be known. If not known, as is often the case, then the information time/fraction can only be estimated.

For example, in the time-to-event case, Tsiatis^[25] showed that the variance of the log-rank statistics, when calculated over time, grows proportionally to the number of events observed. Hence the information time is equal to the proportion of the maximum number of events expected by the end of a study, with the numerator being the observed number of events at the (calendar) time of interim analysis. For a maximum information trial, the maximum number of events expected by the end of a study is chosen in advance to achieve a desired power, given other design parameters. However, for a maximum duration trial (i.e., when the maximum trial duration is fixed), the maximum information is random. In such a trial, the denominator of the information time can be estimated under either the null or the alternative hypothesis, thus leading to two information time scales.

A natural compromise to overcome this difficulty of uncertain information time scale is first to choose an estimate of the maximum information, either “under the null” or “under the alternative” hypothesis. We then set the information time to be 1 if the proportion exceeds 1 or if the current analysis is the last analysis and the proportion has not yet reached 1 (e.g., see Kim et al.^[35]). The consequence of this compromise is that the Type I error spending function will be altered from the original one that was prespecified in the design.

The following simple example^[35] illustrates the above discussion.

Example: Overrunning and underrunning.

Suppose $K = 2$ analyses were planned for a trial where the one-sided significance level of 0.05 was specified in the protocol. At the design stage, the uniform (linear) Type I error spending function $\alpha(t) = 0.05t$ was chosen for monitoring. The expected total number of events is 200 under the null hypothesis, but is 100 under the alternative. Suppose that there are 50 events at the first analysis.

If we have chosen the information time scale based on the null hypothesis, then $t_1 = 50/200 = 0.25$, and $\alpha(t_1) = 0.0125$. The group sequential boundary b_1 such that $P\{Z_{t_1} \geq b_1\} = 0.0125$ is $b_1 = 2.24$. Suppose, at the final analysis, the number of events is truly 200, as expected under the null hypothesis, the correlation between Z_{t_1} and Z_{t_2} is $(t_1/t_2)^{1/2} = (0.25)^{1/2} = 0.5$. Thus the group sequential boundary b_2 such that $P\{Z_{t_1} < 2.24, Z_{t_2} \geq b_2\} = 0.05 - 0.0125 = 0.0375$ is $b_2 = 1.74$.

However, if the realized number of events at the final analysis turns out to be 100 instead (i.e., an underrunning situation), then the “true” t_1 should be $50/100 = 0.5$ and we should have spent $\alpha(t_1) = 0.025$. But we cannot go back in time because we have already adopted $b_1 = 2.24$ for the test at the first analysis. What we can do is to recognize that 1) at $t_1 = 0.5$ (not 0.25), we used $\alpha(t_1) = 0.0125$; and 2) the

correlation between Z_{t_1} and Z_{t_2} is $(t_1/t_2)^{1/2} = (0.5)^{1/2}$. Thus the group sequential boundary b_2 such that $P\{Z_{t_1} < 2.24, Z_{t_2} \geq b_2\} = 0.05 - 0.0125 = 0.0375$ is $b_2 = 1.70$.

From the fact that $\alpha(0.5) = 0.0125 < 0.025$ (see Point 1 above), we see that the Type I error spending function was no longer the uniform (linear) spending function. Instead, it is a convex function, running under the uniform (linear) spending function.

The overrunning case is the opposite when 100 events under the alternative hypothesis was used to estimate t_1 , but it turned out that 200 events occurred at the final analysis. The consequence is that the linear spending function is altered to a concave function running over it.

Other Group Sequential Procedures: Stochastic Curtailment and Bayesian Methods

Extending the repeated significance testing procedure, a host of other group sequential procedures, frequentist or Bayesian, have been proposed in the literature. These procedures do not necessarily specify formal boundaries. They can be outlined by the following principal steps at each interim analysis: 1) divide the outcome space (the treatment–difference scale) into a region where the test treatment is considered superior and a region where the control is considered superior; and 2) guide the trial to stop if the “chance” is minimal that the control is superior (in which case we would use the test treatment), or the same is true for the test treatment (in which case we would use the control treatment). The key is that this “chance” may be measured by the conditional,^[36] posterior,^[37,38] or predictive^[39] probability of what would be the final outcome given the data at the interim stage. We then set to find the regions in step 1 by requiring certain prespecified operating characteristic. The latter two (posterior and predictive) approaches are considered preferable by Bayesian statisticians.^[40] Lin et al.^[41] advanced a general theory on stochastic curtailment for censored survival data, which provided an analytical basis of several previously proposed methods based on conditional power.

OTHER MONITORING MATTERS

Data Monitoring Committee

In order to ensure objectivity and to safeguard the blinding of the study, trials that use a group sequential design often have a DMC to evaluate the interim analyses and relevant external information. The ICH guidelines (E9, Sections 4.5 and 4.6) also call for an external Independent Data Monitoring Committee (IDMC) to be formed for monitoring comparisons of efficacy and/or safety outcomes for trials that have major public health significance. The composition of a DMC membership usually involves disciplines in the following areas: clinical,

laboratory, epidemiology, biostatistics, data management, and ethics.^[42] The organization structure for a DMC to function varies from trial to trial. For example, government-sponsored collaborative trials usually have a different reporting structure from industry-sponsored trials. More references on the DMC issues can be found in Friedman and DeMets,^[43] Fleming and DeMets,^[44] DeMets et al.,^[45] Armstrong and Furberg,^[46] and Freidlin et al.^[47]

When a DMC reviews interim data, not only the outcome measures (primary and secondary) are involved, but also the recruitment, baseline variables, adverse effects, and compliance. An early termination of a trial is a serious matter, and considerations of such cover a wide range issues of baseline comparability, quality of the data, internal and external consistency of the results, risk and benefit ratio, length of the follow-up, public impact, and, of course, the probability of Type I and Type II errors. Group sequential boundaries often serve as an important “guide-line”—not an absolute “rule”—in the consideration.

Confidence Intervals Following A Group Sequential Test

In addition to the hypothesis testing, the calculation of a confidence interval for the difference of treatment effects θ is often an important part of data analysis. A problem analogous to an inflation of the Type I error arises when estimating the population parameters upon a termination of a group sequential procedure because the study is stopped preferentially when extreme data have been observed. Thus the usual fixed sample estimators are biased toward the extremes.

The duality of hypothesis testing and confidence interval estimation, however, is helpful for the problem. Basically, the confidence interval of θ involves an inversion of the acceptance region of the test statistic, which, in turn, in the sequential testing setting, involves the boundary crossed, the stopping stage ($i = M$), and the cumulative/partial sum of the observations at the stopping stage ($S_i = S_M$). Because the acceptance region is based on “nonextreme” results for the sufficient statistic (i, S_i), clearly an ordering on the outcome space of (i, S_i) is necessary. However, the best choice for such an ordering is not as clear, as pointed out by Emerson and Fleming,^[48] because the densities for the group sequential statistics lack a monotone likelihood ratio, so the theory regarding the uniformly most powerful tests and the uniformly most accurate confidence bounds does not apply. For the binomial parameter, Jennison and Turnbull^[49] proposed a procedure based on an intuitive outcome space ordering, in which results corresponding to earlier termination are regarded as more extreme than those that terminate later. This approach was then adopted to the normal mean case by Tsiatis et al.^[50] Chang and O’Brien^[51] investigated an ordering based on the likelihood ratio test for the binomial case, then Chang^[52] applied the same approach to the normal mean case. In both cases, the

likelihood ratio ordering of the sample space was shown to give shorter confidence intervals, but Emerson and Fleming^[48] reported that there were occasions in which the confidence set based on the likelihood ratio was not an interval. Emerson and Fleming^[48] proposed the ordering based on sample mean. Confidence intervals based on the sample mean ordering agree with the test results for all practical group sequential test designs, in the sense that a $(1 - 2\alpha)$ confidence interval will not include hypotheses that are rejected by a one-sided level- α group sequential test. For random group sizes, Emerson and Fleming also recommended the approximate confidence intervals based on the sample mean ordering as well as the approximation to the bias-adjusted mean method of Whitehead.^[53]

OTHER RECENT DEVELOPMENTS

Early Acceptance/Rejection of the Null Hypothesis, Type I and Type II Error Spending Function Approaches, and Group Sequential Equivalence Trials

Group sequential methods allowing an early acceptance of the null hypothesis for futility is as important and as popular as that for an early rejection of the null hypothesis in the sense of ethics and economics. DeMets and Ware^[54,55] have proposed modifications to the Pocock and the O’Brien–Fleming designs to test one-sided hypotheses with an allowance for an early stopping in favor of the null. Gould and Pecore^[56] proposed a procedure that takes the cost function into account and pointed out that, although these procedures are less powerful than the ones that allow early rejection only, the saving of cost can be much greater. Lan and Friedman,^[57] when emphasizing the one-sided null hypothesis as being a “harmful” direction for the test treatment, suggested a combination of Type I error spending function for the upper boundary and a conditional probability (stochastic curtailment) approach for the lower boundary. An extension of the idea of the flexible Type I error spending function is naturally the use of both Type I and Type II error spending functions in constructing the rejection and acceptance boundaries.^[58,59] Recently, the idea of both acceptance and rejection boundaries has also been extended to group sequential equivalence trials by Whitehead.^[60]

Sequential Monitoring Trials with Multiple (Survival) Endpoints

In addition to the interim analyses with repeated measurements previously mentioned (“Partial Sum Process with Independent Increments”), the design of group sequential clinical trials with multiple endpoints in general or in survival outcomes has also been considered by other authors. Tang et al.^[61] addressed multivariate normal distribution

based on the global test of O'Brien.^[62] Previously generated group sequential boundaries can be used for the combined test statistic. Lin^[63] proposed a sequential non-parametric testing for detecting the stochastic ordering of two multivariate distributions, with an emphasis on multiple time-to-event endpoints, based on the weighted sum of linear rank statistics. Stopping boundaries that preserve the overall significance level and that may lead to earlier trial termination than procedures based on univariate survival endpoints were computed. Williams^[64] extended the method of Lin by developing approaches for constructing repeated joint confidence regions for the hazard ratios, in the spirit of Jennison and Turnbull,^[65,66] based on inverting linear rank statistics in terms of the parameter of interest. Technically, in all procedures, the basic building blocks are the same as that laid out earlier in "Statistical Methods," namely, finding the joint distribution (asymptotically, if necessary) of the test statistic over time by forming the test statistic at each time as a sum of i.i.d. random variables. The property of partial sum with independent increment structure can then be utilized to fit in the flexible Type I (and Type II) spending function framework. Of course, as in the nonsequential case, for the procedure to be meaningful, the endpoints being combined need to be biologically related and to demonstrate treatment effects in the same direction.

Monitoring Design Specifications to Extend a Trial

The traditional group sequential methods, as just reviewed, have been focused on an early termination of a trial for ethical and economic reasons. Recently, considerable interest has also been expressed in possibly extending a trial beyond its originally planned sample size and/or duration, based on interim data. This occurs when the assumed values for some parameters used for the initial sample size and/or duration calculation are uncertain, as is often the case for a new treatment studied for an unfamiliar disease. For example, the within-group variance may be underestimated initially, or treatment effects are less, or patient accrual rate is slower, than anticipated. Allowing for an upward adjustment of the sample size or study duration can prevent a trial from being inconclusive because of underpower after investing considerable resources.

Most of the methods in this area suggest a two-stage design because of practical logistics involved in extending a trial. Wittes and Brittain^[67] termed the first stage/phase an *internal pilot study*. Their procedure was further studied in Refs. [68,69]. For the univariate continuous case, Shih^[70] and Gould and Shih^[71] proposed two procedures to estimate the within-group variance without breaking the treatment codes and to update the sample size needed to secure the planned power of the trial. A computer program is available on the Web.^[72] Shih and Gould^[73] extended the technique to the case of repeated measures where the

key response is the slope. Shih and Long^[74] examined the robustness of the procedure with unequal variances and center effects present in the trial. For the binary case, Gould^[75] suggested using the pooled event rate, and Shih and Zhao^[76] proposed a simple design with a dummy stratification to reestimate the sample size without breaking the treatment code at the interim stage. To keep the treatment code masked is essential for maintaining the integrity of a trial where no IDMC is employed. Because these procedures only increase sample size and/or duration when necessary (otherwise they keep the original size and/or the duration intact) through updating nuisance parameters without unveiling the treatment effects at the interim stage, the Type I error probability is affected minimally, if at all. Gould and Shih^[77] and Whitehead et al.^[78] recently further examined strategies to combine blinded sample size reestimation with the traditional group sequential designs. See recent reviews in Refs. [79,80].

When an IDMC was in place and unblinding was not an issue, Andersen,^[81] Henderson et al.,^[82] Proschan and Hunsberger,^[83] Cui et al.,^[84] and Li et al.^[85] proposed the use of conditional power for updating the design specifications and extending a trial. For more complex group sequential designs (e.g., multiple arm and factorial with arbitrary survival curves), Halpern and Brown^[86] and Natarajan et al.^[87] developed simulation-based programs. Recently, Scharfstein and Tsiatis^[88] proposed to use the interim data to revise the initial guesses in the computer simulation, and to update the study design to attain the maximum information.

REFERENCES

- Schultz, J.R.; Nichol, F.R.; Elfring, G.L.; Weed, S.D. Multiple stage procedure for drug screening. *Biometrics* **1973**, *29*, 293–300.
- Elfring, G.L.; Schultz, J.R. Group sequential designs for clinical trials. *Biometrics* **1973**, *29*, 471–477.
- Jennison, C.; Turnbull, B.W. Statistical approaches to interim monitoring of medical trials: A review and commentary. *Stat. Sci* **1990**, *5*, 299–317.
- Jennison, C.; Turnbull, B.W. *Group Sequential Methods With Applications to Clinical Trials*; Chapman & Hall/CRC: Boca Raton, FL, 2000.
- ICH Guidelines*; <http://www.fda.gov/cder/guidance/guidance.htm>
- Pocock, S.J. Group sequential methods in the design and analysis of clinical trials. *Biometrika* **1977**, *64*, 191–199.
- O'Brien, P.C.; Fleming, T.R. A multiple testing procedure for clinical trials. *Biometrics* **1979**, *35*, 549–556.
- Haybittle, J.L. Repeated assessment of results in clinical trials of cancer treatment. *Br. J. Radiol* **1971**, *44*, 793–797.
- Peto, R.; Pike, M.C.; Armitage, P.; Breslow, N.E.; Cox, D.R.; Howard, S.V.; Mantel, N.; McPherson, C.K.; Peto, J.; Smith, P.G. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *Br. J. Cancer* **1976**, *35*, 585–611.

10. Whitehead, J.; Stratton, I. Group sequential clinical trials with triangular continuation regions. *Biometrics* **1983**, *39*, 227–236.
11. Whitehead, J.; Brunier, H. The double triangular test: A sequential test for the two-sided alternative with early stopping under the null hypothesis. *Seq. Anal.* **1990**, *9*, 117–136.
12. Milton, R.C. Computer evaluation of the multivariate normal integral. *Technometrics* **1972**, *14* (4), 881–889.
13. Schervish, M.J. Multivariate normal probabilities with error bound. *Appl. Stat.* **1984**, *33*, 81–94.
14. *EaSt: A Software Package for the Design and Interim Monitoring of Group Sequential Clinical Trials*; Cytel Software Corporation: Cambridge, MA, 1992.
15. Bruce, A.G.; Emerson, S. *S + SeqStat: Reference Manual*; MathSoft, Inc, 1997. Research Report No. 51.
16. Brunier, H.; Whitehead, J. *PEST 3.0 Operating Manual*; Reading University, 1993.
17. Gu, M.; Lai, T.L. Determination of power and sample size in the design of clinical trials with failure—time endpoints and interim analyses. *Control. Clin. Trials* **1999**, *20*, 423–438.
18. Reboussin, D.M.; DeMets, D.L.; Kim, K.M.; Lan, K.K.G. Computations for group sequential boundaries using the Lan–DeMets spending function method. *Control. Clin. Trials* **2000**, *21*, 190–207.
19. Slud, E.R.; Wei, L.J. Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J. Am. Stat. Assoc.* **1982**, *77*, 862–868.
20. Lan, K.K.G.; DeMets, D.L. Discrete sequential boundaries for clinical trials. *Biometrika* **1983**, *70*, 659–663.
21. Kim, K.; DeMets, D.L. Design and analysis of group sequential tests based on type-I error spending rate function. *Biometrika* **1987**, *74*, 149–154.
22. The Scandinavian Simvastatin Survival Study Group Design and baseline results of the Scandinavian Simvastatin Survival Study of patients with stable angina and/or previous myocardial infarction. *Am. J. Cardiol.* **1993**, *71*, 393–400.
23. Hwang, I.K.; Shih, W.J.; deCani, J.S. Group sequential designs using a family of type I error probability spending functions. *Stat. Med.* **1990**, *9*, 1439–1445.
24. Wang, S.K.; Tsiatis, A.A. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **1987**, *43*, 193–199.
25. Tsiatis, A.A. The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika* **1981**, *68*, 311–315.
26. Tsiatis, A.A. Repeated significance testing for a general class of statistics used in censored survival analysis. *J. Am. Stat. Assoc.* **1982**, *77*, 855–861.
27. Jennison, C.; Turnbull, B.W. Group-sequential analysis incorporating covariates information. *J. Am. Stat. Assoc.* **1997**, *92*, 1330–1341.
28. Scharfstein, D.O.; Tsiatis, A.A.; Robbins, J.M. Semiparametric efficiency and its implication on the design and analysis of group sequential studies. *J. Am. Stat. Assoc.* **1997**, *92*, 1342–1350.
29. Sellke, T.; Siegmund, D. Sequential analysis of proportional hazards model. *Biometrika* **1983**, *70*, 315–326.
30. Wei, L.J.; Su, J.Q.; Lachin, J.M. Interim analyses with repeated measurements in a sequential clinical trial. *Biometrika* **1990**, *77*, 359–364.
31. Lee, J.W.; DeMets, D.L. Sequential comparison of changes with repeated measurements data. *J. Am. Stat. Assoc.* **1991**, *86*, 757–762.
32. Lachin, J.M. Group sequential monitoring of distribution-free analyses of repeated measures. *Stat. Med.* **1997**, *16*, 653–668.
33. Lan, K.K.G.; Zucker, D. Sequential monitoring for clinical trials: The role of information and Brownian motion. *Stat. Med.* **1993**, *12*, 753–765.
34. Lan, K.K.G.; Reboussin, D.M.; DeMets, D.L. Information and information fractions for designing sequential monitoring of clinical trials. *Commun. Stat., Theory Methods* **1994**, *23* (2), 403–420.
35. Kim, K.M.; Boucher, H.; Tsiatis, A.A. *Design and Analysis of Group Sequential Log Rank Tests in Maximum Duration versus Information Trials*; Department of Biostatistics, Harvard School of Public Health, 1993, Technical report.
36. Lan, K.K.G.; Simon, R.; Halperin, M. Stochastically curtailed tests in long term clinical trials. *Commun. Stat., Seq. Anal.* **1982**, *1*, 207–219.
37. Freedman, L.S.; Spiegelhalter, D.J. Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Control. Clin. Trials* **1989**, *10*, 357–367.
38. Fayers, P.M.; Ashby, D.; Parmar, M.K.B. Bayesian data monitoring in clinical trials. *Stat. Med.* **1997**, *16*, 1413–1430.
39. Spiegelhalter, D.J.; Freedman, L.S.; Blackburn, P.R. Monitoring clinical trials: Conditional or predictive power? *Control. Clin. Trials* **1986**, *7*, 8–17.
40. Berry, A. Interim analysis in clinical trials: Classical vs. Bayesian approaches. *Stat. Med.* **1985**, *4*, 521–526.
41. Lin, D.Y.; Yao, Q.; Ying, Z. A general theory on stochastic curtailment for censored survival data. *J. Am. Stat. Assoc.* **1999**, *94*, 510–521.
42. DeMets, D.L. Data monitoring and sequential analysis—an academic perspective. *J. Acquir. Immune Defic. Syndr.* **1990**, (Suppl. 2), S124–S133.
43. Friedman, L.; DeMets, D.L. The Data Monitoring Committee: How it operates and why. *IRB*, **1981**, *3*, 6–8. (April).
44. Fleming, T.R.; DeMets, D.L. Monitoring of clinical trials: Issues and recommendations. *Control. Clin. Trials* **1993**, *14*, 183–197.
45. DeMets, D.L.; Ellenberg, S.S.; Fleming, T.R.; Childress, J.F.; Mayer, K.H.; Pollard, R.B.; Rahal, J.J.; Walters, L.; O'Fallon, J.; Whitley-Williams, P.; Straus, S.; Sande, M.; Whitley, R.J. The data an safety monitoring board and acquired immune deficiency syndrome (AIDS) clinical trials. *Control. Clin. Trials* **1995**, *16*, 408–421.
46. Armstrong, P.W.; Furberg, C.D. Clinical trials data and safety monitoring boards: The search for a constitution. *Circulation* **1994**, *1* (Sess. 6).
47. Freidlin, B.; Korn, E.L.; George, S.L. Data Monitoring Committee and interim monitoring guidelines. *Control. Clin. Trials* **1999**, *20*, 395–407.
48. Emerson, S.S.; Fleming, T.R. Parameter estimation following group sequential hypothesis testing. *Biometrika* **1990**, *77*, 875–892.
49. Jennison, C.; Turnbull, B.W. Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. *Technometrics* **1983**, *25*, 49–58.

50. Tsiatis, A.A.; Rosner, G.L.; Mehta, C.R. Exact confidence intervals following a group sequential test. *Biometrics* **1984**, *40*, 797–803.
51. Chang, M.N.; O'Brien, P.C. Confidence intervals following group sequential tests. *Control. Clin. Trials* **1986**, *7*, 18–26.
52. Chang, M.N. Confidence intervals for a normal mean following a group sequential test. *Biometrics* **1989**, *45*, 247–254.
53. Whitehead, J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika* **1986**, *73*, 573–581.
54. DeMets, D.L.; Ware, J.H. Group sequential methods for clinical trials with a one-sided hypothesis. *Biometrika* **1980**, *67*, 651–660.
55. DeMets, D.L.; Ware, J.H. Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika* **1982**, *69*, 661–663.
56. Gould, A.L.; Pecore, V.J. Group sequential methods for clinical trials allowing early acceptance of H_0 and incorporating costs. *Biometrika* **1982**, *69*, 75–80.
57. Lan, K.K.G.; Friedman, L. Monitoring boundaries for adverse effects in long-term clinical trials. *Control. Clin. Trials* **1986**, *7*, 1–7.
58. Pampallona, S.; Tsiatis, A.A. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *J. Stat. Plan. Inference* **1994**, *42*, 19–35.
59. Chang, M.N.; Hwang, I.K.; Shih, W.J. Group sequential designs using, both type I and type II error probability spending functions. *Commun. Stat., Theory Methods* **1998**, *27* (6), 1323–1339.
60. Whitehead, J. Sequential designs for equivalence studies. *Stat. Med.* **1996**, *15*, 2703–2715.
61. Tang, D.I.; Gnecco, C.; Geller, N.L. Design of group sequential clinical trials with multiple endpoints. *J. Am. Stat. Assoc.* **1989**, *77*, 862–868.
62. O'Brien, P.C. Procedure for comparing samples with multiple endpoints. *Biometrics* **1984**, *40*, 1079–1087.
63. Lin, D.Y. Nonparametric sequential testing in clinical trials with incomplete multivariate observations. *Biometrika* **1991**, *78*, 123–131.
64. Williams, P. Sequential monitoring of clinical trials with multiple survival endpoints. *Stat. Med.* **1996**, *15*, 2341–2357.
65. Jennison, C.; Turnbull, B.W. Repeated confidence intervals for group sequential clinical trials. *Control. Clin. Trials* **1984**, *5*, 33–45.
66. Jennison, C.; Turnbull, B.W. Interim analyses: The repeated confidence interval approach (with discussion). *J. R. Stat. Soc., B* **1989**, *51*, 305–361.
67. Wittes, J.; Brittain, E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Stat. Med.* **1990**, *9*, 65–72.
68. Zucker, D.; Wittes, J.; Schabenberger, O.; Brittain, E.; Prochan, M. Internal pilot studies: I. Type I error rate of the naïve t -test. *Stat. Med.* **1999**, *18*, 3481–3491.
69. Wittes, J.; Schabenberger, O.; Zucker, D.; Brittain, E. Internal pilot studies: II. Comparison of various procedures. *Stat. Med.* **1999**, *18*, 3493–3509.
70. Shih, W.J. Sample Size Reestimation in Clinical Trials. In *Biopharmaceutical Sequential Statistical Applications*; Peace, K.E., Ed.; Marcel Dekker: New York, 1992, 285–301.
71. Gould, A.L.; Shih, W.J. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Commun. Stat., Theory Methods* **1992**, *21* (10), 2833–2853.
72. Wang, W. *REEST.ZIP*; 1999; 3–21. StatLib: <http://lib.stat.cmu.edu/DOS/general/REEST.ZIP>.
73. Shih, W.J.; Gould, A.L. Re-evaluating design specifications of longitudinal clinical trials without unblinding when the key response is rate of change. *Stat. Med.* **1995**, *14*, 2239–2248.
74. Shih, W.J.; Long, J. Blinded sample size re-estimation with unequal variances and center effects in clinical trials. *Commun. Stat., Theory Methods* **1998**, *27* (2), 395–408.
75. Gould, A.L. Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. *Stat. Med.* **1992**, *11*, 55–66.
76. Shih, W.J.; Zhao, P.L. Design for sample size re-estimation with interim data for double-blind clinical trials with binary outcomes. *Stat. Med.* **1997**, *16*, 1913–1923.
77. Gould, A.L.; Shih, W.J. Modifying the design of ongoing trials without unblinding. *Stat. Med.* **1998**, *17*, 89–100.
78. Whitehead, J.; Whitehead, A.; Todd, S.; Bolland, K.; Sooriyachchi, M.R. Mid-trial design reviews for sequential clinical trials. *Stat. Med.* **2001**, *20*, 165–176.
79. Shih, W.J. Commentary: Sample size re-estimation—journey for a decade. *Stat. Med.* **2001**, *20*, 515–518.
80. Gould, A.L. Sample size re-estimation: Recent developments and practical considerations. *Stat. Med.* **2001**, *20*, 2625–2643.
81. Andersen, P.K. Conditional power calculation as an aid in the decision whether to consider a clinical trial. *Control. Clin. Trials* **1987**, *8*, 67–74.
82. Henderson, G.W.G.; Fisher, S.G.; Weber, L.; Hammermeister, K.E.; Sethi, G. Conditional power for arbitrary survival curves to decide whether to extend a clinical trial. *Control. Clin. Trials* **1991**, *12*, 304–313.
83. Proschan, M.A.; Hunsberger, S.A. Design extension of studies based on conditional power. *Biometrics* **1995**, *51*, 1315–1324.
84. Cui, L.; Hung, J.; Wang, S.J. Modification of samples size in group sequential clinical trials. *Biometrics* **1999**, *55*, 853–857.
85. Li, G.; Shih, W.J.; Xie, T.; Lu, J. A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics* **2002**, *3*, 277–287.
86. Halpern, J.; Brown, B.W. A computer program for designing clinical trials with arbitrary survival curves and group sequential testing. *Stat. Med.* **1993**, *14*, 109–122.
87. Natarajan, R.; Turnbull, B.W.; Slate, E.H.; Clark, L.C. A computer program for sample size and power calculation in the design of multiple-arm and factorial clinical trials with survival time endpoints. *Comput. Methods Programs Biomed.* **1996**, *49*, 137–147.
88. Scharfstein, D.O.; Tsiatis, A.A. The use of simulation and bootstrap in information-based group sequential studies. *Stat. Med.* **1998**, *17*, 75–87.

Group Sequential Tests and Variance Heterogeneity: Clinical Trials

Keumhee Carriere Chough

University of Alberta, Edmonton, Alberta, Canada

Abdulkadir Hussein

University of Windsor, Windsor, Ontario, Canada

Taesung Park

Seoul National University, Seoul, South Korea

Abstract

Group Sequential Tests and Variance Heterogeneity in Clinical Trials In clinical trials, investigators conduct a series of interim analyses of such summary statistics as means and variances in order to decide whether to terminate the trial early. The focus of this entry, is a discussion of the variance heterogeneity problem and its impact on the Type I error of group sequential test procedures in clinical trials.

INTRODUCTION

In clinical trials, investigators conduct a series of interim analyses of such summary statistics as means and variances in order to decide whether to terminate the trial early. If a treatment is deemed beneficial to patients, it is more ethical not to wait until the trial ends to provide patients with a better treatment.^[1–4] In multiarmed trials involving g populations/treatments, investigators must not only conduct an analysis of the variances in the data but also deal with the usual multiple comparison problems associated with this process.^[5,6]

In addition, the trial can suffer from the variance heterogeneity problem.^[7] Variance heterogeneity problem can occur in any empirical investigation and it is important for practitioners to be fully aware of the consequences.^[8] When such a problem is identified in the trials, investigators have dealt with the problem by transforming the data to achieve variance homogeneity. If transforming the data is undesirable or produces unsatisfactory results, some studies have suggested the use of nonparametric tests.

In this entry, we discuss the variance heterogeneity problem and its impact on the Type I error of group sequential test procedures in clinical trials. We discuss an alternative that is based on approximation theory while using common procedures.^[7] Because group sequential procedures typically involve small samples at each interim analysis and given that most current methods do not allow for corrections for heterogeneity, the impact on the type I error of the procedure can be quite significant.^[7]

We present an application of the significance level method^[1] of group sequential procedures for such

situations. We illustrate the application by using Pocock^[2]- and O'Brien–Fleming^[3]-type group sequential boundaries to compare the means of continuous treatment outcomes. This entry is organized as follows. In Group Sequential Procedures the Pocock and O'Brien–Fleming methods and the significance level approach are defined. The section Group Sequential Procedures Under Variance Heterogeneity introduces an approximate degrees of freedom correction method for the significance level approach to account for heterogeneity. The next section extends to multiarmed trials and discusses contrasts and trend tests as well as multiple comparison problems. Concluding remarks are provided in the section on related issues.

GROUP SEQUENTIAL PROCEDURES

In a sequential clinical trial comparing two treatments, suppose a total of K interim analyses are scheduled to test the two treatment means

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_0 : \mu_1 \neq \mu_2 \quad (1)$$

and σ_1^2, σ_2^2 are known or unknown nuisance parameters. We denote $\rho = \sigma_1/\sigma_2$. Consider responses $(x_{11}, x_{12}, \dots, x_{1n_{1k}}, \dots)$ from the distribution $N(\mu_1, \sigma_1^2)$ for treatment A, and $(x_{21}, x_{22}, \dots, x_{2n_{2k}}, \dots)$ from the distribution $N(\mu_2, \sigma_2^2)$ for treatment B. At the k th interim analysis, let $n_k = n_{1k} + n_{2k}$ be the total cumulative number of observations on both treatment arms and let the summary statistics $\bar{x}_{1k}, \bar{x}_{2k}, s_{1k}^2, s_{2k}^2, s_{pk}^2$ represent, respectively, sample means, sample variances (unbiased), and the sample pooled estimator of variance

based on the n_{1k} and n_{2k} samples available at the k th interim investigation.

In general, at the k th interim analysis, the ratio

$$T_k = \frac{\hat{\delta}_k}{SE(\hat{\delta}_k)} \quad (2)$$

is used for testing purposes, where $\hat{\delta}_k = \bar{x}_{1k} - \bar{x}_{2k}$ and $SE(\hat{\delta}_k)$ is the estimated standard error of $\hat{\delta}_k$.

A fixed-sample testing procedure corresponds to a single analysis with $K = 1$. In this case, if the variances are known, the ratio becomes the normal Z-score. Hence, we reject H_0 at level α if $|T_1|$ exceeds $z(\alpha/2)$, the upper $(100\alpha/2)$ th percentile point of the standard normal distribution. On the other hand, if we can assume $\rho = 1$ so that the two variances have common but unknown value, then we can pool the sample variances to get a robust estimator of the common population variance by using $SE(\hat{\delta}_1) = s_{p1} \sqrt{1/n_1 + 1/n_2}$ in the ratio of Eq. (2). In this case the ratio Eq. (2) has the exact t -distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom. Therefore, we reject H_0 at level α if $|T_1|$ exceeds $t_{\nu}(\alpha/2)$, the upper $(100\alpha/2)$ th percentile point of a student's t -distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom. This test is known to be uniformly most powerful (UMP) under the given assumption.

The original group sequential procedures of Pocock^[2] and O'Brien–Fleming^[3] were developed for the known variance case with evenly incremented group sizes (i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and $n_{1k} = n_{2k} = mk$ for some constant m). To account for multiple testing, Pocock^[2] and O'Brien and Fleming^[3] suggested to reject H_0 at the k th interim analysis if the Z-score exceeds $c_k = C_p(K, \alpha)$ or $c_k = \sqrt{K/k} C_{\text{OBF}}(K, \alpha)$, respectively, where $C_p(K, \alpha)$ and $C_{\text{OBF}}(K, \alpha)$ are constants depending on α and K , obtained via numerical integrations that exploit the independent increment structure of the interim Z-scores (see also Armitage^[9]).

Furthermore, when the variances are equal but unknown, an approach originally suggested by Pocock and known as the “significance level approach” can be used.^[2] In the significance level approach, to account for estimating the denominator of Eq. (2), H_0 is rejected at level α at the k th interim analysis stage if $|T_k|$ exceeds $t_{\nu_k}(\alpha_k)$ with $SE(\hat{\delta}_k) = s_{pk} \sqrt{1/n_{1k} + 1/n_{2k}}$ where $\alpha_k = 1 - \Phi(c_k)$, c_k are the Pocock^[2] or O'Brien–Fleming^[3] critical boundaries for the known variance case, and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal and $\nu_k = n_k - 2$. This approach is applicable to a variety of different types of alternative hypotheses and can also account for deviations from the condition of equally spaced group sizes as long as these deviations are not too severe.^[4]

Last, for large samples, the semiparametric efficiency theory applies.^[10] The semiparametric efficiency theory^[10] states that the ratios in Eq. (2) at consecutive interim analyses, using a nonpooled standard error estimate for $\hat{\delta}_k$, have an independent increment structure and are jointly

normal for large samples. Thus, the O'Brien–Fleming and Pocock critical values, c_k , developed for the known variance case, can still be used even with nonpooled sample variances. See also Kittelson and Emerson^[11] and Jennison and Turnbull.^[12] Therefore, a procedure based on the large sample theory would reject H_0 , if the $|T_k|$ in Eq. (2) with $SE(\hat{\delta}_k) = \sqrt{(s_{1k}^2/n_{1k}) + (s_{2k}^2/n_{2k})}$ exceeds c_k obtained by either the Pocock^[2] or O'Brien–Fleming^[3] methods.

GROUP SEQUENTIAL PROCEDURES UNDER VARIANCE HETEROGENEITY

When variance heterogeneity is present, we cannot assume $\rho = 1$, and the variance estimator cannot be pooled. Therefore, the ratio in Eq. (2) does not have an exact t -distribution. There is no UMP test whose distribution is free of the nuisance parameter ρ .^[13,14] This problem is known as the Behrens–Fisher (BF) problem. Hussein and Carriere^[7] proposed a modification of the Pocock^[2] and O'Brien–Fleming^[3]-type group sequential t -test procedures^[4] to make the procedures suitable for use under variance heterogeneity in small sample interim analyses. Although there are no serious problems when the sample size is large or the design is balanced,^[15] modification is usually needed for small sample trials. Note that $\rho = 1$ is a default assumption in the most commonly used group sequential software, PASS^[16] and S + SeqTrial.^[17]

Many investigators have considered approximate solutions to variance heterogeneity problems in comparing two sample means. The Welch^[13] approximation utilizes the fact that the distribution of the random variable $(SE(\hat{\delta}))^2 / \sigma^2$ is approximated by $e\chi_{\nu'}^2 / \nu'$. Then, by equating their first two moments, we solve for e and ν' . Welch^[13] found that $e = 1$ and

$$\nu'^{-1} = c^2(n_1 - 1)^{-1} + (1 - c)^2(n_2 - 1)^{-1} \quad (3)$$

with $c = (s_1^2/n_1) / (SE(\hat{\delta}))^2$ where n_1 and n_2 are the sizes of the two samples. Hence, the ratio in Eq. (2) will have an approximate t -distribution with ν' degrees of freedom, which depends on an estimate of ρ from the sample. This idea has been extended in various contexts and settings.^[18–23]

In group sequential trials using the significance level approach with planned K analyses, Welch's approximation can easily apply at each k interim analysis. Hussein and Carriere^[7] used the significance level approach of Pocock^[2–4] along with Welch's correction^[13] to the degrees of freedom of the two-sample t -statistic under variance heterogeneity. The procedure using the significance level approach with Welch's correction would reject H_0 at level α if $|T_k| \geq t_{\nu_k}(\alpha_k)$ with $SE(\hat{\delta}_k) = \sqrt{(s_{1k}^2/n_{1k}) + (s_{2k}^2/n_{2k})}$, and

$\alpha_k = 1 - \Phi(c_k)$. A rearrangement of Eq. (3) leads to the approximate degrees of freedom at the k th analysis as

$$\nu'_k = \frac{\left(\left(\frac{s_{1k}^2}{n_{1k}} \right) + \left(\frac{s_{2k}^2}{n_{2k}} \right) \right)^2}{\left(\frac{s_{1k}^4}{(n_{1k}^2(n_{1k}-1))} \right) + \left(\frac{s_{2k}^4}{(n_{2k}^2(n_{2k}-1))} \right)} \quad (4)$$

for $k = 1, \dots, K$. Note that for equal sample sizes and equal sample variances, it can be shown that the ν'_k reduces to $\nu_k = n_k - 2$. However, this correction generally leads to a narrower confidence interval when the variances are actually equal.^[13]

Hussein and Carriere^[7] reported the performance of their modified group sequential method using the Pocock and O'Brien–Fleming procedure in simulation, and compared the results with those of the large sample method mentioned above. Both balanced and unbalanced designs were examined to see the effect of having large variances associated with a small sample size (negative association) and large variances with a large sample size (positive association). The case of $\rho = \sigma_1^2 / \sigma_2^2 < 1$ corresponds to the design where the larger variance is associated with the larger sample, while the case of $\rho > 1$ corresponds to the design where the larger variance is associated with the smaller sample. The modified procedure was more accurate than the method based on large sample theory, regardless of how well balanced the design is and how large the degree of variance heterogeneity.

Specifically, when an investigator uses the large sample method, the empirical size deviates quite considerably from the nominal level, especially for small samples. This is true even for the case of equal variances in balanced designs. However, the inaccuracy in this case may be purely due to the use of a Z -value when the critical value from a t -distribution should have been used for small sample cases. The inflation of the empirical size becomes worse as the number of interim analyses, K , increases, and as ρ deviates from 1 for balanced cases, or as ρ increases for unbalanced cases (i.e., when small samples are associated with large variances). Especially when using Pocock's procedure, the deviation of the empirical Type I error from the nominal level is greater than those described above when $K > 1$. This remains true even for a total sample size as large as $n_K = 480$. This may be because Pocock's procedure has narrower boundaries at the early stages of analyses when the sample is small and the large sample theory is more vulnerable.

On the other hand, the empirical sizes of the Welch corrected method are in excellent agreement with the nominal level for all K and ρ , whether or not the design is balanced.^[7] For a small sample test using Welch's correction, designs with $\rho = n_{1k}/n_{2k}$ were found to be optimal, in the sense that they maximize power.^[24] The performance of the large

sample method improves as the sample size increases and is comparable to small sample tests using Welch's correction. However, the approach using Welch's correction consistently produced excellent results. The empirical size based on the modified method is virtually identical to the nominal level, even at a small total sample size of $n_K = 60$. Also, the procedure using the assumption of $\rho = 1$ should not be used as it often produces extremely conservative results. See Hussein and Carriere^[7] for details.

EXTENSION TO MULTIARMED TRIALS

An immediate extension of the two-sample comparison is to multiarmed comparison problems with more than two treatments. Suppose that the accumulating data are from parallel g groups. That is, we have x_{ij} , $i = 1, \dots, g$, $j = 1, \dots, n_i$. At the k th interim analysis there are n_{ik} cumulative observations collected on the i th arm and, hence, a total number of subjects of $n_k = \sum_{i=1}^g n_{ik}$. Furthermore, the observations of the i th arm are assumed to be normally distributed with mean μ_i and variance σ_i^2 .

In such multiarmed trials, investigators may be interested in testing the global omnibus hypothesis that the means of all arms are equal

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_g \quad (5)$$

with some variance σ_i^2 . If the arms are homogenous (i.e., σ_i^2 are equal but unknown), then the usual analysis of variance principles apply and a group sequential procedure would reject H_0 at the k th interim analysis if

$$F_k = \frac{1}{(g-1)s_{pk}^2} \sum_{i=1}^g n_{ik} (\bar{x}_{ik} - \bar{x}_k)^2 \geq c_k \quad (6)$$

for any $k = 1, \dots, K$. Here, $\bar{x}_{ik}$, $\bar{x}_k$, and s_{pk}^2 are the sample mean for the i th arm, the grand sample mean for all arms and the pooled estimate of the common variance, all based on the cumulative data up to the k th interim analysis, respectively. At each k , the F_k follows an F -distribution with $(g-1)$ and $(n_k - g)$ degrees of freedom.

For equal group sizes, the monitoring boundary values, c_k , of Pocock or O'Brien–Fleming are obtained by using either the recursive integration formulae in Jennison and Turnbull^[25] or the significance level approach.^[1] In the latter case, c_k is given by $c_k = F_{g-1, (n_k-g)}^{-1}(1 - \alpha'_k)$ where $f_{\nu_1, \nu_2}^{-1}(1 - \alpha')$ denotes the $(1 - \alpha')$ -percentile point of F distribution with ν_1, ν_2 degrees of freedom, and $\alpha'_k = 1 - P(\chi_{g-1}^2 \leq c'_k)$ where c'_k are boundaries obtained for the case of common and known variance, i.e., for monitoring a χ^2 test statistic. The exact c'_k for monitoring a chi square with degrees of freedom up to 5 are tabulated in Jennison and Turnbull.^[25]

When the arms are heterogenous, the investigator can adopt the Welch approximate test (see, for example,

Welch^[26]) for the above global hypothesis. At the k th interim analysis, we reject the null hypothesis (5) if

$$F_k^w = \frac{\sum_{i=1}^g \left(\frac{\hat{w}_{ik}}{(g-1)} \right) (\bar{x}_{ik} - \bar{x}_{\cdot k})^2}{\left\{ 1 + \frac{2(g-2)}{g^2-1} \sum_{i=1}^g \frac{1}{n_{ik}-1} \left(1 - \frac{\hat{w}_{ik}}{\sum \hat{w}_{ik}} \right)^2 \right\}} \geq c_k \quad (7)$$

where $\hat{w}_{ik} = n_{ik} / s_{ik}^2$, n_{ik} is the cumulative sample in the i th treatment arm up to the k th interim analysis stage and s_{ik}^2 is the nonpooled within-arm variance estimator. The above statistic, F_k^w in Eq. (7), follows an approximate F_{ν_1, ν_2} distribution with degrees of freedom $\nu_1 = g - 1$ and

$$\nu_2 = \left\{ \frac{3}{g^2-1} \sum_{i=1}^g \frac{1}{n_{ik}-1} \left(1 - \frac{\hat{w}_{ik}}{\sum \hat{w}_{ik}} \right)^2 \right\}^{-1} \quad (8)$$

Therefore, the monitoring constants c_k can be obtained either by recursive integration or by using the significance level approach.

Example 1: Consider a trial comparing three treatments ($g = 3$). The investigator is interested in testing hypothesis (5) using the OBF-type of boundary and considers $K = 4$ interim analyses with equally spaced group sizes and overall significance level $\alpha = 0.05$. The variances of the three treatments are not assumed to be equal. Then from Jennison and Turnbull^[25] (Table 2b), we find $C_{\text{OBF}}(g-1, K, \alpha) = C_{\text{OBF}}(2, 4, 0.05) = 6.20$. Therefore, the exact monitoring boundaries for a χ^2 with two degrees of freedom are $c'_k = 6.20\sqrt{4/k}$, $k=1, 2, 3, 4$. Using Welch's statistic, we reject H_0 at the k th interim analysis if Eq. (7), based on the collected data, holds with $c_k = F_{2, \nu_2}^{-1}(1 - \alpha'_k)$, $\alpha'_k = 1 - P(\chi^2_2 \leq 6.20\sqrt{4/k})$ and ν_2 is obtained from Eq. (8) above. Note that, not only the statistic in the left-hand side of Eq. (7) but also c_k has to be computed from data, since ν_2 is a data-dependent quantity that cannot be computed ahead of time.

RELATED ISSUES

Multiple Comparisons

Once the omnibus test Eq. (5) is rejected, we are interested in performing multiple comparisons of the g arms to identify where the differences come from. A great deal of methodology has been developed for the fixed-sample case. In the context of group sequential procedures, only large sample methods have been suggested to date. Hellmich^[5] provides an excellent review for applying various multiple comparison group sequential procedures. In this entry, we

describe two approaches to the multiple testing problem that can easily utilize a combination of approximate solutions to the variance heterogeneity problem and the significance level approach. These procedures, suggested by Follmann, Proschan, and Geller^[6] are essentially pairwise comparison methods. They strongly control the family-wise Type I error rate in the sense that the probability of rejecting at least one of the true pairwise null hypotheses does not exceed the nominal α .

Suppose that all pairwise comparisons of g arms are of interest. We have $H_0^{ij} : \mu_i = \mu_j$, with $1 \leq i, j \leq g$ null hypotheses to be tested. To test each of these hypotheses, the usual t -test

$$|T_k^{ij}| = \left| \frac{\hat{\delta}_k^{ij}}{\text{SE}(\hat{\delta}_k^{ij})} \right|$$

is used at the k th interim analysis where $\hat{\delta}_k^{ij} = \bar{x}_{ik} - \bar{x}_{jk}$ is the difference between sample means for the i th and j th arm based on data collected up to the k th interim analysis stage. When the variances are heterogeneous, the standard error of $\hat{\delta}_k^{ij}$ is the nonpooled estimator, $\text{SE}(\hat{\delta}_k^{ij}) = \sqrt{(s_{ik}^2/n_{ik}) + (s_{jk}^2/n_{jk})}$, as previously described.

The first and simplest approach to controlling the family-wise type I error is to monitor each pairwise test statistic separately with the Bonferroni adjusted α . If we fix the nominal type I error to be α then the nominal for each pairwise comparison will be $Q^{\text{Bon}} = \alpha/I$ where $I = g(g-1)/2$ is the number of all possible pairwise hypotheses to be tested. Thus, each of these hypotheses is treated as an independent two-sample group sequential test, and its monitoring boundaries are obtained using the methods of Pocock, O'Brien–Fleming or others who have suggested appropriate solutions. This method is obviously very conservative and is expected to be low in power.

The second approach, as originally developed by Follmann, Proschan, and Geller^[6] is to treat the set of all test statistics for all pairwise comparisons and from all interim stages as one multivariate vector. Follmann, Proschan, and Geller^[6] showed that this vector is asymptotically multivariate normal with a certain covariance matrix. They assumed that, at each interim analysis, all pairwise hypothesis are evaluated using the same critical value (seeking equal evidence), all group variances are the same, and sample sizes are evenly distributed across arms and interim analyses. Then, under these assumptions, they obtained Pocock- and OBF-type boundaries (see also Table I of Hellmich^[5]) by simulating the jointly normal vectors.

We can adapt the above approaches to the multiple comparison problem with variance heterogeneity by using the tabled constants^[5,6] as follows: reject H_0^{ij} at the k th interim analysis if $|T_k^{ij}| \geq t_{\nu_k}(\alpha'_k)$ where $\alpha'_k = 1 - \Phi^{-1}(c_k)$ and c_k is the boundary obtained from the table in Follmann, Proschan, and Geller^[6] for the k th interim

analysis and for the appropriate number of arms g and number of interim analyses K . The degrees of freedom, ν_k^{ij} , can be set as those of Welch's correction, calculated using the formula Eq. (4) given in "Extension to Multiarmed Trials."

Using the Follmann–Proschan–Geller approach, if inferior arms are dropped during the course of the trial, one has to make sure that information fraction (i.e., the ratio of the actual number to the planned number of patients for the current arms) is not less than that at the preceding interim analysis and the information time must not decrease. In particular, this is satisfied if the total number of planned subjects is evenly distributed across arms and interim analysis. Thus, inferior arms may be dropped and the procedure may be based only on samples from the remaining arms using the original boundaries. However, this can cause the procedure to be conservative and reduce power. To alleviate this problem, Follmann, Proschan, and Geller^[6] propose a sequential rejection procedure, which strongly controls the type I error for three or less number of arms (in the case of all pairwise comparisons) and for any number of arms (in the case of comparisons to a common control). After dropping inferior arms, boundaries are recalculated based on the current number of arms and the originally planned number of interim analyses (see the following example). The Bonferroni approach is not affected because each pairwise comparison is designed separately. Note that, because of the use of the significance level approach, the above procedure allows for unequally incremented group sizes. In other words, the fraction of patients recruited between each two interim analyses need not be exactly $1/K$ of the total number of patients that the investigator plans to have for the study, although such a deviation from the planned equally spaced groups should be kept to a minimum.

Example 1: Consider all pairwise comparisons among $g = 4$ treatment arms. Suppose we plan $K = 4$ interim analyses using OBF boundaries and $\alpha = 0.05$. Thus, $I = 4(4 - 1)/2 = 6$ different pairs of treatments are to be compared. Whether using the Bonferroni adjustment method or the Follmann–Proschan–Geller method,^[6] each pair of treatments (i, j) is monitored separately and the hypothesis rejected at the k th interim analysis if its corresponding nonpooled t -test statistic $|t_k^{ij}|$ exceeds $t_{\nu_k^{ij}}(\alpha'_k)$ where $\alpha'_k = 1 - \Phi^{-1}(c_k)$. If the investigator uses the Bonferroni adjustment, then a total nominal $\alpha^{\text{Bon}} = 0.05/6 = 0.0083$ is to be used for each pairwise comparison. Therefore, $c_k = C_{\text{OBF}}(K = 4, \alpha^{\text{Bon}} = 0.0083) \sqrt{4/k} = 2.6706 \sqrt{4/k}$ for $k = 1, 2, 3, 4$. The OBF constants $C_{\text{OBF}}(K, \alpha)$ are tabulated in Jennison and Turnbull^[25] or can be calculated using software.^[16,17] On the other hand, an investigator using the Follmann–Proschan–Geller method,^[6] which is essentially Tukey's pairwise comparison method, obtains $C_{\text{OBF}}(g, K, \alpha) = C_{\text{OBF}}(4, 4, 0.05) = 2.60$ from Table 1 of Hellmich^[5] (corresponding to

$g = A = 4, R = K = 4$ and all pairwise $\alpha = 0.05$). Thus, $c_k = 2.60 \sqrt{4/k}$ for $k = 1, 2, 3, 4$. Note that the constant C_{OBF} depends on the number of arms g , the number of analyses K , and the level α .

Now we remove inferior arms while continuing to use the originally planned boundaries in the above. However, a sequential rejection approach can be used in some situations to maintain power, as the following example illustrates.

Example 2: Consider the previous example with $g = 4$ and $K = 4$ with one of them a control. Now suppose that three treatments are compared to the control so that $I = 3$ hypotheses, $H_0^{1j}; j = 2, 3, 4$, have to be tested. Using OBF-type of boundary tabulated in Hellmich,^[5] and the sequential rejection procedure,^[6] we have $C_{\text{OBF}}(4, 4, 0.05) = 2.40$. Hence, we reject H_0^{1j} at the k th interim analysis if $|T_k^{1j}| \geq \nu_k^{1j}(\alpha'_k)$ where $\alpha'_k = 1 - \Phi^{-1}(c_k)$, $c_k = 2.40 \sqrt{4/k}$, $k = 1, 2, 3, 4$ and ν_k^{1j} are the Welch's degrees of freedom calculated from data at the k th interim analysis. For example, if arm 2 is deemed inferior and dropped at the first interim analysis ($k = 1$), we obtain $C_{\text{OBF}}(3, 4, 0.05) = 2.26$ and the monitoring boundaries $c_k = 2.26 \sqrt{4/k}$ at subsequent stages $k = 2, 3, 4$ for $g = 3$ arms and $K = 4$ interim analyses.

Contrasts and Trend Tests

Assuming that one of the global hypotheses of interest is $H_0: \sum_{i=1}^g a_i \mu_i = \mu_0$ where a_i and μ_0 are known constants with $\sum_i a_i = 0$, the test statistic at the k th interim analysis is

$$T_k = \frac{\sum_{i=1}^g a_i \bar{x}_{ik} - \mu_0}{\sqrt{\sum_{i=1}^g a_i^2 s_{ik}^2 / n_{ik}}}$$

In a fixed sample design, if the variances are homogeneous, this statistic based on a pooled sample variance follows a t -distribution with $n - g$ degrees of freedom. The trend tests that are often used in pharmacology studies fall into this category of linear hypothesis, with $a_i = i - \sum_{i=1}^g i / g$ and $\mu_0 = 0$.

In the heterogeneous case, the investigator can use Welch's generalization of the two-sample heterogeneous problem to the multisample linear hypothesis.^[27,28] When adopting this correction, one rejects H_0 at the k th interim analysis if $|T_k| \geq t_{\nu_k}(\alpha_k)$ where $\alpha_k = 1 - \Phi(c_k)$, c_k is the boundary value for the equal and known variance case and the degrees of freedom for the t -distribution are derived from Welch^[13] and are given by

$$\nu_k = \frac{\left[\sum_i a_i^2 s_{ik}^2 \right]^2}{\sum_i a_i^2 s_{ik}^4 / (n_{ik} - 1)} \quad (9)$$

Example 1: If an investigator has three treatment arms with unknown variances and wishes to test whether the means exhibit a linear trend, he or she can choose the contrast $(a_1, a_2, a_3) = (-1, 0, 1)$ so that $H_0: -1 \times \mu_1 + 0 \times \mu_2 + 1 \times \mu_3 = 0$ vs. $H_a: -1 \times \mu_1 + 0 \times \mu_2 + 1 \times \mu_3 \neq 0$. Suppose we plan four interim analyses and equally incremented sample sizes at each interim analysis (for example, at each k , $\sum_{i=1}^3 n_{ik} = 3mk$ for some constant m), a nominal $\alpha = 0.05$, and OBF monitoring boundaries. For this setup if the variances are equal and known, the upper four critical points for the OBF boundary are $c_k = C_{\text{OBF}}(K, \alpha) \times \sqrt{k/K} = C_{\text{OBF}}(4, 0.05) \sqrt{k/4}$ where the C_{OBF} is read from a table.^[1] Thus, the four positive boundary points are $c_k = 2.024 \sqrt{k/4}$ for $k = 1, 2, 3, 4$. Now, since the variances are possibly heterogenous, we reject H_0 at the k th interim analysis if

$$|T_k| = \left| \frac{\sum_{i=1}^3 a_i \bar{x}_{ik}}{\sum_{i=1}^3 a_i^2 s_{ik}^2 / n_{ik}} \right| \geq t_{\nu_k}(\alpha'_k)$$

where $\alpha'_k = 1 - \Phi(c_k)$ using the c_k found above. The degrees of freedom ν_k are calculated using Eq. (9).

CONCLUSIONS

This entry deals with clinical trials where sequential testing procedures are applied to compare summary statistics in several interim investigations. Most such clinical trials currently use the large sample method to test parameters of interest, and consequently do not produce valid study results unless the sample size is large. In fact, the available sample sizes are often very small.

We have focused on small sample problems in group sequential procedures, which can be accompanied by the variance heterogeneity problem. We have discussed Welch's corrected degrees of freedom method in the significance level approach^[4] of Pocock and O'Brien–Fleming-type group sequential methods for sequentially testing the equality of the means of two normally distributed populations. This modification produced excellent results in most cases, as the modification kept the Type I error rate at a nominal level, even for small samples. Various investigators have also reported that Welch's approximate degrees of freedom method is robust against mild departures from the normality assumption.^[8,18–23]

Even at a relatively large total sample size of $n_K = 480$, the large sample method is not entirely satisfactory, especially when using Pocock's procedure with large K . Therefore, it is important to consider modifying the significance level approach, for example, by using the method suggested by Welch. In this entry, we also discussed extensions to multiarmed trials and some linear hypothesis tests.

ACKNOWLEDGEMENT

This work was funded in part by grants from the Natural Science and Engineering Research Council of Canada to K.C. Carriere and Abdulkadir Hussein, and the National Research Laboratory Program of the Korean Ministry of Science and Technology to Taesung Park.

REFERENCES

1. Jennison, C.; Turnbull, B.W. *Group Sequential Methods with Applications to Clinical Trials*; Chapman & Hall/CRC Press: Boca Raton, FL, 2000.
2. Pocock, J.S. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. **1977**, *64*, 191–199.
3. O'Brien, P.C.; Fleming, T.R. A multiple testing procedure for clinical trials. *Biometrics*. **1979**, *35*, 549–556.
4. Jennison, C.; Turnbull, B.W. On group sequential tests for data in unequally sized groups and with unknown variance. *J. Statist. Plann. Inference*. **2001**, *96*, 263–288.
5. Hellmich, M. Monitoring clinical trials with multiple arms. *Biometrics*. **2001**, *57*, 892–898.
6. Follmann, D.A.; Proschan, M.A.; Geller, N.L. Monitoring pairwise comparisons in multiarmed clinical trials. *Biometrics*. **1994**, *50*, 325–336.
7. Hussein, A.; Carrière, K.C. On group sequential procedures under variance heterogeneity statistical methods in medical research. **2005**, *14*, 121–128.
8. Hussein, A.; Carriere, K.C. Robustness of procedures for the Behrens–Fisher problems: extension to bivariate normal mixtures. *Comm. Statist. Part B Simulation Comput.* **2001**, *30*, 831–845.
9. Armitage, P. *Sequential Medical Trials*; Oxford: Blackwell, 1975.
10. Scharfstein, D.; Tsiatis, A.; Robins, J. Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *J. Am. Statist. Assoc.* **1997**, *92*, 1342–1350.
11. Kittelson, J.; Emerson, S. A unifying family of group sequential test designs. *Biometrics*. **1999**, *55*, 874–882.
12. Jennison, C.; Turnbull, B.W. Group-sequential analysis incorporating covariate information. *J. Am. Statist. Assoc.* **1997**, *92*, 1330–1341.
13. Welch, B.L. Specification of rules for rejecting too variable a product, with particular reference to an electric lamp problem. *J. R. Statist. Soc. Suppl.* **1936**, *3*, 9–48.
14. Banerjee, S. On confidence interval for two-mean problem based on separate estimates of the variances and tabulated values of t -table. *Sankhya*. **1961**, *23*, 359–378.
15. Wang, H.; Chow, S.-C. A practical approach for comparing means of two groups without equal variance assumption. *Statist. Med.* **2002**, *21*, 3137–3151.
16. Hintze, J. *NCSS and PASS. Number Cruncher Statistical Systems*; Kaysville, UI, 2001.
17. Emerson, S. *S + SeqTrial Technical Overview*; MathSoft: Seattle, WA, 2000.
18. Bhoj, D.D. An approximate solution to the Behrens–Fisher problem. *Biom. J.* **1993**, *35*, 635–640.

19. Yao, Y. An approximate degrees of freedom solution to the multivariate Behrens–Fisher problem. *Biometrika*. **1965**, 52, 139–147.
20. Kim, S.-J. A practical solution to the Behrens–Fisher problem. *Biometrika*. **1992**, 79, 171–176.
21. Scheffé, J. Practical solutions of the Behrens–Fisher problem. *J. Am. Statist. Assoc.* **1970**, 65, 1501–1509.
22. de la Rey, N.; Nel, D.G. A comparison of the significance levels and power functions of several solutions to the multivariate Behrens–Fisher problem. *S. Afri. Statist. J.* **1993**, 27, 129–148.
23. Algina, J.; Oshima, T.; Tang, K.L. Robustness of Yao’s, James’, and Johansen’s tests under variance-covariance heteroscedasticity and nonnormality. *J. Educ. Statist.* **1995**, 16, 125–139.
24. Lee, A. Optimal sample sizes determined by two-sample Welch’s *t*-test. *Comm. Statist. Part B Simulation Comput.* **1992**, 21, 689–696.
25. Jennison, C.; Turnbull, B.W. Exact calculations for sequential *t*, χ^2 and *F* tests *Biometrika*. **1991**, 78, 133–141.
26. Welch, B.L. On the comparison of several mean values: an alternative approach. *Biometrika*. **1951**, 38 (3/4), 330–336.
27. Welch, B.L. The generalization of “student’s” problem when several different population variances are involved. *Biometrika*. **1947**, 34, 28–35.
28. Welch, B.L. On the comparison of several mean values: an alternative approach. *Biometrika*. **1951**, 38, 330–336.

Human Drug Abuse Trials

Qianyu Dang

CDER, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

Human drug abuse study is an important part of evaluating a drug's effect on the human central nervous system (CNS) to produce subjective responses related to abuse potential. There are three types of pharmacodynamics (PD) clinical studies—the human abuse potential (HAP) study, the human abuse study for abuse-deterrent opioids, and the human drug withdrawal study. All of them play vital roles in a nation's effort to curb illegal use of prescription opioid drugs and other controlled substances.

Keywords: Prescription opioids; Crossover study; Abuse potential studies; Visual analogue scales.

INTRODUCTION

Human drug abuse study is an important part of evaluating a drug's effect on the human central nervous system (CNS) to produce subjective responses related to abuse potential. There are three types of pharmacodynamics (PD) clinical studies—the human abuse potential (HAP) study, the human abuse study for abuse-deterrent opioids, and the human drug withdrawal study. In April, 2015, the U.S. Food and Drug Administration (FDA) published the guidance on “Abuse-Deterrent Opioids—Evaluation and Labeling.”^[1] FDA issued the industry guidance of “Assessment of Abuse Potential of Drugs”^[2] in January, 2017. As of March 2017, there is no FDA guidance for human drug withdraw study.

HUMAN ABUSE STUDY FOR ABUSE-DETERRENT OPIOIDS

Prescription opioid drugs such as oxycodone, hydrocodone, and oxymorphone are primarily used for pain management. However, abuse of these drugs for recreational purpose has become a serious public health crisis in the United States. An important effort to curb opioid drug abuse is to develop relatively safer opioid analgesics that are formulated to deter abuse while still provide pain management for patients. Numerous new technologies have been developed to create new opioid drug formulations to deter abuse. The abuse-deterrence effect needs to be fully evaluated before the product is approved by the regulatory authority with such claims. The premarket evaluations can be divided into three categories:

1. Laboratory-based in vitro manipulation and extraction studies (category 1)

2. Pharmacokinetic studies (category 2)
3. Clinical abuse potential studies (category 3)

The human abuse study is part of category 3 studies. The need and design of category 3 studies may depend on the results of category 1 and category 2 studies.

A human abuse study for abuse-deterrent opioid drugs is typically a randomized, double-blind, placebo-controlled, and positive-controlled crossover study.^[3] Study participants are usually experienced recreational drug users familiar with the opioid drug effects. The study is composed of a prequalification phase to screen participants and assessment phase in which only qualified subjects can participate. The route of administration (intranasal, oral, or intravenous), dose selection, manipulation mode, and sample preparation for the study should be carefully chosen based on factors including epidemiological data of abuse history as well as category 1 and category 2 study findings.

In both phases, the primary tools to assess the subjective drug effects are standardized instruments include visual analogue scales (VAS) and Profile of Mood States. VAS scores are used to measure drug liking, drug high, drug-taking again, overall drug liking, good effects, and bad effects. The measure can be unipolar (0–100, with 0 to be neutral) or bipolar (0–100, with 50 to be neutral) scale. Bipolar scale is recommended in 2015 FDA guidance for primary measure of drug-liking. Drug-liking is also the primary end point of interest in most drug abuse studies.

The purpose of prequalification phase is to ensure selected subjects can differentiate between placebo and an original immediate-release formulation of the same opioid being tested in an abuse-deterrent formulation in the next phase. Subjects qualified in the prespecified acceptable range of both placebo and positive control drugs will be selected to participate in the assessment phase.

The assessment phase usually has multiple doses of positive control, test drug, and placebo. It could also include different ways of manipulation of opioid drugs such as crushing and chewing. The study is validated through comparison of positive control (C) and placebo (P) while the efficacy of abuse-deterrent effect is proven through the comparison between the abuse-deterrent test drug (T) and the positive control without abuse-deterrent properties. The primary analysis is testing the differences of means for the $E_{\max}$ of primary measures. $E_{\max}$ is defined as the maximum of responses measured in the predefined time frame after drug dosing. The purpose is to prove that T has less abuse potential than C; so it is an efficacy study. The analysis population is usually the subjects who complete all treatments and have responses of primary measures within the prespecified time interval before and after drug $T_{\max}$. Descriptive statistics include tables and graphs displaying the outcomes of interest for each treatment group as well as the differences among them.^[4]

The primary statistical analysis of drug abuse studies for abuse-deterrent opioids includes the test of the following hypotheses with regard to primary outcomes such as drug-liking, drug-high, drug-taking again, etc.:^[5]

1. *The test for deterrent effect:* Comparing the means of C and T:

$$H_0 : \mu_c - \mu_T \leq \delta_1 \text{ versus } H_1 : \mu_c - \mu_T > \delta_1 \quad (1)$$

where the test margin of $\delta_1 = \delta^* \times (\mu_c - 50)$, $0 < \delta^* < 1$. In other words, the test margin has to be greater than 0.

2. *The validation test:* Comparing the means of C and P:

$$H_0 : \mu_c - \mu_p \leq \delta_2 \text{ versus } H_1 : \mu_c - \mu_p > \delta_2 \quad (2)$$

where the test margin of $\delta_2 \geq 15$.

Both tests are two sided with the significance level to be 2.5%. All the test margins need to be predefined. The choice of δ^* is critical and should be predetermined based on preclinical and other information of the innovative formulation's abuse deterrent-effect. A closed testing procedure should be applied by starting with a prespecified and clinically meaningful δ^* . If the result of hypothesis test 1 is statistically significant, then test the hypothesis again with a 0.05 increment of δ^* until an insignificant result is achieved.

Statistical methods such as mixed models can be used to implement the hypothesis testing described above. Models should include variables that may affect how the product is used and also other related confounders (e.g., geographic variability and demographic characteristics) as fixed effects. If the model assumptions are not met, then paired t-test approaches may be considered before applying non-parametric test of median differences.

The secondary statistical analysis is based on the percentage reduction of drug-liking for test formulation T from C relative to P. For each subject, the percentage reduction is defined as:

$$\% \text{reduction} = \frac{c_i - t_i}{c_i - p_i} \times 100\%, i = 1, 2, \dots, n \quad (3)$$

Here c_i , t_i , and p_i are the $E_{\max}$ drug-liking bipolar VAS score for subject I in three treatment arms, respectively. In the situation of large Placebo response of p_i , the following definition should be applied instead:

$$\% \text{reduction} = \frac{c_i - t_i}{c_i - 50} \times 100\%, i = 1, 2, \dots, n \quad (4)$$

In case $c_i \leq 50$, the reduction will be assigned as 0.

There are two types of statistical tests for percentage reduction of abuse potential, responder analysis, and median percentage reduction. Either one can be used to support the results from primary analysis. A responder is a participant who had a minimum $\delta^* 100\%$ of reduction in primary outcomes as defined above. For example, the hypothesis test of half of the participants who are responders with at least 10% reduction is:

$$H_0 : p^* \leq 50\% \text{ versus } H_1 : p^* > 50\% \quad (5)$$

Where p^* is the percentage of responders with 10% or more reduction.

The other approach is testing the median of percentage reduction (ptr) among all subjects in the analysis data set. For example, the hypothesis test for median percentage reduction to be at least 10% is:

$$H_0 : \text{median(ptr)} \leq 10\% \text{ versus } H_1 : \text{median(ptr)} > 10\% \quad (6)$$

The Wilcoxon-signed rank test can be applied if the percentage reduction distribution is symmetrical. Otherwise, the sign test can be used. All tests and their test parameters need to be prespecified and tested at 2.5% significance level.

HAP STUDY

Under the Controlled Substances Act [CSA; (21 U.S.C. 811(b), 811(c), 812(c))], any new drug product with the potential to be abused need to be assessed for HAP. The results can be used to propose drug scheduling. The results of human drug abuse potential studies, combined with other preclinical and clinical evaluations, are used for the determination of drug scheduling by The Drug Enforcement Administration.

The design of HAP study is very similar to abuse study for abuse-deterrent opioids. Typically it includes a prequalification phase and a treatment phase. Study participants are also experienced recreational drug users. The prequalification phase includes C and P. The same C drug will be used in the treatment phase. Only subjects with acceptable responses for both P and C drugs will be eligible for the treatment phase. The treatment phase usually has multiple doses of test drug, C drug and P arms. It is commonly designed as a randomized, double-blind, double-dummy, placebo- and active-controlled, crossover study. A balanced William square design is also recommended. The primary outcomes of HAP study may include the VAS score at $E_{\max}$ of drug-liking (at the moment), drug-high, as well as overall drug-liking and drug-taking again at 12 and 24 hrs after dosing. Other subjective measures, safety and physiological measures, and abuse-related adverse events will be used to determine whether the study shows any signal of abuse. Bipolar VAS is also recommended for the primary measurement of drug-liking.

The analysis population are also subjects as defined in the abuse-deterrent study. The statistical analysis begins with descriptive statistics making up tabulations and graphs to show the responses of interest for each treatment and for each paired difference among treatments. Since it is basically a safety study, the null hypothesis should be that the test drug has abuse potential and the alternative hypothesis should be that the test drug has no abuse potential. Following hypotheses need to be tested:

1. *The validation test: C has less abuse potential than P:*

$$H_0 : \mu_c - \mu_p \leq \delta_1 \text{ versus } H_a : \mu_c - \mu_p > \delta_1 \quad (7)$$

where $\delta_1 > 0$.

2. *Test drug has greater or equal abuse potential than C:*

$$H_0 : \mu_c - \mu_T \leq \delta_2 \text{ versus } H_a : \mu_c - \mu_T > \delta_2 \quad (8)$$

where $\delta_2 \geq 0$.

3. *Test drug has greater abuse potential than P:*

$$H_0 : \mu_T - \mu_p \geq \delta_3 \text{ versus } H_a : \mu_T - \mu_p < \delta_3 \quad (9)$$

where $\delta_3 > 0$.

The actual values of δ_1 , δ_2 , and δ_3 depend on abuse potential measures, the route of drug administration, and the type of C drug, respectively. All the margins need to be prespecified.

Since most studies are crossover with repeated measures from the same subject under different treatment arms, statistical models such as mixed model can be used in analysis. If the model assumptions are not met, then paired t-test or non-parametric approaches may be considered depending on the actual data distribution. HAP is a safety study and statistical significance of the test must be achieved on all doses; no multiplicity adjustment is required.

CONCLUSIONS

Drug abuse potential refers to a medical drug that produces positive psychoactive effects, which could lead to its intentional non-medical use. Human drug abuse study is a vital part of the systematic evaluation of drug abuse potential that includes chemical, animal, human PK, human PD, and post-marketing epidemiological studies. It is also an important tool to evaluate the true effect of a new formulation of opioid pain killers developed to deter non-medical uses. This area is under extensive scientific research as we understand more about the science of these drug products.

ACKNOWLEDGMENTS

I would like to express my gratitude to FDA controlled substance statistical review team members Ling Chen, Wei Liu, and Anna Sun for their contribution to the field of human drug abuse studies through their research and review work.

REFERENCES

1. Abuse-Deterrent Opioids — Evaluation and Labeling Guidance for Industry. 2015. Available at: <http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm334743.pdf> (accessed December 2016).
2. Assessment of Abuse Potential of Drugs. 2017. Available at: <https://www.fda.gov/ucm/groups/fdagov-public/@fdagov-drugs-gen/documents/document/ucm198650.pdf> (accessed March 2017).
3. Chen, L.; Tsong, Y. Design and analysis for drug abuse potential studies: issues and strategies for implementing a crossover design. *Drug Inf. J.* **2007**, *41* (4), 481–489.
4. Chen, L.; Bonson, K.R. An equivalence test for the comparison between a test drug and placebo in human abuse potential studies. *J. Biopharm. Stat.* **2013**, *23* (2), 294–306.
5. Chen, L.; Wang, Y. Heat map displays for data from human abuse potential crossover studies. *Drug Inf. J.* **2012**, *46*, 701–707.

Hypothesis Testing

Devan V. Mehrotra and Ivan S. F. Chan

Merck Research Laboratories, West Point, Pennsylvania, U.S.A.

Abstract

The goal of hypothesis testing is to make a data-based decision concerning the a priori hypothesized value(s), or range of values, of the parameter(s). This entry details several methods of hypothesis testing.

INTRODUCTION

A common objective of any scientific experiment is to make valid inferences about a parameter (or parameters) of a well-defined target population, on the basis of a sample of observed values from the population. For example, consider a randomized clinical trial designed to compare the effectiveness of two antihypertensive drugs, designated test (T) and control (C), in lowering systolic blood pressure after 12 weeks of treatment. For such a trial, the parameter of interest is usually $\theta = \mu_T - \mu_C$, where μ_i denotes the population mean response for drug i . In other words, μ_i is the (hypothetical) mean reduction in systolic blood pressure that would be observed if the entire target population (defined by the inclusion and exclusion criteria for the trial) could be evaluated after 12 weeks of treatment with drug i . In this example, “compare the effectiveness” could mean that the researcher is interested in estimating $\theta = \mu_T - \mu_C$ and/or seeking to demonstrate beyond a reasonable doubt that $\theta > 0$, i.e., the test drug is more effective than the control drug, on average.

In general, two separate but implicitly related approaches are commonly used to make statements about population parameters: estimation and *hypothesis testing*. In the former, an attempt is made to estimate the population parameter(s) of interest with high precision. In the latter, which is the focus here, the purpose is to make a data-based decision concerning the a priori hypothesized value(s), or range of values, of the parameter(s). Hypothesis testing permeates all branches of statistical science, and is one of the most widely used tools in scientific research. For a captivating historical development of this topic, including delightful vignettes about pioneers of hypothesis testing such as Karl Pearson, Ronald Fisher, Jerzy Neyman, and Egon Pearson, see Salsburg.^[1]

NEYMAN–PEARSON FORMULATION OF HYPOTHESIS TESTING

There are four basic elements of a hypothesis test: the null hypothesis (H_0), the alternative hypothesis (H_a), the test

statistic (T), and the rejection region of the test. In practice, H_a is often referred to as the research hypothesis, while H_0 is the “straw man” that the researcher wishes to reject by presenting evidence (sample data) against it. In our clinical trial example, we would commonly test $H_0: \theta = 0$ vs. $H_a: \theta > 0$. The test statistic, a function of the observed data, is used to choose between H_0 and H_a . The rejection region is a (possibly discontinuous) range of values of T that would be rarely observed under H_0 . Accordingly, in the classic Neyman–Pearson^[2] formulation of hypothesis testing, if the observed value of T falls in the rejection region of the test, then H_0 is rejected; otherwise, it is accepted.

Two types of error are possible in this decision process. A type I error is one in which a true H_0 is incorrectly rejected, while a type II error results from failing to reject H_0 when it is false. The probabilities of making type I and type II errors are labeled α and β , respectively, and the maximum value of α is called the size of the rejection region or, more commonly, the test size. In order to determine these error probabilities, we must first determine, either exactly or approximately, what the probability distribution of the test statistic would be if H_0 was true, the so-called null distribution of T . In general, it is important to control the test size at a low level, such as 2.5% or 5%. The power of the test, $1 - \beta$, is the probability of rejecting H_0 when it is false. For a given α , the rejection region of the test should be chosen so as to minimize the probability of a type II error, i.e., to maximize power. If there exists a particular test (defined by the test statistic and the rejection region) with greater power than any other test of the same size, the corresponding test is called a *most powerful* test. It should be noted that the concept of a most powerful test only applies when we are dealing with *simple* null and alternative hypotheses. For instance, in our clinical trial example, this could mean testing $H_0: \theta = 0$ versus, $H_a: \theta = 2$. If the alternative hypothesis is not simple, but *composite*, such as $H_a: \theta > 0$, then the chosen test is called *uniformly most powerful* if it is most powerful over the entire range of the parameter value(s) defined by the alternative hypothesis. Note that for composite hypotheses such as $H_a: \theta > 0$, the power will be

a function of θ , with power increasing as θ becomes more positive. In general, the “farther” H_a is from H_0 , the greater will be the power of the test. The power to reject H_0 can be increased by gathering more evidence (increasing sample size), and by restricting the range of potential values of the parameter(s) of interest under H_a . To clarify the latter point, suppose that the control treatment in our clinical trial example is a placebo, and that three increasing doses of the test treatment ($T_1 < T_2 < T_3$) are being studied. In addition, suppose we can assume a priori that if the test treatment is better than placebo, then a higher dose will be at least as effective as a lower dose, in the T_1 – T_3 range. The null hypothesis would be $H_0: \mu_C = \mu_{T_1} = \mu_{T_2} = \mu_{T_3}$, and a possible alternative is $H_a: \mu_i > \mu_C$ for at least one $i \in \{T_1, T_2, T_3\}$. However, this alternative does not reflect our a priori assumption of a monotonically nondecreasing treatment effect with increasing dose. Our point of clarification is that power could be increased in this situation by using a test statistic that is sensitive to detecting a restricted (and more meaningful) alternative hypothesis, namely $H_a: \mu_C \leq \mu_{T_1} \leq \mu_{T_2} \leq \mu_{T_3}$ with at least one strict inequality.

FISHER’S P -VALUE AND SIGNIFICANCE TESTING

Despite the popularity of the Neyman–Pearson formulation of hypothesis testing, it has been criticized since its inception on various grounds by several leading statisticians, including Ronald Fisher and David Cox. Recall that in the Neyman–Pearson formulation, H_0 is *accepted* if the test statistic does not fall in the rejection region. Fisher^[3–5] argued repeatedly that hypothesis tests are cogent for the *rejection* of hypotheses, but by no means for their *acceptance*. He promoted the use of P -values to quantify the magnitude of evidence *against* (not *for*) the null hypothesis. The P -value is the probability that the value of the test statistic would be at least as extreme as that observed if H_0 were true. For instance, in our clinical trial example, assuming that larger values of the test statistic would provide stronger evidence against H_0 , the P -value would be calculated as $P = P_{H_0}(T \geq T_{\text{obs}})$, where T_{obs} is the observed value of the test statistic for the data at hand.

If the null distribution of T is continuous (e.g., standard normal), then the P -value has a uniform (0,1) distribution under H_0 . Hence if we prespecify a number x , $0 \leq x \leq 1$, and reject H_0 if and only if $P \leq x$, then in the long run we will incorrectly reject H_0 with relative frequency x . Accordingly, in the Neyman–Pearson formulation of hypothesis testing, with α being viewed as the long-run probability of making a type I error, H_0 is rejected if $P \leq \alpha$, and accepted otherwise. Unfortunately, there are some problems with this commonly used decision process. For instance, when $\alpha = 0.05$, $P = 0.0499$ and

$P = 0.0501$ would lead to different decisions, which is rather unappealing. Moreover, $P = 0.0001$ and $P = 0.0499$ would be accorded the same interpretation under the Neyman–Pearson framework.

Cox,^[6] in agreement with Fisher, objected to the rigidity of the Neyman–Pearson decision process. He proposed that the P -value serve as a descriptive tool in a broader “significance testing” analysis, rather than as a decision maker to accept or reject H_0 . He cautioned that the interpretation of the P -value should take into account, among other aspects, the design of the experiment (including sample size), as well as the precision and magnitude of the estimated parameter(s) of interest. In Cox’s view, for example, a P -value of 0.045 in a smaller experiment could be deemed more “significant” than the same P -value in a larger experiment, all else being equal. See Cox^[6] for “hypothesis dividing” and “hypothesis refining” applications of P -values.

HYPOTHESES IN SUPERIORITY, NONINFERIORITY, AND EQUIVALENCE TRIALS

Returning to our clinical trial example, if it was a *superiority* trial, the purpose would be to demonstrate that the test drug is more effective than the control drug, on average. The null and alternative hypotheses would be $H_0: \mu_T - \mu_C \leq 0$ and $H_a: \mu_T - \mu_C > 0$, respectively. If it was a *superiority* trial, the purpose would be to demonstrate that the test drug is more effective than the control drug by at least a clinically significant amount Δ , on average. The null and alternative hypotheses would be $H_0: \mu_T - \mu_C \leq \Delta$ and $H_a: \mu_T - \mu_C > \Delta$, respectively. If it was a *noninferiority* trial, the purpose would be to demonstrate that the test drug is either more effective than the control drug, or less effective by at most Δ , on average. The null and alternative hypotheses would be $H_0: \mu_T - \mu_C \leq -\Delta$ and $H_a: \mu_T - \mu_C > -\Delta$, respectively.

Noninferiority trials are commonly undertaken in situations where the test and control drugs are expected to be approximately equally effective, but the former may be preferable because of reasons such as a better tolerability/safety profile, or a more convenient dosing schedule. The choice of the noninferiority bound Δ is a challenging issue in practice, and must be resolved prior to the start of the trial. In general, the choice of Δ should be based on statistical reasoning as well as clinical and regulatory judgments. More general discussions on the choice of the noninferiority bound can be found in Temple,^[7] Jones et al.,^[8] Ebbutt and Frith,^[9] Röhmle,^[10] and the ICH E10 Guideline.^[11] Both superiority and noninferiority hypotheses are implicitly associated with one-sided tests. Interestingly, however, researchers often report two-sided P -values. This can be confusing in the absence of an accompanying point or interval estimate of the parameter

of interest (here, $\theta = \mu_T - \mu_C$), because the direction of the treatment effect is masked by a two-sided P -value.

Equivalence trials are ubiquitous in pharmacokinetic applications. For example, in a typical average bioequivalence trial, the objective is to show that the test and control drugs (often two formulations of the same drug) have “equivalent” pharmacokinetic profiles, as judged by the area under the plasma-concentration time curve (AUC). The parameter of interest is usually $\theta = \mu_T/\mu_C$, where μ_i denotes the true geometric mean AUC for drug i , and we test $H_0: \theta \leq \theta_l$ or $\theta \geq \theta_u$ vs. $H_a: \theta_l < \theta < \theta_u$. As in a noninferiority trial, the equivalence bounds θ_l and θ_u must be prespecified before the trial. In practice, these limits are dictated by regulatory agencies such as the Food and Drug Administration (FDA), with common values being $\theta_l = 0.80$ and $\theta_u = \theta_l^{-1} = 1.25$. Equivalence trials inherently involve two-sided hypotheses. Interestingly, however, they are commonly analyzed using a two one-sided testing approach (TOST) in which the P -value, to support a conclusion of equivalence, is the larger of the P -values from each of the two one-sided tests (Schuirmann^[12]). The validity of this testing approach has been documented by Berger,^[13] and TOST is currently the recommended approach in regulatory submissions (ICH E9 Guideline^[14]).

BAYESIAN HYPOTHESIS TESTING

In practice, problems arise if the P -value is used to infer something about the veracity or believability of the null. For instance, some researchers incorrectly use the P -value as representing the probability that the null hypothesis is true, and this can lead to interesting paradoxes (Lindley,^[15] Shafer^[16]). Another criticism of the P -value is that it is not very useful when the sample sizes are very large. The reason is that because the null is rarely *exactly* true, increasing the sample size will eventually lead to a small P -value, and hence a rejection of the null hypothesis.

The Bayesian approach to hypothesis testing is a well-established alternative to the classic (frequentist) P -value-based inference, albeit at the “cost” of requiring a prior distribution for the hypothesized parameter(s). In the Bayesian framework, the P -value is replaced with the posterior probability that H_0 is true. If π_0 denotes the prior probability favoring H_0 , then the posterior probability that H_0 is true can be calculated as $[1 + B^{-1}(1 - \pi_0)\pi_0^{-1}]^{-1}$, where B is the “Bayes factor.” Note that B is the ratio of the posterior odds of H_0 to its prior odds. A value of $B = 1/10$ means that H_a is supported 10 times as much by the data as is H_0 . For many years, the acceptability of Bayesian hypothesis testing was limited due (in part) to lack of computational convenience in calculating B . However, recent advancements in computing and development of efficient numerical algorithms have removed that hurdle; see Berger^[17] for details on available Bayesian software. Insightful discussions on

Bayesian hypothesis testing can be found in several articles, including Berger and Delampady,^[18] Casella and Berger,^[19] Kass and Raftery,^[20] and Berger et al.^[21]

CONCLUSION

In summary, we have described the four basic elements of a hypothesis test, discussed type I and type II errors, reviewed the concepts of test size, power, and P -values, and briefly contrasted frequentist with Bayesian hypothesis testing. In any research endeavor, it is important that the postulated null and alternative hypotheses accurately reflect the intended goal of the investigation. Providing a correct statistical answer to the wrong question is often referred to as type III error, and should be avoided at all costs. Readers interested in gaining more in depth knowledge about hypothesis testing are referred to Lehmann.^[22]

REFERENCES

1. Salsburg, D. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*; WH Freeman and Company: New York, 2001.
2. Neyman, J.; Pearson, E. On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika* **1928**, *20A*, 175–240.
3. Fisher, R.A. *Statistical Methods for Research Workers*; Oliver and Boyd: Edinburgh, 1925.
4. Fisher, R.A. Statistical tests. *Nature* **1935**, *136*, 474–475.
5. Fisher, R.A. *Statistical Methods and Scientific Inference*; Oliver and Boyd: Edinburgh, 1959.
6. Cox, D.R. The role of significance tests. *Scand. J. Statist* **1977**, *4*, 49–70.
7. Temple, R. Problems in interpreting active control equivalence trials. *Accountability in Research* **1996**, *4*, 267–275.
8. Jones, B.; Jarvis, P.; Lewis, J.A.; Ebbutt, A.F. Trials to assess equivalence: The importance of rigorous methods. *Br. Med. J.* **1996**, *313*, 36–39.
9. Ebbutt, A.F.; Frith, L. Practical issues in equivalence trials. *Stat. Med.* **1998**, *17*, 1691–1701.
10. Röhm, J. Therapeutic equivalence investigations: Statistical considerations. *Stat. Med.* **1998**, *17*, 1703–1714.
11. ICH E10 Guideline. Choice of Control Group and Related Issues in Clinical Trials. International Conference on Harmonization (ICH), July 2000.
12. Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm* **1987**, *15*, 657–680.
13. Berger, R.L. Multiparameter hypothesis testing and acceptance sampling. *Technometrics* **1982**, *24*, 295–300.
14. ICH E9 Guideline. Statistical Principles for Clinical Trials. International Conference on Harmonization (ICH), February 1998.
15. Lindley, D.V. A statistical paradox. *Biometrika* **1957**, *44*, 187–192.

16. Shafer, G. Lindley's paradox (with discussion). *J. Am. Stat. Assoc.* **1982**, 77, 35–351.
17. Berger, J. Bayesian analysis: A look at today and thoughts of tomorrow. *J. Am. Stat. Assoc.* **2000**, 95, 1269–1276.
18. Berger, J.; Delampady, M. Testing precise hypotheses. *Stat. Sci.* **1987**, 2, 317–352.
19. Casella, G.; Berger, R.L. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Am. Stat. Assoc.* **1987**, 82, 106–111.
20. Kass, R.E.; Raftery, A.E. Bayes factors. *J. Am. Stat. Assoc.* **1995**, 90, 773–795.
21. Berger, J.; Boukai, B.; Wang, W. Unified frequentist and Bayesian testing of precise hypotheses. *Stat. Sci.* **1997**, 12, 133–160.
22. Lehmann, E.L. *Testing Statistical Hypothesis*, 2nd Ed.; Wiley: New York, 1986.

Imputation with Item Nonrespondents

Hansheng Wang

Peking University, Beijing, People's Republic of China

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

Missing values in studies are commonly encountered. An item nonrespondent refers to a subject who fails to answer some (not all) of the items. This entry is focused on the approaches clinical researchers can take in dealing with item nonrespondents, looking closely at statistical properties.

INTRODUCTION

In clinical research, an instrument (or questionnaire) consisting of a number of items (or questions) is usually used to quantitate a subject's behavior and the ability to function in day-to-day activities as perceived by the subject. For example, the EORTC QLQ-C30 questionnaire is commonly used as a reliable instrument to measure the quality of life of patients with cancer. On the other hand, the positive and negative syndrome scale (PANSS) is useful for assessing positive symptoms (e.g., delusions, hallucinatory behavior, and hostility), negative symptoms (e.g., emotional/passive social withdrawal and poverty of thought and spontaneous activity), and general psychopathology (e.g., anxiety and depression) in schizophrenic patients and patients with schizoaffective disorder. For each item, the patient (if self-administered) or the investigator is asked to provide a rating score from 0 (none) to k (severe). A self-administered instrument is often used because it can reflect how a patient feels and may have less bias because patients feel that their answers are more confidential. However, a self-administered instrument is expected to have a larger variability due to the unstable condition of the patient. Based on patient ratings, several statistical methods for assessment of drug effects have been proposed (see, e.g., Refs. [1–3]).

OVERVIEW

In practice, missing values are commonly encountered, which can be classified into two categories, namely, unit nonrespondent and item nonrespondent (see, e.g., Wang^[4,5]). A unit nonrespondent refers to a subject who fails to answer all of the items (questions). An item nonrespondent refers to a subject who fails to answer some

(not all) of the items. In practice, because unit, nonrespondents do not provide any information regarding the treatment effect, they are usually excluded from the analysis under the assumption of missing completely at random. On the other hand, because item nonrespondents do provide partial information, it is not acceptable to exclude these subjects from analysis.

In clinical research, two approaches are commonly suggested for handling item nonrespondents. The first approach is to simply ignore all the item nonrespondents from the analysis. Although this method is statistically valid under the assumption of “missing completely at random,” it suffers from decreasing power/efficiency and from excluding too many evaluable patients from the analysis. The second approach is to impute the missing item and then calculate the total score based on the imputed data. As indicated in Shao and Wang,^[6] imputation methods in clinical research commonly include last observation carry forward (LOCF) or endpoint analysis, mean/median imputation, regression with covariates, etc. To examine the statistical properties of imputation with item nonrespondents, for illustration purpose, we will focus only on mean imputation method.

In the next section, the mean imputation approach will be described under a commonly employed statistical model. Also included in this section are the statistical properties of the estimators derived from the model based on the imputed data, in the section “Statistical inference,” two procedures including the so-called linearization procedure and the jackknife method proposed by Rao and Shao^[7] for estimation of the asymptotic variance of the derived estimators are discussed. Also included are statistical tests based on the mean imputation approach. The section “An Example” gives an example concerning the study of agitated behavior in schizophrenic patients to illustrate the use of the mean imputation method. A brief concluding remark is given in the last section.

STATISTICAL MODEL

Single-Imputation Class

In clinical research, the treatment groups are usually considered as imputation classes. The imputation will be carried out within each imputation class. For simplicity, we first consider the case of a single-imputation class. The results can be generalized to multiple-imputation classes as discussed at the end of this section.

Let $x_{ij} \in S_j$, $i = 1, \dots, n$; $j = 1, \dots, m$ be the score of the j th item from the i th subject. S_j is the finite sample space for x_{ij} . Note that S_j need not to be the same for different j . The observations from the same subject (e.g., $x_{i1}, \dots, x_{im}$) are usually correlated. However, the vectors $(x_{i1}, \dots, x_{im})'_{i=1, \dots, n}$ are assumed to be independent and identically distributed (i.i.d.) for different i . The total score of the i th subject is defined to be

$$z_i = \sum_{j=1}^m x_{ij}$$

The parameter of interest is $\mu = E(z_i) = \sum_j \mu_j$, where $\mu_j = E(x_{ij})$. Because there are missing scores, denote y_{ij} by missing scores of x_{ij} . That is, $y_{ij} = 0$ means missing and $y_{ij} = 1$ indicates not missing. As a result, $(y_{i1}, \dots, y_{im})'_{i=1, \dots, n}$ are usually correlated for a fixed i but are i.i.d. random vectors for different i . It is assumed that $E(y_{ij}) = p_j$.

For any given j , the observed sample mean for the j th item is given by

$$\bar{x}_{-j}^* = \frac{\sum_{i=1}^n x_{ij} y_{ij}}{\sum_{i=1}^n y_{ij}}$$

As a result, if x_{ij} is missing (i.e., $y_{ij} = 0$), it is replaced by x_{-j}^* . Let x_{ij}^* be the value of the j th item of the i th subject after imputation. It can be verified that

$$x_{ij}^* = y_{ij} x_{ij} + (1 - y_{ij}) x_{-j}^*$$

Hence, the total score of the i th subject after imputation is given by

$$z_i^* = \sum_{j=1}^m x_{ij}^* = \sum_{j=1}^m (y_{ij} x_{ij} + (1 - y_{ij}) x_{-j}^*)$$

It is common practice to treat the imputed values as the observed values and to obtain estimators designed based on complete data sets for estimation of the parameter of interest. In the situation where there are no missing values, the sample mean of z_i is a natural estimator for μ . Consequently,

the sample mean of z_i^* becomes our estimator for μ , which is given by

$$\begin{aligned} \hat{\mu}^I &= \frac{1}{N} \sum_{i=1}^n z_i^* = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m (y_{ij} x_{ij} + (1 - y_{ij}) x_{-j}^*) \\ &= \sum_{j=1}^m \left(\frac{\frac{1}{n} \sum_{i=1}^n x_{ij} y_{ij}}{\frac{1}{n} \sum_{i=1}^n y_{ij}} \right) \end{aligned}$$

The following theorem can be obtained.

Theorem 1. Assuming single-imputation class, based on the proposed mean imputation procedure,

$$\sqrt{n}(\hat{\mu}^I - \mu) \rightarrow_d N(0, \sigma^2)$$

where “ $\rightarrow_d$ ” denotes “converge in distribution” and σ^2 is given by

$$\sigma^2 = \text{var} \left(\sum_{j=1}^m \frac{1}{p_j} (x_{ij} y_{ij} - \mu_j y_{ij}) \right)$$

Proof:

By Taylor's expansion and Slutsky's theorem, it follows that

$$\begin{aligned} \sqrt{n}(\hat{\mu}^I - \mu) &= \sum_{j=1}^m \sqrt{n} \left(\frac{\frac{1}{n} \sum_{i=1}^n x_{ij} y_{ij}}{\frac{1}{n} \sum_{i=1}^n y_{ij}} - \mu_j \right) \\ &= \sum_{j=1}^m \sqrt{n} \left(\frac{1}{p_j n} \sum_{i=1}^n [(x_{ij} y_{ij} - \mu_j p_j) - \mu_j (y_{ij} - p_j)] \right) \\ &\quad + o_p(1) = \frac{1}{\sqrt{n}} \sum_{j=1}^m \left(\sum_{i=1}^n \frac{1}{p_j} (x_{ij} y_{ij} - \mu_j y_{ij}) \right) \\ &\quad + o_p(1) \rightarrow_d N(0, \sigma^2) \end{aligned}$$

where σ^2 is defined in Theorem 1.

Multiple-Imputation Classes

Assuming there are H imputation classes, let $x_{h,ij}$ be the score of the j th item from the i th subject in the h th imputation class. The quantities $\mu_h, y_{h,ij}, \mu_{h,j}, z_{h,i}, p_{h,j}$ are similarly defined.

For multiple-imputation classes, the parameter of interest is usually a linear combination of μ_h , where $\mu_h = E(x_{h,ij})$. Let w_h denote the coefficient. The parameter of interest is given by

$$\mu = \sum_{h=1}^H w_h \mu_h$$

In clinical research study, μ usually denotes the treatment effect of interest and w_h s are usually the coefficients of contrast satisfying $\sum_h w_h = 0$. For example, in a placebo-controlled clinical trial, the number of imputation classes equals the number of treatment groups, which is two [treatment ($h = 1$) and placebo ($h = 2$)]. w_h s are then chosen to be $w_1 = 1$ and $w_2 = -1$. The parameter of interest is $\mu = \mu_1 - \mu_2$, which is the treatment effect.

For multiple-imputation classes, mean imputation can be carried out within each imputation class. The sample mean of the total score after imputation ($\hat{\mu}_h^t$) is used as an estimator for μ_h . The estimator of μ is given by

$$\hat{\mu}^t = \sum_h w_h \hat{\mu}_h^t \quad (1)$$

Similarly, the following results can be obtained.

Theorem 2. Assuming multiple-imputation classes and $n_h/n \rightarrow p_h$ for some $p_h > 0$, based on the mean imputation procedure described above, we have

$$\sqrt{n}(\hat{\mu}^t - \mu) \rightarrow_d N(0, \sigma^2)$$

Where σ^2 is given by

$$\sigma^2 = \sum_{h=1}^H \frac{w_h^2}{p_h} \text{var} \left(\sum_{j=1}^m \frac{1}{p_{h,j}} (x_{h,ij} y_{h,ij} - \mu_{h,j} y_{h,ij}) \right)$$

Proof:

Similar to the proof of Theorem 1, we have

$$\begin{aligned} \sqrt{n}(\hat{\mu}^t - \mu) &= \sqrt{n} \sum_h w_h (\hat{\mu}_h^t - \mu_h) \\ &= \sum_h \frac{w_h}{\sqrt{n_h} h} \sqrt{n_h} (\hat{\mu}_h^t - \mu_h) \\ &= \sum_h \frac{w_h}{\sqrt{p_h}} \sqrt{n_h} (\hat{\mu}_h^t - \mu_h) + o_p(1) \\ &\rightarrow N \left(0, \sum_h \frac{w_h^2}{p_h} \sigma_h^2 \right) \end{aligned}$$

where σ_h^2 is the asymptotic variance of $\sqrt{n_h}(\hat{\mu}_h^t - \mu_h)$, which can be estimated according to Theorem 1.

STATISTICAL INFERENCE

Variance Estimates

Although the sample variance of $\sum_{j=1}^m (1/p_{h,j})(x_{h,ij} y_{h,ij} - \mu_{h,j} y_{h,ij})$ can be used to estimate its true variance, $p_{h,j}$ and $\mu_{h,j}$ are unknown. One way to overcome this problem is to substitute $p_{h,j}$ and $\mu_{h,j}$ with their consistent estimators (e.g., $x_h^*, \dots, j$ and $1/n \sum_i y_{h,ij}$, respectively). Since this procedure is

based on Taylor's expansion or the linearization of $\hat{\mu}^t$, we denote this estimator by $\hat{\sigma}_{\text{linear}}^2$.

The above procedure is relatively complicated although it requires less computation. As an alternative, the jackknife procedure proposed by Rao and Shao^[7] can be used to estimate the asymptotic variance. Referring to the proofs of Theorem 1 and Theorem 2 in the Appendix, we note that $\hat{\mu}^t$ is the smooth function of $1/n \sum x_{h,ij} y_{h,ij}$ and $1/n \sum y_{h,ij}$. Both of them are sample means of i.i.d. random variables. As a result, Rao and Shao's procedure can be applied here to estimate the asymptotic variance. More specifically, for any given h and i , let $\hat{\mu}_{h,-i}^t$ be an estimate of μ . It can be obtained by first constructing a reduced data set by dropping the i th subject in the h th stratum. Then, perform the proposed mean imputation procedure to the reduced data set. Finally, $\hat{\mu}_{h,-i}^t$ is obtained by (1) but based on the reduced data set. Then, the jackknife estimate of σ^2 is given by

$$\hat{\sigma}_{\text{jack}}^2 = \sum_{h=1}^H \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} (\hat{\mu}_{h,-i}^t - \hat{\mu}_h^t)^2$$

Testing Hypothesis

In clinical research, the statistical hypotheses of interest can usually be written as

$$H_0: \mu = \mu_0 \quad \text{and} \quad H_1: \mu \neq \mu_0$$

where μ_0 is a prespecified value. One simple example is $\mu_0 = 0$ for a two-group parallel design comparing treatment effects (recall $H = 2$, $w_1 = 1$, and $w_2 = -1$). The test statistic is given by

$$Z = \frac{\hat{\mu}^t - \mu_0}{\hat{\sigma}}$$

where $\hat{\sigma}$ is an estimate for σ . It could be $\hat{\sigma}_{\text{jack}}$ or $\hat{\sigma}_{\text{linear}}$. Refer to the proofs of Theorem 1 and Theorem 2 given in the Appendix. It can be noted that $\hat{\mu}^t$ is a smooth function of $1/n \sum x_{h,ij} y_{h,ij}$ and $1/n \sum y_{h,ij}$, which are sample means of i.i.d. random variables. As a result, according to central limit theorem and Slutsky's theorem, Z is asymptotically normally distributed with mean 0 and variance 1 under the null hypothesis. Therefore, for a given significance level α , the null hypothesis can be rejected if $|Z| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $(1 - \alpha/2) \times 100\%$ percentile of a standard normal distribution.

A Simulation Study

A simulation study was conducted to evaluate the finite sample performance of the proposed method. The objective of the simulation study is multifold. It includes the following.

1. Note that the estimators derived by both imputation and complete units are consistent for μ . As a result, it is of interest to compare their relative efficiency in terms of standard deviation.
2. Two methods (linearization and jackknife) are proposed for estimation of the standard deviation of $\hat{\mu}'$. Although they are asymptotically equivalent to each other, it is of interest to compare their finite sample performance by comparing them with the true standard deviation.
3. The proposed Z-test for the hypothesis of interest is based on large sample theory. As a result, the empirical significance level of the test may not be exactly the same as the nominal level. Therefore, it is of interest to evaluate the finite sample performance of the proposed Z-test based on both $\hat{\sigma}_{\text{linear}}$ and $\hat{\sigma}_{\text{jack}}$.

Consider a two-arm parallel randomized clinical trial comparing a test treatment with a placebo. There are two imputation classes ($H = 2$), $w_1 = 1$, $w_h = -1$, and the parameter of interest is the treatment effect $\mu = w_1\mu_1 + w_2\mu_2 = \mu_1 - \mu_2$.

For illustration purposes and for the sake of convenience, we fix $m = 5$ and $n = 20$, which means there are a total of 20 items (questions), and for each question the rating score ranges from 1 to 5. We also assume the score of each item is marginally distributed with probabilities of 0.3, 0.1, 0.3, 0.1, and 0.2. Note that it is not necessary for each item to have the same sample space and the same distribution.

Let n_1 and n_2 denote the sample sizes for group 1 and group 2, respectively. They range from 20 to 40. For a given subject, we first generate a continuous standard uniform random number U_0 . For any $j \leq n$, we generate another standard uniform random number R_j independent of U_0 . Then, define $U_j = U_0$ if $R_j \leq 0.5$. Otherwise, let U_j be another standard uniform random number independent of both U_0 and R_j . Compare U_j with the cumulative distribution function of the score of the item to determine the value of the j th item. More specifically,

$$x_{h,j} = \begin{cases} 1 & \text{if } U_j \leq 0.3 \\ 2 & \text{if } 0.3 < U_j \leq 0.4 \\ 3 & \text{if } 0.4 < U_j \leq 0.7 \\ 4 & \text{if } 0.7 < U_j \leq 0.8 \\ 5 & \text{if } 0.8 < U_j \end{cases}$$

where “ $\Leftarrow$ ” means “defines to be.” There are two important properties of $x_{h,j}$ generated by this procedure. First, U_j is still a marginally standard uniform random variable. As a result, $x_{h,j}$ is distributed with probabilities of 0.3, 0.1, 0.3, 0.1, and 0.2. Second, all the U_j from the same subject have probability 0.5 to share the same U_0 . Consequently, they are correlated. The purpose of this is to reflect the

fact that the observations from the same subject should be correlated in some way.

In order to determine if the score of an item is missing or not, we generate a standard uniform random variable U_0 for each subject. Generate another standard uniform random variable R_j for each $j \leq n$. Then, define $U_j = U_0$ if $R_j \leq 0.5$. Otherwise, let U_j be another standard uniform random number independent of both U_0 and R_j . Then, the score of the j th item is set to be missing if $U_j > p_h$ otherwise it is observed, where p_h is a prespecified value. As a result, for a given item, it has a marginal probability p_h to be observed. In the simulation, we consider p_h to be 0.80, 0.85, 0.90, 0.95, and 1.00. As a result, the missing of all items from the same subject are correlated with each other but independent of the value of the item.

For a given set of parameters (n_1 , n_2 , p_h), 10,000 simulation runs were carried out. For each simulation, a data set mimics a two-arm parallel trial, and is generated according to the specified parameters. For the generated data set, μ is estimated by both imputation method ($\hat{\mu}'$) and complete units only ($\hat{\mu}$). Their sample means are reported. Their sample standard deviations are used as estimates for their true standard deviation. For the same data set, both jackknife and linearization methods are used to estimate the standard deviation of $\hat{\mu}'$. Their sample means are used to estimate the true values of those two estimates. For the same data set, the Z-tests are performed by using both $\hat{\sigma}_{\text{jack}}$ and $\hat{\sigma}_{\text{linear}}$. The empirical p value is estimated by the proportion of the data sets rejecting the null hypothesis. Table 1 summarizes the simulation results.

Table 1 findings are summarized below.

1. The efficiency (in terms of standard deviation) of the point estimate based on imputed data is improved as compared with the method of using the completers only.
2. Both jackknife and linearization provide reasonably good estimates for the standard deviation of $\hat{\mu}'$. The jackknife estimator seems to perform slightly better than the estimate obtained by linearization.
3. The empirical level of both Z-tests are reasonably close to the nominal level 5%. However, the Z-test using $\hat{\sigma}_{\text{jack}}$ is slightly better than the Z-test using $\hat{\sigma}_{\text{linear}}$.

AN EXAMPLE

To illustrate the application of the proposed mean imputation with item nonrespondents, consider an example concerning the study of treatment effect on controlling agitated behavior of a drug product in schizophrenic patients and patients with schizoaffective disorders. A randomized, parallel-group clinical trial comparing two treatments was conducted in patients with mild to moderate schizophrenia or schizoaffective disorder to help control their agitated behavior. Twenty patients were randomized and completed

Table 1 Simulation Results ($m = 5$ and $n = 20$)

n_1	n_2	p_h	Complete units based				Imputation based			
			$\hat{\mu}^I$	σ	$\hat{\mu}$	σ	Jackknife		Linearization	
							$\hat{\sigma}_{\text{jack}}$	p	$\hat{\sigma}_{\text{linear}}$	p
20	20	0.8000	0.7619	14.9606	0.0284	5.1977	5.1614	0.0597	5.1391	0.0609
		0.8500	0.1983	12.9868	0.0709	5.0866	5.1062	0.0557	5.0931	0.0559
		0.9000	-0.0739	9.3859	-0.0263	5.0755	5.0546	0.0546	5.0485	0.0552
		0.9500	0.1001	6.7447	0.0668	5.0279	5.0095	0.0587	5.0062	0.0587
		1.0000	0.0887	5.0813	0.0887	5.0813	4.9653	0.0635	4.9653	0.0635
20	30	0.8000	0.5605	14.6527	-0.0474	4.7282	4.7090	0.0567	5.1506	0.0367
		0.8500	0.1758	11.9008	0.0561	4.6332	4.6568	0.0564	5.1000	0.0371
		0.9000	0.0922	8.3996	0.0432	4.6271	4.6099	0.0586	5.0513	0.0373
		0.9500	0.0027	6.0743	0.0302	4.5251	4.5757	0.0533	5.0140	0.0343
		1.0000	-0.0289	4.5519	-0.0289	4.5519	4.5375	0.0569	4.9724	0.0360
20	40	0.8000	0.3131	14.0286	0.0088	4.5208	4.4748	0.0600	5.1697	0.0305
		0.8500	0.1110	11.1785	-0.0310	4.4393	4.4214	0.0571	5.1095	0.0300
		0.9000	0.0685	8.1120	0.0021	4.4150	4.3736	0.0573	5.0576	0.0310
		0.9500	-0.0792	5.7790	-0.0777	4.3477	4.3306	0.0606	5.0081	0.0306
		1.0000	0.0778	4.3305	0.0778	4.3305	4.3007	0.0598	4.9721	0.0291
30	30	0.8000	0.2106	13.9243	-0.0540	4.2101	4.2314	0.0502	4.2209	0.0508
		0.8500	0.1861	10.6147	-0.0280	4.1717	4.1854	0.0526	4.1797	0.0527
		0.9000	-0.0367	7.4941	0.0073	4.1271	4.1335	0.0525	4.1309	0.0530
		0.9500	0.0389	5.4454	-0.0039	4.1258	4.0919	0.0553	4.0911	0.0553
		1.0000	0.0283	4.0690	0.0284	4.0690	4.0602	0.0553	4.0602	0.0553
30	40	0.8000	0.4015	13.3119	-0.0112	3.9533	3.9597	0.0536	4.2282	0.0396
		0.8500	-0.0337	9.9220	-0.0091	3.9083	3.9156	0.0548	4.1834	0.0408
		0.9000	-0.0291	7.0364	-0.0068	3.8735	3.8665	0.0541	4.1342	0.0408
		0.9500	0.0070	5.1251	-0.0164	3.8366	3.8333	0.0549	4.0983	0.0391
		1.0000	0.0486	3.7901	0.0486	3.7901	3.7975	0.0548	4.0601	0.0399
40	40	0.8000	0.0155	12.5925	-0.0331	3.6945	3.6690	0.0577	3.6627	0.0586
		0.8500	0.0626	9.0309	-0.0264	3.6302	3.6248	0.0525	3.6216	0.0528
		0.9000	0.0285	6.3868	0.0412	3.5928	3.5842	0.0534	3.5823	0.0535
		0.9500	0.0573	4.6978	0.0412	3.5562	3.5499	0.0556	3.5493	0.0552
		1.0000	0.0413	3.5004	0.0414	3.5004	3.5129	0.0535	3.5129	0.0535

the study of 12 weeks. An instrument consisting of 20 items was used to quantitate the patient's agitated behavior. The scale of each item ranges from 1 (none) to 5 (severe). The results at endpoint are given in Table 2.

Although, the trial recruited 20 patients for each treatment group, there were only eight unit respondents per group. The analysis based on complete units estimates the treatment effect $\hat{\mu} = -3.875$ with standard deviation $\hat{\sigma} = 5.662$. The corresponding Z statistic is given by $Z = 3.875/5.662 = -0.684$. The p value is given by 0.494, which is not significant (< 0.05).

On the other hand, the analysis based on the proposed mean imputation approach estimates the treatment effect $\hat{\mu}' = -10.767$ with standard deviation $\hat{\sigma}_{\text{jack}} = 4.862$ and $\hat{\sigma}_{\text{linear}} = 4.859$. The corresponding Z statistics are given by 2.214 and 2.216, respectively, with p values given by 0.027 and 0.027, respectively, which are significant.

In this example, the analysis based on unit respondents only may lose too much power/efficiency and consequently may be misleading. The mean imputation approach, on the other hand, is more efficient than the analysis using complete units.

Table 2 Data Listing

No	TRT	Complete	Itemized scores (1–20)																			
1	1	1	3	1	3	1	2	1	1	1	2	1	1	5	1	1	5	1	1	1	1	5
2	1	0	4	1	3	1	1	3	1	1	1	4	1	1	*	1	1	1	1	1	3	3
3	1	0	5	5	1	1	1	1	1	1	1	1	2	5	4	*	5	1	1	1	2	1
4	1	1	3	3	3	3	2	3	3	3	3	4	2	3	3	2	3	3	3	4	3	3
5	1	0	*	1	1	2	1	3	1	1	1	5	3	3	*	1	1	1	1	5	1	1
6	1	0	*	3	3	*	*	3	3	*	3	*	*	*	3	*	3	3	*	*	*	*
7	1	1	1	1	1	3	1	3	3	1	2	1	1	1	5	1	1	1	3	1	1	1
8	1	1	3	3	5	3	3	1	3	1	3	3	4	1	3	3	3	1	5	3	3	3
9	1	0	1	3	2	2	2	*	3	2	2	2	2	2	*	2	2	*	2	3	2	2
10	1	0	2	3	3	2	2	2	2	1	3	2	2	2	3	2	*	2	1	3	2	2
11	1	0	5	1	2	2	*	2	4	2	2	5	3	3	*	2	2	3	2	5	5	*
12	1	1	1	4	2	5	3	2	1	1	1	1	3	1	3	1	1	1	3	5	3	5
13	1	1	1	1	5	1	1	3	1	1	1	1	1	1	3	3	1	3	3	1	1	1
14	1	1	3	2	3	3	3	5	3	3	3	1	3	2	3	3	3	3	5	3	3	3
15	1	1	4	4	4	2	4	4	4	5	4	1	1	1	1	3	4	4	3	1	4	3
16	1	0	1	4	1	1	2	1	1	1	1	*	1	4	3	3	4	1	2	1	1	5
17	1	0	5	2	1	5	5	3	5	5	5	5	4	1	*	5	5	5	5	5	5	5
18	1	0	3	5	3	3	5	3	3	3	3	*	3	3	*	1	3	1	3	4	*	3
19	1	0	5	3	1	1	1	3	1	5	1	1	3	4	1	1	5	2	*	1	1	1
20	1	0	1	1	1	1	1	2	3	3	1	4	1	*	1	1	1	3	1	3	1	1
21	2	1	3	5	3	3	3	1	1	3	3	3	3	3	3	3	3	3	1	4	3	3
22	2	0	3	2	2	2	1	1	5	1	*	*	1	5	1	2	1	3	1	*	1	3
23	2	0	1	3	5	*	4	3	3	*	3	3	3	3	5	5	3	3	4	4	3	3
24	2	1	3	4	3	3	4	2	3	1	4	4	2	3	3	3	3	5	1	3	2	4
25	2	1	2	4	2	2	2	4	2	2	3	2	4	3	1	1	2	2	2	2	5	2
26	2	1	4	2	4	5	5	2	2	2	4	4	1	4	1	4	4	4	4	3	4	4
27	2	0	4	2	2	2	*	2	2	2	4	2	2	3	1	5	1	2	1	5	3	5
28	2	0	4	4	2	*	4	*	3	4	4	5	4	4	4	2	4	4	4	3	2	4
29	2	1	3	3	3	3	3	3	3	3	1	3	3	3	2	3	2	3	3	5	1	1
30	2	0	2	1	4	1	3	3	3	4	3	*	3	*	2	1	3	3	3	3	4	3
31	2	0	5	5	4	*	1	5	4	5	3	3	1	5	5	5	*	5	5	5	5	5
32	2	1	5	2	2	2	3	2	4	2	2	2	2	2	2	2	2	3	4	2	4	4
33	2	1	1	1	1	1	1	1	1	2	3	5	1	5	2	5	4	3	2	1	1	3
34	2	0	1	3	2	1	1	5	1	*	1	1	1	1	1	3	3	4	1	*	1	3
35	2	0	4	5	3	4	4	1	4	4	1	4	*	2	4	4	1	2	4	4	4	4
36	2	0	4	5	5	5	5	5	5	5	5	5	5	5	5	5	1	5	5	5	5	*
37	2	0	*	1	1	5	1	1	1	1	2	1	1	*	3	1	2	1	2	5	1	3
38	2	0	5	5	5	3	3	5	5	5	5	5	2	5	*	2	5	5	5	5	4	5
39	2	1	1	4	2	2	1	1	3	1	1	1	1	1	1	1	3	2	1	1	5	1
40	2	0	3	3	*	3	3	3	3	1	3	3	3	3	5	4	3	3	3	5	3	3

*indicates missing value.

No = subject number.

TRT = treatment group.

Complete denotes whether the unit is complete (1, complete and 0, not complete).

CONCLUSION

In summary, we propose a mean imputation approach to handle item nonrespondents when the total score is the parameter of interest. The derived estimator is consistent and asymptotically normal. The jackknife method is useful for estimation of the standard deviation of the proposed estimator. The proposed method is not only statistically valid but also more efficient as compared with the method based on complete units only.

REFERENCES

1. Tandon, P.K. Application of global statistics in analysing quality of life. *Stat. Med.* **1990**, *9*, 819–827.
2. Olschewski, M.; Schumacher, M. Statistical analysis of quality of life data in cancer clinical trials. *Stat. Med.* **1990**, *9*, 749–763.
3. Chow, S.C.; Ki, F. On statistical characteristics of quality of life assessment. *J. Biopharm. Stat.* **1994**, *4*, 1–17.
4. Wang, H. Two-Way Contingency Tables with Conditionally and Marginally Imputed Data; Ph.D. Thesis; University of Wisconsin-Madison, 2001.
5. Wang, H. Imputation in Clinical Research. In *Encyclopedia of Biopharmaceutical Statistics*; Chow, S.C.; Ed.; Marcel Dekker, Inc.: New York, 2003.
6. Shao, J.; Wang, H. Sample correlation coefficients based on survey data under regression imputation. *J. Am. Stat. Assoc.* **2002**, *97*, 544–552.
7. Rao, J.N.K.; Shao, J. Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **1992**, *79*, 811–822.

Imputation: Clinical Research

Hansheng Wang

Peking University, Beijing, People's Republic of China

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

This entry summarizes the most commonly used imputation methods and investigates their statistical properties, and also discusses recent developments.

INTRODUCTION

Missing values or incomplete data are commonly encountered in clinical research and are studied by many authors.^[1–3] Basically, the causes of missing values in a study can be classified into two categories. The first category includes the reasons that are not directly related to the study. For example, a patient may be lost to follow-up because he/she moves out of the area. This category of missing values can be considered as *missing completely at random*. The second category includes the reasons that are related to the study. For example, a patient may withdraw from the study due to treatment-emergent adverse events. In practice, it is not uncommon to have multiple assessments from each subject. Subjects with all observations missing are called unit nonrespondents. Because unit nonrespondents do not provide any useful information, these subjects are usually excluded from the analysis. On the other hand, the subjects with some, but not all, observations missing are referred to as item nonrespondents. In practice, excluding item nonrespondents from the analysis is considered against the *intent-to-treat* (ITT) principle and, hence, not acceptable. In clinical research, the primary analysis is usually conducted based on ITT population, which includes all randomized subjects with at least posttreatment evaluation. As a result, most item nonrespondents may be included in the ITT population. In practice, excluding item nonrespondents may seriously decrease power/efficiency of the study.

To account for item nonrespondents, two methods are commonly considered. The first method is the so-called likelihood-based method. Under a parametric model, the marginal likelihood function for the observed responses is obtained by integrating out the missing responses. The parameter of interest can then be estimated by the maximum likelihood estimator (MLE). Consequently, a corresponding test (e.g., likelihood ratio test) can be constructed. The merit of this method is that the resulting statistical procedures are usually efficient. The drawback

is that the calculation of the marginal likelihood could be difficult. As a result, some special statistical or numerical algorithms are commonly applied for obtaining the MLE. For example, the expectation-maximization (EM) algorithm is one of the most popular methods for obtaining the MLE when there are missing data. The other method for item nonrespondents is imputation. Compared with the likelihood-based method, the method of imputation is relatively simple and easy to apply. The idea of imputation is to treat the imputed values as the observed values and then apply the standard statistical software for obtaining consistent estimators. However, it should be noted that the variability of the estimator obtained by imputation is usually different from the estimator obtained from the complete data. In this case, the formulas designed to estimate the variance of the complete data set cannot be used to estimate the variance of estimator produced by the imputed data. As an alternative, two methods are considered for estimation of its variability. One is based on Taylor's expansion. This method is referred to as the "linearization method." The merit of the linearization method is that it requires less computation. However, the drawback is that its formula could be very complicated and/or nontrackable. The other approach is based on resampling method (e.g., bootstrap and jackknife). The drawback of the resampling method is that it requires an intensive computation. The merit is that it is very easy to apply. With the help of a fast-speed computer, the resampling method has become much more attractive in practice.

Note that imputation is not only popular in clinical research, it is also very popular in many other statistical fields such as sample survey. However, the imputation methods in clinical research are more diversified due to the complexity of the study design relative to sample survey. As a result, the statistical properties of many commonly used imputation methods in clinical research are still unknown, while most imputation methods used in sample survey are well studied. Hence, the imputation methods in clinical research provide a unique challenge

and also an opportunity for the statisticians in the area of clinical research. In what follows, we will summarize the most commonly used imputation methods and investigate their statistical properties. Recent development will also be discussed.

LAST OBSERVATION CARRY FORWARD/ENDPOINT ANALYSIS

Last observation carry forward (LOCF) and endpoint analysis (EPA) are probably the most commonly used imputation methods in clinical research. For illustration purposes, one example is described below.

Consider a randomized, parallel-group clinical trial comparing r treatments. Each patient is randomly assigned to one of the treatments. According to the protocol, each patient should undergo s consecutive visits. Let y_{ijk} be the observation from the k th subject in the i th treatment group at visit j . The following statistical model is usually considered.

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \text{ where } \epsilon_{ijk} \sim N(0, \sigma^2) \quad (1)$$

where μ_{ij} represents the fixed effect of the i^{th} treatment at visit j . If there are no missing values, the primary comparison between treatments will be based on the observations from the last visit ($j = s$) because this reflects the treatment difference at the end of the treatment period. However, it is not necessary that every subject completes the study. Suppose that the last evaluable visit $j^* < s$ for the k th subject in the i th treatment group. Then the value of y_{ij^*k} can be used to impute t_{isk} . After imputation, the data at endpoint are analyzed by the usual analysis of variance (ANOVA) model. We will refer to the procedure described above as LOCF.

Note that the method of LOCF is usually applied according to the ITT principle. The ITT population include all randomized subjects. In clinical research, although the LOCF is commonly employed, the statistical motivation and the consequence are not clear. In what follows, we attempt to provide two different approaches to understand the statistical properties of LOCF.

Bias-Variance Trade-Off

The objective of a clinical study is usually to assess the safety and efficacy of a test treatment under investigation. Statistical inferences on the efficacy parameters are usually obtained. In practice, a sufficiently large number of sample size is required to obtain a reliable estimate and to achieve a desired power for establishment of the efficacy of the treatment. The reliability of an estimator can be evaluated by bias and by variability. A reliable estimator should have a small or zero bias with small variability. Hence, the estimator based on LOCF and the estimator

based on completers are compared in terms of their bias and variability.

For illustration purposes, we focus on only one treatment group with two visits. Assume that there are a total of $n = n_1 + n_2$ randomized subjects, where n_1 subjects complete the trial, while the remaining n_2 subjects only have observations at visit 1. Let y_{ik} be the response from the k th subject at the i th visit and $\mu_i = E(y_{ik})$. The parameter of interest is μ_2 . The estimator based on completers is given by

$$\bar{y}_c = \frac{1}{n_1} \sum_{k=1}^{n_1} y_{i2k}$$

On the other hand, the estimator based on LOCF can be obtained as

$$\bar{y}_{\text{LOCF}} = \frac{1}{n} \left(\sum_{i=1}^{n_1} y_{i2k} + \sum_{i=n_1+1}^n y_{i1k} \right)$$

It can be verified that the bias of $\bar{y}_c$ is 0 with variance $\sigma^2 = n_1$, while of the bias of $\bar{y}_{\text{LOCF}}$ is $n_2(\mu_1 - \mu_2)/n$ with variance $\sigma^2/(n_1 + n_2)$.

As noted, although LOCF may introduce some bias, it decreases the variability. In a clinical trial with multiple visits, usually, $\mu_j \approx \mu_s$ if $j \approx s$. This implies that the LOCF is recommended if the patients withdraw from the study at the end of the study. However, if a patient drops out of the study at the very beginning, the bias of the LOCF could be substantial. As a result, it is recommended that the results from the analysis based on LOCF be interpreted with caution.^[4]

Hypothesis Testing

In practice, the LOCF is viewed as a pure imputation method for testing the null hypothesis of

$$H_0 : \mu_{1s} = \dots = \mu_{rs}$$

where μ_{ij} are defined in Eq. 1.

Shao and Zhong^[5] provided another look of statistical properties of the LOCF under the above null hypothesis. More specifically, they partitioned the total patient population into s subpopulations according to the time when the number of patients drop out from the study. Note that in their definition, the patients who complete the study are considered a special case of “drop out” at the end of the study. Then μ_{ij} represents the population mean of the j th subpopulation under treatment i . Assume that the j th subpopulation under the i th treatment accounts for $p_i \times 100\%$ of the overall population under the i th treatment. They argue that the objective of the intend-to-treat analysis is to test the following hypothesis test

$$H_0 : \mu_1 = \dots = \mu_r \quad (2)$$

where $\mu_i = \sum_j^s 1 p_{ij} \mu_{ij}$. Based on the above hypothesis, Shao and Zhong^[5] indicated that the LOCF bears the following properties:

1. In the special case of $r = 2$, the asymptotic ($n_i \rightarrow \infty$) size of the LOCF under H_0 is $\leq \alpha$ if and only if

$$\lim \left(\frac{n_2 \tau_1^2}{n} + \frac{n_1 \tau_2^2}{n} \right) \leq \lim \left(\frac{n_1 \tau_1^2}{n} + \frac{n_2 \tau_2^2}{n} \right)$$

where

$$\tau_i^2 = \sum_{j=1}^s p_{ij} (\mu_{ij} - \mu_i)^2$$

The LOCF is robust in the sense that its asymptotic size is α if $\lim(n_1/n) = (n_2/n)$ or $\tau_1^2 = \tau_2^2$. Note that, in reality, $\tau_1^2 = \tau_2^2$ is impractical unless $\mu_{ij} = \mu_i$ for all j . However, $n_1 = n_2$ (as a result $\lim(n_1/n) = \lim(n_2/n)$) is very typical, in practice. The above observation indicates in such a situation ($n_1 = n_2$) that LOCF is still valid.

2. When $r = 2$, $\tau_1^2 \neq \tau_2^2$, and $n_1 \neq n_2$, the LOCF has an asymptotic size smaller than α if $(n_2 - n_1)\tau_1^2 < (n_2 - n_1)\tau_2^2$ or larger than α if “ $<$ ” in Eq. 3 is replaced by “ $>$.”
3. When $r \geq 3$, the asymptotic size of the LOCF is generally not α except for some special case (e.g., $\tau_1^2 = \tau_2^2 = \dots = \tau_r^2 = 0$).

Because the LOCF usually does not produce a test with asymptotic significance level α when $r \geq 3$, Shao and Zhong^[5] proposed the following testing procedure based on the idea of poststratification. The null hypothesis H_0 should be rejected if $T > \chi_{1-\alpha, r-1}^2$, where $\chi_{1-\alpha, r-1}^2$ is the $(1 - \alpha)$ th quantile of a chi-square random variable With $r - 1$ degrees of freedom and

$$T = \sum_{i=1}^r \frac{1}{\hat{V}_i} \left(\bar{y}_{i-} - \frac{\sum_{i=1}^r \bar{y}_{i-} / \hat{V}_i}{\sum_{i=1}^r 1 / \hat{V}_i} \right)^2$$

$$\hat{V}_i = \frac{1}{n_i(n_i - 1)} \sum_{j=1}^s \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{i-})^2$$

Under the Eq. model 1 and the null hypothesis of 2, this procedure has the exact type I error α .

MEAN/MEDIAN IMPUTATION

Missing ordinal responses are also commonly encountered in clinical research. For those types of missing data, mean or median imputation is commonly considered. Let x_i be the ordinal response from the i th subject, where $i = 1, \dots, n$. The parameter of interest is $\mu = E(x_i)$. Assume that x_i for $i = 1, \dots, n_1 < n$ are observed and the rest are missing.

Median imputation will impute the missing response by the median of the observed response (i.e., x_i , $i = 1, \dots, n_1$). The merit of median imputation is that it can keep the imputed response within the sample space as the original response by appropriately defining the median. The sample mean of the imputed data set will be used as an estimator for the population mean. However, as the parameter of interest is population mean, the median imputation may lead to biased estimates.

As an alternative, mean imputation will impute the missing value by the sample mean of the observed units (i.e., $1/n_1 \sum_{i=1}^{n_1} x_i$). The disadvantage of the mean imputation is that the imputed value may be out of the original response sample space. However, it can be shown that the sample mean of the imputed data set is a consistent estimator of population mean. Its variability can be assessed by jackknife method proposed by Rao and Shao.^[6]

In practice, usually, each subject will provide more than one ordinal response. The summation of those ordinal responses (total score) is usually considered as the primary efficacy parameter. The parameter of interest is the population mean of the total score. In such a situation, mean/median imputation can be carried out for each ordinal response within each treatment group.

REGRESSION IMPUTATION

The method of regression imputation is usually considered when covariates are available. Regression imputation assumes a linear model between the response and the covariates. The method of regression imputation has been studied by various authors.^[7,8]

Let y_{ijk} be the response from the k th subject in the i th treatment group at the j th visit. The following regression model is considered:

$$y_{ijk} = \mu_i + \beta_i x_{ij} + \epsilon_{ijk} \quad (3)$$

where x_{ij} is the covariate of the k th subject in the i th treatment group. In practice, the covariates x_{ij} could be demographic variables (e.g., age, sex, and race) or patient's baseline characteristics (e.g., medical history or disease severity). Model 4 suggests a regression imputation method. Let $\hat{\mu}_i$ and $\hat{\beta}_i$ denote the estimators of μ_i and β_i based on complete data set, respectively. If y_{ijk} is missing, its predicted mean value $y_{ijk}^* = \hat{\mu}_i + \hat{\beta}_i x_{ij}$ is used to impute. The imputed values are treated as true responses and the usual ANOVA is used to perform the analysis.

MARGINAL/CONDITIONAL IMPUTATION FOR CONTINGENCY TABLES

In an observational study, two-way contingency tables can be used to summarize two-dimensional categorical data.

Each cell (category) in a two-way contingency table is defined by a two-dimensional categorical variable (A, B) , where A and B take values in $\{1, \dots, a\}$ and $\{1, \dots, b\}$ respectively. Sample cell frequencies can be computed based on the observed responses of (A, B) from a sample of units (subjects). Statistical interest includes the estimation of cell probabilities and testing hypotheses of goodness of fit or the independence of the two components A and B . In an observational study, there can be more than one strata. It is assumed that within a strata, sampled units independently have the same probability π_A to have missing B and observed A , π_B to have missing A and observed B , π_C to have observed A and B . (The probabilities π_A , π_B , and π_C may be different in different imputation classes.) As units with both A and B missing are considered as unit nonrespondent, they are excluded in the analysis. As a result, without loss of generality, it is assumed that $\pi_A + \pi_B + \pi_C = 1$.

For a two-way contingency table, it is very important for an appropriate imputation method to keep imputed values in the appropriate sample space. Whether in calculating the cell probability or in testing hypotheses (e.g., testing independence or goodness of fit), the corresponding statistical procedures are all based on the frequency counts of a contingency table. If the imputed value is out of the sample space, additional categories will be produced which is of no practical meaning. As a result, two hot deck imputation methods are thoroughly studied by Shao and Wang.^[9]

Simple Random Sampling

Consider a sampled unit with observed $A = i$ and missing B . Two imputation methods were studied by Shao and Wang.^[8] The marginal (or unconditional) random hot deck imputation method imputes B by the value of B of a unit randomly selected from all units with observed B . The conditional hot deck imputation method imputes B by the value of B of a unit randomly selected from all units with observed B and $A = i$. All nonrespondents are imputed independently.

Point Estimation

After imputation, the cell probabilities p_{ij} can be estimated using the standard formulas in the analysis of data from a two-way contingency table by treating imputed values as observed data. Denote these estimators by $\hat{p}_{ij}^I$ where $i = 1, \dots, a$, and $j = 1, \dots, b$. Let

$$\hat{p}^I = (\hat{p}_{11}^I, \dots, \hat{p}_{1b}^I, \dots, \hat{p}_{a1}^I, \dots, \hat{p}_{ab}^I)'$$

and

$$p = (p_{11}, \dots, p_{1b}, \dots, p_{a1}, \dots, p_{ab})'$$

where $p_{ij} = P(A = i, B = j)$. Intuitively, marginal random hot deck imputation leads to consistent estimators of $p_i = P(A = i)$ and $p_j = P(B = j)$, but not p_{ij} . Shao and

Wang^[8] showed that $\hat{p}^I$ under conditional hot deck imputation are consistent, asymptotically unbiased, and asymptotically normal.

Theorem 1. Assume that $\pi_C > 0$. Under conditional hot deck imputation,

$$\sqrt{n}(\hat{p}^I - p) \rightarrow_d N(0, MPM' + (1 - \pi_C)P)$$

Where $P = \text{diag}\{p\} = pp'$ ($\text{diag}\{a\}$ is a diagonal matrix whose i th diagonal element is the i th component of the vector a),

$$M = \frac{1}{\sqrt{\pi_C}} \left(I_{a \times b} - \pi_A \text{diag}\{p_{B|A}\} I_a \otimes U_b \right) - \pi_B \text{diag}\{p_{A|B}\} U_a \otimes I_b,$$

$$p_{A|B} = (p_{11}/p_{.1}, \dots, p_{1b}/p_{.b}, \dots, p_{a1}/p_{.1}, \dots, p_{ab}/p_{.b})',$$

$$p_{B|A} = (p_{11}/p_{1.}, \dots, p_{1b}/p_{1.}, \dots, p_{a1}/p_{a.}, \dots, p_{ab}/p_{a.})'$$

Where I_a denotes an a -dimensional identity matrix, U_b denotes a b -dimensional square matrix with all components being 1, and $\otimes$ is the Kronecker product.

Goodness-of-Fit Test

A direct application of Theorem 1 is to obtain a Wald-type test for goodness of fit. Consider the null hypothesis of the form $H_0: p = p_0$, where p_0 is a known vector. Under H_0 ,

$$X_W^2 = n(\hat{p}^* - p_0^*)' \Sigma^{*-1} (\hat{p}^* - p_0^*) \rightarrow_d \chi_{ab-1}^2$$

where χ_v^2 denotes a random variable having the chi-square distribution with v degrees of freedom, $\hat{p}^*(p_0^*)$ is obtained by dropping the last component of $\hat{p}^I(p_0)$, and $\hat{\Sigma}^*$ is the estimated asymptotic covariance matrix of $\hat{p}^*$, which can be obtained by dropping the last row and column of $\hat{\Sigma}$, the estimated asymptotic covariance matrix of $\hat{p}^I$.

Noting the fact that the computation of $\hat{\Sigma}^{*-1}$ is complicated, Wang and Shao proposed a simple correction of the standard Pearson chi-square statistic by matching the first-order moment, an approach developed by Rao and Scott.^[10] Let

$$X_G^2 = n \sum \frac{(\hat{p}_{ij}^I - p_{ij})^2}{p_{ij}}$$

It is noted that under conditional imputation the asymptotic expectation of X_G^2

$$\frac{1}{\pi_C} (ab + \pi_A^2 a + \pi_B^2 b - 2\pi_A a - 2\pi_B b - 2\pi_A \pi_B + 2\pi_A \pi_B \delta) - \pi_C ab + (ab - 1)$$

Let (see equation at bottom of the page below). Then the asymptotic expectation of X_G^2/λ is $ab - 1$, which is the first-order moment of a standard chi-square variable with $ab - 1$ degrees of freedom. Thus, X_G^2/λ can be used just like a normal chi-square statistic to test the goodness of fit. However, it should be noted that this is just an approximated test procedure which is not asymptotically correct.

According to a Shao and Wang's simulation study, this test performs reasonably well with moderate sample sizes.

TESTING FOR INDEPENDENCE

Testing for the independence between A and B can be performed by the following chi-square statistic when there is no missing data

$$X^2 = n \sum_{ij} \frac{(\hat{p}_{ij}^I - \hat{p}_i^I \hat{p}_j^I)^2}{\hat{p}_i^I \hat{p}_j^I} \rightarrow d \chi_{(a-1)(b-1)}^2$$

It is of interest to know what the asymptotic behavior of the above chi-square statistic is under both marginal and conditional imputation. It is found that under the null hypothesis of A and B are independent and conditional hot deck imputation

$$X^2 \rightarrow d(\pi_C^{-1} + 1 - \pi_C) \chi_{(a-1)(b-1)}^2$$

and under marginal hot deck imputation

$$X_{MI}^2 \rightarrow d \chi_{(a-1)(b-1)}^2$$

Results Under Stratified Simple Random Sampling

When Number of Strata is Small

Stratified samplings are also commonly used in medical study. For example, a large epidemiology study is usually conducted by several large centers. Those centers are usually considered as strata. For those types of study, the number of strata is not very large; however, the sample size within each strata is very large. As a result, imputation is usually carried out within each strata.

Within the h th stratum, we assume that a simple random sample of size n_h is obtained and samples across strata are obtained independently. The total sample size is $n = \sum_{h=1}^H n_h$ where H is the number of strata and n_h is the sample size in stratum h . The parameter of interest is the overall cell probability vector $p = \sum_{h=1}^H w_h p_h$, where w_h is the h th stratum weight. The estimator of p based on conditional imputation is given by $\hat{p}^I = \sum_{h=1}^H w_h \hat{p}_h^I$. Assume that $n_h = n \rightarrow p$ as $n \rightarrow \infty$, $h = 1, \dots, H$. Then a direct application of Theorem 1 leads to

$$\sqrt{n}(\hat{p}^I - p) \rightarrow d N(0, \Sigma)$$

where

$$\Sigma = \sum_{h=1}^H \frac{w_h^2}{p_h} \Sigma_h$$

and Σ_h is the Σ in Theorem 1 but restricted to the h th stratum.

When Number of Strata is Large

In a medical survey, it is also possible to have the number of strata (H) very large, while the sample size within each strata is small. A typical example is that if a medical survey is conducted by family, then the family can be considered as a strata and all the members within the family become the samples from this strata. In such a situation, the method of imputation within stratum is impractical because it is possible that within a stratum, there are no completers. As an alternative, Wang and Shao^[9] proposed the method of imputation across strata under the assumption that $(\pi_{hA}, \pi_{hB}, \pi_{hC})$, where $h = 1, \dots, H$, is constant. More specifically, let $n_{h,ij}^C$ denote the number of completers in the h th stratum such that $A = i$ and $B = j$. For a sampled unit in the k th imputation class with observed $B = j$ and missing A , the missing value is imputed by i according to the conditional probability

$$p_{ij} \Big| B, k = \frac{\sum_h w_h n_{h,ij}^C / n_h}{\sum_h w_h n_{h,j}^C / n_h}$$

Similarly, the missing value of a sampled unit in the k th imputation class with observed $A = i$ and missing B can be imputed by j according to the conditional probability

$$p_{ij} \Big| A, k = \frac{\sum_h w_h n_{h,ij}^C / n_h}{\sum_h w_h n_{h,i}^C / n_h}$$

Again, $\hat{p}^I$ can be computed by ignoring imputation classes and treating imputed values as observed data. The following result establishes the asymptotic normality of

$$\lambda = \frac{\frac{1}{\pi_C} (ab + \pi_A^2 a + \pi_B^2 b - 2\pi_A a - 2\pi_B b - 2\pi_A \pi_B + 2\pi_A \pi_B \delta) + (ab - 1)}{ab - 1}$$

based on the method of conditional hot deck imputation across strata.

Theorem 2. Let $(\pi_{hA}, \pi_{hB}, \pi_{hC}) = (\pi_A, \pi_B, \pi_C)$ for all h . Assume further that $H \rightarrow \infty$ and that there are constants c_j for $j = 1, \dots, 4$, such that $n_h \leq c_1$, $c_2 \leq H w_h \leq c_3$, and $p_{h,ij} > c_4$ for all h . Then

$$\sqrt{n}(\hat{p}^I - p) \rightarrow d N(0, \Sigma)$$

where Σ is the limit of

$$n \left(\sum_h \frac{w_h^2}{n_h} \Sigma_h + \Sigma_A + \Sigma_B \right)$$

Details regarding the expression of $\Sigma_h, \dots, \Sigma_A$, and Σ_B and their procedure can be found in Wang.^[9]

CONCLUSION

Missing values or incomplete data are commonly encountered in clinical research. How to handle the incomplete data is always a challenge to the statisticians in practice. Imputation as one of very popular methodology to compensate for the missing data is widely used in biopharmaceutical research. As compared to its popularity, however, its theoretical properties are far away from well understood. Further research is needed. More information regarding the prevention and treatment of missing data in clinical trials can be found in.^[11]

ACKNOWLEDGEMENTS

Minor updates have been made by the Editor-in-Chief in order to reflect developments since publication of the *Encyclopedia of Biopharmaceutical Statistics, Third Edition*.

REFERENCES

1. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*; Wiley: New York, 1987.
2. Schafer, J.L. *Analysis of Incomplete Multivariate Data*; Chapman and Hall: London, 1997.
3. Kalton, G.; Kasprzyk, D. The treatment of missing data. *Sury. Methodol.* **1986**, *12*, 1–16.
4. Cheng, B.; Chow, S.C. Validity of LDCE. In *Encyclopedia of Biopharmaceutical Statistics*, 2nd ed.; Chow, S.C., Ed.; Marcel Dekker, Inc: New York, 2003; 1023–1029.
5. Shao, J.; Zhong, B. Last Observation Carry-Forward and Intention-to-Treat Analysis. *Stat. Med.* **2003**, *22*, 2429–2441.
6. Rao, J.N.K.; Shao, J. Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **1992**, *79*, 811–822.
7. Srivastava, M.S.; Carter, E.M. The maximum likelihood method for non-response in sample surveys. *Surv. Methodol.* **1986**, *12*, 61–72.
8. Shao, J.; Zhong, H. Sample correlation coefficients based on survey data under regression imputation. *J. Am. Stat. Assoc.* **2002**, *97*, 544–552.
9. Wang, H. Two-Way Contingency Tables with Marginally and Conditionally Imputed Nonrespondents. Ph.D. Thesis; Department of Statistics, University of Wisconsin-Madison, 2001.
10. Rao, J.N.K.; Scott, A.J. On simple adjustments to chi-square tests with sample survey data. *J. Annal. Stat.* **1987**, *15*, 1–12.
11. NRC. The Prevention and Treatment of Missing Data in Clinical Trials. National Research Council, Division of Behavioral and Social Sciences and Education, Committee on National Statistics, Panel on Handling Missing Data in Clinical Trials, National Academies Press, Washington, DC, 2010.

In Vitro Testing: Bioequivalence

Hansheng Wang

Peking University, Beijing, People's Republic of China

Ying Zhang

AIG American General, Neptune, New Jersey, U.S.A.

Jun Shao

University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

It is widely recognized that, as opposed to *in vivo*, *in vitro* methods are less variable, easier to control, and more likely to detect differences between products if they exist. This entry details several analysis methods for *in vitro* bioequivalence testing as well as discussing some of the difficulties associated with profile analysis.

INTRODUCTION

Bioequivalence testing is considered as a surrogate for the clinical evaluation of the therapeutic equivalence of drug products based on the *Fundamental Bioequivalence Assumption* that when two drug products (e.g., a brand name drug and its generic copy) are equivalent in bioavailability, they will reach the same therapeutic effect.^[1]

In other words, two drugs are called bioequivalent if they are equivalent in bioavailability. In practice, most drugs' active moiety is absorbed into the blood system before it takes action. As a result, the pharmacokinetic (PK) parameters taken from the bloodstream—such as the area under the plasma concentration curve (AUC), the maximum concentration ($C_{\max}$), and the time to the maximum concentration ($T_{\max}$)—are usually considered to characterize the bioavailability of the drug. This type of bioequivalence is called *in vivo* bioequivalence. However, for some locally acting drug products, such as nasal aerosols (e.g., metered dose inhalers) and nasal sprays (e.g., metered dose spray pumps) that are not intended to be absorbed into the bloodstream, bioavailability may be assessed by measurements intended to reflect the rate and the extent to which the active ingredient or active moiety becomes available at the site of action. For these local delivery drug products, the U.S. Food and Drug Administration (FDA) indicates that bioequivalence may be assessed, with suitable justification, by *in vitro* bioequivalence studies alone (21 CFR 320.24). Although it is recognized that *in vitro* methods are less variable, easier to control, and more likely to detect differences between products if they exist, the clinical relevance of the *in vitro* tests or the magnitude

of the differences in the tests is not clearly established until a guidance on bioavailability and bioequivalence studies for nasal aerosols and nasal sprays for local action has been issued by the FDA.^[2]

STUDY DESIGN AND DATA COLLECTION

For the assessment of *in vitro* bioequivalence, the 1999 FDA guidance^[2] requires that *in vitro* testing of emitted dose uniformity, droplet size distribution, spray pattern, plume geometry, priming/repriming, and tail-off profile be done to demonstrate comparable delivery characteristics between two drug products. However, the 2003 FDA draft guideline^[3] indicates that *in vitro* bioequivalence can be established through the following tests for single actuation content through container life, droplet size distribution by laser diffraction, drug in small particles/droplets, or particle/droplet size distribution by cascade impactor, drug particle size distribution by microscopy, spray pattern, plume geometry, and priming and re-priming. In this section, a brief description of study design and some of these tests are given.

Study Design

According to the FDA,^[2] three lots/sublots from each product are required to be tested for *in vitro* emitted dose uniformity, droplet size distribution, spray pattern, plume geometry, priming/repriming, and tail-off profile. For each *in vitro* test, 10 samples are randomly drawn from each lot. Samples are randomized for *in vitro* tests. The analysts will

not have access to the randomization codes. An automated actuation station with a fixed settings (actuation force, dose time, return time, and hold time) is usually used for the *in vitro* tests.

Priming, Emitted Dose Uniformity, Priming/Repriming, and Tail-Off Profile

Based on the FDA comment, the priming, emitted dose uniformity, priming/repriming, and tail-off tests may be tested in the following setting. Three individual lots of test product and reference product are evaluated. For each lot, 10 samples are then tested for pump priming, unit spray content through life, and tail-off studies. Then additional samples for each lot are evaluated for the prime hold study (reprime study).

For each sample unit, spray samples are collected for sprays 1–8 and analyzed in order to determine the minimum number of actuations required before the pump delivers the labeled dose of drug (sprays 1–8). To characterize emitted dose uniformity at the beginning of unit life, spray 9 is collected. Sprays 10–15 are wasted by the automatic actuation station. Spray 16 is collected in the middle of unit life. Sprays 17–20 are wasted. Sprays 21–23 are collected at the end of the unit life. Additional sprays after spray 23 are collected and analyzed to determine the tail-off profile.

Ten additional samples are drawn randomly from each lot of drug product for the pump prime hold study. For each unit, the first 12 sprays (sprays 1–12) are wasted. Sprays 13 and 14 are collected as fully primed sprays. The unit is then stored undisturbed for 24 hr. Within each lot, five samples are placed in the upright position while the other five are placed in a side position. After that, sprays 15–17 are collected. The unit is then stored undisturbed in its former position for another 24 hour. After that, the doses emitted by sprays 18–20 are collected. All spray samples are weighted in order to obtain repriming characteristics.

Spray Pattern

A spray pattern produced by a nasal spray pump evaluates in part the integrity and the performance of the orifice and pump mechanism in delivering a dose to its intended site of deposition. Measurements can be made on the diameter of the horizontal intersection of the spray plume at different distances from the actuator tip. Spray patterns are usually measured at three distances (e.g., 1, 2, and 4 cm) at both the beginning (sprays 8–10) and the end (sprays 17–19) of unit life. As a result, a total of six spray patterns is collected for each sample unit. For each spray pattern image, the diameters (the longest and shortest diameters) and the ovality (which is defined by the ratio of the longest to the shortest diameters) are measured.

Droplet Size Distribution

For a test of droplet size distribution, methods of laser diffraction and cascade impaction are commonly used. These methods are briefly described below.

Laser Diffraction

For a test of droplet size distribution using laser diffraction particle analyzer, each sample unit is first primed by actuating the pump eight times using an automatic actuation station. Droplet size distribution is then determined at three distances (e.g., 3, 5, and 7 cm) from the laser beam and at the beginning, the middle, and the end of unit life. At each distance, three measurements of delay times (plume formation, start of dissipation, and intermediate measurements) and overall evaluation are used to characterize the droplet size. As a result, a total of 36 measurements is recorded for each sample unit.

Cascade Impaction Technique

When the spray pump is actuated in the nasal cavity, a fine mist of droplets is generated. Droplets that are $>9 \mu\text{m}$ in diameter are considered nonrespirable and are therefore useful for nasal deposition. As recommended in the FDA 1999 Guidance, the data should be reported as follows:

- Group 1: Adaptor (expansion chamber, i.e., 5-L flask), rubber gasket, throat, and Stage 0.
- Group 2: Stage 1.
- Group 3: Stage 2 to filter.

Each sample unit is first primed by actuating the pump seven times using an automatic actuation station. Droplet size distribution is then determined at the beginning and the end of the life of the sample. Thus a total of six groups of results is reported for each spray unit.

Plume Geometry

Plume geometry is performed on the nasal spray plume that is allowed to develop into an unconstrained space that far exceeds the volume of nasal cavity. It represents a frozen moment in spray plume development that is viewed from two axes perpendicular to the axis of plume development. The samples should be actuated vertically. Prime the pump with 10 actuations until a steady fine mist is produced from the pump. A fast-speed video camera is placed in front of the sample bottle and starts recording. Repeat the test by rotating the actuator 90° to the previous actuator placement so that two side views are at 90° to each other (two perpendicular planes) and, relative to the axis of the plume of the spray, are captured when actuated into space.

Spray plumes are characterized at three stages: early upon formation, as the plume starts dissipate, and at some intermediate time. Longest vertical distance (LVD), widest horizontal distance (WHD), and plume angle (ANG) are recorded and analyzed.

REGULATORY REQUIREMENTS AND STATISTICAL METHODS

Noncomparative Analysis

For each *in vitro* test, the FDA requires that a noncomparative analysis be performed. Noncomparative analysis refers to the statistical summarization of the bioavailability data by descriptive statistics. As a result, means, standard deviations, and coefficients of variation (CVs) in percentage of the six *in vitro* tests should be documented. More specifically, the overall sample means for a given formulation should be averaged over all samples (e.g., bottles/canisters), life stages (except for priming and repriming evaluations), and lots or batches. In addition to the overall means, means at each life stage for each batch averaged over all bottles/canisters and for each life stage averaged over all lots (or batches) should be presented. For profile data, means, standard deviations, and percent CVs should be reported for each stage. The between-lot (or batch), within-lot (or batch) between-sample (e.g., bottle or canister), and within-sample (e.g., bottle or canister) between-life stage variability should be evaluated through appropriate statistical models.

Bioequivalence Limit

The following formula is given in the 1999 FDA redetermining the bioequivalence (BE) limit

$$\frac{(\text{average BE limit in natural log scale})^2 + \text{variance terms of offset}}{\text{scaling variance}}$$

As it can be seen, in order to obtain the BE limit, there are three quantities that need to be specified. They are (i) average BE limit, (ii) variance terms offset, and (iii) scaling variance, respectively. The 1999 FDA guidance indicates that the final specification of those parameters should be based on the results of the ongoing simulation study. However, the following values are recommended in the FDA's draft guidance.

As a result of the low variability of *in vitro* measurements, at the present time, the FDA recommends that the ratio of geometric means should fall within 0.90 and 1.11. As a result, a value of 0.90 is recommended as the average BE limit for *in vitro* data.

The objective of variance terms offset is to allow some difference among the total variances that may be inconsequential. As a result of the low variability of *in vitro*

measurements, the FDA recommends that a value of 0 should be taken based on the guidance of population and individual bioequivalence.^[4] In practice, however, a value of 0.01 may be accepted by the FDA for variance terms offset depending upon the nature of the drug products under investigation.

The purpose of scaling variance is to adjust the BE criterion depending on the reference product variance. When the reference variance is greater than the scaling variances, the limit is widened. On the other hand, the limit is narrowed when reference variance is less than scaling variance. The FDA indicates that the choice of the scaling variance should be at least 0.1.

As a result, the specification of 0.90 for the average BE limit—0.0 for the variance offset and 0.10 for scaling standard deviation—gives the following BE limit:

$$\theta_{BE} = \frac{\log(0.9)^2 + 0}{0.1^2} = 1.11$$

Nonprofile Analysis

The FDA^[3] indicates that the *in vitro* bioequivalence of nasal aerosols and sprays can be established by seven bioequivalence tests. They are classified as either the nonprofile analysis or the profile analysis. Nonprofile analysis applies to emitted dose or spray content uniformity, through container life, droplet size distribution, spray pattern, and priming/repriming. The criterion for nonprofile *in vitro* bioequivalence is described as follows.

Let y_T , y_R , and y'_R be independent *in vitro* bioavailabilities, where y_T is from the test product and y_R and y'_R are from the reference product. The two products are said to be *in vitro* bioequivalent if $\theta < \theta_{BE}$, where:

$$\theta = \frac{E(y_R - y_T)^2 - E(y_R - y'_R)^2}{\max\{\sigma_0^2, E(y_R - y'_R)^2/2\}} \quad (1)$$

θ_{BE} a prespecified BE limit and σ_0^2 is a prespecified constant. Values of σ_0^2 and θ_{BE} can be found in the FDA guidance. According to the FDA guidance, *in vitro* bioequivalence can be claimed if the hypothesis that $\theta \geq \theta_{BE}$ is rejected at the 5% level of significance, provided that the ratio of geometric means between drug products is within 0.90 and 1.11.

Let m_T and m_R be the number of canisters from the test product and the reference product, respectively, and one observation from each sample (e.g., bottle or canister) is obtained. The data can be described by the following model:

$$y_{jk} = \mu_k + \varepsilon_{jk}, \quad j = 1, \dots, m_k \quad (2)$$

where $k = T$ for the test product, $k = R$ for the reference product, μ_T and μ_R are fixed product effects, ε_{jk} 's are

independent random measurement errors distributed as $N(0, \sigma_k^2)$, $k = T, R$. Under model 2, the parameter θ in Eq. 1 is equal to:

$$\theta = \frac{(\mu_T - \mu_R)^2 + \sigma_T^2 - \sigma_R^2}{\max\{\sigma_0^2, \sigma_R^2\}} \quad (3)$$

As a result, $\theta < \theta_{BE}$ is equivalent to where ζ is the linearized BE parameter given by:

$$\zeta = (\mu_T - \mu_R)^2 + \sigma_T^2 - \sigma_R^2 - \theta_{BE} \max\{\sigma_0^2, \sigma_R^2\} \quad (4)$$

The bioequivalence can be concluded by constructing an approximate 95% upper bound for ζ . If the 95% upper bound is less than 0, *in vitro* bioequivalence is established, or otherwise rejected. In order to obtain the approximate 95% upper bound, the FDA^[2,3] recommends the following procedure proposed by Hyslop et al.,^[5] which was originally developed for the establishment of *in vivo* individual bioequivalence under a 2×4 crossover design. Let ζ_U denote the 95% upper bound for ζ and:

$$\begin{aligned} \hat{\delta} &= \bar{y}_T - \bar{y}_R \\ s_k^2 &= \frac{1}{m_k - 1} \sum_{j=1}^{m_k} (y_{jk} - \bar{y}_k)^2 \\ \bar{y}_k &= \frac{1}{m_k} \sum_{j=1}^{m_k} y_{jk} \end{aligned}$$

Then, it follows that:

$$\tilde{\zeta}_U = \hat{\delta}^2 + s_R^2 - s_T^2 \theta_{BE} \max\{\sigma_0^2, s_R^2\} + \sqrt{U}$$

where U is the sum of the following three quantities:

$$\begin{aligned} &\left[\left(|\hat{\delta}| + z_{0.95} \sqrt{\frac{S_T^2}{m_T} + \frac{S_R^2}{m_R}} \right)^2 - \hat{\delta}^2 \right] \\ &S_T^4 \left(\frac{m_T - 1}{\chi_{0.05; m_T - 1}^2} - 1 \right)^2 \end{aligned}$$

and

$$(1 + c\theta_{BE})^2 S_R^4 \left(\frac{m_R - 1}{\chi_{0.95; m_R - 1}^2} - 1 \right)^2$$

where $c = 1$ if $S_R^2 \geq \sigma_0^2$, $c = 0$ if $S_R^2 < \sigma_0^2$, z_a is the a th quantile of the standard normal distribution, and $\chi_{r,a}^2$ is the a th quantile of the central chi-square distribution with r degrees of freedom. As suggested by the FDA, *in vitro* bioequivalence can be claimed if $\tilde{\zeta}_U < 0$. This procedure is recommended by the FDA guidance.

Profile Analysis

As indicated in the FDA^[2] profile analysis using a confidence interval approach should be applied to cascade impactor (CI) or multistage liquid impinger (MSLI) for particle size distribution. As indicated in the 1999 FDA guidance, equivalence may be assessed based on chi-square differences. The idea is to compare the profile difference between test product and reference product samples to the profile valuation between reference product samples. More specifically, let y_{ijk} denote the observation from the j th subject's i th stage in the k th treatment. Given a sample (j_0) from test product and two samples (j_1, j_2) from reference products and assuming that there are a total of S stages, the profile distance between test and reference is given by:

$$d_{TR} = \sum_{i=1}^S \frac{(y_{ij0T} - 0.5(y_{ij1R} + y_{ij2R}))^2}{(y_{ij0T} - 0.5(y_{ij1R} + y_{ij2R}))^2}$$

Similarly, the profile variability within reference is defined to be:

$$d_{RR} = \sum_{i=1}^S \frac{(y_{ij1R} - y_{ij2R})^2}{0.5(y_{ij1R} + y_{ij2R})}$$

For a given triplet sample of (Test, Ref. [1], Ref. [2]), the ratio of d_{TR} and d_{RR} (i.e., $rd = d_{TR}/d_{RR}$) can then be used as a bioequivalence measure for the triplet samples between the two drug products. For a selected sample, the 95% upper confidence bound of $E(rd) - E(d_{TR}/d_{RR})$ is then used as a bioequivalence measure for the determination of bioequivalence. In other words, if the 95% upper confidence bound is less than the bioequivalence limit, then we claim that the two products are bioequivalent.

The FDA^[2] recommends a bootstrap procedure to construct the 95% upper bound for $E(rd)$. The procedure is described below. Assume that the samples are obtained in a two-stage sampling manner. In other words, for each treatment (test or reference), three lots are randomly sampled. Within each lot, 10 samples (e.g., bottles and canisters) are sampled. The following paragraph is quoted from the 1999 FDA guidance regarding the bootstrap procedure to establish profile bioequivalence.

For an experiment consisting of three lots each of test and reference products, and with 10 canisters per lot, the lots can be matched into six different combinations of triplets with two different reference lots in each triplet. The 10 canisters of a test lot can be paired with the 10 canisters of each of the two reference lots in $(10 \text{ factorial})^2 = (3,628,800)^2$ combinations in each of the lot triplets. Hence a random sample of the N canister pairing of the six Test–Ref. [1]–Ref. [2] lot triplets is needed. rd is estimated by the sample mean of the rd s calculated for the triplets in 10 selected samples of N .

In the sample guidance, the FDA recommends that a value of 500 should be taken for N .

RECENT DEVELOPMENT

Nonprofile Analysis

As indicated earlier, the FDA^[2] requires at least 30 samples to be taken from each of the test and the reference drug products (i.e., $m_k = 30$). However, $m_k = 30$ may not be enough to achieve a desired power for the bioequivalence test. In practice, there are two options to increase the power. One is to increase the sample size and the other one is to increase the replicates per sample. Increasing the sample size can certainly increase the power, but in some situations, obtaining replicates from each bottle or canister may be more practical and/or cost-effective. However, how to perform the statistical analysis based on replicates becomes a problem of interest.

Chow et al.^[6] proposed the following method to assess *in vitro* bioequivalence when a replicate from each bottle or canister is available. Suppose that there are n_k replicates from each bottle or canister for product k . Let y_{ijk} be the i th replicate in the j th canister under product k . Let b_{jk} be the between-bottle or between-canister variation and let e_{ijk} be the within-bottle or within-canister measurement error. Then:

$$y_{ijk} = \mu_k + b_{jk} + e_{ijk}, \quad i = 1, \dots, n_k, \quad j = 1, \dots, m_k \quad (5)$$

where b_{jk} 's and e_{ijk} 's are independent, $b_{jk} \sim N(0, \sigma_{Bk}^2)$, and $e_{ijk} \sim N(0, \sigma_{Wk}^2)$. Under model 5, the total variances σ_T^2 and σ_R^2 in Eqs. 3 and 4 are equal to $\sigma_{BT}^2 + \sigma_{WT}^2$ and $\sigma_{BR}^2 + \sigma_{WR}^2$, respectively (i.e., the sums of between-bottle or between-canister and within-bottle or within-canister variances). The parameter θ in Eq. 1 is still given by Eq. 3 and $\theta < \theta_{BE}$ if and only if $\zeta < 0$, where ζ is given in Eq. 4.

Under model 5, an approximate 95% upper bound for ζ is given by:

$$\hat{\zeta}_U = \hat{\delta}^2 + s_{BT}^2 + (1 - n_T^{-1})s_{WT}^2 - s_{BR}^2 - (1 - n_R^{-1})s_{WR}^2 - \theta_{BE} \max\{\sigma_0^2, s_{BR}^2 + (1 - n_R^{-1})s_{WR}^2\} + \sqrt{U}$$

where U is the sum of the following five quantities:

$$\left[\left(|\hat{\delta}| + z_{0.95} \sqrt{\frac{s_{BT}^2}{m_T} + \frac{s_{BR}^2}{m_R}} \right)^2 - \hat{\delta}^2 \right] s_{BT}^4 \left(\frac{m_T - 1}{\chi_{0.05; m_T - 1}^2} - 1 \right)^2$$

$$(1 - n_T^{-1})^2 s_{WT}^4 \left(\frac{m_T(n_T - 1)}{\chi_{0.05; m_T(n_T - 1)}^2} - 1 \right)^2$$

$$(1 + \theta_{BE})^2 s_{BR}^4 \left(\frac{m_R - 1}{\chi_{0.95; m_R - 1}^2} - 1 \right)^2$$

and

$$(1 + c\theta_{BE})^2 (1 - n_R^{-1})^2 s_{WR}^4 \left(\frac{m_R(n_R - 1)}{\chi_{0.95; m_R(n_R - 1)}^2} - 1 \right)^2$$

and $c = 1$ if $s_{BR}^2 + (1 - n_R^{-1})s_{WR}^2 \geq \sigma_0^2$ and $c = 0$ if $s_{BR}^2 + (1 - n_R^{-1})s_{WR}^2 < \sigma_0^2$. *In vitro* bioequivalence can be claimed if $\hat{\zeta}_U < 0$.

If the difference between model 2 and model 5 is ignored and the confidence bound $\hat{\zeta}_U$ with m_k replaced by $m_k n_k$ (instead of $\hat{\zeta}_U$) is used, then the asymptotic size of the test procedure is not 5%.

Profile Analysis

The bootstrap procedure described in the section “Bioequivalence Limit” has received much attention and criticism since it was introduced by the FDA. The major criticisms are described below.

First, the statistical properties of this procedure are unknown. It includes two aspects. One is that the statistical model, which should be used to describe the profile data, is not clearly defined in the FDA guidance. In addition, even under an appropriate statistical model, the statistical properties of the bootstrap procedure are still unknown. More specifically, is the bootstrap sample mean a consistent estimator for $E(rd)$? Is the 95% percentile of the bootstrap samples an appropriate 95% upper bound for $E(rd)$? These questions are not addressed in the FDA guidance.

Second, no criteria are given regarding the passage or failure of the bioequivalence study. This is the issue that confuses most researchers/scientists in practice. After the conduction of a valid trial and an appropriate statistical analysis following the FDA guidance, the sponsor still cannot tell if its product has passed or failed the bioequivalence test. This is a direct consequence of our first point (i.e., the statistical properties of the recommended bootstrap procedure are unknown).

Third, the simulation study using different random number generation schemes may produce contradictory results. It is possible for a good product to fail the bioequivalence test simply because of “bad luck.” It is also possible for a bad product to pass the bioequivalence test with an “appropriate” choice of random number generation scheme. As a result, researchers/scientists tends to reply

more on the descriptive statistics of the two drug products (treatment and reference) in order to assess their bioequivalence instead of the bootstrap procedure. The proposed bootstrap procedure recommended by FDA is not as reliable as it should be.

As a result, further research of profile analysis becomes a problem of interest in practice. More specifically, the questions of interest include: (i) What statistical model should be used to describe the profile data? (ii) Is $E(rd)$ defined by the FDA as a good parameter for characterizing the bioequivalence between test and reference products? Can we define the test-to-reference distance and reference-to-reference variability differently? (iii) What BE limit should be used? (iv) What statistical procedure should we use to evaluate the *in vitro* bioequivalence between the two products based on appropriate model, parameter, and bioequivalence criterion?

Sample Size

The FDA^[2,3] indicates that an *in vitro* bioequivalence study should be based on testing a suitable number of bottle or canisters from each of three batches (lots) of the test and reference drug products. The number of bottles (canisters) to be studied for each batch (lot) should be no less than 10. As a result, a minimum of 30 samples is required for each of the test and the reference drug products.

However, the FDA's requirement may not yield sufficient power for the establishment of *in vitro* bioequivalence between drug products. Chow et al.^[6] propose a procedure for determining sample sizes as follows.

In practice, it is commonly preferred to choose $m = m_T = m_R$ and $n = n_T = n_R$ so that the power of the bioequivalence test reaches a given level β (say 80%) when the unknown parameter vector $\psi = (\delta, \sigma_{BT}^2, \sigma_{BR}^2, \sigma_{WT}^2, \sigma_{WR}^2)$ is set at some initial guessing value ψ for which the value of ζ (denoted by $\hat{\zeta}$) is negative. Let U be given in the definition of $\hat{\zeta}_U$ and let $U\beta$ be the same as U but with 5% and 95% replaced by $1 - \beta$ and β , respectively. Because:

$$P(\hat{\zeta}_U < \zeta + \sqrt{U} + \sqrt{U\beta}) \approx \beta$$

the power of the bioequivalence test $P(\hat{\zeta}_U < 0)$ is approximately larger than β if $\zeta + \sqrt{U} + \sqrt{U\beta} \leq 0$. Let $\bar{U}$ and $\tilde{U}_\beta$ be U and $U\beta$, respectively, with $(\delta, s_{BT}^2, s_{BR}^2, s_{WT}^2, s_{WR}^2)$ replaced by. Then, the sample sizes $m = m_T = m_R$ and $n = n_T = n_R$ that produce a test with a power of approximately β should satisfy:

$$\bar{\zeta} + \sqrt{\bar{U}} + \sqrt{\tilde{U}_\beta} \leq 0 \quad (6)$$

As a result, having a large m and a small n is an advantage when mn , the total number of observations for one

treatment, is fixed. Thus the optimal sample size and the replicates combination can be determined by the following procedure:

- Step 1. Set $m = 30$ and $n = 1$. If Eq. 6 holds, stop, and the required sample sizes are $m = 30$ and $n = 1$; otherwise, go to Step 2.
- Step 2. Let $n = 1$ and find a smallest integer m^* such that Eq. 6 holds. If $m^* \leq m_+$ (the largest possible number of canisters in a given problem), stop, and the required sample sizes are $m = m^*$ and $n = 1$; otherwise, go to Step 3.
- Step 3. Let $m = m_+$ and find a smallest integer n^* such that Eq. 6 holds. The required sample sizes are $m = m_+$ and $n = n^*$.

If in practice it is much easier and inexpensive to obtain more replicates than to sample more canisters, then Steps 2–3 in the previous procedure can be replaced by:

- Step 2'. Let $m = 30$ and find a smallest integer n^* such that Eq. 6 holds. The required sample sizes are $m = 30$ and $n = n^*$.

Because selecting sample sizes according to Eq. 6 only produces a test with approximate power β , we conduct a simulation study to examine the actual power corresponding to the selected sample sizes according to Steps 1–3 (or Steps 1 and 2'). That is, for a given combination of parameter values in Table 1, we select sample sizes m^* and n^* according to Steps 1–3 or Steps 1 and 2' with $\beta = 80\%$, and then compute the actual power p corresponding to the selected m^* and n^* by 10,000 simulations. Note that m_+ is set to be ∞ in the simulation so that Step 3 is not needed.

The performance of the above sample size determination procedure is evaluated by simulation. Table 1 shows the selected sample sizes for a nominal power of 80%, as reported by Chow et al.^[6] The results show that the selected sample sizes produce a test with power $\geq 75\%$, except for two cases where the power is about 73%.

CONCLUSION

For the assessment of *in vitro* bioequivalence, the FDA^[2,3] requires that *in vitro* testing of emitted dose uniformity, droplet size distribution, spray pattern, plume geometry, and priming and re-priming be done to demonstrate comparable delivery characteristics between two drug products. Those tests can be divided into two categories. They are, namely, profile and nonprofile analysis. For profile analysis, a statistical procedure similar to the one for testing individual bioequivalence^[5] are adopted. However, for profile analysis, no satisfactory statistical procedure is available for establishment of *in vitro* bioequivalence. Further study is needed.

Table 1 Selected Sample Sizes m^* and n^* and the Actual Power p when $\theta_{BE} = 1.125$ and $\sigma_0 = 0.2$ (10,000 simulations)

σ_{BT}	σ_{BR}	σ_{WT}	σ_{WR}	δ	Step 1	Step 2	Step 2'		
					P	m^*, n^*	P	m^*, n^*	P
0	0	0.25	0.25	0.0530	0.4893	55, 1	0.7658	30, 2	0.7886
				0	0.5389	47, 1	0.7546	30, 2	0.8358
				0.4108	0.6391	45, 1	0.7973	30, 2	0.8872
		0.50	0.50	0.2739	0.9138	—	—	—	—
				0.1061	0.4957	55, 1	0.7643	30, 2	0.7875
				0	0.5362	47, 1	0.7526	30, 2	0.8312
0.25	0.25	0.25	0.25	0.0750	0.4909	55, 1	0.7774	30, 3	0.7657
				0	0.5348	47, 1	0.7533	30, 2	0.7323
				0.4405	0.5434	57, 1	0.7895	30, 3	0.8489
		0.50	0.50	0.2937	0.8370	—	—	—	—
				0.1186	0.4893	55, 1	0.7683	30, 2	0.7515
				0	0.5332	47, 1	0.7535	30, 2	0.8091
0.50	0.25	0.25	0.50	0.1186	0.4903	55, 1	0.7660	30, 4	0.7586
				0	0.5337	47, 1	0.7482	30, 3	0.7778
0.25	0.50	0.25	0.25	0.2937	0.8357	—	—	—	—
				0.50	0.5016	55, 1	0.7717	30, 4	0.7764
				0	0.5334	47, 1	0.7484	30, 3	0.7942
		0.50	0.50	0.5809	0.6416	45, 1	0.7882	30, 2	0.7884
				0.3873	0.9184	—	—	—	—
				0.3464	0.6766	38, 1	0.7741	30, 2	0.8661
0.50	0.50	0.25	0.50	0.1732	0.8470	—	—	—	—
				0.3464	0.6829	38, 1	0.7842	30, 2	0.8045
				0.1732	0.8450	—	—	—	—
		0.50	0.50	0.1500	0.4969	55, 1	0.7612	30, 3	0.7629
				0	0.5406	47, 1	0.7534	30, 2	0.7270

In Step 1, $m^* = 30$, $n^* = 1$.

REFERENCES

1. Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability (and Bioequivalence Studies)*, 3rd Ed.; Chapman and Hall/CRC, Taylor & Francis: New York, 2008.
2. FDA. *Guidance for Industry on Bioavailability and Bioequivalence Studies for Nasal Aerosols and Nasal Sprays for Local Action*; Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, MD, 1999.
3. FDA. *Guidance for Industry on Bioavailability and Bioequivalence Studies for Nasal Aerosols and Nasal Sprays for Local Action*; Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, MD, 2003.
4. FDA. *Guidance for Industry: Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Considerations*; Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, MD, 2000.
5. Hyslop, T.; Hsuan, F.; Holder, D.J. A small sample confidence interval approach to assess individual bioequivalence. *Stat. Med* **2000**, *19*, 2885–2897.
6. Chow, S.C.; Shao, J.; Wang, H. In vitro bioequivalence testing. *Stat. Med* **2003**, *22*, 55–68.

In Vitro Testing: Dissolution Profile Comparison

Yi Tsong, Pradeep M. Sathe, and Vinod P. Shah

U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.

Abstract

Dissolution test is one of the most valuable in vitro tests used to assure drug product quality. This entry offers theoretical considerations, various approaches, and similarity criterion to aide in the comparison of dissolution profiles.

INTRODUCTION

Dissolution test is one of the most important in vitro tests that is used to assure the drug product quality. A dissolution profile comparison and assurance of similar profile is also an important step in assuring product performance in the presence of a change and in providing biowaivers for lower strength of products from a given manufacturer.

For the assessment of postapproval changes in a drug product such as (1) scale-up, (2) manufacturing site, (3) component and composition, and (4) equipment and process changes, a comparison of dissolution profiles between prechange and postchange products was first recommended in the SUPAC-IR guidance.^[1] This provided a measure of product similarity using dissolution profile characteristics. Later this was also recommended in the other Food and Drug Administration (FDA) Guidances.^[2–4] Changes approved based on a meaningful dissolution data are often accepted as adequate to assure appropriate release and therefore desired in vivo performance. In addition, dissolution profile similarity is also used to provide biowaivers of lower strength of immediate- and modified-release dosage forms as long as the higher strength meets the bioequivalence requirements and the lower strengths show formulation proportionality to the highest strength, and employ the same release mechanism in the case of modified-release dosage forms.^[5]

Dissolution profiles may be considered similar by virtue of 1) overall (or global) profile similarity and 2) uniform similarity at every dissolution sample point.

MEASURES OF DIFFERENCE OR SIMILARITY BETWEEN TWO DISSOLUTION PROFILES

In dissolution profile comparison, the dissolution measurements of both the test and reference batches should be

made under identical conditions and the dissolution sample points for both profiles should be the same, e.g., for immediate release products, 15, 30, 45, and 60 min and for controlled-release products, 1, 3, 5, and 8 hr. Let

μ_{ti}, μ_{ri} = the population percent dissolutions measured at t_i , the i th ($i = 1, \dots, P$) time point of the test and reference products, respectively, with $t_0, \mu_{r(0)}, \mu_{r(0)} = 0$
 $D_i = |\mu_{ti} - \mu_{ri}|$

Several measures of the difference or similarity of two profiles were proposed in the last decade.^[6]

1. Maximum distance

$$D_{\text{Max}} = \text{Max}_{i=1 \text{ to } P} D_i$$

2. Difference of area under the profiles

$$D_{\text{AUC}} = \sum_{i=1}^P [(\mu_{ti} + \mu_{t(i-1)}) - (\mu_{ri} + \mu_{r(i-1)})](t_i - t_{i-1})/2$$

D_{Max} is defined as a measure of the maximum difference of two curves and it is a measure of uniform similarity. D_{AUC} , on the other hand, is an insensitive global measure of the difference between the two mean dissolution profiles (by ignoring the difference between the two mean profiles) at individual time points. Other measurements defined as functions of the difference at each time point are either derived or modified from the following two basic metrics of difference between two mean dissolution curves.

3. Mean distance

$$D^1 = \sum_{i=1}^P |\mu_{ti} - \mu_{ri}|/P$$

The views expressed in this article are those of the authors and do not represent the official position of the U.S. Food and Drug Administration.

A generalized form of the mean distance can be written as

$$D^1 = \sum_{i=1}^P w_i |\mu_{ti} - \mu_{ri}| / P$$

where w_i is the weight for $|\mu_{ti} - \mu_{ri}|$.

The following three measures are modifications of D^1 with different weights.

(3.1) Area between the profiles

$$D_{ABC} = \sum_{i=1}^{p-1} |\mu_{ti} - \mu_{ri}| \left[(t_{i+1} + t_i)/2 - (t_i + t_{i-1})/2 \right]$$

which uses $w_i = [(t_{i+1} + t_i)/2 - (t_i + t_{i-1})/2]P$.

D_{ABC} is insensitive if the point of cross of the two dissolution profiles cannot be identified.

(3.2) Difference factor f_1

$$f_1 = \left\{ \left[\sum_{i=1}^P |\mu_{ti} - \mu_{ri}| \right] / \left[\sum_{i=1}^P \mu_{ri} \right] \right\} \times 100$$

proposed by Moore and Flanner^[7] can be considered as a weighted D_1 with $w_i = 100P/(\sum_{i=1}^P \mu_{ri})$. The difference factor f_1 is a ratio of the absolute difference of the test and reference normalized with respect to the sum of the mean percent dissolutions of the reference lot. This measure is therefore a scaled difference of the two profiles. Because it is normalized with respect to the reference means, the measure is sensitive to the reference product dissolution profile. In other words, with exactly the same size of difference between two profiles, the measure can give different values when the reference profiles are different.

(3.3) First-order Rescigno index

$$D_{R(1)} = \left\{ \left[\sum_{i=1}^P |\mu_{ti} - \mu_{ri}| \right] / \left[\sum_{i=1}^P (\mu_{ti} + \mu_{ri}) \right] \right\}$$

The measure originally proposed by Rescigno^[8] was in general form. The first- and second-order indices are more frequently referenced in the literature. This measure suffers the similar sensitivity as f_1 by normalizing D_1 with respect to $[\sum_{i=1}^P (\mu_{ti} + \mu_{ri})]/P$, the mean of the sum of the two dissolution values.

4. Mean squared distance

$$D^2 = \left[\sum_{i=1}^P (\mu_{ti} - \mu_{ri})^2 / P \right]^{1/2}$$

A few other measures take the general form of a mean squared difference. These are represented as follows:

$$D^{2*} = \left[\sum_{i=1}^P w_i (\mu_{ti} - \mu_{ri})^2 / P \right]^{0.5}$$

where w_i is the weight for $(\mu_{ti} - \mu_{ri})^2$.

With squared terms, these measures are more sensitive to large differences at any time point compared to the measures that use absolute value terms. The second-order Rescigno index and the f_2 measures are two examples of this class.

(4.1) Second-order Rescigno Index^[8]

$$D_{R(2)} = \left\{ \left[\sum_{i=1}^P (\mu_{ti} - \mu_{ri})^2 \right] / \left[\sum_{i=1}^P (\mu_{ti} + \mu_{ri})^2 \right] \right\}^{0.5}$$

uses $w_i = P/[\sum_{i=1}^P (\mu_{ti} + \mu_{ri})^2]$.

$D_{R(2)}$ has the same sensitivity problem as the first-degree Rescigno index.

(4.2) Similarity factor f_2 ^[7]

$$f_2 = 50 \log \left\{ \left[1 + (1/P) \sum_{i=1}^P (\mu_{ti} - \mu_{ri})^2 \right]^{-0.5} \times 100 \right\}$$

As a function of $[1 + (1/P) \sum_{i=1}^P (\mu_{ti} - \mu_{ri})^2]^{-0.5}$, f_2 increases as $\sum_{i=1}^P (\mu_{ti} - \mu_{ri})^2$ decreases. In contrast to D_2 , through mathematical transformation, f_2 measures the similarity of the two profiles. In practice, the measures defined in (1) to (4.2) are estimated by substituting $\bar{x}_{ti}$ and $\bar{x}_{ri}$, the sample means, for μ_{ti} and μ_{ri} , respectively, in the formulae. These estimates are not normalized by the covariance or variance of percent dissolution and do not have a known sampling distribution or asymptotic sampling distribution.

A traditional measure of the difference with covariance normalization may be given (in the multivariate form) as follows.

5. Standardized mean squared distance (or Mahalanobis distance)^[9]

$$D_M = \sqrt{[(\mu_t - \mu_r)' \Sigma_{\text{pooled}}^{-1} (\mu_t - \mu_r)] n/2}$$

where $\Sigma_{\text{pooled}} = (\Sigma_t + \Sigma_r)/2$ is the covariance matrix pooled across both batches, $\mu_t = (\mu_{t1}, \mu_{t2}, \dots, \mu_{tP})$ is the vector of the mean dissolution of the test batch, μ is defined the same way, and Σ_t and Σ_r are the covariance matrices of test and reference batches, respectively.

Note that the dissolution curves of different products are most discriminative at the time points at which dissolution rate accelerates. By assigning weight to the dissolution difference at the sample time when the dissolution is almost complete, the measures may become insensitive to the difference between two dissolution curves. Hence, selecting the weight needs to be carefully considered with respect to the selection of sample points.

SELECTION OF MEASURES

For regulatory considerations, the FDA CDER Immediate Release Drug Dissolution Profile Comparison Working Group evaluated all potential measures and determined to limit further evaluation to mean squared distances D^2 , similarity factor f_2 , and standardized mean squared distance D_M . The maximum distance measure was not considered due to its oversensitivity to failing any similarity requirement based on a single large distance at any sample point. On the other hand, mean distance, difference between the area between profiles, and area between profiles are not sensitive enough to detect large dissolution difference at a particular sample point. Both Rescigno indices and the difference factor f_1 were not considered because they are denominator sensitive. The statistical properties of similarity factor f_2 and standardized mean squared distance D_M will be discussed in the following sections.

DISSOLUTION PROFILE SIMILARITY TESTING BASED ON D_M

D_M is estimated by substituting x_t , x_r , the sample means and S_{pooled} , S_t , S_r for the corresponding parameters in the formula. With the assumption of equal variance–covariance matrices and multivariate normal assumption of x_t and x_r , the statistical properties of the estimate of D_M were well studied in the literature of statistical multivariate analysis.^[10,11] The multivariate confidence region has often been used to describe the variation of estimation of D_M . The decisions of similarity can be made with appropriate comparison of the confidence limits of the metric with a prespecified similarity limit ΔD_M . The 90% confidence limits of D_M can be obtained from the multivariate confidence region of the expected values of x_t and x_r , under multivariate normal assumption.

The $(1 - \alpha)100\%$ confidence region of all differences, y , satisfies the following equation,

$$T^2 = K(y - (x_t - x_r))' S_{\text{pooled}}^{-1} (y - (x_t - x_r)) \leq F_{P, 2n-P-1, 90}$$

where $K = [n^2/(2n)](2n - P - 1)/[(2n - 2)P]$, $F_{P, 2n-P-1, 90}$ is the 90th percentile of F distribution with degrees of freedom $(P, 2n - P - 1)$. It is obvious that this requires $2n$ to be greater than $P + 1$.

The $(1 - \alpha)100\%$ confidence interval (D_M^l , D_M^u) of D_M is obtained using the Lagrange multiplier method^[10] such that

$$D_M^u = \text{Max} \left(\sqrt{y_1^{*'} S_{\text{pooled}}^{-1} y_1^{*} n/2}, \sqrt{y_2^{*'} S_{\text{pooled}}^{-1} y_2^{*} n/2} \right)$$

$$D_M^l = \text{Min} \left(\sqrt{y_1^{*'} S_{\text{pooled}}^{-1} y_1^{*} n/2}, \sqrt{y_2^{*'} S_{\text{pooled}}^{-1} y_2^{*} n/2} \right)$$

where

$$y_1^{*} = (x_t - x_r) \left(1 + \sqrt{F_{P, 2n-P-1, 90} / [K(x_t - x_r)' S_{\text{pooled}}^{-1} (x_t - x_r)]} \right)$$

$$y_2^{*} = (x_t - x_r) \left(1 - \sqrt{F_{P, 2n-P-1, 90} / [K(x_t - x_r)' S_{\text{pooled}}^{-1} (x_t - x_r)]} \right)$$

Dissolution profile similarity is concluded if $D_M^u < \Delta D_M$.

DISSOLUTION PROFILE SIMILARITY TEST BASED ON SIMILARITY FACTOR f_2

Moore and Flanner^[7] proposed the similarity factor f_2 to measure the similarity between two expected dissolution profiles. The simplicity of the similarity factor generated considerable interest and was therefore recommended for dissolution profile comparison in the FDA's Guidance for Industry.^[1-5] Subsequently, many examples of the application of f_2 were published in the literature.^[12,13] These applications were often carried out using point estimate by simply substituting $\bar{x}_{ti}$ and $\bar{x}_{ri}$ for μ_{ti} and μ_{ri} without considering data variation. Most of the criticisms^[14-16] on the use of f_2 were made without realizing that f_2 represents the parameter and is not a statistical testing procedure itself. The statistical properties of f_2 and the similarity testing criteria based on f_2 were, however, delineated later.^[17,18]

Theoretical Considerations

To illustrate the applications of similarity factor f_2 , consider the dissolution profiles of the two batches generated using P number of sample points. Let $(\mu_{r1}, \mu_{r2}, \dots, \mu_{rP})$ represent the cumulative dissolution measures at P time points on the reference profile and $(\mu_{t1}, \mu_{t2}, \dots, \mu_{tP})$ be the corresponding P measures on the test profile. The distances between the two profiles at these P time points are $(|\mu_{r1} - \mu_{t1}|, |\mu_{r2} - \mu_{t2}|, \dots, |\mu_{rP} - \mu_{tP}|)$. The distances at the P time points may be combined into one measure, $D^2 = \sqrt{\{\sum_{i=1}^P [(\mu_{r1} - \mu_{t1})^2 + (\mu_{r2} - \mu_{t2})^2 + \dots + (\mu_{rP} - \mu_{tP})^2]\}}$, which is the square root of the sum of squared differences.

$$f_2 = 50 \log \left\{ \left[1 + (1/P) \sum_{i=1}^P (\mu_{ti} - \mu_{ri})^2 \right]^{-1/2} \times 100 \right\} \\ = 50 \log \{ [1 + (D^2)^2]^{-1/2} \times 100 \}$$

where log is the logarithm based on 10.

If Δf_2 is a required similarity limit for f_2 , then the two dissolution profiles are similar if $f_2 > \Delta f_2$.

Estimation and Confidence Interval of Similarity Factor

Let $(\bar{x}_{r1}, \bar{x}_{r2}, \dots, \bar{x}_{rP})$ and $(\bar{x}_{t1}, \bar{x}_{t2}, \dots, \bar{x}_{tP})$ be the sample mean of n ($= 12$ generally) tablets measured at the P time points of reference and test products, respectively. With these sample means, f_2 is estimated as follows

$$\hat{f}_2 = 50 \log \left\{ \left[1 + (1/P) \sum_{i=1}^P (\bar{x}_{ti} - \bar{x}_{ri})^2 \right]^{-1/2} \times 100 \right\}$$

Assuming that the expected value of $\hat{f}_2$ equals f_2 , i.e., $E(\hat{f}_2) = f_2$, for an assurance of 95% correct decision, similarity of the two profiles can be concluded if the lower limit of the 90% confidence interval of $E(\hat{f}_2) > \Delta f_2$. Both the exact and the asymptotic distribution of $\hat{f}_2$ are complicated to derive. Alternately, the 90% confidence interval can be simulated through the bootstrap method^[19] as given by Shah et al.^[17] and Ma et al.^[18] To make the method easy to use by nonstatisticians, Shah et al. proposed to use $\hat{f}_2$ only when the data variation is small (i.e., with CV < 10% at any point).

Bias of the Estimate of Similarity Factor

Assuming that there are n tablets in both the test and reference batches, consider the expected value of $[(1/P) \sum_{i=1}^P (\bar{x}_{ti} - \bar{x}_{ri})^2]$,

$$\begin{aligned} E \left[(1/P) \sum_{i=1}^P (\bar{x}_{ti} - \bar{x}_{ri})^2 \right] \\ = E \left[(1/P) \left\{ \sum_{i=1}^P [(\bar{x}_{ti} - \bar{x}_{ri}) - (\mu_{ti} - \mu_{ri})]^2 + \sum_{i=1}^P (\mu_{ti} - \mu_{ri})^2 \right\} \right] \\ = (1/P) \left[\sum_{i=1}^P (\mu_{ti} - \mu_{ri})^2 + \sum_{i=1}^P (\sigma_{ti}^2 + \sigma_{ri}^2)/n \right] \end{aligned}$$

where σ_{ti}^2 and σ_{ri}^2 are the variances of percent dissolution measured at the i th time point of the test and reference batches, respectively.

When n becomes large, the expected value of the mean squared differences between sample means $E[(1/P) \sum_{i=1}^P (\bar{x}_{ti} - \bar{x}_{ri})^2]$ becomes very close to $(1/P) [\sum_{i=1}^P (\mu_{ti} - \mu_{ri})^2]$, the mean squared difference between population means. Hence, $\hat{f}_2$ is an asymptotically unbiased estimator of f_2 .

On the other hand,

$$\begin{aligned} E(\hat{f}_2) &= E \left\{ 50 \log \left[\left(1 + (1/P) \sum_{i=1}^P (\bar{x}_{ti} - \bar{x}_{ri})^2 \right)^{-1/2} \times 100 \right] \right\} \\ &= E \left\{ 100 - 25 \log \left[1 + (1/P) \sum_{i=1}^P (\bar{x}_{ti} - \bar{x}_{ri})^2 \right] \right\} \\ &\approx 100 - 25 \log \left\{ 1 + E \left[(1/P) \sum_{i=1}^P (\bar{x}_{ti} - \bar{x}_{ri})^2 \right] \right\} \\ &= 100 - 25 \log \left\{ 1 + (1/P) \left[\sum_{i=1}^P (\mu_{ti} - \mu_{ri})^2 + \sum_{i=1}^P (\sigma_{ti}^2 + \sigma_{ri}^2)/n \right] \right\} \end{aligned}$$

which is

$$\begin{aligned} &< 100 - 25 \log \left\{ 1 + (1/P) \sum_{i=1}^P (\mu_{ti} - \mu_{ri})^2 \right\} \\ &= 50 \log \left\{ \left[1 + (1/P) \sum_{i=1}^P (\mu_{ti} - \mu_{ri})^2 \right]^{-1/2} \times 100 \right\} \\ &= f_2 \end{aligned}$$

This implies that the $\hat{f}_2$ is a conservative estimate of f_2 . This conservative bias makes the use of $\hat{f}_2$ without taking into consideration the variance when the variance is small. The bias was more thoroughly studied and discussed based on simulation results by Ma et al.^[18]

SIMILARITY CRITERION

It is clear that one of the criteria for similarity test is that the dissolution profiles of the prechange batches should be responsible. The similarity limit or region is determined as the tolerable difference in the dissolution profile between two batches of the prechange product. Ideally, such a limit or region is determined empirically, e.g., using the confidence limits of the 99% or 95% confidence interval of individual dissolution profile difference (in terms of the metric to be used) between any two prechange batches. However, often the similarity limit is based on chemist's experience. For example, two profiles can be determined as similar if the average difference at all time points is no more than a given percentage, say, 10% or 15%. The proposed similarity limits of average of 10% difference across all time points was based on an unofficial survey carried out by the FDA Dissolution Profile Comparison Working Group. The simple limit was selected with the objection to select a measure that is reasonably sensitive to a single

large difference. The similarity limit is then transformed to the corresponding value of the measures. For example, with 10% average difference at all time points, the similarity limits for the corresponding difference or similarity measures are

$$\Delta f_2 = 50 \log \left\{ \left[1 + (1/P) \sum_{i=1}^P (10)^2 \right]^{-0.5} \times 100 \right\} = 49.89 \approx 50$$

$$\Delta D_M = \sqrt{[(10)' \Delta_{\text{pooled}}^{-1} (10)]} \approx \sqrt{[(10)' S_{\text{pooled}}^{-1} (10)]}$$

where Δ_{pooled} is the pooled covariance/ $\sqrt{n}$.

Note that both measures are sensitive to the selections of time points. For example, at the time point when the dissolution is not taking place or near saturated, the difference between two profiles is small. Measuring dissolution at such time points will make D_M small and f_2 large. In addition, the parametric confidence region based on D_M is limited to $2n - P - 1 > 0$. The number of time points is therefore limited.

MODELING APPROACH

The dissolution profile is often studied by fitting some empirical physicochemical model to mathematically describe the course of dissolution of a product.^[20–23] for immediate-release drug products, the following sigmoid curves were often found to fit the dissolution data well.

1. Exponential model

$$x(t) = x_{\text{max}} (1 - e^{-at})$$

where x_{max} is the maximum dissolvable quantity.

2. Probit model

$$x(t) = x_{\text{max}} A \Phi(\alpha + \beta A \log(t))$$

where $\Phi(y)$ is the standard normal variable.

3. Gompertz model

$$x(t) = x_{\text{max}} \exp[-\alpha e^{-\beta \log(t)}]$$

4. Logistic curve

$$x(t) = x_{\text{max}} A e^{(\alpha + \beta \log(t))} / \{1 + e^{(\alpha + \beta \log(t))}\}$$

5. Weibull curves

$$x(t) = x_{\text{max}} \{1 - \exp[-\alpha t^\beta]\}$$

where $x(t)$ is percent dissolved at time t divided by 100 and x_{max} is the maximum dissolution.

The parameter x_{max} in the above models is often set to be 100 unless the data clearly indicate otherwise.

Because α in each model determines the dissolution or undissolved proportion at $t = 1$, it is often described as location or scale parameter, whereas because β in each model determines the dissolution rate per unit of time, it is often described as the acceleration or shape parameter. In practical application, when the product has a large lag time t_0 , t in the models needs to be replaced by $t - t_0$.

Sathe et al.^[24,25] proposed the model-dependent approach with the following steps:

1. Establish a mathematical model describing dissolution data using the prechange product. For each tablet sampled from the batch, obtain the parameter values by fitting the selected model to the data.
2. Using the model selected based on a good fitting of the prechange product, determine the parameters of each tablet of the postchange product
3. Find the difference in the means of the parameters of the two products. Compare the upper limit(s) of the 90% confidence intervals of the difference of the means with a prespecified similarity limit for determining similarity.

In comparison to either D_M or f_2 measures, the model-dependent approach is more robust with respect to the choice of sampling points when the number of time points are large. It is recommended for dissolution curves that are sampled at many time points. It is also applicable for data of postchange products collected at different time points. The difficult part of this approach is the determination of the similarity limits.

ALTERNATE APPROACHES

Other than the above three measures recommended in the FDA Guidance, many statistical procedures or measures were proposed in the last decade.^[26–30] Ma et al.^[18,26] thoroughly studied the statistical properties of ΔD^1 (estimate of D^1 by substituting $(\bar{x}_{it}, -\bar{x}_{rt})$ for $(\mu_{it}, -\mu_{rt})$) and concluded that the exact or approximate distribution of the measure is difficult and complicated to derive. In addition, ΔD^1 is also a biased estimate of D^1 with the size of bias often smaller than when using f_2 for f_2 . However, the bias is minimal when the number of tablets is no less than 12. The advantage of using f_2 over D^1 is its ability to pick up the difference between the profiles. It can pick up any large difference between two profiles but it is not as sensitive as D_{max} .

Chow and Ki^[27] proposed a time series procedure to take into consideration the correlation among dissolution values at various time points and the correlation among the baskets used in testing. The procedure tests against a function of the Q value, which is the dissolution specification of the USP. As pointed out in the FDA Guidance, the

Q value is appropriate only for quality control and should not be used for dissolution profile comparison.

Gill^[29] proposed the application of split-plot trend analysis for dissolution profile comparison. The procedure is a hypothesis testing procedure such that by rejecting a null hypothesis of equal parameter one can conclude that the profiles are dissimilar. The concept of "concluding same by showing lack of power to reject a null hypothesis of sameness" has long been faulted in the statistical literature.

CONCLUSION

Dissolution test is one of the most valuable *in vitro* tests used to assure drug product quality. Similar dissolution profile assures product sameness and product performance in the presence of certain Scale-Up and Post Approval Changes. Several statistical approaches were studied by various authors for dissolution profile comparison. Similarity factor f_2 was found to be the simplest and most widely applicable method for this purpose, and is generally recommended for dissolution profile comparison in regulatory guidances. The statistical properties of f_2 and estimation of confidence interval have been delineated.

There is an average 10% difference across all time points in two dissolution profiles when substituted in the equation results in a f_2 value of ~50. f_2 value between 50–100 is considered to indicate product sameness and therefore similar product performance. Dissolution profile comparison is useful in providing biowaivers for lower strength of products from a given manufacturer, and also for *in vivo* studies based on the biopharmaceutics classification system.

REFERENCES

1. The U.S. Food and Drug Administration Guidance. *Immediate Release Solid Oral Dosage Forms: Scale-Up and Postapproval Changes: Chemistry, Manufacturing, and Controls, In Vitro Dissolution Testing, and In Vivo Bioequivalence Documentation (SUPAC-IR)*, Rockville, MD, 1995.
2. The U.S. Food and Drug Administration Guidance. *Modified Release Solid Oral Dosage Forms: Scale-Up and Postapproval Changes: Chemistry, Manufacturing, and Controls, In Vitro Dissolution Testing, and In Vivo Bioequivalence Documentation (SUPAC-MR)*, Rockville, MD, 1997.
3. The U.S. Food and Drug Administration Guidance. *Dissolution Testing of Immediate Release Solid Dose Dosage Forms*, Rockville, MD, 1997.
4. The U.S. Food and Drug Administration Guidance. *Extended Release Oral Dosage Forms: Development, Evaluation and Application of In Vitro/In Vivo Correlations*, Rockville, MD, 1997.
5. The U.S. Food and Drug Administration Guidance. *Bioavailability and Bioequivalence Studies for Orally Administered Drug Products-General Considerations*, Rockville, MD, 2000.
6. Tsong, Y.; Hammerstrom, T.; Sathe, P.; Shah, V.P. Comparing Two Dissolution Data Sets for Similarity. *The 1996 Proceedings of the Biopharmaceutical Section of the American Statistical Association*; ASA: Alexandria, VA, 1997, 129–134.
7. Moore, J.W.; Flanner, H.H. Mathematical comparison of curves with an emphasis on in-vitro dissolution profiles. *Pharm. Technol.* **1996**, 20 (6), 64–74.
8. Rescigno, A. Bioequivalence. *Pharm. Res.* **1992**, 9/7, 925–928.
9. Tsong, Y.; Hammerstrom, T.; Sathe, P.M.; Shah, V.P. Statistical assessment of mean difference between two dissolution data sets. *Drug Inf. J.* **1996**, 30, 1105–1112.
10. Johnson, R.A.; Wichern, D.W. *Applied Multivariate Analysis*; Prentice-Hall, Inc: Englewood Cliff, NJ, 1989.
11. Shah, V.P.; Yamamoto, L.A.; Shuirmann, D.; Elkins, J.; Skelly, J.P. Analysis of in vitro dissolution of whole vs. half controlled release theophylline tablets. *Pharm. Res.* **1987**, 4, 416–419.
12. Polli, J.E.; Rekhi, G.S.; Shah, V.P. Methods to compare dissolution profiles. *Drug Inf. J.* **1996**, 30, 1113–1120.
13. Polli, J.E.; Rekhi, G.S.; Augsburger, L.L.; Shah, V.P. Methods to compare dissolution profiles and a rationale for wide dissolution specifications for metoprolol tartrate tablets. *J. Pharm. Sci.* **1997**, 86, 690–700.
14. Liu, J.P.; Chow, S.C. Statistical issues on the FDA conjugated estrogen tablets bioequivalence guidance. *Drug Inf. J.* **1996**, 30, 881–889.
15. Liu, J.P.; Ma, M.C.; Chow, S.C. Statistical evaluation of similarity factor f_2 as a criterion for assessment of similarity between dissolution profiles. *Drug Inf. J.* **1997**, 31, 1255–1271.
16. Ju, H.L.; Liaw, S. On the assessment of similarity of drug dissolution profiles—a simulation study. *Drug Inf. J.* **1997**, 31, 1273–1289.
17. Shah, V.P.; Tsong, Y.; Sathe, P.; Liu, J.P. In vitro dissolution profile comparison: Statistics and analysis of the similarity factor f_2 . *Pharm. Res.* **1998**, 15, 889–896.
18. Ma, M.C.; Wang, B.B.C.; Liu, J.P.; Tsong, Y. Assessment of similarity between dissolution profiles. *J. Biopharm. Stat.* **2000**, 10 (2), 229–249.
19. Efron, B.; Tibshirans, R.J. *An Introduction to the Bootstrap*; Chapman and Hall: New York, 1993.
20. Langenbucher, F. Linearization of dissolution rate curves by the Weibull distribution. *J. Pharm. Pharmacol.* **1972**, 24, 979–981.
21. Kervinen, L.; Yliruusi, J. Modelling S-shaped dissolution curves. *Int. J. Pharm.* **1993**, 92, 115–122.
22. Tsong, Y.; Hammerstrom, H. Statistical issues in drug quality control based on dissolution testing. *Proc. Biopharm. Sect. Am. Stat. Assoc.* **1994**, 295–300.
23. Tsong, Y.; Hammerstrom, T.; Chen, J.J. Multipoint dissolution specification and acceptance sampling rule based on profile modeling and principal component analyses. *J. Biopharm. Stat.* **1997**, 7 (3), 423–439.
24. Sathe, P.M.; Tsong, Y.; Shah, V.P. In-vitro dissolution profile comparison: statistics and analysis, model dependent approach. *Pharm. Res.* **1996**, 13 (12), 1799–1803.

25. Sathe, P.M.; Tsong, Y.; Shah, V.P. In Vitro Dissolution Profile Comparison and IVIVR: Carbamazepine Case. In *In Vitro–In Vivo Correlation*; Young, D.; Devane, J.D.,; Butler, J., Eds.; Plenum Publishing Corp: New York, 1996.
26. Ma, M.C.; Liu, R.P.; Liu, J.P. Statistical Evaluation of dissolution similarity. *Stat. Sin.* **2000**, *9* (4), 1011–1028.
27. Chow, S.C.; Ki, F.Y.C. Statistical comparison between dissolution profiles of drug products. *J. Biopharm. Stat.* **1997**, *7* (2), 241–258.
28. O'Hara, T.; Dune, A.; Kinahan, A.; Cunningham, S.; Stark, P.; Devane, J. Review of Methodologies for the Comparison of Dissolution Profile Data. In *In Vitro–In Vivo Correlation*; Young, D.,; Devane, J.D.,; Butler, J., Eds.; Plenum Publishing Corp: New York, 1996.
29. Gill, J.L. Repeated measurements: Split-plot trend analysis versus analysis of difference. *Biometrics* **1988**, *44*, 289–297.
30. Leeson, L.J. In vitro/in vivo correlation. *Drug Inf. J.* **1995**, *29*, 903–915.

In Vitro Testing: Micronucleus Test

Wherly P. Hoffman

Department of Biometrics, Elan Pharmaceuticals Inc., South San Francisco, California, U.S.A.

Michael L. Garriott (Retired), and Cindy Lee

Eli Lilly and Company, Indianapolis, Indiana, U.S.A.

Abstract

Genetic toxicology tests are among the early studies conducted to assess the safety profile of a compound. One of the tests currently being examined is the *in vitro* micronucleus (IVM) test. This entry discusses the design, statistical analyses, and interpretation of results of the IVM test.

Keywords: Comparison test; *In vitro* micronucleus test; Trend test.

INTRODUCTION

Genetic toxicology tests are among the early studies conducted to assess the safety profile of a compound. They are designed to determine whether the compound can interact with DNA and lead to the production of gene mutations or chromosomal breakage. Some tests can also detect compounds that interact with components of the mitotic spindle apparatus. Although the tests required for registration of new pharmaceuticals are clearly defined, research investigating modifications to existing tests and development of new tests continues. One of the tests currently being examined is the *in vitro* micronucleus (IVM) test.^[1–4] The design, statistical analyses, and interpretation of results of the IVM will be discussed in this article. The section “Background” includes information pertaining to the relevance of the IVM to the toxicology safety profile. “Design” describes the statistical elements of the design for the test. “Statistical Analyses” discusses two statistical methods for analyzing data from the test. “Results—An Example” presents the results of an example, and “Conclusion” provides concluding remarks to summarize this article.

BACKGROUND

This section will briefly summarize why the IVM is being developed and how the test can be conducted. The procedure is described without great detail and is intended only to provide the reader with an understanding of how the data are obtained. This background will then lead into the subsequent sections on design and statistical methods.

Why the *In Vitro* Micronucleus Test

The rationale for the development of genetic toxicology tests and the reasons they are included in safety assessment are discussed in another chapter—“Ames Test.” Because different genetic toxicology tests measure different endpoints and because no single test is capable of detecting all of the various types of genetic damage that can occur, a battery of tests is conducted to assess safety. For pharmaceuticals, this battery generally consists of: (i) the Ames test, (ii) the *in vitro* chromosome aberration test, and (iii) the *in vivo* micronucleus test.^[5] The Ames test measures gene mutation in bacteria. The chromosome aberration test measures chromosomal damage in Chinese hamster ovary (CHO) cells, Chinese hamster lung (CHL) cells, or human lymphocytes (HL) grown in tissue culture. The micronucleus test is performed in mice or rats and assesses both chromosomal damage and interaction with the mitotic spindle apparatus, as micronuclei are either chromosomal fragments or whole chromosomes. Currently, an IVM test employing the same cells as the chromosome aberration test is often used very early in development as a screen for the more labor-intensive and more expensive chromosome aberration test. Additionally, this test is expected to become an acceptable alternative to the aberration test when submitting the compound for regulatory approval because of the similarity of qualitative results from the two tests.^[1,6,7] Therefore, it is important to consider statistical analyses of the data generated in conducting IVM.

In Vitro Micronucleus Test

As noted earlier, the IVM uses various immortal cell lines (CHO, CHL) or primary cell cultures of HL grown in various tissue culture mediums. There are several protocol

variations that have been examined by different laboratories, including the use of cytochalasin B and the need for an extended treatment in the absence of exogenous metabolic activation. Cytochalasin B blocks cell division and leads to an accumulation of binucleated cells.^[8,9] Micronuclei are then evaluated only in the binucleated cells to ensure that the cells divided following exposure to the test article. However, some researchers question the need for cytochalasin B when using immortal cell lines. An extended treatment in the absence of metabolic activation is used in the *in vitro* chromosome aberration assay to improve its utility for detecting spindle poisons; however, this aspect of the test is controversial and has led to an investigation of the need for this type of treatment in the IVM.^[4] An OECD guideline for the IVM is currently being developed with input from investigators who have evaluated the possible protocol variations and will ultimately determine the procedure to be used in the conduct of the test.^[10]

The procedure we used for the investigation of the IVM at Eli Lilly and Company was similar to that of Curry et al.^[11] and is briefly described as follows. CHO cells were seeded onto eight-well chamber slides 24 hours prior to dosing. Generally, each chemical was tested with eight concentrations that were equally spaced on a log scale, and each concentration was tested on two separate slides. These concentrations were determined from a preliminary cytotoxicity test. The highest concentration included was based on a general rule of 50% cell survival. Treatments were for 4 hours in the absence and in the presence of metabolic activation and for 24 hours in the absence of activation. Because micronuclei may occur spontaneously, a vehicle control slide was always included for each of the three treatments. One thousand (1000) binucleated cells per well were examined and the number of micronucleated, binucleated (MNBN) cells per 2,000 binucleated cells per concentration, including the vehicle control, was recorded. If a compound is not capable of causing chromosomal damage or interfering with the function of the mitotic spindle, then there should be no significant increase in the incidence of micronucleated cells compared to the number of micronucleated cells observed in the vehicle control.

Design

The number of micronuclei in 2,000 binucleated cells is the response to be statistically evaluated. It is the sum of two counts of 1,000 binucleated cells from each duplicate chamber on two separate slides. Three test conditions were included to ensure a thorough examination of each chemical tested: (i) a 4-hour exposure without metabolic activation, (ii) a 24-hour exposure without metabolic activation, and (iii) a 4-hour exposure with metabolic activation. The different exposure conditions were necessary because some chemicals are not directly mutagenic but are metabolized to chemicals that are mutagenic, while other chemicals may delay cell division. A control and up to eight concentration

levels are usually evaluated, and they are typically equally spaced on a log scale. For example, the evaluation of hydrogen peroxide under the third test condition included a control and eight concentration levels 0.16, 0.5, 1.6, 5, 16, 50, 160, and 320 µg/mL. As the survival of the cells will eventually decrease due to cytotoxicity as the concentration increases, it is important to have enough cells alive and available for the development of micronuclei for a proper evaluation of the compound effect. The concentrations should be high enough to allow induction of micronuclei, yet not so high as to kill much more than 50% of the cells. The acceptable lowest cell survival rate is around 45% for inclusion of the concentration levels for statistical analysis. Additional chambers of cells are included solely for obtaining the survival information for each concentration under each of the three testing conditions.

Statistical Analyses

The statistical analysis is conducted for each of the three test conditions separately. The compound is considered to have produced an effect if there is a statistically significant increase in the incidence of micronucleated cells in any of the three test conditions. Each binucleated cell examined is either a micronucleated cell with a fixed probability, p , or it is not, and the cells are independent from each other. The number of micronucleated cells, X , among a fixed number of cells, n , follows a binomial distribution as

$$X \sim B(n, p)$$

Therefore, for the i th concentration, $i = 1, 2, \dots, k$, the number of micronuclei, X_i , out of 2,000 binucleated cells is distributed as

$$X_i \sim B(n_i, p_i), \quad \text{for } i = 1, 2, \dots, k$$

where $n_i = 2,000$, and p_i being a fixed and unknown probability.

The evaluation of the compound effect on increases in the number of micronuclei out of 2,000 binucleated cells can be formulated as follows:

$$H_0 : p_1 = p_2 = \dots = p_k$$

vs.

$$H_a : p_1 \leq p_2 \leq \dots \leq p_k \text{ with at least one inequality.}$$

A linear contrast L_k in the proportions of a control, p_1 , and all concentration levels, $p_2, p_3, \dots, p_k$, can be specified to test the above hypotheses.^[12] Under the null hypothesis of no compound effect, the contrast should be 0, while under the alternative hypothesis of an increasing trend, the contrast should be positive. For example, for $k = 9$, the linear contrast, L_9 , is

$$L_9 = -4p_1 - 3p_2 - 2p_3 - 1p_4 + 0p_5 + 1p_6 + 2p_7 + 3p_8 + 4p_9$$

This test procedure can be carried out using PROC GENMOD in SAS.^[13] The binomial proportions are expressed by counts and total counts in a generalized linear model with a logit link function. The logit function is a one-to-one function that maps a proportion, p , from (0,1) to $(-\infty, \infty)$,

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

The concentration levels are included as a class variable in the model to reflect the log-scaled concentration levels. The fitted model is a saturated one because the number of parameters is the same as the number of observed proportions. The contrasts for increasing trends are defined by replacing the proportions in the above with their corresponding logits. These two formulations are equivalent because of the one-to-one correspondence between the proportion and the logit of the proportion. Contrasts for linear trends are evaluated by the likelihood ratio chi-square test. The resulting p values from the CONTRAST statements in PROC GENMOD are for two-sided tests. To determine the sign of the “slope” of the overall linear trend across all doses, one can obtain the estimate of the trend using the ESTIMATE statement. If the estimate of the trend is positive, then the one-sided p value is half of the p value reported from the CONTRAST statement. If the estimate of the trend is negative, then the one-sided p value is 1 minus half of the p value reported from the CONTRAST statement.

An alternate approach to testing for increasing trends in the proportions is to perform pairwise comparisons to the control. This approach is for testing any increase in the proportion at any concentration level compared to the control,

$$H_0 : p_1 = p_i, \quad i = 2, 3, \dots, k$$

vs.

$$H_a : p_1 < p_i \text{ for at least one } i = 2, 3, \dots, k$$

This pairwise comparison approach can be carried out using the GENMOD procedure as described for the trend test by modifying the contrasts to reflect the difference between the proportions from the i th concentration and the control

$$L_i = -1p_1 + 1p_i, \quad i = 2, 3, \dots, k$$

Similar to the trend test, the resulting p values from the CONTRAST statements in PROC GENMOD are for

two-sided tests. If the proportion of the i th concentration is greater than the proportion of the control, then the one-sided p value is half of the p value reported from the CONTRAST statement. Otherwise, the one-sided p value is 1 minus half of the p value reported from the CONTRAST statement.

For the pairwise comparison test, the type I error rate is inflated if no multiplicity adjustment is made. An adjustment by the Bonferroni approach^[14] can be carried out using PROC MULTTEST in SAS.^[15] The individual p values are multiplied by the number of the comparisons performed on the data. If the product is greater than 1, then the adjusted p value is defined to be 1. Therefore, a significant effect is declared for a compound only if the smallest unadjusted p value is below the predetermined type I error rate divided by the number of the comparisons made. In both the trend test and the pairwise comparison test, a one-sided test with a type I error rate of 0.05 is typically used.

Results—an Example

The results of an IVM test on hydrogen peroxide with cytochalasin B for the three test conditions are presented in Table 1. The numbers of micronucleated cells along with the p values from the trend tests as well as the unadjusted p values from the pairwise comparison tests are reported in Table 1. As six pairs of proportions are compared under the first two test conditions, each unadjusted p value was compared against 0.0083 as one-sixth of the familywise type I error rate of 0.05. The numbers of micronucleated cells are plotted against the ordinal concentration levels in Figs. 1–3.

For the 4- and 24-hour tests without metabolic activation, the concentration levels were from 0 up to 5 $\mu\text{g/mL}$, while the levels were up to 320 $\mu\text{g/mL}$ for the test with metabolic activation. This difference was driven by the criterion of having a survival rate of about 50% in the highest concentration to ensure a valid test design. In other words, to ensure that the cells were exposed to reasonably high concentrations of the chemical. For both the trend test and the pairwise comparison test, a highly significant effect was detected for either a 4- or 24-hour exposure to 5 $\mu\text{g/mL}$ of hydrogen peroxide without metabolic activation. However, no statistical significance was detected under the test condition of a 4-hour exposure with metabolic activation.

CONCLUSION

For validation purposes, 16 chemicals were evaluated by Garriott et al.^[4] using the two statistical approaches for assessing increases in the number of micronucleated cells. These 16 chemicals included non-genotoxins and direct- and indirect-acting genotoxins with various mechanisms

Table 1 Summary of *p* Values from Statistical Analyses on the Numbers of Micronucleated Cells in 2,000 Binucleated Cells Tested in Hydrogen Peroxide with Cytochalasin B

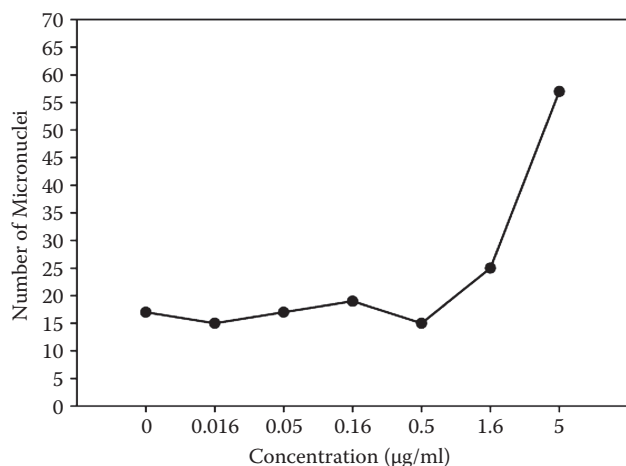
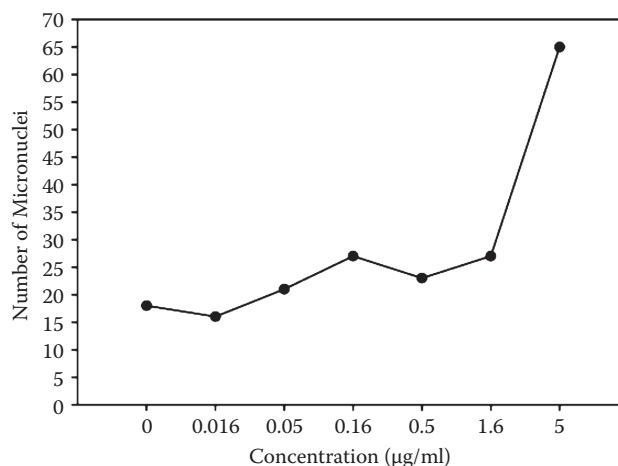
Concentration (µg/mL)	4-hr Exposure without metabolic activation			24-hr Exposure without metabolic activation			4-hr Exposure with metabolic activation		
	Count	Trend test <i>p</i> value	Pairwise comparison unadjusted <i>p</i> value	Count	Trend test <i>p</i> value	Pairwise comparison unadjusted <i>p</i> value	Count	Trend test <i>p</i> value	Pairwise comparison unadjusted <i>p</i> value
0	17			18			21		
0.016	15	NT	0.639	16	NT	0.635	–		
0.05	17	NT	0.500	21	NT	0.315	–		
0.16	19	NT	0.369	27	NT	0.088	12	NT	0.943
0.5	15	NT	0.639	23	NT	0.216	11	NT	0.964
1.6	25	NT	0.107	27	NT	0.088	13	NT	0.917
5	57	<0.001 ^a	<0.001 ^b	65	<0.001 ^a	<0.001 ^b	17	NT	0.743
16							20	NT	0.562
50							23	NT	0.381
160							15	NT	0.843
320							19	0.125	0.625

NT: Not tested.

For the 4- and 24-hour tests without metabolic activation, the concentration levels were from 0 up to 5 µg/mL while for the 4-hour test with metabolic activation, the concentration levels were 0 and 0.16 up to 320 µg/mL.

^aStatistically significant increasing trend detected at the 0.05 level (*p* value ≤ 0.05).

^bStatistically significant increases detected at the 0.05 level (unadjusted *p* value ≤ 0.0083 = 0.05/6).

**Fig. 1** Number of micronuclei out of 2,000 binucleated cells after a 4-hour exposure to hydrogen peroxide without metabolic activation and in the presence of cytochalasin B**Fig. 2** Number of micronuclei out of 2,000 binucleated cells after a 24-hour exposure to hydrogen peroxide without metabolic activation and in the presence of cytochalasin B

of action and different levels of genotoxic potency. There was total agreement between the two statistical tests in declaring statistical significance of the chemicals for all three metabolic activation conditions in the presence of cytochalasin B. In general, for chemicals that cause chromosomal damage and interact with the mitotic spindle apparatus, the number of micronuclei usually increases as the concentration increases. Otherwise, the number of

micronuclei fluctuates around the control range and exhibits no positive trend. As the trend test is more powerful in detecting increasing trends than the pairwise comparison test, it is the natural choice for analysis of data from the IVM test for routine screening of chemicals. However, one can also consider using the trend test as the primary test for detecting monotonically increasing trends while keeping the pairwise comparison test as a backup test for

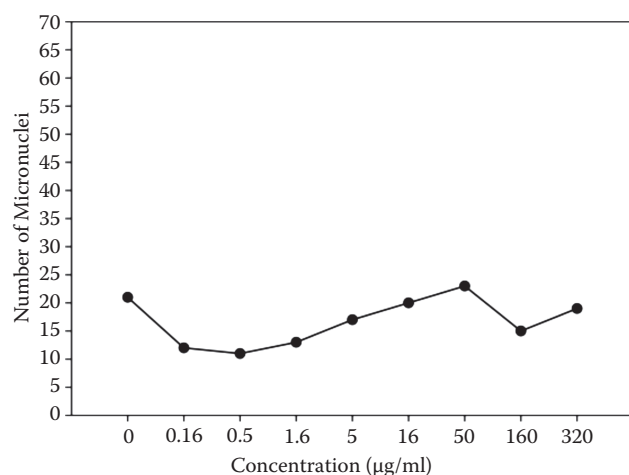


Fig. 3 Number of micronuclei out of 2,000 binucleated cells after a 4-hour exposure to hydrogen peroxide with metabolic activation and in the presence of cytochalasin B

cases when the trend is not significant but there appear to be significant increases at lower concentrations.

Although concurrent controls are always included in the IVM test and also in the evaluation of the effect of the chemical, variation from test to test can be assessed from historical control counts. It will be beneficial for data checking and interpretation of the test results if reference ranges for the control counts are established.

Other statistical analyses for evaluating an increasing trend in the binomial proportions include the Cochran-Armitage trend test^[16,17] and logistic regression modeling with the actual or log concentration levels as a regressor. The trend test carried out as described in this article is reasonably straightforward and resulted in consistent conclusions on most of the chemicals. On occasions when the trend test may appear to be either too sensitive or too insensitive, historical data should be referenced to help interpret the observed results.

REFERENCES

1. Miller, B.; Potter-Locher, F.; Seelbach, A.; Stopper, H.; Utesch, D.; Madle, S. Evaluation of the in vitro micronucleus test as an alternative to the in vitro chromosomal aberration assay: Position of the GUM working group on the in vitro micronucleus test. *Mutat. Res.* **1998**, *410*, 81–116.
2. Kirsch-Volders, M.; Sofuni, T.; Aardema, M.; Albertini, S.; Eastmond, D.; Fenech, M.; Ishidate, Jr., M.; Lorge, E.; Norppa, H.; Surrallés, J.; von der Hude, W.; Wakata, A.

Report from the in vitro micronucleus assay working group. *Environ. Mol. Mutagen.* **2000**, *35*, 167–172.

3. Matsushima, T.; Hayashi, M.; Matsuoka, A.; Ishidate, Jr., M.; Miura, K.; Shimizu, H.; Suzuki, Y.; Morimoto, K.; Ogura, H.; Mure, K.; Koshi, K.; Sofuni, T. Validation of the in vitro micronucleus test in a Chinese hamster lung cell line (CHL/IU). *Mutagenesis* **1999**, *14*, 569–580.
4. Garriott, M.L.; Phelps, J.B.; Hoffman, W.P. A protocol for the in vitro micronucleus test: I. Contributions to the development of a protocol suitable for regulatory submissions from an examination of 16 chemicals with different mechanisms of action and different levels of activity. *Mutat. Res.* **2002**, *517*, 123–134.
5. Muller, L.; Kikuchi, Y.; Probst, G.; Schechtman, L.; Shimada, H.; Sofuni, T.; Tweats, D. ICH-harmonised guidances on genotoxicity testing of pharmaceuticals: Evolution, reasoning and impact. *Mutat. Res.* **1999**, *436*, 195–225.
6. Wakata, A.; Sasaki, M. Measurement of micronuclei by cytokinesis-block method in cultured Chinese hamster cells: Comparison with types and rates of chromosome aberrations. *Mutat. Res.* **1987**, *190*, 51–57.
7. International Conference on Harmonisation. Draft Guidance on S2(R1) Genotoxicity Testing and Data Interpretation for Pharmaceuticals Intended for Human Use. *Fed. Regist.* **2008**, *73*, 16024–16025.
8. Fenech, M.; Morley, A. Measurement of micronuclei in lymphocytes. *Mutat. Res.* **1985**, *147*, 29–36.
9. Krishna, G.; Kropko, M.; Theiss, J. Use of the cytokinesis-block method for the analysis of micronuclei in V79 Chinese hamster lung cells: Results with mitomycin C and cyclophosphamide. *Mutat. Res.* **1989**, *222*, 63–69.
10. Organisation for Economic Cooperation and Development (OECD). Draft Proposal for a New Guideline 487: In Vitro Mammalian Cell Micronucleus Test (MNvit), Version 5, November 2, 2009.
11. Curry, P.; Balwierz, P.; Ivett, J. An in situ micronucleus assay. *Environ. Mol. Mutagen.* **1998**, *31* (Suppl. 29), 3.
12. Neter, J.; Wasserman, W.; Kutner, M.H. Analysis of Factor Effects. *Applied Linear Statistical Models*, 2nd Ed. Richard, D., Ed.; Irwin Inc: Homewood, IL, 1985; 570–572.
13. SAS Institute Inc. *The GENMOD Procedure. SAS/STAT User's Guide, Version 8*; SAS Institute Inc.: Cary, NC, 1999, Vol. 2, 1363–1464.
14. Miller, Jr., R.G. *Simultaneous Statistical Inference*; McGraw-Hill: New York, 1966.
15. SAS Institute Inc. *The MULTTEST Procedure. SAS/STAT User's Guide, Version 8*; SAS Institute Inc.: Cary, NC, 1999, Vol. 2, 2311–2358.
16. Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics* **1955**, *11*, 375–386.
17. Cochran, W.G. Some methods for strengthening the common chi-square tests. *Biometrics* **1954**, *10*, 417–451.

Individual Bioequivalence

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Jen-pei Liu

Division of Biometry, Department of Agronomy, and Institute of Epidemiology, National Taiwan University, Taipei; and Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan

Abstract

This entry outlines the limitation of average bioequivalence for the assessment of drug interchangeability, reviews the decision rules, and statistical methods for evaluation of individual bioequivalence, and provides a comprehensive review of FDA guidelines.

INTRODUCTION

The current concept for the assessment of bioequivalence (BE) is based on the *Fundamental Bioequivalence Assumption* that when two formulations of the same drug product or two drug products (e.g., a brand name drug and its generic copy) are equivalent in the rate and the extent of drug absorption, it is assumed that they will reach the same therapeutic effect or that they are therapeutically equivalent.^[1] Pharmacokinetic (PK) responses, such as area under the plasma or blood concentration–time curve (AUC) and maximum concentration ($C_{\max}$), are usually considered to assess the rate and the extent of drug absorption. The U.S. Food and Drug Administration (FDA) requires that evidence of BE in average bioavailabilities in terms of some primary PK responses such as AUC and $C_{\max}$ between the two formulations of the same drug product or the two drug products be provided.^[2,3] This type of BE is referred to as *average bioequivalence* (ABE).

In the medical community, as more generic drug products become available in the marketplace, it is of great concern whether a number of generic drug products of the same brand name drug can be used safely and interchangeably. Basically, drug interchangeability can be classified as drug prescribability or drug switchability. Drug prescribability is defined as the physician's choice for prescribing an appropriate drug product for the physician's patients between a brand name drug product and a number of generic drug products of the brand name drug product that have been shown to be bioequivalent to the brand name drug product. The underlying assumption of drug prescribability is that the brand name drug product and its generic copies can be used interchangeably in terms of efficacy and safety of the drug product. However, ABE does not take into account the equivalence in variability of bioavailability. A relatively large intra-subject variability

of a test drug product (e.g., a generic drug product) compared to that of the reference drug product (e.g., its brand name drug product) may present a safety concern. To overcome this disadvantage, in addition to providing evidence of ABE, it is recommended that BE in variability of bioavailabilities between drug products be established. This type of BE is called *population bioequivalence* (PBE). In practice, although PBE is often considered for the assessment of drug prescribability, it does not fully address drug switchability because of the possible existence of subject-by-formulation interaction.

Drug switchability is related to the switch from a drug product (e.g., a brand name drug product) to an alternative drug product (e.g., a generic copy of the brand name drug product) within the same subject, whose concentration of the drug product has been titrated to a steady, efficacious, and safe level. As a result, drug switchability is considered more critical than drug prescribability in the study of drug interchangeability for patients who have been on medication for a while. To ensure drug switchability, it is recommended that BE be assessed within individual subjects. This type of BE is known as *individual bioequivalence* (IBE). In "Limitations of Average Bioequivalence," the limitation of ABE for the assessment of drug interchangeability is outlined. The concepts, decision rules, and statistical methods for evaluation of IBE (as described in the 2001 FDA guidance) are given in the section "Individual Bioequivalence." In "Review of the 2001 FDA Guidance," a comprehensive review of the FDA guidances will be provided. A brief summary is given in the "Conclusion."

LIMITATIONS OF AVERAGE BIOEQUIVALENCE

Under ABE, two formulations of the same drug or two drug products are claimed bioequivalent if the ratio of the

means of the primary PK responses between two formulations is within (80%, 125%), with a 90% assurance^[3] It follows that under an ABE criterion, a generic drug product can be used as a substitution for a brand name drug if it has been shown to be bioequivalent to the brand name drug based on average bioavailability. However, the FDA does not indicate that two generic drug products of the same brand name drug can be used interchangeably, even when they are shown to be, on average, bioequivalent to the same brand name drug. In addition, the FDA does not require that the evidence of ABE among generic copies of the same brand name drug be provided. As more generic drugs become available in the marketplace, it is very likely that a patient may switch from one generic drug to another. Therefore an interesting question to physicians and patients is whether the brand name drug and its generic copies can be used safely and interchangeably.

Chen^[4] pointed out that the ABE criterion for the assessment of BE has limitations for addressing drug interchangeability, especially for drug switchability. These limitations include: (i) ABE focuses only on the comparison of population average between the test product and the reference drug product; (ii) the distribution of the primary PK responses of interest is not taken into account, nor is the intrasubject variance of the drug product under study; and (iii) the ABE criterion ignores the subject-by-formulation interaction, which has an impact on drug switchability. As a result, Chen^[4] suggested that the ABE criterion be switched to the IBE approach to overcome these disadvantages.

Many practitioners, however, indicated that the impact of the foregoing limitation on safety might be negligible because little clinical evidence was observed in the past decade to demonstrate that the generic drugs approved based on the ABE criterion were unsafe. In addition, if the limitations of the ABE criterion do have an impact on safety, it is recommended that the decision rule (i.e., the one-fits-all BE limits) of ABE be modified before switching to a totally new, and yet much more complicated, approach of IBE. For example, if the safety of the drug product under study is a great concern, we may lower the upper BE limit from, say, 125–115%. Similarly, we may adjust the lower BE limit to ensure the efficacy. This flexibility not only reflects the nature of the drug product under the study, but also allows a more accurate and reliable assessment of BE between drug products. The adjustment of BE limits of ABE should be flexible according to drug classification, which may depend upon either (i) the intrasubject variability and the therapeutic window of the drug product, or (ii) the population difference ratio (PDR) and/or the individual difference ratio (IDR) as specified in the FDA guidance.^[5]

Chow and Liu^[6] proposed to perform a meta-analysis for an overview of ABE. The proposed meta-analysis provides an assessment of BE among generic copies of an innovative drug that can be used as a tool in monitoring the performance of the approved generic copies of the

innovative drug. In addition, it provides more accurate estimates of intersubject and intrasubject variability of the drug products.

INDIVIDUAL BIOEQUIVALENCE

Individual Bioequivalence Criterion

As indicated earlier, switchability evaluates drug interchangeability within the same patients through the concept of IBE. Individual bioequivalence is originally motivated by the so-called “75/75 rule,” which claims BE if at least 75% of individual subject ratios (i.e., relative individual bioavailability of the generic drug product to the innovative drug product) are within (75%, 125%) limits. Along with this line, Anderson and Hauck^[7] proposed the concept of testing individual equivalence ratio (TIER). The idea is to test IBE based on the dichotomization of continuous PK metrics by calculating the *P* value for at least the observed number of subjects who fall within the BE limit with the minimum proportion of the population in which two drug products must be equivalent in order to claim IBE.

It should be noted that no universal definition of IBE exists that is uniformly accepted by researchers from the regulatory agency, the academia, and the pharmaceutical industry. For example, IBE may be established based on the comparison between distributions within each subject, or it could be based on the distribution of the difference or ratio within each subject.^[8] In addition to average bioavailability, we may also consider the assessment for the variability of bioavailability and the variability due to subject-by-formulation interaction. This disaggregated approach to evaluation of IBE can be assessed by means of intersection–union test (IUT) principle.^[9] Under the disaggregated approach, IBE is concluded if and only if all of the three hypotheses are rejected at the α level of significance. Although the disaggregate approach is a level α test, it can be very conservative.

On the other hand, IBE can also be evaluated using the aggregated approaches that are derived from the distribution of either the difference or the ratio within each subject. Under the aggregated approach, the criteria for assessment of IBE can be classified as probability-based and moment-based criteria (e.g., Refs. [7,10–14]). To address drug switchability, the FDA stressed that the comparisons of interest should be expressed in terms of the ratio of the expected squared difference between the test (T) product and the reference (R) product (administered to the same subject) and the expected squared difference between two administrations of the reference product to the same subject. The square root of this ratio is called the IDR:

$$\begin{aligned} \text{IDR} &= [E(T - R)^2 / E(R - R)^2]^{1/2} \\ &= [(\mu_T - \mu_R)^2 + \sigma_D^2 + (\sigma_{WT}^2 + \sigma_{WR}^2) / 2\sigma_{WR}^2]^{1/2} \end{aligned}$$

where μ_T and μ_R are the averages of the test and reference drug products, respectively; σ_{WT}^2 and σ_{WR}^2 are the intra-subject variability of the test and reference drug products, respectively; and σ_D^2 is the variance as a result of the subject-by-formulation interaction.

Based on the IDR, the U.S. FDA proposed the following aggregated, mixed-scaled, moment-based, one-sided individual bioequivalence criterion (IBC):

$$IBC = \left[(\mu_T - \mu_R)^2 + \sigma_D^2 + 0 \sigma_{WT}^2 + \sigma_{WR}^2 / \max(\sigma_{WR}^2, \sigma_{WO}^2) \right]$$

where σ_{WO}^2 is a constant that can be adjusted to control the probability of passing IBE.

It follows that the hypothesis for the evaluation of IBE can be formulated as:

$$H_0 : IBC \geq \theta_1 \text{ vs. } H_a : IBC < \theta_1$$

where θ_1 is the IBE limit.

When $\sigma_{WR}^2 > \sigma_{WO}^2$, the IBC used is scaled to the variability of the reference formulation and is referred to as the reference-scaled IBC. On the other hand, when $\sigma_{WR}^2 \leq \sigma_{WO}^2$, it is called the constant-scaled IBC. The reasons for the mixed-scaled IBC recommended in the 2001 FDA guidance are twofold. First, in general, the reference-scaled IBC effectively widens the BE limits for more variable reference products. Secondly, for drug products with low variability but a wide therapeutic range, the reference-scaled IBC can unnecessarily narrow the BE limit. In this case, the constant-scaled IBC is used.

When the reference-scaled IBC is used, the relationship between IDR and IBC can be described by the following equation:

$$IDR = (IBC/2 + 1)^{1/2}$$

As a result, assuming that the maximum allowable IDR is 1.25, the 2001 FDA guidance recommends that $\sigma_{WO} = 0.2$. On the other hand, the determination of θ_1 should be based on the consideration of both ABE criterion and variance terms in the IBE criterion as formulated as follows:

$$\theta_1 = \left[(\ln 1.25)^2 + \varepsilon_1 \right] / \sigma_{WR}^2$$

where ε_1 is the variance allowance factor. The U.S. guidance suggests that the allowance for the variance terms $\sigma_{WT}^2 - \sigma_{WR}^2$ and σ_D^2 be 0.02 and 0.03, respectively. The choice of σ_D^2 being 0.03 implies that when $\mu_T - \mu_R = 0$, the probability that individuals would have their average ratio outside ABE limits (0.8, 1.25) is around 80%. Consequently, the 2001 FDA guidance recommends that $\sigma_{WR} = 0.2$ and:

$$\begin{aligned} \theta_1 &= \left[(\ln 1.25)^2 + \varepsilon_1 \right] / \sigma_{WR}^2 \\ &= [0.0498 + 0.05] / 0.04 \\ &= 2.495 \\ &\approx 2.5 \end{aligned}$$

The decision rule recommended by the 2001 FDA guidance is to conclude IBE if and only if the 95% upper confidence bound for IBC is less than 2.5 and the point estimate of the geometric mean ratio of the test drug product to the reference drug product falls within 80–125%.

The 2001 FDA guidance recommends a statistical method proposed by Hylop et al.^[15] for the construction of the 95% confidence bound for the linearized IBC, which is given as:

$$LIBC = (\mu_T - \mu_R)^2 + \sigma_D^2 + (\sigma_{WT}^2 + \sigma_{WR}^2) - 2.5 \times \max(\sigma_{WR}^2, 0.04)$$

and the corresponding hypothesis is reformulated as:

$$H_0 : LIBC \geq 0 \text{ vs. } H_a : LIBC < 0$$

It follows that IBE is concluded if and only if the 95% upper confidence bound for LIBC is less than 0 and the point estimate of the geometric mean ratio of the test drug product to the reference drug product falls within 80–125%.

Statistical Method

The 2001 FDA guidance provides a detailed description of the statistical steps for the construction of 95% confidence bound for LIBC under the two-sequence and four-period (2×4) replicated crossover design (TRTR, RTRT).^[15] It is suggested that the methods of moments and the restricted maximum likelihood method for the estimation of the mean and variance parameters in IBE approach be used. The recommended statistical procedure for the 2×4 replicated crossover design is outlined below (see also Ref. [16]).

Consider the following statistical model that assumes a four-period design with an equal replication of T and R in each of s sequences with an assumption of no or equal carryover effects:

$$Y_{ijkl} = \mu_k + \gamma_{ikl} + \delta_{ijk} + \varepsilon_{ijkl}$$

where $i = 1, \dots, s$ indicates sequence; $j = 1, \dots, n_i$ indicates subject within sequence i ; $k = R, T$ indicates treatment; $l = 1, 2$ indicates replicates on treatment k for subjects within sequence i . Y_{ijkl} is the response of replicate l on treatment k for subject j in sequence i ; γ_{ikl} represents the fixed effect of replicate l on treatment k in sequence i ; δ_{ijk} is the random subject effect for subject j in sequence i on treatment k ; and ε_{ijkl} is the random error for subject j within sequence i on replicate l of treatment k . The ε_{ijkl} s are assumed to be mutually independent and identically distributed as:

$$\varepsilon_{ijkl} \sim N(0, \sigma_{WK}^2)$$

for $i = 1, \dots, s$; $j = 1, \dots, n_i$; $k = R, T$; and $l = 1, 2$. Also, the random subject effects $\delta_{ij} = (\mu_R + \delta_{ijR}, \mu_T + \delta_{ijT})'$ are assumed to be mutually independent and distributed as:

$$\delta_{ij} \sim N_2 \left[\begin{pmatrix} \mu_R \\ \mu_T \end{pmatrix}, \begin{pmatrix} \sigma_{BR}^2 & \rho\sigma_{BT}\sigma_{BR} \\ \rho\sigma_{BT}\sigma_{BR} & \sigma_{BT}^2 \end{pmatrix} \right]$$

The following constraint is applied to the nuisance parameters to avoid an overparameterization of the model for $k = R, T$:

$$\sum_{i=1}^s \sum_{l=1}^2 \gamma_{ikl} = 0$$

In order to estimate the linearized criteria, define:

$$\begin{aligned} I_{ij} &= Y_{ijT} - Y_{ijR} \\ T_{ij} &= Y_{ijT1} - Y_{ijT2} \\ R_{ij} &= Y_{ijR1} - Y_{ijR2} \end{aligned}$$

for $i = 1, \dots, s$ and $j = 1, \dots, n_p$ where:

$$Y_{ijT} = \frac{1}{2}(Y_{ijT1} + Y_{ijT2}) \text{ and } Y_{ijR} = \frac{1}{2}(Y_{ijR1} + Y_{ijR2})$$

Compute the formulation means and the variances of I_{ij} , T_{ij} , and R_{ij} , pooling across sequences, and denote these variance estimates by M_1 , M_T , and M_R , respectively, where:

$$\begin{aligned} \hat{\mu}_k &= 1/s \sum_{l=1}^s \bar{Y}_{ik} \quad (k = R, T) \text{ and } \hat{\Delta} = \hat{\mu}_T - \hat{\mu}_R \\ \bar{Y}_{ik} &= \frac{1}{n} \sum_{j=1}^{n_j} \frac{1}{2} \sum_{l=1}^2 Y_{ijkl} \\ M_1 &= \hat{\sigma}_1^2 = \frac{1}{n_I} \sum_{i=1}^s \sum_{j=1}^{n_j} (I_{ij} - \bar{I}_i)^2 \\ n_1 &= n_T = n_R = \left(\sum_{i=1}^s n_i \right) - s \\ M_T &= \hat{\sigma}_{WT}^2 = \frac{1}{2n_T} \sum_{i=1}^s \sum_{j=1}^{n_j} (T_{ij} - \bar{T}_i)^2 \\ M_R &= \hat{\sigma}_{WR}^2 = \frac{1}{2n_R} \sum_{i=1}^s \sum_{j=1}^{n_j} (R_{ij} - \bar{R}_i)^2 \end{aligned}$$

Then, the linearized criteria can be estimated by:

a. Reference scale:

$$\hat{n}_1 = \hat{\Delta}^2 + M_1 + 0.5M_T - (1.5 + \theta_1)M_R$$

b. Constant scale:

$$\hat{n}_2 = \hat{\Delta}^2 + M_1 + 0.5M_T - 1.5M_R - \theta_1\sigma_{W0}^2$$

and the subject-by-formulation interaction variance component can be estimated by:

$$\hat{\sigma}_D^2 = \hat{\sigma}_1^2 - \frac{1}{2}(\hat{\sigma}_{WT}^2 + \hat{\sigma}_{WR}^2)$$

As a result, a 95% upper confidence bound for the reference-scaled criterion can be obtained as:

$$Hn_1 = \sum E_q + \left(\sum U_q \right)^{1/2}$$

where E_q and U_q are defined and given in the following table:

$H_q = \text{confidence bound}$	$E_q = \text{point estimate}$	$U_q = (H_q - E_q)^2$
$H_D = \left(\left \bar{\Delta} \right + t_{1-\alpha, n-s} \times \left(\frac{1}{S^2} \sum_{j=1}^s n_j^2 M_1 \right)^{1/2} \right)^2$	$E_D = \hat{\Delta}^2$	U_D
$H_1 = \frac{(n-s)M_1}{\chi_{\alpha, n-s}^2}$	$E_1 = M_1$	U_1
$H_T = \frac{0.5(n-s)M_T}{\chi_{\alpha, n-s}^2}$	$E_T = 0.5M_T$	U_T
$H_R = \frac{-(1.5 + \theta_1)(n-s)M_R}{\chi_{1-\alpha, n-s}^2}$	$E_R = -(1.5 + \theta_1)M_R$	U_R
$H_{n1} = \sum E_q + \left(\sum U_q \right)^{1/2}$		

Similarly, a 95% upper confidence bound for the constant scale is given by:

$$H_{n2} = \sum E_q - \theta_1\sigma_{W0}^2 + \left(\sum U_q \right)^{1/2}$$

where E_q and U_q are specified in the following table:

$H_q = \text{confidence bound}$	$E_q = \text{point estimate}$	$U_q = (H_q - E_q)^2$
$H_D = \left(\left \bar{\Delta} \right + t_{1-\alpha, n-s} \times \left(\frac{1}{S^2} \sum_{j=1}^s n_j^2 M_1 \right)^{1/2} \right)^2$	$E_D = \hat{\Delta}^2$	U_D
$H_1 = \frac{(n-s)M_1}{\chi_{\alpha, n-s}^2}$	$E_1 = M_1$	U_1
$H_T = \frac{0.5(n-s)M_T}{\chi_{\alpha, n-s}^2}$	$E_T = 0.5M_T$	U_T
$H_R = \frac{-(1.5)(n-s)M_R}{\chi_{1-\alpha, n-s}^2}$	$E_R = -(1.5)M_R$	U_R
$H_{n2} = \sum E_q - \theta_1\sigma_{W0}^2 + \left(\sum U_q \right)^{1/2}$		

REVIEW OF THE 2001 FDA GUIDANCE

In October 2002, the U.S. FDA issued a draft guidance on *Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Considerations*.^[3] This guidance presents a unified approach to meet the requirements set forth in 21 CFR Part 30 as applied to dosage forms intended for oral administration. As a result, this guidance is to replace a number of previously issued FDA drug-specific BE guidances, including the guidance on statistical approaches issued in July 1992.^[2] In addition, in January 2001, the U.S. FDA issued a guidance on *Statistical Approaches to Establishing Bioequivalence*.^[5] This guidance provides the BE criteria and their corresponding statistical procedures for ABE, PBE, and IBE as present in the guidance for general considerations. In what follows, we provide a comprehensive review of the FDA guidance on *Statistical Approach to Assessing Bioequivalence* from both scientific/statistical and practical points of view (see also Ref. [17]).

Aggregated Criteria Vs. Disaggregated Criteria

The 2001 FDA guidance recommends aggregated criteria as described earlier for the assessment of IBE. In addition to the average of bioavailability, the IBE criterion takes into account the variability of bioavailability, as well as the variability because of the subject-by-formulation interaction. Although the FDA requires that the point estimate of the geometric mean ratio between test product and reference product be within 80–125% when using the IBE approach under the proposed aggregated criteria, it is not clear whether the IBE criterion is superior to the ABE criterion in the assessment of drug exchangeability. In other words, it is not known whether or not IBE implies ABE under aggregated criteria. Hence, the question of particular interest to pharmaceutical scientists is whether the proposed IBE criterion can really address the drug-to-drug interchangeability (i.e., drug switchability).

Liu and Chow^[8] suggested that disaggregated criteria be implemented for the assessment of drug interchangeability. In addition to ABE, the disaggregated criteria may consider the following hypothesis testing for assessing BE in the variability of bioavailability and variability due to the subject-by-formulation interaction:

$$H_0 : \sigma_{WT}^2 / \sigma_{WR}^2 \geq \Delta_v \text{ vs. } H_0 : \sigma_{WT}^2 / \sigma_{WR}^2 < \Delta_v$$

and

$$H_0 : \sigma_D^2 \geq \Delta_s \text{ vs. } H_a : \sigma_D^2 < \Delta_s$$

where Δ_v is the BE limit for the ratio of intrasubject variability and Δ_s is an acceptable limit for variability due to subject-by-formulation interaction. We conclude IBE if and only if the 90% confidence interval for μ_T/μ_R is within

(80%, 125%), and both the 95% upper confidence limit for $\sigma_{WT}^2/\sigma_{WR}^2$ is less than Δ_v and the 95% upper confidence limit for σ_D^2 is less than Δ_s . Under the disaggregated criteria, it is clear that IBE implies ABE.

In practice, it is then of interest to examine the relative merits and disadvantages of the FDA-proposed aggregated criteria and disaggregated criteria described early for the assessment of drug interchangeability. In addition, it is also of interest to compare the aggregated and disaggregated criteria of PBE and IBE with the conventional ABE criterion in terms of the consistencies and inconsistencies in concluding BE for regulatory approval.

Interpretation of the Criteria

The guidance suggests that a logarithmic transformation be performed for all BE measures (e.g., AUC and C_{max}). For ABE, the interpretation on both the original and the log scale is straightforward and easily understood, because the difference in arithmetic means on the log scale is the ratio of geometric means on the original scale. For log transformation, two drug products are concluded ABE if the ratio of the average bioavailabilities on the original scale is between 80% and 125%. This statement is the same as saying that the two drug products are concluded ABE if the difference between the arithmetic means on the log scale is between $-0.2231 (= \ln(0.8))$ and $0.2231 (= \ln(1.25))$. On the other hand, the interpretation of the aggregate criteria for IBE is not straightforward and clear on both scales. For example, the exponentials of the subject-by-formulation interaction and intrasubject variability on the log scale do not correspond to the subject-by-formulation interaction and intrasubject variability on the original scale. In addition, more detailed explanations and sufficient reasons should be given as to why $\sigma_{WT}^2 - \sigma_{WR}^2 = 0.02$ and $\sigma_D^2 = 0.03$ are chosen for θ_1 and their interpretation on the original scale.

Masking Effect

The goal for evaluating BE is to assess the similarity of the distributions of the PK measures obtained either from the population or from individuals in the population. However, under aggregate criteria, different combinations of values for the components of the aggregate criterion can yield the same value. In other words, BE can be reached by two totally different distributions of PK measures. At the 1996 Advisory Committee meeting, it was reported that the data sets from the FDA's files show that a 14% increase in average (ABE allows only 80–125%) is offset by a 48% reduction in the variability, and the test passes IBE but fails ABE. The current requirement of the 2001 FDA guidance on IBE cannot alleviate the above situation because it only requires that the point estimates of the geometric test/reference mean be within 80–125% in addition to meeting the 95% confidence bound for IBE being less than 2.5.

Power and Sample Size Determination

For the proposed aggregated criterion, it is desirable to have sufficient statistical power to declare IBE if the value of the aggregated criterion is small. On the other hand, we would not want to declare IBE if the value is large. In other words, a desirable property for the assessment of BE is that the power function of the statistical procedure is monotone decreasing function. However, because different combinations of values of the components in the aggregated criteria may reach the same value, the power function for any statistical procedures based on the aggregated criteria is not a monotone decreasing function. Another major concern is how the aggregated IBC will affect the sample size determination based on the power analysis. Unlike ABE, the relationship between the power and the parameters are complicated and there exists no closed form for the power function of the statistical procedure proposed by Hylop et al.^[15] for IBE. As a result, the 2001 FDA guidance states that the sample size for individual BE should be based on simulated data.

Chow et al.^[16] indicated that for a given set of parameter $(\Delta, \rho, \sigma_{BT}, \sigma_{BR}, \sigma_{WT}, \sigma_{WR})$, under the assumption that:

$$\eta = \Delta^2 + \sigma_{BT}^2 + \sigma_{BR}^2 - 2\rho\sigma_{BT}\sigma_{BR} + \sigma_{WT}^2 - \sigma_{WR}^2 - \theta_{IBE} \max\{\sigma_0^2, \sigma_{WR}^2\} < 0$$

The n (the sample size per sequence) needed to achieve a given power p can be estimated by the first integer (n) satisfying the following inequality:

$$\eta + \sqrt{U_p} + \sqrt{U_{0.95}} < 0$$

where $U_p = U_D + U_1 + U_T + U_R$. More details regarding sample size calculation for IBE can be found in Chow and Shao^[18] and Chow et al.^[16]

Note that different combinations of $\mu_T - \mu_R$, σ_D , σ_{WT} , and σ_{WR} could lead to the same value of LIBC and yet the sample size required for achieving the desired power may be different. For example, combinations of $\mu_T - \mu_R$, σ_D , σ_{WT} and $\sigma_{WR} = (0.2, 0, 0.2, 0.2)$, $(0, 0.2, 0.15, 0.15)$, and $(0, 0.2, 0.2, 0.2)$ will give the same value of LIBC (i.e., -0.0598) but different sample sizes per sequence 52, 20, and 59, respectively, are required.^[16]

Two-Stage Test Procedure

To apply the IBC for the assessment of IBE, the 2001 FDA guidance suggests that the constant scale be used if the observed estimator of σ_{WR} is smaller than 0.2. However, statistically, the observed estimator of σ_{WR} being smaller than 0.2 does not mean that σ_{WR} is smaller than 0.2. A test on the null hypothesis that σ_{WR} is smaller than 0.2 should be performed. As a result, the proposed statistical procedure for the assessment of IBE becomes a two-stage test

procedure. It is then recommended that the overall type I error rate and calculation of power be adjusted accordingly.

Study Design

The 2001 FDA guidance recommends that two replicated designs, i.e., (ITRTR, RTRT) and (TRT, RTR), be used for the assessment of IBE. It is not clear whether the two replicated crossover designs are the optimal design in terms of power or variance in 2×4 and 2×3 replicated crossover designs with respect to the aggregated criteria. In addition, the 2001 FDA guidance indicates that the 2×4 (TRTR, RTRT) design be recommended, while the 2×3 (TRT, RTR) design is preferred as an alternative. Several questions are raised. First, it is not clear what the relative efficiency of the two designs is if the total number of observation is to be fixed? Secondly, it is not clear how these two designs compare to other 2×4 and 2×3 replicated designs such as (TRRT, RTRT) and (TTRR, RRTT) designs, and (TRR, RRT) and (TTR, RRT) designs. Finally, it may be of interest to study the relative merits and disadvantages of these two designs as compared to other designs, such as Latin square designs and four-sequence and four period designs.

Other issues regarding the proposed replicated designs include the following: (i) it will take a longer time to complete; (ii) the subject's compliance may be a concern; (iii) it is likely to have a higher dropout rate and missing values, especially in 2×4 designs; and (iv) there is little in the literature on statistical methods that deals with dropouts and missing values in a replicated crossover design setting.

Outlier Detection

No formal statistical procedure for the detection of outliers for the standard 2×2 crossover designs or for the 2×3 and 2×4 replicated crossover design has been suggested in the 2001 FDA guidance. For a valid statistical assessment, the procedures proposed by Chow and Tse^[19] and Liu and Weng^[20] should be used. These proposed statistical procedures for outlier detection in BE studies were derived under crossover designs that incorporate the correlations with the same subject. The guidance provides little or no discussion regarding the treatment of identified outliers.

CONCLUSION

The rationale for switching from ABE to IBE lacks (i) convincing clinical evidence regarding the limitations of ABE for addressing drug interchangeability, and (ii) a valid scientific/statistical justification for IBE. It is then recommended that clinical evidence for the limitations of ABE and scientific/statistical justification of IBE be carefully evaluated before the IBE criteria and corresponding statistical procedure be implemented for the replacement of ABE as the standard requirement for the assessment of BE.

To provide convincing clinical evidence regarding the limitations of ABE for addressing drug interchangeability, it would be helpful to study the incidence rates concerning the failure or critical safety issue of approved generic drug products in the past decade. A high incidence rate may indicate that there is a deficiency in the current regulation of ABE. If the incidence is relatively low and is considered clinically acceptable, then there is no need to switch from ABE to IBE, especially when clinical performance is unknown. On the other hand, to provide scientific/statistical justification for the use of IBE, it is suggested that the probabilities of consistencies and inconsistencies for BE assessment of ABE and IBE be evaluated.^[8] If the probability of consistency between ABE and IBE is high and the probability of inconsistency is low, then there is no need to switch from ABE to IBE because ABE and IBE will basically reach the same conclusion regarding BE. If there is a relatively high probability of inconsistency between ABE and IBE with a relatively low probability of consistency between ABE and IBE, then the impact on clinical performance for the switch from ABE to IBE is necessarily provided. In this case, instead of switching from ABE to IBE, we might tighten the BE limits for ABE to achieve the same result as IBE.

In addition to efficacy and safety, the identity, strength, purity, and stability of approved generic drug products are important drug characteristics that have an impact on average bioavailability, variability of bioavailability, and variability due to subject-by-formulation interaction and consequently drug interchangeability. As a result, it is suggested that the regulatory requirements for post-approval equivalence in the laboratory development and the manufacturing process be established regardless of the switch from ABE to IBE.

More information regarding the relative merits and disadvantages of ABE and IBE can be found in a special issue of *Clinical Pharmacology Therapy and Toxicology on Bioequivalence Assessment: Methods and Applications*,^[21] a special issue of *Drug Information Journal on Bioavailability and Bioequivalence*,^[22] the proceedings of FIP Bio-International '96 Bioavailability, Bioequivalence and Pharmacokinetics, a special issue of *Journal of Bio-pharmaceutical Statistics*,^[23] a special issue of *Statistics in Medicine in individual Bioequivalence*,^[14] and a reference book on statistical design and analyses for the evaluation of BE.^[1]

REFERENCES

1. Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*, 3rd Ed.; Chapman and Hall/CRC, Taylor & Francis: New York, 2009.
2. FDA. *Draft Guidance on Statistical Procedures for Bioequivalence Studies Using a Standard Two-Treatment Crossover Design*; The U.S. Food and Drug Administration: Rockville, MD, 1992.
3. FDA. *Guidance for Industry: Bioavailability and Bioequivalence Studies for Orally Administered Drug Products-General Considerations*; The U.S. Food and Drug Administration: Rockville, MD, 2002.
4. Chen, M.L. Individual bioequivalence—a regulatory update. *J. Biopharm. Stat.* **1997**, *5*, 5–11.
5. FDA. *Guidance for Industry: Statistical Approaches to Establishing Bioequivalence*; The U.S. Food and Drug Administration: Rockville, MD, 2001.
6. Chow, S.C.; Liu, J.P. Meta-analysis for bioequivalence review. *J. Biopharm. Stat.* **1997**, *7*, 97–111.
7. Anderson, S.; Hauck, W.W. Consideration of individual bioequivalence. *J. Pharmacokinet. Biopharm.* **1990**, *18*, 259–273.
8. Liu, J.P.; Chow, S.C. Some thoughts on individual bioequivalence. *J. Biopharm. Stat.* **1997**, *7*, 41–48.
9. Berger, R.L. Multiparametric hypothesis testing and acceptance sampling. *Technometrics*. **1982**, *24*, 295–300.
10. Esinhart, J.D.; Chinchilli, V.M. Extension to the use of tolerance intervals for assessment of individual bioequivalence. *J. Biopharm. Stat.* **1994**, *4*, 39–52.
11. Sheiner, L.B. Bioequivalence revisited. *Stat. Med.* **1992**, *11*, 1777–1788.
12. Schall, R.; Luus, R.E. On population and individual bioequivalence. *Stat. Med.* **1993**, *12*, 1109–1124.
13. Holder, D.J.; Hsuan, F. Moment-based criteria for determining bioequivalence. *Biometrika*. **1993**, *80*, 835–846.
14. Chow, S.C.; Liu, J.P. Individual bioequivalence. *Stat. Med. (Spec. Issue)*. **2000**, *19* (20).
15. Hylop, T.; Hsuan, F.; Holder, D.J. A small-sample confidence interval approach to assess individual bioequivalence. *Stat. Med.* **2000**, *19*, 2885–2897.
16. Chow, S.C.; Shao, J.; Wang, H. Individual bioequivalence testing under 2×3 crossover designs. *Stat. Med.* **2002**, *21*, 629–648.
17. Chow, S.C. Individual bioequivalence—a review of FDA draft guidance. *Drug Inf. J.* **1999**, *33*, 435–444.
18. Chow, S.C.; Shao, J. *Statistics in Drug Research—Methodology and Recent Development*; Marcel Dekker, Inc.: New York, 2002.
19. Chow, S.C.; Tse, S.K. Outliers detection in bioavailability/bioequivalence. *Stat. Med.* **1990**, *9*, 549–558.
20. Liu, J.P.; Weng, C.S. Detection of outlying data in bioavailability/bioequivalence studies. *Stat. Med.* **1992**, *10*, 1375–1389.
21. Steinijans, V.W.; Schulz, H.-U. Bioequivalence assessment: methods and applications. *Clin. Pharmacol Ther. Toxicol. (Spec. Issue)* **1992**, *30* (Suppl. I).
22. Chow, S.C. Bioavailability and bioequivalence. *Drug Inf. J. (Spec. Issue)* **1995**, *29* (3), 793–1067.
23. Chow, S.C. Bioavailability and bioequivalence. *J. Biopharm. Stat. (Spec. Issue)* **1997**, *7* (1), 1–204.

Encyclopedia of Biopharmaceutical Statistics, 4th Edition

Volume III

Instrument–Population Bioequivalence
Pages 1167—1746

Instrument–
Kaplan–Meier

Kappa–Logistic

Logistic–
Meta-Analysis

Microarray–
Multicollinearity

Multidimensional–
Multi-Regional

Multi-Regional–Onset

Optimal–
Pharmaceutical

Pharmacodynamic–
Population



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Instrument Development and Validation

Cindy Rodenberg, James T. Kuznicki, and Gloria H. Yiu

The Procter & Gamble Co., Cincinnati, Ohio, U.S.A.

Abstract

This entry provides an overview of both qualitative and quantitative approaches to the development and validation of psychometric instruments.

Keywords: Development; Psychometric instrument; Validation.

INTRODUCTION

A psychometric instrument can provide a standardized and objective means of collecting data on subjective states or events. Literally thousands of such instruments exist, ranging from ad hoc instruments composed by investigators for one time use up to rigorously developed and validated instruments used over decades for research and evaluation.^[1–5] Despite the large number, variety, and widespread availability of these instruments, investigators frequently face the need to develop new ones. This need arises because a proper development and validation of the existing instruments was achieved for a specific purpose and relative to a defined population. While the existing pool of instruments will often contain one that has been developed in the population of interest for the purpose desired, new research questions will just as often require new instruments for measurement. Development for a defined purpose and in a defined population is the only way to ensure a proper sampling and a valid measurement of the content of the subjective state, behavior, or disease to be measured.

OVERVIEW

This article provides an overview of both qualitative and quantitative approaches to the development and validation of psychometric instruments.

The approach to inventory development discussed in the first edition of this article had three major steps. The first step was the patient generation of items or questions for the inventory and their arrangement into groups of apparently related items, called domains, for potential inclusion in the final inventory. The second step was the qualitative content validation of the items and domains. The third step was the quantitative psychometric item analysis, which was used to select an optimum subset

of items and to determine domain structure. Methods to assess the developed instrument's reliability and validity were additionally discussed.

Two major additions have been made. The first is the use of Confirmatory Factor Analysis (CFA) to assess how well a hypothesized factor structure fits a new set of data. This new set of data may be a new sample from the same population, in order to confirm the identified factor structure during the exploratory phase, or may represent a different population in order to assess whether the factor structure fits this new population.

Additionally, a major advance in the area of measurement development that will also be discussed in this update is Item Response Theory (IRT). The first edition to this article was restricted to classical test theory methods while this updated version will expand into IRT methods.

Where appropriate, examples will be provided from the development of the Profile of Female Sexual Function© (PFSF), a multinational, patient-generated inventory for the measurement of hypoactive sexual desire (hypoactive sexual desire disorder, HSDD; see DSM-IV, Ref. [6]), or low desire, in menopausal women. This instrument was developed and validated among menopausal women having HSDD for the purpose of measuring the response to therapy in clinical trials of novel therapeutic agents. Although a large number of sexual function measurement tools exist,^[2] before the present work was done, no properly developed instrument for HSDD was in existence. Since the first edition, articles detailing the development and validation of the PFSF have been published.^[7,8]

The discussion on instrument development in this article focuses on the qualitative and quantitative aspects of questionnaire development, validation and refinement. Discussions of other important topics in instrument development, such as scale selection, recall period, response set, sampling, and related issues, can be found in many of the references cited here.

ITEM GENERATION

The objective of item generation is to establish a preliminary inventory that has face and content validity. This process must ensure the proper identification of the characteristics of the disease to be measured and an adequate sampling of their content to allow a valid measurement.

Face validity is that characteristic of an item which, “on the face of it,” appears to be relevant to the outcome it intends to capture. For example, if sexual desire is being measured, then items should look like they are addressing the characteristics of sexual desire and not some other concept, such as marital happiness. In most cases, this is a necessary characteristic for a valid instrument. However, there are instances where this characteristic may not be desired or needed. For example, instruments ascertaining a sensitive subject (e.g., substance abuse) may try to elicit responses in a more indirect manner if it is felt that a subject will not answer direct questions truthfully.^[9]

Content validity refers to the instrument property of a proper sampling of items, presentation, and the measurement of all aspects of the disease or other states relevant to the patient. Items should reflect the medical diagnosis of the disease as well as patients’ thoughts, feelings, and emotions as affected by the disease. This is generally a qualitative process that speaks to the sampling and the presentation of the items in an inventory.^[1,9] A detailed discussion of item generation and domain identification can be found in Nunnally and Bernstein,^[9] Juniper et al.,^[10] and McHorney,^[11] among others. Although this is generally addressed qualitatively, a quantitative method referenced in Streiner and Norman 4th edition^[12] is the content validity index (CVI). Here a group of experts rate the relevance of an item to a domain on a scale of 1, totally irrelevant, to 4, extremely relevant. CVI is the proportion of experts who rate the item a 3 or 4. Guidelines for acceptance of an item are given in Lynn.^[13] See Haynes et al.^[14] for a comprehensive guide for determining content validity.

There are numerous methods for item development. Two publicly available item pools are the International Personality Item Pool (IPIP) (<http://ipip.ori.org>) and PROMIS (Patient-Reported Outcomes Measurement Information System). [15–17] IPIP provides both items and scales measuring personality traits such as assertiveness, anger, and compassion. PROMIS is an initiative of the NIH to develop item banks to measure patient reported outcomes in pain, fatigue, emotional distress, physical functioning, social role participation, and general health perception.

Other methods of item generation include the development based on literature review and an understanding of the disease, the discussions among experts or a direct development of items by experts, the factor analytic methods applied to data from previously existing instruments, and the direct development by patients. The patient identification of domains and the generation of inventory items have numerous advantages.^[11,18] Patient-generated

items are: (i) directly reflective of patients’ experiences and perspectives and therefore have a high degree of face validity; (ii) more likely to be acceptable to patients in terms of their content and reading level; and (iii) more likely to ensure that investigators and patients have a single and common understanding of the item and domain meanings.

The use of literature and professional input alone is often not sufficient and may suffer from the phenomenon exemplified by the game of “telephone” in which the meaning of a message changes as it is passed along from person to person. Thus the subtle modifications that occur in patient reports when relayed from patient to physician to expert to literature text might inadvertently, but importantly, alter the original meaning. This problem is reduced if *verbatim* statements from patients can be adapted to inventory items.

The patient generation of items through focus groups, followed by individual patient interviews to establish item content and to clarify meaning, is increasingly becoming the standard practice.^[11] In both types of interviews, a professional interviewer leads the patient or patients through a detailed discussion of their experiences with the disease and/or their reaction to items. The interview is constructed to generate discussion that can be gleaned for potential items. These interviews should avoid any evaluation. Evaluation during patient interviews can inhibit discussion, especially of personal topics such as sexuality, and can therefore allow important, but sensitive, aspects of a disease to go undiscovered. Constituting groups with patients who have the desired characteristics of the instrument’s target population in common, but who are unfamiliar with one another, will also facilitate an open discussion.^[11] Transcripts of the interview are later reviewed and words and phrases that patients used to describe their experiences are identified for potential use as inventory items.

Patients typically provide a wealth of information concerning their disease or psychological state. For example, in the development of the PFSF, over 450 potential items were generated. Common statements heard from women suffering from hypoactive sexual desire were “I wanted to avoid sex” and “I felt sexually unattractive.” In addition, the amount of time spent on a certain set of symptoms by patients, or the emotion or degree of animation in the discussion can provide valuable clues as to the importance of various symptoms, feelings, attitudes, and emotions. A discussion of the degree to which a symptom or feeling has changed after the emergence of the disease can provide an important indication of its likely response to therapeutic interventions in the future.

Items should be reviewed for language level and items with an unusually high level should be eliminated or revised to an easier reading level; a common rule is that of a 6th grader or 12-year-old. Colloquialisms and slang terms should be eliminated unless absolutely needed for understanding.^[19] Compound items should be split into unitary items. For example, a compound item such as “I felt no

desire for sexual activity and could not become aroused” should be made into two unitary items “I felt no desire for sexual activity” and “I could not become aroused.” It may be beneficial to include items that express the same meaning with different terms to determine which will work best among the patients. A web-based tool, QUAID (<http://www.psyg.memphic.edu/quaid/html>) was developed by Graesser et al.^[20] to identify types of problems such as vague terms, complex syntax, or items with too high memory load.

Items that are extracted from the interviews are grouped together on the basis of content similarity in preliminary domains that are the precursors of the ultimate subscales of the final instrument. The domains should reflect any generally accepted medical criteria of the disease, as well as include concepts identified by patients even if these concepts are not generally recognized as important at the current time by medical or psychological science. Items that express different levels of the same feeling or symptom should also be included to provide a sufficient gradation along a continuum if the instrument is intended to measure response to a treatment. Foreexample, the three items “I was uninterested in sexual activity,” “I felt sexual desire,” and “I felt a high degree of sexual desire” represent a continuum of sexual desire. If each is measured using a six-point rating scale, a total score (after reversing item scoring for the negatively worded item) of 18 points represents a continuum along which a patient can express changes in sexual desire. IRT methods are discussed later but provide a more sophisticated method to assess “item difficulty”; a characteristic where more of a trait is needed in order for a person to respond in a positive direction. IRT models are being used to develop shortened instruments by selecting a set of questions whose “item difficulty” span the range of the underlying trait but are not redundant with one another.

In the development of the PFSF, the initial review was able to classify the 450 potential items obtained for the PFSF into 12 preliminary domains that seemed to be measuring distinct underlying constructs. Optimally, each domain will include enough items to result in the potential for reliable measurement of the underlying construct after item reduction by statistical techniques is completed (see the section “Quantitative Instrument Development”), but few enough

to avoid fatigue by those completing the inventory. The final review of the initial 450 items generated for the sexual function instrument resulted in 83 items arranged in seven domains, each containing from 5 to 19 items.

QUALITATIVE CONTENT VALIDATION

The objectives of the qualitative content validation are: to maximize the degree to which the items adequately represent the patients’ symptoms and experiences; to produce items that have clear, unitary, and unambiguous meanings; and, in the case of multilanguage inventories, to maintain similar meaning across translations. The Cognitive Interview Technique,^[21,22] or Cognitive Laboratory, is commonly used for qualitative validation. A cognitive interview is a one-on-one interview between a trained professional interviewer and a patient. The interviewer reviews the thought processes the patient went through in determining what the patient’s answer to each inventory item should be. In so doing, the patient reveals what the items mean to one’s self. Unlike item generation, the process is evaluative in nature and results in the elimination and/or the refinement of items.

In practice, a moderator asks the subject to read the instructions on the questionnaire and then to fill in responses. The subject is additionally requested to mark any words, phrases, or sections that were confusing or difficult to answer. The moderator then walks the subject through the questionnaire sentence by sentence and the subject describes one’s thought processes and understanding as the instructions are read and as the items are answered.

The cognitive interview process is an iterative one. Modifications resulting from earlier interviews are reviewed in subsequent interviews. The interviews identify items with multiple or ambiguous meanings, or meanings different from what the investigator intends. If the instrument is translated into languages beyond the original one, cognitive interviews serve to identify and to correct altered meanings resulting from the translation process. Such difficulties will occur even if items are carefully screened to avoid slang, colloquialisms, and inappropriately high language levels. [Table 1](#) includes examples of problematic

Table 1 Examples of Problematic Items and Domains

Item: “I feel like a sexual person”	After translation in French: “I feel like a prostitute”
Item: “I felt relaxed about sex”	Some patients reported feeling relaxed about the subject of sex but not about actually having sex
Item: “I felt apathetic about sex”	Patients did not know the meaning of the word “apathetic”
Item: “It took forever to get aroused”	Patients who did not have sex over the past month answered “never”
Domain: Spontaneity	Unintended meaning: Women with low libido reported that sexual activity was often spontaneous because they rarely thought about or planned for it
Anticipated meaning: A tendency of women with normal libido to engage in sexual activity with little or no planning	
Domain: Attractiveness	Ambiguous meaning: Patients were unclear as to whether the domain referred to how attractive they themselves felt, how attractive they found men, or if they felt men were attracted to them
Anticipated meaning: A measure of how attractive the patient felt	

items and domains identified by the cognitive interview process during the development of the PFSF.

While it is unlikely that ambiguity and unintended meaning can ever be reduced to zero, to the extent that this goal is achieved, the content of the instrument becomes increasingly valid for its intended purpose. The cognitive interview process ensures that an instrument measures what it is intended to measure and does so independently of its statistical properties.

QUANTITATIVE INSTRUMENT DEVELOPMENT

The third step in this approach uses psychometric item reduction and multivariate analysis techniques to identify the underlying factors that describe the characteristics of the disease that are meaningful to the patients and to determine the subset of items that reliably measures each factor (i.e. the measurement model). The processes are executed among patients drawn from the same population as those who generated and reviewed the items and domains, and are the target population for use with the instrument. These methods are widely discussed in the psychometric literature.^[9,11,23–25]

METHODS OF ANALYSIS

Developing a good scale involves selecting items that are good indicators of factors and ensuring factors are well-measured (i.e. developing a good measurement model). Although many steps have been incorporated along the way to provide items that will perform well, there still may be issues.

Hence, we've broken this process into two sets of analyses: (i) analyses to reduce the number of items in the instrument by eliminating those with poor data quality; and (ii) analyses to identify a good measurement model. Although these analyses are discussed in a sequential fashion, in practice, the process is iterative, with later analyses sometimes necessitating a review of decisions based on earlier results. Additionally, although items performing poorly may warrant immediate elimination, for others, decisions as to whether or not to retain the item may hinge on the combined results of multiple analyses as well as criteria involving clinical considerations.

DATA QUALITY

Missing data rates and item frequency distributions are examined to identify problematic items. High missing data rates on an item may indicate that the item was not understandable or relevant. If the inventory deals with a sensitive subject, for example, items with high missing data rates may represent a topic that patients do not feel comfortable

responding to, perhaps because the item is viewed as too personal or intrusive. A lack of understanding of the item may be another reason for high missing data rates. Additionally, overall high missing data rates may indicate patient fatigue, which could indicate an overly lengthy instrument.

Items with high “endorsement frequency”^[23] (i.e., items having a high percentage of subjects with floor or ceiling responses or one particular response, e.g. exceeding 80%) should be considered for elimination. Assuming that a high response to a particular item indicates a lack of that aspect of the disease, subjects scoring at the highest response level would not be able to move in a more positive direction with respect to that item after treatment. Therefore retention of that item may reduce the sensitivity of the instrument to clinical intervention. For example, 48% of the patients responded either “seldom” or “never” to the phrase “I feel sexually unattractive” and hence this item was deleted.

Items having a large proportion of subjects with floor responses or one particular response may also warrant elimination. Although a subject scoring at the lowest response would have a large potential for improvement, a high percentage of subjects with that response may indicate that the item difficulty level is such that movement to the next response would take an inordinate amount of change, again reducing the sensitivity of the instrument to detect therapeutic changes in the positive direction. A typical sample of patients represents various degrees of the severity of the disease. Sets of items that adequately describe a range of disease levels will likewise allow patients' responses to more accurately describe their particular level of the disease, and to reflect movement indicating either an improvement or a worsening of their condition.

Other methods are discussed by Juniper et al.^[10] They suggest eliminating items based on a composite score, termed “impact,” consisting of the proportion of patients in a sample experiencing an item multiplied by the mean rating of importance. Items having a low “impact score” are deleted. For example, items that are experienced by a large proportion of patients but have low importance would be eliminated from the instrument. Another method they suggest is to eliminate items not changing with known changes in the disease state or after treatment with a known effective therapy. This depends upon an effective therapy already existing for the condition or a disease state of relatively short duration with a clear onset and a known duration.

DETERMINATION OF FACTOR STRUCTURE

One approach to determine domains is to use judgment based on clinical experience or scientific knowledge, or domains in already established instruments. Each of these

approaches has its legitimate use, especially if a theory in the area is developed enough to feel confident with these decisions. However, the scientific and medical conceptualization of disease states evolves and changes over time, as does the patients' conceptualizations and ways of expressing what is important.^[11] Even individual words and items can change in meaning or relevance. All of these factors can contribute to the invalidity of an instrument or item. Neither factor analysis (discussed below) nor scientific and clinical considerations alone can adequately guard against this difficulty. Typically, both are useful to develop a sound measurement tool, which is patient-relevant and reflects state-of-the-art science.

Instruments typically are developed so that the items within a domain are homogeneous. In other words, they each measure some facets of the same underlying construct. Streiner and Norman^[24] note that without this property, a domain may be tapping a number of traits. Consequently, the exact effect of treatment will be indeterminate. Analyses that establish scale homogeneity are discussed here.

Factor analytic methods can be used to assess the underlying dimensionality of a set of items and to identify groups of related items that are hypothesized to measure unitary underlying constructs. In practice, the constructs are referred to as factors or domains. The factor analysis model expresses a set of items as a continuous function of factors and residuals. Exploratory factor analyses is an exploratory procedure which does not pose any constraints on the relationship between the items and the factors but only restricts the number of factors. Confirmatory factor analysis will be discussed below where constraints are placed on this relationship; forcing items to relate to particular factors. A nice overview of EFA and CFA as well as a potential strategy for instrument development is given by Muthen and Muthen in their Mplus short course notes, Topics 1 (continuous outcome variables) and 2 (categorical outcome variables).^[26]

Initially, a principal components analysis is run in order to determine if all items, or which items, relate to a single general factor. The first component of an unrotated principal components analysis identifies those items that tend to share a common underlying attribute or construct in the dataset. In the case of the PFSF, items should revolve around sexuality. Items that do not have a meaningful correlation (e.g., correlates less than 0.4) with the first unrotated principal component are removed. Inventories will frequently incorporate both positively and negatively worded items as a means of countering acquiescent responding,^[23,25,27,28] in which patients show a tendency to agree with all items regardless of content. In that case, it is possible that positive and negative items will distribute across the first two principal components, with one component containing positively worded items and the other containing negatively worded items, even though both measure the same underlying general construct. In

that case, items correlating with the first two unrotated principal components should be retained.

Principal components analysis is followed by a rotated factor analysis using (e.g., a varimax and/or oblique rotation). This analysis is done to identify factors (i.e., linear combinations of items) that account for distinct portions of the variability in the overall data set. A scree plot of the eigenvalues—a measure of the amount of variability extracted by a factor, corresponding to each factor, by a factor number—can be used to aid in determining the number of factors to be rotated. One option, commonly used, is to base the decision on the number of factors with eigenvalues greater than 1. The justification for this is that any factor retained must contribute more than the total variance added by a single variable. A discussion of this topic as well as other methods of factor analysis can be found in Comrey and Lee.^[29]

Items that correlate highly with a given factor are said to load on that factor and can be viewed as a measurement scale for the underlying construct represented by the factor. Items showing complex correlations across multiple dimensions (i.e., loading on numerous factors) or items having small loadings across all dimensions (e.g., factor loadings <0.4 on all factors) should be eliminated from the instrument.

A portion of the factor analysis table from the factor analysis performed on the PFSF is given in [Table 2](#). Items and domains have been selected for illustrative purposes and do not represent the entire factor analysis table. The first four items show unique loadings on the second factor, later to be labeled “Sexual Desire,” and illustrate a pattern of factor loadings indicating that the items are all primarily measuring the same underlying factor. Certain items considered originally to be capturing separate domains of arousal and orgasm were found to load on the same factor (Factor 4). As will be discussed later, two domains, “Sexual Arousal” and “Orgasm,” were later constructed based on clinical perspective. Lastly, for a number of items in the original Pleasure domain, unique loadings on the first factor were obtained. However, the item “It was very easy to have sex” was deleted from the instrument because of relatively low and multiple factor loadings, suggesting some degree of ambiguity. The second item, “Having sex was a chore,” although initially thought to represent “Sexual Pleasure,” was later found to load to a greater extent on the “Sexual Responsiveness” domain and was therefore moved to this domain.

In parallel to the above analyses, item-total correlations can be used to build and confirm scale homogeneity, as well as changes in internal consistency as individual items are removed. As further discussed in “Instrument Validation” below, these properties are important in establishing scale homogeneity. The item-total correlation measures the correlation between an item and its domain total score and is assumed to reflect the extent to which the item measures the same underlying construct as do the other items in the

Table 2 Example of a Rotated Factor Analysis

Original domain item	Factor 1: Pleasure	Factor 2: Desire	Factor 3: Responsiveness	Factor 4: Arousal/orgasm
<i>Desire</i>				
I felt sexual desire	0.25	0.73	0.14	0.07
Desire item 2	0.22	0.73	0.37	0.01
Desire item 3	0.20	0.76	0.29	0.12
Desire item 4	0.21	0.74	0.18	0.18
<i>Arousal</i>				
It was very difficult for me to become aroused	0.29	0.46	0.14	0.66
Arousal item 2	0.24	0.37	0.26	0.61
Arousal item 3	0.25	0.24	0.30	0.65
<i>Orgasm</i>				
It took a lot of work for me to reach orgasm	0.27	−0.01	0.05	0.84
Orgasm item 2	0.32	0.11	0.03	0.83
Orgasm item 3	0.44	0.26	−0.07	0.62
<i>Pleasure</i>				
It was very easy to have sex	0.40	0.26	0.48	0.16
Having sex was a chore	0.21	0.15	0.74	0.12

domain. The item-total correlation is corrected for overlap by removing that item from the sum of the other items when computing the domain score. Items not correlating significantly with their own domain total (e.g., item-total correlation corrected for overlap <0.4) are considered for elimination. Additionally, items correlating more with the total domain score of another domain should be eliminated or moved to the other domain. The principle behind this analysis is similar to that of the principal component and rotated factor analysis and provides complementary results.

Another approach, which can be used to build scale homogeneity, is to consider changes in the reliability (e.g., Cronbach's α) of the domain as it is computed with and without each item.^[24] If a significant increase in reliability is noted when a particular item is excluded, then that item should be eliminated.

The determination of domains can, and often must, be done in conjunction with nonstatistical considerations. For example, as shown in Table 2, factor analyses done on the PFSF items found that those appearing to measure sexual arousal and those measuring orgasm loaded on a single factor. Sexual arousal and orgasm are viewed by patients as separate entities. Further, the DSM-IV5 lists arousal and orgasm disorders as distinct diseases. These considerations made separating the two sets of items into two inventory domains very reasonable.

The construction of domains by the use of factor analysis and item-total correlations will result in homogeneous or nearly homogeneous domains. As mentioned earlier, the more similar the items are to each other, the more likely they are to measure the same construct. However, as items become increasingly similar, they are also more likely to measure

the same facet of the underlying construct. Such items add little to the domain, and high degrees of redundancy may decrease the content validity of the domain. Factor analysis and item-total correlations do not address this issue. Redundant items within a domain can be identified by inter-item correlations. As a rule of thumb, items that correlate highly (e.g., interitem correlation >0.75) are considered redundant. Redundant items can also influence the results of the factor analyses, leading to trivial factors or bloated specific factors. Hence looking at inter-item correlations both before and after the factor analysis is useful.

CONFIRMATORY FACTOR ANALYSIS

Once the final measurement model is identified, CFA methods can be used to assess whether the factor model established above fits a new sample from the same population or potentially to investigate its fit in a new population. Brown et al.^[30] provides an in-depth discussion of CFA modeling and its advantages over EFA modeling.

The final PFSF included 37 items allocated to seven domains; Sexual Desire (9 items), Sexual Arousal (3 items), Sexual Self-image (4 items), Orgasm (4 items), Sexual Concerns (3 items), Sexual Pleasure (7 items), and Sexual Responsiveness (7 items).^[7,8]

A 7-factor CFA model was fit in Mplus using an independent sample of data from 2 large phase III clinical studies to assess the efficacy and safety of a testosterone patch versus placebo on Sexual Desire and other aspects of sexual functioning.^[31,32] The majority of study patients were from the United States ($n = 1020$ of 1094 subjects)

and hence initial interest was to confirm the factor structure in this set of patients.

To assess the fit of the model the comparative fit index (CFI) and Tucker Lewis index (TLI) were used with values above 0.9 indicating an acceptable fit.^[33,34] The root-mean-square error of approximation (RMSEA) and Standardized Root Mean Square Residual (SRMR) were also examined with values less than 0.08 considered acceptable.^[35]

The CFA model produced a significant chi-square value (chi-square = 3403, df = 608, $p < 0.0001$). Marsh and Hau^[36] suggest using the Chi/df index over the Chi-square alone. Values less than 5 indicate reasonable fit while values between 2 and 3 indicate good fit. In this case, the Chi/df was slightly greater than 5, with a value of 5.5.

Values of 0.90 and 0.89 were observed for CFI and TLI, respectively. The RMSEA was 0.069 with 90% CI: (0.067, 0.071) and SRMR was 0.071. All factor loadings were highly statistically significant. Hence the data fit a seven-factor CFA reasonably well.

Even though an acceptable fit was observed, further examination for the purpose of illustration was performed. Examination of the modification indices (measure of change in chi-square associated with freeing the value of the parameter) indicated the possibility that item 4 on the Sexual Self-Image scale (“Unhappy about yourself sexually”) may cross-load with the Sexual Concerns domain as well as the potential for some method effects (e.g. due to positive and negative wording of items). For example, items 5 (“I lacked sexual desire”) and 7 (“I was uninterested in sex”) of the Sexual Desire domain had a large modification index indicating correlated residuals. In this case, these two items were negatively balanced while the remaining questions for that domain were positively balanced. The ability to model correlated residuals and hence address method effects is one of the major advantages discussed by Brown^[30] of CFA over EFA.

A model was fit where item 4 of the Sexual Self-Image scale was eliminated and 8 pairs of error terms were allowed to co-vary. In this case, the minimum criteria for acceptance were observed for all the fit indices.

The above analyses treat the response scale as continuous. However, the items in the PFSF utilized a 6-point Likert scale ranging from 1 for Always to 6 for Never (high score associated with high sexual functioning). In this case, categorical EFA and CFA models are more appropriate. Categorical EFA and CFA models can be run in Mplus.^[26] Polychoric correlation matrices are used rather than Pearson correlation matrices for categorical EFA. 2-parameter logistic or normal ogive IRT models are utilized to model the probability of a particular response given a person's underlying latent trait value.

Analyses on the data above were repeated treating the items as 4-point categorical variables where response options 4, 5, 6 (e.g. “I lacked sexual desire” responses of “sometimes”, “seldom”, or “never”) were collapsed due to very few patients with those responses. In this case, similar

results were seen as those discussed above with continuous data again suggesting reasonable fit to a 7-factor model.

Muthen and Muthen^[26] indicate that treating the data as categorical has less to do with the number of response categories but rather the presence of floor and ceiling effects. Treating the variables as continuous can lead to attenuated correlations and factors that reflect item difficulty extremeness.

ITEM RESPONSE THEORY

The value of eliminating redundant items and of including items with varying degrees of difficulty have been discussed previously. A tool that allows assessment of these characteristics for a set of items is item response theory (IRT).

Item response theory (IRT) models are mathematical functions relating the level of an individual's latent trait with the probability of observing a particular response. Hays et al.^[37] provide a nice overview of IRT basics, common models, advantages provided by IRT models, as well as challenges in applying IRT methods. For IRT models, unidimensionality is assumed. Because this may not hold strictly, analyses to assess whether the scale is “essentially” unidimensional have been considered.^[17,38]

Figure 1 below shows the Item Characteristic Curve (ICC) for items 1, 3 and 8 of the Sexual Desire domain based on a 1-dimensional CFA of all 9 items. The probability of endorsement, in this case, responding “Sometimes,” “Often,” “Very Often,” or “Always” (e.g. item 1 “I felt like having sex”) for a given level of Sexual Desire is plotted for each item based on a 2-parameter logistic model.

In this model there are 2 parameters of interest. The discrimination parameter (slope of the ICC curve) considers the change in probability associated with changes in the level of the underlying trait. Items with higher discrimination will have steeper slopes. In this example, item 8 (“I got warm all over just thinking about sex”) discriminates less than questions 1 and 3.

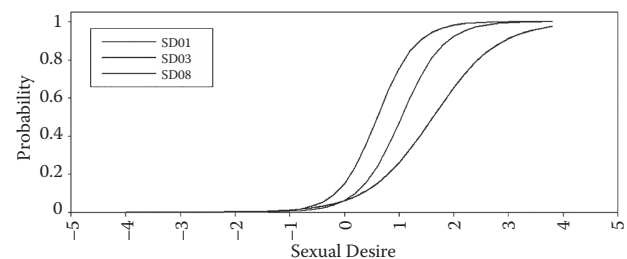


Fig. 1 Item Characteristic Curves corresponding to items 1, 3, and 8. Probability corresponds to the probability of responding “Sometimes”, “Often”, “Very Often”, or “Always” (e.g. item 1 “I felt like having sex”) for a given level of Sexual Desire

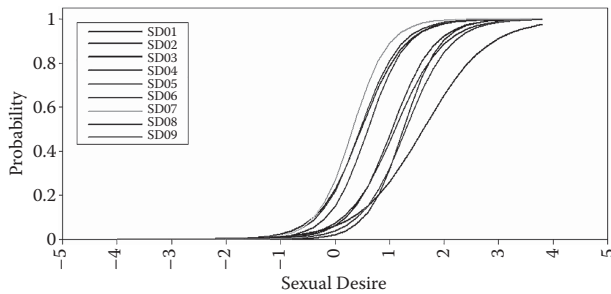


Fig. 2 Item Characteristic Curves for all 9 Sexual Desire Items. For positive balanced items (1–4, 6, 8–9) the y-axis corresponds to the probability of responding “Sometimes”, “Often”, “Very Often”, or “Always” (e.g. item 1 “I felt like having sex”) for a given level of Sexual Desire. For negative balanced items (5 and 7) the y-axis corresponds to the probability of responding “Never”, “Seldom”, “Sometimes”, or “Often” (e.g. item 5 “I lacked sexual desire”) for a given level of Sexual Desire

The second parameter is the difficulty parameter and corresponds to the value of the latent trait at which the probability of endorsement is 50%. In this case, item 8 is the most “difficult” item since it takes more of the latent trait, i.e. Sexual Desire, for it to be endorsed, while 1 is the easiest since it takes less of the latent trait to be endorsed.

Figure 2 illustrates the ICC for all 9 items of the Sexual Desire domain. In this case, one can see that for some items, e.g. items 4 and 5, 3 and 6, and 2 and 9, similar discriminations and difficulties in measuring Sexual Desire are obtained.

Although not discussed here, an extension of the 2-parameter logistic model for ordered categorical variables is given by the graded response model. In this case, each item is described by one slope parameter and $k-1$ between category threshold parameters where k refers to the number of response categories.^[37]

Evaluating the discrimination and difficulty parameters can provide information to allow for shorter yet reliable scales to be assembled. Numerous other uses for IRT models in scale construction, such as test equating, assessment of appropriateness of response options, or assessing differential item bias to name a few, can be found in the articles already referenced.

SCORING

A common algorithm for summarizing domain and inventory data is to compute a simple average or sum of the items in the domain. A second possibility is the weighted average of the items in the domain where the weights are the factor loadings from the factor analysis. Unless the factor analysis producing the weights is based on a very large sample, these estimates may be unreliable and specific to the sample from which they were derived. Other possible weights include item-total or item-criterion

correlations, but these correlations, like factor loadings, can vary from study to study because of sampling error. The differential weighting of items generally adds little value in terms of increased reliability beyond what the number and the quality of items provide [9]. A simple average with unit weight is therefore usually appropriate. In the case of missing responses to items, a common approach is to impute the average of the nonmissing items for the missing items, in a domain, prior to computing the score.

INSTRUMENT VALIDATION

After the items and domains have been identified, the psychometric properties of the instrument are assessed to determine if the instrument meets established criteria of reliability and validity. While the major indices of reliability and validity can be computed on the data collected during the item reduction process, such calculations may overestimate the true value as similar analyses may have been used to develop the instrument. To demonstrate the generalizability of the instrument beyond the data from which it was generated, the inventory must be tested among an independent sample of patients drawn from the same target population.

One method to obtain an independent dataset is to withhold a subsample of data from patients in the initial development and item reduction analyses and later use this sample to examine the properties of the reduced instrument. However, this procedure may be influenced by the fact that the withheld data that are later analyzed for psychometric properties consist of item responses collected in the context of the unreduced instrument, thereby allowing the possibility of context effects.^[21] Therefore the use of an independent study to establish instrument properties is the preferred approach.

Psychometric properties considered necessary for a valid and reliable instrument include good test-retest reliability and internal consistency, and good content, concurrent, and construct validity. The assessment of an instrument’s validity is an ongoing process in which hypotheses regarding the constructs measured by the instrument are repeatedly generated and tested. It is this accumulation of evidence, supporting the instrument’s usefulness as a measurement tool of the intended disease state, which establishes its validity. Definitions for validity and reliability can be found in numerous texts.^[9,23–25]

VALIDITY

Validity refers to the property of measuring what is intended to be measured. Three forms of validity that are part of a validated instrument are content, concurrent, and

construct validity. Content validity refers to the adequate sampling and presentation of relevant disease characteristics. This form of validity was discussed in earlier sections concerning qualitative validation and therefore will not be further discussed in this section.

CONCURRENT VALIDITY

Concurrent validity is established for a new instrument by demonstrating a good correlation with an already existing tool that is widely accepted as measuring the same construct(s). For example, a clinical evaluation by a physician might be considered the “gold standard” diagnosis. Or a well-recognized and accepted diagnostic criterion that is widely used by practitioners in the area may exist. In the case that the existing tool is continuous, a Pearson or Spearman’s correlation coefficient can be computed.^[39] In the case of a dichotomous diagnostic tool, the area under the Receiver Operating Characteristics curve^[40] can be used to establish a good relationship.

If no such “gold standard” exists, one option is to compare the new instrument to existing instruments that measure similar or related constructs. This is referred to as convergent validity and is a form of construct validity (discussed below). For example, all instruments in the same general area of sexual function should show some degree of relationship. Furthermore, all instruments or domains designed to measure sexual desire should show a somewhat closer relationship to one another than they do to other domains of sexuality, even if they have been developed on varying populations.

An example of convergent validity is shown in [Table 3](#). These data show the relationship between data on the domains of the PFSF and the Derogatis Interview for Sexual Function-Self Report© (DISF-SR21). The DISF-SR is a gender-keyed brief outcome measure to assess general sexual functions. The instrument measures five domains of sexual function: sexual cognition and fantasy, sexual arousal, sexual behavior and experiences, orgasm, and

sexual drive and relationship.^[41] Correlations of domains that were expected prospectively to be strongly related are shown in bold print. The overall pattern of correlations in this table is supportive of the convergent validity of the PFSF.

CONSTRUCT VALIDITY

Construct validity is the extent to which an instrument or its domains measure the underlying constructs they are purported to represent. One way to examine construct validity is to determine if the data collected by the instrument reflect hypotheses generated by a consideration of the underlying constructs. To the extent that hypotheses are developed and supported by the instrument, the instrument’s construct validity is accrued.

One form of construct validity is referred to as “known groups validity” or discriminant validity. It refers to the instrument’s ability to detect differences between groups “known” or expected to differ. For example, the PFSF is purported to characterize the subjective sexual feelings and symptoms of women with HSDD. Women diagnosed with HSDD are therefore expected to score lower on the domains of the PFSF than women with normal libido. [Fig. 3](#) shows the mean responses on the domains of the PFSF from the three original validation studies conducted in the United States, Canada, Europe, and Australia (data on file). A repeated-measures analysis of covariance showed that women with HSDD scored significantly lower than normal women in all regions studied ($p < 0.001$). Of note also is the similarity of profiles for the three general regions studied, hence supporting the expectation from the qualitative content validation that the same content would be captured over various regions.

Similar to convergent validity discussed above, divergent validity can additionally be used to establish construct validity. Just as one would expect two measures assessing the same construct to be moderately or strongly related,

Table 3 Correlation of the PFSF with the DISF-SR

PFSF domain	DISF-SR domain				
	Sexual cognition/ fantasy	Sexual arousal	Sexual behavior/ experiences	Orgasm	Drive and relationship
Sexual desire	0.64	0.76	0.67	0.69	0.82
Arousal	0.49	0.70	0.56	0.80	0.82
Orgasm	0.47	0.66	0.47	0.87	0.70
Sexual pleasure	0.51	0.73	0.56	0.83	0.81
Sexual concerns	0.45	0.67	0.55	0.74	0.78
Sexual responsiveness	0.50	0.67	0.63	0.66	0.79
Sexual self-image	0.46	0.66	0.60	0.66	0.73

PFSF, Profile of Female Sexual Function; DISF-SR, Derogatis Interview for Sexual Function—Self-Report.

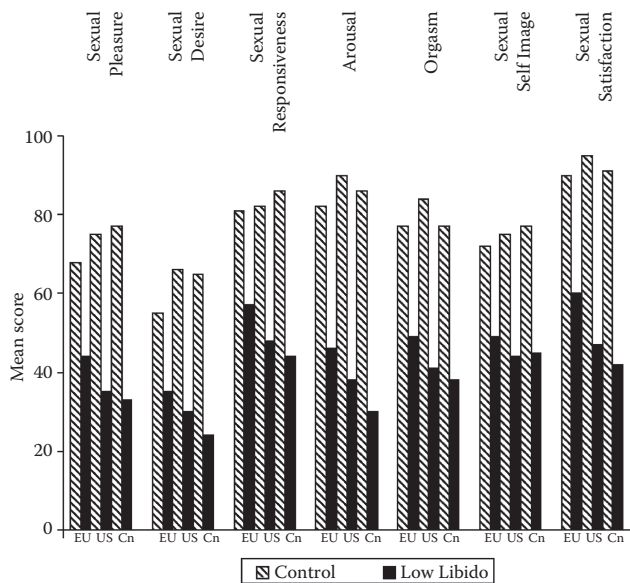


Fig. 3 The PFSF—discrimination by domain across geographies. EU = Europe/Australia; US = United States; Cn = Canada

one would expect measures that are conceptually independent to be unrelated, or only weakly related. The establishment of low correlations between independent measures is known as divergent validity. For example, Rosen et al.^[42] established divergent validity for the International Index of Erectile Function (IIEF) by showing low correlations of the IIEF with the Locke-Wallace Index of Marital Satisfaction Scale. The observed lack of relationship provides confidence that the IIEF measures erectile function and not some aspect of marital adjustment.

Convergent and divergent validity were discussed above relative to the domain level. Additionally, convergent and divergent validity can be established at the item level through the use of item-total correlations. An instrument has good convergent and divergent validity if items correlate to a greater extent with their own domain than with the other domains. Multitrait/Multi-item Analysis Program—Revised (MAP-R)^[43] computes the percentage of correlations in which an item correlates significantly more with its own domain than with other domains and classifies an instrument as having achieved “scaling success” if this occurs for at least 80% of the item-total correlations within each domain and overall.

For instruments that are intended for use in clinical trials with drug therapy, an important form of construct validity is responsiveness to change or sensitivity to the effects of treatment. As discussed in “Quantitative Instrument Development,” the responsiveness of an item to change can be a vital aspect of instrument development. Rosen et al.^[42] demonstrated the IIEF to be sensitive and specific to changes after treatment by showing statistically significant changes from baseline in domain scores for a

subgroup of subjects self-rated as responders while no statistically significant changes were found for the subgroup of subjects self-rated as nonresponders.

TEST-RETEST RELIABILITY AND INTERNAL CONSISTENCY

Reliability refers to the reproducibility of a measure on separate measurement occasions. Ideally, an individual's score on a measurement tool will remain constant over time if no intervention or other relevant event that might be expected to affect measurement outcome occurs. Scores on successive testing can be affected in several ways. For example, a true change in the underlying condition can occur with treatment, or spontaneously in the absence of treatment. Alternatively, a change in a patient's score might occur spuriously as a result of prior exposure to the instrument. The demonstration of test-retest reliability provides the investigator with increased assurance that any changes in scores from one testing to the next are the result of a true change in the underlying disease and are not a spurious result related to repeat testing.

An intraclass correlation coefficient (ICC) corresponding to a one-way random effects model, $\rho = [(\sigma_s^2)/(\sigma_e^2 + \sigma_s^2)]$, can be used to estimate test-retest reliability. This coefficient represents the correlation between two responses from the same individual. Shrout and Fleiss^[44] discuss other models for estimating ICC and give the assumptions underlying these models. Bartko^[45] compares these approaches and shows that the ICC from a one-way random effects model is the most conservative. Alternatively, a Pearson correlation coefficient can be used to estimate test-retest reliability. However, this estimate is less conservative than ICC in that a constant change between repeated sampling of the patients can lead to a large Pearson correlation even when subjects' responses have not remained constant.

Another form of reliability can be estimated by Cronbach's α coefficient. Cronbach's α coefficient estimates the internal consistency or scale homogeneity of a domain. It is a function of the number of items in a domain and the average interitem correlation. Cronbach's α values between 0.75 and 0.94 are desirable for population-based instruments, and values at or above 0.95 are desirable for instruments to be used for measurement or diagnosis with individuals.^[46]

Table 4 illustrates test-retest reliability and internal consistency estimates for the PFSF from an independent confirmatory validation study involving menopausal, low-libido women and age-matched, normal-libido controls. The ICCs indicate consistent responses across a 4-week time interval. Additionally, Cronbach's α values indicate good domain homogeneity for all the domains.

Table 4 Test-Retest Reliability and Internal Consistency of the PFSF

Domain	Test-retest reliability		Internal consistency	
	Low libido	Normal libido	Low libido	Normal libido
Sexual desire	0.73	0.76	0.91	0.93
Arousal	0.63	0.68	0.92	0.81
Orgasm	0.75	0.82	0.92	0.88
Sexual pleasure	0.76	0.80	0.95	0.93
Sexual concerns	0.58	0.74	0.86	0.74
Sexual responsiveness	0.83	0.91	0.92	0.89
Sexual self-image	0.65	0.78	0.83	0.85

CONCLUSION

The development and the validation of psychometric instruments require both qualitative and quantitative methods to optimize content validity and to establish acceptable psychometric and statistical properties. A comprehensive approach to achieve this objective was discussed in this article. The approach to instrument development discussed in this article consists of three major steps: (i) preliminary item and domain generation; (ii) qualitative content validation; and (iii) quantitative item reduction analyses. The incorporation of the methods discussed in the first two steps, in particular the cognitive interview process, is important to achieve a complete sampling of the important characteristics of the disease, adequate patient understanding, and the robustness to language variation. The quantitative methods are designed to identify sensitive items that are capable of providing good-quality data and stable data and domains that measure known unique aspects of the disease. These properties enhance the ability of the instrument to determine the exact effects of treatment.

The validation of an instrument requires a demonstration of acceptable reliability and validity characteristics. Standard analyses for establishing and assessing reliability and validity were given in this article.

ACKNOWLEDGMENTS

We would like to acknowledge Dr. Colleen McHorney, Dr. Leonard Derogatis, Dr. John Rust, and Dr. Susan Golombok for their contributions to the development of the PFSF and their direction in determining the appropriate methodology.

REFERENCES

1. *Measures of Personality and Social Psychological Attitudes*; Academic Press: New York, 1991.
2. Davis, C.M.; Yarber, W.L.; Bauserman, R.; Schreer, G.; Davis, S.L. *Handbook of Sexuality-Related Measures*; Sage Publications: London, 1998.
3. PAR Catalog of Professional Testing Resources **2002**, 25 (1).
4. Psychological Corporation. The Catalog for Psychological Assessment Products; **2002**.
5. MAPI Research Institute. ProQolid (Patient-Reported Outcomes and Quality of Life Instrument Database); <http://proqolid.org/>
6. *Diagnostic and Statistical Manual of Mental Disorders*, 4th Ed; American Psychiatric Association: Washington, DC, 1994.
7. McHorney, C.A.; Rust, J.; Golombok, S.; Davis, S.; Bouchard, C.; Brown, C.; Basson, R.; Sarti, C.D.; Kuznicki, J.; Rodenberg, C.; Derogatis, L. Profile of female sexual function: a patient-based, international, psychometric instrument for the assessment of hypoactive sexual desire in oophorectomized women. *Menopause* **2004**, *11* (4), 474–483.
8. Derogatis, L.R.; Rust, J.; Golombok, S.; Bouchard, C.; Nachtigall, L.; Rodenberg, C.; Kuznicki, J.; McHorney, C.A. Validation of the profile of female sexual function (PFSF) in surgically and naturally menopausal women. *J. Sex Marital Therapy* **2004**, *30*, 25–36.
9. Nunnally, J.C.; Bernstein, I.H. *Psychometric Theory*, 3rd Ed; McGraw-Hill: New York, 1994.
10. Juniper, E.F.; Guyatt, G.H.; Jaeschke, R. *How to Develop and Validate a New Health-Related Quality of Life Instrument. Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed; Lippincott-Raven: Philadelphia, PA, 1996, 49–56.
11. McHorney, C.A. Methodological and psychometric issues in health status assessment across populations and applications. *Adv. Med. Sociol.* **1994**, *5*, 281–304.
12. Streiner, D.L.; Norman, G.R. *Health Measurement Scales: A Practical Guide to Their Development and Use*, 4th Ed.; Oxford University Press: Oxford, 2008.
13. Lynn, M.R. Determination and quantification of content validity. *Nurs. Res.* **1986**, *35*, 382–385.
14. Haynes, S.N.; Richard, D.; Kubany, E.S. Content validity in psychological assessment: A functional approach to concepts and methods. *Psychol. Assess.* **1995**, *7*, 238–247.
15. Cella, D.; Yount, S.; Rothrock, N.; Gershon, R.; Cook, K.; Reeve, B.; Ader, D.; Fries, J.F.; Bruce, B.; Rose, M. PROMIS Cooperative Group. The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap cooperative Group During Its First Two Years. *Med. Care* **2007**, *45*, s3–s11.

16. DeWalt, D.A.; Rothrock, N.; Yount, S.; Stone, A.A. PROMIS Cooperative Group. Evaluation of item candidates: The PROMIS qualitative item review. *Med. Care* **2007**, *45*, s12–s21.
17. Reeve, B.B.; Hays, R.D.; Bjorner, J.B.; Cook, K.F.; Crane, P.K.; Teresi, J.A.; Thissen, D.; Revicki, D.A.; Weiss, D.J.; Hambleton, R.K.; Liu, H.; Gershon, R.; Reise, S.P.; Lai, J.S.; Cella, D. PROMIS Cooperative Group. Psychometric evaluation and calibration of health related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med. Care* **2007**, *45*, s22–s31.
18. McHorney, C.A. Health status assessment methods for adults: Past accomplishments and future challenges. *Annu. Rev. Public Health* **1999**, *20*, 309–335.
19. Thomas, C.; Pouget, C.; Nadjar, A.; Derogatis, L. Cultural adaptation of the DISF-SR into French, German, Finnish and Polish. *Qual. Life Newsl.* **2000**, *24*, 11.
20. Graesser, A.C. Question Understanding Aid (QUAID): A web facility that tests question comprehensibility. *Public Opin. Q.* **2006**, *70*, 3–22.
21. Sudman, S.; Bradburn, N.M.; Schwarz, N. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*; Jossey-Bass: San Francisco, CA, 1996.
22. Jobe, J.B.; Mingay, D.J. Cognitive laboratory approach to designing questionnaires for surveys of the elderly. *Public Health Rep.* **1990**, *105*, 518–524.
23. Rust, J.; Golombok, S. *Modern Psychometrics*, 2nd Ed; Routledge: London, 1999.
24. Streiner, D.L.; Norman, G.R. *Health Measurement Scales: A Practical Guide to their Development and Use*, 2nd Ed; Oxford University Press: Oxford, 1995.
25. Robinson, J.P.; Shaver, P.R.; Wrightsman, S. *Criteria for Scale Selection and Evaluation. Measures of Personality and Social Psychological Attitudes*; Academic Press: New York, 1991; 1–16.
26. Muthen and Muthen: Mplus Short Course Notes, Topics 1 and 2; <http://www.mplus.com>
27. Ray, J.J. Reviving the problem of acquiescent response bias. *J. Soc. Psychol.* **1983**, *121*, 81–96.
28. Fugita, S.S.; Crittenden, K.S. Towards culture- and population-specific norms for self-reported depressive symptomology. *Int. J. Soc. Psychiatry* **1990**, *36* (2), 83–92.
29. Comrey, A.L.; Lee, H.B. *A First Course in Factor Analysis*; Lawrence Earlbaum Associates: Hillsdale, NJ, 1992.
30. Brown, T.A. *Confirmatory Factor Analysis for Applied Research*; The Guilford Press: New York, 2006.
31. Simon, J.; Braunstein, G.; Nachtigall, L.; Utian, W.; Katz, M.; Miller, S.; Waldbaum, A.; Bouchard, C.; Derzko, C.; Buch, A.; Rodenberg, C.; Lucas, J.; Davis, S. Testosterone patch increases sexual activity and desire in surgically menopausal women with hypoactive sexual desire disorder. *J. Clin. Endocrinol. Metab* **2005**, *90*, 5226–5233.
32. Buster, J.; Kingsberg, S.A.; Aguirre, O.; Brown, C.; Breaux, J.G.; Buch, A.; Rodenberg, C.A.; Wekselman, K.; Casson, P. Testosterone patch for low sexual desire in surgically menopausal women: a randomized trial. *Obstet Gynecol* **2005**, *105*, 944–952.
33. Bentler, P.M. On the fit of models to covariances and methodology to the bulletin. *Psychol. Bull.* **1992**, *112*, 400–404.
34. Hatcher, I. *A Step-by-STEP Approach to using the SAS System for Factor Analysis and Structural Equation Modeling*; SAS Institute Inc: Cary, NC, 1994.
35. Browne, M.W.; Cudeck, R. Alternative ways of assessment model fit. In *Testing Structural Equation Models*; Bollen, K.A.; Long, J.S., Eds.; SAGE: Newbury Park, CA, 1993.
36. Marsh, H.W.; Hau, K.T. Assessing goodness of fit: is parsimony always desirable. *J. Exp. Educ.* **1996**, *64*, 364–390.
37. Hays, R.D. Item response theory and health outcomes measurement in the 21st century. *Med. Care* **2000**, *38*, II-28–II-42.
38. Reise, S.P.; Morizot, J.; Hays, R.D. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Qual. Life Res.* **2007**, *15*, 19–31.
39. Sheskin, D.J. *Handbook of Parametric and Nonparametric Statistical Procedures*; CRC Press LLC: Boca Raton, FL, 1997.
40. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a Receiver Operating Characteristics (ROC) curve. *Radiology* **1982**, *143*, 29–36.
41. Derogatis, L.R. The Derogatis Interview for Sexual Functioning (DISF/DISF-SR): An introductory report. *J. Sex Marital Ther.* **1997**, *23*, 291–304.
42. Rosen, R.C.; Riley, A.; Wagner, G.; Osterloh, I.H.; Kirkpatrick, J.; Mishra, A. The International Index of Erectile Function (IIEF): A multidimensional scale for assessment of erectile dysfunction. *Urology* **1997**, *49* (6), 822–830.
43. Hays, R.D.; Hayashi, T. Beyond internal consistency reliability: Rationale and user's guide for multitrait analysis program on the microcomputer. *Behav. Res. Meth. Instrum. Comput.* **1990**, *22* (2), 167–175.
44. Shrout, P.E.; Fleiss, J.L. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **1979**, *86* (2), 420–428.
45. Bartko, J.J. On various intraclass correlation reliability coefficients. *Psychol. Bull.* **1976**, *83* (5), 762–765.
46. McHorney, C.A.; Tarlov, A.R. The use of health status measures for individual patient level applications: Problems and prospects. *Qual. Life Res.* **1994**, *3*, 43–44.

Integrated Summary Report

Laura J. Meyerson

Biogen, Inc., Cambridge, Massachusetts, U.S.A.

Abstract

An integrated summary report is a compilation of all the evidence for efficacy and safety from the data collected in all completed clinical trials. This entry breaks down each individual component of the integrated summary report.

INTRODUCTION

There are many reasons to integrate and to summarize all the data from a clinical trial program. Each clinical trial in the program is unique in its objective and design. Some are small safety studies among normal volunteers, while others are efficacy trials in a large patient population. The primary reason to create an integrated summary is to compare and to contrast all the various study results and to arrive at one consolidated review of the benefit/risk profile. A second and important reason is to reach a defensible statistical conclusion, through an exploration of the integrated data, that no competing alternative hypothesis that can reasonably account for the observed findings exists. Third, pooling the data from various studies enables the examination of trends in rare subgroups of patients, such as the elderly, those with differing disease states (mild vs. severe), and those with comorbidities at baseline. Last, providing such a summary in the new drug application is required by the Food and Drug Administration (FDA) and other international authorities.

OVERVIEW

An integrated summary report is a compilation of all the evidence for efficacy and safety from the data collected in all completed clinical trials. There are two integrated summary reports: the integrated summary of efficacy (ISE) and the integrated summary of safety (ISS). Both are required for all new drug applications in the United States.

The analysis approach for the ISE is substantially different from that for the ISS. The ISS summarizes all data collected from the clinical trial program, including normal volunteers and patients. The ISE includes data only from those clinical trials that present some evidence of efficacy, either through surrogate markers or through well-designed tests for efficacy. The ISE requires a more prospective approach to analysis, whereas the ISS approach is more sensitive because it allows a retrospective search through

all the data to expose possibly rare but important side effects. As a result, pooling data across studies is required in the ISS to obtain more reliable estimates of the incidence of rare adverse events. A pooled analysis in the ISE is not required, but it can be useful for estimating a more precise treatment effect. However, pooling is generally not useful for substantiating evidence of efficacy, without side-by-side study results each showing evidence of efficacy separately. Therefore this entry has two sections. The first section is devoted to the ISE, the second pertains to the ISS.

There are a number of areas of specific interest that arise within the integrated summary report. A description of the demographic and baseline clinical features of the population treated during the course of the clinical trial program is necessary. Key questions concerning efficacy need to be addressed by considering the results of relevant trials and by highlighting the degree to which they reinforce or contradict each other. For example, when is it useful to pool trials to obtain a more precise estimate of the treatment effect? All the safety information available from the entire database of all studies should be summarized. How can the data be thoroughly searched so that any potential safety concern is identified?^[1]

INTEGRATED SUMMARY OF EFFICACY

Dose Rationale

One of the most important sections of the ISE is that on dose selection. Efficacy information at all doses studied should be analyzed in such a way that a dose or dosing interval can be recommended. Information that is needed to support a new drug application should identify an appropriate starting dose, as well as how to adjust dosage to the needs of a particular patient.^[1] The maximum dosage beyond which there is no likely benefit or that would produce unacceptable side effects should also be identified. It is important to identify the lowest dose exhibiting a clinically important effect. To know for certain that one has identified the minimally

effective dose, it is often necessary to study a dose that has no clinical benefit compared to placebo. This section of the ISE should contain the information from the dose response studies, that is, controlled information on the minimum, maximum, and average dose from the response curve for efficacy. It should also contain any other information from other doses. Safety needs to be brought into the discussion because choosing the dose is always a risk/benefit decision.

In order to provide proper dose–response information, there should be at least one prospective dose–response study in the ISE that explores more than three doses. In principle, being able to detect a statistically significant difference using pairwise comparisons between doses is not necessary in such a trial, if a statistically significant trend (upward slope) across doses can be established using all doses. The more doses explored in this study, the better the dose–response curve will be characterized. It should be demonstrated, however, that the dose to be recommended has a statistically significant and clinically meaningful effect. This can be done in the dose–response trials or in other adequate and well-controlled trials.

The entire database should be examined extensively for possible dose–response effects. Limitations of study design should be considered. For example, the titration scheme designs should not be pooled with fixed-dose designs. Many trials will titrate to the desired response. Weaker responders would then receive the higher doses. This may lead to an inverted U-shaped dose–response curve. Despite the different designs, all clinical data should be examined for dose–response information.

The entire database should be examined to explore possible differences in subgroups of patients based on baseline differences. Age, gender, and race should always be examined. Other important covariates may be baseline disease severity and baseline comorbidities. Height, weight, renal function, and lean body mass are important in looking at the dose–response data as a function of body size and metabolism.

Efficacy Data

The International Conference on Harmonization guideline^[2] says that individual clinical trials to demonstrate

efficacy must be performed and must be large enough to satisfy their objectives. Additional valuable information may also be gained by summarizing a series of clinical trials that address essentially identical key efficacy questions. This can be done in the ISE. The main results of such a set of studies should be presented in an identical form to permit comparison across studies, focusing on estimates plus confidence limits. The use of meta-analysis to combine these estimates of treatment effect across studies of the same dose regimen is often useful because it allows a more precise overall estimate of the treatment effect at those doses. However, only under exceptional circumstances would a meta-analysis be the most appropriate way to demonstrate efficacy via an overall hypothesis test.

Figure 1 shows the clinical trials that are summarized in the ISE. In this program, phase IIB and phase III studies are the focus of the ISE. These studies should be planned while taking into account that they will be combined and presented together for an overall estimate of the benefit of the therapy. They should have similar endpoints (primary and secondary), similar populations, and some overlap in the dose regimens. Phase IIB is the study that demonstrates “proof of principle,” usually through a statistically significant increase in efficacy with an increase in dose. There should be two Phase III studies—one to demonstrate a statistically significant and clinically meaningful benefit, and the other to replicate that demonstration.

Meta-Analysis

In pharmaceutical development, it is known at the time of positive Phase II results that a new drug application is likely to be filed. It is at this time that the researchers should start planning for the ISE. In the planning of the adequate and well-controlled trials, one should incorporate any needs that the ISE plan may require. It is at this time that the decision to do meta-analyses should be made in a prospective manner. The individual studies can then be designed with pooling as a goal. Which studies will be pooled and which studies will not can be decided prior to obtaining the results. Then pooling will not be influenced by the

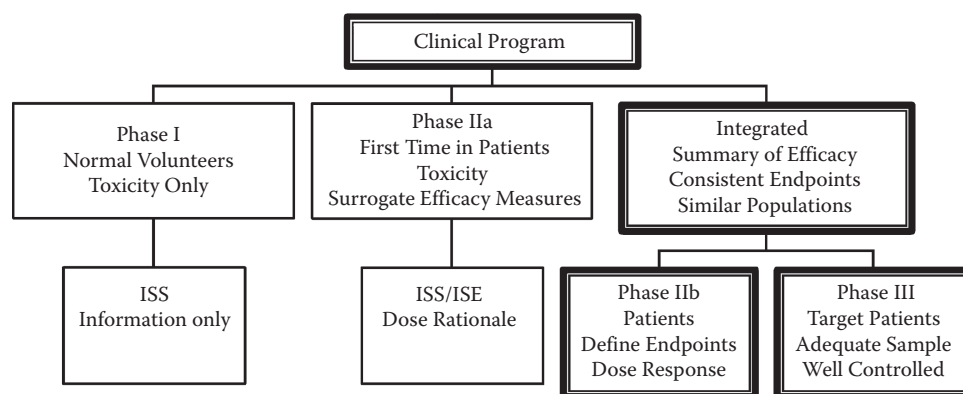


Fig. 1 The integration of efficacy information to support a marketing application

results. The decision to undertake a meta-analysis after the data from all the efficacy studies have been reviewed cannot be made to save any negative or borderline results that were obtained in the individual studies. In such a case, the researchers could pick and choose studies so that the results favor the treatment being proposed. A simple principle is to perform a meta-analysis in a manner that is consistent with the conduct of a single well-designed randomized clinical trial.^[3]

Flather et al.^[4] address why we do meta-analysis: “Even large randomized clinical trials may not answer specific questions reliably because of weaknesses in their design or, more commonly, because they are not large enough to detect the moderate but medically important treatment effects that can be expected realistically.” The primary reason for performing meta-analysis is to obtain more reliable estimates of the treatment effects. Secondary reasons for performing meta-analysis are (1) to summarize formally the available information on a particular treatment, and (2) to generate hypotheses that may be tested in future trials. For example, in the development of a treatment for ischemic stroke, each individual study could be powered for functional status, while the meta-analysis could be powered for mortality (which would generally take larger numbers). In this case, the randomized controlled trials are powered for differences in endpoints that are more sensitive (possibly, surrogates), while meta-analysis is powered for the less sensitive clinical endpoints. The pooling may allow the detection of clinically significant changes on clinical endpoints because of the increase in sample size. This is similar to the optimal information size discussed in Pogue and Yusuf.^[5]

In pooled analysis, all appropriate clinical trials should be included to avoid bias. Studies should be excluded if not of a consistent, similar design eligibility criteria, trial treatment regimens, concomitant medications, definitions of outcome events, or length of follow-up. Studies with incongruent results should not be pooled but should be examined for an explanation of their differences. Given the retrospective nature of most meta-analyses and the multiple comparisons that are carried out, many researchers believe that a *p* value of 0.05 is not stringent enough. Levels of 0.01 or even 0.001 should be routinely considered.^[6]

After the decision to do phase III studies, develop a prospective protocol for your meta-analysis by defining specific hypotheses (primary and secondary) to be tested:

Frame a question that is medically relevant and biologically sensible.

Define the study population and relevant subgroups.

Define treatment regimens to be included.

Define outcomes of interest.

Identify trials that will be pooled in a prospective manner.

Use only properly randomized trials.

Include all patients randomized in each trial.

Prespecify valid statistical methods to combine data.

Determine the primary endpoint as well as the effect size necessary to be clinically meaningful.

The ISE should have a description of the entire efficacy database’s demographics and baseline disease characteristics. Subgroup analyses for age, race, gender, and other characteristics should be looked at by pooling data over different studies of the same dose regimen. The intent-to-treat population should be the focus; however, if there is a large number of discontinued patients or major protocol violators, the results from a “clean” subpopulation, valuable for efficacy, should be presented as well.

The pooled results should always be presented in combination with the individual study results. This allows the examination of the degree of heterogeneity between the studies.^[7]

INTEGRATED SUMMARY OF SAFETY

The summary of safety data is an important and very large part of a new drug application. It is important to investigate the safety data thoroughly through pooling to search for rare adverse events. For example, if an adverse event rate is truly only 1%, then more than 300 patients are needed to observe it with 95% confidence. This is from “the rule of three,” which states: If none of *n* patients shows the event of interest, then the upper 95% confidence limit is approximately $3/n$. So if the event were truly rare, say 0.01%, then 30,000 patients would have to be observed, which is larger than most randomized clinical trials. It is also important to estimate accurately the expected rate of any common adverse event associated with the use of the drug.

The International Conference on Harmonization guideline^[2] states that all safety data need to be examined thoroughly to uncover any indications of potential toxicity, and to follow up any indications by searching for an associated supportive pattern of observations. The combination of the safety data from all human exposures to a drug should provide the main source of information, because its larger sample size provides the best chance of detecting the rarer adverse events and, perhaps, of estimating their approximate incidence. However, incidence data are difficult to evaluate without a natural comparator group; so a section on the examination of safety from the controlled studies alone is particularly useful. The results from the controlled studies should be combined separately for placebo and active controlled studies.

The components of an ISS should include the extent of exposure, characteristics of the population, deaths, dropout analyses for serious or potentially serious adverse events, rates of common adverse events, as well as clinical laboratory results. All indications of potential toxicity arising from statistical exploration of the data should be reported. The evaluation of the reality of these potential

adverse effects should take into account the issue of multiplicity arising from the numerous statistical comparisons made. The evaluations should make use of survival analysis methods to exploit the potential relationship of adverse events incidence to duration of exposure and follow-up. The risk associated with identified adverse events should be appropriately quantified to allow a proper assessment of the risk/benefit relationships.

There are at least three reasons to pool data across studies for the safety summary. One goal is to improve the precision of the incidence estimates. This is especially important for rare events. A second goal is to improve the statistical power for detecting any risk factors that may be associated with the use of the drug and the adverse effect. A final goal is to generate hypotheses about risk that may be explored in future studies.^[8]

It is appropriate to consider carefully the pooling strategy. It is more appropriate to pool studies of a similar design. One needs to consider the patient population, dosing regimens, duration of exposure, and study methods prior to pooling. Furthermore, pooling may not give a meaningful incidence rate if the incidence rates in the individual studies differ dramatically. The differences may be important predictors of the event and should not be ignored. An example of this may be a drug for which phototoxicity was observed. Outpatient studies may reflect the adverse event, while inpatient studies would not. Therefore they should not be combined. If the incidence rates are comparable, then pooling to get a more precise estimate is appropriate.^[8]

Once the pooling has been decided and appropriate patient samples have been obtained for estimating incidence, it is important to consider important predictive factors, such as dose, plasma level, duration of treatment, concomitant medications, age, sex, and concurrent illness, among others.^[8] These factors should be looked at in combination, as well. For example, elderly patients with severe disease may be more prone to a particular side effect, especially when treated simultaneously with a particular concomitant medication. These three-way interactions are difficult to detect and require large sample sizes, and it is exactly the type of information that the clinician is interested in.

CONCLUSION

During the design of a clinical program, careful attention should be paid to the uniform definition and collection of measurements that will facilitate a subsequent interpretation of the series of trials, particularly if the data are likely to be combined across trials. A common

dictionary for recording medication details, medical history, and adverse events should be selected. A common definition of primary and secondary variables is nearly always worthwhile for replication of results. The manner in which key variables are collected, the timing of assessments, the handling of protocol violations, and the definition of prognostic factors should all be kept compatible. To change these items from trial to trial makes it difficult to summarize them into a consistent story. Changes should be made only when there is a compelling reason to do so.^[2]

Integrated summaries are the ultimate conclusion for the benefit/risk ratio of the drug therapy in humans. These are the endproduct for a typically long period during which clinical studies are performed. Therefore it is important to plan the summaries early in the clinical development program. In this way, each clinical study can be planned so that it will fit clearly into the ISE and/or the ISS and will thus be easier to design. If we "begin with the end in mind,"^[9] the plan for the integrated summaries should be written in advance and should evolve as new results are obtained.

REFERENCES

1. In *Statistical Principles for Clinical Trials*. Draft 2C, International Conference on Harmonization, November 13, 1996.
2. In *Dose Response Information to Support Drug Registration; Guideline; Availability*. Notice, Federal Register (IV), International Conference on Harmonization 1994, 59 (216).
3. Pogue, J.; Yusuf, S. Overcoming the limitations of current meta-analysis of randomised controlled trials. *Lancet* **1998**, *351*, 915–916.
4. Flather, M.D. Strengths and limitations of meta-analysis: Larger studies may be more reliable. *Control. Clin. Trials* **1997**, *18*, 568–579.
5. Pogue, J.; Yusuf, S. Cumulating evidence from randomized trials: Utilizing sequential monitoring boundaries for cumulative meta-analysis. *Control. Clin. Trials* **1997**, *18*, 580–593.
6. Furberg, C.D.; Morgan, T.M. Lessons from overviews of cardiovascular trials. *Stat. Med.* **1987**, *6*, 295–303.
7. Egger, M.; Smith, G.D.; Phillips, A.N. Meta-analysis: Principles and procedures. *BMJ* **1997**, *315*, 1533–1537.
8. *Draft Guidance: Conducting a Clinical Safety Review of a New Product Application and Preparing a Report on the Review*; U.S. Department of Health and Human Services, FDA, CDER, 11/96.
9. Covey, S.R. *The Seven Habits of Highly Effective People*; Franklin-Covey: Provo, UT, 1986.

Intention-to-Treat Analyses (ITT)

Ronald K. Knickerbocker

Pfizer Global Research and Development, Ann Arbor, Michigan, U.S.A.

Abstract

Intention to treat (ITT) is a principle used for primary analyses in most randomized clinical trials testing new therapeutic agents. This entry breaks down the statistical and scientific issues concerning an intention-to-treat analyses.

INTRODUCTION

Randomization of therapy ensures that a therapy assignment is based on chance alone. This means that the baseline characteristics (measured and unmeasured) of the therapy groups should be comparable, with any differences due to chance. Thus, differences in trial outcomes among the groups should be due only to therapy assignment. Most statistical tests of significance require randomization as a fundamental principle. Therefore, randomization is the foundation for statistical inference of prospective clinical trials comparing therapy regimens.^[1]

Intention to treat (ITT) is a principle used for primary analyses in most randomized clinical trials testing new therapeutic agents. In a randomized clinical trial, subjects are randomly assigned to different therapies or regimens. The ITT principle requires that any comparison among therapy groups in a randomized clinical trial is based on results for all subjects in the therapy groups to which they were randomly assigned.^[2] This principle helps maintain the benefits of randomization. An alternative is to analyze only a subset of the subjects, such as those deemed to be compliant. However, the subset of compliant subjects may be different from the entire randomized population in ways that can lead to incorrect conclusions. The Coronary Drug Project provides an example of the possible bias introduced when looking only at the compliant.^[3]

If the randomized subjects in a clinical trial follow the therapy regimen assigned and complete the procedures specified in the protocol exactly, then the ITT analysis is the natural analysis for looking at all data from all randomized subjects. Unfortunately, in most randomized clinical trials, a combination of predictable and unpredictable circumstances leads to a subset of subjects who are unable to follow the intended protocol exactly. These circumstances include subjects entering the trial when ineligible, subjects who do not follow their assigned therapy regimen, subjects erroneously given or assigned the incorrect

therapy, subjects who refuse trial procedures, subjects taking disallowed concomitant medication, and subjects who terminate trial participation prior to completion of the trial. These factors (along with others) lead to subsets of randomized subjects at the end of the trial who may no longer be comparable or representative of the population under study.

REGULATORY REQUIREMENTS AND GUIDELINES

The International Conference on Harmonization (ICH) Guidelines on Statistical Principles for Clinical Trials^[4] contain extensive comments on ITT principles in Section 5.2, including the following discussion:

The intention-to-treat principle implies that the primary analysis should include all randomized subjects. Compliance with this principle would necessitate complete follow-up of all randomized subjects for study outcomes. In practice this ideal may be difficult to achieve, for reasons to be described. In this document the term “full analysis set” is used to describe the analysis set which is as complete as possible and as close as possible to the intention-to-treat ideal of including all randomized subjects. Preservation of the initial randomization in analysis is important in preventing bias and in providing a secure foundation for statistical tests. In many clinical trials the use of the full analysis set provides a conservative strategy. Under many circumstances, it may also provide estimates of treatment effects which are more likely to mirror those observed in subsequent practice.

There are a limited number of circumstances that might lead to excluding randomized subjects from the full analysis set, including the failure to satisfy major entry criteria (eligibility violations), the failure to take at least one dose of trial medication and the lack of any data postrandomization. Such exclusions should always be justified. Subjects who fail to satisfy an entry

criterion may be excluded from the analysis without the possibility of introducing bias only under the following circumstances:

1. The entry criterion was measured prior to randomization.
2. The detection of the relevant eligibility violations can be made completely objectively.
3. All subjects receive equal scrutiny for eligibility violations. (This may be difficult to ensure in an open-label study, or even in a double-blind study if the data are unblinded prior to this scrutiny, emphasizing the importance of the blind review.)
4. All detected violations of the particular entry criterion are excluded.

In some situations, it may be reasonable to eliminate from the set of all randomized subjects any subject who took no trial medication. The intention-to-treat principle would be preserved despite the exclusion of these patients provided, e.g., that the decision of whether or not to begin treatment could not be influenced by knowledge of the assigned treatment. In other situations, it may be necessary to eliminate from the set of all randomized subjects any subject without data postrandomization. No analysis is complete unless the potential biases arising from these specific exclusions, or any others, are addressed.

When the full analysis set of subjects is used, violations of the protocol that occur after randomization may have an impact on the data and conclusions, particularly if their occurrence is related to treatment assignment. In most respects, it is appropriate to include the data from such subjects in the analysis, consistent with the intention-to-treat principle. Special problems arise in connection with subjects withdrawn from treatment after receiving one or more doses who provide no data after this point, and subjects otherwise lost to follow-up, because failure to include these subjects in the full analysis set may seriously undermine the approach. Measurements of primary variables made at the time of the loss to follow-up of a subject for any reason, or subsequently collected in accordance with the intended schedule of assessments in the protocol, are valuable in this context; subsequent collection is especially important in studies where the primary variable is mortality or serious morbidity. The intention to collect data in this way should be described in the protocol. Imputation techniques, ranging from the carrying forward of the last observation to the use of complex mathematical models, may also be used in an attempt to compensate for missing data. Other methods employed to ensure the availability of measurements of primary variables for every subject in the full analysis set may require some assumptions about this subject's outcomes or a simpler choice of outcome (e.g., success/failure). The use of any of these strategies should be described and justified in the statistical section of the protocol, and the assumptions underlying any mathematical models employed should be clearly explained. It is also important to demonstrate the robustness of the corresponding results

of analysis especially when the strategy in question could itself lead to biased estimates of treatment effects.

Because of the unpredictability of some problems, it may sometimes be preferable to defer detailed consideration of the manner of dealing with irregularities until the blind review of the data at the end of the trial, and, if so, this should be stated in the protocol.

The Center for Drug Evaluation and Research (CDER) of the Food and Drug Administration (FDA) has released a Guidance for the Format and Content of the Clinical and Statistical Sections of an Application^[5] that discusses ITT analyses. The subsection on data sets analyzed (in the "Efficacy Results" section) contains the following:

Exactly which patients are included in the effectiveness analysis should be precisely defined, e.g., all patients with any effectiveness observation or with a certain minimum number of observations, only patients completing the trial, all patients with an observation during a particular time window, only patients with a specified degree of compliance, etc. It should be clear, if not defined in the study protocol, when, *relative* to study completion, and how, inclusion/exclusion criteria were developed. As a general rule, even if the applicant's preferred analysis is based on a reduced subset of the patients with data, there should be an additional "intent-to-treat" analysis using all randomized patients.

The subsection on demographic and baseline features of individual patients and comparability of treatment groups contains the following statement:

If the data sets in the "intent-to-treat" analysis and the applicant's preferred analysis are substantially different, comparability of treatment groups should be examined and any other reasons for such differences should be discussed.

In the section on statistical/analytical issues, the subsection on handling of dropouts or missing data states:

The results of a clinical trial should be assessed not only for the subset of patients who completed the study, but also for the entire patient population randomized (the intent-to-treat analysis). Several factors need to be considered and compared for the treatment groups in analyzing the effects of dropouts: the reasons for the dropouts, the time to dropout, and the proportion of dropouts among treatment groups at various time points.

Procedures for dealing with missing data, e.g., use of imputed data, should be described. Detailed explanation should be provided as to how such imputations were done and what underlying assumptions were made.

The FDA also provides guidance for some therapeutic areas that refer to preferences regarding ITT analyses. For example, the draft guidelines for preclinical and clinical evaluation of agents used in the prevention or treatment of postmenopausal osteoporosis^[6] states that "Plans for imputing missing data (both fracture follow-up assessment

and BMD data) should be explained in the protocol, and one or more supplementary intent-to-treat analyses using imputed data should be performed.” Such guidelines are an important resource for studies that might support the registration of a therapy being studied.

STATISTICAL ISSUES

When designing a study, sample size calculations should consider all aspects of the primary analyses. When ITT principles are used, calculation of sample size should include assumptions about the effect of factors such as noncompliance and crossover to other active therapies that are available. These factors will generally attenuate the assumed therapy effect and lead to the need for more subjects.^[7]

SCIENTIFIC ISSUES

It should be the goal of any randomized clinical trial to develop a protocol that can be reasonably followed by at least a large proportion of the subjects in all therapy groups.^[8] However, it is fairly common to see a degree of partial or complete noncompliance with the assigned therapy regimen (it is important to note that compliance may involve more than simply taking the medication, but also following specific instructions related to use [e.g., timing of multiple doses, relationship to food intake, or the need to remain upright]). In particular, subjects may not be able to tolerate side effects perceived to be related to therapy, and may stop taking medication or switch to other therapies. Subjects may also become noncompliant because of a perceived lack of efficacy of the therapy. It is also possible for a subject to be given a different therapy than that assigned, due to error; thus, subjects may take their study medication but still be noncompliant with the *assigned* therapy. In extreme cases, noncompliance may make study inference impossible.

In cases where subjects do not follow the protocol, a natural question is whether such subjects should be excluded from analysis because they do not seem to provide information relevant to the therapy regimens being studied. The ITT principle does not allow comparisons that exclude randomized subjects, regardless of their compliance with the therapy regimens and procedures of the trial, although practically such exclusions may be necessary in cases such as subjects missing all postbaseline measurements. This ensures that the randomization is protected (i.e., all groups have comparable baseline characteristics and that any differences besides therapy are because of chance), and precludes the possibility of bias as a result of selectively excluding subjects from therapy groups, which may lead to systematic differences among the groups that are attributed to factors other than therapy assignment.

In studies with an inactive control (e.g., placebo) or where the therapy under study is more effective than the control, the ITT analysis will usually lead to attenuated estimates of the true therapy, particularly when noncompliance is limited to subjects taking nothing or other therapies under study instead of the assigned therapy. In particular, ITT analyses will generally underestimate the effects of an efficacious therapy regimen. The ITT analysis will also generally underestimate any effects of a therapy on side effects. When active alternatives are available and are used by trial participants, effects on estimates of efficacy or side effect profile are generally not predictable and depend on factors such as the effectiveness of the alternative therapy and the characteristics of the noncompliant trial participants. It should also be noted that ITT analyses can increase variability, which might mask differences in equivalence trials.^[9]

It is intuitive to want to compare therapy groups based on the subjects who complete and are fully compliant with the protocol (completers or evaluable analysis).^[10] If subjects are removed from analysis in a completely random fashion, excluding them should have no effect on estimates of response or therapy effect. However, excluding nonrandomly selected subsets of subjects from an analysis may introduce bias in the results, even if membership in the subset to be excluded is equally distributed among the therapy groups, because the subset may have different characteristics than the randomized population. For example, it has been demonstrated that subjects compliant with study medication may respond better than noncompliant subjects in both treated and placebo groups.^[3] In the case where the subjects removed from analysis are not representative of the randomized populations, but are similar among therapy arms, estimates of therapy effects will usually be unbiased, but inference will no longer be generalizable to the intended study population.

Once therapy groups are compared based on subsets of subjects not solely defined by the randomization, it is impossible to show that the subset of subjects from which the inferences are being drawn are comparable. This leads to the potential introduction of bias and limits the generalizability of the trial.^[11] An exception to this rule is removing subjects found to be ineligible based on objective data that was known or could be determined in the same manner for all subjects prior to randomization.^[12] However, if the subject is found to be ineligible based on postrandomization data, it is possible that therapy assignment could have affected the availability of the information and, thus, that subject should be included in the ITT analysis.

Intention-to-treat analyses are often considered to be more generalizable to clinical practice. In particular, when subjects in a therapy group stop taking the drug or are not fully compliant with the regimen, the response rate of the therapy declines, as it would in clinical practice. In contrast, with most objective measures, not taking study medication should have minimal effect on placebo control groups, provided that the subjects are followed to study completion. However, this advantage of generalizability

can be lost when subjects are allowed to switch to or add on other effective therapies. In this situation, active therapy arms may still be reflective of what happens to subjects who initiate therapy in clinical practice, and should not have a large effect on the therapy arm if the other therapies are similar in efficacy to that being tested. However, the placebo group will generally have a higher observed response rate than what would be seen with a group treated with placebo alone. Therefore, effective therapies will be penalized by the additional active medication use in the placebo arm. This can lead to the need for larger studies to compensate for the attenuated therapy effect because of bias in the placebo arm. It should also be noted that such crossovers may lead to overestimates of the therapy effect when other medications available are more effective than the test therapy, particularly when more subjects in the test arm take the additional agent. Careful consideration should be given to minimizing the use of additional active agents in placebo-controlled clinical trials when alternative therapies (or the therapy being studied) are available.

Finally, it should be noted that some subjects end up with missing data even when efforts are made to follow all randomized subjects. For example, subjects may terminate participation in the study prior to completion, miss or refuse procedures, or move away from the clinical center. When no postbaseline information is collected for a randomized subject, the subject may be excluded or some value indicating "failure" may be imputed. In most cases, only some of the data are missing for a subject. Even in a strict ITT analysis, data will need to be imputed for these subjects. Many methods exist for imputing missing data. Common methods include carrying forward the last value for subjects leaving the study and treating those subjects as failures.^[13] Other authors have proposed methods for imputing endpoints based on earlier data obtained in the study.^[14]^[15] The primary method for imputation should be documented prior to unblinded data analysis. Several different imputation methods can be used to confirm the primary results, evaluate the sensitivity of the primary method used, and examine the robustness of the therapy effect.

CONCLUSION

Statisticians generally agree that following the ITT principle reduces bias and is appropriate for most trials. However, there are many important decisions that need to be considered during the planning phase of data analysis, including the strictness of the application of the principle; handling

of ineligible subjects who were erroneously randomized; biasing factors, such as noncompliance, concomitant medication use, and crossover to active medications; methods for imputing missing data; and the effect of the analysis on safety or other endpoints where equivalence might be expected. Karl Peace has written that a goal of clinical research in the development of new drugs is to design protocols and conduct clinical trials so that the evaluable-only and ITT analyses are the same.^[11]

REFERENCES

1. Byar, D.P.; Simon, R.M.; Friedewald, W.T.; Schlesselman, J.J.; Demets, D.L.; Ellenberg, J.H.; Gail, M.H.; Ware, J.H. Randomized clinical trials: Perspectives on some recent ideas. *N. Engl. J. Med.* **1976**, *295*, 74–80.
2. Gillings, D.; Koch, G. The application of the principle of intention-to-treat to the analysis of clinical trials. *Drug Inf. J.* **1991**, *25*, 411–424.
3. Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project. *N. Engl. J. Med.* **1980**, *303*, 1038–1041.
4. FDA. *Guidelines on Statistical Principles for Clinical Trials (E9)*; International Conference on Harmonization (ICH), 1998.
5. FDA. *Guideline for the Format and Content of the Clinical and Statistical Sections of an Application*; 1988.
6. FDA. *Draft Guidelines for Preclinical and Clinical Evaluation of Agents Used in the Prevention or Treatment of Postmenopausal Osteoporosis*; 1994.
7. Lee, Y.J.; Ellenberg, J.H.; Hirtz, D.G.; Nelson, K.B. Analysis of clinical trials by treatment actually received: Is it really an option? *Stat. Med.* **1991**, *10*, 1595–1605.
8. Friedman, L.M.; Furberg, C.D.; Demets, D.L. *Fundamentals of Clinical Trials*; Mosby-Year Book: St. Louis, 1996.
9. Ellenberg, J.H. *Encyclopedia of Biostatistics*; Armitage, P.; Colton, T., Eds.; Wiley: Chichester, 1998, 2056–2060.
10. Ellenberg, J.H. Intent-to-treat analysis versus as-treated analysis. *Drug Inf. J.* **1996**, *30*, 535–544.
11. Fisher, L.; Dixon, D.O.; Herson, J.; Frankowski, R.; Hearron, M.S.; Peace, K.E. *Statistical Issues in Drug Research and Development*; Peace, K., Ed.; Marcel Dekker: New York, 1990, 331–350.
12. Gail, M.H. Eligibility exclusions, losses to follow-up, removal of randomized patients and uncounted events in cancer clinical trials. *Cancer Treat.* **1985**, *69*, 1107–1113.
13. Gould, A.L. A new approach to the analysis of clinical drug trials with withdrawals. *Biometrics* **1980**, *36*, 721–727.
14. Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*; Wiley: New York, 1987.
15. Efron, B. Missing data and the bootstrap. *J. Am. Stat.* **1994**, *89*, 463–474.

Interchangeability: Generic Drugs

Jiayin Zheng and Shein-Chung Chow

Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, U.S.A.

Mengdie Yuan

Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

With more and more generics becoming available in the marketplace, the safety/efficacy concerns may arise as the result of interchangeably use of approved generics. However, bioequivalence assessment for regulatory approval among generics of the innovative drug product is not required. In practice, approved generics are often used interchangeably without any mechanism of safety monitoring. In this article, based on indirect comparisons, we introduce several methods to assessing bioequivalence and interchangeability between generics. The applicability of the methods and the similarity assumptions were discussed, as well as the inappropriateness of directly adopting adjusted indirect comparison to the field of generics' comparison. Besides, some extensions were given to take into consideration the important topics in clinical trials for bioequivalence assessments, e.g., multiple comparisons and simultaneously testing bioequivalence among three generics. Two real examples, the studies of malaria generics and HIV/AIDS generics prequalified by the WHO, were used to demonstrate the use of the method.

Keywords: Adjusted indirect comparison; Average bioequivalence; Fiducial probability; Geometric mean ratio; Indirect comparison.

INTRODUCTION

Average bioequivalence assessment for regulatory review and approval between a generic and a brand-name drug is well established and widely used in the pharmaceutical industry.^[1-5] Before a generic can be approved by the United States (US) Food and Drug Administration (FDA), the sponsor is required to conduct a bioequivalence study to demonstrate that the generic is bioequivalent to the brand-name drug in terms of the rate and extent of drug absorption. Under the Fundamental Bioequivalence Assumption, two drug products are considered therapeutic equivalent if they are shown to be bioequivalent in drug absorption profile. An approved generic can then be used as a substitute of the brand-name drug. In particular, the free interchangeability is generally assumed for small-molecule generics, resulting in the substitution of various generics in pharmacies. Since small-molecule drugs are generally safe and effective, these assumptions and practices usually work. However, important exceptions exist.

When several generics are bioequivalent to the same brand-name drug, it is not obvious that they are bioequivalent to each other. Anderson and Hauck^[6] demonstrated that although with two or three generics, the probability of such lack of bioequivalence is fairly low, the concern

becomes substantial with five, six, or more products on the market. As more and more generics become available in the marketplace, and the fact that approved generics may be used interchangeably, it is a concern whether (1) the quality of approved generics is consistent and (2) the interchangeability among the approved generics is safe and efficacious. For example, with respect to the pharmacokinetic (PK) mean responses, if an approved generic A falls on the bioequivalence lower end (say 80%) of the brand-name drug and another approved generic B is on the higher end (say 120%), the relative change in response of a subject who switches from A to B could lead to a drastic change (say 50%) in blood concentration that could potentially raise a safety concern. Also, strong safety concerns were repeatedly raised about the generic-to-generic substitution of, e.g., antiepileptics^[7,8] and immunosuppressant products.^[9] However, bioequivalence assessment among generics is not required by regulatory agencies such as the FDA,^[2,10] and there is often a lack of head-to-head comparative trials between all available generics of the same brand-name drug to ensure the interchangeability between them.

As the settings of the original bioequivalence trials are not identical, using the original data directly to test the bioequivalence between generics is not recommended. To deal with this difficulty, the following two approaches

may be useful: (1) calibrate the original data based on the summary statistics of the brand-name drug from both bioequivalence trials, and then analyze the calibrated data; and (2) use the results from different trials to estimate the relative bioavailabilities between generics through indirect comparison. The latter can be a practical way, since the original study data are always unavailable and only the summary data are often accessible. Following this idea, several approaches of indirect comparison in investigating the bioequivalence between generics have been employed and reviewed.^[11–16] Among them, the adjusted indirect comparison was considered by Garcia-Arieta et al.^[12] as the simplest and most suitable method for bioequivalence studies. The adjusted indirect comparison is a useful tool for assessing the bioequivalence between two generics of the same brand-name drug, and partly preserves the power of randomized controlled trials. Nonetheless, it requires the original studies be sufficiently powered (>80%) and the point estimate difference between generics not exceed the 7% difference.^[12] To get a better performance, Garcia-Arieta et al.^[12] also extended the acceptance limit from $\pm 20\%$ to $\pm 30\%$, which seems arbitrary.

For clinical practice, the bioequivalence and interchangeability between generics make sense only if the condition that both generics are bioequivalent to the corresponding brand-name drug holds. However, directly adopting the adjusted indirect comparison does not take this condition into consideration. In view of this, the restricted confidence interval of logarithmic geometric mean ratio (GMR) for two generics was proposed,^[17] under the framework of indirect comparisons with constraints. Furthermore, the fiducial probability^[18] of bioequivalence between generics can be derived. In this article, we focused on bioequivalence testing of a basket of generics and proposed a normality-based method to screen average bioequivalence pairs of generics with certain assurance. Assume that there are two generics, denoted by G_A and G_B , respectively, which have been shown to be bioequivalent to the same brand-name drug (denoted by B_R). The available data are from two bioequivalence trials, one comparing G_A and B_R and the other one comparing G_B and B_R . The same criterion used for testing the average bioequivalence of a generic and the brand-name drug is adopted to address the testing of average bioequivalence between two generics. That is, if the 90% confidence interval of the GMR between two generics falls into a prespecified interval, say (80.00%, 125.00%), we claim that the two generics are average bioequivalent. The confidence interval of GMR is constructed based on the 90% confidence intervals from the original bioequivalence studies under normality assumption.

In the next section, we briefly discussed the adjusted indirect comparison for testing bioequivalence between generics, outlined the method of fiducial probability, and provided a detailed description under the conventional two-sequence, two-period crossover design. In Section “Some Extensions,” some extensions for the fiducial

probability method including multiple comparisons were presented. Then Section “Similarity Assumptions and Determining the Bioequivalence Limits” discussed the important assumptions of similarity and the determination of the bioequivalence limits. Two examples of malaria generics and HIV/AIDS generics were provided to demonstrate the use of the method in Section “Real Example Analysis.” In Section “Summary,” some concluding remarks were given.

RESTRICTED CONFIDENCE INTERVALS OF GMR BETWEEN GENERICS

In this section, we discuss the inappropriateness of directly applying adjusted indirect comparison to testing generics’ bioequivalence, and then introduce the fiducial probability method of restricted confidence intervals of GMR between two generics, using the normality-based statistics from two independent bioequivalence trials. The applicability of this method depends on similarity assumptions, which will be discussed in Section “Similarity Assumptions and Determining the Bioequivalence Limits.” To better illustrate the method, we provided detailed solutions under 2×2 crossover design without carryover effect in the [Appendix](#).

Denote (L_A, U_A) and (L_B, U_B) as the $1 - 2\alpha$ confidence intervals of $\mu_A - \mu_R$ and $\mu_B - \mu_R$, respectively, where μ_A, μ_B , and μ_R are the true logarithmic geometric means of G_A, G_B , and B_R , respectively. (L_A, U_A) and (L_B, U_B) are obtained from two independent bioequivalence trials with normality-based statistical method. Assume that the two trials are similar so that the relative effect estimated by the trial of G_A versus B_R is generalizable to subjects in the trial of G_B versus B_R , and the relative effect estimated by the trial of G_B versus B_R is generalizable to subjects in the trial of G_A versus B_R . Denote $(\delta_L, \delta_U) = (\log(0.8), \log(1.25))$ as the bioequivalence limits. The approval of both generics requires that both (L_A, U_A) and (L_B, U_B) fall within (δ_L, δ_U) . Based on (L_A, U_A) , (L_B, U_B) , the statistical methods used for obtaining these two confidence intervals, and the similarity assumptions, we construct the restricted confidence intervals of $\mu_A - \mu_B$ by indirect comparison.

Adjusted Indirect Comparison for Generics’ Bioequivalence

The method of adjusted indirect comparison can be described as follows: in the trial of G_A versus B_R , assume that $\mu_A - \mu_R$ was estimated as $\hat{\nu}_{AR}$. Similarly, in the trial of G_B versus B_R , $\mu_B - \mu_R$ was estimated as $\hat{\nu}_{BR}$.

Under the similarity assumption, based on indirect comparison, we get a point estimator of $\mu_A - \mu_B$ as $\hat{\nu}_{AR} - \hat{\nu}_{BR}$, of which the variance was estimated by Var_{AB} . Under the normality assumption, the $1 - 2\alpha$ confidence interval of $\mu_A - \mu_B$ for adjusted indirect comparison can be expressed as

$$\hat{v}_{AR} - \hat{v}_{BR} \pm z_{\alpha} / t_{\alpha}(df) \cdot \sqrt{\widehat{Var}_{AB}}, \quad (1)$$

where $z_{\alpha} / t_{\alpha}(df)$ is the α quantile of standard normal distribution or that of the Student's t -distribution with the degree of freedom df .

From the viewpoint of Gwaza et al.,^[13] among the available methods for performing indirect comparisons, the adjusted indirect comparison is the simplest and most suitable method for bioequivalence studies, because it uses publicly available data, and partly preserves the power of randomized controlled trials. For clinical practice, the bioequivalence and interchangeability between generics make sense only if both generics are bioequivalent to the corresponding brand-name drug. Directly adopting adjusted indirect comparison may not be appropriate since this method derives the confidence interval of $\mu_A - \mu_B$, ignoring the fact that the confidence intervals for $\mu_A - \mu_R$ and $\mu_B - \mu_R$ are contained with (δ_L, δ_U) , resulting in a narrower confidence interval and overestimation of clinical meaningful bioequivalence between generics. To mitigate this limitation, constraints on the indirect comparison were imposed and the fiducial probability type of restricted confidence intervals of $\mu_A - \mu_B$ was proposed,^[17] under the framework of indirect comparisons with constraints.

Restricted Confidence Intervals of $\mu_A - \mu_B$

In this section, we present the fiducial probability type of restricted confidence intervals of $\mu_A - \mu_B$, which is based on indirect comparisons under the constraint of bioequivalence between the generics and the brand-name drug.

Without loss of generality, in the trial of G_A versus B_R , assume that $\mu_A - \mu_R$ was estimated as $\hat{v}_{AR}$, of which the variance was estimated by $\widehat{Var}(\hat{v}_{AR})$. (L_A, U_A) can be expressed as $\hat{v}_{AR} \pm t_{\alpha}(df_A) \sqrt{\widehat{Var}(\hat{v}_{AR})}$, where $t_{\alpha}(df_A)$ is the α quantile of the Student's t -distribution with the degree of freedom df_A . Note that $t_{\alpha}(df_A)$ is the α quantile of standard normal distribution when $df_A = \infty$. Similarly, in the trial of G_B versus B_R , $\mu_B - \mu_R$ was estimated as $\hat{v}_{BR}$, of which the variance was estimated by $\widehat{Var}(\hat{v}_{BR})$, and $(L_B, U_B) = \hat{v}_{BR} \pm t_{\alpha}(df_B) \sqrt{\widehat{Var}(\hat{v}_{BR})}$. Then we obtain the fiducial distribution of $\mu_l - \mu_R$ as $\hat{v}_{lR} \pm t(df_l) \sqrt{\widehat{Var}(\hat{v}_{lR})}$, where $l=A$ or B , and $t(df_l)$ is the Student's t -distribution with the degree of freedom df_l . Denote f_l as the probability density function of the fiducial distribution of $\mu_l - \mu_R$. It is reasonable to assume that f_A and f_B are statistically independent. Based on f_A and f_B , calculate the following probability:

$$pr\{\delta_L \leq \mu_l - \mu_R \leq \delta_U, l = A, B \text{ and } -\Delta \leq \mu_A - \mu_B \leq \Delta\},$$

which is denoted as q_{Δ} . When $\Delta = 2\delta_U, q_{2\delta_U}$ is equal to $pr\{\delta_L \leq \mu_l - \mu_R \leq \delta_U, l = A, B\}$, since $\{\delta_L \leq \mu_l - \mu_R \leq \delta_U,$

$l = A, B\}$ is a subset of $\{2\delta_L \leq \mu_A - \mu_B \leq 2\delta_U\}$. For any q that $0 < q \leq q_{2\delta_U}$, we can find the minimal Δ satisfying

$$pr\{\delta_L \leq \mu_l - \mu_R \leq \delta_U, l = A, B \text{ and } -\Delta \leq \mu_A - \mu_B \leq \Delta\} \geq q.$$

Denote this minimal Δ by Δ_q . Then we get a restricted confidence interval of $\mu_A - \mu_B$ as $(-\Delta_q, \Delta_q)$ with the confidence level of q . Denote (δ'_L, δ'_U) ($\delta'_L = -\delta'_U$) as the target bioequivalence limits for $\mu_A - \mu_B$. Given the confidence level of $1 - 2\beta$, if the resulting $\Delta_{1-2\beta} \leq \delta'_U$, the clinical meaningful bioequivalence between G_A and G_B can be concluded. In other words, in this case, we have

$$pr\{\delta_L \leq \mu_l - \mu_R \leq \delta_U, l = A, B \text{ and } \delta'_L \leq \mu_A - \mu_B \leq \delta'_U\} \geq 1 - 2\beta.$$

Conversely, we can simply calculate the fiducial probability $pr\{\delta_L \leq \mu_l - \mu_R \leq \delta_U, l = A, B \text{ and } \delta'_L \leq \mu_A - \mu_B \leq \delta'_U\}$ (denoted as $q_{\delta'_U}$) and then compare $q_{\delta'_U}$ with the prespecified confidence level. If $q_{\delta'_U}$ is greater than the prespecified confidence level, we conclude that the clinical meaningful bioequivalence between G_A and G_B holds.

The fiducial probability can be expressed as

$$\int_{\delta_L}^{\delta_U} \int_{\delta_L \vee (x + \delta'_L)}^{\delta_U \wedge (x + \delta'_U)} f_A(x) f_B(y) dy dx, \quad (2)$$

and numerical integration can be used to get an accurate estimation.

SOME EXTENSIONS

In this section, we give some extensions of the fiducial probability method: getting p value to accommodate multiple comparison, a natural extension to the comparison of three generics.

Multiple Comparisons

We may encounter the issue of multiple comparisons (or multiple testing) when conducting the pairwise comparisons. In multiple comparisons, usually we adjust the significance level in some way, so that the probability of observing at least one significant result due to chance remains below the desired significance level. Specifically in practical applications, the family-wise error rate (FWER) and false discovery rate (FDR)^[19] are often considered and controlled with appropriate methods.

FWER ensures that the probability of committing a single false rejection is bounded by α . It focuses on a "family" of statistical tests and is appropriate when a single false positive in this family is a problem. The Bonferroni correction is the most common way to achieve this, which sets the significance cut-off at α / m (here m is the number of comparisons). The Bonferroni correction tends to be too

conservative, leading to a high rate of false negative or lower power, especially when the tests are correlated or m is large.

FDR is defined as the proportion of actual false positive among all discoveries (significant results).^[20] The Benjamini–Hochberg procedure^[19] is a powerful technique for controlling the FDR. It works by estimating some rejection region so that on average, $FDR < \alpha$, where α is a prespecified target value of FDR. This procedure can be briefly described as follows: put all p -values in order, from smallest to largest; the smallest one has a rank of $i = 1$, then the next smallest has $i = 2, 3, \dots$; compare each individual p -value to its own critical value, $(i\alpha)/m$, where m is the total number of tests; denote the largest p -value that satisfies $p < (i\alpha)/m$ by P_{max} ; all p -values no larger than P_{max} are significant. Although independence of test statistics was assumed in the work of Benjamini and Hochberg,^[19] addressing positive dependence by Benjamini and Yekutieli^[21] essentially assured users that this simple procedure was safe to use in many situations arising in practice, and the modification to general dependence is often not needed.^[20] Therefore, the Benjamini–Hochberg procedure is still expected to perform well in this article's case which is under the pairwise comparison setting, a specific situation of dependence. The prespecified value of the FDR should be chosen cautiously. In bioequivalence studies for generics, missing out some pairs that are actually bioequivalent will not lead to serious consequences. On the other hand, the cost of getting a false positive (leading to safety/efficacy issues) can be huge. Thus, in the case of this article, controlling a low FDR (say, 0.1) is recommended. Another approach proposed by Storey,^[22,23] the positive FDR (pFDR), deals with FDR from a different philosophy and in an opposite way. It works by first fixing the significance region, and then estimating the corresponding FDR α .

In the context of this article, we recommend using the Benjamini–Hochberg procedure for controlling the FDR. The target value of the FDR should be small. Sometimes, a “Benjamini–Hochberg adjusted p -value” is used. The adjusted p -value for a test is either the raw p -value times m/i or the adjusted p -value for the next higher raw p -value, whichever is smaller. If the adjusted p -value is smaller than the target FDR, the test is significant. In the real examples of the next section, this Benjamini–Hochberg adjusted p -value was used for controlling the FDR.

No matter which method is adopted, p -value in each single test is required for multiple testings. From Section “Restricted Confidence Intervals of $\mu_A - \mu_B$,” the p -value for the method of fiducial probability can be expressed as

$$1 - q_{\delta'_U} = 1 - pr\{\delta_L \leq \mu_l - \mu_R \leq \delta_U, l = A, B \\ \text{and } \delta'_L \leq \mu_A - \mu_B \leq \delta'_U\}$$

Simultaneous Comparison of Three Generics

We also considered the extension to accommodate the comparison of more than two generics simultaneously. In practice, other than pairwise comparison, the comparison of a basket of generics as a whole may arise. Here we consider testing average bioequivalence of three generics as a whole in terms of geometric mean. Denote G_A , G_B , and G_C as the three generics corresponding to the same brand-name drug B_R , with their logarithmic geometric means denoted by μ_A, μ_B, μ_C , and μ_R . Denote (L_A, U_A) , (L_B, U_B) , and (L_C, U_C) as three confidence intervals of logarithmic geometric means from three bioequivalence trials of G_A versus B_R , G_B versus B_R , and G_C versus B_R , respectively.

Denote f_l as the probability density function of the fiducial distribution of $\mu_l - \mu_R$. The fiducial probability $P\{\mu_l - \mu_R \in (\delta'_L, \delta'_U), l = A, B, C \text{ and } \mu_A - \mu_B, \mu_A - \mu_C, \mu_B - \mu_C \in (\delta'_L, \delta'_U)\}$ (denoted as $q_{\delta'_U}$) can be expressed as

$$\int_{\delta'_L}^{\delta'_U} \int_{\delta'_L \vee (x + \delta'_L)}^{\delta'_U \wedge (x + \delta'_U)} \int_{\delta'_L \vee (x + \delta'_L) \vee (y + \delta'_L)}^{\delta'_U \wedge (x + \delta'_U) \wedge (y + \delta'_U)} f_A(x) f_B(y) f_C(z) dz dy dx.$$

SIMILARITY ASSUMPTIONS AND DETERMINING THE BIOEQUIVALENCE LIMITS

Similarity Assumption

The methods for indirect comparison should be applied cautiously. Both clinical similarity and methodological similarity should be considered when using the methods of restricted confidence interval based on indirect comparison.^[11] First, the important conditions of the trials that could bias the estimated effect should be comparable. Patients should be similar enough in the bioequivalence trials under comparison, such that the relative effect estimated by the trial of G_A versus B_R is generalizable to patients in the trial of G_B versus B_R and vice versa. For PK/pharmacodynamic bioequivalence research, patient characteristics, the mode of drug administration, and parameter measurement should be alike between the trials involved. Second, methodological quality of the trials should be sufficiently similar. If the trials for comparison are similarly biased, it is easy to mathematically prove that the results of indirect comparison are unbiased.^[12] In addition, trials with different designs might not be comparable as the formulation effect might not be expected to be the same between them. However, Garcia-Arieta et al.^[12] considered results from conventional 2×2 crossover designs and replicate designs as combinable.

Compared with indirect comparison of efficiency trials, confidence in the clinical similarity and methodological quality of bioequivalence studies is often ensured since the basic designs of such studies are generally consistent.^[12,16]

For example, in bioequivalence trials, the participants' characteristics are frequently alike that healthy adult volunteers within the age of 18–55 years are commonly defined. Besides, randomization of subjects in the allocation of sequence are always guaranteed.

Determining the Bioequivalence Limits

Unlike the bioequivalence trials between generics and the brand-name drug that are well designed with sufficient power, indirect comparison between generics often has larger variability and reduced precision, leading to a low level of power where the bioequivalence limits are set to be the same as in the bioequivalence trial (say ($\log(0.80)$, $\log(1.25)$)). In order to enlarge the power, as recommended by Garcia-Arieta et al.,^[12–14] a slightly wider bioequivalence interval (say ($\log(0.70)$, $\log(1.00/0.70)$)) may be used for indirect comparison, as appropriate. However, cautions should be taken when relaxing the bioequivalence limits as the type I error may not be controlled below a prespecified level.

REAL EXAMPLE ANALYSIS

In this section, we used the method of fiducial probability with $\delta'_U = \delta_U$ to investigate the relative bioavailability of different generics. Two examples were illustrated to demonstrate the use of the fiducial probability method.

Malaria Generics

As in Gwaza et al.,^[13] we selected the data from three bioavailability/bioequivalence studies conducted independently, which are available in WHO Public Assessment Reports (WHOPARs) at <https://extranet.who.int/prequal/> (<https://extranet.who.int/prequal/sites/default/files/documents/MA052part6v1.pdf>, <https://extranet.who.int/prequal/sites/default/files/documents/MA062part6v1.pdf>, and <https://extranet.who.int/prequal/sites/default/files/documents/MA064part6v2.pdf>). The studies compared fixed dose combination (FDC) artemether/lumefantrine 20/120 mg tablets in adult healthy volunteers with the same brand-name drug submitted to the WHO Prequalification of Medicines Programme. These generic tablets were all compared with the brand-name drug, the WHO-approved artemether/lumefantrine 20/120 mg tablets (Coartem® / Riamet®) from Novartis Pharma (Basel, Switzerland), and had been prequalified by the WHO. The three studies enrolled 55, 64, and 58 adult males, respectively, and conducted single-center, open-label, randomized, two-period, two-treatment, two-sequence, crossover studies under non-fasting conditions. Two types of measures, the area under the time–concentration curve (AUC_{0-t} and AUC_{0-inf}) and the drug peak concentration (C_{max}), were

investigated for bioequivalence between generics. In the study performed by Cipla, an AUC truncated at 72 h was reported. Therefore, three potential pairs of generics were tested for bioequivalence, resulting in 18 comparisons after multiplying the six measures (3/3 for artemether/lumefantrine). We adjusted the p -value to control the target FDR using the Benjamini–Hochberg procedure.

Table 1 listed the 90% confidence intervals of PK parameters of artemether and lumefantrine from those reports. Based on these confidence intervals and the same size in each study, we calculated the fiducial probability for each pairwise comparison and the Benjamini–Hochberg adjusted p -value. The results were shown in the same table. Using the fiducial probability method with $\delta'_U = \delta_U = \log(125.00\%)$ and the target FDR of 0.1, all comparisons were significant and thus considered as bioequivalent except three pairs for C_{max} : Ajanta versus Ipca, Ipca versus Cipla of artemether, and Ajanta versus Cipla of lumefantrine. The results were consistent with those obtained by adjusted indirect comparison method with the conventional 80%–125% acceptance range in the work of Gwaza et al.^[13] Although three pairs of comparisons were not significant, two of them would be significant if the target FDR was set to be 0.2. The nonsignificant results might be partially attributed to the reduced precision of the fiducial probability method.

HIV/AIDS Generics

Combivir® (Lamivudine/Zidovudine) is an antiviral medication containing a combination of 150 mg lamivudine and 300 mg zidovudine. It belongs to reverse transcriptase inhibitors and helps keep the HIV virus from reproducing in the human body. Tablets are indicated for the treatment of HIV-1 infection in combination with at least one other antiretroviral agent. Several bioequivalence studies have been conducted to compare generic lamivudine/zidovudine 150/300 mg tablet with the brand-name drug Combivir® (Lamivudine/Zidovudine) 150/300 mg tablet. We selected three studies that were sufficiently similar to investigate the bioequivalence between generics, using the same method as in the previous example. The data were obtained from WHOPAR available at <https://extranet.who.int/prequal/> (<https://extranet.who.int/prequal/sites/default/files/documents/HA286part6v2.pdf>, <https://extranet.who.int/prequal/sites/default/files/documents/HA291part6v1.pdf>, and <https://extranet.who.int/prequal/sites/default/files/documents/HA392part6v1.pdf>). Each study was randomized, open-label, two-treatment, two-period, two-sequence, single-dose, and crossover. They were all conducted in healthy human adult subjects under fasting conditions performed between 2005 and 2007. The number of subjects that completed the study and were utilized for analysis to establish PK parameters and to assess bioequivalence were

Table 1 Results for the bioequivalence tests of PK parameters of artemether and lumefantrine in FDC generics

		The 90% confidence intervals for the ratio of PK parameters		
		Ajanta	Ipca	Cipla
Artemether	C_{max}	86.3–99.1	101.0–118.9	86.7–103.7
	AUC_{0-t}	93.3–104.7	102.4–116.6	90.1–106.9
	AUC_{0-inf}	93.5–104.6	102.8–117.0	90.1–106.9
Lumefantrine	C_{max}	80.3–95.4	87.6–99.5	89.3–108.2
	AUC_{0-t}	82.1–97.3	87.3–103.2	88.7–108.1
	AUC_{0-inf}	82.5–97.8	87.6–103.0	88.7–108.1
		The fiducial probability with $\delta'_U = \delta_U = \log(1.25)$ and the Benjamini–Hochberg adjusted p -value		
		Ajanta vs Ipca	Ajanta vs Cipla	Ipca vs Cipla
Artemether	C_{max}	0.795 (0.2048)*	0.996 (0.0238)	0.857 (0.1517)*
	AUC_{0-t}	0.989 (0.0255)	0.999 (0.0055)	0.962 (0.0623)
	AUC_{0-inf}	0.989 (0.0255)	0.999 (0.0055)	0.957 (0.0623)
Lumefantrine	C_{max}	0.955 (0.0623)	0.894 (0.1197)*	0.992 (0.0255)
	AUC_{0-t}	0.975 (0.0459)	0.943 (0.0688)	0.991 (0.0255)
	AUC_{0-inf}	0.980 (0.0405)	0.950 (0.0642)	0.992 (0.0255)

*Not significant by the type I method with the target FDR of 0.1.

62, 31, and 43, respectively. Three bioavailability characteristics C_{max} , AUC_{0-t} , and AUC_{0-inf} for bioequivalence evaluation were taken. Table 2 summarizes the results and the 90% confidence intervals from the reports. All comparisons were significant and thus were taken as bioequivalent except the pair of Strides versus Matrix for C_{max} of Zidovudine, with an adjusted p -value of 0.1344. The adjusted p

value of that pair was slightly larger than the target value of 0.1, partially due to the insufficient sample sizes of the original studies. The results indicated that those WHO pre-qualified Lamivudine/Zidovudine 150/300mg products are not only bioequivalent with the brand-name drug Combivir® in terms of AUC and C_{max} , but also bioequivalent between themselves with respect to AUC and C_{max} .

Table 2 Results for the bioequivalence tests of PK parameters of Lamivudine and Zidovudine in FDC generics

		The 90% confidence interval for the ratio of PK parameters		
		Ranbaxy	Strides	Matrix
Lamivudine	C_{max}	96.8–109	90.63–110.52	82.4–101.7
	AUC_{0-t}	97.8–105	96.31–108.29	88.0–100.6
	AUC_{0-inf}	98.2–105	96.42–107.92	88.8–100.7
Zidovudine	C_{max}	85.5–108	81.15–112.51	83.8–114.1
	AUC_{0-t}	93.7–102	92.77–105.41	97.3–109.7
	AUC_{0-inf}	93.7–102	92.63–105.15	97.4–109.5
		The fiducial probability with $\delta'_U = \delta_U = \log(1.25)$ and the Benjamini–Hochberg adjusted p -value		
		Ranbaxy vs Strides	Ranbaxy vs Matrix	Strides vs Matrix
Lamivudine	C_{max}	0.996 (0.0057)	0.928 (0.0834)	0.929 (0.0834)
	AUC_{0-t}	1.000 (0.0000)	0.999 (0.0016)	0.995 (0.0072)
	AUC_{0-inf}	1.000 (0.0000)	1.000 (0.0010)	0.997 (0.0041)
Zidovudine	C_{max}	0.909 (0.0966)	0.926 (0.1344)*	0.886(0.1344)*
	AUC_{0-t}	1.000 (0.0001)	1.000 (0.0004)	0.999 (0.0012)
	AUC_{0-inf}	1.000 (0.0001)	1.000 (0.0004)	0.999 (0.0012)

*Not significant by the type I method with the target FDR of 0.1.

SUMMARY

The article discussed the issue of bioequivalence and interchangeability between generics that had been shown to be bioequivalent to the same brand-name drug, based on average bioequivalence in terms of GMR. Indirect comparison is a valuable way to address it. However, the recommended adjusted indirect comparison may be inappropriate from the aspect of clinical meaningful values for the parameters. The method of fiducial probability takes the clinical meaningful value into account. Extensive simulation studies not shown in this article demonstrated that this method has more power than the adjusted indirect comparison, and it controls the overall type I error below a prespecified level. The passing rate (power) is high when the true GMR is close to 1 and data noises are not large. The passing rate decreases when the true GMR departs away from 1 and approaches the bioequivalence limits. With a moderate GMR, e.g., 0.9, the method has a satisfactory power when the original studies were adequately powered. The method can be extended to simultaneous comparison of three generics and multiple testings. The applicability of the fiducial probability method requires the similarity assumptions, which should be cautiously examined in advance.

DISCLAIMER

This article has been reviewed by the FDA and determined not to be consistent with the Agency's views or policies. It reflects only the views and opinions of the authors.

Appendix: Under 2×2 Crossover Design without Carryover Effects

Let Y_{ijk} be the log transformation of the PK response of interest of the i th subject in the j th period and k th sequence of the trial. As suggested by the FDA, the following statistical model is useful in describing Y_{ijk} :

$$Y_{ijk} = \mu + F_l + P_j + Q_k + S_{ikl} + e_{ijk},$$

where μ is the overall mean; P_j is the fixed effect of the j th period, where $j = 1, 2$, and $\sum_{j=1}^2 P_j = 0$; Q_k is the fixed effect of the k th sequence, where $k = 1, 2$, and $\sum_{k=1}^2 Q_k = 0$; F_l is the direct fixed effect of the l th drug formulation ($FT + FR = 0$) when $j = k$ ($l = T$, the test formulation; otherwise, $l = R$, the reference (brand-name) formulation); S_{ikl} is the random effect of the i th subject in the k th sequence under drug formulation l ; and $S_{ik} = (S_{ikT}, S_{ikR})$, $i = 1, \dots, n_k$, $k = 1, 2$, are independently and identically distributed (*i.i.d*) bivariate normal random vectors with mean (0,0) and an unknown variance covariance matrix

$$\begin{pmatrix} \sigma_{BT}^2 & \rho^\sigma BT^\sigma BR \\ \rho^\sigma BT^\sigma BR & \sigma_{BR}^2 \end{pmatrix}.$$

e_{ijk} 's are the independent random errors distributed as $N(0, \sigma_{wl}^2)$, and S_{ik} 's and e_{ijk} 's are independent. Note that σ_{BT}^2 and σ_{BR}^2 are between-subject variances and σ_{WT}^2 and σ_{WR}^2 are within-subject variances, and that $\sigma_{TT}^2 = \sigma_{BT}^2 + \sigma_{WT}^2$ and $\sigma_{TR}^2 = \sigma_{BR}^2 + \sigma_{WR}^2$ are called total variances for the test and reference formulations, respectively.

The average bioequivalence index is $\nu = F_T - F_R = \mu_T - \mu_R$. Let $\bar{y}_{jk}$ be the sample average of the observations in the j th period and the k th sequence. Under the assumed statistical model, $\bar{y}_{11} - \bar{y}_{21} \sim N\left(\nu + P_1 - P_2, \frac{\tau^2}{n_1}\right)$ and $\bar{y}_{12} - \bar{y}_{22} \sim N\left(-\nu + P_1 - P_2, \frac{\tau^2}{n_2}\right)$, where $\tau^2 = \sigma_{BT}^2 + \sigma_{BR}^2 - 2\rho^\sigma BT^\sigma BR + \sigma_{WT}^2 + \sigma_{WR}^2 = \sigma_{TT}^2 + \sigma_{TR}^2 - 2\rho^\sigma BT^\sigma BR$. Consequently, we have the point estimators of ν and τ^2 as $\hat{\nu} = \frac{1}{2}(\bar{y}_{11} - \bar{y}_{21} - \bar{y}_{12} + \bar{y}_{22})$ and $\hat{\tau}^2 = \frac{(n_1 - 1)s_{D1}^2 + (n_2 - 1)s_{D2}^2}{n_1 + n_2 - 2}$, where s_{Dk}^2 is the sample variance based on the differences $\{y_{i1k} - y_{i2k}, i = 1, \dots, n_k\}$, $k = 1, 2$.

Denote $c = \frac{1}{4}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$, the point estimators satisfy: $\hat{\nu} \sim N(\nu, c\tau^2)$ and $\frac{(n_1 + n_2 - 2)\hat{\tau}^2}{\tau^2} \sim \chi^2(n_1 + n_2 - 2)$. Then, the $1 - 2\alpha$ confidence lower limit and upper limit of ν are given as $L = \hat{\nu} - Z_\alpha \sqrt{c\hat{\tau}^2}$, $U = \hat{\nu} + Z_\alpha \sqrt{c\hat{\tau}^2}$, where $Z_\alpha = t_\alpha(n_1 + n_2 - 2)$ is the α quantile of t -distribution with degree of $n_1 + n_2 - 2$.

Now assume that both bioequivalence trials of G_A versus B_R and G_B versus B_R are 2×2 crossover designs without carryover effects. The sample sizes of the two trials were n_A and n_B , respectively. (L_A, U_A) and (L_B, U_B) are the $1 - 2\alpha$ confidence intervals of $\mu_A - \mu_R$ and $\mu_B - \mu_R$, respectively. Denote the estimators of (ν, τ) by $(\hat{\nu}_A, \hat{\tau}_A)$ for G_A and $(\hat{\nu}_B, \hat{\tau}_B)$ for G_B . Thus, we have

$$\begin{aligned} \hat{\nu}_A &= \frac{(L_A + U_A)}{2}, \\ \sqrt{c_A \hat{\tau}_A^2} &= \frac{U_A - L_A}{-2 \times t_\alpha(n_A - 2)}, \\ \hat{\nu}_B &= \frac{(L_B + U_B)}{2}, \\ \sqrt{c_B \hat{\tau}_B^2} &= \frac{U_B - L_B}{-2 \times t_\alpha(n_B - 2)}. \end{aligned}$$

Denote $\sqrt{c_A \hat{\tau}_A^2}$ and $\sqrt{c_B \hat{\tau}_B^2}$ by S_A and S_B , respectively. To test the clinically meaningful average bioequivalence between G_A and G_B based on the restricted confidence interval with fiducial probability, first we obtain the fiducial distributions of $\mu_A - \mu_R$ and $\mu_B - \mu_R$. Taking $(\hat{v}_A, \hat{v}_B, S_A, S_B)$ as constants, we have $\mu_A - \mu_R \sim \hat{v}_A + t(n_A - 2)S_A$ and $\mu_B - \mu_R \sim \hat{v}_B + t(n_B - 2)S_B$, where $t(n_A - 2)$ and $t(n_B - 2)$ are the Student's t -distribution with the degrees of freedom $n_A - 2$ and $n_B - 2$, respectively. The fiducial probability $q_{\delta'_U} = \text{pr}\{\delta'_L \leq \mu_l - \mu_R \leq \delta'_U, l = A, B \text{ and } \delta'_L \leq \mu_A - \mu_B \leq \delta'_U\}$ can be expressed as

$$\begin{aligned} & \int_{\delta'_L}^{\delta'_U} \int_{\delta'_L \vee (x + \delta'_L)}^{\delta'_U \wedge (x + \delta'_U)} \frac{1}{S_A S_B} f_A\left(\frac{x - \hat{v}_A}{S_A}\right) f_B\left(\frac{y - \hat{v}_B}{S_B}\right) dy dx \\ &= \int_{\delta'_L}^{\delta'_U} \frac{1}{S_A} f_A\left(\frac{x - \hat{v}_A}{S_A}\right) \cdot \left[F_B\left(\frac{(\delta'_U \wedge (x + \delta'_U)) \vee \delta'_L - \hat{v}_B}{S_B}\right) \right. \\ & \quad \left. - F_B\left(\frac{(\delta'_L \vee (x + \delta'_L)) \wedge \delta'_U - \hat{v}_B}{S_B}\right) \right] dx, \end{aligned}$$

where f_A and f_B are the probability density functions of $t(n_A - 2)$ and $t(n_B - 2)$, respectively, and F_B is the cumulative distribution function of $t(n_B - 2)$. We can obtain $q_{\delta'_U}$ by numerical integration. If $q_{\delta'_U}$ is greater than the prespecified confidence level, we conclude that generic G_A and generic G_B are average bioequivalent in terms of geometric mean.

REFERENCES

1. Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*, 3rd Ed.; CRC Press: Boca Raton, FL, 2009.
2. FDA. *Guidance for Industry: Statistical Approaches to Establishing Bioequivalence*; Center for Drug Evaluation and Research, U.S. Food and Drug Administration: Rockville, MD, 2001.
3. Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* **1987**, *15* (6), 657–680. doi: 10.1007/BF01068419.
4. Chow, S.C.; Endrenyi, L.; Lachenbruch, P.A.; Mentre, F. Scientific factors and current issues in biosimilar studies. *J. Biopharm. Stat.* **2014**, *24*, 1138–1153. doi: 10.1080/10543406.2014.948961.
5. Chow, S.C.; Song, F.Y.; Chen, M. Some thoughts on drug interchangeability. *J. Biopharm. Stat.* **2016**, *26*, 178–186. doi: 10.1080/10543406.2015.1092027.
6. Anderson, S.; Hauck, W.W. The transitivity of bioequivalence testing: Potential for drift. *Int. J. Clin. Pharmacol. Ther.* **1996**, *34* (9), 369–374.
7. Bialer, M.; Midha, K.K. Generic products of antiepileptic drugs (AEDs): A perspective on bioequivalence and interchangeability. *Epilepsia* **2010**, *51*, 941–950. doi: 10.1111/j.1528-1167.2010.02573.x.
8. Privitera, M. Generic antiepileptic drugs: Current controversies and future directions. *Epilepsy. Curr.* **2008**, *8*, 113–117. doi: 10.1111/j.1535-7511.2008.00261.x.
9. van Gelder, T.; Gabardi, S. Methods, strengths, weaknesses, and limitations of bioequivalence tests with special regard to immunosuppressive drugs. *Transplant. Int.* **2013**, *26* (8), 771–777. doi: 10.1111/tri.12074.
10. FDA. *Draft Guidance for Industry: Bioequivalence Studies with Pharmacokinetic Endpoints for Drugs Submitted under an ANDA*; Center for Drug Evaluation and Research, U.S. Food and Drug Administration: Rockville, MD, 2013.
11. Song, F. What is indirect comparison. *Hayward Med. Commun.* **2009**, 1–6.
12. Garca-Arieta, A.; Potthast, H.; Leufkens, H.; Welink, J.; Gordon, J.; Gwaza, L.; Maliepaard, M.; Stahl, M. Assessment of the interchangeability between generics. *GaBI J.* **2016**, *5* (2): 55–59. doi: 10.5639/gabij.2016.0502.015.
13. Gwaza, L.; Gordon, J.; Welink, J.; Potthast, H.; Hansson, H.; Stahl, M.; Garca-Arieta, A. Statistical approaches to indirectly compare bioequivalence between generics: A comparison of methodologies employing artemether/lumefantrine 20/120 mg tablets as prequalified by WHO. *Eur. J. Clin. Pharmacol.* **2012**, *68*, 1611. doi: 10.1007/s00228-012-1396-1.
14. Gwaza, L.; Gordon, J.; Welink, J.; Potthast, H.; Leufkens, H.; Stahl, M.; Garca-Arieta, A. Adjusted indirect treatment comparison of the bioavailability of who-prequalified first-line generic antituberculosis medicines. *Clin. Pharmacol. Ther.* **2014**, *96* (5), 580–588. doi: 10.1038/clpt.2014.144.
15. Gwaza, L.; Gordon, J.; Potthast, H.; Welink, J.; Leufkens, H.; Stahl, M.; Garca-Arieta, A. Influence of point estimates and study power of bioequivalence studies on establishing bioequivalence between generics by adjusted indirect comparisons. *Eur. J. Clin. Pharmacol.* **2015**, *71* (9), 1083–1089. doi: 10.1007/s00228-015-1889-9.
16. Endrenyi, L.; Thfalusi, L. Adjusted indirect comparisons between generics bioequivalence and interchangeability. *GaBI J.* **2016**, *5* (2), 53–54. doi: 10.5639/gabij.2016.0502.014.
17. Zheng, J.Y.; Chow, S.C.; Yuan, M.D. On assessing bioequivalence and interchangeability between generics based on indirect comparisons. *Stat. Med.* **2017**, *36* (19), 1978–2993. doi: 10.1002/sim.7326.
18. Fisher, R.A. The fiducial argument in statistical inference. *Ann. Hum. Genet.* **1935**, *6* (4), 391–398.
19. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **1995**, *57* (1), 289–300.
20. Benjamini, Y. Discovering the false discovery rate. *J. R. Stat. Soc. Series B Stat. Methodol.* **2010**, *72* (4), 405416. doi: 10.1111/j.1467-9868.2010.00746.x.
21. Benjamini, Y.; Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **2001**, *29*, 11651188.
22. Storey, J.D. A direct approach to false discovery rates. *J. R. Stat. Soc. Series B Stat. Methodol.* **2002**, *64* (3), 479498. doi: 10.1111/1467-9868.00346.
23. Storey, J.D.; Tibshirani, R. Statistical significance for genomewide studies. *PNAS* **2003**, *100* (16), 9440–9445. doi: 10.1073/pnas.1530509100.

Interim Analysis

Paul Gallo

Novartis Pharmaceuticals Corporation, East Hanover, New Jersey, U.S.A.

Abstract

By interim analysis, we refer to any analysis, summarization, or presentation of clinical trial data which involves separate results produced according to treatment arm (which may be fully identified or coded, e.g., A/B), and which is produced prior to formal completion of the trial and database finalization. This entry details procedural issues and concerns as well as benefits of using interim analysis.

Keywords: Clinical trial; Interim analysis; Stopping rule.

INTRODUCTION

By *interim analysis*, we refer to any analysis, summarization, or presentation of clinical trial data which involves separate results produced according to treatment arm (which may be fully identified or coded, e.g., A/B), and which is produced prior to formal completion of the trial and database finalization. Interim analyses have, in recent years, become quite a commonplace feature of clinical trial protocols; in fact, it is rare to find a long-term clinical trial which does not include some schedule for interim examination of accruing data. Statistical and procedural guidelines for performing interim analyses are relatively recent and are, in fact, still evolving to some extent. For trials which utilize interim analyses, designing and executing these analyses properly can be very important for maintaining the integrity of the trial and full interpretability of its results.

MOTIVATION FOR THE USE OF INTERIM ANALYSES

Interim analyses can be performed for a variety of reasons. The primary motivation is to protect the well-being of patients in the study; this particularly applies in longer-term studies involving mortality or serious morbidity endpoints, or trials which include patients with severe disease conditions. Monitoring of accruing data can demonstrate whether or not there are safety concerns associated with the trial for which modification or termination might be called for. With respect to the primary endpoint of the trial, if a standard of proof for a difference between study treatments can be established, then it will often be unethical to continue the trial and expose patients in the inferior arm(s) to ineffective treatment regimens. Even if there is

not definitive evidence of a difference between treatments, if one treatment arm is trending worse than another (particularly, if this is a new or experimental treatment, or one with possible safety concerns), this might also ethically warrant termination of the trial.

Interests of the study sponsor can be a motivating factor for performing interim analyses: Prior to the planned conclusion of a trial, if a standard of proof for the effectiveness of a treatment can be achieved, or if it can be shown that an experimental treatment would not demonstrate benefit if the trial were to continue (“futility”), then savings of time and resources can be realized by early termination (a futility judgment may also entail an ethical component, similarly as in the previous paragraph: if a study will not achieve its objectives, it may not be appropriate to continue to treat patients within the trial framework).

Interim analyses may be used to verify that a trial is being conducted as planned and that there are not operational or compliance issues in the conduct of the trial that need to be addressed. Interim analyses might be performed to attempt to verify the statistical assumptions underlying the trial design (however, accuracy of such assumptions can in some cases be adequately assessed based upon blinded data). A role for interim analyses that has been a topic of much recent interest involves their use in *adaptive trial designs*, i.e., study designs in which changes may be implemented in some aspect of the conduct of the trial as it proceeds, based on accruing unblinded results (see, e.g., Ref. [1]). Adaptive trials generally require the use of specialized statistical methodology.

Occasionally, an interim analysis may be performed for purposes external to the trial. It may not be an objective of the analysis to modify the ongoing trial in any way; rather, it may be desired to obtain information, perhaps for the design of other trials or for other business purposes contingent on some knowledge which can be obtained from the

accruing results of the trial in progress. (It should be kept in mind that this can be controversial in a *confirmatory* trial, i.e., one which aims to provide definitive evidence of efficacy, due to the potential for introducing bias, for reasons we will discuss later).

The potential enthusiasm for the use of interim analyses is understandable. Through judicious use of interim analyses, exposure of patients to ineffective or unsafe treatments can be reduced, sponsor resources can be saved or refocused more efficiently, and the pace at which effective treatments reach the marketplace can be expedited. Such enthusiasm must be carefully guided, however. There are well-known statistical and procedural issues and concerns associated with interim analyses which require that they be implemented carefully and cautiously so that they do not impair the integrity of the trial and its ability to answer the scientific questions for which it was designed.

STATISTICAL ISSUES

The primary statistical issue associated with interim analysis involves protection of the significance level for the trial endpoint. If comparative treatment results can potentially be examined at a number of timepoints throughout the trial, then there may well be multiple opportunities to stop the trial and claim a significant difference among treatment arms. (Note: This may apply even if the stated objectives of the interim analyses did not consider the possibility of stopping and claiming an effect.) It is widely understood that if one were to perform a significance test on a primary endpoint at various stages of a trial and, furthermore, were to use a stopping rule based on conventional nominal significance levels unadjusted for the multiple looks, the overall significance level could become highly inflated. For example, envision a situation in which data could be examined at four timepoints during a trial (including the final analysis), with equal amounts of information accruing between each of these looks. If the trial would be stopped whenever the primary endpoint reached significance at a conventional 5% level (two-sided), the true overall significance level would, in fact, be nearly 13%.^[2] It follows that significance criteria must be determined so that the overall false positive rate across all examinations of the data is maintained at the desired level.

In this sense, interim analysis adjustment is basically a multiplicity issue. Under general assumptions, the correlation structure of the different test statistics is known—basically, it is determined by the “common” data among test statistics constructed at different timepoints—so that the desired overall significance level can be achieved. A variety of procedures have been described in the literature to address this issue so that practitioners have quite a broad and flexible arsenal of tools from which to choose to implement plans which best address the objectives of the interim analyses and maintain the integrity of the study as a whole.

Many of the more well-known procedures are mentioned and described briefly in a later section.

As mentioned previously, interim analyses may be performed for reasons which do not envision that any action will be taken in the current trial, e.g., to obtain information to be used in the design of other trials or other such purposes. Such analyses may not even involve data on the main trial endpoints. Analyses such as these, in which no actions relative to the current trial are envisioned, are sometimes referred to as *administrative* interim analyses. From a purely statistical standpoint, interim analyses which do not allow the possibility of stopping and making a formal claim should not require any adjustment of statistical criteria for the final analysis. Interim analyses performed solely to address whether there might be safety concerns which could warrant termination may also fall into this category. However, it should be considered that any review of comparative interim data could potentially set in motion a chain of events which could lead to further unblinding, and unplanned actions might possibly be taken. For example, results might be so compelling, perhaps unexpectedly positive, so that termination might be warranted on ethical grounds. In this sense, perhaps no interim analysis can be guaranteed to be completely administrative; it is somewhat a matter of judgment based on the particular details of a situation whether formal statistical adjustment of significance criteria should be employed for interim analyses which are considered to be administrative. (Another option, suggested in Ref. [3], is to impose a very trivial statistical adjustment for interim analyses which do not have the intent of stopping for positive efficacy; this thus maintains the operating characteristics of the study while providing an operational definition for an efficacy assessment in case one becomes necessary.)

Utilization of statistical methods for interim analyses can lead to complications which are somewhat unique to the interim analysis context. One of these involves estimation: if parameter and interval estimates at the conclusion of a trial are computed “naively”, i.e., unadjusted for the actual structure of interim examinations of the data and prespecified stopping rule, point estimates will be biased and confidence intervals will not provide the correct coverage. Methods have thus been developed to produce more appropriate versions; some of these are referred to in a later section.

PROCEDURAL ISSUES AND CONCERNS

The main procedural issue associated with interim analyses involves confidentiality of results: The integrity of the trial is best ensured if knowledge of interim results is carefully protected and, in particular, remains unknown to those involved in conducting the trial. Arguments underlying these concerns are well described in a U.S. FDA

guidance document.^[4] To briefly summarize the rationale: knowledge of how trial results are trending can potentially have unknown and subtle effects on the conduct of the remainder of the trial, which could bias the overall results and raise doubts about their interpretability. For example, such knowledge could potentially affect investigator attitudes and enthusiasm and, depending on the particular situation, might have an impact on the types of patients recruited, on specific details of the treatment regimen, on endpoint judgments, etc.

In addition, there is an issue regarding the ability of trial personnel to manage the study in an objective manner. While a trial is ongoing, trial management personnel will at times consider actions based on trial experiences or on emerging external information, perhaps of the type that validly lead to protocol amendments or changes to the statistical analysis plan. Such decisions must be made objectively, and not in an advantageous manner advised by knowledge of the interim results. In order to avoid possible biases or appearances of bias, it is thus clearly highly desirable to avoid having individuals with knowledge of interim results be party in any way to such activities and decisions. It is desirable that statisticians with responsibilities in the trial should remain blinded to interim results and individual patient treatment codes for several reasons: because they may be involved with data cleaning decisions and activities, because they are likely to be involved in decisions regarding finalization of the study analysis plan, and because they may have reason to be in contact with investigators or other personnel at the study sites who must remain blinded.

Thus, the integrity of a trial and the ability to validly interpret its results are best ensured by strictly controlling access to the results of any interim analyses which are performed. Inquisitiveness on the part of investigators, sponsor representatives, or other trial personnel is *not* a valid justification for being made aware of interim results. Further regulatory guidance on this issue is provided in Ref. [5], and in the International Conference on Harmonisation Document E9: “Statistical Principles for Clinical Trials,” which states:

The execution of an interim analysis should be a completely confidential process because unblinded data and results are potentially involved. All staff involved in the conduct of the trial should remain blind to the results of such analyses, because of the possibility that their attitudes to the trial will be modified and cause changes in the characteristics of patients to be recruited or biases in treatment comparisons. This principle may be applied to all investigator staff and to staff employed by the sponsor except for those who are directly involved in the execution of the interim analysis.

Of course, *some* individuals must be aware of interim analysis results in order to make recommendations or decisions relating to the objectives of the analysis.

Commonly, evaluation of interim analysis results is performed by a *Data Monitoring Committee (DMC)*. This is a group of individuals possessing the necessary expertise and experience (e.g., clinical, statistical, other disciplines relevant to the situation) charged with reviewing the interim results and making recommendations based upon them. Its members are ideally independent of all other trial activities other than as needed to carry out their duties on the DMC, and they are free of perceptions of conflict of interest. To the extent that statisticians and programmers must be unblinded to treatment codes in order to produce analysis results for a DMC to review, it is desirable that additional staff without other trial responsibilities be named to perform these tasks, for the reasons alluded to above.

When formal statistical boundaries or stopping rules are in place, it is important to keep in mind that these may not necessarily be strictly adhered to, but rather may be interpreted as more of a guideline.^[7] Statistical rules, of course, tend to be governed in large part by the principle of ensuring that the specified significance level for a primary endpoint is protected; however, a practical reality in clinical trials is that there is a much more complex set of objectives and available information to be taken into account when an action with ramifications as critical to a trial as its continuation or termination is being considered. Thus, a DMC will often consider the totality of information available to it which might include: results from safety variables or other endpoints, perhaps from a risk/benefit viewpoint; the ability of the trial to provide more definitive answers to important scientific questions if continued; new scientific information from external sources; etc. (A related statistical issue involves *overrunning*, which refers to obtaining additional data after a formal stopping criterion has been achieved. This can occur either due to a judgment of the type described above that it is appropriate for the stopping criterion to be overruled, or simply because the lag in obtaining data for an interim analysis causes valid data to be received after a decision has been made to stop the trial; see Ref. [8].)

Another issue which arises in trials with interim analyses involves the trade-off between the completeness/quality of interim data and the need to make timely decisions. If data used in an interim analysis does not accurately reflect the status of the trial at a fairly recent point in time, then patient welfare might be jeopardized, or other decisions related to the interim analysis objectives might be made less efficiently. Very often, it is difficult, if not impossible, to ensure that interim data be both timely and of the quality one would envision for a finalized trial database, depending on the nature of the endpoint and its assessment and on factors such as whether there is a formal endpoint adjudication process. Balancing the conflicting attributes of cleanliness of interim data vs. timeliness of review is very much a case-by-case decision for which no single recommendation will apply uniformly; this is an important

consideration to be taken into account when developing interim analysis plans and procedures.

In trials with interim analyses, the criteria for success in the final analyses are generally inextricably linked with the details of the interim analysis methodology used. It is thus important that the statistical details of an interim analysis plan be thoroughly specified in the protocol (or in an amendment, if a need arises for an unplanned interim analysis), and subsequent changes to the plan should also be documented prior to implementation. Processes for performing interim analyses, restricting access to treatment codes and comparative results, and the composition and operating procedures of the DMC are also important to specify in advance because, as described above, these factors can have important implications for the validity and interpretability of trial results. Final study reports will generally include analyses of all data obtained; that is, data received subsequent to a decision to terminate a trial should be included in analyses in the trial report.

Because of the potential for interim analyses to introduce bias as we have described, there may be added motivation to examine final trial data to see if there is evidence of inconsistency before and after interim analyses. A recent CHMP guidance document^[9] describes how, in any trial with interim analyses, apparent differences before and after interim analyses in the characteristics of patients enrolled or the results obtained may raise concerns about the interpretation of the overall results because of the possibility that operational bias has been introduced. These concerns may be particularly acute in adaptive trials, where the nature of the changes implemented may raise further questions about the interpretability of aggregate results. The same guidance document suggests that “simple rejection of a global null hypothesis across all stages of the trial may not be sufficient to establish a convincing treatment effect” and that methods should be pre-specified to address whether stages can be justifiably combined.

Useful discussions of some logistic, procedural, and regulatory issues associated with interim analyses can be found in Refs. [10–12]; for more comprehensive descriptions of DMC roles, responsibilities, and operational processes, see Refs. [13–15]; for discussions of these issues specific to adaptive trials, see Refs. [16,17].

DEVELOPMENT OF INTERIM ANALYSIS STATISTICAL METHODOLOGY

As mentioned above, quite a rich literature has been produced during the past few decades addressing statistical and procedural aspects of performing interim analyses in clinical trials, and much methodology has been developed. Descriptions of the development and evolution of the current methodology can be found in Refs. [18,19]; a brief but by no means complete summary is presented here.

Early ad hoc procedures describing interim analysis critical values or stopping rules were proposed by Haybittle^[20] and Peto et al.^[21] For a pairwise treatment comparison, a typical version of this type of approach proposed that a standardized normal deviate critical value of $z = 3.0$ could be used as a stopping rule at any interim analysis, without adjustment of the final analysis criteria. This illustrated the fact that for such conservative boundaries, any significance level inflation was likely to be trivial; however, such approaches obviously were quite limited with regard to implementation of efficient procedures with desired operating characteristics.

Research and interest in this topic took off with the introduction of the discrete group sequential boundaries proposed by Pocock^[22] and O’Brien and Fleming^[23] based upon the idea of repeated significance tests due to Armitage et al.^[2] For interim analyses equally spaced in terms of the amount of data accrued, Pocock boundaries on a standard normal test statistic scale corresponded to an identical z score for each interim analysis. On this same z scale, O’Brien-Fleming boundary values were governed by an inverse square-root relationship, were more conservative at early looks, and decreased to a final analysis critical value only slightly above that nominally corresponding to the desired overall significance level. As a simple illustration, for a four-look scheme with a one-sided 0.025 level, Pocock boundary values would be $z = 2.36$ at each stage, whereas the O’Brien-Fleming boundary values, in order, would be $z = 4.05, 2.86, 2.34$, and 2.02 (the corresponding single-testing criterion is of course the familiar $z = 1.96$). Under alternative hypotheses, Pocock boundaries could have a smaller expected sample size; O’Brien-Fleming boundaries paid a minimal “interim analysis penalty” and tended to yield a maximum sample size only slightly larger than the single-analysis design with equivalent level and power (the use of O’Brien-Fleming-like boundaries has been, and remains, quite popular in practice).

These early interim analysis boundary proposals were described in situations in which the analyses were equally spaced in terms of the amount of information available between them so that the number and timing of the analyses needed to be known in advance. This obviously was a constraint that was quite restrictive, given some of the practical realities of clinical trials such as the need to schedule analyses at specified calendar times, the variable nature of data accrual, etc. Lan and DeMets^[24] proposed the use of an α -spending function, or *use function*, to address these difficulties. In this approach, one only needed to prespecify the overall significance level and a monotone function which described the rate at which it was “used up.” At any particular interim analysis timepoint, the boundary would be determined according to the past and current schedule of analyses and not by future information. Among the spending functions described by Lan and DeMets were versions which approximately generalized the boundaries originally proposed by Pocock and O’Brien-Fleming.

Subsequent research resulted in extensions of these approaches into areas such as the following: accommodation of different data structures, most notably time-to-event data;^[25,26] the incorporation of lower stopping boundaries, corresponding to failure to reject the null hypothesis;^[27,28] and extension to general classes of spending functions, some of which explicitly incorporated power considerations.^[29–31]

A different methodological approach was developed by Whitehead and colleagues.^[32,33] In this approach, letting θ denote a measure of the true difference between two treatments, the boundaries are determined by the relationship between two statistics: Z , a measure of the observed difference (Note: This is not a standard normal deviate); and V , a measure of the amount of information about θ contained in the data, which will be approximately proportional to the sample size. Mathematically, Z is the efficient score for θ , and V is Fisher's information. Straight-line boundaries based upon these two statistics have been described and investigated as well as triangular boundaries corresponding to upper and lower stopping limits. This approach is fully flexible in terms of the number and timing of the analyses, allows study designs to have desired operating characteristics, and can be used to provide valid and unbiased inferences.

Another approach which has been used to assist in interim analysis continuation decisions is *stochastic curtailment*.^[34] According to this approach, conditional probabilities of outcomes at the scheduled end of the trial are calculated based upon the observed interim data, and decisions are made based upon these (typically, this conditioning assumes the prespecified null and alternative hypotheses, although the approach is not restricted to these). For example, a futility judgment might be made and a trial may be stopped, if the conditional chance of ultimate success, given the alternative hypothesis, is very low. In what might be considered an extension of this concept, the use of *predictive probabilities* has also been proposed.^[35,36] In these Bayesian or semi-Bayesian approaches, the trial outcome probabilities are calculated based upon current beliefs about the unknown parameters determined from the observed interim data itself (possibly combined with a priori assumptions). Still another approach—that of *repeated confidence intervals*^[37,38]—has some similarity to these in the sense that it can be used to provide information to assist in decision making, which is independent of any particular stopping boundary or monitoring schedule. Recent research considers additional levels of flexibility in trials with interim analyses, allowing trial modification through the application of adaptive design techniques. Generally, study stages are defined by the points at which interim analyses are performed to consider adaptations, and within-stage inferences or estimates are combined using pre-specified weights or combination functions; see, e.g., Refs. [39–44].

Correct parameter and interval estimation in the group sequential trial context has been addressed by a number of authors.^[45–48]

SOFTWARE

Much of the methodology associated with interim analyses is quite complex to implement computationally. A number of commercial software packages are available to facilitate the design and the utilization of interim analysis plans and to permit the construction of stopping boundaries and the production of valid analysis results for a variety of situations. These include *East* (Cytel Software, 2008)^[49] and *S + SeqTrial*, a module of the *S-Plus* package (Insightful, 2002),^[50] both of which incorporate a variety of methodological approaches; and *PEST4* (MPS Research Unit, 2000)^[51] which is based upon the straight-line and triangular boundaries approach of Whitehead and colleagues. The location of on-line FORTRAN programs that can be used to implement the α -spending function approach, and a description of their usage, can be found in Ref. [52].

CONCLUSION

There are clearly important benefits that can be achieved by judicious use of interim analyses in clinical trials. Properly planned and implemented, these can enable trials to achieve their objectives in a more ethical and efficient manner. The benefits accruing from interim analyses are, of course, not “free.” Interim analyses will generally require more intensive planning during the design phase of the trial: the specific interim analysis objectives and methods, and the processes governing the performance of the analyses and implementation of decisions, will all need to be carefully considered. A DMC will generally be required, along with additional staff to perform the analyses. Timely data collection and cleaning as the trial proceeds will take on added importance. Including interim analyses may cause trials to require more patients or have longer duration than they would have otherwise; on the other hand, there is also the potential for trials to be substantially shortened. When not planned and implemented carefully, interim analyses may impair rather than enhance trials, by raising questions about the integrity and interpretability of their results. Statistical methodology, software, and clinical and regulatory perspectives relating to interim analyses seem likely to continue the evolution evidenced in recent years.

REFERENCES

1. Chow, S.C.; Chang, M. *Adaptive Design Methods in Clinical Trials*; Chapman & Hall/CRC: Boca Raton, FL, 2007.

2. Armitage, P.; McPherson, C.K.; Rowe, B.C. Repeated significance tests on accumulating data. *J. R. Stat. Soc., A* **1969**, *132*, 235–244.
3. Koch, G.G.; Davis, S.M.; Anderson, R.L. Methodological advances and plans for improving regulatory success for confirmatory studies. *Stat. Med.* **1998**, *17*, 1675–1690.
4. US Food and Drug Administration. *Guidance for Clinical Trial Sponsors: Establishment and Operation of Clinical Trial Data Monitoring Committees*. Rockville MD 2006.
5. Committee for Medicinal Products for Human Use. *Guideline on Data Monitoring Committees*. London, 2006.
6. International Conference on Harmonisation Expert Working Group. *ICH Harmonised Tripartite Guideline: Statistical Principles for Clinical Trials*; Federal Register 1998, *63*, 49583–49598.
7. DeMets, D.L. Stopping guidelines vs. stopping rules: a practitioner's point of view. *Commun. Stat., Theory Methods* **1984**, *13*, 2395–2418.
8. Whitehead, J. Overrunning and underrunning in sequential clinical trials. *Control Clin. Trials* **1992**, *13*, 106–121.
9. Committee for Medicinal Products for Human Use. *Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned With an Adaptive Design*. London, 2007.
10. Fleming, T.R.; DeMets, D.L. Monitoring of clinical trials. *Control Clin. Trials* **1993**, *14*, 183–197.
11. Facey, K.M.; Lewis, J.A. The management of interim analyses in drug development. *Stat. Med.* **1998**, *17*, 1801–1809.
12. DeMets, D.L.; Califf, R.; Dixon, D.; Ellenberg, S.; Fleming, T.; Held, P.; Julian, D.; Kaplan, R.; Levine, R.; Neaton, J.; Packer, M.; Pocock, S.; Rockhold, F.; Seto, B.; Siegel, J.; Snapinn, S.; Stump, D.; Temple, R.; Whitley, R. Issues in regulatory guidelines for data monitoring committees. *Clin. Trials* **2004**, *1*, 162–169.
13. Ellenberg, S.S.; Fleming, T.R.; DeMets, D.L. *Data Monitoring Committees in Clinical Trials: A Practical Perspective*; Wiley: Chichester, 2002.
14. Grant, A.M.; Altman, D.G.; Babiker, A.B.; Campbell, M.K.; Clemens, F.J.; Darbyshire, J.H.; Elbourne, D.R.; McLeer, S.K.; Parmar, M.K.B.; Pocock, S.J.; Spiegelhalter, D.J.; Sydes, M.R.; Walker, A.E.; Wallace, S.A.; the DAMOCLES study group. Issues in data monitoring and interim analysis of trials. *Health Technol. Assess.* **2005**, *9* (7).
15. Herson, J. *Data and Safety Monitoring Committees in Clinical Trials*; Chapman & Hall/CRC: Boca Raton, 2009.
16. Gallo, P. Confidentiality and trial integrity issues for adaptive designs. *Drug. Inf. J.* **2006**, *40*, 445–449.
17. Herson, J. Coordinating data monitoring committees and adaptive clinical trial designs. *Drug. Inf. J.* **2008**, *42*, 297–301.
18. Jennison, C.; Turnbull, B.W. *Group Sequential Methods With Applications to Clinical Trials*; Chapman & Hall/CRC: London, 2000.
19. Hwang, I.K. Overview of the Development of Sequential Procedures. In *Biopharmaceutical Sequential Statistical Application*; Peace, K.E.; Ed.; Marcel Dekker, Inc.: New York, 1992, 3–15.
20. Haybittle, J.L. Repeated assessment of results in clinical trials of cancer treatment. *Br. J. Radiol.* **1971**, *44*, 793–797.
21. Peto, R.; Pike, M.C.; Armitage, P.; Breslow, N.E.; Cox, D.R.; Howard, S.V.; Mantel, N.; McPherson, K.; Peto, J.; Smith, P.G. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *Br. J. Cancer* **1976**, *34*, 585–612.
22. Pocock, S.J. Group sequential methods in the design and analysis of clinical trials. *Biometrika* **1977**, *64*, 191–199.
23. O'Brien, P.C.; Fleming, T.R. A multiple testing procedure for clinical trials. *Biometrics* **1979**, *35*, 549–556.
24. Lan, K.K.G.; DeMets, D.L. Discrete sequential boundaries for clinical trials. *Biometrika* **1983**, *70*, 659–663.
25. Slud, E.; Wei, L.J. Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J. Am. Stat. Assoc.* **1982**, *77*, 862–868.
26. Tsiatis, A.A. Repeated significance testing for a general class of statistics used in censored survival analysis. *J. Am. Stat. Assoc.* **1982**, *77*, 855–861.
27. DeMets, D.L.; Ware, J.H. Group sequential methods for clinical trials with a one-sided hypothesis. *Biometrika* **1980**, *67*, 651–660.
28. Emerson, S.S.; Fleming, T.R. Symmetric group sequential designs. *Biometrics* **1989**, *45*, 905–923.
29. Hwang, I.K.; Shih, W.J.; DeCani, J.S. Group sequential designs using a family of Type I error probability spending functions. *Stat. Med.* **1990**, *9*, 1439–1445.
30. Wang, S.K.; Tsiatis, A.A. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **1987**, *43*, 193–199.
31. Pampallona, S.; Tsiatis, A.; Kim, K. Interim monitoring of group sequential trials using spending functions for the Type I and Type II error spending probabilities. *Drug Inf. J.* **2001**, *35*, 1113–1121.
32. Whitehead, J.; Stratton, I. Group sequential clinical trials with triangular continuation regions. *Biometrics* **1983**, *39*, 227–236.
33. Whitehead, J. *The Design and Analysis of Sequential Clinical Trials*, 2nd Ed. Rev.; Wiley: Chichester, 1997.
34. Lan, K.K.G.; Simon, R.; Halperin, M. Stochastically curtailed tests in long-term clinical trials. *Commun. Stat., Seq. Anal.* **1982**, *1*, 207–219.
35. Spiegelhalter, D.J.; Freedman, L.S.; Blackburn, P.R. Monitoring clinical trials: conditional or predictive power?. *Control Clin. Trials* **1986**, *7*, 8–17.
36. Freedman, L.S.; Spiegelhalter, D.J. Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Control Clin. Trials* **1989**, *10*, 357–367.
37. Jennison, C.; Turnbull, B.W. Repeated confidence intervals for group sequential tests. *Control. Clin. Trials* **1984**, *5*, 33–45.
38. Jennison, C.; Turnbull, B.W. Interim analyses: the repeated confidence interval approach (with discussion). *J. R. Stat. Soc., B* **1989**, *51*, 305–361.
39. Bauer, P.; Kohne, K. Evaluation of experiments with adaptive interim analyses. *Biometrics* **1994**, *50*, 1029–1041.
40. Cui, L.; Hung, H.M.J.; Wang, S.J. Modification of sample size in group sequential clinical trials. *Biometrics* **1999**, *55*, 853–857.
41. Wassmer, G.; Eisebitt, R.; Coburger, S. Flexible interim analyses in clinical trials using multistage adaptive test designs. *Drug Inf. J.* **2001**, *35*, 1131–1146.
42. Mehta, C.R.; Tsiatis, A.A. Flexible sample size considerations using information-based interim monitoring. *Drug Inf. J.* **2001**, *35*, 1095–1112.

43. Muller, H.H.; Schafer, H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **2001**, *57*, 886–891.
44. Bretz, F.; Koenig, F.; Brannath, W.; Glimm, E.; Posch, M. Adaptive designs for confirmatory clinical trials. *Stat. Med.* **2009**, *28*, 1181–1217.
45. Tsiatis, A.A.; Rosner, G.L.; Mehta, C.R. Exact confidence intervals following a sequential test. *Biometrics* **1984**, *40*, 797–803.
46. Whitehead, J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika* **1986**, *73*, 573–581.
47. Emerson, S.S.; Fleming, T.R. Parameter estimation following group sequential hypothesis testing. *Biometrika* **1990**, *77*, 875–892.
48. Todd, S.; Whitehead, J.; Facey, K.M. Point and interval estimation following a sequential clinical trial. *Biometrika* **1996**, *83*, 453–461.
49. Cytel Statistical Software and Services. *East: Software for Advanced Clinical Trial Design, Simulation, and Monitoring, Version 5.2*; Cytel Software Corporation: Cambridge, MA, 2008.
50. Insightful Corporation. *S + SeqTrial 2 User's Manual*; Insightful Corporation: Seattle, WA, 2002.
51. MPS Research Unit. *PEST 4: Operating Manual*; The University of Reading: Reading, 2000.
52. Reboussin, D.M.; DeMets, D.L.; Kim, K.; Lan, K.K.G. Computations for group sequential boundaries using the Lan-DeMets spending function method. *Control Clin. Trials* **2000**, *21*, 190–207.

International Conference on Harmonization (ICH)

Jen-pei Liu

Division of Biometry, Department of Agronomy, and Institute of Epidemiology, National Taiwan University, Taipei, Taiwan; Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan

Shein-Chung Chow

Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

The International Conference on Harmonization (ICH) was organized to provide an opportunity for important initiatives to be developed by regulatory authorities, as well as to provide an opportunity for industry associations to promote international harmonization of regulatory requirements. This entry provides the guidelines to harmonize technical procedures.

Keywords: Efficacy; Guidelines; Quality; Safety.

INTRODUCTION

In drug research and development, health authorities in different countries have very similar-although different requirements for approval of the commercial use of drug products. As a result, pharmaceutical companies may repeatedly have to conduct similar studies or prepare different documents of the same pharmaceutical product for regulatory submission. Therefore there is a strong industrial and regulatory interest and a need in the international harmonization of requirements for drug research and development so that information generated in one country or area would be acceptable to other countries or areas. Consequently, the International Conference on Harmonization (ICH) of Technical Requirements for the Registration of Pharmaceuticals for Human Use was organized to provide an opportunity for important initiatives to be developed by regulatory authorities, as well as to provide an opportunity for industry associations to promote international harmonization of regulatory requirements.

ORGANIZATION

Currently, however, the ICH, is concerned only with tripartite harmonization of technical requirements for the registration of pharmaceutical products among three regions: the European Union, Japan, and the United States. Basically, the organization of the ICH consists of two representatives (one from a health authority and one from the pharmaceutical industry) from each of these three regions. As a result, the ICH organization consists of six parties: the European Commission of the European Union; the European Federation of Pharmaceutical Industries and Associations (EFPIA); the Japanese Ministry

of Health, Labor, and Welfare (MHLW); the Japanese Pharmaceutical Manufacturers Association (JPMA); the U.S. Food and Drug Administration (FDA); and the Pharmaceutical Research and Manufacturers of America (PhRMA). The ICH steering committee was established in April 1990: (i) to determine policies and procedures; (ii) to select topics; (iii) to monitor progress; and (iv) to oversee the preparation of biannual conferences. Each of the six parties has two seats on the ICH steering committee. The ICH steering committee also includes observers from the World Health Organization, the Canadian Health Protection Branch, and the European Free Trade Area, which have one seat each on the committee. In addition, two seats of the ICH steering committee are given to the International Federation of Pharmaceutical Manufacturers Association (IFPMA), which represents a research-based pharmaceutical industry from 56 countries outside the regions covered by the ICH. Note that the IFPMA has responsibilities to run the ICH Secretariat at Geneva, Switzerland, and to coordinate the preparation of documentation.

GUIDELINES

In order to harmonize technical procedures, the ICH has issued a number of guidelines and draft guidelines. After the ICH steering committee selected the topics, the ICH guidelines were initiated by a concept paper and went through a five-step review process given in [Table 1](#). [Table 2](#) provides a list of currently available ICH guidelines or draft guidelines. [Table 3](#) gives the table of contents for the ICH guideline on General Considerations for Clinical Trials. In addition, [Tables 4–6](#) give the tables of contents of the ICH guidelines for Good Clinical Practices: Consolidated

Table 1 Review Steps for the ICH Guidelines*Step 1*

1. A harmonized topic is identified.
2. An expert working group (EWG) is formed.
3. Each party has a topic leader and a deputy.
4. A rapporteur for the EWG is selected.
5. Other parties are represented on EWG as appropriate.
6. A guideline, policy statement, and “points to consider” are produced.
7. Scientific issues are agreed upon.
8. Sign-off and submission to the ICH steering committee.

Step 2

1. Review of the ICH document by the steering committee.
2. Sign-off by all six parties.
3. Formal consultation in accord with regional requirements.

Step 3

1. A regulatory rapporteur is appointed.
2. Comments are collected and reviewed across all three regions.
3. Step 2 draft is revised.
4. Sign-off by EWG regulatory members.

Step 4

1. Forwarding to steering committee.
2. Review and sign-off by three regulatory members of the ICH.
3. Recommendation for adoption to regulatory bodies.

Step 5

1. Adoption of recommendations by regulatory agencies.
2. Incorporation into domestic regulations and guidelines.

Table 2 The ICH Guidelines or Draft Guidelines*Safety*

1. S1A The Need for Long-Term Rodent Carcinogenicity Studies of Pharmaceuticals
2. S1B Testing for Carcinogenicity of Pharmaceuticals
3. S1C Dose Selection for Carcinogenicity Studies of Pharmaceuticals
4. S1C(R) Guidance on Dose Selection for Carcinogenicity Studies of Pharmaceuticals: Addendum on a Limit Dose and Related Notes
5. S2A Specific Aspects of Regulatory Genotoxicity Tests for Pharmaceuticals
6. S2B Genotoxicity: A Standard Battery for Genotoxicity Testing of Pharmaceuticals
7. S2(R1) Genotoxicity Testing and Data Interpretation for Pharmaceuticals Intended for Human Use
8. S3A Toxicokinetics: The Assessment of Systemic Exposure in Toxicity Studies
9. S3B Pharmacokinetics: Guidance for Repeated Dose Tissue Distribution Studies

10. S4A Duration of Chronic Toxicity Testing in Animals (Rodent and Nonrodent Toxicity Testing)
11. S5(R2) Detection of Toxicity to Reproduction for Medicinal Products
 - S5A Detection of Toxicity to Reproduction for Medicinal Products
 - S5B Detection of Toxicity to Reproduction for Medicinal Products: Addendum on Toxicity to Male Fertility
12. S6 Preclinical Safety Evaluation of Biotechnology—Derived Pharmaceuticals
13. S7A Safety Pharmacology Studies for Human Pharmaceuticals
14. S7B Nonclinical Evaluation of the Potential for delayed Ventricular Repolarization (QT Interval Prolongation) by Human Pharmaceuticals
15. S8 Immunotoxicity Studies for Human Pharmaceuticals
16. S9 Nonclinical Evaluation for Anticancer Pharmaceuticals

Efficacy

1. E1A The Extent of Population Exposure to Assess Clinical Safety for Drugs Intended for Long-Term Treatment of Non-Life-Threatening Conditions
2. E2A Clinical Safety Data Management: Definitions and Standards for Expedited Reporting
3. E2B Data Elements for Transmission of Individual Case Safety Reports
4. E2B(R) Clinical Safety Data Management: Data Elements for Transmission of Individual Case Safety Reports
5. E2C(R1) Clinical Safety Data Management: Periodic Safety Update Reports for Marketed Drugs
 - E2C Clinical Safety Data Management: Periodic Safety Update Reports for Marketed Drugs
 - E2C Addendum to ICH E2C Clinical Safety Data Management: Periodic Safety Update Reports for Marketed Drugs
6. E2D Postapproval Safety Data Management: Definitions and Standards for Expedited Reporting
7. E2E Pharmacovigilance Planning
8. E2F Development Safety Update Report
9. E3 Structure and Content of Clinical Studies
10. E4 Dose–Response Information to Support Drug Registration
11. E5 Ethnic Factors in the Acceptability of Foreign Clinical Data
12. E6 Good Clinical Practice: Consolidated Guideline
13. E7 Studies in Support of Special Populations: Geriatrics
14. E8 General Considerations for Clinical Trials
15. E9 Statistical Principles for Clinical Trials
16. E10 Choice of Control Group in Clinical Trials
17. E11 Clinical Investigation of Medicinal Products in the Pediatric Population

(Continued)

Table 2 The ICH Guidelines or Draft Guidelines (*Continued*)

18. E12A Principles for Clinical Evaluation of New Antihypertensive Drugs.	Appendix 4 Maintenance Procedures for Updating
19. E14 Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs	12. Q4B Evaluation and Recommendation of Pharmacopoeial Texts for Use in the International Conference on Harmonisation Regions
Questions and Answers	• Annex 1: Residue on Ignition/Sulphated Ash General Chapter
20. E15 Pharmacogenomics Definitions and Sample Coding	• Annex 2: Test for Extractable Volume of Parenteral Preparations General Chapter
<i>Joint safety/efficacy</i>	• Annex 3: Test for Particulate Contamination: Subvisible Particles General Chapter
1. M2 eCTD: Electronic Common Technical Document Specification	• Annex 4A: Microbiological Examination of Non-Sterile Products: Microbial Enumeration Tests General Chapter
• M2: eCTD Specification Questions and Answers and Change Requests	• Annex 4B: Microbiological Examination of Non-Sterile Products: Tests for Specified Micro-organisms General Chapter
• Companion Document: Current Q & As and Change Requests	• Annex 4C: Microbiological Examination of Non-Sterile Products: Acceptance Criteria for Pharmaceutical Preparations and Substances for Pharmaceutical Use General Chapter
2. M3(R2) Nonclinical Safety Studies for the Conduct of Human Clinical Trials and Marketing Authorization for Pharmaceuticals	• Annex 5: Disintegration Test General Chapter
3. M3 Nonclinical Safety Studies for the Conduct of Human Clinical Trials for Pharmaceuticals	• Annex 6: Uniformity of Dosage Units General Chapter
4. M4: Common Technical Document for the Registration of Pharmaceuticals for Human Use	• Annex 7: Dissolution Test General Chapter
• M4: Organization of the CTD	• Annex 8: Sterility Test General Chapter
• M4 Granularity Annex	13. Q5A Viral Safety Evaluation of Biotechnology Products Derived From Cell Lines of Human or Animal Origin
• M4: The CTD—General Questions and Answers	14. Q5B Quality of Biotechnology Products: Viral Safety Evaluation of Biotechnology Products Derived from Cell Lines of Human or Animal Origin
• M4: The CTD—Quality	15. Q5B Quality of Biotechnology Products: Analysis of the Expression Construct in Cells Used for Production of rDNA Derived Protein Products
• M4: The CTD—Quality Questions and Answers/ Location Issues	16. Q5C Quality of Biotechnology Products: Stability Testing of Biotechnological/ Biological Products
• M4: The CTD—Efficacy	17. Q5D Quality of Biotechnological/Biological Products: Derivation and Characterization of Cell Substrates Used for Production of Biotechnological/Biological Products
• M4: The CTD—Safety	18. Q5E Comparability of Biotechnological/Biological Products Subject to Changes in Their Manufacturing Process
• M4: The CTD—Safety Appendices	19. Q6A International Conference on Harmonisation; Guidance on Q6A Specifications: Test Procedures and Acceptance Criteria for New Drug Substances and New Drug Products: Chemical Substances
5. M5 International Conference on Harmonisation: Draft Guidance on M5 Data Elements and Standards for Drug Dictionaries	20. Q6B Specification: Test Procedures and Acceptance Criteria for Biotechnological/ Biological Products
6. Submitting Marketing Applications According to the ICH/ CTD Format: General Considerations	21. Q7A Good Manufacturing Practice for Active Pharmaceutical Ingredients
<i>Quality</i>	22. Q8 Pharmaceutical Development
1. Q1A Stability Testing of New Drug Substances and Products	23. Q8(R1) Pharmaceutical Development Revision 1
2. Q1A(R2) Stability Testing of New Drug Substances and Products	24. Q9 Quality Risk Management
3. Q1B Photostability Testing of New Drug Substances and Products	25. Q10 Pharmaceutical Quality System
4. Q1C Stability Testing for New Dosage Forms	
5. Q1D Bracketing and Matrixing Designs for Stability Testing of New Drug Substances and Products	
6. Q1E Evaluation of Stability Data	
7. Q2A (R1) Test on Validation of Analytical Procedures	
• Q2A (R1) Test on Validation of Analytical Procedures	
• Q2B Validation of Analytical Procedures: Methodology	
8. Q3A(R) Impurities in New Drug Substances	
9. Q3B(R) Impurities in New Drug Products	
10. Q3C Impurities: Residual Solvents	
11. Q3C Tables and List	

Table 3 Table of Contents for the Guideline on General Considerations for Clinical Trials

1.	Objectives of This Document
2.	General Principles
2.1	Protection of Clinical Trial Subjects
2.2	Scientific Approach in Design and Analysis
3.	Development Methodology
3.1	Considerations for the Development Plan
3.1.1	Nonclinical Studies
3.1.2	Quality of Investigational Medicinal Products
3.1.3	Phases of Clinical Development
3.1.4	Special Considerations
3.2	Considerations for Individual Clinical Trials
3.2.1	Objectives
3.2.2	Design
3.2.3	Conduct
3.2.4	Analysis
3.2.5	Reporting

Table 4 Table of Contents for the ICH Guideline on Good Clinical Practices: Consolidated Guidelines

Introduction	
1.	Glossary
2.	The Principles of ICH GCP
3.	The Institutional Review Board/Independent Ethnic Committee (IRB/IEC)
3.1	Responsibilities
3.2	Composition, Functions, and Operations
3.3	Procedures
3.4	Records
4.	Investigators
4.1	Investigator's Qualifications and Agreements
4.2	Adequate Resources
4.3	Medical Care of Trial Subjects
4.4	Communication with IRB/IEC
4.5	Compliance with Protocol
4.6	Investigational Products
4.7	Randomization Procedures and Unblinding
4.8	Informed Consent of Trial Subjects
4.9	Records and Reports
4.10	Progress Reports
4.11	Safety Reporting
4.12	Premature Termination or Suspension of a Trial
4.13	Final Report(s) by Investigator/Institution
5.	Sponsor
5.1	Quality Assurance and Quality Control
5.2	Contract Research Organization
5.3	Medical Expertise
5.4	Trial Design
5.5	Trial Management, Data Handling, Recording Keeping and Independent Data Monitoring Committee
5.6	Investigator Selection
5.7	Allocation of Duties and Functions
5.8	Compensation to Subjects and Investigators
5.9	Financing
5.10	Notification/Submission to Regulatory Authority(ies)

5.11	Confirmation of Review by IRE/IEC
5.12	Information on Investigational Product(s)
5.13	Manufacturing, Packaging, Labeling, Coding of Investigation Product(s)
5.14	Supplying and Handling Investigational product(s)
5.15	Record Access
5.16	Safety Information
5.17	Adverse Drug Reaction Reporting
5.18	Monitoring
5.19	Audit
5.20	Noncompliance
5.21	Premature Termination or Suspension of a Trial
5.22	Clinical Trial/Study Reports
5.23	Multicenter Trials
6.	Clinical Trial Protocol and Protocol Amendment(s)
6.1	General Information
6.2	Background Information
6.3	Trial Objectives and Purpose
6.4	Trial Design
6.5	Selection and Withdrawal of Subjects
6.6	Treatment of Subjects
6.7	Assessment of Efficacy
6.8	Assessment of Safety
6.9	Statistics
6.10	Direct Access to Source Data/Documents
6.11	Quality Control and Quality Assurance
6.12	Ethics
6.13	Data Handling and Record Keeping
6.14	Financing and Insurance
6.15	Publication
6.16	Supplements
7.	Investigator's Brochure
7.1	Introduction
7.2	General Considerations
7.3	Contents of the Investigator's Brochure
7.4	Appendix 1
7.5	Appendix 2
8.	Essential Documents for the Conduct of a Clinical Trial
8.1	Introduction
8.2	Before the Clinical Phase of the Trial Commences
8.3	During the Clinical Conduct of the Trial
8.4	After Completion or Termination of the Trial

Table 5 Table of Contents for the ICH Guideline on the Structure and Content of Clinical Study Reports

Introduction	
1.	Title Page
2.	Synopsis
3.	Table of Contents for the Individual Clinical Study Report
4.	List of Abbreviations and Definition of Terms
5.	Ethics
6.	Investigators and Study Administrative Structure
7.	Introduction
8.	Study Objectives
9.	Investigational Plan
10.	Study Patients
11.	Efficacy Evaluation

(Continued)

Table 5 Table of Contents for the ICH Guideline on the Structure and Content of Clinical Study Reports (*Continued*)

-
- | | |
|-----|--|
| 12. | Safety Evaluation |
| 13. | Discussion and Overall Conclusions |
| 14. | Tables, Figures, and Graphs Referred to But Not Included in the Text |
| 15. | Reference List |
| 16. | Appendices |
-

Table 6 Table of Contents for the ICH Guideline on Statistical Principles for Clinical Trials

-
- | | |
|------|--|
| I. | Introduction |
| 1.1 | Background and Purposes |
| 1.2 | Scope and Direction |
| II. | Considerations for Overall Clinical Development |
| 2.1 | Study Context |
| 2.2 | Study Scope |
| 2.3 | Design Techniques to Avoid Bias |
| III. | Study Design Considerations |
| 3.1 | Study Configuration |
| 3.2 | Multicenter Trials |
| 3.3 | Type of Comparisons |
| 3.4 | Group Sequential Designs |
| 3.5 | Sample Size |
| 3.6 | Data Capture and Processing |
| IV. | Study Conduct |
| 4.1 | Study Monitoring |
| 4.2 | Changes in Inclusion and Exclusion Criteria |
| 4.3 | Accrual Rates |
| 4.4 | Sample Size Adjustment |
| 4.5 | Interim Analysis and Early Stopping |
| 4.6 | Role of Independent Data Monitoring Committee |
| V. | Data Analysis |
| 5.1 | Prespecified Analysis Plan |
| 5.2 | Analysis Sets |
| 5.3 | Missing Values and Outliers |
| 5.4 | Data Transformation/Modification |
| 5.5 | Estimation, Confidence Interval, and Hypothesis Testing |
| 5.6 | Adjustment of Type I Error Rate and Confidence Levels |
| 5.7 | Subgroup, Interactions, and Covariates |
| 5.8 | Integrity of Data and Computer Software |
| VI. | Evaluation of Safety and Tolerability |
| 6.1 | Scope of Evaluation |
| 6.2 | Choice of Variables and Data Collection |
| 6.3 | Set of Subjects to Be Evaluated and Presentation of Data |
| 6.4 | Statistical Evaluation |
| 6.5 | Single Study vs. Integrated Summary |
| VII. | Reporting |
| 7.1 | Evaluation and Reporting |
| 7.2 | Summarizing the Clinical Database |
-

Annex 1 Glossary

Guidelines, for Structure and Content of Clinical Study Reports, and for Statistical Principles for Clinical Trials, respectively. From these tables, it can be seen that these guidelines are not only for the harmonization of design, conduct, and analysis, and for a report for a single clinical trial, but also for a consensus in protecting and maintaining the scientific integrity of the entire clinical development plan of a pharmaceutical entity.

CONCLUSION

ICH serves as a platform to facilitate harmonization of technical requirements for the registration of pharmaceutical products among regions and between health authority and industry. It is a step forward in the direction of globalization and in reduction of development cost for many needed drugs.

Item Response Theory

Wen-Hung Chen

North Potomac, Maryland, U.S.A.

Joseph C. Cappelleri

Pfizer Inc., New York, New York, U.S.A.

Abstract

This article serves as an introduction to the item response theory (IRT) methodology for researchers who are interested in the development of healthcare measures to collect health status data for use in clinical practice, health-care research, and clinical trials. Although IRT is relatively new to health-care field, its use has increased rapidly because it provides a detailed analysis of the item performance that enables the selection of well-performed items to be included in the measure. The basic concept of IRT is to describe the relationship between the items and the underlying magnitude of the trait believed to be governing the responses to these items. Topics included in this article are history of the IRT, its basic assumptions, well-known IRT models, and an illustration of its application. Only the unidimensional IRT models are discussed in this article as they are more commonly used health-care measures. IRT is especially suitable when the number of items is large. However, it also requires a large sample size to yield robust findings. The growing experience and continuing research will lead to improvement and wider application of the IRT methodology in the health-care field even further.

Keywords: Clinical outcome assessment; Graded response model; Partial credit model; Patient-reported outcome; Psychometrics; Rasch model.

INTRODUCTION

This article serves as an introduction to the item response theory (IRT) methodology for researchers who are interested in the development of health-care measures. Health-care researchers develop measures to collect health status data for use in clinical practice, health-care research, and clinical trials. Hailed as the modern or advanced psychometric methodology, IRT has gained a foothold in developing health-related measures. IRT provided a detailed analysis of the item performance that enables the selection of well-performed items to ensure and enhance the measurement properties of the health-care measures to be developed.

The article starts with a background and a brief history of the IRT (Sections “Background” and “Brief History of IRT”, respectively), followed by the description of its basic assumptions and terminology including the discussion of model fit, item fit, and sample size considerations (Section “Basic Assumptions and Terminology”). Some well-known IRT models are then introduced (Section “IRT Models”). Only the unidimensional IRT models are discussed in this article as they are more commonly used in the development of health-related measures. The article concludes with the application of the IRT for the development of instruments for use in health care (Section “Application of IRT in Healthcare Assessment”).

BACKGROUND

Hailed as the modern or advanced psychometric methodology, IRT has gained a foothold in developing instruments in the field of health care. Some instruments are used in clinical practice or in health-care research as tools for gathering health status data, or for screening, diagnosing, or monitoring progress on specific diseases. Some instruments are developed to measure patients’ preference or satisfaction with medical treatments. The IRT methodology is of special interest for researchers whose goal is to develop instruments for use as study endpoints for clinical trials to evaluate treatment outcomes.

In order to demonstrate that a treatment is effective or safe, the instruments, called clinical outcome assessments (COAs), have to be reliable, valid, and sensitive to detect change.^[1] COAs measure a specific concept (i.e., what is being measured by an assessment, such as pain intensity) within a particular context of use.^[1] Specifically, COAs are used to capture key symptoms and psychological and functional impacts directly from patients themselves or indirectly from observers and clinicians regarding a disease condition or a treatment. This information is critical for the regulatory review of a new treatment. Such information can also assist health technology assessors and health-care payers in determining the value of a new treatment. To this end, COAs

are increasingly being used as the key study endpoints in clinical trials across a wide variety of therapeutic areas.

This article, intended to be illustrative and to provide general understanding, focuses on the use of the IRT methodology in the development of the COAs. However, the basic principles of applying IRT to the development instruments are the same regardless of whether the purpose of an instrument is for monitoring general health status, for screening or diagnosing a specific disease, or for evaluating COA in a clinical study. Whether it is developing a new instrument, modifying an existing instrument, or adapting an existing instrument for a new population, the psychometric analysis should start with examining the performance of individual items. The goal is to demonstrate that the items perform well psychometrically and together they assess what the instrument is intended to measure. Another goal of psychometric analysis is to develop the scoring algorithm for the instrument.

The IRT methodology provides a detailed analysis of item performance that is helpful in selecting items with desired measurement properties that are expected to constitute a valid and reliable instrument. Under IRT, the relationship between the items and the underlying latent (unobservable) trait or attribute of interest (e.g., depressive symptoms, pain, physical functioning) does not vary according to the study population. Also, an estimate of the standard error on an item characteristic is conditional and available for each trait level.^[2–4] In addition, with IRT, the items and the trait estimates of the individuals are placed on the same measurement scale, which allows the evaluation of whether the specific set of items adequately covers the range of the underlying traits of the target population.

As an illustration, Fig. 1 shows the distributions of the underlying trait (labeled as “Severity” with logit as the scale unit) of two groups of subjects (stable vs. exacerbation) in the upper panel, where higher values indicate individuals with higher levels of severity. The range of

“Severity” that the items assessed is shown in the lower panel, where higher values indicate the items that are more likely to match those with higher level of severity. The figure shows that the items cover a wide range of the scale, where the majority of the subjects are located on the scale from -3.5 to 3.5 . The figure also shows the separation of the underlying severity between the two groups of subjects as measured by the instrument.

The basic concept of IRT is to describe the relationship between the items and the underlying magnitude or severity of the trait believed to be governing the responses to these items. Specifically, nonlinear mathematical functions are used to describe the probability of observing a particular item response with a given level of the underlying trait. An IRT model is said to be unidimensional when it involves only a single underlying trait. A multidimensional IRT model assumes there is more than one underlying trait. Most applications of the IRT in health-related measures use unidimensional IRT models. It is not just because it is simpler and more practical, but oftentimes it is a reasonable assumption and the findings are interpretable and useful. As such, we will only discuss unidimensional IRT models that are commonly used in the development of instruments in this article.

BRIEF HISTORY OF IRT

The earliest conceptualization of the IRT may be traced back to Thurstone’s article “A Method of Scaling Psychological and Educational Tests”.^[6] Frederic Lord further formulated the concept with a cumulative normal distribution function to describe the relationship between the probability of a correct response to an item and a latent variable.^[7] Around the same time, the Danish mathematician Georg Rasch^[8] and the Austrian sociologist Paul Lazarsfeld^[9] have also proposed parallel models independently. Although the

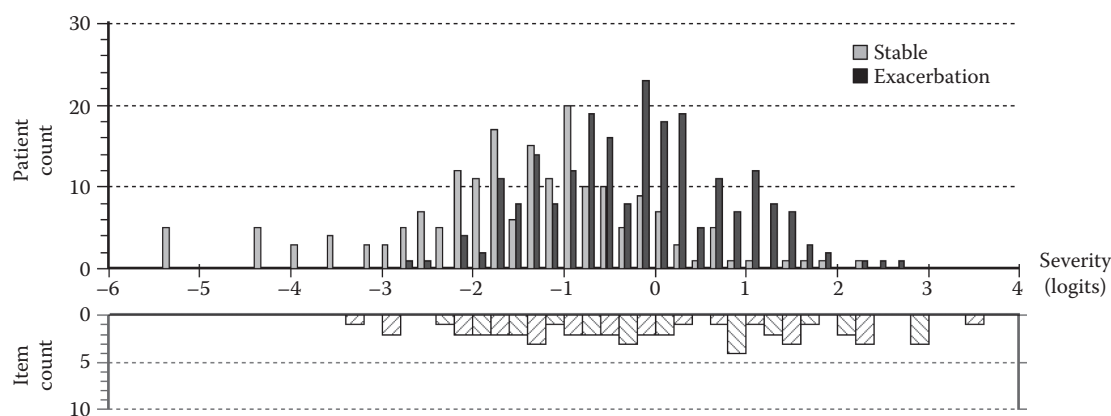


Fig. 1 Score distribution for patients and items by state (stable, acute). Severity on the x (horizontal) axis is measured on the logit scale. Negative logits indicate less severity, and more positive logits indicate more severity. The top histogram shows the frequency distribution of subjects with stable and exacerbating conditions; the bottom histogram shows the range of severity on the set of items measured. From Jones et al.^[5] © 2011 Elsevier. All rights reserved

theoretical foundation of the IRT was established in the 1950s and 1960s, it did not gain immediately recognition of its usefulness. The main reason was the lack of a proper statistical estimation procedure for the item response models. IRT emerged as a challenger to the then dominant classical test theory by the publication of the Lord and Novick's *Statistical Theory of Mental Test Scores*.^[10] Birnbaum's proposal^[11] to replace the cumulative normal distribution function with logistic function has made the computation much more manageable. The logistic function has since become the operational model for IRT analysis. Several estimation methods have been developed since.^[12–16] Various IRT models have been introduced at the same time.^[17–21] Riding along with the advance of the computer technology, the application of IRT took off into the 1980s and 1990s and has become a common practice in educational testing.

The widespread and successful uses of IRT in educational testing did not go unnoticed by researchers who were also seeking for a better way to develop measures for use in health-care practice, health-care research, and clinical studies. The increasing interest was fueled by the growing emphasis on evidence-based medicine where the effectiveness of medical treatments should be evaluated based on how patients survive, feel, and function. The outcomes movement gained momentum with the publication of Ellwood's "Shattuck Lecture—Outcomes Management. A Technology of Patient Experience"^[22] in 1988 and the creation of the "Agency for Health Care Policy and Research" in 1989. Studies of patient-focused outcomes were recommended and health-related outcome measures were in demand.^[23] Collection of patient-focused data, such as patient-reported outcomes (PROs), necessitated the development of well-defined and reliable instruments, and their development process should be psychometrically sound and efficient.

It was only natural that researchers in health care borrowed the knowledge and experiences from the educational testing.^[23,24] Publications of PRO data analysis using IRT can be found as early as in 1996.^[25] Revicki and Cella's 1997 article formally introduced the potential of IRT for use in health outcome assessment.^[26] The first decade of the 21st century saw the explosion of the IRT application in the development of generic health status instruments, as well as disease-specific outcome assessments. Today, as in the educational testing, IRT has become a major tool for analyzing the performance and psychometric properties of the items in the development of healthcare assessments.

BASIC ASSUMPTIONS AND TERMINOLOGY

Unidimensionality

The unidimensional IRT model means that only a single underlying trait is measured by the items that constitute the instrument. That is, the instrument measures a single construct—the underlying concept of interest—such as

pain, asthma symptoms, physical functioning, or emotional well-being. A person's level on this single underlying trait gives rise to the responses to the items. Normally, whether the items belong to a single construct is the first thing that is tested before fitting the item response data to an IRT model. A common approach to test unidimensionality is through exploratory factor analysis or confirmatory factor analysis. Sometimes, however, data are fitted to an IRT model first, and then the dimensionality of the item residuals is examined using principal component analysis. The items are unidimensional if there are as many principal components of the item residuals as there are items. If unidimensionality is not supported, grouping of items should be identified and broken into subscales. In this case, the IRT model should be fit to the individual subscale that is unidimensional.

Local Independence

IRT also assumes that items that constitute the instrument are locally independent. Local independence means that the responses to a pair of items are uncorrelated when the underlying trait has been controlled (or accounted) for. In other words, except random error, the underlying trait is the only factor that governs the item responses. Item pairs are locally dependent because there is at least one other factor influencing the item responses. Local dependent indices and the correlation of the item residuals are the common approaches used to detect pairs of locally dependent items.^[27]

When examining the correlation of item residuals, the findings of correlated residuals indicate that the items are either locally dependent or are multidimensional. In this situation, the item content should be reviewed to understand the underlying causes to resolve the issue. Multidimensionality occurs when two items measure different constructs between them. Local dependence is when the two items measure the same construct but another nuisance factor(s) also affect(s) how the items are answered. When unidimensionality assumption is violated, the two items should be divided into two unidimensional subscales, if both constructs are relevant. When local independence assumption is violated, one item of the item pair should be removed for the instrument.

Item Characteristic Curve and Categorical Probability Curve

As mentioned earlier, IRT is a collection of mathematical equations that are used to describe the relationship between the probability of endorsing a given item response and the underlying trait being measured. When there are only two response categories (i.e., items with dichotomous responses), this relationship can be graphically depicted as probability curve and is commonly called an item characteristic curve or a trace line. Each of the two response categories has its own characteristic curve. However, it is sufficient to show only the curve of the response category that increases in probability because the other curve is its

mirror image with decreasing probability. Figure 2 shows the item characteristic curves of two dichotomous items. It also shows the discrimination and location parameters that are described in the next section.

When an item has more than two response categories (i.e., a polytomous item), the relationship between the response categories and the underlying trait is depicted in the same way, except that the probability curves of all response categories are plotted. They are often called the categorical response curves. Figure 3 shows the categorical response curves of an item with five response categories, labeled 0–4. As shown in the figure, each response category has a certain range on the underlying trait where it is more likely to be endorsed.

Item Parameters

The shapes of the item characteristic curve and the categorical probability curve, and their locations, are determined by the parameters embedded in IRT models. Because

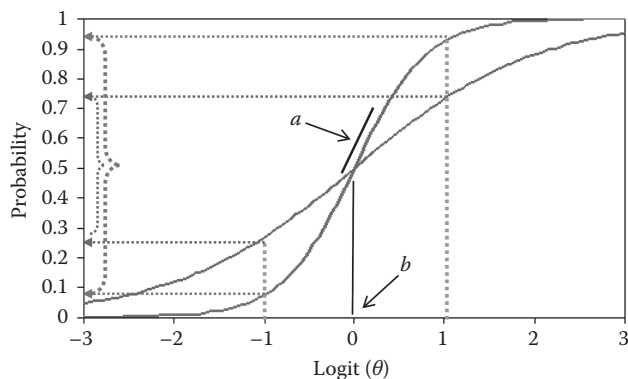


Fig. 2 Item characteristic curves of two dichotomous items. The two items displayed here have the same location parameter, denoted as b , but have different discrimination parameters, denoted as a . An item with a steeper slope has a higher discrimination parameter and can better differentiate subjects at different trait levels than the other item with a lower discrimination parameter

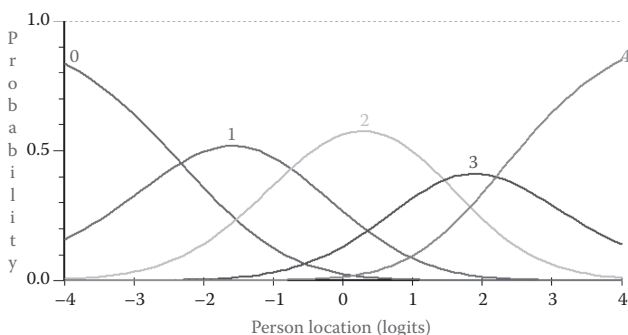


Fig. 3 Categorical probability curves of a polytomous item. The item has five response categories. Their probabilities of being endorsed are shown across the range of the underlying traits

these parameters are used to describe the items, they are also called the item parameters. IRT models differentiate themselves by how these parameters are formulated or constrained. In health-care application, two types of item parameters are most common and useful: discrimination parameter and location parameter. Discrimination parameter, also called the slope parameter, describes how well an item can differentiate individuals at different trait levels. The discrimination parameter represents the strength of the association between the item and the underlying trait; the higher the value of the discrimination parameter, the stronger the association. The stronger the association the greater the item is able to discriminate between the low and high levels of the underlying trait.

For instance, Fig. 2 shows two items with different discrimination parameters. The item with the higher discrimination parameter can differentiate individuals at different trait levels better than the item with the lower one. In the majority of IRT models, each item only has a single discrimination parameter. Among them, some models constrain the discrimination parameter to be the same for all items, such as the Rasch model^[8] and the partial credit model.^[19]

The location parameters describe the location on the underlying trait where the item responses have a specific probability. For dichotomous items, the location parameter is usually the location of where the probability is the same for either one response (i.e., 0.5). An exception is the three-parameter logistic (3PL) model where an additional guessing parameter changes the probability at the location of the location parameter. For polytomous items, the location parameters indicate either where, on the trait level, the probabilities are the same between adjacent responses (e.g., partial credit model and generalized partial credit model)^[19,20] or where the cumulative probability of a given response and any responses above it is 0.5 (e.g., graded response model).^[17] Figure 4 shows the item characteristic curves of two dichotomous items with different location parameters.

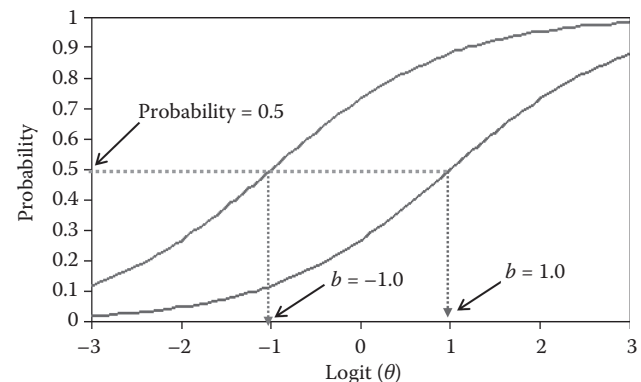


Fig. 4 Item characteristic curves of two dichotomous items with different location parameters. Location parameters indicate the level of underlying trait where the probability of endorsing the given response is 0.5

Model Fit

As with any use of mathematical models, it is important to assess the fit of the data to the model. The fit of a model can be compared using indices of the overall model fit. One straightforward procedure is to compare the observed total score distribution to the model-predicted score distribution. Their distributions could be visually compared or the difference between them could be assessed with a statistical test such as a chi-square test. Tests are also available for comparing the fit between different models.

Item Fit

Item fit indices can help identify items that may be retained, excluded, or replaced. To assess the item fit, the key element is the residual. The residual is the difference between an observed proportion and a model-predicted (expected) proportion of a given response. Item fit indices are constructed from the residuals, and the smaller the residuals, the better the item fit. A complementary way to assess the item fit in IRT models is to examine the ordinal response categories of the items. The order of response categories should monotonically increase from low to high along the underlying trait. That is, individuals with more of the underlying trait should endorse higher response categories, while those with less of the underlying trait should endorse lower categories. Items with response categories that do not follow this pattern are considered disordered and should be examined further.

Sample Size Considerations

IRT models require large samples to obtain *stable* parameter estimates. Several factors are involved in sample size estimation and no definitive answer can be given. A rule of thumb is that a larger sample size is warranted when the number of item parameters to be estimated is large. The factors that will affect the number of item parameters to be estimated include the choice of IRT model, the number of response categories, and the number of items. Sample sizes of >100 for the basic Rasch model and 250 for the graded response model are usually sufficient for estimating the item parameters. But a sample size of up to 500 for the graded response model is recommended for stable parameter estimates.

Another factor to consider is the objective of the IRT analysis. A large sample size may not be needed to obtain an initial or descriptive picture of item performance, provided that a presentative sample is obtained. If the purpose is to obtain more robust estimates of the item parameters and use them to calculate person scores, generally a large sample size is required. Ideally, respondents should be spread fairly uniformly over the range of the underlying traits. At the minimum, there should be at least some people responding to each of the categories of every item to allow the IRT model to be fully estimated.

In this section, we highlight the matters on the model fit, item fit, and sample size considerations. More detail on these topics can be found elsewhere^[28,29] and in the references cited therein.

IRT MODELS

The unidimensional and locally independent assumptions allow the likelihood function for these IRT models to take the relatively simple form of a product over items:

$$L_i(\theta | \mathbf{Z}_i) = \prod_{j=1}^J p(z_{ij} | \theta) \quad (1)$$

where $\mathbf{Z}_i = (z_{i1}, \dots, z_{iJ})'$ represents the i th person's response pattern on an instrument composed of J items. The generic likelihood function of the IRT analysis is the same regardless of the IRT model used. It differs only in the specific form of the function $p(z_{ij} | \theta)$, where a particular IRT model is assumed. The commonly used IRT models are briefly described here in terms of their mathematical formulas.

Rasch Model

The simplest IRT model is the Rasch model or one-parameter logistic (1PL) model. Although the two models differ in their parameterization, they are interchangeable as the results of one can be obtained through linear transformation of the other. The Rasch model takes the following form:

$$P(X_j = 1 | \theta_i, b_j) = \frac{e^{\theta_i - b_j}}{1 + e^{\theta_i - b_j}} = \frac{1}{1 + e^{-\theta_i + b_j}}. \quad (2)$$

Again, the letter b denotes the location parameter. The Rasch model is applicable only to dichotomous items. In addition, the Rasch model assumes that all items are equally discriminating that is reflected by a constant slope parameter of 1.

2PL Model

On the other hand, the 2PL model does not assume that all items are equally discriminating. Therefore, each item has its own discrimination parameter. This discrimination parameter is denoted by the letter a . The following equation presents the formula of the 2PL model:

$$P(X_j = 1 | \theta_i, a_j, b_j) = \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}} = \frac{1}{1 + e^{-a_j(\theta_i - b_j)}}. \quad (3)$$

As mentioned earlier, the discrimination parameter is also called the slope parameter, because it determines the flatness or steepness of the item characteristic curves.

The bigger the value of the a parameter, the steeper the slope, and the higher the discriminatory power of the item. Like the Rasch model, the 2PL model is also only applicable to dichotomous items.

Partial Credit Model

The most useful IRT models for use in the instrument development are perhaps the polytomous IRT models. This distinction is owing to the fact that, in a typical health-care instrument, the items are often constructed with multiple response categories on an ordinal scale such as a Likert scale, verbal rating scale, or numeric rating scale. For example, it is more informative to know whether a patient is having no, mild, moderate, or severe pain than knowing whether a patient is having or not having pain.

For items that have more than two ordered response categories, the Master's partial credit model is often used.^[19] The Master's partial credit model is the polytomous extension of the Rasch model. Under the Master's model, the probability of a response in category k for an item with m_j categories can be written as follows:

$$P(X_{ij} = k | \theta_i, b_{j1}, b_{j2}, \dots, b_{j, m_j-1}) = \frac{\sum_{v=1}^k (\theta_i - b_{jv})}{1 + \sum_{g=1}^{m_j-1} \sum_{v=1}^g (\theta_i - b_{jv})} \quad (4)$$

and

$$P(X_{ij} = 0 | \theta_i, b_{j1}, b_{j2}, \dots, b_{j, m_j-1}) = \frac{1}{1 + \sum_{g=1}^{m_j-1} \sum_{v=1}^g (\theta_i - b_{jv})}. \quad (5)$$

As Eq. 4 shows, under the partial credit model, each item has $m - 1$ number of b parameters. However, unlike the b parameter in the Rasch or 2PL model, for the partial credit model, the b parameters are not the location where the probability is 0.5. Instead, the b parameters are where the probability is the same for two adjacent responses. Thus, the b parameters in the partial credit models are called the item-step (or item-threshold step) parameters to distinguish them from the location parameters. A special case of the partial credit model, the Andrich's rating scale model^[30] also assumes that the item-threshold steps are equal across all items. These three models—Rasch model, partial credit model, and rating scale model—can be called the Rasch family of models.

Generalized Partial Credit Model

The other commonly used polytomous IRT models are the generalized partial credit model^[20] and the graded response model.^[17] The generalized partial credit model, as its name

indicated, is a general form of the partial credit model. Specifically, in the generalized partial credit model, the discrimination parameters are not constrained to be unity; instead, each item has its own discrimination parameter like the 2PL model. That is, in addition to the $m - 1$ number of location parameters, each item has one additional discrimination parameter. The generalized partial credit model takes the following form:

$$P(X_{ij} = k | \theta_i, a_j, b_{j1}, b_{j2}, \dots, b_{j, m_j-1}) = \frac{e^{\sum_{v=1}^k a_j (\theta_i - b_{jv})}}{1 + \sum_{g=1}^{m_j-1} e^{\sum_{v=1}^g a_j (\theta_i - b_{jv})}} \quad (6)$$

and

$$P(X_{ij} = 0 | \theta_i, a_j, b_{j1}, b_{j2}, \dots, b_{j, m_j-1}) = \frac{1}{1 + \sum_{g=1}^{m_j-1} e^{\sum_{v=1}^g a_j (\theta_i - b_{jv})}}. \quad (7)$$

Graded Response Model

Samejima's^[17] graded response model is perhaps the most familiar polytomous IRT model that is used in the instrument development in health care. Formulated conceptually differently, the graded response model first dichotomizes the multiple responses to responses less than k versus responses of k or higher. The probability of observing a response of k or higher is monotonically increasing as the underlying trait increases. The logistic function used to describe this probability is the same as the IRT models for dichotomous item responses (i.e., 2PL). The probability of observing a response of k is then the difference between the probability of observing a response of k or higher and the probability of observing a response of $k + 1$ or higher:

$$P(X_{ij} = k | \theta_i, a_j, b_{j1}, b_{j2}, \dots, b_{j, m_j-1}) = \frac{1}{1 + e^{-a_j (\theta_i - b_{jk})}} - \frac{1}{1 + e^{-a_j (\theta_i + b_{j(k+1)})}} \quad (8)$$

and

$$P(X_{ij} = 0 | \theta_i, a_j, b_{j1}, b_{j2}, \dots, b_{j, m_j-1}) = 1 - \frac{1}{1 + e^{-a_j (\theta_i + b_{j1})}}. \quad (9)$$

With the graded response model, for each item, the item parameters are the discrimination parameter, a , and the $m - 1$ b_{ik} parameters. These b_{ik} parameters, called the category threshold parameters, show the point on the scale of the latent trait at which the probability passes 50% for endorsing a response of k or higher. The higher the category threshold parameter, the higher the level of the underlying trait required to pass 50% probability.

Choosing an IRT Model

The characteristics of the Rasch family of models are that they assume all items have uniform discriminatory power. Under the Rasch family of models, persons are ordered the same in terms of predicted responses regardless of the location of the items. Similarly, items are ordered the same in terms of predicted responses regardless of the underlying trait of the person. From the Rasch family of models' perspective, the discrimination parameter represents a person-by-item interaction, with the item curves crossing, making the ordering of persons or items no longer invariant. Under the Rasch family of models, only the b parameters need to be estimated. The b parameter of the Rasch family of models is similar to the percent correct or average score in the classical test theory. In IRT models, in order to estimate the item parameters, one has to assume the values or distribution of the underlying trait (i.e., θ). Unique to the Rasch family of models, the raw sum scores of the persons are the sufficient statistics of the θ . That is, the θ can be estimated sufficiently with the raw sum score alone. This property makes the sample size requirement smaller for the Rasch family of models than that for other IRT models.

Perhaps the greatest advantage of the Rasch family of models from the perspective of the users is that there exists a one-to-one correspondence between the observed sum score and the IRT θ score. For example, this property would enable clinicians and patients to directly link the PRO score to the content of the instrument and, therefore, facilitate interpretation of the score. With a direct correspondence of the observed sum score to the instrument's IRT scale score, the patient's condition can be immediately assessed without waiting for the assessment result to be reported back after a complex calculation of the IRT scale score. This is a very attractive property especially when a paper form of the instrument is used. However, the advantage of speedy scoring of the Rasch family of models is diminished nowadays by the ever-present personal electronic device, such as a tablet or a smartphone, where the complex calculation can be performed in real time.

The strength of the Rasch family of models is also its limitation. Uniform discrimination of all items may not be supported for some instruments or for certain conditions. For example, while patients with lung cancer may experience three major symptoms—fatigue, chronic pain, and shortness of breath—the symptom shortness of breath may be more discriminating between mild and severe patients than the symptom fatigue. In this situation, IRT models in the form of generalized partial credit model and graded response model, which do not assume uniformity across items with respect to the discrimination parameter, may be more appropriate. The advantage of these IRT models is that they sometimes fit the data better depending on the strength of the association between the items and the underlying trait. This property allows the further evaluation of

which items are more relevant than the other items. On the other hand, having to estimate the extra discrimination parameters for each of the individual items means a larger sample size is necessary.

Another advantage of the non-Rasch family IRT models is that they are suitable for use to develop computerized adaptive testing (CAT).^[2,31] In CAT, the estimate of an individual's score can be obtained with fewer items than normally would by using a fixed-length instrument. This is achieved by selecting the next item to be administered based on the estimated score from the responses of preceding items. By selecting items that are most informative and closest to the estimated level of the underlying trait, fewer items are needed to reach the desired precision of the score. The disadvantage is that CAT requires expertise in both IRT methodology and computer system implementation. In addition, to take the advantage of a CAT, an item bank with a large number of items is also needed. If an instrument has only few items, a fixed-length form may be more economical and feasible.

Choosing an IRT model requires good understanding of the underlying construct that the instrument is intended to measure, as well as the objective of measuring that construct. The feasibility (e.g., sample size) and the intended use (e.g., generic vs. disease-specific) are also the factors to be considered. Ultimately, the decisive factor should be whether the IRT model of choice can produce interpretable results for decision-making purposes. As mentioned earlier, items with multiple category responses are often used in health-care assessments because they yield more useful information. In this case, polytomous IRT models should be used, and the choice of IRT model comes down to the partial credit model, rating scale model, generalized partial credit model, or graded response model. Choosing between the generalized partial credit model and the graded response model is less of an issue as the two models generally yield very similar results and won't affect the findings; the choice often is based on the researcher's familiarity with each of the models.

APPLICATION OF IRT IN HEALTH-CARE ASSESSMENT

The growing emphasis on the patient-focused care and treatment development has accelerated the demand and use of well-developed, reliable, and fit-for-purpose health-care assessments. These assessments take many forms; they may be oral or written, or they may require performance. These assessments may be administered electronically or on paper, and may be self-reported or administered by an interviewer. The assessments may be reported by the patients themselves, clinicians, or observers such as patients' caregivers. Regardless of the form, administration mode, or type of assessments, the measures must be valid and the results must be interpretable, especially when

critical health-care decisions are to be based on the results of these assessments.

To develop a new assessment that is well defined and reliable, one starts with the basic unit of an assessment: the items themselves. Well-planned assessment development typically begins with defining the context-of-use that includes the identification of the target population, concepts of interest, and intended use. Frequently, developing a conceptual framework and an endpoint model is the best way to define and visualize the context of use of an assessment.^[1] The next step is to determine which appropriate format, administration mode, and type of assessment are appropriate for the intended context of use. The next step in the development process is to create the items.

Once the items created are deemed measuring the intended concepts by exhibiting content validity, quantitative data should be collected in order to examine the performance of these items. First, from the item response frequency distribution, we know how the response categories are used by the individuals and how many missing item responses there are. Through examination of the IRT unidimensional assumption we know whether the items measure the same construct or multiple constructs. This information will help confirm the conceptual framework or how it should be modified.

If the Rasch family of model is used, through the item model fit statistics we will know which items do not conform to the uniform discrimination property. If non-Rasch family polytomous IRT models are used, the discrimination parameters inform about the strength of the association of the items with the underlying trait, which provide insights into how well each of the items discriminates for various levels of the underlying trait.

Next, by examining the categorical probability curves, we are able to discover whether the response categories are ordered correctly on the underlying trait. The lower response categories should occupy the lower range on the trait scale, and the higher response categories the higher range. The upper panel of Fig. 5 shows the categorical probability curves where the response categories are positioned in the correct monotonic order. In contrast, the lower panel of Fig. 5 shows the curve of response category 1 is completely overlapped with the curve of response category 0, indicating that one of these categories is ordered incorrectly.

Finally, the item-step parameters or the category threshold parameters show the range on the underlying trait that the items measure and whether there are gaps on the scale that we do not have items to measure them. This step can also be aided by plotting the test information function curve. Test information function curve helps visualize where the measurement on the underlying trait is most precise and where it is not. Altogether, through the IRT application, we obtain the information that enables us to select or remove items, or to add new items, so that the assessment covers the target range on the underlying trait with desirable precision.

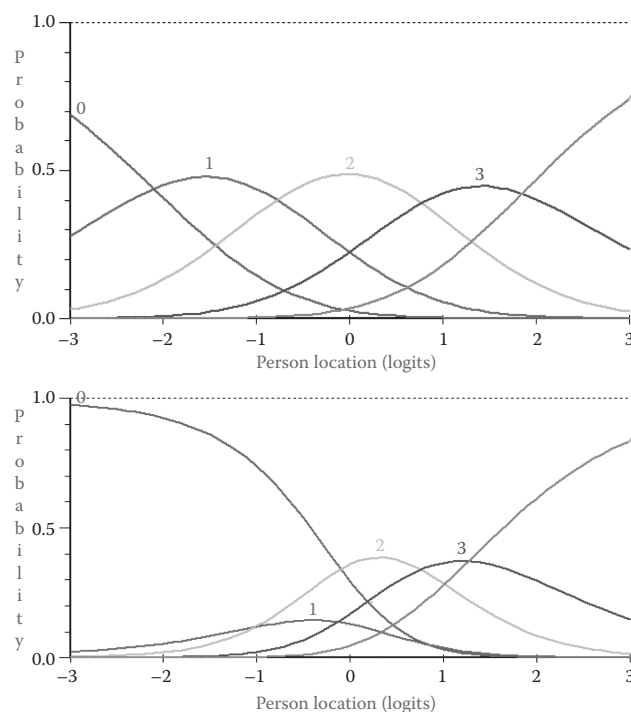


Fig. 5 Categorical probability curves of two hypothetical items. The upper panel shows an item with correctly ordered response categories. The lower panel shows an item with incorrectly ordered response categories

CONCLUSION

This article serves as an introduction to the IRT methodology for researchers who are interested in the development of health-care measures. IRT provides a detailed analysis of the item performance that enables the selection of well-performed items to ensure and enhance the measurement properties of the health-care measures to be developed. IRT is especially suitable when the number of items is large. However, it also requires a large sample size to yield robust findings. Choice of an IRT model depends on the construct to be measured, the purpose of developing the tool, and the practical considerations such as the sample size. Its use in the health outcome assessment has increased rapidly. The growing experience and continuing research will lead to improvement and wider application of the IRT methodology in the health-care field. This, in turn, will lead to better tools for assessing health outcomes and health status and, ultimately, help better communication of treatment benefit and decision-making for the patients and clinicians.

REFERENCES

1. US Department of Health and Human Services Food and Drug Administration. *Guidance for Industry: Patient-reported Outcome Measures: Use in Medical Product Development*

- to Support Labeling Claims; Department of Health and Human Services Food and Drug Administration: Silver Spring, MD, 2009. www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf.
2. Hay, R.D.; Morales, L.S.; Reise, S.P. Item response theory and health outcome measurement in the 21st century. *Med. Care* **2000**, *38* (9), II28–II42.
 3. Cappelleri, J.C.; Lundy, J.J.; Hays, R.D. Overview of classical test theory and item response theory for the quantitative assessment of items in the developing patient-reported outcomes measures. *Clin. Ther.* **2015**, *36* (5), 648–662.
 4. Nguyen, T.H.; Han, H.R.; Kim, M.T.; Chan, K.S. An introduction to item response theory for the patient-reported outcome measurement. *Patient* **2014**, *7* (1), 23–35.
 5. Thurstone, L.L. A method of scaling psychological and educational tests. *J. Educ. Psychol.* **1925**, *16*, 433–451.
 6. Jones, P.W.; Chen, W.H.; Wilcox, T.K.; Sethi, S.; Leidy, N.K. Characterizing and quantifying the symptomatic features of COPD exacerbations. *Chest* **2011**, *139* (6), 1389–1394.
 7. Lord, F.M. *A Theory of Test Scores* (Psychometric Monograph No.7); Psychometric Society: Richmond, VA, 1952.
 8. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright; The University of Chicago Press: Chicago, IL, 1960/1980.
 9. Lazarsfeld, P.F. The logical and mathematical foundations of latent structure analysis. In *The American Soldier: Studies in Social Psychology in World War II*, Vol. IV. Measurement and Prediction; Ed. Stouffer, S.A.; Princeton University Press: Princeton, NJ, 1950, 362–412.
 10. Lord, F.M.; Novick, R.M. *Statistical Theories of Mental Test Scores*; Addison-Wesley: Reading, MA, 1968.
 11. Birnbaum, A. Some latent traits models and their use in inferring an examinee's ability. In *FM Lord & MR Novick, Statistical Theory of Mental Test Scores*; Addison-Wesley: Reading, MA, 1968.
 12. Bock, R.D.; Lieberman, M. Fitting a response model for n dichotomously scored items. *Psychometrika* **1970**, *35* (2), 179–197.
 13. Bock, R.D.; Aitkin, M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **1981**, *46* (4), 443–459.
 14. Shilling, S.; Bock, R.D. High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika* **2005**, *70*, 533–555.
 15. Cai, L. High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika* **2010**, *75*, 33–57.
 16. Cai, L. Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *J. Educ. Behav. Stat.* **2010**, *35*, 307–335.
 17. Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika. Monograph No. 17, 34* (4, Pt. 2), **1969**.
 18. Bock, R.D. Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. *Psychometrika* **1972**, *37*, 29–51.
 19. Masters, G.N. A Rasch model for partial credit scoring. *Psychometrika* **1982**, *47*, 149–174.
 20. Muraki, E.A. Generalized partial credit model: Application of an EM algorithm. *Appl. Psychol. Meas.* **1992**, *16*, 159–176.
 21. Roberts, J.; Donoghue, J.; Laughlin, J. A general item response theory model for unfolding unidimensional polytomous responses. *Appl. Psychol. Meas.* **2000**, *24*, 3–32.
 22. Ellwood, P.M. Shattuck lecture—Outcomes management. A technology of patient experience. *N. Engl. J. Med.* **1988**; *318*, 1549–1556.
 23. McHorney, C.A. Generic health measurement: Past accomplishments and a measurement paradigm for the 21st century. *Ann. Intern. Med.* **1997**, *127*, 743–750.
 24. Hambleton, R.K.; Swaminathan, H.J.; Rogers, H.J. *Fundamentals of Item Response Theory*. Sage Publications: Newbury Park, CA, 1991.
 25. Cella, D.F.; Dineen, K.; Arnason, B.; Reder, A.; Webster, K.A.; Karabatsos, G.; Chang, C.H.; Lloyd, S.; Steward, J.; Stefanski, D. Validation of the functional assessment of multiple sclerosis quality of life instrument. *Neurology* **1996**, *47*, 129–139.
 26. Revicki, D.A.; Cella, D.F. Health status assessment for the twenty-first century: Item response theory, item banking, and computer adaptive testing. *Qual. Life Res.* **1997**, *6*, 595–600.
 27. Chen, W.H.; Thissen, D. Local dependence indices for item pairs using item response theory. *J. Educ. Behav. Stat.* **1997**, *22*, 265–289.
 28. Cappelleri, J.C.; Zou, K.H.; Bushmakina, A.G.; Alvir, J.M.J.; Alemayehu, D.; Symonds, T. *Patient-Reported Outcomes: Measurement, Implementation and Interpretation*. Chapman & Hall/CRC Press: Boca Raton, FL, 2013.
 29. Embretson, S.E.; Reise, S.P. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates: Mahwah, NJ, 2000.
 30. Andrich, D. A rating formulation for ordered response categories. *Psychometrika* **1978**, *43*, 561–573.
 31. Wainer, H.; Dorans, N.J.; Eignor, D.; Flaugh, R.; Green, B.F.; Mislevy, R.J.; Steinberg, L.; Thissen, D. *Computerized Adaptive Testing: A Primer*, 2nd Ed.; Lawrence Erlbaum Associates: Mahwah, NJ, 2000.

IVRS/IWRS for Randomization

Mon-Gy Chen

Department of Biostatistics, Amgen, Thousand Oaks, California, U.S.A.

Abstract

The interactive voice response system (IVRS) and the interactive web response system (IWRS) have applications ranging from widely used banking services to clinical trials. This entry describes the application of IVRS/IWRS to patient randomization and drug management in clinical trials.

Keywords: Randomization; Telephony; Voice response; Web response.

INTRODUCTION

The rapid development of the Internet and web browsers as well as computer telephony technology around the turn of the twenty-first century has enabled the automatic interchange of information between a computer and a digital phone or a web-based system. It has many applications in industry. The interactive voice response system (IVRS) and the interactive web response system (IWRS), for example, have applications ranging from widely used banking services to clinical trials.

The IVRS uses a phone as an input device and responds with a computer-controlled, high-quality digitized human voice whereas the IWRS uses the Internet and a web browser. The multilingual computer response capability enables IVRS/IWRS to function worldwide 24 hours a day and to collect real-time online quality data. These features make IVRS/IWRS an ideal tool for patient randomization and drug management for both national and international clinical trials.

OVERVIEW

This entry describes the application of IVRS/IWRS to patient randomization and drug management in clinical trials. We start with a description of the traditional randomization and drug management procedure, followed with a description of the application of IVRS/IWRS to mapping patient randomization to the drug management system, and end with a description of a stand-alone patient randomization system, including a minimization procedure.

For our discussion, “randomization” is defined as a process by which each subject has an equal chance or a prespecified probability of being assigned to one of the treatments in a study. This process generates a randomization schedule containing patient numbers and the corresponding treatment assignments.

The drug management system involves several activities: packaging of clinical supplies, shipping them from a warehouse to study sites, and dispensing them to patients by a third party. Unless otherwise specified, a double-blind randomized study is assumed. The terms “drug” and “treatment” are used interchangeably.

TRADITIONAL APPROACH TO RANDOMIZATION AND DRUG MANAGEMENT

In a double-blind randomized clinical trial, much pre-study preparation needs to be completed before a site can enroll and randomize the first eligible patient. One such preparation involves the design and generation of a randomization schedule and a plan to manage and ensure an adequate drug supply throughout the study. The traditional approach, which links a patient number to the clinical supply containing the treatment for that patient, is described as follows:

1. The statistician generates the randomization schedule according to the specifications in the final protocol. The end product is a schedule containing patient identification numbers and the corresponding treatment assignments. The total number of patients generated in the randomization schedule depends on the sample size required for the study and the projected number of dropouts.
2. A copy of the validated randomization schedule is sent to the Clinical Supply Group.
3. The Clinical Supply Group prepares and packages the drug supply according to the randomization schedule. In general, an extra drug supply of at least 20% is recommended to cover any potential drug wastage during the study.
4. Individual patient kits, each of which contains the treatment assigned by the randomization process,

are prepared. Each kit is labeled with a patient number only. The treatment code is either printed on the blinded portion of a two-part label or kept in a sealed envelope and secured by a third party at the site or at a central location. In either case, the seal or the blind for a given patient can be broken only in case of a medical emergency and for that patient only.

5. A prespecified, fixed quantity of drug supply is shipped to a given site before the first patient is randomized. If a permuted-block randomization is used within a site, the drug supply sent to a site needs to be in multiples of block size. A permuted block of size 4 in a two-treatment randomization results in two repeats of the two treatments in any random order.
6. When the first patient becomes eligible at a site, the kit labeled Patient Number 1 is dispensed by the pharmacist or designee at that site. The next patient receives the kit labeled Patient Number 2. This process continues until the last patient has been randomized at the site or until study termination.
7. If the number of patients enrolled at a site exceeds the drug supply, then an additional drug supply needs to be shipped in a timely fashion before the next patient is ready for randomization. On the other hand, if a site enrolls fewer patients than the drug supply, then the remaining drugs will be wasted.

The process works as follows. Assume a study has a randomized, double-blind, parallel-group design. Two treatments, A and B, are to be evaluated in a total of 16 patients at two study sites. A permuted-block design with a block size of 4 is used. That is, for every four patients randomized, two are assigned to treatment A and two are assigned to treatment B. According to these specifications, a randomization schedule is generated (Table 1).

The Clinical Supply Group, upon receiving the randomization schedule, packages eight kits containing drug A, labeled Patient Number 1, 4, 5, 8, 9, 11, 14, or 15, and the other eight kits containing drug B, labeled Patient Number 2, 3, 6, 7, 10, 12, 13, or 16. Because each center plans to randomize eight patients, the kits labeled Patient Numbers 1 to 8 are shipped to site 1 and the kits labeled Patient

Numbers 9–16 are sent to site 2. Note that the number of patients enrolled at each site has to be in multiples of the block size to maintain treatment balance.

The first patient randomized at site 1 is dispensed the kit labeled Patient Number 1 and therefore receives treatment A. The first randomized at site 2 is dispensed the kit labeled Patient Number 9 and therefore receives treatment A. The second patient randomized at sites 1 (Patient Number 2) and 2 (Patient Number 10) receive treatment B. This process continues until the enrollment is complete or the study is terminated for other reasons.

In this example, each patient in the study has a unique patient number. Alternatively, each site can number the patient starting with 1 according to the chronological order of entry. In this case, the unique patient number for the study consists of the site number and the patient number. Applying this method, Patient Numbers 9–16 at site 2 in the example would now become Patient Numbers 201–208. The leading digit “2” identifies the site number and the second and third digits are for numbering the patients within a site. Thus the patients at site 1 will now have Patient Numbers 101–108. In this numbering system, it is important to have both the site number and the patient number preprinted on the label of the drug supply.

PROBLEMS ASSOCIATED WITH THE TRADITIONAL APPROACH

A number of problems arise with the traditional approach to randomization and drug management. One concerns the wastage of drugs as a result of the inability to share drug supplies between sites; another is the inability to ensure treatment balances across sites.

In our example, suppose that only two patients were randomized, one from each site. Only two kits would be used. Because 14 of the total 16 kits would not be used, the overall wastage for the study would be 87.5% (14/16). According to the randomization schedule, both patients enrolled would be assigned treatment A. Ideally, when a study enrolls two patients; it is preferable to have both treatments represented. In order to accommodate this type of treatment balance across sites, a central randomization is needed. Under the traditional approach, it would be impossible to balance the treatment codes across all sites. Because the rate of enrollment in clinical trials varies (some sites enroll patients rapidly whereas others may enroll a few or none at all), the traditional approach does not have a mechanism to allow for adjusting the drug supply when a site does not perform (i.e., drug wastage) or a site over enrolls (i.e., drug shortage).

These problems are readily resolved by using IVRS/IWRS. The wastage problem can be overcome by assigning patient numbers independent of the supply kits and dynamically matching them through IVRS/IWRS at the time of randomization. The IVRS/IWRS, which is accessible to

Table 1 Randomization Schedule

Patient	Treatment	Patient	Treatment
1	A	9	A
2	B	10	B
3	B	11	A
4	A	12	B
5	A	13	B
6	B	14	A
7	B	15	A
8	A	16	B

all sites at all times, is ideal for providing central randomization to check treatment balance across sites. The following section describes the use of IVRS/IWRS in drug management and patient randomization.

HOW DOES THE INTERACTIVE VOICE/WEB RESPONSE SYSTEM WORK?

The interactive voice/web response system has the capability of linking information from different tables stored in the relational database and performing queries using common identifiers. With this capability, a randomization schedule such as [Table 1](#) can be separated into many different tables and connected by IVRS/IWRS at the time of randomization. This mapping process allows IVRS/IWRS the flexibility to perform many special functions unmatched by the traditional randomization approach. The following section explains some of these features and the connection of tables by IVRS/IWRS.

DRUG SUPPLY AND PACKAGING

As already mentioned, if the drug supply can be packaged separately, independent of the patient numbers and the randomization schedule, with only the minimal quantity shipped to a site when the site is ready to randomize the first patient, the savings in the drug supply can be enormous. The Clinical Supply Group can package the drug as soon as the total supply of each drug type is known.

Let us apply this approach to our earlier example, in which 16 kits need to be packaged. The Clinical Supply Group, without the knowledge of the randomization schedule, can package the drug in two groups: eight kits for treatment A and eight kits for treatment B. The kits can be labeled with any numbering system unrelated to the patient number. The simplest method is to label the eight kits containing treatment A with consecutive numbers d01–d08 and the eight kits containing treatment B with numbers d09–d16. However, this method, although straightforward, risks revealing the treatment codes through clever guesswork.

An alternative approach is to scramble the kit numbers before treatment assignment. One such sequence for 16 kits might be d02, d05, d10, d06, d13, d09, d12, d08, d16, d011, d07, d03, d014, d04, d15, and d01. The next step is to assign the first eight numbers to kits containing treatment A and the last eight kits containing treatment B. After the assignment, the kit numbers within each treatment group can be sorted in an ascending order for easy packaging and assembling. The resulting kit numbers are displayed in [Table 2](#).

After the kit numbers have been assigned, the Clinical Supply Group can package the drugs accordingly and store them in a warehouse ready for shipment. At this

Table 2 Kit Numbers

Treatment A	Treatment B
d02	d01
d05	d03
d06	d04
d08	d07
d09	d11
d10	d14
d12	d15
d13	d16

point, the kits are not associated with any patient number or any study site. When a site is anticipating the first eligible patient, a minimal quantity containing equal numbers of treatment A and treatment B is shipped to the site. The minimal quantity of an initial shipment depends on a number of factors—it must take into account the enrollment rate at each site, and, preferably, it must be in multiples of the block size in the case of a permuted-block design. If a site can enroll four patients a day, the initial shipment has to be eight kits or more to ensure an adequate drug supply. In our example, the enrollment is assumed to be one patient per week at all sites; therefore an initial shipment of four kits, two treatment A and two treatment B, is acceptable. The first site to request shipment receives kit numbers d01, d02, d03, and d05, and the next site receives kit numbers d04, d06, d07, and d08; the contents, however, remain blinded to the site. In theory, an initial shipment of two, one for each treatment, is sufficient; however, this is not recommended in practice because of the potential risk of revealing the drug codes by the staff at the site.

As soon as the file containing kit numbers is available, as in [Table 2](#), it is loaded into the IVRS/IWRS relational database, which is accessible through a phone or website. When a site is ready for randomization, IVRS/IWRS notifies the warehouse to ship the initial supply to the site. The kit numbers and the associated site number are recorded in the IVRS/IWRS database. The system is then ready to manage (i.e., dispense or resupply) the drug supply.

When the drug supply is low, IVRS/IWRS will trigger the warehouse to resupply the site. For a slow-enrolling site, the trigger point may be one of each treatment type remaining at the site. For a fast-enrolling site, the trigger point depends on the enrollment rate. For example, if a site enrolls four patients a day, the trigger point has to be at least four kits remaining at a site, two of each drug type. The quantity of resupply also depends on the enrollment rate and how close it is to the study completion. When the study approaches its completion, no more resupply is needed. The trigger point and the quantity of resupply must be established when the IVRS system is being developed for the study.

RANDOMIZATION SCHEDULE

In the traditional approach, the randomization schedule (see [Table 1](#)) is used for packaging, shipping, and dispensing; a patient number is associated with a kit. With IVRS/IWRS, the same schedule is used, but the process is more flexible and has additional features.

One advantage of using IVRS/IWRS is its ability to separate patient numbers from the randomization codes and to match them dynamically at the time of randomization. The randomization codes serve as a bridge between the treatment assignment and any patient numbering system. This bridge is formed at the time of randomization. For example, [Table 3](#) uses the same randomization schedule as [Table 1](#), but the patient numbers are now randomization codes; these codes are not associated with patient numbers.

Assume that the patients are sequentially numbered at each site, starting with 1. When the first patient at site 1 (Patient 101) is ready for randomization, the site calls IVRS or enters the required information on the web page, which would then search through [Table 3](#) stored in the IVRS/IWRS database for the next available code and the treatment it is linked to. In this case, code 1 is available; hence Patient 101 is assigned to treatment A. If the next patient is from site 2 (Patient 201), this patient is linked to code 2 by IVRS/IWRS and treatment B is assigned. Should the study be terminated at this point, there will be two patients from the two sites, one assigned to treatment A and the other to treatment B.

Thus a central randomization is achieved. In this type of randomization, all the randomization codes are shared by all patients from all sites. The end result is a central randomization, with a balanced treatment across all sites but not necessarily within one site.

If, however, the study is designed to have treatment balance within a given center, then the IVRS/IWRS database needs to be constructed so that codes 1–8 are reserved for site 1 and codes 9–16 for site 2. With this new structure, the second patient in our example (Patient 201) would be linked to code 9, receiving treatment A, instead of code 2, receiving treatment B.

Table 3 Randomization Codes

Code	Treatment	Code	Treatment
1	A	9	A
2	B	10	B
3	B	11	A
4	A	12	B
5	A	13	B
6	B	14	A
7	B	15	A
8	A	16	B

MAPPING BY INTERACTIVE VOICE/WEB RESPONSE SYSTEM

After the clinical supply list (see [Table 2](#)) and the random assignment of treatment (see [Table 3](#)) are separated and available, they are stored in the IVRS/IWRS relational database. The IVRS/IWRS can query the database and connect these tables at the time of randomization and respond to the caller with real-time randomization and drug assignment. This mapping process is described as follows:

1. To ensure system security, an entry from any site must first pass stringent system and study-specific security measures (e.g., site identification, password, caller identification) to continue the dialog with IVRS/IWRS. For reference, a sample IVRS script is given in the [appendix](#) to this entry.
2. Through the phone keypad or the web page, any pre-specified study and patient information can be collected, which may include site and patient numbers, patient eligibility, and other patient characteristics, into the IVRS/IWRS database.
3. As soon as the eligibility is verified, IVRS/IWRS searches the randomization schedule ([Table 3](#)) for the next available randomization code and the corresponding treatment.
4. Once the treatment is identified, IVRS/IWRS then searches the drug supply list ([Table 2](#)) for the next available kit number containing the treatment at that site.
5. The IVRS, through the voice response system, informs the caller which kit number to dispense, without revealing the treatment code. In case of IWRS, the dispensing information is delivered via an e-mail or on the screen.
6. The randomization procedure is complete when the caller confirms the kit number by means of the phone or by e-mail. The entire process takes no more than two minutes.
7. The IVRS/IWRS stores the site, patient, and kit numbers as well as the treatment assignment in the database.
8. The IVRS/IWRS checks the drug supply at the site to assess the need for resupply.

Let us now illustrate by extending our earlier example. [Fig. 1](#) provides a graphic presentation of the mapping process for this example. Let us assume the study calls for a central randomization and a patient numbering system by site. The randomization schedule in [Table 3](#) and the kit list in [Table 2](#) are stored in the IVRS/IWRS database. Because of a slow enrollment rate, an initial shipment of four kits is provided to each site: kits d01, d02, d03, and d05 are sent to site 1, and d04, d06, d07, and d08 are sent to site 2. The trigger point for resupply is set at the time when two kits (one of each treatment) still remain at the site.

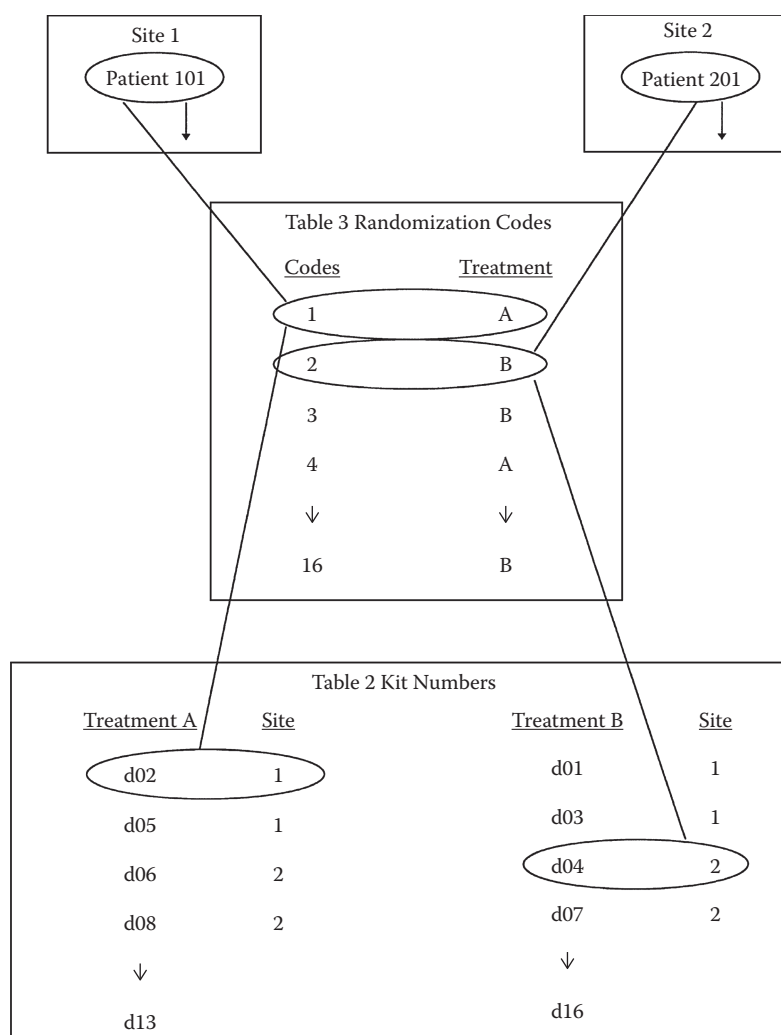


Fig. 1 Relational database for randomization and drug assignment

If the first entry is from site 1 (Patient Number 101), IVRS/IWRS matches this patient to the first available randomization code in Table 3, i.e., code 1, which is linked to treatment A. The IVRS/IWRS then searches Table 2 for the next available kit containing treatment A in site 1, i.e., d02. The IVRS/IWRS informs the site to dispense kit d02 without revealing its contents. The double-blind status of the study is therefore maintained. After confirmation that kit d02 has been dispensed, IVRS/IWRS checks the trigger point for resupply. Because the site still has three kits, IVRS/IWRS will not generate a resupply notice.

Applying the same approach, if the next patient comes from site 2 (Patient Number 201), IVRS/IWRS maps this patient to code 2 in Table 3 and hence assigns treatment B and kit d04. Again, no replacement kit is needed.

If the study is terminated after two patients have been randomized, a maximum of eight kits would have been shipped to the sites and only two would have been used. The number of kits wasted would be six, which would be much lower than the 14 kits under the traditional approach in the earlier example. The percentage of wastage would

be 37.5% (6/16) vs. 87.5% (14/16). In addition to saving drug supply, the treatment would be balanced (one treatment A and one treatment B) by the central randomization capability of IVRS/IWRS although only two patients were enrolled. In traditional randomization, treatment balance across sites can happen only by chance, not by design.

In all of the examples given, the size of the study (i.e., number of patients and sites) is kept small for easy illustration and understanding. In real studies, the numbers of sites and patients are much larger, which generally would lead to a much greater saving in drug supply.

PATIENT RANDOMIZATION USING THE INTERACTIVE VOICE/WEB RESPONSE SYSTEM

In addition to the dynamic mapping of patient number, randomization code, drug assignment, and drug supply, IVRS/IWRS also performs stand-alone patient randomization for various randomization methods. The most commonly used

methods in clinical trials include nonstratified randomization, stratified randomization, and minimization. The implementation of these methods using the IVRS/IWRS is described next.

NONSTRATIFIED RANDOMIZATION

Whether it is a permuted-block or a simple randomization, the implementation is straightforward in a nonstratified randomization because it involves only one randomization schedule. The IVRS/IWRS, upon verification of the entry and the collection of patient eligibility, connects the patient to the next available treatment code in the randomization schedule.

This type of randomization is commonly used in randomized large-scale, open-label studies. Large-scale studies would have a greater chance of achieving treatment balance within a stratum because of sufficient number of patients enrolled in each stratum. In open-label trials, treatment assignment is known to both patient and investigator; therefore it is likely to be subject to selection bias (e.g., the investigator may choose to assign the active treatment rather than a placebo to a more severely ill patient). The IVRS/IWRS is especially valuable in these studies because the treatment is assigned at the moment when the phone call is completed or the web page information is submitted and therefore the selection bias is avoided.

STRATIFIED RANDOMIZATION

One of our earlier examples involved treatment balance within each study site. Study site is considered a stratification factor in this case because of potential “institution effect” in carrying out the trials. Many clinical trials, especially in cancer research, involve multiple prognostic factors that could affect treatment outcome. Traditionally, treatment balance within a stratum is achieved by generating a permuted-block randomization schedule for each stratum. All such schedules are stored in the IVRS/IWRS database. The IVRS/IWRS, upon collecting patient information, first matches the stratum and then searches for the next available randomization code in that stratum.

For example, in the treatment of stress incontinence, the baseline incontinence episode frequency (IEF) and gender are thought to have impact on the treatment outcome; therefore treatment balance within each stratum defined by baseline IEF and gender is deemed necessary. Because there are three levels of baseline IEF (mild, moderate, and severe) for each gender, a total of six baseline strata is considered for randomization. A randomization schedule is generated for each of the six strata, and all six schedules are stored in the IVRS/IWRS database.

When a site makes an entry to randomize a patient, IVRS/IWRS first collects the stratification information through the phone or the webpage. If the patient is female with moderate baseline IEF, IVRS/IWRS will search for the next available randomization code in the randomization schedule for the stratum of female with moderate baseline IEF, and will assign the corresponding treatment to that patient.

One difficulty of using this procedure is that the number of strata increases geometrically with the number of prognostic factors and their levels. For example, three prognostic factors, each with four levels, will generate 64 strata ($4 \times 4 \times 4$). As the number of strata increases, the ratio between the total number of patients and the number of strata becomes smaller. In other words, the number of patients in each stratum could be so small that in most strata the first permuted block of treatments would not be completely assigned. The result would be a considerable imbalance of treatment assignments across all stratification factors.

Although this procedure has traditionally been used, it does have limitations in balancing treatments simultaneously across several prognostic factors. Minimization procedures that overcome these difficulties have since been proposed by Taves^[1] and expanded by Pocock and Simon.^[2] The implementation of a minimization procedure by IVRS/IWRS is described and illustrated in an example in the next section. The manner in which stratification can be incorporated into the analysis needs to be determined prior to the use of this method.

MINIMIZATION

Minimization is a method of assigning patients to treatments in order to minimize the differences between the treatment groups in the number of patients as well as in patient characteristics and prognostic factors. Although the minimization method is mathematically straightforward, it had not been widely used for decades following its inception because of the difficulties in implementing it on a routine basis without an interactive computer. Therefore the introduction of IVRS/IWRS has made this dynamic procedure a popular one for clinical trials in recent years because it assigns treatment on an ongoing basis and does not require a prespecified randomization schedule. This procedure is best illustrated by the following example.

Suppose that a trial for controlling stress incontinence is conducted in five centers and that there are two treatment groups (active treatment and placebo) to be balanced with regard to baseline IEF, gender, and centers. A total of 93 patients has already been entered. The distribution of these patients by the three stratification factors is given in [Table 4](#). This table is stored in the IVRS/IWRS relational database. The IVRS/IWRS collects

Table 4 Number of Patients by Stratification Factors for the First 93 Patients

Stratification factor	Active	Placebo
Sex		
Male ^a	24	25
Female	23	21
Baseline IEF		
Mild	19	20
Moderate	20	19
Severe ^a	8	7
Site		
1 ^a	7	8
2	9	8
3	13	12
4	14	12
5	4	6

^aStrata for Patient 94.

the stratification factors associated with the next patient: Patient 94 is from site 1 and is a male with severe baseline IEF. For each treatment group, IVRS/IWRS adds up the number of patients who had already been randomized and who shared the same baseline characteristics as Patient 94 and assigns the patient to the group having the lower total.^[3] In this example, the sum for the active group is 39 (24 + 8 + 7) and for the placebo group is 40 (25 + 7 + 8). Therefore Patient 94 is assigned to the active group, thus minimizing the imbalance in the number of patients between the two treatment groups across all the stratification factors. Note that the marginal total is used for decision making in this example simply for easy illustration; other criteria for decision making can be found in other references.^[1,3]

In this example, Patient 94 is assigned to the active drug with a probability of 1. Alternatively, a less “deterministic” approach can be applied, which assigns the new patient to the treatment with a lower total with a probability of less than 1 (e.g., 0.75 or 0.95). For this type of randomization, IVRS/IWRS uses a uniform random-number generator to perform the task. Although the random numbers generated by a computer are in fact pseudorandom numbers, they are able to serve the purpose of randomization without compromising the integrity of the process. Let us extend our current example and assume that the probability is set at 0.80. A list of uniform random numbers, ranging from 0.00–0.99, is generated. The IVRS/IWRS searches the list for the next random number and assigns Patient 94 to active treatment if the next available random number is between 0.00 and 0.79, inclusively, and to placebo otherwise.

As a general practice, minimization should not be applied until a given number of patients has been randomized using a permuted-block design so that the risk of

revealing the treatment codes of the first few enrolled can be avoided.

OTHER INTERACTIVE VOICE RESPONSE SYSTEM APPLICATIONS

There are several additional capabilities of IVRS/IWRS besides randomization and drug management. One is the collection of real-time quality data in clinical trials. For example, the patient diary in clinical trials has been a cumbersome, handwritten process subject to human error and missing data. Using IVRS/IWRS, quality data can be obtained directly from the patient. The IVRS/IWRS can also remind the patient to enter the diary data so that missing data can be avoided, and a high patient compliance can also be expected. This has proven to be a major cost savings because of the ease of data processing as well as the accelerated availability of quality data.

In addition, in long-term-safety studies, IVRS/IWRS can collect and provide real-time crucial safety information, such as serious adverse events and mortality data. The fast and accurate information provided by IVRS/IWRS has proven to be valuable for monitoring the safety of clinical trials. Other operational aspects of clinical trials, such as patient enrollment and withdrawal status, can also be collected through IVRS/IWRS. This type of real-time data is very useful for strategic planning (e.g., enrollment may need to increase if the dropout rate is higher than expected in the protocol).

CONCLUSION

The interactive voice/web response system is an application based on the rapidly growing Internet and web-based system together with the computer telephony technology. Its multilingual capability and its ability to function without human intervention provide services at any time and place. The computer-based technology makes worldwide access possible, therefore making it an ideal tool for central randomization and drug management in clinical research. This entry has described the application of IVRS/IWRS in randomization and drug management. One of the great advantages of using IVRS/IWRS in managing the clinical drug supply is the reduction of drug wastage. Our example demonstrated a reduction in drug wastage from 87.5% under the traditional approach to 37.5% using IVRS/IWRS. In addition to saving drug supply, IVRS/IWRS has the ability to centrally randomize the patients from all sites so that treatment balance can be achieved across all sites for the entire study. The IVRS/IWRS technology can also provide clinical researchers with a new and cost-effective method for collecting real-time quality data and for providing critical monitoring information at any time and any place a digital phone or the Internet/web browser is available.

Appendix: Sample Interactive Voice Response System Script for Randomization and Drug Management

General

- 1.1 IVRS: Please enter your identification number.
- 1.2 CALLER: XXXXXXXX
- 1.3 IVRS: Please enter your access code.
- 1.4 CALLER: XXXX

Language is determined by IVRS from identification number and access code.

- 1.5 IVRS: Welcome to the X Company automated randomization and drug assignment system for the BONE study. For computer-automated help at any time, please press the “star” key, or to speak with the help desk, please press 0.

Upon leaving the system:

- 1.6 IVRS: Thank you for using the X Company automated randomization and drug assignment system for the BONE study. Good-bye.

These prompts are played for every caller. They are used to determine who the caller is and verify security clearance to main menu.

Main Menu

These prompts provide the caller with a list of system options.

- 2.1 IVRS: To screen a patient, press 1.
- 2.2 IVRS: To record a patient screen failure, press 2.
- 2.3 IVRS: To randomize a patient, press 3.
- 2.4 IVRS: To assign drug, press 4.
- 2.5 IVRS: To record a patient discontinuation, press 5.
- 2.6 IVRS: To confirm a drug shipment, press 6.
- 2.7 IVRS: To review patient information, press 7.
- 2.8 IVRS: To exit, press 9.

REFERENCES

- 1. Taves, D.R. Minimization: A new method of assigning patients to treatment and control groups. *Clin. Pharmacol. Ther.* **1974**, *15*, 443–453.
- 2. Pocock, S.J.; Simon, R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* **1975**, *31*, 103–115.
- 3. Reed, J.V.; Wickham, E.A. Practical experience of minimization in clinical trials. *Pharm. Med.* **1988**, *3*, 349–359.

Japan: Ministry of Health, Labour and Welfare and Pharmaceutical Administration

Kiyohito Nakai

Ministry of Health, Labour and Welfare, Tokyo, Japan

Abstract

Under the MHLW, a wide range of affairs from clinical trials and approval review to measures for safety in the postmarketing stage is now handled to secure and improve the health and livelihood of people in Japan. This entry details the role of the MHLW in the lives of Japanese citizens and the Japanese pharmaceutical industry.

INTRODUCTION

Advances in science and technology and changes in domestic and international situations have made pharmaceutical affairs even more complicated. As the social needs and demands arose, the system of pharmaceutical administration in Japan has been dramatically changed in the last decade, including the amendment of the Pharmaceutical Affairs Laws in 1996 and the establishment of the Pharmaceutical and Medical Devices Evaluation Centre (PMDEC) under the National Institute of Health Science in 1997.

In January 2001, when the Japanese government carried out the reorganization of government ministries and agencies, the Ministry of Health and Welfare and the Ministry of Labour were merged to form the Ministry of Health, Labour and Welfare (MHLW). In the new system under the MHLW, a wide range of affairs from clinical trials and approval review to measures for safety in the postmarketing stage is now handled to secure and improve the health and livelihood of people in Japan.

MINISTRY OF HEALTH, LABOUR AND WELFARE

Function

In order to guarantee and improve the health and livelihood of the people, the MHLW is responsible for improving and promoting the social welfare, social security, public health, and the conditions and environment of labor. Furthermore, the MHLW is responsible for extending aid to the war repatriated, wounded, and sick soldiers

and the families of the war dead and nonrepatriated soldiers, and for settling pending affairs in the former army and navy.^[1]

Organization

As shown in [Fig. 1](#), the MHLW consists of the Minister's Secretariat, the Director General for Policy Planning and Evaluation, and 11 bureaus (the Health Policy Bureau; the Health Service Bureau; the Pharmaceutical and Medical Safety Bureau; the Labour Standards Bureau; the Employment Security Bureau; the Human Resources Development Bureau; the Equal Employment, Children and Families Bureau; the Social Welfare and War Victims' Relief Bureau; the Health and Welfare Bureau for the Elderly; the Health Insurance Bureau; and the Pension Bureau). Among the bureaus in the MHLW, the Pharmaceutical and Medical Safety Bureau (PMSB) and the Health Policy Bureau (HPB) are principally responsible for pharmaceutical administration.

A large number of establishments such as institutes, national hospitals, quarantine stations, and social welfare facilities are associated with the MHLW. The institutions and organizations of importance in pharmaceutical administration and regulation include the Pharmaceutical Affairs and Food Sanitation Council (PAFSC), the National Institute of Health Science (NIHS), and the National Institute of Infectious Diseases (NIID). In addition, the Organization for Pharmaceutical Safety and Research (OPSR), which is not an organization of the MHLW but a semigovernmental body, has a very important role in pharmaceutical administration and regulation.

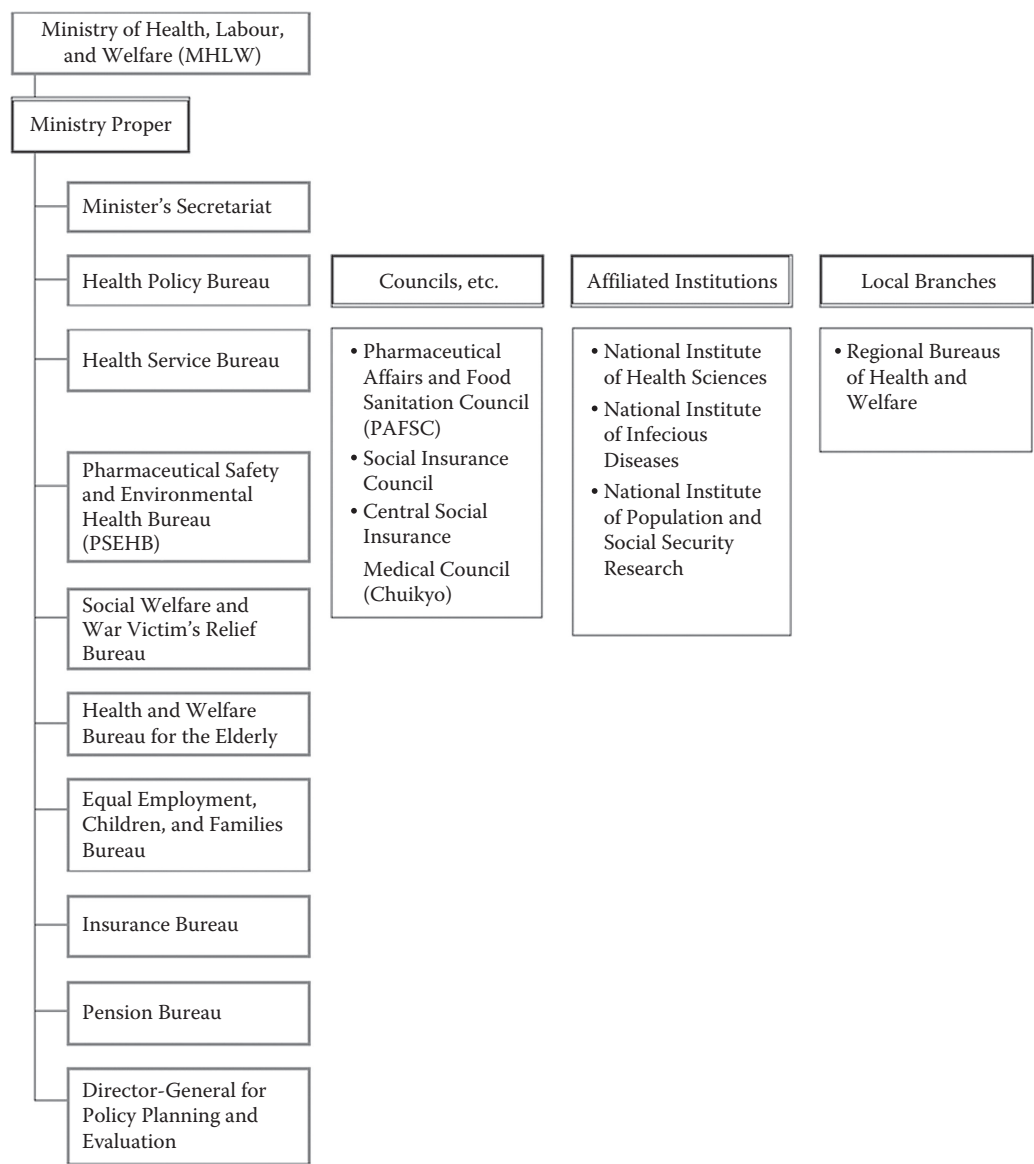


Fig. 1 Organization of Ministry of Health, Labour and Welfare (Health-related organizations only)

ADMINISTRATIVE ORGANS RESPONSIBLE FOR PHARMACEUTICAL AFFAIRS

Pharmaceutical and Medical Safety Bureau

The PMSB was established in July 1997 to provide a stricter approval review system for drugs and to improve measures for safety of medical and pharmaceutical supplies and facilities. This bureau is composed of five divisions (the General Affairs Division, the Evaluation and Licensing Division, the Safety Division, the Compliance and Narcotic Division, and the Blood and Blood Products Division) and the Department of Food Sanitation. The PMSB is in charge of affairs regarding the quality and safety of drugs, quasi-drugs, cosmetics, medical devices,

and other health-related products; the safety management of medical facilities such as hospitals and clinics; and the control of narcotics, psychotropics, cannabis, opium, stimulants, and a variety of poisonous and deleterious substances.

Health Policy Bureau

The HPB is in charge of planning policies for the realization of a high-quality, effective system for offering medical services and is composed of seven divisions. Among the divisions of the HPB, the Economic Affairs Division and the Research and Development Division are involved in pharmaceutical affairs. These divisions are in charge of the management of the pharmacomedical supply system

and the planning of measures for research and development of drugs.

Pharmaceutical Affairs and Food Sanitation Council

The PAFSC is an advisory body to the Minister for Health and Welfare, consisting of specialists in fields such as medical, pharmaceutical and biological sciences, statistics, and social sciences. It is divided into two councils (the Pharmaceutical Affairs Council and the Food Sanitation Council), both of which are subdivided into several committees and their subcommittees. The PAFSC conducts the investigation into and consideration with regard to pharmaceutical affairs and food sanitation, at the request of the Minister for Health, Labour and Welfare, including revision of the Japanese Pharmacopoeia, establishment of standards for pharmaceuticals, and the approval of new drug and medical devices.

National Institute of Health Sciences

The NIHS, formerly called the National Institute of Hygienic Sciences, is responsible for conducting basic research to ensure the quality, efficacy, and safety of a wide range of products that directly and indirectly affect the human health. To ensure the policies of the MHLW that reflect the results of such research, NIHS fulfills its mission of improving the quality of life of Japan's citizens.

Pharmaceutical and Medical Evaluation Centre

Under the NIHS, the PMDEC was established in 1997. PMDEC has a very important role in the Japanese pharmaceutical administration system, especially for new drug-approval procedures. The PMDEC is in charge of reviewing approval applicants such as drugs, quasi-drugs, cosmetics, and medical devices and conducting the reexamination and reevaluation of drugs and medical devices. In the review procedure in the PMDEC, review teams are organized from experts in such fields as medicine, pharmacology, veterinary science, and statistics and perform an approval review of applicants.

National Institute of Infectious Diseases

The NIID was formerly the National Institute of Health. This institute mainly conducts researches on the pathogenicity, etiology, prevention, and treatment of infectious diseases. In addition, the NIID is involved in researches and examinations and establishes the standard of biological products such as vaccines and blood products.

Organization for Pharmaceutical Safety and Research

The OPSR^a was established in 1979 as a semigovernmental body authorized by the Minister of Health and Welfare in accordance with the Law of the Organization for Pharmaceutical Safety and Research. Among the six departments of the OPSR, the Product Reviews Department, Clinical Trials Guidance Department, and Compliance Review Department are in charge of the consultation and reviews of clinical trials and investigations into the suitability of data with the reliability standards, GLP, GCP, and GPMSP, which are submitted with pharmaceutical affairs applications. These businesses are very closely connected with the process of drug approval under the commission of the MHLW and in accordance with the Pharmaceutical Affairs Law.

In addition to these businesses, the OPSR is responsible for reviewing equivalency between medicines applied for approval and medicines that have already been approved (equivalency review), providing prompt relief to individuals with health conditions such as disease and disablement, and death caused by adverse reactions when making proper use of a licensed drug, and for supporting the promotion of researches in pharmaceutical technology and orphan drug development.

PHARMACEUTICAL REGULATION IN JAPAN

Laws Related with Pharmaceutical Affairs

The administration of pharmaceutical affairs are regulated by several laws as follows: the Pharmaceutical Laws, the Pharmacists Laws, the Law of the Organization for Pharmaceutical Safety and Research, the Blood Collection and Blood Donation Services Control Law, the Poisonous and Deleterious Substances Control Law, the Narcotics and Psychotropics Law, the Cannabis Control Law, the Opium Law, and the Stimulants Control Law. Among these laws, the Pharmaceutical Affairs Law provides fundamental regulations and schemes for pharmaceutical activities to ensure the quality, efficacy, and safety of drug and medical devices and also to promote R&D activities, especially for orphan drugs.

DRUG DEVELOPMENT

Development of a new drug is regulated and managed under several organizations including the PMSB in the MHLW, the

^aThe PMDEC and OPSR will be merged and a new organization will be established by April 2004 as part of a reorganization of the Japanese pharmaceutical administration system.

PMDEC under NIHS, the OPSR, and the PAFSC. It can be divided into three stages: the clinical trial stage, the approval stage and the postapproval stage. A summary of the measures to ensure the safety and quality of the drugs in each stage of the drug development process is shown in Fig. 2.

Clinical and Nonclinical Studies of Drugs

A developed drug undergoes several nonclinical studies such as physiological studies, toxicity studies in accordance with the Standards for Conduct of Nonclinical Studies on the Safety of Drugs [also called good laboratory practice (GLP)], and pharmacological and pharmacokinetic studies until notification of clinical trials is submitted to the MHLW. The clinical trials in the premarketing stage, divided into Phase I, Phase II and Phase III studies, must be in accordance with the Standards for Conduct of Clinical Trials [also called good clinical practice (GCP)].

Consultation of Clinical Trials

A company that tries to get an approval in Japan can ask for advice on clinical trials from OPSR, which is supported by PMDEC and PAFSC. The OPSR provides four kinds of advice, such as pre-first clinical trial notification advice, end of Phase 2 advice, preapproval application advice, and individual advice.

A company requests advice on the clinical trial and provides related data. The OPSR gives an advice to the company, and delivers the minutes of the meeting that shows the discussion between the OPSR and the company. The

company can use the minutes as a part of the documents for drug application. A flowchart of consultation of clinical trials is shown in Fig. 3.

Review of Clinical Trial Design

When a product under investigation is intended to be used in humans for the first time in Japan, the sponsor has to submit a clinical-trial notification to the PMDEC 30 days before the trial starts. The PMDEC sends the notification to the OPSR. The OPSR discusses with the sponsor and reviews the notification from the point of view of safety and the protection and well-being of human subjects involved in the trials.

Drug Approval

For manufacturing and importing drugs in Japan, an approval is necessary from the Minister for Health, Labour and Welfare. The general process of drug approval is shown in Fig. 4.

According to the Pharmaceutical Affairs Law, data concerning a drug intended to be manufactured/imported is required to be submitted to the PMDEC. The data must comply with the standards specified in the ordinances of the MHLW such as the GLP, the GCP, and the Standards for the Reliability of Application Data. Drugs are classified into “prescription drugs” and “nonprescription drugs,” both of which are divided into those with new ingredients and those with already approved ingredients. From the data submitted by an applicant, each of the drugs

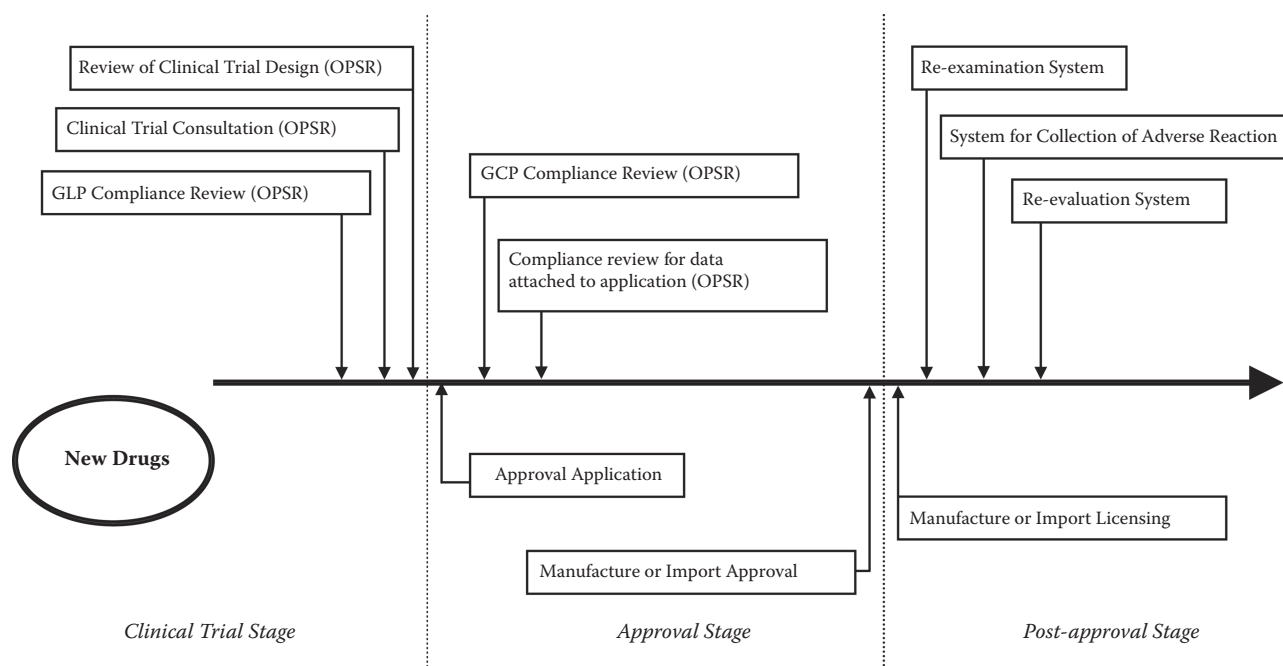


Fig. 2 Development of drugs. Drug development is divided into three stages (clinical trial stage, approval stage, and postapproval stage), throughout which the safety and quality of a new drug are regulated

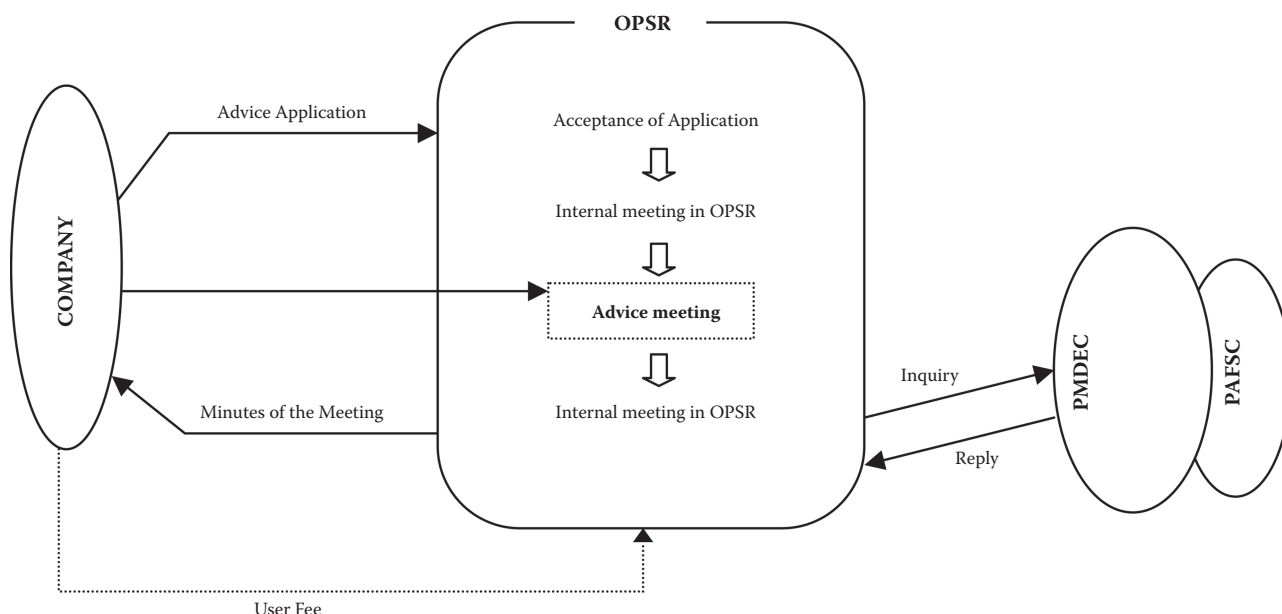


Fig. 3 Consultation of clinical trials

applied for approval is examined in detail with regard to its quality, efficacy, and safety by a team of specialists in the PMDEC, and after it is reviewed whether or not it has complied with the reliability standards, GLP and GCP, by the OPSR. After completion of the team review, the application is investigated and discussed by the committee and the pharmaceutical affairs council of the PAFSC. Finally, decisions for approval of drugs is made by the MHLW. The applicant receives an approval after these procedures are completed.

In addition to the approval of drug, the license to manufacture/import the drug is necessary for the company that tries to manufacture/import the drug. The license is examined when the application comes through whether or not the applicant is qualified for the manufacture/importation of the drug regarding their facilities and personnel, in compliance with the Regulations for Manufacturing Control and Quality Control of drugs [also called good manufacturing practice (GMP)] by the prefectural government or the Minister for Health and Welfare.

In general, the periods required for an approval of prescription drugs, nonprescription drugs, and drugs used for in vitro diagnostic tests, shortened from April 2000, are 12, 10, and 6 months, respectively.

POSTMARKETING SURVEILLANCE OF DRUGS

To assure the quality, efficacy, and safety of drugs, all approved drugs are kept under surveillance after coming into the market. The Good Post-Marketing Surveillance Practice (GPMSP) is applied as a standard for the proper implementation and reliability of the postmarketing

surveillance data. The GPMSP embodies the provisions to which pharmaceutical companies such as manufacturers and importers are subject in the organization and management of postmarketing surveillance. The measures in the postapproval stage are classified into the reexamination system, the reevaluation system, and the adverse drug reaction (ADR) reporting system (Fig. 2).

Reexamination System

Because of limitations on the data by premarketing clinical trials (referred to as Phase I/II/III studies) submitted for review at the time of approval of a new drug, the new drugs on the market are further examined for reconfirmation of the safety of drugs in practical use during the period of 4–10 years after their approval. The postmarketing reexaminations (referred to as Phase IV study) are obligatory for pharmaceutical companies that obtained approval of a new drug. The data of reexaminations submitted by pharmaceutical companies are reviewed by the PMDEC, and a final decision regarding the drug in question is made by the MHLW.

Reevaluation System

The reevaluation of drugs is performed systematically and periodically as already approved drugs in need of reconsideration according to the current status of the medical and pharmaceutical sciences. In addition to the periodic reevaluation system, an ad hoc reevaluation can be designed, for example, when the safety of drugs showing a certain efficacy becomes a subject of discussion.

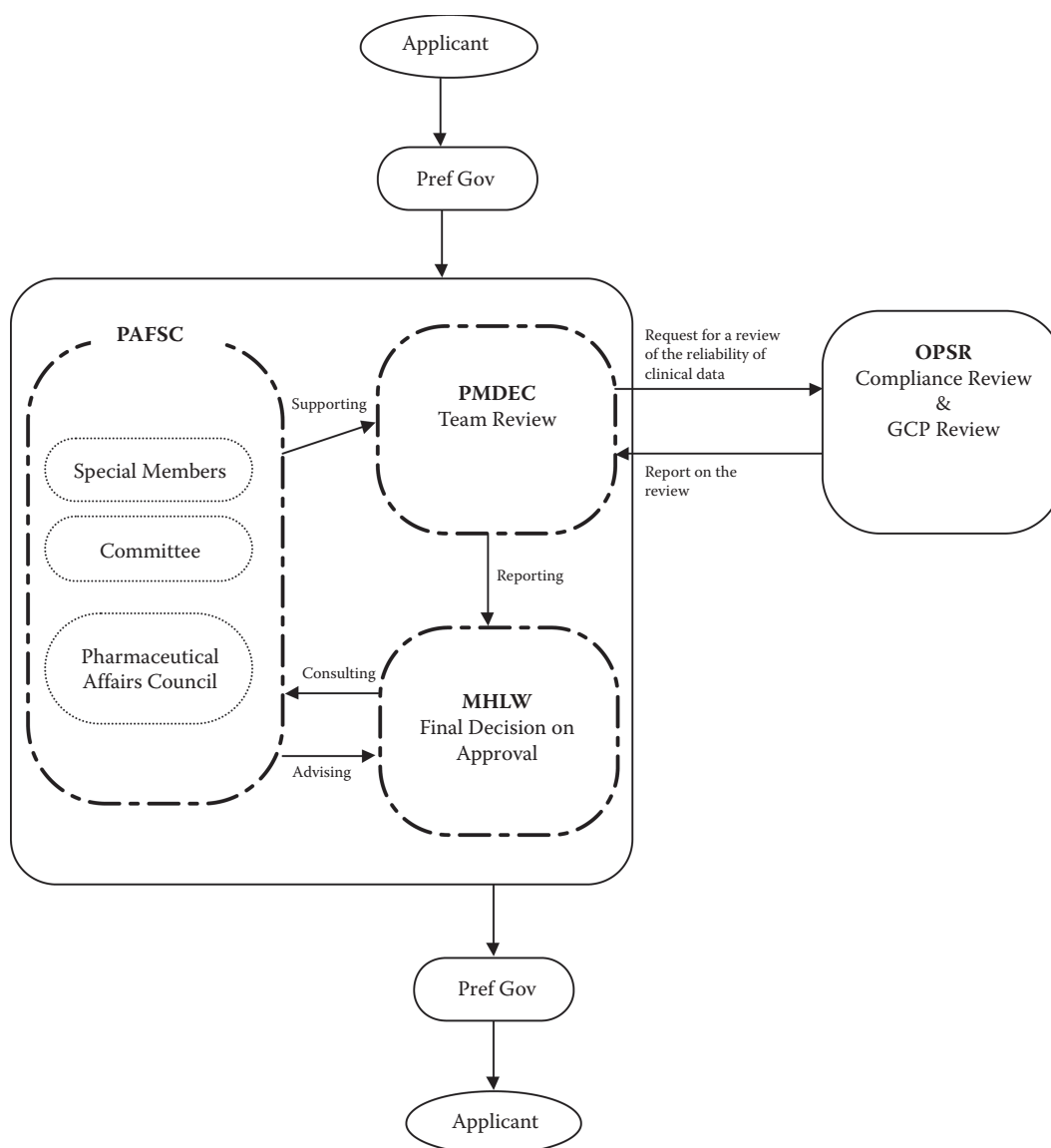


Fig. 4 Process of drug approval. Pref Gov Prefectural Government; PAFSC: The Pharmaceutical Affairs and Food Sanitation Council; PMDEC: The Pharmaceuticals and Medical Devices Evaluation Centre; MHLW: The Ministry of Health, Labour and Welfare; OPSR: The Organization for Pharmaceutical Safety and Research; GCP: Good Clinical Practice

Adverse Drug Reaction Reporting System

The adverse drug reaction reporting system is the program for collecting and reporting information on the cases of any suspected adverse drug reactions and infections as a result of the use of drugs, including death, disability, disease, and congenital abnormalities. This reporting system is composed from pharmaceutical companies' reports under the Pharmaceutical Affairs Laws and with all medical facilities and pharmacies as monitoring facilities. The information on drug safety is also exchanged with various countries by the World Health Organization (WHO) International Drug Monitoring Programme.

SUPPLY OF DRUG INFORMATION

The information on the quality, efficacy, and safety of drugs provided by the postmarketing surveillance is reflected in some administrative measures taken by the MHLW, including suspension of manufacture and/or marketing of the drug in question, revocation of its approval, partial changes in the medication, and revision of its precautions. In addition, the information on drug safety is fed back from the MHLW to the people involved in medical and pharmaceutical services such as doctors, dentists, and pharmacists. It is also supplied in the form of a package insert leaflet or a drug safety report for emergency, termed doctor letter by pharmaceutical companies.

INTERNATIONAL HARMONIZATION IN DRUG DEVELOPMENT

To speed up the process of supplying excellent drugs developed in the world to the patients, it is essential to avoid unnecessary duplicate studies of drugs for their quality, efficacy, and safety. Achieving this goal will not only make approval review faster and efficient but also promote research and development, and urgent measures are required to realize this objective.

Under such circumstances, Japan, the United States, and the European Union (EU) have organized the International Conference of Harmonization of Technical Requirements for Registration of Pharmaceutical for Human Use (ICH) with a view to the harmonization of quality, efficacy, and safety requirements as the main topics, based on cooperation with the respective regulatory agencies and pharmaceutical industry associations. The MHLW is one of the members of the ICH Steering Committee that designs the topics of the ICH.^[2,3]

As results of harmonization of the requirements for registration progress in the ICH, the mutual acceptance of the data is increased. Actually, the clinical and nonclinical data from the United States or EU have been used as data for new drug application in Japan. Therefore, ICH guidelines have contributed to speeding up the drug development process in Japan. Furthermore, it is expected that clinical and nonclinical data that are collected from research and development processes in Japan would be more widely used in the United States or EU.

An agreement was reached on guidelines for the acceptance of data from foreign clinical studies as a result of ICH. In Japan, several guidelines related to this topic were issued. According to the guideline, when data from clinical trials performed in a foreign country are used, first the data are checked to assure that they comply with Japanese regular requirements, and then reviewed regarding whether the drug is affected by ethnic factors or not. When necessary, a bridging study is performed, and when it is concluded that the outcome of clinical study in the foreign country can be extrapolated to the Japanese population, the foreign data can be accepted. In these procedures for acceptance of data from foreign countries, the importance of consultation for clinical trials, which is conducted by the OPSR, is increased.

STATISTICAL PROGRAMS IN DRUG DEVELOPMENT

In March 1992, the MHLW published the guideline for statistical analysis of clinical studies, which shows examples of statistical misusing methods and indicate the most appropriate methods at that time.^[4,5]

The ICH guideline "Statistical consideration in the design of clinical trials" was published in November 1998

to replace the former guideline mentioned above.^[6] This guideline is intended to minimize the bias from the results of clinical trials and to maximize the precision of clinical trial data by proposing approaches. The guideline states that the actual responsibility for all statistical work related to a clinical trial should be done by statisticians who have appropriate qualifications and experiences. The participation of statisticians is intended to verify if the statistical principles have been appropriately applied in the trial to support drug development with other clinical trial experts. It also noted the importance of the statisticians' participation in the clinical trial from the stage of planning the trial and emphasized that the prespecification of protocol such as a documentation of the method of analysis is needed.

In addition to the ICH guideline, "Question & answer related to the guideline" is also issued as an attachment to the ICH guideline, with the objective of supporting the understanding of the guideline. It contains details that are not mentioned in the guideline. For example, it is clearly mentioned that 2.5% significance is used in the one-sided test and 5% significance is used in the two-sided test. It also mentions that several centers with a large number of subjects per center are preferable to a large number of centers with few subjects per center. It is difficult to say how many subjects are needed per center because the number of subjects depends on the target diseases and the treatment; however, 10 subjects per center are described in it as one example.

ACKNOWLEDGEMENTS

Minor updates have been made by the Editor-in-Chief in order to reflect developments since publication of the *Encyclopedia of Biopharmaceutical Statistics, Third Edition*.

REFERENCES

1. Ministry of Health, Labour and Welfare; <http://www.mhlw.go.jp>.
2. English Regulatory-Information Working Group. *Pharmaceutical Administration and Regulations in Japan 1999–12*; Japanese Pharmaceutical Manufacturers Association, 1999.
3. Pharmaceutical Review System Study Group. *Pharmaceutical Administration in Japan*, 10th Ed.; Yakuji Nippo Ltd, 2001.
4. *Pharmaceutical Administration and Regulations in Japan*. Japan Pharmaceutical Manufacturers Association. <http://www.jpma.or.jp/english/parj/whole.html>, 2017.
5. Ministry of Health and Welfare. *Guideline for Statistical Analysis of Clinical Study Result. Notification No. 20*, The New Drug Division; 1992.
6. Ministry of Health and Welfare. *Statistical Principles for Clinical Trials. Notification No. 1047*, Evaluation and Licensing Division; 1998.

Kaplan–Meier Estimator

Hongyu Jiang

Harvard University, Boston, Massachusetts, U.S.A.

Jason Fine

University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.

Abstract

The Kaplan–Meier is a nonparametric estimator of distribution function of a time to event variable in the presence of independent censoring. This entry provides the method of formulation for the estimator and addresses its relation to other estimators, and gives an example to illustrate the methodology.

INTRODUCTION

The Kaplan–Meier^[1] is a nonparametric estimator of distribution function of a time to event variable in the presence of independent censoring. The estimator is also referred to as the Product-Limit estimator because it was first derived as the limiting case of the classical life table estimator of Böhmer^[2] in the actuarial science literature for grouped data. However, the Product-Limit principle was not well known until Kaplan and Meier derived it using the maximum likelihood arguments in 1958.^[1]

In medical studies, time to a particular event is often used as a marker of the intervention effect. For example, in lung cancer clinical trial, time from start of treatment to death is often used to evaluate whether new therapy is superior to the standard treatment in prolonging survival. In HIV clinical trials, a frequently used endpoint in comparing antiretroviral treatments is time from start of treatment to viral rebound.^[3] The event of interest may be a marker of disease progression and is sometimes referred to as a surrogate endpoint for the failure time.^[4] In a randomized trial, event times are assumed to be independently distributed, where event times under different treatments may follow different distributions. By estimating the survival distributions in the treatment arms, one may gain insight into the treatment effect.

FORMULATION OF THE ESTIMATOR

To fix the notation, let T denote the event time variable. Let $\{T_i, i = 1, \dots, n\}$ denote the independent event times from n subjects whose common survival function $S(t) = \Pr(T > t)$ is of interest. If all event times are observed, the nonparametric estimate of the survival function is simply the empirical estimator, $\hat{S}(t) = \sum_{i=1}^n I\{T_i > t\}/n$. An advantage of the nonparametric estimator is that no assumption is

needed with respect to the underlying distribution. Parametric alternatives based on the Weibull distributions may lead to biased estimation of $S(t)$.

Fortunately, observing event times on all subjects is not realistic in most medical studies because of limited follow-up time. This results in right censoring, where the event time is only observed if it occurs before the end of follow-up, otherwise, it is censored and known to be larger than the last follow-up time, the so-called censoring time. Denote the censoring variable by C , assumed to be independent of T . Let the minimum of T and C be X . The observed data for the i th subject is X_i and $D_i = I\{T_i \leq C_i\}$, where $I\{\cdot\}$ is the indicator function. By convention, if T_i and C_i have the same value, then T_i is observed.

Notice that the times $\{X_i, i = 1, \dots, n\}$ may have ties. Let $t_1 < t_2 < \dots < t_m$ be the m ordered, unique event times, d_j be the number of subjects who have events at time t_j , and r_j be the number of subjects who have not experienced events before time t_j , i.e., at risk at t_j , where $j = 1, \dots, m$. One can easily verify that

$$d_j = \sum_{i=1}^n I\{X_i = t_j \text{ and } D_i = 1\} \text{ and} \\ r_j = \sum_{i=1}^n I\{X_i \geq t_j\}$$

The Kaplan–Meier estimator of $S(t)$ is defined on the range of the observed data, $[0, \tau_n]$, where $\tau_n = \max(X_i, i = 1, \dots, n)$:

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{j: t_j \leq t} \left(1 - \frac{d_j}{r_j}\right) & \text{if } t_1 \leq t \leq \tau_n \end{cases} \quad (1)$$

For $t > \tau_n$, if there is no censoring at the largest observed time τ_n , then $\hat{S}(t) = 0$, otherwise, $\hat{S}(t)$ is undefined. Efron^[5]

proposed estimating $S(t)$ beyond τ_n by 0, while Gill^[6] suggested $\hat{S}(\tau_n)$. The versions are equivalent in large samples. Klein^[7] showed that in small samples, Gill's estimator has smaller bias in the tail of the distribution.

As with the empirical distribution function, the Kaplan–Meier estimator is a step function. However, it only jumps at uncensored X_i . If there is no censoring ($\sum_{i=1}^n D_i = n$), then the estimator reduces to exactly the empirical estimator.

RELATION TO OTHER ESTIMATORS

The standard approach to deriving the Kaplan–Meier estimator is by maximizing the generalized (or nonparametric) likelihood.^[1,8–10] It is the limit of the life table estimator and has close connections with some other well-known nonparametric estimators for the distribution function with censored data.

Nelson–Aalen Estimator

Instead of estimating the survival function, the Nelson–Aalen estimator^[11,12] estimates the cumulative hazard function $H(t) = -\ln[S(t)]$ by

$$\hat{H}(t) = \sum_{j: t_j \leq t} \frac{d_j}{r_j}$$

where d_j/r_j may be viewed as an estimate of the instantaneous hazard function, $dH(t)/dt$, at time t_j . Although it is not a consistent estimate of $dH(t_j)/dt$, the sum of all such estimates for $t_j \leq t$ is consistent for $H(t)$. An alternative estimator of $S(t)$ can be obtained as $\tilde{S}(t) = \exp[-\hat{H}(t)]$. The equivalence between $\tilde{S}(t)$ and $\hat{S}(t)$ in large samples can be roughly illustrated as follows. From Eq. 1, $\hat{S}(t)$ may be written as

$$\hat{S}(t) = \exp \left[\sum_{j: t_j \leq t} \ln \left(1 - \frac{d_j}{r_j} \right) \right]$$

For large n , d_j/r_j is small for all t_j . Hence $\ln(1 - d_j/r_j) \approx -d_j/r_j$. It follows that $\hat{S}(t) \approx \tilde{S}(t)$. A more rigorous derivation shows that the Kaplan–Meier estimator and the Nelson–Aalen estimator are related through the product integral expression:

$$\hat{S}(t) = \prod_{u \leq t} [1 - d\hat{H}(u)]$$

Self-Consistent Estimator

The self-consistent estimator was first proposed by Efron^[5] for right-censored data. To estimate the survival function, a self-consistency equation is constructed analogously to that for the empirical estimator without censoring. In a

sample with size n , we estimate the survival probability at t by counting the number of subjects not experiencing the event by t and divide this number by n . Each subject whose event time is beyond t is counted as 1. A subject censored at a time x before t has $S(t)/S(x)$ chance of not failing by t and is counted as that fraction. The resulting self-consistent estimating equation is

$$S(t) = \frac{1}{n} \left(\sum_{i=1}^n I\{X_i > t\} + (1 - D_i) \sum_{i=1}^n \frac{S(t)}{S(X_i)} \right) \quad (2)$$

It has been shown that Efron's^[5] version of the Kaplan–Meier estimator is the unique solution to the above equation. The self-consistent estimating Eq. 2 can be used to derive an iterative procedure which is a special case of the EM algorithm for estimating the survival function. This idea has been generalized to construct nonparametric estimators from complicated censored, truncated, and grouped data^[13] which reduce to the Kaplan–Meier estimator in the special case of right censoring.

The self-consistent estimator may be interpreted via a redistribute-to-the-right algorithm, also proposed by Efron.^[5] The algorithm starts with an empirical estimator that puts probability mass $1/n$ at each observed time. Iterating from the smallest censored X_i to the largest censored X_i , the mass is removed and redistributed in equal portions to all points to the right of the censored time. With a finite sample, this algorithm converges to Efron's version of the Kaplan–Meier estimator.

INFERENCE ON SURVIVAL FUNCTION

Variance Estimation

The variance of the Kaplan–Meier estimator at time t , $\sigma^2(t)$, can be estimated by Greenwood's formula:^[14]

$$\hat{\sigma}^2(t) = \hat{S}(t)^2 \sum_{j: t_j \leq t} \frac{d_j}{r_j(r_j - d_j)} \quad (3)$$

When there is no censoring in the data, Eq. 3 reduces to the standard binomial variance estimator, $n^{-1}\hat{S}(t)[1 - \hat{S}(t)]$, for the empirical distribution. An alternative way of estimating the variance of $\hat{S}(t)$ is proposed by Aalen and Johansen:^[15]

$$\tilde{\sigma}^2(t) = \hat{S}(t)^2 \sum_{j: t_j \leq t} \frac{d_j}{r_j^2}$$

Klein showed that both variance estimators tend to underestimate the true variance of the Kaplan–Meier estimator for small to moderate samples. On average, Greenwood's formula tends to have smaller bias. A more conservative variance estimator is proposed by Peto et al.^[16]

Confidence Limits and Bands

In large samples, $\hat{S}(t)$ is approximately normally distributed with variance $\sigma^2(t)$. Hence a $100(1 - \alpha)\%$ pointwise confidence interval of $S(t)$ is

$$\hat{S}(t) \pm Z_{\alpha/2} \hat{\sigma}(t) \quad (4)$$

where $Z_{\alpha/2}$ is the $(1 - \alpha/2)$ th quantile of the standard normal distribution. However, the normal approximation to the distribution of $\hat{S}(t)$ may not work well in small samples, resulting in coverage probability much smaller than $(1 - \alpha)$. To achieve coverage closer to the nominal probability, we may construct confidence intervals for transformed $S(t)$ and then transform back. Two choices of transformation were suggested by Borgan and Liestøl.^[17] One is the log-minus-log transformation, $\ln\{-\ln[S(t)]\}$, and the resulting confidence interval for $S(t)$ is

$$\left[\hat{S}(t)^{\exp\left\{\frac{Z_{\alpha/2}\hat{\sigma}(t)}{\hat{S}(t)\ln\hat{S}(t)}\right\}}, \hat{S}(t)^{\exp\left\{-\frac{Z_{\alpha/2}\hat{\sigma}(t)}{\hat{S}(t)\ln\hat{S}(t)}\right\}} \right]$$

The other is the arcsine-square-root transformation, $\arcsin(S(t)^{1/2})$, and the corresponding confidence interval for $S(t)$ is

$$\left[\sin^2 \left\{ \max(0, \arcsin(\hat{S}(t)^{1/2}) - 0.5Z_{\alpha/2} \frac{\hat{\sigma}(t)}{(\hat{S}(t)(1 - \hat{S}(t))^{1/2})} \right\}, \sin^2 \left\{ \min\left(\frac{\pi}{2}, \arcsin(\hat{S}(t)^{1/2})\right) + 0.5Z_{\alpha/2} \frac{\hat{\sigma}(t)}{(\hat{S}(t)(1 - \hat{S}(t))^{1/2})} \right\} \right]$$

Besides normal approximation, confidence limits may also be obtained by inverting a likelihood ratio statistic,^[18] which also achieves good coverage probability in small samples, but may be computationally demanding.

With pointwise confidence intervals, we are able to make inference about survival probability at a single time point. Sometimes, however, it may be of interest to study the simultaneous behavior of the survival curve over an interval, say $[t_L, t_U]$. That is, we may want to find upper and lower confidence bands such that $\{S(t), t \in [t_L, t_U]\}$ is guaranteed to lay entirely within the bands with $1 - \alpha$ level of confidence. Statistically, we want to specify two random functions $L(t)$ and $U(t)$ satisfying

$$\Pr(L(t) \leq S(t) \leq U(t), \text{ for } t \in [t_L, t_U]) = 1 - \alpha \quad (5)$$

Because it can be shown that $\sqrt{n}\{\hat{S}(t) - S(t)\}$ converges weakly to a mean-zero Gaussian process with a simple Brownian motion structure,^[6,19] simultaneous inference on $S(t)$ over an interval is possible analytically. Two types of confidence bands have been proposed.

The first is equal precision (EP) bands which are proportional to the pointwise confidence limits at all t .^[20]

The width of each pointwise confidence interval for $S(t)$ (or transformed $S(t)$) on the interval $[t_L, t_U]$ is inflated by a constant κ such that the new limits satisfy Eq. 5. The value κ is determined by the sample size, the estimated survival probabilities and the estimated variances at t_L and t_U . A minor restriction is that the EP bands can only be obtained on a subinterval of $[t_L, t_U]$, where t_l and t_m are as defined previously. With moderate sample sizes, the non-transformed EP bands achieve coverage probability that is much lower than the nominal level, but the problem can be solved by constructing transformed EP bands.

Another method was developed by Hall and Wellner (HW).^[21] The HW bands inflate the pointwise confidence limits by factors that depend not only on the sample size, and the point and variance estimates of $S(t_L)$ and $S(t_U)$, but also on the point and variance estimates of $S(t)$. Hence the bands are not proportional to the pointwise confidence limits. Typically, HW bands are tighter than EP bands. Both nontransformed and transformed HW bands yield close to the nominal coverage probability.

The inflation coefficients for both EP and HW confidence bands are tabulated in Ref. [22], pp. 439–456.

AN EXAMPLE

A data set from an acute leukemia clinical trial^[23] will be used to illustrate the methodology. In the study, 21 pairs of children with acute leukemia matched by remission status were accrued. In each pair, one child was randomly assigned to the drug 6-mercaptopurine (6-MP) and the other to the control treatment. Patients were followed from randomization until their leukemia returned (relapse) or until the end of the study. The following are the observed ordered survival times (in weeks) where a superscript + means the observed time is a censored time.

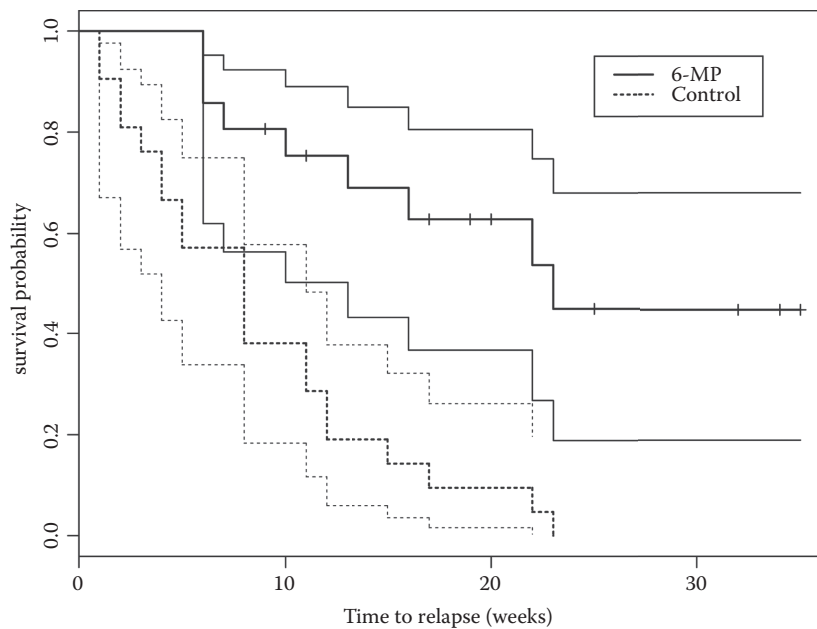
6-MP arm:	6, 6, 6 ⁺ , 6 ⁺ , 7, 9 ⁺ , 10, 10 ⁺ , 11 ⁺ , 13, 16, 17 ⁺ , 19 ⁺ , 20 ⁺ , 22, 23, 25 ⁺ , 32 ⁺ , 32 ⁺ , 34 ⁺ , 35 ⁺ ;
Control arm:	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23.

Let S_1 and S_2 denote the survival functions of subjects in the 6-MP arm and in the control arm, respectively. Because there is no censoring in the control arm, the Kaplan–Meier estimator of S_2 is just the empirical estimator. The more complicated computation of the Kaplan–Meier estimator of S_1 is illustrated in Table 1.

The estimates of the survival curves of time to relapse for the two arms are plotted in Fig. 1. The thick solid line is the Kaplan–Meier estimator of the survival function for the 6-MP arm, and the thin solid lines are the pointwise upper and lower 95% confidence intervals based on the log-minus-log transformation. Similarly, the thick dotted

Table 1 Computation of the Kaplan–Meier Estimator of S1 (the Survival Function of Time to Relapse in the 6-MP Arm)

Distinct event times: t_j	Number of events at t_j : d_j	Number at risk at t_j : r_j	Kaplan–Meier estimator $\hat{S}_1$	Greenwood variance estimator $\hat{\sigma}^2(\hat{S}_1)$
6	3	21	$[1 - (3/21)] = 0.857$	$0.857^2[3/(21 \times 18)] = 0.0058$
7	1	17	$0.857[1 - (1/17)] = 0.807$	$0.807^2[0.0079 + (1/(17 \times 16))] = 0.0076$
10	1	15	$0.807[1 - (1/15)] = 0.753$	$0.753^2[0.0116 + (1/(15 \times 14))] = 0.0093$
13	1	12	$0.753[1 - (1/12)] = 0.690$	$0.690^2[0.0164 + (1/(12 \times 11))] = 0.0114$
16	1	11	$0.690[1 - (1/11)] = 0.628$	$0.628^2[0.0240 + (1/(11 \times 10))] = 0.0130$
22	1	7	$0.628[1 - (1/7)] = 0.538$	$0.538^2[0.0330 + (1/(7 \times 6))] = 0.0164$
23	1	6	$0.538[1 - (1/6)] = 0.448$	$0.448^2[0.0569 + (1/(6 \times 5))] = 0.0181$

**Fig. 1** Comparison of survival curves of time to relapse by treatment arms

line and the thin dotted lines are the Kaplan–Meier estimate and its pointwise confidence limits, respectively, for the control arm. It is easy to see that the patients receiving 6-MP treatment have much higher survival probabilities than the patients receiving the control treatment.

OTHER ISSUES

Descriptive Statistics

The most commonly looked at summary statistic with censored data is the median, or 0.5 quantile of the distribution. Unlike the mean, it is less sensitive to the right tail of the survival distribution, where estimation may be unstable. The median survival time η can be estimated by

$$\hat{\eta} = \inf \{t : \hat{S}(t) \leq 0.5\}$$

In large samples, $\hat{\eta}$ is approximately normally distributed and the analytical form of its variance can be evaluated explicitly. However, variance estimation involves the density function of T at the median, which is difficult with censored data. To solve this problem, Brookmeyer and Crowley^[24] constructed confidence intervals for η by inverting the pivotal $(\hat{S}(\eta) - 0.5)$ without estimating the variance of $\hat{\eta}$.

The p th quantile of the survival distribution, η_p , can be similarly estimated by

$$\hat{\eta}_p = \inf \{t : \hat{S}(t) \leq 1 - p\}$$

Occasionally, the mean survival time (life expectancy), $\mu = \int_0^\infty S(t)dt$, may be of interest. However, μ cannot be reliably estimated with censored data when the support of T is larger than that of C . It is only possible to estimate the mean survival time restricted to an interval $[0, u]$, where u

is no bigger than τ_n , the largest observed time. The point and variance estimates of the restricted mean μ_u are:

$$\hat{\mu}_u = \int_0^u \hat{S}(t) dt;$$

$$\hat{\text{var}}(\hat{\mu}_u) = \sum_{j=1}^m \left[\int_{t_j}^u \hat{S}(t) dt \right]^2 \frac{d_j}{r_j(r_j - d_j)}$$

Hypothesis Testing

In the one-sample problem, testing whether the survival probability at a fixed time t_0 equals a hypothesized value b at α significance level is straightforward, simply by checking if b lays inside the $(1 - \alpha)$ confidence interval of $S(t_0)$. Testing whether a survival curve has a prespecified shape can be performed by checking whether it lies within the confidence bands. However, this approach tends to be conservative and is not widely used. For testing the hypothesis $H_0: S(t) = S_0(t)$, where $S_0(t)$ is a known continuous function, the standard approach is the logrank test for the equivalent hypothesis $H_0: \lambda(t) = \lambda_0(t)$, where $\lambda(t)$ and $\lambda_0(t)$ are the corresponding hazard functions defined as $\lambda(t) = -dS(t)/S(t)$ and $\lambda_0(t) = -dS_0(t)/S_0(t)$, respectively.^[25] Alternatively, the hypothesis can be tested using Cramér–von Mises-type or Kolmogorov–Smirnov-type test statistics based on the Kaplan–Meier estimator.^[25–27]

In the two-sample problem, the most commonly used test for comparing two continuous distribution functions is again the logrank test, which is most powerful if the proportional hazard alternative is true. However, when the proportional hazard assumption is violated, as occurs with crossing hazards, the logrank test may no longer be optimal. As in the one-sample case, two-sample Cramér–von Mises or Kolmogorov-type test statistics, which are sensitive to those alternatives, may be constructed using the Kaplan–Meier estimators.^[26,27] A study of the tests' small sample performance can be found in Ref. [28].

All the above-mentioned two-sample tests are rank-based. Another alternative approach is the so-called weighted Kaplan–Meier statistics, which is not a rank-based test, but is nonparametric. It reduces to a two-sample t -test for the equality of the mean survival times with uncensored data.^[29,30]

Informative Censoring

Random censoring is a necessary condition for the validity of the Kaplan–Meier estimator. That is, T and C are independent. In clinical trials with staggered accrual, subjects who have not experienced the event of interest by study closure have their event times right-censored at their total duration on study. This censoring is often referred to as

administrative censoring and is unlikely to be correlated to the outcome of interest. Thus the censoring time can be assumed to be independent or uninformative of the event time of interest. In other cases, the assumption may not be satisfied. Some patients may discontinue participation because of the reasons related to the unobserved failure time of interest. For example, patients may drop out of a study because the treatment is not working, e.g., they may be too sick to make the clinical visits. This is often referred to as informative dropout.

Care should be exercised when interpreting the estimated survival probabilities in the presence of potentially informative censoring because the Kaplan–Meier estimator may no longer be consistent. One way of dealing with this complication is by modeling the correlation between the failure time variable and the censoring variable, and then adjust the marginal estimate by the estimated dependence structure.^[31] The resulting estimator may be compared with the naive Kaplan–Meier curve, which is valid if T and C are, in fact, independent.

CONCLUSION

The Kaplan–Meier estimator is a simple but very useful nonparametric estimator of survival function with right-censored time-to-event data when the censoring is noninformative. Its large sample property and small sample performance have been thoroughly studied in literature. The estimation and graphical display of this estimator and its confidence interval or bands have been routine in analyzing right-censored failure time data.

REFERENCES

1. Kaplan, E.L.; Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **1958**, *53*, 457–481.
2. Böhmer, P.E. Theorie der unabhängigen wahrscheinlichkeiten. *Rapports Mém. P.v. Septième Cong. Int. Actuaire* **1912**, *2*, 327–343. Amsterdam.
3. Gilbert, P.B.; DeGruttola, V.; Hammer, S.M. Virologic and regimen termination surrogate endpoints in AIDS clinical trials. *J. Am. Med. Assoc.* **2001**, *285*, 777–784.
4. Prentice, R.L. Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat. Med.* **1989**, *8*, 431–440.
5. Efron, B. The Two Sample Problem with Censored Data. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*; Prentice-Hall: New York, 1967, Vol. 4, 831–853.
6. Gill, R.D. Censoring and Stochastic Integrals. *Mathematical Center Tracts*; Mathematisch Centrum: Amsterdam, 1980, 124.
7. Klein, J.P. Small-sample moments of some estimators of the variance of the Kaplan–Meier and Nelson–Aalen estimators. *Scand. J. Stat.* **1991**, *18*, 333–340.

8. Kiefer, J.; Wolfowitz, J. On a theorem of Hoel and Levine on extrapolation. *Ann. Math. Stat.* **1965**, *27*, 887–906.
9. Johansen, S. The product limit estimator as maximum likelihood estimator. *Scand. J. Stat.* **1978**, *5*, 195–199.
10. Scholz, F.W. Towards a unified definition of maximum likelihood. *Can. J. Stat.* **1980**, *8*, 193–203.
11. Nelson, W. Theory and applications of hazard plotting for censored failure data. *Technometrics* **1972**, *14*, 945–965.
12. Aalen, O. Nonparametric inference for a family of counting processes. *Ann. Stat.* **1978**, *6*, 534–545.
13. Turnbull, B.W. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. R. Stat. Soc., Ser. B* **1976**, *38*, 290–295.
14. Greenwood, M. The Errors of Sampling of the Survivorship Tables. *Reports on Public Health and Statistical Subjects*; HMSO: London, 1926, Vol. 33. Appendix 1.
15. Aalen, O.O.; Johansen, S. An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scand. J. Stat.* **1978**, *5*, 141–150.
16. Peto, R.; Pike, M.C.; Armitage, P.; Breslow, N.E.; Cox, D.R.; Howard, S.; Mantel, N.; McPherson, K.; Peto, J.; Smith, P. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II: Analysis and examples. *Br. J. Cancer* **1977**, *35*, 1–39.
17. Borgan, Ø.; Liestøl, K. A Note on confidence intervals and bands for survival curve based on transformations. *Scand. J. Stat.* **1990**, *17*, 35–41.
18. Cox, D.R.; Oakes, D. *Analysis of Survival Data*; Chapman Hall: London, 1984, 51, 52.
19. Breslow, N.; Crowley, J. A large sample study of the life table and product limit estimates under random censorship. *Ann. Stat.* **1974**, *2*, 437–453.
20. Nair, V.N. Confidence bands for survival functions with censored data: A comparative study. *Technometrics* **1984**, *14*, 265–275.
21. Hall, W.J.; Wellner, J.A. Confidence bands for a survival curve from censored data. *Biometrika* **1980**, *67*, 133–143.
22. Klein, J.P.; Moeschberger, M.L. *Survival Analysis Techniques for Censored and Truncated Data*; Springer-Verlag: New York, 1997, 100–103.
23. Freireich, E.J.; Gehan, E.; Frei, E.; Schroeder, L.R.; Wolman, I.J.; Anbari, R.; Burgert, E.O.; Mills, S.D.; Pinkel, D.; Selawry, O.S.; Moon, J.H.; Gendel, B.R.; Spurr, C.L.; Storrs, R.; Haurani, F.; Hoogstraten, B.; Lee, S. The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy. *Blood* **1963**, *21*, 699–716.
24. Brookmeyer, R.; Crowley, J.J. A confidence interval for the median survival time. *Biometrics* **1982**, *38*, 29–41.
25. Andersen, P.K.; Borgan, Ø.; Gill, R.D.; Keiding, N. *Statistical Models Based on Counting Processes*; Springer-Verlag: New York, 1993, 391, 392, 399, 400.
26. Koziol, J.A. A two sample Cramér–von Mises test for randomly censored data. *Biom. J., J. Math. Methods Biosci.* **1978**, *20*, 603–608.
27. Koziol, J.A.; Yuh, Y.S. Omnibus two-sample test procedures with randomly censored data. *Biom. J., J. Math. Methods Biosci.* **1982**, *24*, 745–752.
28. Schumacher, M. Two-sample tests of Cramér–Von Mises and Kolmogorov–Smirnov type for randomly censored data. *Int. Stat. Rev.* **1984**, *52*, 263–281.
29. Pepe, M.S.; Fleming, T.R. Weighted Kaplan–Meier statistics: A class of distance tests for censored survival data. *Biometrics* **1989**, *45*, 497–507.
30. Pepe, M.S.; Fleming, T.R. Weighted Kaplan–Meier statistics: Large sample and optimality considerations. *J. R. Stat. Soc., Ser. B Methodol.* **1991**, *53*, 341–352.
31. Fine, J.P.; Jiang, H.; Chappell, R. On semi-competing risks data. *Biometrika* **2001**, *88*, 907–919.

Kappa Coefficients: Medical Research

Helena Chmura Kraemer

Stanford University School of Medicine, Stanford, Connecticut, U.S.A.

Abstract

A “kappa-type” statistic, generally, is one formed by linearly recalibrating some statistic to have the value zero when the data on which it is based are randomly generated. This entry discusses the mathematical basis for kappa coefficients in medical research and addresses some of the criticisms they face in terms of validity and reliability.

INTRODUCTION

A “kappa-type” statistic, generally, is one formed by linearly recalibrating some statistic to have the value zero when the data on which it is based are randomly generated, and one, when the generation is, in some sense, optimal. However, the kappa coefficients (population parameters) most commonly and usefully used in medical research are of only two types:

- Intraclass kappa, which is used to indicate the reproducibility, reliability, or precision of a binary clinical measure, such as presence/absence of a diagnosis, success/failure in response to treatment, positive/negative result of a medical test, or presence/absence of a risk factor.
- Weighted kappa, which is used to indicate the validity or accuracy of one such binary clinical measure assessed against a binary criterion or “gold standard.”

OVERVIEW

Kappa coefficients have had a checkered history. They have been “introduced” several times over, and the reasons for their continued use have little to do with the reasons for their introduction. Originally, they were introduced as percentage agreement corrected for chance, which only stimulated arguments about how “agreement” or “chance” should be defined, and whether such correction was at all necessary. Kappa coefficients have been applied not only to binary measures, but to more than two nonordered categories,^[1] to multiple diagnoses,^[2] as well as to ordered categories, with questionable success.

Theoretical papers have occasionally produced erroneous results, leading to Fleiss’s classic comment:^[3] “Many human endeavors have been cursed with repeated failures before final success is achieved. The scaling of Mount Everest is one example. The discovery of the Northwest

Passage is a second. The derivation of a correct standard error for kappa is a third.” (p. 974). Consequently, there are many papers in the research literature presenting questionable results in substantive as well as methodological areas, and predictably, papers questioning the use and value of kappa coefficients. Indeed, the introduction of kappa has been described as the major impediment to the development of high quality psychiatric diagnoses.^[4]

In the following discussion, the primary focus is on the types of kappa coefficients that have proved valuable in medical research, namely the intraclass and weighted kappas for binary measures. Then we will discuss how kappa fits within the context of other measures available for the same tasks. Finally, we will discuss some of the controversies surrounding kappa coefficients.

MATHEMATICAL BASIS

Let X_{sr} denote a binary random variable, for convenience one that takes on the values 1 and 0, that represents a binary classification of subject s ($s = 1, 2, \dots$) from a population of subjects, and r a rater or rating ($r = 1, 2, \dots$) from a population of possible or potential ratings. For subject s , the probability that a randomly selected rating will have $X_{sr} = 1$ is p_s , where p_s has some (usually unknown) distribution from 0 to 1 over the population of subjects. The population mean $E(p_s) = P$, the prevalence or incidence of whatever X indicates, and the population variance is σ_p^2 . The intraclass kappa is defined as:

$$\kappa_p = \sigma_p^2 / PP',$$

where $P' = 1 - P$.

For example, X_{sr} might be a diagnosis or a medical test result (positive = 1, negative = 0) for a disorder. Then, a very common simple mathematical model is one where it is assumed that the probability of having the disorder is π . Those who have the disorder have $p_s = \text{Se}$ (Se is a constant

“sensitivity”), and those without $p_s = 1 - \text{Sp}$ (Sp a constant “specificity”). This posits a two-point distribution for p_s that is particularly easy to work with. In this case:

$$\begin{aligned} P &= \pi \text{Se} + \pi'(1 - \text{Sp}), \quad \text{and} \\ \sigma_p^2 &= \pi\pi'(\text{Se} + \text{Sp} - 1)^2 \\ \kappa_p &= \pi\pi'(Se + Sp - 1)^2 / PP' \end{aligned}$$

Note that in this model, as π approaches either 0 or 1, κ_p approaches zero. This becomes important in some of the controversies surrounding kappa coefficients.

However, while mathematically convenient and useful for exploring insights about reliability or validity, such simple models seldom well represent real clinical applications. The distribution of p_s can be estimated by taking multiple independent measures for each subject using M randomly selected raters or ratings from the population of potential ratings for each subject. The proportion of positive ratings for subject s is an unbiased estimator of p_s . With a large enough sample of subjects, one can approximate the distribution of p_s from their sample estimates. When this has been done for either medical diagnoses or tests, it appears that p_s is more likely to have a continuous U-shaped distribution between 0 and 1, than, for example, the two-point distribution described above. That is, there are likely to be many subjects whose signs, symptoms, and response unambiguously indicate $X = 0$ ($p_s = 0$) or $X = 1$ ($p_s = 1$) to the rater, but p_s for the remaining subjects will range across the entire spectrum from 0 to 1. To make strong, but unwarranted, assumptions about the distribution of p_s in a model is problematic, and can seriously mislead. Kappa coefficients require no such strong assumptions.

THE INTRACLAS KAPPA: REPRODUCIBILITY OR RELIABILITY

Consider the problem of the “independent second opinion.” The probability that a randomly selected subject evaluated by two independent raters sampled from the population of raters will both be positive is $E(p_x^2) = P^2 + \sigma_p^2 = P^2 + \kappa_p PP'$. The probability that the two disagree is $2PP'(1 - \kappa_p)$, and the probability that both are negative is $P'^2 + \kappa_p PP'$. Then:

$$\begin{aligned} \kappa_p &= \text{Prob}(\text{second positive if first positive}) \\ &\quad - \text{Prob}(\text{second positive if first negative}) \end{aligned}$$

and

$$\begin{aligned} \kappa_p &= \text{Prob}(\text{second negative if first negative}) \\ &\quad - \text{Prob}(\text{second negative if first positive}) \end{aligned}$$

Thus when $\kappa_p = 0$, the first opinion has no predictive value for the second, whereas when $\kappa_p = 1$, the first always perfectly predicts the second. In general, κ_p quantitatively indicates how closely the first and second opinions will

agree, an interpretable measure of reproducibility of measurement.

Moreover, the “reliability” of a measure is classically defined as the percentage of the observed variance (here PP') representing true variance (here σ_p^2). Thus the definition of the intraclass kappa corresponds exactly to the classical definition of the reliability of a measure.

When $M = 2$, κ_p is particularly easy to estimate. If the proportion of the N subjects sampled on which both raters agree is A , and the proportion of all $2N$ ratings that are positive is $\hat{P}$ (an unbiased estimator of P), then the maximum likelihood estimator of κ_p is:

$$\begin{aligned} & (A - \hat{P}^2 - \hat{P}'^2) / (1 - \hat{P}^2 - \hat{P}'^2) \\ & \hat{P}' = 1 - \hat{P} \end{aligned}$$

Here the origins of kappa as the statistic percentage agreement (A) corrected for chance are clearly seen. While it is true that this was the original motivation for introduction of the sample kappa coefficient, the argument for the continued use of the intraclass kappa as a coefficient of reproducibility or reliability is not based on this computation, but (1) on its interpretability as a measure of the predictive value of a first measure for a second, and (2) that the intraclass kappa satisfies the classical definition of reliability of a measure. Finally, as we will see, intraclass kappa play an important role in influencing the precision of estimation of other parameters and in power of testing hypotheses.^[5]

Asymptotically, the variance of the intraclass kappa with two raters and N subjects is given by:^[6]

$$N \text{ variance} = (1 - \kappa_p) \left[(1 - \kappa_p)(1 - 2\kappa_p) + \frac{\kappa_p(2 - \kappa_p)}{2PP'} \right]$$

However, because the sample distribution of the estimator of intraclass kappa approaches a normal distribution very slowly (as is also true for the weighted kappa), it is recommended that either jackknife or bootstrap procedures be used to generate confidence interval or tests^[6] in preference to asymptotic theory.

When there are $M > 2$ raters/ratings, the same intraclass kappa can be estimated as follows: Let s be the number of positive ratings of the M given, and f_s the proportion of the N subjects with s positive ratings. Then:^[7]

$$\kappa_p = 1 - M \sum f_s(s/M)(1 - s/M)/(M - 1)\hat{P}\hat{P}'$$

where:

$$\hat{P} = \sum f_s s / M$$

While there are no fixed standards for what constitutes adequate reliability of measurement, useful “rules of thumb” have been suggested.^[8] A kappa greater than 0.8 is considered “almost perfect”; 0.61–0.8 as “substantial,”

0.41–0.6 as moderate; 0.2–0.41 as “fair,” and below 0.2 as “slight.” In fact, there are relatively few medical diagnoses that are “almost perfect,” the best are either “substantial” or “moderate” and many are only “fair.”^[9,10]

The reliability of a binary measure can almost always be improved with a little effort. One might standardize the protocols for testing, and the conditions of testing or measurement, to reduce the inconsistency of the subjects’ response. One might clarify the measurement protocol to raters, train the raters, and monitor their performance to reduce rater error.

If all else fails, one might use consensus ratings, i.e., one might have M raters per subject, and combine these M ratings into one. If one simply uses the number of positive ratings among the M ratings, one moves from a binary rating to an interval one. The kappa coefficient would not then be the appropriate reliability measure; the intraclass correlation coefficient would be a better choice.^[11–13] Then, the reliability of that interval measure would approximately follow the Spearman–Brown rule,^[14,15] where κ_p is the coefficient of reliability of a single binary rating,^[16] the reliability of a mean of M ordinal measures is approximately given by:

$$M\kappa_p / [1 + (M - 1)\kappa_p]$$

What this means is that if the reliability coefficient of a single binary rating is κ_p , but this is assumed inadequate, to raise the reliability of an interval consensus rating to R ; one would need approximately $R(1 - \kappa_p)/(1 - R)\kappa_p$ raters. Thus, for example, to raise the reliability from $\kappa_p = 0.4$ to $R = 0.8$, one would need $(0.8 \cdot 0.6)/(0.2 \cdot 0.4) = 6$ independent raters per subject.

The problem in medical applications is that to make decisions about inclusion/exclusion from studies, whether to treat or not, whether to hospitalize or not, an interval consensus does not well serve. One still needs a binary measure. A binary consensus measure of M raters can be generated by selecting a cut-point, J , and declaring a consensus result positive if J or more of the M raters are positive. How many raters (M) would be needed and what cut-point (J) should be used to achieve acceptable reliability is a difficult problem that can only be approached empirically.^[17,18] Intuitively, many feel that J should be just larger than $M/2$, i.e., the binary consensus of M raters should be positive if the majority agree that it is positive. Unfortunately, this is one of those occasions when intuition misleads, for this majority rule does not always yield the most reliable binary consensus for M raters. Neither does the unanimity rule, i.e., that the binary consensus is positive if all M raters agree it is positive, or that the binary consensus is negative if all M raters agree it is negative. The optimal cut-point depends strongly on the unknown distribution of p_s in the population, and can vary widely from one clinical situation to another.

Concern about the reliability of binary measures in medical research and practice is well founded. Unreliability of diagnosis, for example, means that when a patient seeks an independent second or third diagnosis, the diagnoses will not agree. That is, of course, of concern both to medical consumers and to their physicians for it increases the risk of poor clinical decision making and consequently, unnecessary, medical costs. In research applications, unreliability of measurement attenuates the precision of estimators and reduces the power of statistical tests. Thus, not only does poor reliability necessitate the use of much large samples than might otherwise be necessary to detect clinical effects, thereby increasing the cost of research studies, but even when such effects are detected, the size of effects may be attenuated, and the accuracy of their estimation poor.

WEIGHTED KAPPA: VALIDITY

Suppose now that there are two binary variables— X_{rs} , as defined above, is a “gold standard” or a binary criterion, and Y_{rs} , which is another binary variable that is to be “blindly” assessed against that “gold standard.” For example, Y may be a medical test result, or a diagnosis, or a risk factor, and X may be the presence/absence of the corresponding disorder. For subject s , the probability that a randomly selected rating will have $Y_{sr} = 1$ is q_s , where q_s has some (usually unknown) distribution from 0 to 1 over the population of subjects. The population mean $E(q_s) = Q$ the prevalence or incidence of whatever Y indicates, and the population variance is σ_p^2 . The intraclass kappa for Y is defined as $\kappa_q = \sigma_p^2/QQ'$, where $Q' = 1 - Q$.

Now the probability that both $X_{rs} = 1$ and $Y_{rs} = 1$ (a “true positive” result) is:

$$E(p_s q_s) = PQ + \rho(\kappa_p \kappa_q)^{1/2} (PP'QQ')^{1/2}$$

that $X_{rs} = 1$ and $Y_{rs} = 0$ (a “false negative”) is:

$$E(p_s (1 - q_s)) = PQ' - \rho(\kappa_p \kappa_q)^{1/2} (PP'QQ')^{1/2}$$

that $X_{rs} = 0$ and $Y_{rs} = 1$ (a “false positive”) is:

$$E((1 - p_s) q_s) = P'Q - \rho(\kappa_p \kappa_q)^{1/2} (PP'QQ')^{1/2}$$

and that both $X_{rs} = 0$ and $Y_{rs} = 0$ (a “true negative”) is:

$$E((1 - p_s)(1 - q_s)) = P'Q' + \rho(\kappa_p \kappa_q)^{1/2} (PP'QQ')^{1/2}$$

Here ρ is the product moment correlation coefficient between p_s and q_s , in the population sampled. It should be noted here that whether ρ approaches zero, or either κ_p or κ_q approaches zero, one achieves the same results—the association between X and Y will appear random. This

is why unreliability attenuates all effect sizes measuring association.

When a binary measure is evaluated against a criterion, the costs or medical consequences of a false positive and a false negative may not be the same. If, for example, one were considering breast self-examination to detect breast cancers, the primary concern would be false negatives, for a false negative might lead to a fatal delay in initiating treatment, while a false positive would mean only some unnecessary anxiety and an unnecessary trip to the doctor. On the other hand, with mammography to detect breast cancers, there would be much greater concern about false positives, for that might mean initiation of unnecessary surgical biopsy, hospitalization, anesthetization—that themselves would put a cancer-free patient at unnecessary cost and risk. Acknowledgment of such possibilities led to the proposal for the weighted kappa^[6,19] that determines where the clinical costs of using Y instead of X for decision making lies between using a random number generator and that when using X itself.

The weighted kappa has the form:^[6]

$$\kappa(r) = \rho\sigma_p\sigma_q / (PQ'r + P'Qr')$$

where r is an index ($0 \leq r \leq 1$, $r' = 1 - r$) reflecting the relative clinical importance of false negatives vs. false positives. When $r = 1$, all emphasis is placed on avoiding false negatives (as for a screening test); when $r = 0$, all emphasis is placed on avoiding false positives (as for a definitive medical test). The most common choice of weight is when $r = 1/2$, when equal emphasis is placed on avoiding both types of errors. This coefficient is easily estimated from a 2×2 table relating the binary measure to its binary criterion, where TP, TN, FP, and FN denote the observed proportion of true positives, true negatives, false positives, and false negatives, respectively,

$$\kappa(r) = (TP \cdot TN - FP \cdot FN) / (\hat{p}(1 - \hat{q})r + (1 - \hat{p})\hat{q}r')$$

When there is careful consideration of the relative importance of false positives and false negatives in a clinical research situation, the most common choice is $r = 1/2$. In fact, this choice of $r = 1/2$ is so common that $\kappa(1/2)$ is often referred to as “the” kappa coefficient, as if there were no others. Moreover, when $r = 1/2$, $\kappa(1/2)$ is often called “percentage agreement corrected for chance.” However, calling the situation when both X and Y equal 1 or both X and Y equal 0 “agreement” is a misnomer, because X and Y measure different things. Here “chance” means $PQ + P'Q'$, the probability that X and Y are both either 1 or 0 if a random number generator were used to assign a proportion P to have $X = 1$ and Q to have $Y = 1$. It should be emphasized that using $r = 1/2$ should reflect a decision about the relative importance of the two types of errors, not be an automatic choice, for that can only mislead medical decision making.

WEIGHTED KAPPA IN THE CONTEXT OF MEASURES OF 2×2 ASSOCIATION

Some of the reservations expressed about kappa coefficients simply relate to the fact that these are relatively new indices compared to other measures of 2×2 association such as Odds Ratio, risk ratios, risk differences, and the Phi coefficient. Researchers with experience in evaluating risk factors, or medical tests, or diagnoses have long realized that one can choose among the various traditional measures of 2×2 association and often pick one that makes the association seem weak, and another that makes the same association seem strong. Such arbitrariness leads to comments such as: “lies, damn lies, and statistics.” It also makes the introduction of yet new measures of 2×2 association such as kappas, which may introduce even more discrepancy, seem injudicious. However, introduction of the weighted kappas actually clarifies the situation.

Table 1 presents a 2×2 table representing the probabilities of the various outcomes of two binary measures in a population. Below the table are the definitions of a range of common measures of 2×2 association: weighted kappa, sensitivity, specificity, predictive values, risk ratios, risk differences, and Odds Ratio (where there are five mathematically equivalent definitions). They are not all similarly scaled. Phi, the risk differences, and the weighted kappas have value 0 for random decision making and 1 when there are no errors; the risk ratios and the Odds Ratio have value 1 for random decision making and are infinite when there are no errors. Sensitivity, specificity, and the predictive values do not have a fixed null value, but equal 1 when there are no errors. The scaling differences are one source of problems, for it is difficult, for example, to know whether Odds Ratio or risk ratios equal 10 are large (much larger than 1!) or small (much smaller than infinite!).

Table 1 also lists the relationship of each index to the weighted kappas. It can be readily seen that all the measures, except the Phi coefficient and the Odds Ratio, either directly (risk differences) or recalibrated (sensitivity, specificity, predictive values, risk ratios) equal a weighted kappa for some value of r . This eliminates most of the arbitrariness of 2×2 measures of association, because it clarifies that different relative importance attached to false positives and false negatives explains much of the inconsistencies between different measures of 2×2 association. The two exceptions are the Phi coefficient and the Odds Ratio.

The Phi coefficient poses little problem. If $\kappa(r) = 0$ for any r , then $\kappa(r) = 0$ for all r , and $\Phi = 0$. If there are no errors, $\kappa(r) = 1$ for all r , and $\Phi = 1$ as well. If $P = Q$, the $\kappa(r)$ is a constant $\kappa(\cdot)$ for all r , and $\Phi = \kappa(\cdot)$. In general, the Phi coefficient is the geometric mean of the extreme kappa coefficients, while $\kappa(1/2)$ is their harmonic mean. Thus, frequently, Φ and $\kappa(1/2)$ are approximately equal.

Table 1 2×2 Association and Definitions of Associated Terms, Measures of 2×2 Association and Correlation Coefficients

	$Y = 1$	$Y = 0$	
$X = 1$	True positive (TP)	False negative (FN)	$P = a + b$
$X = 0$	False positive (FP)	True negative (TN)	$P' = 1 - P = c + d$
	$Q = a + c$	$Q' = 1 - Q = b + d$	1.0
Weighted kappas (r = any value between 0 and 1):			
$\kappa(r) = (TP\ TN - FP\ FN) / (PQ'r + P'Qr')$, $r' = 1 - r$.			
Definitions:			
Phi = $(TP\ TN - FP\ FN) / (PP'QQ')^{1/2}$			
Sensitivity of Y to X (Se) = TP/P			
Specificity of Y to X (Sp) = TN/P'			
Predictive value of a :			
Positive test (PVP) = TP/Q			
Negative test (PVN) = TN/Q'			
Risk ratios:			
$RR_1 = Se / (1 - Sp)$			
$RR_2 = Sp / (1 - Se)$			
$RR_3 = PVP / (1 - PVN)$			
$RR_4 = PVN / (1 - PVP)$			
Risk differences:			
$RD_1 = Se + Sp - 1$			
$RD_2 = PVP + PVN - 1$			
Odds ratio:			
$OR = (TP\ TN) / (FP\ FN)$			
$= (Se\ Sp) / ((1 - Se)(1 - Sp))$			
$= (PVP\ PVN) / ((1 - PVP)(1 - PVN))$			
$= RR_1\ RR_2 = RR_3\ RR_4$			
Relationship to weighted kappas			
Phi = $(\kappa(0)\kappa(1))^{1/2}$			
$(Se - Q) / (1 - Q) = \kappa(1)$			
$(Sp - Q') / Q = \kappa(0)$			
$(PVP - P) / P' = \kappa(0)$			
$(PVN - P') / P = \kappa(1)$			
$(RR_1 - 1) / (RR_1 + P' / P) = \kappa(0)$			
$(RR_2 - 1) / (RR_2 + P / P') = \kappa(1)$			
$(RR_3 - 1) / (RR_3 + Q' / Q) = \kappa(1)$			
$(RR_4 - 1) / (RR_4 + Q / Q') = \kappa(0)$			
$RD_1 = \kappa(P')$			
$RD_2 = \kappa(Q)$			
$OR - 1 = (P'\kappa(1) + P\kappa(0)) / (PP'(1 - \kappa(0))(1 - \kappa(1)))$			
$= (Q\kappa(1) + Q'\kappa(0)) / (QQ'(1 - \kappa(0))(1 - \kappa(1)))$			

Among the other measures of 2×2 association, the Phi coefficient is particularly important, because the 2×2 chi square test is most commonly used to test the null hypothesis of association equals $N\phi^2$ and:

$$\phi^2 = \kappa(1)\kappa(0) = \rho^2(\kappa_p, \kappa_q)$$

From this, it is obvious that the power of the test to detect association depends on: 1) the association between the true values of X and Y (ρ) and 2) the reliabilities of measurement of X and of Y (κ_p and κ_q).

That leaves the Odds Ratio, where there is a serious problem. When there is random association between X and Y , $\kappa(r) = 0$ for all r ; Odds Ratio equal 1. Similarly, when there are no errors, neither false positives nor false negatives, then $\kappa(r) = 1$ for all r , and the Odds Ratio equals infinity. Thus far, there are no discrepancies. However, suppose there were no false negatives, but many false positives. Then $\kappa(1) = 1$, but $\kappa(r) < 1$ for $r \neq 1$, and Odds Ratio is infinite. That $\kappa(1) = 1$ here is logical because $r = 1$

indicates that only the absent false negatives are of clinical importance. Similarly, when there are no false positives, but there are false negatives, $\kappa(0) = 1$, and $\kappa(r) < 1$ for all $r \neq 0$, but once again Odds Ratio is infinite. Again, here $\kappa(0) = 1$ is logical because $r = 0$ indicates that only the absent false positives are of clinical importance. Thus the Odds Ratio is infinite if either kind of error is absent, even if absent error is the one that clinically matters little, and the kind of error that is present but ignored is the one that matters a great deal. This is troublesome for the use and interpretation of the Odds Ratio in situations when clinical decision making can be misled.

When $P = Q = 1/2$, all weighted kappas are equal, $\kappa(r) = \kappa(\cdot)$. Then $Y = (\sqrt{OR} - 1) / (\sqrt{OR} + 1)$ equals $\kappa(\cdot)$ (Y is called "Yule's Index"). Thus when $P = Q = 1/2$, $\kappa(\cdot) = 0.2, 0.4, 0.6, 0.8$ correspond to Odds Ratios of 2, 5, 16, 81. That an Odds Ratio of about 2 would be considered "slight" corresponds poorly to interpretations of the Odds Ratio in the research literature, where an Odds Ratio of 2 often generates considerable scientific interest. This begins to suggest that the standards used to assess the magnitude

of the Odds Ratio are much less demanding than those for assessing kappa coefficients.

When $P = Q \neq 1/2$, again all weighted kappas are equal, $\kappa(r) = \kappa(\cdot) = \text{Phi}$. Then

$$\text{OR} - 1 = \kappa(\cdot) / PP'(1 - \kappa(\cdot)^2)$$

and Yule's index is now larger than $\kappa(\cdot)$, actually approaching 1, whenever P approaches either 0 or 1. Thus, for example, when $P = Q = 0.01$ and $\kappa(\cdot) = 0.2$, Odds Ratio = 32.6, and Yule's Index = 0.702. The latter suggest very strong association when all the weighted kappas indicate slight association. Which is right?

Suppose, in this case, that Y were a medical test to detect a fatal condition that could be averted with aggressive treatment, where the treatment would be costly and risky, but would prevent death. Then for every 10,000 people, 100 would have the condition. Of the 100 who have the condition, 21 would be detected by the test, and the remaining 79 would die of the condition undetected. Those 21 would be aggressively treated and saved. Of the 9900 who do not have the condition, 9821 would test negative, and 79 positive. Those 79 would also be treated aggressively because they would be thought to have the condition, incurring the cost and risk of the treatment, totally unnecessarily. Of the 100 subjects treated, 79 are treated unnecessarily. Thus one benefits 21 patients and harms 79. If the test itself carried cost or risk, the remaining 9900 whose fate is unaffected by the test would also incur some harm. In that case, one would benefit 21 patients and harm 9979. The Odds Ratio here of 32.7 ($R = 0.70$) would seem overly enthusiastic, and $\kappa(\cdot) = 0.2$ more reasonable.

In general, when association is nonrandom and either P or Q is extreme (near 0 or 1), Odds Ratio tends to be very large, when all other measures of association, most particularly the weighted kappas, are much smaller. As illustrated above, when one examines the consequences of using Y instead of X in such cases, it is often clear that the Odds Ratio tends to exaggerate association.

Clearly, anything that improves the reliability of X or of Y will improve the validity of Y assessed against X , regardless of what measure is used. In addition, one might also implement some action that would change the value of Q , for the validity of Y assessed against X , whatever the value of r , depends on the relative size of P and Q . Frequently, binary Y is defined by some scales exceeding a cut-point, e.g., systolic blood pressure greater than 120 mm Hg, or smoking more than 1 cigarette per day, or scoring less than 100 on an IQ test. In such cases, one can negotiate the cut-point to achieve the optimal $\kappa(r)$ for whatever value of r is appropriate to the context. Generally, best results are obtained when r is near 1 if $Q > P$, when r is near 0 if $Q < P$, and when r is near 1/2 if Q is approximately equal to P .

PROBLEMS WITH KAPPA?

The Base Rate Problem

One recurrent issue is the so-called "base rate problem" with kappa coefficients, i.e., that both the intraclass kappa for fixed X and weighted kappas for a fixed X and Y change when applied to different populations (differing P). Such change is a fact, but why it is regarded as a problem specific to kappa coefficients is not clear. Every correlation coefficient, for example, is likely to change as one moves from one population to another.

The genesis of labeling this as a problem for kappa in particular seems to be based on the mistaken assumptions: 1) that the sensitivity and specificity of Y to X are fixed constants, that p_s for each subject is either a constant Se or a constant Sp ; and 2) that the "gold standard" is pure gold (perfect reliability). If these assumptions were true, then regardless of which subpopulation were sampled (whatever the value of P), the Odds Ratio would equal $(\text{Se} \cdot \text{Sp}) / ((1 - \text{Se})(1 - \text{Sp}))$, a fixed constant. However, as with the two-point model, kappa coefficients would vary from one population to another depending on the value of P . However, almost no "gold standards" used in medical research are pure gold, having perfect reliability. Furthermore, as pointed out earlier, the two-point distribution model seldom, if ever, holds. Consequently, Odds Ratio varies from one subpopulation to another, just like kappa coefficients and other measures of association.

Multicategory Kappa

When both intraclass and weighted kappas were introduced, a multicategory version was proposed, i.e., a kappa useful when there were more than two unordered categories (e.g., ethnic group, country of birth).^[19] Such kappas have proved problematic. For example, the intraclass kappa when there are $C > 2$ categories can be expressed as:^[16]

$$\kappa_{\text{mult}} = \sum P_c P'_c \kappa_c / \sum P_c P'_c$$

where P_c ($c = 1, 2, \dots, C$) is the prevalence of category c in the population, $P'_c = 1 = P_c$, and κ_c is the intraclass kappa for the binary measure $X_c = 1$ for all subjects in category c , and $X_c = 0$ otherwise. If there is one category with high kappa and its P_c is near 0.5 (to maximize $P_c P'_c$), even if all the other categories are completely indistinguishable from each other (κ'_c 's near zero), the multicategory kappa would be very high. This works in the other direction as well; the multicategory kappa may be very near zero even if all the categories (relatively rare ones) are perfectly distinguishable, and one common category is unreliable. Consequently, the multicategory kappa may be misleading because it attributes to the categorical measure, as a whole, a reliability that may only apply to one of its categories. For

this reason, it is preferable to examine the individual binary kappas (κ_c , $c = 1, 2, \dots, C$) in assessing the reliability of a multicategory measure, as well as the cross-classifications to understand where errors might occur and why, rather than to use the multicategory kappa.

Kappa for Ordered Categories

Kappa coefficients have also been used for ordered categories (i.e., ordinal measures with a small number of responses, such as 5-point scales). In addition to the problem with multiple unordered categories (cited above), use of kappa here ignores the ordinal nature of the data. It has been pointed out that if one assigns scores to the ordered categories, and uses as a weight in the weighted kappa the absolute difference between the scores, the weighted kappa is then equivalent to an intraclass correlation coefficient.^[11,20] In fact, it is some form of intraclass correlation coefficient that is preferable for use with ordinal measures over the use of any kappa coefficient.

Interchanging the Intraclass and Weighted Kappa Coefficients

Kirk and Kuchins^[4] blamed the “discovery” of the kappa coefficients [specifically $\kappa(1/2)$] for impeding the progress in developing psychiatric diagnosis. Their argument was that the emphasis that kappa coefficients place on reliability assessment obscured the much more important issue of the validity of diagnosis. In fact, while the intraclass kappa is a measure of reliability, the weighted kappa [e.g., $\kappa(1/2)$] is a measure of validity. Thus it is not kappa per se that distracted focus from validity. In fact, it was the realization that one cannot achieve validity unless one already has reasonable reliability. There was a judgment call that psychiatric diagnosis was not then yet reliable enough.

However, this type of misunderstanding has been quite common. In fact, $\kappa(1/2)$ is often inappropriately used as a measure of reliability, not validity. Only if the design of the study is such that $p_s = q_s$ for all subjects in the population, does $\kappa(1/2)$ estimate the same parameter as does the intraclass kappa, because then, $\rho = 1$, $\sigma_p = \sigma_q$, and $P = Q$. Even then, the sample estimate of $\kappa(1/2)$ would be less efficient than the sample estimate of intraclass kappa for the same population parameter.

In short, intraclass kappas and weighted kappas play different but interrelated roles in medical research applications, and the best coefficient should be chosen for each particular context in which it is to be used.

CONCLUSION

Intraclass and weighted kappas play important roles in medical research: intraclass kappa as a measure of

reproducibility or reliability, weighted kappas as unique measures of validity.

The major argument for the intraclass kappa as a measure of reliability is that it corresponds exactly with the classical definition of reliability, that it is easily interpretable clinically as well, and that it plays a recognizable role in the precision of parameter estimation and statistical hypothesis testing.

The uniqueness of the weighted kappa as a measure of validity lies in the weight that defines it, which reflects the relative clinical importance of false negatives to false positives in the context of application. It can be shown that most other measures of 2×2 association relate to one or another weighted kappa, thus clarifying that the discrepancies long noted between various measures of 2×2 association primarily relate to a choice of weights to false positives and false negatives, often a choice unknowingly made by the user. The major exception is the Odds Ratio, which often tends to exaggerate association.

It is true that kappa coefficients must move past their historical origins as “percentage agreement corrected for chance,” a description that ill characterizes what either kappa coefficient is designed to do. Moreover, many claims have been mistakenly made, either about kappa itself, or contrasts between other measures of association (Odds Ratio in particular) and kappa. Most important is the claim that Odds Ratio is constant over populations (untrue), while the kappa coefficients vary (true): the base-rate problem. In many cases, kappa coefficients have been misused, either for more than two categories, or for ordinal measures, or the intraclass kappa has been used in situations when the weighted kappa should have been used or the weighted kappa when the intraclass kappa should have been used.

Finally, the process of considering what relative weight to attach to false positives and to false negatives is not an easy one, nor one that clinicians, researchers, or policy-makers are accustomed to make. Nevertheless, the basis of medical decision making require such considerations so as to distinguish between accurate and inaccurate medical tests, to correctly identify risk factors for disorders, to identify those likely to respond to a treatment, and other similar tasks. Thus kappa coefficients are invaluable in the contexts for which they are designed—i.e., the reliability and validity of binary measures, particularly in the contexts of medical test evaluation, risk factor evaluation and evaluation of outcomes in randomized clinical trials.

REFERENCES

1. Brenner, H.; Kliebsch, U. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology* **1996**, 7 (2), 199–202.
2. Kraemer, H.C. Extensions of the kappa coefficient. *Biometrics* **1980**, 36, 207–216.

3. Fleiss, J.L.; Nee, J.C.M.; Landis, J.R. Large sample variance of kappa in the case of different sets of raters. *Psychol. Bull.* **1979**, *86* (5), 974–977.
4. Kirk, S.A.; Kutchins, H. *The Selling of the DSM: The Rhetoric of Science in Psychiatry*; Aldine De Gruyter: New York, **1992**.
5. Kraemer, H.C. Measurement of reliability for categorical data in medical research. *Stat. Methods Med. Res.* **1992**, *1*, 183–199.
6. Bloch, D.A.; Kraemer, H.C. 2×2 kappa coefficients: Measures of agreement or association. *Biometrics* **1989**, *45*, 269–287.
7. Fleiss, J.L. *Statistical Methods For Rates and Proportions*; John Wiley & Sons: New York, 1981.
8. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174.
9. Koran, L.M. The reliability of clinical methods, data and judgments, part 1. *N. Engl. J. Med.* **1975**, *293*, 642–646.
10. Koran, L.M. The reliability of clinical methods, data and judgments, part 2. *N. Engl. J. Med.* **1975**, *293*, 695–701.
11. Bartko, J.J. The intraclass correlation coefficient as a measure of reliability. *Psychol. Rep.* **1966**, *19*, 3–11.
12. Bartko, J.J.; Carpenter, W.T. On the methods and theory of reliability. *J. Nerv. Ment. Dis.* **1976**, *163*, 307–317.
13. Bartko, J.J. On various intraclass correlation reliability coefficients. *Psychol. Bull.* **1976**, *83*, 762–765.
14. Spearman, C. Correlation calculated from faulty data. *Br. J. Psychol.* **1910**, *3*, 271–295.
15. Brown, W. Some experimental results in the correlation of mental abilities. *Br. J. Psychol.* **1910**, *3*, 296–322.
16. Kraemer, H.C. Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika* **1979**, *44* (4), 461–472.
17. Kraemer, H.C.; Periyakoil, V.S.; Noda, A. Tutorial in biostatistics: Kappa coefficients in medical research. *Stat. Med.* **2002**, *21*, 2109–2129.
18. Kraemer, H.C. How many raters? Toward the most reliable diagnostic consensus. *Stat. Med.* **1992**, *11*, 317–331.
19. Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213–229.
20. Fleiss, J.L.; Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.* **1973**, *33*, 613–619.

Kruskal-Wallis Test for Ordered Categorical Data: Power Calculation

Chunpeng Fan and Donghui Zhang

Department of Biostatistics and Programming, Sanofi US Inc., Bridgewater, New Jersey, U.S.A.

Abstract

Although the Kruskal–Wallis test has been widely used to analyze ordered categorical data, power and sample size methods for this test have been investigated to a much lesser extent when the underlying multinomial distributions are unknown. This entry generalizes the power and sample size procedures proposed by Fan, Zhang, and Zhang for continuous data to ordered categorical data, when estimates from a pilot study are used in the place of knowledge of the true underlying distribution. A myelin oligodendrocyte glycoprotein-induced experimental autoimmune encephalomyelitis mouse study is used to demonstrate the application of the methods.

Keywords: Ties; Ordinal data; Rank test.

INTRODUCTION

As the non-parametric equivalent of the parametric one-way analysis of variance (ANOVA) model, the rank-based Kruskal–Wallis (1952) test has been widely investigated in statistical science and widely used in pharmaceutical research when the normal residual assumption of its parametric equivalent does not hold. An especially useful scenario of the Kruskal–Wallis test is when the responses are ordinal categorical data, where the normal residual assumption almost never holds. Although the size of the Kruskal–Wallis test has been extensively investigated (see, for example, Iman and Davenport^[1]), the power and sample size calculations have been investigated to a much lesser extent.^[2] The power and sample size procedures proposed by Fan, Zhang, and Zhang^[2] for continuous data can be generated to ordered categorical data, when estimates from a pilot study are used to replace the unknown true underlying distribution. A myelin oligodendrocyte glycoprotein (MOG)-induced experimental autoimmune encephalomyelitis (EAE) mouse study is used to demonstrate the application of the methods.

KRUSKAL–WALLIS TEST FOR ORDERED CATEGORICAL DATA

Let $F_1, \dots, F_K$ denote $K \geq 2$ unknown population cumulative distribution functions, from which K groups of

independent random samples of sizes $N_1, \dots, N_K$ are drawn, respectively. The null and the alternative hypotheses for the Kruskal–Wallis test are

$$\begin{aligned} H_0 : F_1 = \dots = F_K = F^0 \text{ versus} \\ H_a : \text{at least one } F_i \text{ is not equal to } F^0 \end{aligned} \quad (1)$$

Assume an ordered categorical response has C possible values with C being a positive integer. Consider independent random samples $\{X_{ij} \in \{1, \dots, C\} : i = 1, \dots, K, j = 1, \dots, N_i\}$ collected in experiments involving K treatment groups. Let R_{ij} be the rank of the j^{th} observation in the i^{th} sample in the combined sample and $R_i = \sum_{j=1}^{N_i} R_{ij}$. The Kruskal–Wallis test statistic for H_0 versus H_a in Eq. 1 is:

$$S_{\text{KW}} = \frac{\sum_{i=1}^K N_i \left(\frac{R_i}{N_i} - \frac{N+1}{2} \right)^2}{\frac{1}{N-1} \sum_{i=1}^K \sum_{j=1}^{N_i} \left(R_{ij} - \frac{N+1}{2} \right)^2} \quad (2)$$

where $N = \sum_{i=1}^K N_i$ denotes the total sample size.^[3]

Note that when no ties are present, S_{KW} becomes:

$$S_{\text{KW}}^* = \frac{12}{N(N+1)} \sum_{i=1}^K N_i \left(\frac{R_i}{N_i} - \frac{N+1}{2} \right)^2 \quad (3)$$

which is the Kruskal–Wallis test statistic used in the study by Fan, Zhang, and Zhang.^[2]

GENERAL POWER AND SAMPLE SIZE FORMULA

Fan, Zhang, and Zhang^[2] derived the asymptotic properties of S_{KW}^* under the alternative H_a . Theorem 1 below generalizes the asymptotic results to S_{KW} , and its proof can be found in Fan and Zhang.^[4]

THEOREM 1. Under alternative hypotheses H_a in Eq. 1, suppose there exists a distribution F^0 such that $F_i \rightarrow_d F^0$ as $N \rightarrow \infty$ for all $i=1, \dots, K$. Assume $\sigma^2 = \text{Var}\{[F^0(X-) + F^0(X)]/2\} > 0$ with X being a random variable from F^0 . Suppose $\lambda_i = N_i/N \geq \lambda_0$ for all $i \leq K$ and a fixed $\lambda_0 > 0$. Then, as $N \rightarrow \infty$, $(Z_1, \dots, Z_K)^T \rightarrow_d N(O, \sigma^2 \Sigma)$, where $Z_i = N^{-3/2}(R_i - ER_i)$ and $\Sigma = (\lambda_i I\{i=i'\} - \lambda_i \lambda_{i'})_{K \times K}$, for $i, i'=1, \dots, K$. Moreover, the Kruskal-Wallis test statistic S_{KW} approximately has the non-central chi-square distribution $\chi_{K-1}^2(\tau)$, where the non-centrality $\tau = \sum_{i=1}^K N_i \{ER_i/N_i - (N+1)/2\}^2 / (N^2 \sigma^2)$ and a consistent estimate of τ can be obtained by replacing σ^2 with its consistent estimate $s^2/N^2 = \sum_{i=1}^K \sum_{j=1}^{N_i} \{R_{ij} - (N+1)/2\}^2 / (N^2(N-1))$.

Let $p_{ik} = \Pr(X_{ki} < X_{kj}) + \Pr(X_{ki} = X_{kj})/2$ for $i, k=1, \dots, K$. Because $R_i = \sum_{j=1}^{N_i} R_{ij}$, we have $ER_i/N_i - (N+1)/2 = N \sum_{k \neq i} \lambda_k (p_{ik} - 1/2)$ and

$$\tau = \frac{N}{\sigma^2} \sum_{i=1}^K \lambda_i \left\{ \sum_{k \neq i} \lambda_k (p_{ik} - 1/2) \right\}^2 \quad (4)$$

When all p_{ik} s and σ^2 are known, for a study with fixed N and $\lambda_1, \dots, \lambda_K$, we can calculate τ and obtain the power by

$$\text{Power}^{\text{known}} = \Pr\{\chi_{K-1}^2(\tau) > \chi_{K-1, 1-\alpha}^2\} \quad (5)$$

where $\chi_{K-1, 1-\alpha}^2$ is the $(1-\alpha)^{\text{th}}$ quantile of the (central) chi-square distribution with $K-1$ degrees of freedom.

For a study with prespecified power $1-\beta$, the required total sample size for all K groups can be calculated by

$$N^{\text{known}} = \frac{\tau_{\alpha, \beta} \sigma^2}{\sum_{i=1}^K \lambda_i \left\{ \sum_{k \neq i} \lambda_k (p_{ik} - 1/2) \right\}^2} \quad (6)$$

where $\tau_{\alpha, \beta}$ is the solution of the equation $\Pr\{\chi_{K-1}^2(\tau) > \chi_{K-1, 1-\alpha}^2\} = 1-\beta$.

POWER AND SAMPLE SIZE FORMULA FOR ORDERED CATEGORICAL DATA

For the ordinal response data, instead of the knowledge about p_{ik} and σ^2 , practitioners may prespecify the probabilities of observing each category in each group. That is, $\{\pi_{ic} = \Pr(X_{ij} = c) : i=1, \dots, K, c=1, \dots, C\}$ may be known or prespecified. In this case, the power formula 5 and sample size formula 6 need to be reparameterized using $\{\pi_{ic} : i=1, K, c=1, C\}$.

To this aim, we can write that for $i \neq k$,

$$p_{ik} = \sum_{l=1}^{C-1} \sum_{j=l+1}^C \pi_{ij} \pi_{kl} + \frac{1}{2} \sum_{l=1}^C \pi_{il} \pi_{kl} \quad (7)$$

And for σ^2 , we denote $\pi_c = \Pr(X=c) = \sum_{i=1}^K \lambda_i \pi_{ic}$ for $c=1, \dots, C$ with random variable X coming from $F^0 = \sum_{i=1}^K F_i/K$. Then $F^0(c) = \sum_{j=1}^c \pi_j$ and

$$\begin{aligned} \sigma^2 &= \text{Var}\left\{ \frac{F^0(X-) + F^0(X)}{2} \right\} \\ &= E\left\{ \frac{F^0(X-) + F^0(X)}{2} \right\}^2 - \left[E\left\{ \frac{F^0(X-) + F^0(X)}{2} \right\} \right]^2 \\ &= \sum_{c=1}^C \pi_c \left\{ \frac{F^0(c-) + F^0(c)}{2} \right\}^2 - \left[\sum_{c=1}^C \pi_c \left\{ \frac{F^0(c-) + F^0(c)}{2} \right\} \right]^2 \\ &= \sum_{c=1}^C \pi_c \left(\sum_{j=1}^{c-1} \pi_j + \frac{1}{2} \pi_c \right)^2 - \left[\sum_{c=1}^C \pi_c \left(\sum_{j=1}^{c-1} \pi_j + \frac{1}{2} \pi_c \right) \right]^2 \end{aligned} \quad (8)$$

Power for the Kruskal-Wallis test can be calculated by applying Eqs. 7 and 8 in Eq. 5. Required sample size by the Kruskal-Wallis test can be calculated by applying Eqs. 7 and 8 in Eq. 6.

When true values of the parameters $\{\pi_{ic} : i=1, \dots, K, c=1, \dots, C\}$ are unknown and the aim is to calculate power and sample size based on a study with observations $\{X_{ij} \in \{1, \dots, C\} : i=1, \dots, K, j=1, \dots, N_i\}$, the power and sample size can be calculated by replacing p_{ik} s and σ^2 by their corresponding estimates in Eqs. 5 and 6, respectively.

For $i, k=1, \dots, K, i \neq k$, we can estimate p_{ik} by:

$$\hat{p}_{ik} = \frac{1}{N_i N_k} \sum_{j=1}^{N_i} \sum_{l=1}^{N_k} \left(I\{X_{ki} < X_{lj}\} + \frac{1}{2} I\{X_{ki} = X_{lj}\} \right) \quad (9)$$

and estimate σ^2 by $\hat{\sigma}^2 = s^2/N^2$ defined in theorem 1. The power and sample size by this method are denoted as $\widehat{\text{power}}^{\text{known}}$ and $\hat{N}^{\text{known}}$, respectively.

Asymptotically, $\hat{\tau}$ converges to its true value τ . However, with finite sample size in the pilot study, $\hat{\tau}$ is a biased estimator of τ . In our proposed method, we apply adjustments to $\hat{\tau}$ to correct such bias. To guarantee the corrected $\hat{\tau}, \hat{\tau}^*$, to be positive, we may assume that $\hat{\tau}^* = \gamma \hat{\tau}$ for some positive scalar γ . Because $\hat{\tau}$ appears in the denominator of the sample size formula 6, we match the median of $\hat{\tau}^*$ to be τ . By setting median($\hat{\tau}^*$) = median($\gamma \hat{\tau}$) = τ , we have $\gamma = \tau / \text{median}(\hat{\tau})$. Since $\hat{\tau} \rightarrow_d \chi_{K-1}^2(\tau)$, median($\hat{\tau}$) can be estimated by median($\{\chi_{K-1}^2(\hat{\tau})\}$). Therefore,

$$\hat{\tau}^* \approx \hat{\gamma} \hat{\tau} = \frac{\hat{\tau}}{\text{median}\{\chi_{K-1}^2(\hat{\tau})\}} \hat{\tau} \quad (10)$$

With such adjustment, the power Eq. 5 becomes:

$$\widehat{\text{Power}}^{\text{propose}} = \Pr\left\{\chi_{K-1}^2(\hat{\tau}^*) > \chi_{K-1,1-\alpha}^2\right\} \quad (11)$$

and the sample size:

$$\hat{N}_{\text{propose}} = \frac{N\tau_{\alpha,\beta}}{\hat{\tau}} = \hat{N}_{\text{known}} \frac{\text{median}\{\chi_{K-1}^2(\hat{\tau})\}}{\hat{\tau}} \quad (12)$$

Numerical results in Fan and Zhang^[4] show that $\widehat{\text{power}}^{\text{propose}}$ and $\hat{N}_{\text{propose}}$ can give appropriate estimated power and sample size, while $\widehat{\text{power}}^{\text{known}}$ and $\hat{N}_{\text{known}}$ are associated with biased results.

AN EXAMPLE OF AN MOG-INDUCED EAE MOUSE STUDY

The MOG-induced murine EAE is a widely accepted animal model for studying the clinical and pathological features of multiple sclerosis, which is an inflammatory demyelinating disease of the central nervous system.^[5,6] A major indication of EAE in the MOG-EAE model is the spinal cord inflammation. A mouse study was carried out to investigate the effect of a compound C in an MOG-EAE model. This study had four treatment groups—vehicle control (Veh, 20 mice), compound C with concentration 3 mg/kg (C03, 20 mice), compound C with 10 mg/kg (C10, 21 mice), and compound C with 30 mg/kg (C30, 21 mice). All 82 mice were dosed daily after active induction of EAE. At the late phase of disease (day 28), all mice were anesthetized and spinal cords were carefully removed. The lumbar spinal cord of each mouse was dissected, and the histological score for inflammation was determined as in the range from 0 (no inflammation) to 4 (inflammation cells spread out and massive). The goal of the analysis is to compare the treatment effects in terms of the inflammation scores, and in the case when such treatment effects are insignificant (significance level $\alpha = 0.05$), to find the minimum required sample size to achieve desired power in detecting the difference with significance level α . The data can be expressed in terms of the number of mice falling into each score category for each treatment, as in Table 1.

Table 1 MOG-EAE mouse model: Frequencies of mice in each score category for each treatment

Treatment	Score					Total
	0	1	2	3	4	
Veh	2	4	7	7	0	20
C03	5	6	8	1	0	20
C10	2	7	9	1	2	21
C30	7	6	5	2	1	21

Table 2 MOG-EAE mouse model: Power for current study and required sample sizes for future studies to achieve higher powers

N	Current study		Future study		
	$\widehat{\text{power}}^{\text{known}}$	$\widehat{\text{power}}^{\text{propose}}$	Target power	$(\hat{N}_{\text{known}})$	$(\hat{N}_{\text{propose}})$
82	0.599	0.483	0.70	101	130
			0.80	125	161
			0.90	163	210

For this score data, instead of the parametric one-way ANOVA model whose normality assumption is not satisfied, the Kruskal–Wallis test is an appropriate method. The p value for the treatment effects by the Kruskal–Wallis test is 0.0677. One potential reason for such insignificance is that the power of the current study is too low. Since the true values of probabilities π_{ic} s are unknown, we use the $\widehat{\text{power}}^{\text{known}}$ and $\widehat{\text{power}}^{\text{propose}}$ to calculate the power. The results are shown in Table 2. Because the observed treatment effect is insignificant, it is of interest to find the minimum required sample size to detect a difference under significance level α and desired power. Table 2 also shows the required sample sizes for target power 0.70, 0.80, and 0.90, respectively, using the two sample size calculation methods $\hat{N}_{\text{known}}$ and $\hat{N}_{\text{propose}}$. The sample size calculations here assume that the future larger scale study has the same sample size proportions as the current study, i.e., $\lambda_1 : \lambda_2 : \lambda_3 : \lambda_4 = 20 : 20 : 21 : 21$.

In this example, with estimated probabilities π'_{ic} s in the current study, formula 5 ($\widehat{\text{power}}^{\text{known}}$) gives a higher power estimate than the proposed adjusted method (formula 11, $\widehat{\text{power}}^{\text{propose}}$). And the appropriate proposed method shows that the power of the current study is as low as 0.483. In terms of the required sample size to achieve power 0.8, the proposed method ($\hat{N}_{\text{propose}}$) suggests a total sample size 161, which is about double the current sample size. But formula 6 ($\hat{N}_{\text{known}}$) gives a much smaller sample size of 125.

SAS® code to realize the power and sample size formulas is available upon request from Dr. C. Fan.

CONCLUSIONS AND DISCUSSIONS

Under H_a , when the true probability mass functions of $F_1, \dots, F_K$ are known, the power and sample size calculation methods, formulas 5 and 6, can be applied. When the probability mass functions are unknown but their estimates can be obtained from a pilot study, the proposed power and sample size calculation methods, formulas 11 and 12, are appropriate.

For the proposed power calculation to be accurate, the sample size in each group in the pilot study should not be too small. The power formulas do not work well with

sample size 60 or 100. The proposed sample size calculation may not be accurate when the power of the pilot study is too small.

In practice, the true values of π_{ic} s are seldom known. Therefore, practitioners may estimate the probability mass functions π_{ic} s from a historical study and use the estimated π_{ic} s to plan future studies by calculating the required sample sizes with Monte Carlo simulations. However, this does not take the variation in $\hat{\pi}_{ic}$ into consideration. Although this is not an issue in the two-sample Wilcoxon test,^{[2][7]} for the K-sample Kruskal–Wallis test with $K \geq 3$, this causes overestimation of the power and underestimation of the required sample size. Since the Monte Carlo power and sample size calculation methods can be used only with true values of π_{ic} s, it is important to have confidence in the source of π_{ic} s when using the methods.

ACKNOWLEDGMENT

This entry previously appeared as “A note on power and sample size calculations for the Kruskal-Wallis test for ordered categorical data” in the *Journal of Biopharmaceutical Statistics* volume 22.^[4] The authors acknowledge the permission granted by Taylor & Francis Group for this republication.

REFERENCES

1. Iman, R.L.; Davenport, J.M. New approximations to the exact distribution of the Kruskal-Wallis test statistic. *Commun. Stat-Theor. M.* **1976**, *5*, 1335–1348.
2. Fan, C.; Zhang, D.; Zhang, C.H. On sample size of the Kruskal-Wallis test with application to a mouse peritoneal cavity study. *Biometrics* **2011**, *67*, 213–224.
3. Conover, W.J.; Iman, R.L. Rank transformations as a bridge between parametric and nonparametric statistics. *Am. Stat.* **1981**, *35*, 124–133.
4. Fan, C.; Zhang, D. A note on power and sample size calculations for the Kruskal-Wallis test for ordered categorical data. *J. Biopharm. Stat.* **2012**, *22*, 1162–1173.
5. Mi, S.; Hu, B.; Hahm, K.; Luo, Y.; Hui, E.S.K.; Yuan, Q.; Wong, W.M.; Wang, L.; Su, H.; Chu, T.-H.; Guo, J.; Zhang, W.; So, K.-F.; Pepinsky, B.; Shao, Z.; Graff, C.; Garber, E.; Jung, V.; Wu, E.X.; Wu, W. LINGO-I antagonist promotes spinal cord remyelination and axonal integrity in MOG-induced experimental autoimmune encephalomyelitis. *Nat. Med.* **2007**, *13*, 1228–1233.
6. Herrero-Herranza, E.; Pardo, L.A.; Goldb, R.; Linker, R.A. Pattern of axonal injury in murine myelin oligodendrocyte glycoprotein induced experimental autoimmune encephalomyelitis: Implications for multiple sclerosis. *Neurobiol. Dis.* **2008**, *30*, 162–173.
7. Zhao, Y.D.; Rahardja, D.; Qu, Y. Sample size calculation for the Wilcoxon-Mann-Whitney test adjusting for ties. *Stat. Med.* **2008**, *27*, 462–468.

Kullback–Leibler Divergence for Evaluating Equivalence

Vladimir Dragalin

GlaxoSmithKline, Collegeville, Pennsylvania, U.S.A.

Abstract

Kullback–Leibler Divergence for Evaluating Equivalence The Kullback–Leibler divergence metric is a viable alternative to that proposed by the FDA for the assessment of bioequivalence. This entry details the benefits of this type of bioequivalence study by providing models for use.

INTRODUCTION

In many applications, we are interested in showing that two independent samples arise from similar underlying populations with distribution functions F and G , respectively. Testing the hypothesis $H_0: F = G$ vs. the alternative $H_a: F \neq G$ is not suitable for the assessment of similarity of populations. Even when the observed p value of a test for H_0 is large, this does not allow for a controlled error rate when assessing similarity. In equivalence trials, the issue is not to show a difference between the two populations, but to demonstrate that one population is not different from the other by more than a prespecified irrelevant amount. The assessment of equivalence is of interest in various fields: psychology,^[1] chemistry,^[2] environmental statistics,^[3] bridging assessment studies,^[4] vaccine consistency,^[5] neurophysiology,^[6] to name a few. Bioequivalence testing has become a challenging field during the last two decades with an increased interest from academia, regulatory agencies, and pharmaceutical industry.^[7]

Bioequivalence trials seek to demonstrate the similarity of the new formulation, such as a new generic product, to the old formulation, such as the innovator's product, based on the surrogate outcome of plasma concentration of the drug and/or its metabolites. A new formulation that demonstrates a much greater or much lower bioavailability is unacceptable.

The area under the concentration curve (AUC), the maximum concentration ($C_{\max}$), and the time to maximum concentration ($T_{\max}$) are commonly used pharmacokinetic characteristics of the tested formulations.

OVERVIEW

In general, bioequivalence represents a statement about distance between distributions of the pharmacokinetic responses on test and reference formulations. Recently, Dragalin et al.^[8] proposed a unified methodology, based on the Kullback–Leibler divergence (KLD) for the assessment

of bioequivalence that encompasses all three aspects: average bioavailability, prescribability, and switchability. To demonstrate bioequivalence, the following hypotheses are tested:

$$H_0: d(f_T, f_R) > d_0 \text{ vs. } H_a: d(f_T, f_R) \leq d_0$$

where f_T and f_R are the appropriate density functions of the observations from test and reference formulation, respectively, $d(f_T, f_R)$ is KLD, and d_0 is the goalpost to be set by a regulator.

If we reject the null hypothesis H_0 on the basis of observations from a trial, then we may conclude that the test and reference formulations are bioequivalent. On the other hand, if we do not reject H_0 , then bioequivalence is not proved. This does not necessarily mean that the two formulations are not bioequivalent; perhaps, more observations are needed to prove it. Equivalently, we may construct a one-sided confidence interval with upper bound d_U for $d(f_T, f_R)$ and check it with d_0 ; if $d_U < d_0$, then accept bioequivalence; otherwise, reject it.

The current worldwide regulatory standard for evaluating bioequivalence is the so-called average bioequivalence (ABE), which compares the means μ_T and μ_R of the distributions of the response variable on the two formulations, test (T) and reference (R), respectively. The ABE criterion is formulated as:^[9]

$$(\mu_T - \mu_R)^2 \leq \vartheta_A$$

where ϑ_A is a goalpost, a standard set by regulatory agencies that defines how “close” the two formulations must be to be declared bioequivalent. Various statistical methods for evaluating the ABE have appeared.^[10–14] The so-called two one-sided tests^[15,16] are now considered to be the standard method.

However, many authors^[17–20] argue that in some situations, it is not sufficient to focus only on population averages of test and reference formulations. Besides means, the variances σ_T^2 and σ_R^2 of the test and reference distributions,

respectively, could be used to evaluate the bioequivalence. Population bioequivalence (PBE) was introduced to evaluate the prescribability aspect of bioequivalence testing and is meaningful in the clinical setting where the two formulations are equally prescribable to a new patient. The PBE criterion is formulated as:^[21,22]

$$\frac{(\mu_T - \mu_R)^2 + \sigma_T^2 - \sigma_R^2}{\sigma_{\max}^2} \leq \vartheta_p \quad (1)$$

where $\sigma_{\max}^2 = \max\{\sigma_0^2, \sigma_R^2\}$ and ϑ_p and σ_0^2 are goalposts.

However, the equality of bioavailability distributions does not imply that the responses to the two formulations within the same individual patient will be sufficiently close. Individual bioequivalence (IBE) of two formulations means that a patient currently on the reference formulation can be switched to the new test formulation with assurance that the new drug will yield comparable safety and efficacy to that of the product for which it is being substituted. The following mixed-scaling criterion is recommended:^[21,22]

$$\frac{(\mu_T - \mu_R)^2 + \sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2}{\sigma_{W\max}^2} \leq \vartheta_i \quad (2)$$

where σ_{WT}^2 and σ_{WR}^2 are within-subject variances for the T and R formulations, respectively, σ_D^2 is the subject-by-formulation interaction variance component, $\sigma_{W\max}^2 = \max\{\sigma_{W0}^2, \sigma_{WR}^2\}$ and ϑ_i and σ_{W0}^2 are goalposts. Because the within-subject variance of each treatment cannot be separately estimated in a two-period, two-sequence crossover design with sequences {TR, RT}, a design in which subjects obtain repeats of T or R are needed. An example of such a *replicate* design is the one with sequences {TRTR, RTRT}; for others, see Ref. [23].

In addition to the comparison of means, both the PBE and IBE criteria compare variances between the test and reference formulations. While the PBE criterion focuses on total variances, the IBE criterion considers within-subject variability and subject-by-formulation interaction.

The general idea by Dragalin et al.^[8] is to use the KLD as a measure of discrepancy between the distributions of the test and reference formulations. The model for the response on T and R is a standard, linear mixed-effects model. The KLD permits a natural hierarchy of variation, allowing bioequivalence to be considered at various levels: within-subject, between-subject, and overall. There is a KLD for each level. Two formulations are declared bioequivalent if the upper bound of $(1 - \alpha)100\%$ confidence interval for the KLD is less than a given goalpost.

This new approach allows a much more elegant treatment of bioequivalence evaluation than the current ad hoc approaches. It overcomes many of the weaknesses of the conventional methods based on Eqs. 1 and 2 and addresses the conceptual issues with current criteria. In particular, the KLD 1) possesses the natural hierarchical property

that $IBE \Rightarrow PBE \Rightarrow ABE$; 2) satisfies the properties of a true distance; 3) is invariant to monotonic transformations of the data; 4) easily generalizes to the multivariate case where equivalence on more than one parameter (e.g., AUC, $C_{\max}$, and $T_{\max}$) is required; and 5) is applicable over a wide range of distributions of the response variable (e.g., those in the exponential family).

In the next section, the general setting is introduced based on a linear mixed model for the pharmacokinetic responses. The KLD criteria for PBE and IBE are presented. The merits of the KLD approach are provided and desirable features are given relative to improving the drawbacks suffered by the conventional criteria. The “Individual Bioequivalence” section focuses on IBE where a measure is proposed based on the KLD between the distributions of the two formulations, but this time at the individual level.

DESIRABLE FEATURES OF BIOEQUIVALENCE CRITERIA AND CONCEPTUAL ISSUES WITH CURRENT ONES

In bioequivalence studies, the following model for observations is commonly accepted. Let X_{ijk} be the k th response ($k = 1, 2, \dots$) for the j th subject on treatment formulation t ($t = T, R$) and

$$X_{ijk} = \mu_t + v_{ij} + \varepsilon_{ijk} \quad (3)$$

where v_{ij} is the (random) mean deviation from the population average μ_t of the t th formulation for the j th subject and ε_{ijk} is the (random) within-subject deviation from the individual subject's response on repeated administrations of the t th formulation. Conditional on the j th subject, v_{ij} and ε_{ijk} are independent with mean zero and variances $\text{Var}(v_{ij}) = \sigma_{Bt}^2$ (the between-subject variance) and $\text{Var}(\varepsilon_{ijk}) = \sigma_{Wt}^2$ (the within-subject variance). In addition, we assume $\text{Cov}(\varepsilon_{ijk}, \varepsilon_{ijk'}) = 0$ for $k \neq k'$. For the j th subject, v_{Tj} and v_{Rj} are correlated with correlation coefficient ρ and the subject-by-formulation interaction variance

$$\sigma_D^2 = \text{Var}(v_{Tj} - v_{Rj}) = \sigma_{BT}^2 + \sigma_{BR}^2 - 2\rho\sigma_{BT}\sigma_{BR}$$

The current Food and Drug Administration (FDA)-proposed criteria for the evaluation of ABE, PBE, and IBE are functions of the means, within-subject, and between-subject variances, and the subject-by-formulation interaction. In essence, each of these criteria is an ad hoc attempt to assess the discrepancy between distributions of the pharmacokinetic responses on test and reference formulations. However, different measures of such discrepancy are used for different types of bioequivalence: ABE is evaluated in terms of the difference in means of T and R; PBE is evaluated in terms of the second moment of the between-subject differences in bioavailability of the T and R formulations with that of R vs. R itself; while the

IBE is evaluated in terms of the second moment of the within-subject differences.

Assume that the response variables are distributed according to normal distributions with means μ_T and μ_R and variances σ_T^2 and σ_R^2 on test and reference formulations, respectively. In this case, the KLD is

$$d(f_T, f_R) = \frac{1}{2} \left\{ (\mu_T - \mu_R)^2 + \sigma_T^2 + \sigma_R^2 \right\} \times \left(\frac{1}{\sigma_T^2} + \frac{1}{\sigma_R^2} \right) - 2 \quad (4)$$

Particularly, in case of model (3), $\sigma_T^2 = \sigma_{BT}^2 + \sigma_{WT}^2$ and $\sigma_R^2 = \sigma_{BR}^2 + \sigma_{WR}^2$.

Under the assumption that the two formulations have the same variance, i.e., $\sigma_T^2 = \sigma_R^2 = \sigma^2$, the KLD becomes

$$d(f_T, f_R) = \frac{(\mu_T - \mu_R)^2}{\sigma^2} \quad (5)$$

which differs from the current (unscaled) ABE measure $(\mu_T - \mu_R)^2$.^[21,22]

We now consider various aspects of the FDA-proposed measures of bioequivalence and indicate advantages possessed by the KLD measure.

Desirable Properties for Distance

Let F and G be two distributions with densities f and g , respectively. The distance (discrepancy, divergence) between f and g is denoted by $\Delta(f, g)$. A true distance should satisfy

1. $\Delta(f, g) \geq 0$ and $\Delta(f, g) = 0$ iff $f = g$
2. $\Delta(f, g) \leq \Delta(f, q) + \Delta(q, g)$ (triangle inequality)
3. $\Delta(f, g) = \Delta(g, f)$ (symmetry)

where q is a third arbitrary density.

Kullback and Leibler^[24] introduced a measure of the “directed divergence” between statistical populations,

$$I(f, g) = E_f \log \frac{f(X)}{g(X)}$$

variously known as Kullback–Leibler information, information for discrimination, I divergence, the error, or the directed divergence. However, $I(f, g)$ is not symmetric; also, it does not satisfy the triangle inequality in general. Jeffreys^[25] considered the symmetric divergence or J divergence:

$$\Delta_1(f, g) = [I(f, g) + I(g, f)]^{1/2} = \left[\int (f(x) - g(x))(\log f(x) - \log g(x)) dx \right]^{1/2}$$

In the normal distributions case, $\Delta_1(f, g)$ satisfies all three distance properties when $E_f(X) = E_g(X)$ or $\text{Var}_f(X) = \text{Var}_g(X)$. Thus $\Delta_1(f, g)$ can be considered as a distance between distributions. However, it is more convenient to work with $d(f, g) = \Delta_1^2(f, g)$ and, although abusing the terminology, we will call the function $d(f, g)$ the Kullback–Leibler divergence.

It is clear from the definition of the KLD that $d(f, g) \geq 0$ with equality if and only if $f(x) = g(x)$ for all x . It is symmetric and convex in the pair (f, g) .

The present approach in bioequivalence trials^[21,22] is to use the following discrepancy measures:

$$\begin{aligned} \text{ABE} : \delta &= (\mu_T - \mu_R)^2 \\ \text{PBE} : \bar{d} &= \frac{(\mu_T - \mu_R)^2 + \sigma_T^2 - \sigma_R^2}{\sigma_R^2} \end{aligned}$$

The discrepancy measure $\bar{d}$ of the PBE criterion violates all three properties of a true distance: $\bar{d} = 0$ not only when $\mu_T = \mu_R$ and $\sigma_T^2 = \sigma_R^2$, but also for a test formulation with a different mean ($\delta > 0$) and $\sigma_T^2 = \sigma_R^2 - \delta$. Notice that δ might be big enough (thus the lack of average bioequivalence is quite evident), but with a small σ_T^2 (relative to σ_R^2), the test formulation can be declared equal to the reference formulation. Moreover, for reference formulations with relatively large variability σ_R^2 , the “distance” $\bar{d}$ between test and reference formulation might even be negative! Furthermore, PBE does not always imply ABE. The criterion for ABE, $(\mu_T - \mu_R)^2 \leq \theta_A$, requires different values for the “regulatory” bound θ_A for formulations with “small,” “moderate,” and “large” variability. On the other hand, the KLD bioequivalence measure $d(f_T, f_R)$ is in standardized units, and both ABE and PBE criteria are unified as

$$d(f_T, f_R) < d_0$$

Scaling and Symmetry

The current criteria for PBE and IBE provide goalposts dependent on variability of the reference formulation and allow scaling for both highly variable and narrow therapeutic range drugs, and reward for reduced variability in the test formulation. However, the scaling mainly corresponds to a modification of the bioequivalence acceptance limits and ought to be an issue independent of IBE, PBE, and ABE. The proposed by FDA mixed-scaling approach (with $\sigma_{\max}^2$ or $\sigma_{W\max}^2$) has a discontinuity at the change points (σ_0^2 or σ_{W0}^2) from constant to reference scaling; for example, if the estimate of σ_R^2 is just above σ_0^2 , the confidence interval will be wider than just below σ_0^2 .

Moreover, the scaling in both the PBE and IBE criteria is asymmetric with respect to T and R. One may argue that in bioequivalence studies, the reference and test formulation cannot be equally treated: It should be shown that the

test is equivalent to the reference, and not the other way around. However, for physicians and patients, bioequivalent drug formulations should be interchangeably used to achieve a similar therapeutic effect. Bioequivalence studies in fact serve as surrogates for clinical trials in evaluation of therapeutic equivalence in efficacy and safety between the innovator product and its generic copies. A new formulation that demonstrates much greater or much lower bioavailability is unacceptable. Furthermore, the bioequivalence evaluation of the two generic formulations may also be of interest. Therefore, in this sense, both formulations should be symmetrically treated. These issues are resolved by the new method based on the KLD.

Interpretation on the Original Scale

The current bioequivalence criteria are usually formulated on the logarithmic scale of the pharmacokinetic characteristic. The ABE for the population means μ_T and μ_R on the logarithmic scale [e.g., $\ln(\text{AUC})$ being normal] can be expressed as the ratio of the population medians on the original scale. However, the aggregate criteria for PBE and IBE on the logarithmic scale have no obvious translation into the natural parameter space. On the other hand, the KLD is invariant to a continuously differentiable monotonic transformation.^[8] This implies that if a bioequivalence property (ABE, PBE, or IBE) defined in terms of the KLD holds for a given data set, then the property remains valid for any monotone transformation of the data.

Design Requirements

An increased burden for the pharmaceutical industry may arise with use of the FDA individual bioequivalence criterion in that fully or partially replicated designs with at least three periods are required and should be performed in study subjects from the general population. Moreover, existing publications do not clearly indicate that the issue of switchability is properly addressed by the proposed IBE criterion.^[26] The KLD approach has the additional advantage of being able to use data from conventional two-way crossover designs (see details in “Design-Driven Individual Bioequivalence Criteria”).

Logical Hierarchy of Conclusions

Another important issue is the desired logical hierarchy of bioequivalence conclusions: IBE should imply PBE, which in turn should imply ABE. Unfortunately, current FDA approaches do not ensure this kind of hierarchy. Two formulations might satisfy the FDA IBE criterion, but fail to be declared PBE. Or, two formulations might satisfy the PBE criterion, i.e., have both similar means and variances, yet cannot be declared bioequivalent based on a comparison of means alone. On the other hand, the KLD-based criteria for ABE, PBE, and IBE possess this desirable property of

logical hierarchy of conclusions: $\text{IBE} \Rightarrow \text{PBE} \Rightarrow \text{ABE}$. (See details in “Individual Bioequivalence.”)

Generalization to Exponential Family of Distributions

The KLD approach may be used for any type (not only normal) of distributions. A similar formula for $d(f_T, f_R)$ can be derived for the general exponential family of distributions (in canonical form):

$$f_\theta(x) = \exp \left\{ \sum_{i=1}^k \theta_i T_i(x) - \psi(\theta) \right\} h(x)$$

where $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ is a k -dimensional parameter, ψ is a real-valued function of the parameters, and T_i are real-valued statistics. Using

$$E_\theta(T_i(X)) = \frac{\partial \psi(\theta)}{\partial \theta_i}$$

straightforward calculations yield

$$d(f_{\theta_T}, f_{\theta_R}) = \sum_{i=1}^k (\theta_{Ti} - \theta_{Ri}) \times \left[\frac{\partial \psi(\theta)}{\partial \theta_i} \Big|_{\theta=\theta_T} - \frac{\partial \psi(\theta)}{\partial \theta_i} \Big|_{\theta=\theta_R} \right]$$

Particular cases are:

- Kullback–Leibler divergence for two exponential distributions: $(\lambda_T - \lambda_R)((1/\lambda_R) - (1/\lambda_T))$.
- Kullback–Leibler divergence for two Bernoulli distributions: $(p_T - p_R) \log((p_T(1 - p_R))/((1 - p_T)p_R))$. [Notice the difference from other discrepancy measures for Bernoulli case such as difference in proportions $(p_T - p_R)$, risk ratio p_T/p_R , and odds ratio $(p_T(1 - p_R)/(1 - p_T)p_R)$.
- Kullback–Leibler divergence for two Poisson distributions: $(\lambda_T - \lambda_R) \log(\lambda_T/\lambda_R)$.

Generalization to Multivariate Case

In most bioequivalence studies (exceptions are Refs. [13,27–29]), different characteristics of the pharmacokinetics and pharmacodynamics, such as AUC, $C_{\max}$, $T_{\max}$, etc. are separately analyzed.

The approach based on the KLD can be easily extended to the multivariate situation when all these characteristics are simultaneously analyzed. Let $\mathbf{X}_{Rj}$ and $\mathbf{X}_{Tj}$ be observations from the reference and test formulations, but this time they are random vectors (of dimension $p \geq 2$) with components corresponding to various characteristics. Again, let f_R and f_T be the corresponding density functions. Then, it

can be verified that under the assumption that f_R and f_T are multivariate normal density functions with mean vectors μ_R and μ_T and positive definite covariance matrices Σ_R and Σ_T , respectively:

$$\begin{aligned} d(f_T, f_R) &= \frac{1}{2} \left\{ (\mu_T - \mu_R)^T (\Sigma_T^{-1} + \Sigma_R^{-1}) (\mu_T - \mu_R) \right. \\ &\quad \left. + \text{tr}(\Sigma_T^{-1} \Sigma_R + \Sigma_T \Sigma_R^{-1}) - 2p \right\} \\ &= \frac{1}{2} \text{tr} \left\{ [(\mu_T - \mu_R)(\mu_T - \mu_R)^T + \Sigma_T + \Sigma_R] \right. \\ &\quad \left. \times [\Sigma_T^{-1} + \Sigma_R^{-1}] \right\} - 2p \end{aligned}$$

where the superscript T to a vector means “transposed” and -1 to a matrix means it is inverse, tr means the trace of a matrix. Note the similarity with the definition of the KLD for PBE in the univariate case.

INDIVIDUAL BIOEQUIVALENCE

If, besides average bioavailability and prescribability, one is also interested in switchability of the test and reference formulations, an individual bioequivalence measure may be proposed that is also based on the KLD, but this time at the individual level.

Suppose that for each subject in the trial, a pair (X_{Rj}, X_{Tj}) of observations is available. Assuming that

$$\begin{aligned} X_{Rj} | v_{Rj} &\sim \mathbf{N}(\mu_R + v_{Rj}, \sigma_{WR}^2) \quad \text{and} \\ X_{Tj} | v_{Tj} &\sim \mathbf{N}(\mu_T + v_{Tj}, \sigma_{WT}^2) \end{aligned}$$

where $\mathbf{N}(\mu, \sigma^2)$ is normal distribution with mean μ and variance σ^2 , we can calculate as in Eq. 4 the conditional KLD between the distributions of X_{Rj} and X_{Tj} , conditioned on the j th subject-related characteristics (v_{Rj}, v_{Tj}) :

$$\begin{aligned} d(f_{T,v}, f_{R,v}) &= \frac{1}{2} \left\{ (\mu_T + v_{Tj} - \mu_R - v_{Rj})^2 + \sigma_{WT}^2 \right. \\ &\quad \left. + \sigma_{WR}^2 \right\} \left(\frac{1}{\sigma_{WT}^2} + \frac{1}{\sigma_{WR}^2} \right) - 2 \end{aligned}$$

Additionally, assuming that (v_{Rj}, v_{Tj}) has a bivariate normal distribution with zero means, we can calculate the unconditional KLD integrating out the effect of (v_{Rj}, v_{Tj}) :

$$\begin{aligned} d(f_T, f_R) &= \frac{1}{2} \left\{ (\mu_T - \mu_R)^2 + \sigma_D^2 \right. \\ &\quad \left. + \sigma_{WT}^2 + \sigma_{WR}^2 \right\} \times \left(\frac{1}{\sigma_{WT}^2} + \frac{1}{\sigma_{WR}^2} \right) - 2 \end{aligned} \quad (6)$$

where $\sigma_D^2 = \text{Var}(v_{Tj} - v_{Rj}) = \sigma_{BT}^2 + \sigma_{BR}^2 - 2\rho\sigma_{BT}\sigma_{BR}$.

We call this the KLD for IBE. Notice that this measure is based on the same parameters as the currently proposed

FDA measure (Eq. 2). However, the KLD for IBE fulfils the desired distance properties 1 and 3, while Eq. 2 fails on all three properties of a true distance.

There is a desired logical hierarchy of bioequivalence conclusions in the KLD setting: $\text{IBE} \Rightarrow \text{PBE} \Rightarrow \text{ABE}$. [Notice that this is not the case with the current criteria for PBE (Eq. 1) and IBE (Eq. 2).] Indeed, let $f_{R,v}(x)$ be the density function for X_{Rj} conditional on the subject-formulation effects (v_{Rj}, v_{Tj}) and similarly $f_{T,v}(x)$ be the conditional density function for X_{Tj} . Also, let $\varphi(v_R, v_T)$ be the joint density of the subject-formulation effects on the same subject. Then, the KLD for IBE can be defined as

$$d(f_T, f_R) = E_\varphi \left[E_{f_{R,v}} \log \frac{f_{R,v}}{f_{T,v}} + E_{f_{T,v}} \log \frac{f_{T,v}}{f_{R,v}} \right]$$

Because of the convexity in the pair (f_T, f_R) of $d(f_T, f_R)$, we have

$$\begin{aligned} d_I(f_T, f_R) &= E_\varphi \left[E_{f_{R,v}} \log \frac{f_{R,v}}{f_{T,v}} + E_{f_{T,v}} \log \frac{f_{T,v}}{f_{R,v}} \right] \\ &\geq E_{f_R} \log \frac{f_R}{f_T} + E_{f_T} \log \frac{f_T}{f_R} = d_P(f_T, f_R) \end{aligned}$$

where $f_i(x) = \int f_{i,v}(x) \varphi(v_R, v_T) dv_R dv_T$. Here we have used indices I and P to distinguish between KLDs for individual (Eq. 6) and population (Eq. 4) bioequivalence. Similarly, we can show the implication $\text{PBE} \Rightarrow \text{ABE}$ from Eqs. 4 and 5.

Therefore IBE is a stronger property than PBE, which in turn is a stronger property than ABE, and the KLD methodology offers the following testing of a priori ordered hypotheses in a stepwise manner for bioequivalence evaluation:

First: Average bioequivalence using criterion based on Eq. 5.
If shown: Population bioequivalence based on Eq. 4.
If also shown: Individual bioequivalence based on Eq. 6.

The generalization of KLD for IBE to the multivariate case is straightforward:

$$\begin{aligned} d(f_T, f_R) &= \frac{1}{2} \text{tr} \left\{ [(\mu_T - \mu_R)(\mu_T - \mu_R)^T + \Sigma_D \right. \\ &\quad \left. + \Sigma_{WT} + \Sigma_{WR}] [\Sigma_{WT}^{-1} + \Sigma_{WR}^{-1}] \right\} - 2p \end{aligned}$$

where Σ_{WR} and Σ_{WT} are covariance matrices for the error vectors ε_R and ε_T , respectively, and $\Sigma_D = \text{Cov}(v_T - v_R)$.

Design-Driven Individual Bioequivalence Criteria

For IBE evaluation using the IBE measure (Eq. 6), one needs at least a three-period crossover design as in the case

of current criteria (Eq. 2).^[20] However, the switchability criterion can also be formulated as follows:

the test formulation T is switchable with the reference R if the KLD between the bivariate distribution of (X_{Rj1}, X_{Tj2}) and the bivariate distribution of (X_{Rj1}, X_{Rj2}) is less than some goalpost d_i .

Under the assumptions that v_{ij} and ε_{ijk} are normally distributed, we obtain the following:

1. (X_{Rj1}, X_{Tj2}) have a bivariate normal distribution with mean $\mu_T = (\mu_R, \mu_T)$ and covariance matrix

$$\Sigma_T = \begin{pmatrix} \sigma_R^2 & \rho_T \sigma_T \sigma_R \\ \rho_T \sigma_T \sigma_R & \sigma_T^2 \end{pmatrix}$$

2. (X_{Rj1}, X_{Rj2}) have a bivariate normal distribution with mean $\mu_R = (\mu_R, \mu_R)$ and covariance matrix

$$\Sigma_R = \begin{pmatrix} \sigma_R^2 & \rho_R \sigma_R^2 \\ \rho_R \sigma_R^2 & \sigma_R^2 \end{pmatrix}$$

where $\sigma_T^2 = \sigma_{BT}^2 + \sigma_{WT}^2$, $\sigma_R^2 = \sigma_{BR}^2 + \sigma_{WR}^2$, $\rho_T \sigma_T \sigma_R / (\sigma_T \sigma_R)$, and $\rho_R = \sigma_{BR}^2 / \sigma_R^2$.

As in the PBE multivariate case, we obtain:

$$d(f_T, f_R) = \frac{1}{2} \left\{ (\mu_T - \mu_R)^T (\Sigma_T^{-1} + \Sigma_R^{-1}) (\mu_T - \mu_R) + \text{tr}(\Sigma_T^{-1} \Sigma_R + \Sigma_T \Sigma_R^{-1}) \right\} - 4 \quad (7)$$

Note that $d(f_T, f_R) = 0$ iff $\mu_T = \mu_R$, $\sigma_T = \sigma_R$, and $\rho_T = \rho_R$. This is another definition of the KLD for IBE driven by both the definition of the switchability and the design used for its assessment. Notice that the two-period, two-sequence, two-formulation design can be used for evaluating individual bioequivalence using the KLD (Eq. 7), while the IBE criteria (Eqs. 2 and 6) require at least three-period replicate designs.^[20]

Any replicate crossover design with at least two realizations within a patient and within a treatment of the forms

1. three-period, two-sequence, two-formulation design {TRT, RTR},
2. four-period, two-sequence, two-formulation design {TRTR, RTRT},

etc. may also be used for this kind of individual bioequivalence trials. Each such design implies a definition of switchability and the respective KLD can be constructed.

To assure drug switchability, it is generally suggested that bioequivalence be assessed within the same individuals. In the simplest case, the observations in such IBE

studies are matched pairs (X_{Tj}, X_{Rj}) . Because the concept of “switchability” actually means “agreement” between the bioavailabilities of the two formulations, the statistical procedures used in agreement problems^[30] are adequate in IBE assessment. The Total Least Square method was proposed in Dragalin and Fedorov^[31] as an alternative methodology for IBE evaluation. The proposed approach is very intuitive and transparent and can be easily extended to allow the parametric comparison of PK curves of the two formulations, not just its parameters. (See also Refs. [32,33].)

OTHER MEASURES OF DISCREPANCY

There are other distances that are functionals of the entire distribution, not only of their first (one or two) moments. The simplest one is the total variation

$$\Delta_2(f, g) = \int_{-\infty}^{\infty} |f(x) - g(x)| dx$$

which is the area between the two density functions. This distance is generalized to the so-called L_p distance. The limiting $p \rightarrow \infty$ case of the L_p distance is the ℓ_1 distance. Czado and Munk^[34,35] considered trimmed Mallow distance for bioequivalence evaluation. In general, for two continuous cdf's, F and G , the trimmed Mallow distance is defined as

$$\Delta_3(F, G; \alpha) = \left\{ \frac{1}{1-2\alpha} \int_{\alpha}^{1-\alpha} |F^{-1}(u) - G^{-1}(u)|^2 du \right\}^{1/2}, \quad \alpha \in [0, 1/2)$$

For the normal case,

$$\Delta_3(F, G; \alpha) = \left\{ \frac{1}{1-2\alpha} \left[(\mu_T - \mu_R)^2 + (\sigma_T - \sigma_R)^2 \times \left(1 - 2 \frac{z_{1-\alpha} \varphi(z_{1-\alpha})}{1-2\alpha} \right) \right] \right\}^{1/2}$$

where $z_{1-\alpha}$ and $\varphi(z_{1-\alpha})$ is the $(1 - \alpha)$ quantile and density, respectively, of the standard normal distribution. The trimming puts slightly more weight on the mean difference than on difference of the scales.

Another well-known discrepancy measure between two distributions is the Hellinger distance:

$$\Delta_4(f, g) = \left\{ \int \left[\sqrt{f(x)} - \sqrt{g(x)} \right]^2 dx \right\}^{1/2}$$

For the normal case,

$$\Delta_4(f, g) = \left\{ 2 \left(1 - \sqrt{\frac{2\sigma}{1+\sigma^2}} \exp \left\{ -\frac{\delta^2}{4(1+\sigma^2)} \right\} \right) \right\}^{1/2}$$

where $\delta = (\mu_T - \mu_R)/\sigma_R$ is the standardized (in standard deviation units of the reference R) difference in means and $\sigma = \sigma_T/\sigma_R$ is the ratio in test and reference intrasubject variability. Notice that the Hellinger distance, as well as the KLD for PBE (Eq. 4), depends only on δ and σ . This means that the goalpost for each of these two discrepancy measures is independent of the variability of the reference drug. In contrast, the goalpost for PBE based on the FDA measure or the trimmed Mallow distance depends on σ_R . Fig. 1 illustrates the contour plots for the four measures for PBE: KLD, FDA, trimmed Mallow, and Hellinger. For fair comparison, the FDA and trimmed Mallow measures are plotted assuming $\sigma_R = 1$. On the x axis it is the standardized difference in means δ , while on y axis it is the ratio in test and reference intrasubject variability σ . One can immediately see the drawbacks of the FDA PBE: An increase of one unit in δ can be offset by a tenfold decrease in intrasubject variability, $\sigma = 0.1$. The other three measures do not allow any mean–variability tradeoff. Also, notice the difference between an aggregate criterion such as KLD and a disaggregate one,^[7] in which standardized difference in means δ and intrasubject variability σ are individually evaluated. In the latter case, the parameter space corresponding to PBE is a rectangle, while in the former it is an ellipse centered at (0,1).

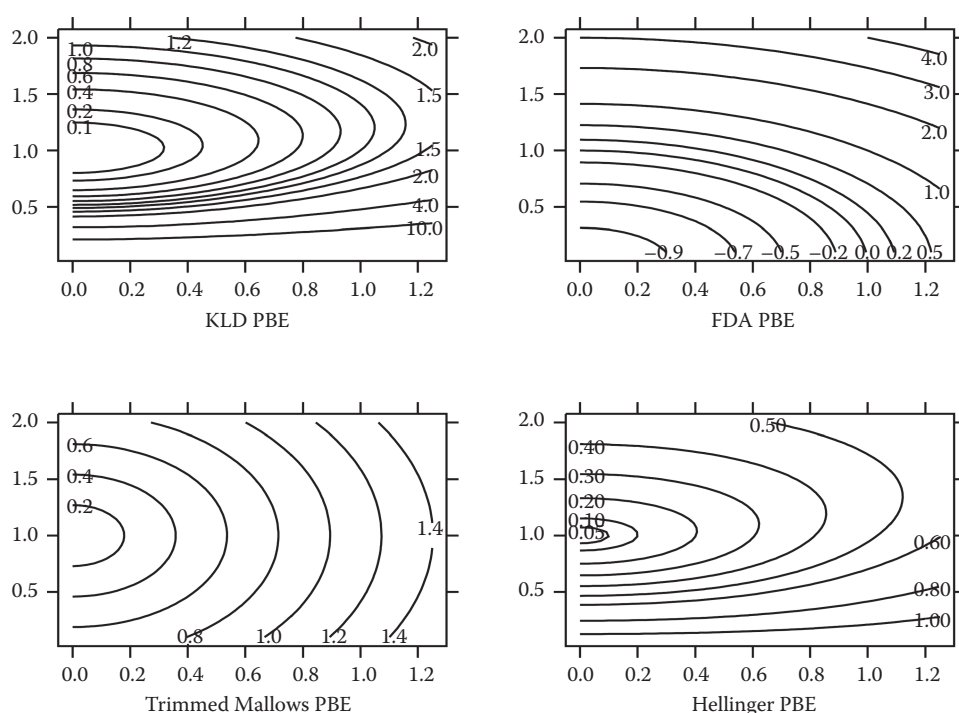


Fig. 1 Contour plots for different discrepancy measures in PBE

SIMULATIONS

Dragalin et al.^[8] present a simulation study for comparing the FDA and KLD metrics in assessing IBE. They consider 16 scenarios for different configurations of the mean difference $\mu_T - \mu_R$, high- and low-variability drugs σ_{WR}^2 , ratio of variances $\sigma_{WT}^2/\sigma_{WR}^2$, and subject-by-formulation interaction σ_D^2 .

The design used in the simulations has 18 subjects on each of the two sequences TRTR and RTRT. The normally distributed data generated in each simulation is based on Eq. 3. Given a particular scenario, the true values of the relevant means and variances were specified and a sample of 144 observations (four observations per subject) was generated.

These simulated data were then used to determine an upper 95% bound for each of the FDA and KLD metrics by bootstrap sampling (2000 bootstrap samples per simulated data set). In each bootstrap sample (performed preserving the number of subjects per sequence^[21]), observations were analyzed using a two-stage (mixed effect, restricted maximum likelihood) linear model including terms for sequence, period, formulation, and period by sequence-within-formulation interaction in the model. Subject within sequence was specified as a random effect, and a heteroscedastic compound symmetric matrix for between-subject variances was assumed in accordance with the original recommendations from the FDA^[21] and from previous research.^[36] Estimates for the parameters of interest were derived and used to estimate the metrics of interest in each bootstrap sample.

The KLD appeared directly sensitive to changes in the means between formulations and the presence of a subject-by-formulation interaction—directly measuring discrepancy between formulations as expected. The FDA metric's behavior appeared insensitive to some factors, consistent with earlier findings,^[36] and conclusions of bioequivalence may be overly dependent on the level of within-subject variation.

To assess how the metrics perform in practice, the AUC and $C_{\max}$ data of 22 data sets from 15 randomized, replicate-design studies were also analyzed. Descriptions of the data sets and estimates derived using the two-stage mixed effect model are given in Ref. [36].

The KLD metric appeared to be a more stringent criterion (i.e., harder to demonstrate IBE) than the FDA metric. The FDA metric by chance finds a slight increase in within-subject test variance unacceptable in the coincident presence of a small subject-by-formulation interaction and in the presence of changes in means between formulations. The KLD assesses these slight changes relative to the overall background noise (as measured by the within-subject variances for test and reference formulations).

The KLD metric rejected IBE for a greater number of data sets than the FDA metric. The problems with inference (observed mean/variance tradeoffs) previously identified for the FDA metric did not appear as prevalent with inference using the KLD, and overall, inference using the KLD appeared more consistent with a procedure that would be expected to be an extension of average bioequivalence.

The FDA Guidance (2001)^[22] recommends the modified large sample (MLS) confidence intervals^[37] for assessment of PBE and IBE. The method proposed in Ref. [37] requires independence of estimators of variance components present in the linear combination defining the FDA PBE and IBE measures. However, under crossover designs usually used in bioequivalence evaluation, the estimators of these variance components are dependent. Using a chi-square representation of a quadratic form of a multivariate normal vector, Lee et al.^[38] extend the MLS method to the situation where estimators of variance components are dependent. It will be interesting to use this new methodology in comparing the KLD with other metrics.

CONCLUSION

The KLD metric is a viable alternative to that proposed by the FDA for the assessment of bioequivalence. It appears to be sensitive to changes in mean response between formulations and to the detection of subject-by-formulation interaction. Identification of differences between individual responses to formulation are clearly identified in low-variability drug products, and bigger quantitative differences must be observed before the KLD flags them as being important for high-variability products.

Formulations are not “rewarded” for being less variable when using the KLD metric while they frequently are when using the FDA metric.^[36] Thus coincident differences in average bioavailability with decreased variance for the test formulation may allow market access under the proposed FDA metric,^[26] and this does happen in practice.^[36] Use of the KLD metric, which does not allow for such a “tradeoff,” is a desirable property in a global marketplace where regulators cannot guarantee that patients will not be switched back and forth from the reference and test products on a daily, weekly, or monthly basis when their prescription is filled at the pharmacist, and formulations are switched to the least-expensive product available in stock. International regulators cannot at present ensure that all patients will be switched to the new formulation consistently, immediately, and permanently. Use of metrics such as that proposed by the FDA may place patient populations at increased risk of side effects or ineffective switching because of marketplace forces beyond regulatory control.

Another benefit of the KLD metric is that it does not necessarily require the use of a three- or four-period replicate design to assess individual bioequivalence. The KLD metric, as a measure of individual bioequivalence, seems to identify more data sets as being of concern than the proposed FDA metric. This appears to be because of the ability of the FDA metric to reward novel formulations for which the within-subject variance is decreased relative to the reference product and the use of the mixed scaling procedure. While the KLD metric may be viewed as being a more “stringent” criterion than the FDA metric, it appears to identify those data sets where changes in formulation are of potential concern.

In conclusion, the results of simulations and retrospective analyses to date indicate that the KLD metric is a potential alternative to the FDA-proposed metric, offering some technical and interpretive advantages and preparing the ground for a reasonable and mathematically tractable methodology for bioequivalence evaluation.

REFERENCES

1. Rogers, J.L.; Howard, K.I.; Vessey, J.T. Using significance tests to evaluate equivalence between experimental groups. *Psychol. Bull.* **1993**, *113*, 553–565.
2. Roy, T. Calibrated nonparametric confidence sets. *J. Math. Chem.* **1997**, *21*, 103–109.
3. McBride, G. Equivalence tests can enhance environmental science and management. *Aust. N. Z. J. Stat.* **1998**, *41*, 19–29.
4. International Conference on Harmonization. *Guidance on Ethnic Factors in the Acceptability of Foreign Clinical Data*, ICH Secretariat (c/o IFPMA): Geneva, 1998. Web site: www.fda.gov/cder/guidance/guidance.htm.
5. Cheuvart, B.; Bollaerts, A. Sample size considerations for assessing vaccine consistency through equivalence. *Drug Inf. J.* **1999**, *33*, 149–151.

6. Munk, A.; Hwang, J.T.G.; Brown, L.D. Testing average equivalence—Finding a compromise between theory and practice. *Biom. J.* **2000**, *5*, 531–552.
7. Chow, S.-C.; Liu, J.-P. *Design and Analysis of Bioavailability and Bioequivalence Studies*, 2nd Ed.; Dekker: New York, 2000, 421.
8. Dragalin, V.; Fedorov, V.; Patterson, S.; Jones, B. Kullback–Leibler divergence for evaluating bioequivalence. *Stat. Med.* **2003**, *22*, 913–930.
9. FDA. *Guidance on Statistical Procedures for Bioequivalence Studies Using a Standard Two-Treatment Crossover Design*; Food and Drug Administration: Rockville, 1992.
10. Westlake, W.J. Symmetric confidence intervals for bioequivalence trials. *Biometrics*. **1976**, *32*, 741–744.
11. Anderson, S.; Hauck, W.W. A new procedure for testing equivalence in bioavailability and other clinical trials. *Commun. Stat., Theory Methods* **1983**, *12*, 2663–2692.
12. Schuirmann, D.J. On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics* **1981**, *37*, 1–617.
13. Berger, R.L.; Hsu, J.C. Bioequivalence trials, intersection–union tests and equivalence confidence sets (with discussion). *Stat. Sci.* **1996**, *11*, 283–319.
14. Bauer, P.; Kieser, M. A unifying approach for confidence intervals and testing of equivalence and difference. *Biometrika*. **1996**, *83*, 934–937.
15. Westlake, W.J. Response to T.B.L. Kirkwood: Bioequivalence testing—A need to rethink. *Biometrics* **1981**, *37*, 589–594.
16. Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* **1987**, *15*, 657–680.
17. Anderson, S.; Hauck, W.W. Consideration of individual bioequivalence. *J. Pharmacokinet. Biopharm.* **1990**, *18*, 259–273.
18. Anderson, S. Individual bioequivalence: A problem of switchability (with discussion). *Biopharm. Rep.* **1993**, *2*, 1–11.
19. Schall, R.; Luus, G.H. On population and individual bioequivalence. *Stat. Med.* **1993**, *12*, 1009–1124.
20. Wang, W. On testing of individual bioequivalence. *J. Am. Stat. Assoc.* **1999**, *94*, 880–887.
21. FDA, *Draft Guidance on In Vivo Bioequivalence Studies on Population and Individual Bioequivalence*; Food and Drug Administration: Rockville, MD, 1997.
22. FDA, *Guidance on Statistical Approaches to Establishing Bioequivalence*; Food and Drug Administration: Rockville, MD, 2001.
23. Hsuan, F.C.; Reeve, R. Assessing individual bioequivalence with high-order cross-over designs: A unified procedure. *Stat. Med.* **2003**, *22*, 2847–2860.
24. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
25. Jeffreys, H. An invariant form for the prior probability in estimation pro. *Proc. R. Soc. Lond., A* **1946**, *186*, 453–461.
26. Barrett, J.S.; Batra, V.; Chow, A.; Cook, J.; Gould, A.L.; Heller, A.; Lo, M.W.; Patterson, S.D.; Smith, B.P.; Stritar, J.A.; Vega, J.M.; Zariffa, N. PhRMA perspective on population and individual bioequivalence. *J. Clin. Pharmacol.* **2000**, *40*, 561–572.
27. Westlake, W.J. Bioavailability and Bioequivalence of Pharmaceutical Formulations. In *Biopharmaceutical Statistics for Drug Development*; Peace, K.E., Ed.; Dekker: New York, 1988, 329–352.
28. Hauck, W.W.; Hyslop, T.F.; Anderson, S.; Bois, F.Y.; Tozer, T.N. Statistical and regulatory considerations for multiple measures in bioequivalence testing. *Clin. Res. Regul. Aff.* **1995**, *12*, 249–265.
29. Wang, W.; Hwang, J.T.G.; Dasgupta, A. Statistical tests for multivariate bioequivalence. *Biometrika* **1999**, *86*, 395–402.
30. Cheng, C.L.; van Ness, J.W. *Statistical Regression with Measurement Error*; Kendall's Library of Statistics. Arnold: London, 1999.
31. Dragalin, V.; Fedorov, V. The total least squares method in individual bioequivalence evaluation. *Biom. J.* **2001**, *43*, 399–420.
32. Tan, C.Y.; Iglewicz, B. Measurement-methods comparisons and linear statistical relationship. *Technometrics* **1999**, *41*, 192–201.
33. Carrasco, J.-L.; Jover, L. Assessing individual bioequivalence using the structural equation model. *Stat. Med.* **2003**, *22*, 901–912.
34. Munk, A.; Czado, C. Nonparametric validation of similar distributions and assessment of goodness of fit. *J. R. Stat. Soc., B.* **1998**, *60*, 223–241.
35. Czado, C.; Munk, A. Assessing the similarity of distributions—Finite sample performance of the empirical Mallows distance. *J. Stat. Comput. Simul* **1998**, *60*, 319–346.
36. Zariffa, N.M.-D.; Patterson, S.D.; Boyle, D.; Hyneck, H. Case studies, practical issues, and observations on population and individual bioequivalence. *Stat. Med.* **2000**, *19*, 2811–2820.
37. Hyslop, T.; Hsuan, F.; Holder, D.J. A small sample confidence interval approach to assess individual bioequivalence. *Stat. Med.* **2000**, *19*, 2885–2897.
38. Lee, Y.; Shao, J.; Chow, S.-C. The modified large sample confidence intervals for linear combinations of variance components: Extension, theory, and application. *J. Am. Stat. Assoc.* **2004**, *99* (466), 467–478.

Laboratory Analyses

Michael J. Klepper

Integrated Safety Systems, Inc., Cary, North Carolina, U.S.A.

Abstract

In clinical trials, laboratory analyses may be used to determine the efficacy of a biopharmaceutical product. This entry describes the most commonly used methods of laboratory analysis as well as some of the scientific issues associated with this type of analysis.

INTRODUCTION

In clinical trials, laboratory analyses may be used to determine the efficacy of a biopharmaceutical product such as in trials evaluating the effectiveness of lipid-lowering drugs where changes in cholesterol levels would be a primary efficacy endpoint of the study. However, analyses of laboratory changes are more often used in the evaluation of biopharmaceutical product safety. Changes in laboratory values observed during clinical trials may signal organ dysfunction/toxicity, hypersensitivity reactions, or important physiological and metabolic changes. Such changes may be due to the investigational drug under study, be a result of an underlying disease process unrelated to the investigational drug, be due to a concomitant medication taken by a patient participating in a clinical trial, may represent a drug interaction, or may be spurious. Moreover, large laboratory changes seen in an individual or a small group of patients may have greater clinical significance than small changes affecting a larger number of patients taking the investigational product. A variety of analyses is used in evaluating laboratory changes, and the results from each analysis give different insights into such changes. In dealing with changes in laboratory values, some analyses may reveal statistically significant changes that are not clinically significant, and conversely clinically relevant changes may not achieve statistical significance. Observed changes in laboratory values during a clinical trial may signal that such changes may be due to a biopharmaceutical product under investigation; however, clinical correlation is important to determine whether this signal is valid and if so whether the change is clinically important.

METHODS

Laboratory Analytes Measured

In clinical trials, laboratory parameters that are collected and analyzed fall into three main categories: hematology, chemistry, and urine.

[Table 1](#) summarizes by category the laboratory parameters that are routinely measured in clinical trials. Depending on the biopharmaceutical product being studied, and the known or theoretical properties and side effects of the product, other laboratory parameters may also be collected. For example, 3-hydroxy-3-methylglutaryl coenzyme A (HMG CoA) reductase inhibitors, also known as statins, are effective agents in lowering cholesterol, low-density lipoprotein (LDL), and very low density lipoprotein (VDRL). These drugs may also damage skeletal muscle, which is often reflected by an increase in creatinine phosphokinase (CPK), a muscle enzyme.^[1] For these reasons, cholesterol, LDL, VDRL, and CPK levels would routinely be measured in clinical trials dealing with statins.

Changes in hematological parameters usually reflect effects on the bone marrow, excessive blood loss such as anemia from bleeding, or increased rate of destruction based on hypersensitivity reactions. Chemistry parameters reflect a variety of different organ functions and other physiological changes including liver and kidney function as well as metabolic changes. Changes in the urine may be specific to the urinary system, e.g., the presence of WBC casts seen in renal disease, or be reflective of a disease process unrelated to the kidney, e.g., the presence of glucose and ketones in the urine in a patient with uncontrolled diabetes. Often a disease will manifest itself with changes seen in more than one laboratory category; for example, a patient with renal failure may be anemic, have elevations of BUN, creatinine, and serum potassium, and have an abnormal urine analysis. Laboratory parameters can be organ-specific or nonspecific; for example, ALT is an enzyme with greatest specificity for the liver, whereas AST is an enzyme that is found in skeletal and cardiac muscle and the liver. Standard laboratory and medical texts should be referred to for more detailed discussions regarding the clinical implications and significance of laboratory test changes.

Table 1 List of Laboratory Analytes by Category

Hematology	Hemoglobin (HBG)
	Hematocrit (HCT)
	Red blood cell (RBC) count
	White blood cell (WBC) count
	Differential WBC count
	Neutrophils
	Lymphocytes
	Eosinophils
	Basophils
	Monocytes
	Platelet count
Chemistry	Alanine aminotransferase (ALT, SGPT)
	Aspartate aminotransferase (AST, SGOT)
	Alkaline phosphatase (ALP)
	Glutamyl transferase (GGT)
	Lactate dehydrogenase (LDH)
	Total bilirubin
	Blood urea nitrogen (BUN)
	Creatinine
	Albumin
	Total protein
	Sodium
	Potassium
	Chloride
	Calcium
	Phosphorus
	Uric acid
Urine	Glucose
	Cholesterol
	Triglycerides
	pH
	Specific gravity
	Presence of glucose, ketones, RBCs, WBCs, bacteria, casts, and crystals

Grouping of Laboratory Analytes

Because a number of different laboratory analytes may reflect similar organ or physiological dysfunction, certain related laboratory analytes when grouped together may aid in the review of laboratory changes. For example, both hemoglobin and hematocrit are related to red blood cells and should be displayed next to each other. Liver abnormalities might show changes in ALT, AST, bilirubin, GGT, LDH, and ALP and therefore these analytes should be displayed together. Renal function often affects both BUN and creatinine and, therefore, these values should also be grouped

together. Although arbitrary, the analytes listed in [Table 1](#) have been grouped in ways that facilitate medical review.

Reference Range of Laboratory Analytes

Each laboratory establishes its reference range for each analyte by obtaining laboratory measurements of subjects believed to be healthy and “normal.” It is assumed that the laboratory values of the healthy and normal population follow a Gaussian distribution and the mean and standard deviation (SD) of each analyte are calculated. Consequently, 95% of the healthy and “normal” population is expected to have values that fall within ± 2 SD of the mean.^[2] For some analytes, other factors such as age and gender affect the reference range. For example, females have a lower reference range for red blood cells, hemoglobin, and hematocrit due to blood loss during menstruation and gender differences in lean body mass. Age affects the reference ranges for the enzyme alkaline phosphatase. This enzyme can be found in a number of tissues especially bone and liver. Because of bone growth reflected by an increase in alkaline phosphatase values, the reference range for this analyte in children is predictably higher than the reference range in adults.

Units of Laboratory Analytes

The International System of Units (Système International Units, SI), which is based on the metric system^[3] is likely to become the global standard. Until this happens, laboratory values collected in clinical trials may report values in conventional (units still used in some United States laboratories) or SI units. In order to pool data across sites and countries such data need to be converted to the laboratory units selected for analysis (see the “Scientific Issues” section).

Types of Laboratory Analyses

In clinical trials, any change from baseline may represent an effect due to the biopharmaceutical product under investigation. The laboratory analyses that are utilized compare the value(s) obtained during the treatment period to those obtained at baseline, with baseline values usually defined as the last measurement before a subject received treatment of the investigational product, placebo, or comparator agent. The International Conference on Harmonization (ICH) and the Food and Drug Administration (FDA) have issued guidances regarding laboratory analyses for clinical studies^[4] and integrated summaries of safety for product registration.^[5–7] The types of analyses utilized for laboratory data obtained from clinical trials generally fall into four main types. The first type includes descriptive statistics such as mean changes, mean percentage changes, and median changes and analyses, which measure central tendency. Another type of analysis examines the number

of patients who have shifts in laboratory values from one category at baseline to a different category postbaseline, i.e., shift tables. Of particular clinical importance are the analyses that evaluate the number of patients with clinically significant laboratory changes, i.e., outliers. The fourth type of analysis examines the incidence of laboratory-related adverse events.

Means, Mean Percentage Change, and Medians

Mean change comparing postbaseline values to baseline values is one of the established methodologies employed in the analysis of laboratory data. Because such changes may be very small, mean percentage change from baseline may be used instead. Standard deviations are also calculated to determine the variability of the data. Means are used for data that are normally distributed otherwise the median is used.^[8] For comparative clinical trials, the difference in change from baseline between treatment groups is compared and a determination is made as to whether or not the difference is statistically significant, i.e., *p* values of ≤ 0.05 . For studies with no comparative group, changes from baseline are described.

Shift Tables

Two types of shift tables can be utilized in assessing categorical changes. The first type of shift table evaluates the number (%) of subjects with a laboratory value below (L) normal or normal (N) at baseline that shifts to a value above normal (H) postbaseline. The number (%) of subjects with above (H) normal or normal (N) values at baseline who had postbaseline values below (L) normal is also calculated. Postbaseline values that remain within the same baseline category, indicating no change, and abnormal baseline values that become normal during the treatment period are usually not included in the shift table analysis in order to highlight the more clinically important changes. An example of this type of shift table is shown in Table 2.

Categorical analyses can also be used to examine the number of subjects in each treatment group with baseline values that reached a certain prespecified change, e.g., the number (%) of subjects with creatinine values that increased 25% from baseline.

Table 2 Example of a Shift Table—Investigational Product

Baseline	On-treatment (<i>N</i> = 100)		
	Low	Normal	High
Low	X	X	4 (4%)
Normal	2 (2%)	X	10 (10%)
High	5 (5%)	X	X

X = shifts that are not clinically meaningful.

Analysis of Outliers

It is also important from a clinical perspective to look at laboratory values that are not only abnormal but fall into the category of clinically significantly abnormal values, i.e., outlier values. Such values may indicate a serious clinical condition that requires further evaluation and treatment. One way of looking at such outliers is by constructing scatterplots. Posttreatment values that indicate no change from baseline values would fall on a 45° line. Values that fall way above or below this line would indicate outlier values. Once identified, such values would have to be explored further. Another method is to use the categorical analysis discussed above to analyze the number (%) of subjects who develop clinically significant values postbaseline. For example, a clinically significant abnormal value for platelets is defined as $\leq 75 \times 10^3/\text{mm}^3$,^[6] the reference range varies depending on what laboratory is used but is generally in the range of $(150 \text{ to } 350) \times 10^3/\text{mm}^3$. The risk of developing spontaneous bleeding episodes that may become life-threatening increases as the platelet count decreases. From a clinical standpoint, it is important to determine how many subjects develop clinically significant decreases in platelet counts. This type of analysis summarizes the number (%) of subjects with one or more clinically significant values for a given analyte. Because change from baseline like the other laboratory analyses discussed above should be the primary focus of the analysis, it is recommended that only those subjects who had normal values at baseline and had one or more clinically significant values during the treatment period be included in the analysis. The Neuropharmacology Division of the FDA has established clinically significant threshold values for some laboratory analytes^[6] that can be used for this type of analysis. Subjects who had an abnormal baseline value could also be included in such an analysis if a clinically significant change for the baseline abnormal value is defined for that analyte, e.g., any value that was 50% greater than the abnormal baseline value. Establishing clinically significant thresholds for those laboratory values abnormal at baseline is usually left to medical judgment and the thresholds chosen should be clearly documented. In the categorical analysis of clinically significant changes, a subject with more than one clinically significant (CS) value for a given analyte is counted only once, but all values for this subject should be shown in a laboratory listing as described below. For trials that include placebo or a comparator control group, the number (%) of subjects in the investigational product group would be compared to the number (%) of subjects who received placebo or a comparator agent. Because the number (%) of such subjects in any given treatment group is expected to be low unless the product under investigation is toxic, testing for statistical significance is usually not appropriate and could be misleading because a low incidence is unlikely to show statistically significant differences. A small difference between

treatment groups, however, may still be clinically important. Table 3 is an example of clinically significant changes in two related analytes AST and ALT. These results would indicate that there may be both a drug and dose effect of the investigational product compared to placebo. Because the changes are seen for both AST and ALT, these findings may be indicative of a liver-related problem.

A listing of all subjects with clinically significant laboratory values is recommended to determine the pattern and clinical relevance of the laboratory change and to provide insight as to whether the change may be due to the investigational product. To illustrate, Table 4 provides three different patterns of clinically significant platelet values. Example A indicates a clinically significant low platelet count at baseline, which indicates a preexisting condition with no clinically relevant change observed during the study suggesting there was no adverse product effect on this variable. Example B shows a normal, baseline platelet value. With each subsequent treatment visit, a successive decrease in platelet count is seen with the value at the final treatment visit exceeding the threshold for a clinically significant low platelet count. The posttreatment value shows the platelet count, although still below normal limits, is improving and approaching normal. Such a pattern as shown in this example would be highly suggestive that the change in the platelet count was due to the investigational product unless there was a plausible alternative explanation, i.e., the patient started taking a concomitant medication known to be toxic to platelets. A spurious abnormally low platelet count is suggested in Example C. By these examples one can see the importance of evaluating listings for all subjects with clinically significant laboratory values. Such listings should include the study protocol/study site number, subject number, the subject's age and gender, and all values obtained throughout the study for that subject for that particular laboratory analyte, including the baseline, treatment, and any posttreatment values (if available). Each

visit should also include the date each value was obtained and the number of days relative to baseline that the value was obtained. All abnormal values should be appropriately flagged as low (L), high (H), and clinically significant (CS).

Laboratory-Related Adverse Events

To ensure a comprehensive review of laboratory data it is also important to examine the incidence of laboratory-related adverse events to determine if there are similar findings found in other laboratory analyses. For instance, if the incidence of the adverse event “hypokalemia” (low potassium) is either statistically or clinically significantly greater in the investigational product group compared to placebo, it is important to determine if similar trends were seen in the other laboratory analyses performed. Such trends might include a mean decrease in potassium, a higher percentage of subjects in the investigational product group with decreased shifts in potassium compared to placebo, or a higher percentage of subjects with clinically significantly low potassium values observed in the investigational group compared to placebo. Further clinical weight is given for those laboratory-related adverse events that were considered to be related to treatment and were serious^[9] and/or resulted in discontinuation from treatment.

SCIENTIFIC ISSUES

The results of any laboratory analysis is dependent on the correctness of the laboratory data analyzed, understanding the limitations of the laboratory analysis employed, and determining the clinical relevance of the findings. Factors affecting the poolability, i.e., integration of data across investigative sites and across studies must also be considered. These issues are discussed next.

Table 3 Summary of Subjects with Clinically Significant Shifts from Baseline

	Dose 1 (n = 50)	Dose 2 (n = 50)	All doses (n = 100)	Placebo (n = 50)
AST $\geq 3 \times$ Upper limit of reference range (10–40 U/L) Number (%)	3 (6%)	10 (20%)	13 (13%)	4 (8%)
ALT $\geq 3 \times$ Upper limit of reference range (5–35 U/L) Number (%)	2 (4%)	13 (26%)	15 (15%)	3 (6%)

Table 4 Examples of Clinically Significant Values that Suggest Different Etiologies

Visit	Example A ($\times 10^3/\text{mm}^3$)	Example B ($\times 10^3/\text{mm}^3$)	Example C ($\times 10^3/\text{mm}^3$)
Baseline	70 (L, CS)	175	175
Treatment week 4	110 (L)	150	150
Treatment week 8	65 (L, CS)	100 (L)	55 (L, CS) Repeat = 155
Treatment week 12 (final visit)	75 (L, CS)	70 (L, CS)	160
Follow-up visit (posttreatment)	None	140 (L)	165

L = below lower limit of reference range [(150–350) $\times 10^3/\text{mm}^3$]; CS = clinically significant.

Laboratory Measurement Errors

It is important to be aware of potential errors that may occur in obtaining laboratory values. Such errors may occur in the collection, processing, transporting, and storage of samples. These include the time of day the sample was obtained (e.g., adrenal hormone production is affected by circadian rhythms); whether the specimen was a fasting or nonfasting sample (relevant for glucose and lipid determinations); venipuncture technique; whether the proper collection tube was used; whether the sample tube was labeled with the correct subject's name; and how the specimen was transported and stored. Other factors include the precision and accuracy of the method used to analyze the sample and whether the sample was measured manually or automatically.^[2] Correct calculation, transcription, data entry, and data conversion are also important factors in assuring accurate laboratory values.

Data-Conversion Issues

When obtaining laboratory values it is important to understand that laboratory reference ranges may change during the course of a clinical trial. This must be taken into account when the data are analyzed because this will impact the determination of normal and abnormal values and may change the numerical threshold for clinically significant values. For instance the Neuropharmacology Division of the FDA defines ALT values ≥ 3 times the upper limit of normal as clinically significant.^[6] If the ALT reference range was initially determined to be 5–40 IU, values above 40 IU would be considered to be abnormally high, and a value of 125 IU would be clinically significant. However, if the reference range changed to 5–45 IU, a value of 42 IU would be normal and a value of 125 IU would no longer be categorized as clinically significantly abnormal. Differences in laboratory reference ranges must also be considered in multicenter studies where each site may be using its own local laboratory. When doing multinational studies, the laboratory units to be used for analysis must be determined, i.e., conventional or SI units. Unit conversion using predefined algorithms must then take place.^[10] Errors in data conversion are not uncommon.

Data-Integration Issues

Pooling of laboratory data across investigative sites is required in multicenter and multinational studies. Furthermore, for product registration proposes, pooling of studies is often required to adequately show the safety profile of the investigational product under review because individual studies may not be large enough to identify important safety trends. Considerations in pooling studies include the following: study design, i.e., double-blind placebo-controlled studies vs. comparator-controlled studies vs. open-label studies; dose of investigational product and

active-controlled agent; number of randomized patients; and the duration of the study. Clearly, pooling large, Phase 3, placebo-controlled studies with similar study durations utilizing similar doses would provide more meaningful information than pooling small, open-label, dose-ranging studies where investigational product was given for short periods of time. Such factors must be taken into account when data are pooled.

Subgroup Analyses

Potential interactions other than the effect of the investigational product on a laboratory parameter should also be explored if there is sufficient data to do so. Subgroup analyses are therefore usually performed in large studies or on pooled data. Common variables for subgroup analyses typically include gender, age, and race. Other variables to be analyzed may be dependent on the pharmacological properties of the product. For example, if there is a suspicion that renal function may influence laboratory changes, a subgroup analysis comparing patients with normal baseline renal function should be compared to patients with abnormal baseline renal function. The information from subgroup analyses is limited and confounding when the subgroups are skewed, e.g., there are considerably more younger patients (≤ 65 years old) than older patients.

Other Analysis Issues

A decision needs to be made regarding what time points should be used for each analysis, i.e., baseline to each visit, baseline to final treatment, or some combination of the two. For shorter-term studies, baseline to final treatment may be adequate. In longer-term studies, looking at mean changes over time is often performed to determine if time may be a factor in any observed changes from baseline. These conventions are often applied to the shift table analysis as well. For shift tables, the lowest and highest postbaseline values regardless of when these occurred might also be considered. Because of the clinical relevance of clinically significant laboratory values, it is recommended that a subject be included in the clinically significant categorical analysis if the subject has one or more values that exceed the predefined threshold regardless of when the observation was made postbaseline. The recommended listing of clinically significant values discussed in the “Methods” section allows the reviewer to determine the time course and pattern of the observed clinically significant value.

Central Laboratories

The use of central laboratories, when available, makes data integration easier and reduces errors in transcription, unit conversion, and reporting inconsistencies. Laboratory samples are typically delivered to a central processing center where the same methodological standards are used.

Reference ranges are established for each analyte. Values below or above the reference range are flagged as such. Furthermore, “panic” values, which are clinically significant and of clinical concern, can be established so that the investigator and sponsor are promptly notified of such values to ensure the proper care of the patient.

Printouts of laboratory results are usually used in place of a separate laboratory case report form. This not only reduces transcription time but also reduces human error in transcribing results. Data tapes are often made available to the sponsor so no additional data entry of the laboratory results is required.

Offsetting the benefits of using a central laboratory are the costs, changes in reference values that occur even with the use of central laboratories especially during trials of longer duration. Furthermore, not all sites participating in a multinational study are likely to be able to use one central laboratory because timely shipment of samples across country borders or overseas is not practical. Even with these limitations the use of central laboratories when available is still preferred.

Study Population Considerations

In interpreting the clinical significance of laboratory changes, it is also important to understand the characteristics of the study population with respect to their underlying diseases and the concomitant medications they are taking. For instance, if a study is being conducted in patients with severe heart failure, it is likely that this study population will be taking diuretics. Diuretic use can affect a number of laboratory parameters including sodium, potassium, calcium, uric acid, and glucose levels.^[11] In such cases, if there is an observed change in one or more of these analytes it has to be determined if the change is due to the investigational product, the underlying disease, or the concomitant medication. This determination can usually be made in comparative trials where there would be no difference in “background noise” between treatment groups assuming that the two treatment groups were balanced with respect to the severity and duration of heart failure, there were no significant differences in the presence of underlying disease or in concomitant use of diuretics, and there were no differences in baseline mean laboratory values. This becomes more of a challenge in open-label trials where there is no comparative group to determine the background noise of the study population.

Clinical Relevance of Laboratory Findings

There are a number of factors that must be taken into account in the proper interpretation of laboratory findings. These factors include spurious laboratory values, changes in some laboratory analytes that are directionally opposite to changes seen in other related analytes, statistically significant findings that are not clinically relevant, and clinically relevant findings that are not statistically significant.

Real Laboratory Findings Vs. Laboratory Error or Artifact

It is important to verify all clinically significantly abnormal laboratory values because some extreme values may simply be due to laboratory error or artifact. For instance, it is not uncommon to find elevated levels of potassium. Large increases in this electrolyte can result in death. Subjects with normal renal function are usually able to excrete excess potassium. Therefore, for any clinically significant elevated potassium value it is important to look at renal function as well. A significantly elevated potassium level in a healthy subject with a normal urinalysis and normal creatinine and BUN values would suggest a laboratory error or artifact. Red blood cells are rich sources of potassium. If a blood sample sits too long or if the red blood cells are damaged in some way, potassium can escape from the red blood cells and settle in the serum where it would increase the serum content of potassium and result in an abnormally high potassium value when measured. Sometimes a laboratory report will indicate that the blood sample was hemolyzed (i.e., the red blood cells were ruptured), and a repeat potassium level should be obtained. Therefore, all clinically significantly abnormal laboratory data should be critically reviewed and repeated, if necessary, before such data are included in any laboratory analysis.

Laboratory data must also be reviewed for the directional change of similar analytes. For example, hemoglobin and hematocrit are both measurements involving red blood cells (RBCs) and consequently should move in the same direction. A patient with anemia would be expected to have both a low hemoglobin value and a low hematocrit value. If the hematocrit was listed as below normal and hemoglobin was listed as above normal, one should suspect a laboratory error or transcription error for one of these analytes. Similarly, in analyzing hemoglobin and hematocrit mean changes, shift tables, and clinically significant changes, each analysis should show similar directions/patterns for these analytes. Therefore, if there was a mean increase for hematocrit but a decrease in mean hemoglobin this would either represent an analysis error, or if no analysis error was detected these disparate findings would suggest that such findings were not clinically meaningful.

Clinical Correlations

In the “Methods” section four different approaches to analyzing laboratory data were summarized. It would not be unusual if all four analyses did not reveal similar trends for any given analyte; however, the more analyses that show similar trends the greater clinical weight there is that the finding is clinically meaningful. Using the example of hemoglobin and hematocrit discussed above, if the mean change analysis indicated that both hemoglobin and hematocrit showed statistically significant decreases from baseline, the other analyses should

Table 5 In-Text Laboratory Analysis Display

Analyte X	Dose 1 (n = 50)	Dose 2 (n = 50)	All doses (n = 100)	Placebo (n = 50)	Comparator (n = 50)
Mean change from baseline to endpoint $\pm$ SD	10.8 $\pm$ 0.3	9.7 $\pm$ 0.8	10.3 $\pm$ 0.5	9.5 $\pm$ 1.2	11.6 $\pm$ 0.2
No. (%) low/normal to high	1 (2%)	2 (4%)	3 (3%)	1 (2%)	5 (10%)
No. (%) high/normal to low	0	1 (2%)	1 (1%)	0	1 (2%)
No. (%) clinically significant changes	0	1 (2%)	1 (1%)	0	2 (4%)

be reviewed for similar trends. If this was a clinically relevant finding one might see a greater percentage of patients with decreases than increases in these analytes in the shift tables. There may also be a higher proportion of subjects with clinically significantly low values. A review of adverse events might show a higher incidence of anemia in the group that received the investigational drug. All these findings would lend weight to the observation that decreases in hemoglobin and hematocrit values were “real” findings.

Table 5 shows a summary of the findings of different laboratory analyses for any given analyte. Such data display help in the review of laboratory findings and in determining the clinical relevance of the laboratory findings.

Clinical Significance vs. Statistical Significance

Statistical analysis of laboratory data for safety determination provides an important tool in aiding the objective interpretation of laboratory findings. Its use is analogous to a filter that blocks out extraneous “random noise” and allows the reviewer to focus on the data that require more in-depth review, helping to identify and quantify any relevant safety findings. However, statistical testing is not always indicated in the analysis of laboratory data.

For instance, a drug that showed frequent toxic effects on the bone marrow would have an unfavorable benefit to risk ratio unless the drug was a chemotherapeutic agent for the treatment of a serious and life-threatening disease such as cancer. Therefore, it would be unexpected and unacceptable to see statistically significant mean decreases in RBCs, WBCs, and platelet counts routinely in a patient population taking drugs for self-limiting conditions such as the common cold. Nevertheless, a small number of individuals, due to hypersensitivity and idiosyncratic reactions, could still be at risk. In one population case control study the incidence of drug-induced aplastic anemia (a serious condition that affects the bone marrow and can lead to severe anemia, life-threatening infections, and bleeding episodes) was quite rare, i.e., estimated to be 0.5 per million per year.^[12] Because these events occur so rarely, mean changes and routine shift tables would unlikely identify any problem. However, if a patient demonstrated the platelet pattern shown in Example B (Table 4) and no alternative explanation could

be found, this single patient's findings would be sufficient to raise suspicion that the biopharmaceutical product under investigation may cause clinically relevant decreases in platelets. The clinical investigation team would then need to be on “heightened alert” for similar cases.

CONCLUSION

In summary, accurate laboratory data, appropriate analysis, and clinical correlation are requisite to ensure the validity of laboratory findings. Having assured this, laboratory analyses play an invaluable role in helping define the safety profile of biopharmaceutical products under investigation.

ACKNOWLEDGMENT

I would like to thank Dr. Carmela Skillman for her critical review of the manuscript and her invaluable assistance in its preparation.

REFERENCES

1. Hardman, J.G.; Limbird, L.E.; Molinoff, P.B.; Ruddon, R.W.; Gilman, A.G. Drugs Used in the Treatment of Hyperlipoproteinemias. In *Goodman and Gilman's The Pharmacological Basis of Therapeutics*; McGraw-Hill: New York, 1996; 875–897.
2. Sacher, R.A.; McPherson, R.A. Principles of Interpretation of Laboratory Tests. In *Widmann's Clinical Interpretation of Laboratory Tests*; F. A. Davis Company: Philadelphia, PA, 2000; 3–27.
3. Paterson, N.; Frazer, S.C. The International System of units (SI units) in medicine. *Br. J. Hosp. Med.* **1975**, *13*, 759–770.
4. *Structure and Content of Clinical Study Reports*; ICH Guideline E3, 1996.
5. *Guideline for the Format and Content of the Clinical and Statistical Sections of an Application*; FDA: CDER, July 1988.
6. *Supplementary Suggestions for Preparing an Integrated Summary of Safety Information in an Original NDA Submission and for Organizing Information in Periodic Safety Updates*, FDA: CDER: Neuropharmacology Division, 1987.

7. *Reviewer Guidance: Conducting a Clinical Safety Review of a New Product Application and Preparing a Report on the Review*; FDA: CDER, 1996. (Draft).
8. Carvounis, C.P. Data Distribution and Probabilities—the Heart of Statistics and Epidemiology. In *Handbook of Biostatistics: A Review and Text*; The Parthenon Publishing Group: New York, 2000, 11–26.
9. *Guideline on Clinical Safety Data Management: Definitions and Standards for Expedited Reporting*; ICH Guideline E2A, 1995.
10. Fischbach, F. Examples of Conversions to Système International (SI) units. In *A Manual of Laboratory and Diagnostic Tests*; Lippincott, Williams and Wilkins: Philadelphia, 2000, 1138–1141.
11. Hardman, J.G.; Limbird, L.E.; Molinoff, P.B.; Ruddon, R.W.; Gilman, A.G. Diuretics. In *Goodman and Gillman's The Pharmacological Basis of Therapeutics*; McGraw-Hill: New York, 1996, 685–713.
12. Patton, W.N.; Duffull, S.B. Idiosyncratic drug-induced haematological abnormalities: Incidence, pathogenesis, management and avoidance. *Drug Safety* **1994**, *11* (6), 445–462.

Latent Class Analysis

Elizabeth S. Garrett-Mayer

Hollings Cancer Center, Medical College of South Carolina, Charleston, South Carolina, U.S.A.

Jeannie-Marie S. Leoutsakos

Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, Maryland, U.S.A.

Abstract

A variable is latent if it cannot be directly measured, but can be inferred through the measurement of other related observable variables. In this entry, different models and methods of latent class analysis are examined, and practical applications and examples are provided.

Keywords: Latent class model; Latent variable; Markov chain Monte Carlo; Major depression; Sensitivity; Validity.

INTRODUCTION

A variable is latent if it cannot be directly measured, but can be inferred through the measurement of other related observable variables. For example, in mental health research, major depression is a latent variable: one cannot directly measure depression, but we can observe symptoms of depression, such as sadness, weight loss, and fatigue. Latent variables can take different forms. They can be continuous, such as intelligence quotient (IQ), or they can be categorical, such as social class. In biomedical research, conceptualizing latent outcome variables as categorical is appealing because the categories can be thought of as diagnostic classes.

The latent class model (LCM) is a method for defining a categorical latent variable based on a set of discrete observed variables. It views the population as a set of subpopulations (or classes) and the observed variables as surrogates that imperfectly measure to which class an individual belongs.^[1] A more mathematical interpretation of the LCM is that it describes the association of the observed variables in a multiway contingency table. In the case of K observed binary variables, the dimension of the contingency table is 2^K . For example, we can fit a LCM using a set of symptoms of depression where symptoms are recorded as “present” or “absent,” and our latent class variable is depression. There are other models for related latent variable situations: for a categorical latent variable and continuous observed variables, we use the latent profile model; for a continuous latent variable and continuous observed variables, we use factor analysis; and for a continuous latent variable and categorical observed variables, we use latent trait models.

Many situations in which a disorder is diagnosed as a syndrome (i.e., a constellation of symptoms) are appropriate

applications for a latent class analysis. However, it is most useful when the observed variables measure diagnoses, rather than continuous scores, and the patterns of the symptoms are thought to contain more information than simply summing up the number of symptoms. The goal of the LCM is to cluster individuals into diagnostic groups, as opposed to grouping the observed symptoms (as in factor analysis).

The clinical utility of the LCM is multifold. Individuals can be clinically diagnosed into latent disease categories. This identification can help allocate resources appropriately and efficiently. The LCM can describe the prevalence of a disease or disorder, facilitating a better understanding of the resources necessary in the community and the potential benefits of early interventions in some instances. Additionally, the LCM can predict prevalence of disease in different populations, which can help in estimation and allocation of services. The LCM has also been used for validation of diagnostic criteria.

A MENTAL HEALTH EXAMPLE: MAJOR DEPRESSION IN THE EPIDEMIOLOGIC CATCHMENT AREA STUDY

Latent class models are prevalent in many areas of medical research, especially in mental health and psychiatric research. For example, LCMs have been used for defining and understanding autism,^[2,3] attention-deficit hyperactivity disorder (ADHD),^[4] alcoholism,^[5,6] bulimia,^[7] depression,^[8–13] harmful behaviors,^[14] Alzheimer’s disease,^[15] schizophrenia,^[16] anxiety and depression in children,^[17] and general psychoses and syndromes,^[18,19] among others. Outside of mental health, LCMs have

been seen in medical research areas such as gerontology,^[20] genetics,^[21] medical care,^[22,23] ophthalmology,^[24] fatigue,^[25,26] and cytometry.^[27] The origin of the LCM in psychiatric research was the validation by Young^[28] of the classification system for schizophrenia developed by Taylor and Abrams.^[29] Since then, the LCM has been used by many researchers for evaluating criterion validity of diagnostic definitions.^[4,9,30,31]

Major depression is a classic example of a latent variable. Moreover, it has been studied and analyzed via LCMs by many researchers.^[9–13,17,30,31] The results of these studies vary largely because of differences in population composition: some are patient populations, while others are epidemiologic samples. The Epidemiologic Catchment Area Study (ECA), which is a multisite, multiwave community sample of individuals, was used in four of these publications,^[9–12] and it will be used here to facilitate discussion of the LCM in practice.

The ECA includes a series of surveys conducted by university-based researchers in five community populations focusing on mental health.^[32] Baseline assessments were made in 1981 (wave 1) in 13,538 respondents using the Diagnostic Interview Schedule (DIS)^[33] and again in 1982 (wave 2). The Baltimore site, which interviewed 3481 individuals in wave 1, followed up its sample between 1993 and 1996 and was able to conduct wave 3 DIS interviews of 1865 of the original 3481 from its 1981 community dwelling cohort. Of the original 3481, 452 could not be located, 878 has died, and 286 did not agree to participate. Validity of the DIS is shown in Ref. [34].

For each symptom in the DIS, individuals were asked if they had ever experienced the symptom. If the respondent answered yes, a series of probe questions followed which assured that the symptom was not trivial. Nontrivial symptoms with no medical explanation are considered plausible psychiatric symptoms, and for these, the respondent was asked to date the most recent occurrence of the symptom. The 25 questions relating to depression in the DIS can be collapsed into 10 symptom groups for depression as defined by the International Classification of Diseases, edition 10 (ICD-10)^[35] and the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV).^[36] For example, all questions in the DIS relating to sleeping problems define one of the symptom groups. Reporting at least one of the sleeping symptoms (i.e., trouble falling asleep, trouble waking, or sleeping too much) is considered a positive symptom response for that symptom group.

The 10 symptom groups and the prevalence of symptoms in the ECA Baltimore site in wave 3 (1993–1996) are in Table 1. The most prevalent symptom group is group 5 (sleeping problems) and the least prevalent is group 6 (guilty/sinful feelings). These 10 symptom groups will be the observed variables that we use to define the latent variable depression in our LCM.

Table 1 Prevalence in Symptom Groups in ECA Wave 3 Depression Data, $N = 1322$

Group	Symptoms	Prevalence
1	Depressed mood Disinterest in sex	0.06
2	Less fun Loss of enjoyment	0.08
3	Reduced energy/fatigued Ideas of self-harm	0.05
4	Want to die Suicidal thoughts Suicide attempts Trouble falling asleep	0.05
5	Trouble waking Sleep too much	0.09
6	Guilty/sinful feelings Reduced concentration	0.02
7	Slow thoughts Indecisive	0.04
8	Feel inferior Lacking self-confidence Loss of appetite	0.03
9	Weight loss Increased appetite Weight gain Slow movement	0.08
10	Fast movement Fidgety	0.04

THE LATENT CLASS MODEL

The LCM provides us with two types of parameters which define the model—(i) class size: a vector indicating the proportion of individuals in each of the classes; and (ii) symptom prevalences: for each class, we obtain a vector of probabilities indicating the prevalence of each of the symptoms in the class.

Notation

In the LCM, the observed variables are often called items, symptoms, or indicators. To specify a LCM with M classes and K observed symptoms, we define Y_i to be a vector of length K corresponding to the K (binary) symptoms of individual i , with 1 indicating presence of symptom and 0 indicating absence. The key model parameters are p and π . p is a K by M matrix of the symptom prevalences by class. Specifically, p_{km} is the probability that an individual in class m will exhibit or report symptom k . The parameter vector π has length M and represents the class sizes. For example, π_m is the proportion of individuals in the population in class m . For a dataset with N individuals, the latent

Table 2 Parameter Estimates in a Three-Class Latent Class Model for Depression

		Class 1	Class 2	Class 3
1	Depressed mood	0.02	0.30	0.78
2	Loss of interest	0.02	0.32	0.91
3	Fatigue	0.01	0.25	0.64
4	Morbid thoughts	0.02	0.24	0.54
5	Sleep	0.03	0.49	0.77
6	Guilt/sin	< 0.01	0.08	0.37
7	Concentration	0.01	0.15	0.78
8	Self-esteem	< 0.01	0.10	0.56
9	Appetite/weight	0.04	0.31	0.78
10	Movement	0.01	0.17	0.73
	Class size	0.88	0.09	0.03

class variable is denoted by the N -vector η , where η_i is subject i 's latent class and can take integer values between 1 and M .

Table 2 provides the parameter estimates of the three-class latent class model for depression. The class size vector π is (0.88, 0.09, 0.03). Class 1 appears to have low prevalence of symptom, with all less than 0.05. Class 3 has the highest prevalences for all symptoms, with a range of 0.37 for guilt/sinful feelings to 0.91 for loss of interest. The prevalences in class 2 range from 0.08–0.49, with most in the 0.15–0.40 range. Based on these results, we can conclude that, assuming that three classes of depression exist, there is a large class of non-depressed individuals (88%) in the population, a small class of depressed individuals (3%), and a class of individuals reporting a few symptoms, but not appearing to be severely depressed (9%).

The Likelihood Function

Now that the model parameters are better understood, it is important to consider the likelihood function of the LCM. The probability that individual i reports a specific symptom pattern y^* can be written as follows:

$$P(Y_i = y^* | p, \pi) = \sum_{m=1}^M \pi_m \prod_{k=1}^K p_{km}^{y_{ik}^*} (1 - p_{km})^{(1-y_{ik}^*)} \quad (1)$$

This is essentially a weighted average of the probability of reporting the pattern in each of the classes where the weights are the class sizes. Using Eq. 1, we can write down the likelihood function for the latent class model by taking the product over N :

$$L(p, \pi | Y) = \prod_{i=1}^N \sum_{m=1}^M \pi_m \prod_{k=1}^K p_{km}^{y_{ik}} (1 - p_{km})^{(1-y_{ik})} \quad (2)$$

where Y is the $N \times K$ data matrix for data on the N individuals and the K observed variables.

Notice that nowhere in the likelihood do we see the latent class variable η . Although it is not a parameter that is estimated via the likelihood function, it does play a role in estimation when using Bayesian estimation approaches where we condition on η . Eq. 3 shows the conditional distribution of the data, given η and the other model parameters.

$$f(Y | \eta, p) = \prod_{i=1}^N \prod_{k=1}^K p_{\eta k}^{y_{ik}} (1 - p_{\eta k})^{(1-y_{ik})} \quad (3)$$

Latent Class Model Assumptions

The LCM has several assumptions. The first, a rather simple assumption, is that each individual belongs to one and only one of the M classes. Second, the number of classes M must be chosen a priori to fit the model. It is most common that one does not know the number of classes, and in that case, you would iteratively fit two-class, three-class, etc. models until you find one that is suitable based on goodness of fit and/or comparison of models. The last assumption to mention is conditional independence: conditional on an individual's latent class, the individual's responses to items are independent. Mathematically,

$$P(Y_{ik}, Y_{ij} | \eta_i) = P(Y_{ik} | \eta_i) P(Y_{ij} | \eta_i) \quad (4)$$

This assumption is obvious from the likelihood: the product over the K symptoms implies that the occurrence of those symptoms is independent, conditional on the class. While the first two assumptions are implicit, the conditional independence assumption is sometimes violated, leading to incorrect inferences. As such, it is an important assumption to check after fitting the LCM. This will be discussed in "Model Evaluation."

The above noted are statistical assumptions. There are some other practical assumptions about the validity of the latent class model worth discussing. Just as in factor analysis and other multivariate techniques, the items that you choose to include in the model are assumed to "define" the underlying construct. For our example of depression, by including exactly these 10 symptom groups, we are assuming that exactly these 10 symptom groups define depression and that we have not neglected to include any other relevant symptoms from the model.

Identifiability

The identifiability of a particular parameter describes the ability of the data to uniquely estimate (or identify) the model parameter. Of the several types of identifiability described in the statistics literature, two types of identifiability are particularly relevant for latent class models: local identifiability and weak identifiability.

Local Identifiability

Recall that a latent class model is a method to describe a multiway contingency table. Given that there are 2^K cells in this contingency table (for binary items), it is not possible to identify a model that has more than 2^K parameters. For this reason, a necessary condition for identifiability of LCMs is $M(K-1) < 2^K$. However, this is not a sufficient condition—there are some models satisfying this inequality that are not identifiable. Goodman^[37] introduced a sufficient condition for identifiability: $M < 2^{(K-1)/2}$. Models satisfying the necessary but not the sufficient condition may or may not be identifiable. Note that this kind of identifiability is what can be called “theoretical” or “asymptotic” identifiability because it is based on the assumption that the dataset is large enough to identify all model parameters. However, LCMs often require very large datasets to be able to estimate all parameters with relative precision.

The notion of satisfying the local identifiability conditions, yet having too small a dataset to identify all model parameters, can be called “empirical identifiability.” Via statistical theory, empirical identifiability is equivalent to having a nonsingular information matrix. In practice, one can calculate the rank of the Jacobian. However, it is often more practical to estimate the model of interest using several starting values in the estimation procedure, as described in the next section. If different starting values lead to different solutions, then empirical identifiability may not have been achieved.

Weak Identifiability

Recall that the Bayesian statistical paradigm is based on the idea that a statistical model is composed of two key quantities: the prior distribution and the likelihood. These are combined in a Bayesian analysis to produce the posterior distribution of the parameters, with which we make statistical inferences. In Bayesian estimation procedures, “weak identifiability” describes the information for parameter estimation provided in the prior distribution and the likelihood: if the prior distribution provides most of the information and the data provide relatively little, then we say that the particular parameter is only weakly identified. This is a slightly different concept than local identifiability in its traditional interpretation. Weak identifiability is defined in the context of Bayesian statistics and occurs when the technical conditions for “theoretical” identifiability (as

mentioned above) are met, but the data provide little information about some parameters so that their posterior and prior distributions are very similar. Weak identifiability also implies that empirical identifiability, as defined in the previous section, is not achieved. Lindley^[38] notes that although a Bayesian model may be only weakly identified, it can still be valid in the sense that the model can still adequately describe the data via its identifiable parameters. It is, however, important to determine which parameters, or functions of parameters, are identifiable and only interpret the model results based on these parameters.

Note that there is a significant difference between the local identifiability discussed above and weak identifiability. Both concepts describe the ability of the data to estimate the parameters, but if we have nonidentifiability in the case where we are estimating the model using a non-Bayesian approach, we cannot estimate the parameters. However, if we choose a Bayesian approach (as described in “Estimation Methods”), we can rely on prior distributions to estimate parameters. But, as noted above, this does not mean that the weakly identified parameters can be interpreted: it simply means that the model is still valid and the identifiable parameters can be interpreted. Moreover, in the case where there is a priori information about the latent classes, either from previous analyses or based on expert opinion, this information can be included in a stronger prior distribution for a parameter. For example, in our depression example, it is well known that the majority of the general population does not suffer from depression. We could use this to assist in modeling the π vector via the prior distribution for π . However, using strong priors should only be carried out when there is good evidence with which to do so.

Sample Size Considerations

One potential limitation of the LCM is that it usually requires a large sample size for model estimation, usually at least several hundred observations and, in some cases, several thousand observations. The size of the dataset required depends essentially on the size of the classes and on the variability of the items in the sample. For example, for evaluating depression, we have an epidemiologic sample which, as expected, has relatively low prevalence of symptoms, so that there is relatively little variability in the symptoms in the sample. Additionally, depression is relatively rare in an epidemiologic setting, and so we have one large class, and then we expect one or more rather small classes. Because of these two characteristics, we need a relatively large sample to detect the depressed classes in our population. If we instead had a patient population in which symptoms were more prevalent (i.e., prevalence for symptoms closer to 0.50) and the “nondepressed” individuals did not account for the large majority of the sample, then we would be able to detect smaller classes with a smaller sample.

A Monte Carlo study, in which many samples are drawn from a hypothesized population with user-specified parameter values, can be conducted. The model is then estimated for each sample, and parameter estimated and standard errors are averaged across the samples. Statistical power for a particular scenario and sample size can then be estimated as the percentage of samples in which the null hypothesis was rejected.^[39]

Posterior Probability of Class Membership

A useful quantity that can be estimated from the model parameters is the posterior probability of membership. This provides, for each possible response pattern, the probability that an individual with the response pattern y^* belongs to each of the classes:

$$P(\eta_i = m | y^*) = \frac{\pi_m \left(\prod_{k=1}^K p_{km}^{y_k^*} (1 - p_{km})^{(1-y_k^*)} \right)}{\prod_{l=1}^M \pi_l \left(\prod_{k=1}^K p_{kl}^{y_k^*} (1 - p_{kl})^{(1-y_k^*)} \right)} \quad (5)$$

Although it may seem complex at first glance, it is a simple implementation of Bayes rule on Eq. 2. For example, using the information provided in Table 2, we can estimate the likely class for an individual reporting only “loss of interest” and “sleep” symptoms. To do this, we use the equation above. For this reported pattern, we find that the individual has a 17% chance of being in class 1, an 83% chance of being in class 2, and less than a 1% chance of being in class 3. For other more common patterns, see Table 3.

A Reparameterization

In some settings, it proves useful to reparameterize the LCM using parameters on the logit scale:

$$p_{km} = \frac{e^{\gamma_{km}}}{1 + e^{\gamma_{km}}}$$

$$\pi_m = \frac{e^{\alpha_m}}{\sum_{m=1}^M e^{\alpha_m}}$$

When using the parameterization above with parameters γ and α , the likelihood function becomes:

$$L(p, \pi | y) = \prod_{i=1}^N \sum_{m=1}^M \frac{e^{\alpha_m}}{\sum_{l=1}^M e^{\alpha_l}} \prod_{k=1}^K \frac{e^{\gamma_{km} y_{ik}}}{1 + e^{\gamma_{km}}} \quad (6)$$

where $\alpha_M = 0$.

For example, in the Mplus software program,^[40] which implements an EM algorithm for model estimation, and in some Bayesian approaches, this reparameterization is used (see Ref. [10] and “Estimation Methods” for more details).

ESTIMATION METHODS

Likelihood-Based Approaches

The most prevalent estimation method for the LCM is a likelihood-based approach. Specifically, expectation-maximization (EM) algorithms are most often used. There are numerous software packages designed to estimate latent class models using this approach: Mplus,^[40] Latent-GOLD, “maximum likelihood latent structure analysis” (MLLSA), “linear structural regression” (LISREL), and others. For a rather comprehensive list of software, suggestions, and links to software, see Ref. [41]. Note that many of these software packages either suggest or require the user to supply “starting values” for the parameters.

A major consideration when using likelihood-based approaches is the potential problem of lack of convergence. This may occur when there are identifiability problems (that is, the dataset is relatively small, or one of the class sizes is very small and there are not enough individuals in this small class to estimate parameters with reasonable precision), or the conditional independence assumption has been violated. It is often suggested that whenever fitting a LCM, one should repeat the estimation procedure several times using different sets of starting values. This will often (but not always) diagnose a problematic model fit.

Another reason that lack of convergence arises is that the parameter estimates may get “stuck” at the boundary of the parameter space during estimation. This is a result of the algorithmic nature of the fitting procedure. In this case, the resulting model fit will show parameter estimates

Table 3 Posterior Probabilities of Class Membership for Common Response Patterns

	Pattern	Class 1	Class 2	Class 3
No symptoms	0000000000	0.99	0.01	< 0.01
All symptoms	1111111111	< 0.01	< 0.01	> 0.99
Somatic symptoms	1101011100	0.01	0.97	0.03
Mood symptoms	0010100011	< 0.01	0.06	0.94
Depressed mood only	1000000000	0.87	0.13	< 0.01

(usually in the p matrix) that have values of exactly 0 or 1. When you encounter estimates such as these, it is unlikely that these are correct. It is probable that the model did not properly converge and should be refit using different starting values.

Markov Chain Monte Carlo Estimation Approaches

The Markov chain Monte Carlo (MCMC) approach to model estimation stems from Bayesian methodology. The general idea is described by Casella and George,^[42] and a more technical treatment can be found in Geman and Geman.^[43] There are many examples of models akin to latent class models which are estimated by MCMC.^[44–48] To see details of the MCMC approach for LCMs, see Ref. [10]. This model fitting approach tends to require more user input and statistical expertise than the likelihood-based approaches. So while there are arguably benefits to the use of the MCMC approach, it is not necessarily a feasible estimation method for those unfamiliar with Bayesian statistics. The MCMC estimation can be performed using the Bayesian package WinBUGS or can be programmed in mathematical or statistical software such as C, Fortran, SAS, Mathematica, Splus, or R.

A difference between the MCMC and the likelihood-based approach is that the likelihood-based software packages usually provide point estimates and standard errors of model parameters (i.e., π and p , or in the parameterized case, α and γ). From these, the user can estimate posterior probabilities of class membership (see “Latent Class Model Assumptions”) but not their standard errors. From the MCMC estimation procedure, the user obtains point estimates and precision estimates of the p and π parameters and has the additional benefit of obtaining point estimates and precision for any function of the parameters. This is a nice feature of the MCMC approach, especially when the goals of the analysis are more than just to obtain the standard model parameters. For example, in “Practical Application,” we use the results from the MCMC to derive sensitivity and specificity estimates (and estimates of their precision) using the MCMC results. This could not be performed using standard likelihood-based software.

In the MCMC approach, the user does not need to specify parameter starting values. However, he does need to choose prior distributions for each parameter. These distributions can be “vague,” essentially providing no a priori information about the parameter estimates, or they can be informative. In some cases, it is preferable to use informative priors. One advantage of the use of priors is that we can avoid the parameter boundary problem that the likelihood-based approach encounters: we can choose priors that do not allow parameters to take values on the parameter boundary.

There are some shortcomings to the MCMC approach. It tends to take longer than the likelihood-based approach

(although this is not universally true and, in many cases, both converge rather quickly depending on the size of the dataset). Additionally, in some cases, the post hoc calculations required to get to the parameters of interest can be substantial and require some programming expertise.

MODEL EVALUATION

Choosing a Model and Assessing Fit

There has been considerable debate about approaches for model selection and for assessing goodness of fit in latent class models. Historically, the most popular approach has been to use likelihood ratio tests or Pearson chi-square statistics to assess fit or choose models. This assumption is not necessarily valid in most cases for a variety of reasons.^[49–52] Some of these reasons include that these statistics do not generally follow a χ^2 distribution in the LCM case, and that a k -class model is not necessarily nested in a $k + 1$ -class model. Additionally, Collins et al.^[50] have shown that these statistics may be either too liberal or too conservative, depending on factors including sample size, number of items, and number of classes. Information criteria, notably the Akaike Information Criteria (AIC)^[53] and the Bayesian Information Criteria (BIC),^[54] are alternative model selection statistics. The BIC is particularly appropriate for LCMs because of its adjustment for sample size, which tends to be large for most latent class analyses. The asymptotic properties of these two information criteria are such that the AIC is not consistent but is efficient for large sample sizes, while the BIC is consistent and not efficient.^[55,56] In simulation studies, the BIC typically outperforms AIC, with the AIC often supporting estimation of a larger number of classes than the BIC.^[57] The parametric bootstrap approach (BLRT) has been found to be a useful alternative to some of the problems arising in the statistics described above. Essentially, this method empirically determines the distributions of test statistics of interest (i.e., the likelihood ratio and the Pearson chi-square). This is described by Ref. [58] and can be seen in practice in Refs. [50,59]. In simulation studies, the BLRT typically outperforms the BIC, though it adds significantly to computation time.^[57]

Lo et al.^[60] have described an approximation to the distribution of the difference in likelihood values between a k -class model and a $k + 1$ -class model (LMR). In simulation studies, the LMR typically performs well, but not as well as the BLRT. Both the BLRT and the information criteria depend on model and distributional assumptions, whereas the LMR does not Ref. [57,60]. Therefore, in cases in which there is concern that the model is misspecified, the LMR might be preferable. It should be noted that Jeffries^[61] has pointed to a potential flaw in the proof supporting the LMR, though its impact (if any) on the utility of the statistic is not known. Graphical techniques for assessing model fit and for model selection have been introduced.^[10]

The Bayes factor, which has a close connection with the BIC,^[62] can be used in slightly different forms.^[63,65] Other methods including parsimony indices and ratios have been reviewed.^[66] Some additional approaches are the goodness-of-fit test^[67] and the Rudas-Clogg-Lindsey index.^[68]

As much as statistical appropriateness is an important consideration of model selection and/or model fit, another important factor in choosing the most appropriate model is the scientific interpretation. As stated in “Latent Class Model Assumptions,” there should be strong theory as a basis for choosing the items to include in the LCM. When choosing between models, one should consider how consistent each of the estimated sets of model parameters are with prior knowledge about the underlying latent variable of interest.

In our depression example, based on graphical techniques^[10] and the BIC, the three-class model was deemed to be the best model and demonstrated sufficient fit.

Assessing Model Assumptions

Recall the conditional independence assumption from “Latent Class Model Assumptions.” It is critical to assess whether or not this assumption has been met, otherwise incorrect inferences may be made. A relatively simple and straightforward approach is to assign individuals to classes based on their posterior probabilities of class membership and then determine whether or not items within classes of individuals appear to be independent. This can be performed in the binary item case by estimating odds ratios between items for individuals in the same class [for a total of $K \times M \times (M - 2)/2$ estimated odds ratios]. If these odds ratios appear to be close to 1, then the conditional independence assumption is met. However, if some of these odds ratios show evidence of being different than 1, then the model should be reconsidered. In cases where the vector of posterior probabilities of class membership does not clearly demonstrate membership in a particular class, simply assigning individuals to the model class may bias tests of model assumptions. Instead, one can multiply impute class membership a number of times and combine the results of the imputations to look for evidence that any of the between-indicator odds ratio are different than one [Ref. 20].

Solutions to violations of the conditional independence assumption include combining items that are highly related into one symptom group and fitting a model with more classes. Precision estimates for these odds ratios can be obtained with either method of estimation discussed.^[10,20]

PRACTICAL APPLICATION

We continue with our example of depression in the ECA study and compare our latent class model results with those of the DSM-IV diagnosis^[36] of major depression.^[10] After

using the MCMC approach for model estimation, we use the results of our MCMC chains to calculate the posterior distributions of sensitivity and specificity and positive and negative predictive values of the DSM-IV criteria, assuming that the latent class model is the “gold-standard” measure of depression. While this assumption may be quite strong, the latent class definition is a relatively unbiased and empirically driven estimate of depression, while the DSM-IV criteria are more subjectively determined by a panel experts. Additionally, it is important to consider that for a disorder to be considered in the DSM-IV, it must be something that is either treatable or would benefit from treatment, and that DSM-IV depression only allows for two diagnostic classes: depressed and nondepressed.

In Fig. 1A, the sensitivity of the DSM-IV diagnosis is shown for each of the three latent classes. The circles represent the estimated sensitivity, and the vertical lines represent the lower 2.5th percentile and the 97.5th percentile of the posterior distribution. (This is analogous to a 95% confidence interval, but not equivalent because of the differences in the estimation procedures.) Recall that the definition of sensitivity is the probability of positive diagnosis given true disease. In our example, we estimate the sensitivity of a DSM-IV diagnosis of depression given each of the three depression classes as defined by the LCM. For example, as we would expect, the probability of DSM-IV diagnosis given class 1 is very low (< 0.01) and the probability of DSM-IV diagnosis given class 3 is quite high (0.90). For class 2, we also obtained a very low sensitivity (0.05). Note that the specificity and the sensitivity for a given class must sum to 1, so that the specificities of classes 1, 2, and 3 are >0.99 , 0.95, and 0.10, respectively. These results imply that the DSM-IV diagnosis of depression is consistent with our class 3 of depression, and that those individuals that we consider to belong to class 2 are rarely identified as depressed by the DSM-IV criteria (only 5% of the time). As a result, the DSM-IV diagnosis has high sensitivity for diagnosing our class 3 depression (0.90) and high specificity (0.05) for classes 1 and 2.

We also consider the PPV and NPV of the DSM-IV diagnosis in our population. In our example, PPV is the probability that an individual who is diagnosed as depressed is truly depressed, and NPV is the probability that an individual who is diagnosed as not depressed is truly not depressed. Recall that, unlike sensitivity and specificity, PPV and NPV are statistics that are population-specific: they depend critically on the prevalence of disease in the population of interest. Our dataset is an epidemiologic sample from the East Baltimore community, and so our results will vary dramatically from a patient population where the prevalence of depression would be expected to be much higher. As above, we calculate the PPV and NPV for each class. For example, for PPV, we estimate the probability of being in each of the classes given a DSM-IV diagnosis. These quantities are shown in Fig. 1C and D. For classes 1, 2, and 3, we see that the PPVs are 0, 0.11, and 0.89, and the

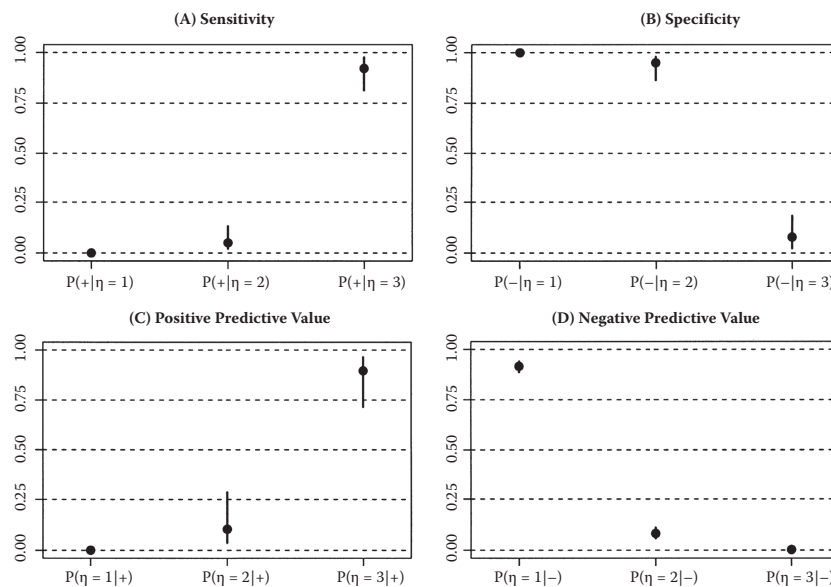


Fig. 1 (A) Sensitivity, (B) specificity, (C) positive predictive value, and (D) negative predictive value of the DSM-IV diagnosis of depression as compared to the latent class model definition

NPVs are 0.91, 0.09, and 0, respectively. These numbers suggest that the PPV for class 3 and the NPV for class 1 are quite good in this population: for an individual who tests positive, the probability of that individual being in class 3 is 0.89, and for an individual who tests negative, the probability of that individual being in class 1 is 0.91. However, for class 2, we see less informative results: both PPV and NPV for class 2 are quite low, implying that individuals in class 2 are not well delineated by this test. Whether a class 2 individual's test result is positive or negative, that individual has about a 10% chance of being in class 2.

This approach is useful because it gives us a reference against which to compare diagnostic criteria. Critics may dislike the assumption that the latent class definition of depression is considered the “gold standard,” but without a true gold standard, the latent class definition can be argued to be our best empirically derived comparator.

CONCLUSION

The latent class model is a useful statistical tool for defining a latent categorical variable based on a set of observed categorical items. LCMs can be especially appropriate for identifying patterns of symptoms of disorders or diseases within populations, for summarizing the distribution of individuals within populations, for deriving diagnostic criteria, and for validation of diagnostic criteria developed by other means. This method has grown in popularity in the past few years because of improvements in accessibility and computational efficiency of latent class estimation procedures. However, there still are issues of estimation and identifiability that should be carefully considered when applying a latent class approach to data.

There are several important further issues to consider. First, the example used here has only considered binary items, but ordinal variables can also be used as observed variables. Second, constraints can be put on parameters in these models, as alluded to in “Identifiability.” For example, if we were interested in a model which includes a class where individuals must report depressed mood, we can constrain the parameter for that class to be 1. Or, if we are willing to acknowledge that there may be some measurement error and that some individuals who actually exhibit depressed mood incorrectly report it, we might want the constraint to be slightly less stringent (i.e., >0.95). Third, there are other models which extend the latent class model by allowing covariates to be associated with the latent class variable.^[20,69] Last, latent class transition models can describe how individuals transition from one class to another over time, and latent growth mixture models can be used to describe latent classes of trajectories over time.^[70,71]

REFERENCES

1. Lazarsfeld, P.F.; Henry, N.W. *Latent Structure Analysis*; Houghton-Mifflin: Boston, MA, 1968.
2. Szatmari, P.; Volkmar, F.; Walter, S. Evaluation of diagnostic criteria for autism using latent class models. *J. Am. Acad. Child Adolesc. Psych.* **1995**, *34*, 216–222.
3. Pickles, A.; Bolton, P.; MacDonald, H.; Bailey, A.; LeCouteur, A.; Sim, C.H.; Rutter, M. Latent-class analysis of recurrence risks for complex phenotypes with selection and measurement error: Atwin and family study of autism. *Am. J. Hum. Genet.* **1995**, *57*, 717–726.
4. Neuman, R.J.; Todd, R.D.; Heath, A.C.; Reich, W.; Hudziak, J.J.; Bucholz, K.K.; Madden, P.A.; Begleiter, H.;

- Porjesz, B.; Kuperman, S.; Hesselbrock, V.; Reich, T. Evaluation of adhd typology in three contrasting samples: A latent class approach. *J. Am. Acad. Child Adolesc. Psych.* **1999**, *38*, 25–33.
5. Bucholz, K.K.; Heath, A.C.; Reich, T.; Hesselbrock, V.M.; Kramer, Jr., J.R.; Nurnberger, J.I.; Schuckit, M.A. Can we subtype alcoholism? A latent class analysis of data from relatives of alcoholics in a multicenter family study of alcoholism. *Alcohol., Clin. Exp. Res.* **1996**, *20*, 1462–1471.
6. Kendler, K.S.; Karkowski, L.M.; Prescott, C.A.; Pedersen, N.L. Latent class analysis of temperance board registrations in Swedish male-male twin pairs born 1902 to 1949: Searching for subtypes of alcoholism. *Arch. Gen. Psychiatry* **1998**, *28*, 803–813.
7. Sullivan, P.F.; Bulik, C.M.; Kendler, K.S. The epidemiology and classification of bulimia nervosa. *Psychol. Med.* **1998**, *28*, 599–601.
8. Sullivan, P.F.; Kessler, R.C.; Kendler, K.S. Latent class analysis of lifetime depressive symptoms in the national comorbidity survey. *Am. J. Psychiatr.* **1998**, *155*, 1398–1406.
9. Garrett, E.S.; Eaton, W.W.; Zeger, S. Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: A latent class model approach. *Stat. Med.* **2002**, *21*, 1289–1307.
10. Garrett, E.S.; Zeger, S.L. Latent class model diagnosis. *Biometrics* **2000**, *56*, 1055–1067.
11. Eaton, W.W.; Dryman, A.; Sorensen, A.; McCutcheon, A. DSM-III major depressive disorder in the community: A latent class analysis of data from the NIMH epidemiologic catchment area study. *Br. J. Psychiatry* **1989**, *155*, 48–54.
12. Chen, L.; Eaton, W.W.; Gallo, J.J.; Nestadt, G. Understanding the heterogeneity of depression through the triad of symptoms, course and risk factors: A longitudinal, population-based study. *J. Affect. Disord* **2000**, *59*, 1–11.
13. Sullivan, P.F.; Prescott, C.A.; Kendler, K.S. The subtypes of major depression in a twin registry. *J. Affect. Disord.* **2002**, *68*, 273–284.
14. Draper, B.; Brodaty, H.; Low, L.F. Types of nursing home residents with self-destructive behaviours: Analysis of the harmful behaviors scale. *Int. J. Geriatr. Psychiatry* **2002**, *17*, 670–675.
15. Lyketsos, C.G.; Sheppard, J.M.; Steinberg, M.; Tschanz, J.A.; Norton, M.C.; Steffens, D.C.; Breitner, J.C. Neuropsychiatric disturbance is Alzheimer's disease clusters into three groups: The cache county study. *Int. J. Geriatr. Psychiatry* **2001**, *16*, 1043–1045.
16. Sham, P.C.; Castle, D.J.; Wessely, S.; Farmer, A.E.; Murray, R.M. Further exploration of a latent class typology of schizophrenia. *Schizophr. Res.* **1996**, *20*, 105–115.
17. Wadsworth, M.E.; Hudziak, J.J.; Heath, A.C.; Achenbach, T.M. Latent class analysis of child behavior checklist anxiety/depression in children and adolescents. *J. Am. Acad. Child Adolesc. Psych.* **2001**, *40*, 106–114.
18. Sullivan, P.C.; Kendler, K.S. Typology of common psychiatric syndromes: An empirical study. *Br. J. Psychiatry* **1998**, *173*, 312–319.
19. Kendler, K.S.; Karkowski, L.M.; Walsh, D. The structure of psychosis: Latent class analysis of probands from the roscommon family study. *Arch. Gen. Psychiatry* **1998**, *55*, 492–499.
20. Bandeen-Roche, K.; Miglioretti, D.; Zeger, S.L.; Rathouz, P. Latent variable regression for multiple discrete outcomes. *J. Am. Stat. Assoc.* **1997**, *92*, 1375–1386.
21. Eaves, L.J.; Silberg, J.L.; Hewitt, J.K.; Rutter, M.; Meyer, J.M.; Neale, M.C.; Pickles, A. Analyzing twin resemblance in multisymptom data: Genetic applications of a latent class model for symptoms of conduct disorder in juvenile boys. *Behav. Genet.* **1993**, *23*, 5–19.
22. Forbes, J.F.; Pickering, R.M. Development of a neonatal case-mix classification. *Med. Care* **1988**, *26*, 1033–1045.
23. Baker, D.; Taylor, H. Inequality in health and health service use for mothers of young children in south west England: Survey team of the avon longitudinal study of pregnancy and child team. *J. Epidemiol. Community Health* **1997**, *51*, 74–79.
24. Bandeen-Roche, K.; Munoz, B.; Tielsch, J.M.; West, S.K.; Schein, O.D. Self-reported assessment of dry eye in a population-based setting. *Invest. Ophthalmol. Vis. Sci.* **1997**, *38*, 2469–2475.
25. Sullivan, P.F.; Smith, W.; Buchwald, D. Latent class analysis of symptoms associated with chronic fatigue syndrome and fibromyalgia. *Psychol. Med.* **2002**, *32*, 881–888.
26. Sullivan, P.F.; Kovalenko, P.; York, T.P.; Prescott, C.A.; Kendler, K.S. Fatigue in a community sample of twins. *Psychol. Med.* **2003**, *33*, 197–201.
27. vanPutten, W.L.; deVries, W.; Reinders, P.; Levering, W.; van der Linden, R.; Tanke, H.J.; Bolhuis, R.L.; Gratama, J.W. Quantification of fluorescence properties of lymphocytes in peripheral blood mononuclear cell suspension using a latent class model. *Cytometry* **1993**, *14*, 86–96.
28. Young, M.A. Evaluating diagnostic criteria: A latent class paradigm. *J. Psychiatr. Res.* **1982–1983**, *17*, 285–296.
29. Taylor, M.A.; Abrams, R. The prevalence of schizophrenia: A reassessment using modern diagnostic criteria. *Am. J. Psychiatry* **1978**, *135*, 945–948.
30. Kendler, K.S.; Eaves, L.J.; Walters, E.E.; Neale, M.C.; Heath, A.C.; Kessler, R.R. The identification and validation of distinct depressive syndromes in a population-based sample of female twins. *Arch. Gen. Psychiatry* **1996**, *53*, 391–399.
31. Sullivan, P.F.; Kessler, R.C.; Kendler, K.S. Latent class analysis of lifetime depressive symptoms in the national comorbidity survey. *Am. J. Psychiatr.* **1998**, *155*, 1398–1406.
32. Eaton, W.W.; Reiger, D.A.; Locke, B.Z. The NIMH epidemiologic catchment area program. *Public Health Rep.* **1981**, *96*, 319–325.
33. Robins, L.N.; Helzer, J.E.; Orvaschel, H. The Diagnostic Interview Schedule. In *Epidemiologic Field Methods in Psychiatry: The NIMH Epidemiologic Catchment Area Program*; Academic Press: New York, 1985.
34. Robins, L.N.; Helzer, J.E.; Croughan, J.; Ratcliff, K.S. National Institute of Mental Health diagnostic interview schedule: Its history, characteristics, and validity. *Arch. Gen. Psychiatry* **1981**, *38*, 381–389.
35. World Health Organization (WHO). *The ICD-10 Classification of Mental and Behavioral Disorders: Clinical Descriptions and Diagnostic Guidelines*; WHO: Geneva, 1992.
36. American Psychiatric Association (APA). *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)*, 4th ed.; APA: Washington, DC, 1994.

37. Goodman, L. Exploratory latent structure analysis using both identifiable and unidentified models. *Biometrika* **1974**, *61*, 215–231.
38. Lindley, D.V. *Bayesian Statistics: A Review*; SIAM, 1971.
39. Muthen, L.K.; Muthen, B.O. How to use a Monte Carlo study to decide on sample size and determine power. *Struct. Equ. Modeling* **2002**, *4*, 599–620.
40. Muthen, L.K.; Muthen, B.O. *Mplus, Version 2.01: The Comprehensive Modeling Program for Applied Researchers*; Muthen & Muthen: Los Angeles, CA, 2001.
41. Uebersax, J. 2003. <http://ourworld.compuserve.com/homepages/jsuebersax/soft.htm>
42. Casella, G.; George, E.I. Explaining the Gibbs sampler. *Am. Stat.* **1992**, *46*, 167–174.
43. Geman, S.; Geman, D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, *6*, 721–741.
44. Phillips, D.B.; Smith, A.F. M. Bayesian Model Comparison via Jump Diffusions. In *Markov Chain Monte Carlo in Practice*; Gilks, W.R.; Richardson, S.; David, S., Eds.; Chapman and Hall: London, 1996, 215–250.
45. Robert, C.P. Mixtures of Distributions: Inference and Estimation. *Markov Chain Monte Carlo in Practice*; Chapman and Hall: London, 1996, 441–464.
46. Roeder, K.; Wasserman, L. Practical Bayesian density estimation using mixtures of normals. *J. Am. Stat. Assoc.* **1997**, *92*, 894–902.
47. Diebolt, J.; Robert, C.P. Estimation of finite mixture distributions through Bayesian sampling. *J. R. Stat. Soc., Ser. B* **1994**, *56*, 363–375.
48. Tanner, M.A.; Wong, W.H. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* **1987**, *82*, 528–550.
49. Read, T.R.C.; Cressie, N.A.C. *Goodness of Fit Statistics for Discrete Multivariate Analysis*; Springer: New York, 1988.
50. Collins, L.M.; Fidler, P.L.; Wugalter, S.E.; Long, J.D. Goodness-of-fit testing for latent class models. *Multivar. Behav. Res.* **1993**, *28*, 375–389.
51. Everitt, B.S. A Monte Carlo investigation of the likelihood ratio test for the number of classes in a latent class analysis. *Multivar. Behav. Res.* **1988**, *23*, 531–538.
52. Aitken, M.; Anderson, D.; Hinde, J. Statistical modelling of data on teaching styles. *J. R. Stat. Soc. Ser. A* **1981**, *144*, 419–461.
53. Akaike, H. Likelihood of a model and information criteria. *J. Econom.* **1981**, *16*, 3–14.
54. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464.
55. Stone, M. Comments on the model selection criteria of Akaike and Schwarz. *J. R. Stat. Soc., Ser. B* **1979**, *41*, 276–278.
56. Raftery, A.E. Choosing models for cross-classification. *Am. Sociol. Rev.* **1986**, *51*, 145–146.
57. Nylund, K.L.; Asparouhov, T.; Muthen, B. Deciding on the number of classes in latent class analysis and growth mixture modeling. A Monte Carlo simulation study. *Struct. Equ. Modeling* **2007**, *14*, 535–569.
58. Noreen, E.W. *Computer Intensive Methods for Testing Hypotheses*; Wiley: New York, 1989.
59. Formann, A.K.; Kohlmann, T. Latent class analysis in medical research. *Stat. Methods Med. Res.* **1996**, *5*, 179–211.
60. Lo, Y.; Mendell, N.R.; Rubin, D.B. Testing the number of components in a normal mixture. *Biometrika* **2001**, *88* (3), 767–778.
61. Jeffries, N.O. A note on “Testing the number of components in a normal mixture”. *Biometrika* **2003**, *90* (4), 991–994.
62. Kass, R.E.; Raftery, A.E. Bayes factors. *J. Am. Stat. Assoc.* **1995**, *90*, 773–795.
63. Kass, R.E.; Wasserman, L. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Stat. Assoc.* **1995**, *90*, 928–934.
64. Aitkin, M. Posterior Bayes factors. *J. R. Stat. Soc., Ser. B* **1991**, *53*, 15–25.
65. Smith, A.F.M.; Spiegelhalter, D.J. Bayes factor and choice criteria for linear models. *J. R. Stat. Soc., Ser. B* **1980**, *42*, 213–220.
66. Mulaik, S.A.; James, L.R.; Van Alstine, J.; Bennett, N.; Lind, S.; Stillwell, C.D. Evaluation of goodness of fit indices for structural equation models. *Psychol. Bull.* **1989**, *105*, 430–445.
67. Reiser, M.; Lin, Y. A Goodness-of-Fit Test for the Latent Class Model When Expected Frequencies Are Small. *Sociological Methodology 1999*; Blackwell: Oxford, 1999, 81–111.
68. Lindsey, B.G.; Rudas, T.; Clogg, C.C. A new index based on mixture models for the analysis of contingency tables. *J. R. Stat. Soc. Ser. B* **1994**, *56*, 623–629.
69. Dayton, C.M.; Macready, G.B. Concomitant-variable latent-class models. *J. Am. Stat. Assoc.* **1988**, *83*, 173–178.
70. Miglioretti, D. Latent transition regression for mixed outcomes. *Biometrics* **2003**, *59*, 710–720.
71. Muthén, B.; Brown, C.H.; Masyn, K.; Jo, B.; Khoo, S.T.; Yang, C.C.; Wang, C.P.; Kellam, S.; Carlin, J.; Liao, J. General growth mixture modeling for randomized preventive interventions. *Biostatistics* **2002**, *3*, 459–475.

Lilly Reference Ranges

Michael G. Wilson

Butler University, Indianapolis, Indiana, U.S.A.

Abstract

The Lilly Reference Ranges are updated sets of nonparametric reference intervals and delta limits for routine clinical laboratory analyses designed specifically for use in clinical trials. This entry provides several tables and examples to give a detailed illustration of how this type of analysis works.

INTRODUCTION

The Lilly Reference Ranges are updated sets of nonparametric reference intervals and delta limits for routine clinical laboratory analytes designed specifically for use in clinical trials. There are 40 sets of adult reference intervals, 5 sets for each of 8 demographic subgroups, for 39 routine clinical laboratory analytes.^[1] The most recent update of the Lilly Reference Ranges is based on a sample of 20,201 adult patients who were being evaluated for entry into a clinical trial. The sets of reference intervals describe the between patient variability while the delta limits summarize the within patient variability for 39 routine clinical laboratory measurements. Here we present the Lilly Reference Ranges and methods for their construction.

OVERVIEW

In clinical laboratory medicine, reference ranges provide a point of comparison. The interpretation of specific observed laboratory test result is a comparative decision-making process. “Health-associated” reference values are most frequently used to screen patients for the presence of disease, to confirm a diagnosis, and to monitor disease progression. However, the Lilly Reference Ranges are used for very different purposes in the clinical trials setting. For this reason, there are sharp contrasts between “health-associated” reference values and the Lilly Reference Ranges.

In a parallel fashion, Thompson et al.^[2] suggested that three major decisions could be made using observed routine clinical laboratory results in clinical trials. Firstly, patients need to be screened for inclusion in clinical trials. Secondly, patients need to be protected during clinical trials. Thirdly, drugs effects need to be assessed during and after the clinical trial.

The two most striking differences between “health-associated” reference intervals and the Lilly Reference Ranges are that they are narrower and wider, respectively,

and that the reference populations are healthy and clinical trial participants, respectively. However sharp these contrasts are, reference intervals share a similar protocol for their definition and determination.^[3]

METHODS

Thompson et al.^[4] published the methodology for determination of Lilly Reference Ranges in 1987. Briefly, the reference sample was selected by predetermining inclusion and exclusion criteria for all patients entering a clinical trial. The 39 most frequently sampled routine analytes were selected for definition. Written consent was obtained from study participants. Specimens were prepared and collected properly and consistently. Reference value data were inspected to evaluate the distribution. Possible outliers were identified. Methods of estimating reference limits were selected. Reference values were analyzed. Populations were subcategorized for separate reference intervals. These steps are discussed in detail below.

Also in 1987, the approved recommendations on the theory of reference values that provided definitions, principles, and procedures for the determination and use of reference values were published by the Expert Panel on the Theory of Reference Values (EPTRV) of the International Federation of Clinical Chemistry and the Standing Committee on Reference Values of the International Council for Standardization in Hematology. These definitions, principles, and procedures provided a rational approach and sound basis for the determination of reference values.^[5–10]

Subsequently, approved guidelines for the definition and determination of reference ranges were published by the National Committee for Clinical Laboratory Standards (NCCLS) in 1995.^[11] The work of the EPTRV proved to be a most useful basis for the development of the NCCLS guidelines. It was the intent of the subcommittee to publish guidelines for determining “health-associated” reference values. Basically, the a posteriori method as described

by the NCCLS protocol guideline was the methodology employed to construct the Lilly Reference Ranges.

Selection of Reference Sample Group

The reference population is a group consisting of all possible reference individuals. The reference sample group is composed of an adequate number of reference individuals who are selected on the basis of well-defined criteria. The researchers wish to draw inferences from the sample and generalize them to the population.

During the selection of reference individuals, careful consideration should be given to the reference population to which the study is attempting to draw an inference. Unlike, “health-associated” reference intervals, which are very useful diagnostically, Lilly Reference Ranges are used for a very different purpose. Clinical trial researchers dealing with routine clinical laboratory measurements are most interested in testing hypotheses concerning the safety of an experimental compound relative to a control compound. The population for which the researchers are attempting to draw an inference is clinical trial participants. This population generally exhibits greater variability in selected routine laboratory parameters than noted in populations of disease-free individuals. Therefore, Thompson et al. included all clinical trial participants, even if they were not subsequently randomized, with an available reference value in the reference sample group.

Reference values are defined to be those values, or test results, obtained from a reference sample group. The empirical distribution of reference values is defined to be the reference distribution and the summary statistics from the reference distribution used to describe the reference distribution is called a reference limit. These reference limits are commonly order statistics. The interval between, and including, the two reference limits is the reference interval. Few reference intervals should be defined as reference ranges. The term “range” should be reserved for describing a set of values defined by the actual minimal and maximal measured values. Lilly Reference Ranges are defined as such because no trimming or censoring of the distribution occurred and because the minimal and maximal values are available for use.

Potential reference values were only those values for reference individuals that were sampled during qualification for enrollment into a clinical trial. These values were the screening and baseline values for the clinical trial that the reference individuals were screened. Analytes that were used as efficacy variables and not routine laboratory safety variables were excluded as reference values. For example, patients in diabetic studies did not contribute glucose values.

Patients were then to be partitioned into 32 subclasses dichotomizing gender, race, age, smoking, and drinking status. Results for glucose values were to be partitioned into fasting and nonfasting status.

The a posteriori sampling process was used for the Lilly Reference Ranges because the process of exclusion and partitioning took place after analyte testing.

Selection of Sample Size and Statistical Methods for Estimating Reference Limits

In general, reference limits can be determined using parametric or nonparametric methods. Lilly Reference Ranges were determined nonparametrically. There were several considerations for this decision. Nonparametric estimation does not require any assumptions about the distribution for a given analyte. Also, it is a methodology that could be applied consistently for all analytes. It is simple to implement and is readily understandable. The disadvantage of using nonparametric estimation is that when the reference distribution is normal Gaussian, the order statistics are slightly biased and more variable. However, large sample sizes were anticipated for the Lilly Reference Ranges and few distributions were thought to be normally distributed.

The reference value data was inspected to evaluate the reference distribution for each analyte and to identify possible outliers. Because an important implicit assumption is that all routine safety reference values come from the same distribution, none were excluded from the reference distribution or subsequent calculations of the Lilly Reference Ranges regardless of how aberrant they appeared.

However, the primary Lilly Reference Ranges was the 98% interval defined by the 1st and 99th percentiles. The 95% and 90% intervals were also calculated.

Delta Limits

All patients who had repeated measurements during their qualification period were included in the determination of the delta values. These patients had both screening and baseline values. These samples were drawn at an interval of 1 or 2 weeks. This interval was typically associated with the placebo baseline or washout period for their clinical trial.

The difference between the screening and baseline reference values was calculated. A distribution of these differences was defined to be the delta distribution. These differences express the within-patient variability. The same methods for the determination of the reference intervals were applied to these difference values for the creation of the delta limits.

Delta limits can be very useful to the researcher. They provide a guideline for the amount of change that can be expected in a given analyte. This guideline can then be used in the interpretation of the observed change due to an experimental compound for a given patient.

Procedures for validating the reference ranges and delta limits as well as calculating the tolerance intervals were also given by Thompson et al.

Appendix A Eli Lilly and Company Adult Reference Ranges in Clinical Trials for Routine Laboratory Test Results ($n = 20,102$)^a

Analyte	Age	Male				Female				Delta limits	
		Caucasian		Non-Caucasian		Caucasian		Non-Caucasian			
		<50	≥50	<50	≥50	<50	≥50	<50	≥50	Low	High
Liver											
ALT	U/L	7–125	5–95	5–121	4–84	5–80	5–80	4–82	4–82	–27	28
AST	U/L	11–79	10–78	11–122	8–84	9–62	9–67	9–66	9–65	–25	22
GGT	U/L	6–194	6–193	7–320	5–274	5–127	5–201	5–161	5–190	–30	29
Alk. phosphatase	U/L	30–153	28–163	32–151	27–174	28–131	31–168	31–138	33–174	–44	28
Bilirubin	μmol/L	3–31	3–29	3–29	3–22	3–22	3–19	2–21	2–19	–10	10
Muscle											
Creatine kinase	U/L	17–640	11–426	27–1105	15–940	19–265	0–267	20–453	10–463	–277	236
Kidney											
Creatinine	μmol/L	62–141	71–168	71–168	71–186	53–124	53–150	53–124	53–150	–35	35
Urea	mmol/L	2.5–8.9	2.9–11.1	2.1–8.9	2.1–12.5	1.8–7.9	2.5–11.4	1.8–7.5	2.5–10.7	–3.2	3.2
Uric acid	μmol/L	167–547	161–583	184–595	208–625	89–470	113–535	107–482	143–619	–125	119
Phosphate(i)	mmol/L	0.72–1.61	0.68–1.49	0.77–1.65	0.68–1.58	0.74–1.65	0.74–1.58	0.74–1.52	0.74–1.61	–0.45	0.42
Calcium	mmol/L	2.07–2.64	2.02–2.62	2.05–2.64	2.05–2.64	2.02–2.62	2.05–2.64	2.02–2.64	2.02–2.69	–0.30	0.30
Electrolytes											
Sodium	mmol/L	135–147	133–148	135–150	134–149	135–147	134–150	135–150	134–152	–8	8
Potassium	mmol/L	3.5–5.5	3.3–5.4	3.0–5.5	3.0–5.4	3.4–5.3	3–3–5.5	3.1–5.3	3.0–5.3	–1.1	1.0
Chloride	mmol/L	96–113	95–112	96–110	95–112	96–114	94–113	98–112	93–111	–9	8
Bicarbonate	mmol/L	21–35	21–35	21–36	21–36	20–33	21–36	20–35	21–36	–7	8
Nutritional											
Glucose, fasting	mmol/L	3.7–10.8	3.9–14.2	3.8–9.5	3.9–16.8	3.7–9.2	3.7–15.0	3.8–15.3	3.9–20.3	–3.1	3.2
Glucose, nonfasting	mmol/L	3.4–9.8	3.8–14.4	3.4–10.8	4.2–20.3	3.6–11.2	4.3–15.2	3.8–14.7	4.2–18.4	–4.8	4.2
Albumin	g/L	36–52	32–50	34–53	32–49	34–50	32–49	33–49	32–50	–7	6
Protein	g/L	62–83	60–82	63–89	61–87	61–82	60–82	63–86	63–86	–11	10
Cholesterol	mmol/L	2.84–8.56	3.41–8.51	2.77–8.46	2.84–8.95	3.10–7.73	3.65–9.10	3.10–7.73	3.78–9.36	–1.78	1.73
HDL cholesterol	mmol/L	0.44–2.02	0.47–2.17	0.47–2.48	0.41–2.48	0.59–2.64	0.57–2.56	0.52–2.51	0.52–2.64	–0.85	0.72
LDL cholesterol	mmol/L	1.50–6.13	1.68–6.08	1.22–6.26	1.42–6.43	1.37–5.46	1.78–6.23	1.16–5.82	1.81–6.813	–1.97	1.78
Triglycerides	mmol/L	0.50–9.53	0.53–8.69	0.45–8.99	0.51–5.05	0.38–6.93	0.58–6.21	0.44–5.59	0.52–4.41	–3.01	2.88

(Continued)

Appendix A Eli Lilly and Company Adult Reference Ranges in Clinical Trials for Routine Laboratory Test Results (*n* = 20,102)^a (Continued)

Analyte	Age	Male				Female				Delta limits	
		Caucasian		Non-Caucasian		Caucasian		Non-Caucasian			
		<50	≥50	<50	≥50	<50	≥50	<50	≥50	Low	High
<i>Erythrocytes</i>											
Hemoglobin (Fe)	mmol/L	7.90–11.29	7.10–11.36	6.83–11.23	6.60–11.00	6.80–10.18	6.60–10.36	6.14–10.10	6.21–10.10	–1.20	1.20
Erythrocytes	TU/L	4.2–6.2	3.9–6.2	3.9–6.4	3.7–6.4	3.7–5.5	3.6–5.7	3.6–5.7	3.5–5.9	–0.7	0.7
Hematocrit	l	0.38–0.55	0.35–0.55	0.35–0.55	0.33–0.54	0.33–0.49	0.33–0.51	0.31–0.49	0.33–0.49	–0.07	0.06
MCV	fL	78–105	77–107	71–104	74–104	76–103	77–105	67–102	73–102	–9	8
MCH (Fe)	fmol	1.60–2.17	1.60–2.20	1.40–2.17	1.40–2.11	1.50–2.10	1.60–2.11	1.30–2.05	1.40–2.05	–0.19	0.20
MCHC (Fe)	mmol/L	19–23	19–23	18–23	18–22	19–23	19–23	18–22	18–22	–2	2
<i>Leukocytes</i>											
Leukoocytes	GI/L	3.6–13.8	3.4–12.8	3.0–13.0	2.9–11.6	3.5–13.4	3.6–12.4	3.2–13.6	3.1–11.4	–4.0	3.7
Bands	GI/L	0–0.62	0–0.54	0–0.27	0–0.32	0–0.75	0–0.55	0–0.37	0–0.20	–0.29	0.26
Neutrophils	GI/L	1.83–9.29	1.82–8.74	1.17–9.24	1.29–7.81	1.70–8.99	1.79–8.52	1.19–8.66	1.31–7.87	–3.38	3.18
Lymphocytes	GI/L	0.88–4.49	0.81–4.32	0.93–4.95	0.82–4–28	0.97–4.20	0.90–4.30	0.94–4.51	0.95–4.47	–1.57	1.54
Monocytes	GI/L	0–0.89	0–0.96	0–0.81	0–0.85	0–0.82	0–0.85	0–0.83	0–0.75	–0.44	0.46
Eosinophils	GI/L	0–0.65	0–0.66	0–0.74	0–0.59	0–0.66	0–0.58	0–0.62	0–0.55	–0.38	0.37
Basophils	GI/L	0–0.18	0–0.18	0–0.15	0–0.14	0–0.17	0–0.17	0–0.16	0–0.17	–0.13	0.13
Platelets	GI/L	136–425	130–483	131–486	123–518	151–501	141–513	144–541	114–559	–93	107
<i>Urine</i>											
Specific gravity		1.006–1.037	1.006–1.035	1.007–1.036	1.006–1.036	1.004–1.036	1.005–1.036	1.005–1.038	1.006–1.037	–0.018	0.017
pH		5.0–8.0	5.0–8.0	5.0–8.0	5.0–8.0	5.0–8.0	5.0–8.0	5.0–8.0	5.0–8.0	–2.0	2.0

^aThe number of adult patients with at least one analyte result is 20,102.

Appendix B Eli Lilly and Company Adult Reference Ranges in Clinical Trials for Routine Laboratory Test Results Smoker/Imbiber (n = 2974)^a

Analyte	Age	Male				Female					
		Caucasian		Non-Caucasian		Caucasian		Non-Caucasian			
		<50	≥50	<50	≥50	<50	≥50	<50	≥50		
Liver											
ALT	U/L	5–143	6–114	6–109	3–77	5–63	7–62	4–112	6–57	–27	28
AST	U/L	11–111	9–104	11–140	11–114	9–55	10–80	8–81	9–62	–25	22
GGT	U/L	8–230	5–167	7–390	6–460	6–168	5–148	7–111	7–98	–30	29
Alk. phosphatase	U/L	31–144	32–137	35–140	36–155	27–136	36–142	35–145	31–235	–44	28
Bilirubin	μmol/L	3–29	3–26	3–29	3–22	3–19	2–17	2–17	2–19	–10	10
Muscle											
Creatine kinase	U/L	16–615	12–474	27–2770	18–600	20–218	0–308	26–316	0–389	–277	236
Kidney											
Creatinine	μmol/L	62–133	71–141	71–168	62–186	53–124	53–115	53–115	53–159	–35	35
Urea	mmol/L	2.1–8.9	2.5–9.6	2.1–8.6	1.4–12.5	1.4–7.5	1.8–8.6	2.0–7.9	2.1–8.2	–3.2	3.2
Uric acid	μmol/L	172–547	167–553	178–607	220–571	95–440	143–476	143–476	143–648	–125	119
Phosphate(i)	mmol/L	0.74–1.58	0.74–1.45	0.77–1.61	0.71–1.71	0.81–1.65	0.84–1.58	0.81–1.65	0.61–1.71	–0.45	0.42
Calcium	mmol/L	2.07–2.62	2.05–2.57	2.05–2.64	2.10–2.69	2.02–2.62	2.05–2.64	2.07–2.62	1.90–2.60	–0.30	0.30
Electrolytes											
Sodium	mmol/L	135–147	134–146	136–149	135–153	135–150	134–151	135–150	128–149	–8	8
Potassium	mmol/L	3.6–5.7	3.4–5.4	3.5–5.3	2.8–5.4	3.5–5.3	3.4–5.5	3.4–6.8	2.9–4.7	–1.1	1.0
Chloride	mmol/L	97–115	96–112	96–110	97–114	98–116	95–117	99–112	90–109	–9	8
Bicarbonate	mmol/L	20–34	21–35	21–32	22–38	19–32	22–36	19–35	20–35	–7	8
Nutritional											
Glucose, fasting	mmol/L	3.6–9.9	3.8–12.1	3.6–8.4	4.1–15.1	3.7–8.0	3.8–8.2	3.9–7.8	4.4–23.3	–3.1	3.2
Glucose, nonfasting	mmol/L	2.2–9.8	2.8–11.9	4.1–8.0	4.2–14.5	0.5–11.2	4.4–13.7	0.9–7.4	4.6–18.4	–4.8	4.2
Albumin	g/L	35–52	33–50	33–51	33–51	34–51	35–50	33–53	32–52	–7	6
Protein	g/L	62–84	61–83	63–88	59–87	60–82	61–80	64–93	60–88	–11	10
Cholesterol	mmol/L	2.97–8.84	3.59–8.56	2.61–8.67	2.92–8.61	3.03–8.25	3.70–9.00	3.54–8.84	3.82–11.12	–1.78	1.73
HDL cholesterol	mmol/L	0.52–2.02	0.42–2.17	0.59–2.64	0.44–2.48	0.70–2.38	0.44–3.10	0.52–2.35	0.08–2.64	–0.85	0.72
LDL cholesterol	mmol/L	1.45–6.15	1.63–6.10	0.75–6.80	1.42–7.16	127–5.46	1.60–5.59	1.06–6.13	2.15–6.15	–1.97	1.78
Triglycerides	mmol/L	0.58–8.24	0.52–8.74	0.53–7.95	0.51–4.14	0.42–5.61	0.51–9.04	0.41–7.69	0.44–3.51	–3.01	2.88

(Continued)

Appendix B Eli Lilly and Company Adult Reference Ranges in Clinical Trials for Routine Laboratory Test Results Smoker/Imbiber (*n* = 2974)^a (Continued)

Analyte	Age	Male				Female							
		Caucasian		Non-Caucasian		Caucasian		Non-Caucasian					
		<50	≥50	<50	≥50	<50	≥50	<50	≥50				
<i>Erythrocytes</i>													
Hemoglobin (Fe)	mmol/L	8.10–11.61	7.60–11.79	6.80–11.42	6.95–11.48	7.02–10.40	7.50–10.98	6.39–10.30	5.80–11.11	–1.20	120		
Erythrocytes	TI/L	4.2–6.1	3.6–6.1	3.6–6.2	3.4–6.9	3.7–5.5	4.0–5.9	3.7–5.6	3.6–6.0	–0.7	0.7		
Hematocrit	l	0.39–0.58	0.37–0.56	0.33–0.56	0.34–0.57	0.34–0.50	0.35–0.53	0.33–0.52	0.31–0.53	–0.07	–0.06		
MCV	fL	78–109	80–111	72–109	75–107	76–105	79–107	71–102	78–104	–9	8		
MCH (Fe)	fmol	1.60–2.23	1.68–2.40	1.40–2.23	1.49–2.17	1.60–2.20	1.70–2.17	1.37–2.05	1.49–2.11	–0.19	0.20		
MCHC (Fe)	mmol/L	19–23	19–23	10–23	19–23	19–23	19–22	19–23	19–22	–2	2		
<i>Leukocytes</i>													
Leukoeyles	GI/L	3.9–14.3	3.6–12.9	3.0–12.9	3.2–11.5	3.6–14.6	4.4–13.3	3.3–14.4	3.1–11.4	–4.0	3.7		
Bands	GI/L	0–0.77	0–0.78	0–0.44	0–0.19	0–1.15	0–1.71	0–0.72	0–0.14	–0.29	0.26		
Neutrophils	GI/L	1.96–9.91	1.84–9.05	0.90–9.15	1.44–8.01	1.71–10.34	2.26–8.79	1.41–8.66	1.60–8.59	–3.38	3.18		
Lymphocytes	GI/L	0.89–4.80	0. and 5–4.15	0.93–4.66	0.75–4.65	0.97–4.47	0.90–4.79	1.01–4.88	0.78–4.79	–1.57	1.54		
Monocytes	GI/L	0–0.95	0–1.13	0–0.81	0–1.00	0–0.83	0–0.93	0–0.84	0–0.66	–0.44	0.46		
Eosinophils	GI/L	0–0.77	0–0.66	0–0.61	0–0.83	0–0.69	0–0.49	0–0.47	0–0.54	–0.38	0.37		
Basophils	GI/L	0–0.18	0–0.18	0–0.14	0–0.16	0–0.18	0–0.19	0–0.18	0–0.18	–0.13	0.13		
Platelets	GI/L	142–424	149–432	148–377	105–574	159–490	121–475	109–489	114–450	–93	107		
<i>Urine</i>													
Specific gravity		1.006– 1.035	1.005– 1.034	1.007– 1.036	1.005– 1.037	1.004– 1.035	1.004– 1.036	1.005– 1.034	1.005– 1.038	–0.018	0.017		
pH		5–9	5–8	5–9	5–7	5–8	5–8	5–7	5–7	–2.0	2.0		

^aThe number of adult patients with at least one analyte result is 2974.

Appendix C Eli Lilly and Company Adult Reference Ranges in Clinical Trials for Routine Laboratory Test Results Smoker/Nondrinker (n = 2037)^a

Analyte	Age	Male				Female					
		Caucasian		Non-Caucasian		Caucasian		Non-Caucasian			
		<50	≥50	<50	≥50	<50	≥50	<50	≥50		
Liver											
ALT	U/L	5–93	4–72	5–95	3–96	4–63	5–61	3–228	4–92	–27	28
AST	U/L	10–60	9–48	11–60	7–84	8–55	10–60	9–139	10–71	–25	22
GGT	U/L	8–194	6–94	8–207	6–101	4–252	7–189	5–181	5–190	–30	29
Alk. phosphatase	U/L	34–138	38–166	36–144	27–194	30–133	39–211	32–121	39–174	–44	28
Bilirubin	μmol/L	3–24	3–21	3–22	3–21	3–17	3–15	2–22	3–15	–10	10
Muscle											
Creatine kinase	U/L	18–670	15–281	30–1481	17–788	0–218	0–297	15–246	0–328	–277	236
Kidney											
Creatinine	μmol/L	53–141	62–150	53–141	71–194	53–115	53–133	53–124	62–124	–35	35
Urea	mmol/L	2.5–9.6	2.5–10.7	2.1–9.3	2.5–12.9	1.8–7.5	2.1–9.6	1.8–7.1	2.5–9.3	–3.2	3.2
Uric acid	μmol/L	167–512	119–541	208–571	190–630	83–470	119–547	50–494	155–529	–125	119
Phosphate(i)	mmol/L	0.68–1.58	0.68–1.49	0.81–1.58	0.61–1.58	0.77–1.58	0.90–1.52	0.77–1.49	0.81–1.52	–0.45	0.42
Calcium	mmol/L	2.10–2.67	2.02–2.64	2.05–2.70	2.10–2.72	2.05–2.60	2.10–2.64	2.05–2.64	1.67–2.64	–0.30	0.30
Electrolytes											
Sodium	mmol/L	135–148	134–146	135–148	125–149	135–150	134–153	134–151	135–149	–8	8
Potassium	mmol/L	3.3–5.3	3.1–5.2	3.0–5.3	3.1–5.2	3.4–5.4	2.8–5.7	3.2–4.9	3.1–5.3	–1.1	1.0
Chloride	mmol/L	94–108	96–112	98–109	94–111	96–113	93–112	99–112	95–107	–9	8
Bicarbonate	mmol/L	19–36	22–34	21–35	22–34	19–31	21–34	21–35	22–34	–7	8
Nutritional											
Glucose, fasting	mmol/L	3.4–8.3	4.2–15.4	2.8–18.9	3.7–25.1	3.6–9.4	3.7–14.4	4.1–10.3	3.9–20.3	–3.1	3.2
Glucose, nonfasting	mmol/L	4.2–9.0	4.3–22.2	0.7–16.5	4.0–16.7	3.9–21.6	4.5–18.8	3.8–10.7	4.8–13.1	–4.8	4.2
Albumin	g/L	34–53	32–49	36–51	34–47	34–51	33–49	35–49	33–50	–7	6
Protein	g/L	61–83	61–81	63–86	63–89	61–81	59–81	62–84	62–85	–11	10
Cholesterol	mmol/L	2.87–8.77	3.54–8.56	2.74–8.74	2.43–8.90	3.21–7.60	4.03–9.39	3.28–7.32	3.70–9.31	–1.78	1.73
HDL cholesterol	mmol/L	0.26–1.78	0.36–2.02	0.52–1.76	0.41–2.80	0.47–2.22	0.54–2.30	0.54–2.30	0.67–2.69	–0.85	0.72
LDL cholesterol	mmol/L	1.81–5.72	1.94–6.71	122–5.99	0.70–5.38	1.89–8.04	1.97–6.15	1.16–5.48	2.25–6.83	–1.97	1.78
Triglycerides	mmol/L	0.72–12.78	0.63–9.03	0.43–3.62	0.45–11.49	0.53–6.25	0.87–15.72	0.53–3.82	0.75–4.30	–3.01	2.88

(Continued)

Appendix C Eli Lilly and Company Adult Reference Ranges in Clinical Trials for Routine Laboratory Test Results Smoker/Nonsmoker (*n* = 2037)^a (Continued)

Analyte	Age	Male				Female				Delta limits	
		Caucasian		Non-Caucasian		Caucasian		Non-Caucasian			
		<50	≥50	<50	≥50	<50	≥50	<50	≥50	Low	High
<i>Erythrocytes</i>											
Hemoglobin (Fe)	mmol/L	8.10–11.61	7.00–11–36	7.57–11.23	6.27–11.61	7.10–10.60	6.90–10.43	6.45–10.80	6.60–10.18	–1.20	1.20
Erythrocytes	TI/L	4.3–6.4	3.9–6.2	4.0–6.4	3.9–6.6	3.6–5.7	3.8–5.9	3.5–6.0	3.8–5.8	–0.7	0.7
Hematocrit	l	0.39–0.56	0.35–0.55	0.37–0.56	0.33–0.56	0.34–0.52	0.36–0.52	0.32–0.54	0.33–0.50	–0.07	–0.06
MCV	fL	78–104	75–106	74–99	73–107	78–105	81–105	70–104	79–103	–9	8
MCH (Fe)	fmol	1.60–2.10	1.40–2.11	1.60–2.05	1.30–2.10	1.60–2.20	1.70–2.11	1.4G–2.17	1.49–2.20	–0.19	0.20
MCHC (Fe)	mmol/L	19–23	19–23	18–22	18–22	19–23	19–22	18–22	18–22	–2	2
<i>Leukocytes</i>											
Leukoocytes	GI/L	4.1–14.9	3.5–13.5	3.1–15.5	2.9–13.1	4.3–14.1	3.9–13.6	3.5–14.8	3.1–12.5	–4.0	3.7
Bands	GI/L	0–1.03	0–0.71	0–0.54	0–0.38	0–0.78	0–0.58	0–0.39	0–0.20	–0.29	0.26
Neutrophils	GI/L	2.18–10.35	2.05–9.05	1.21–9.66	1.25–9.96	1.96–9.84	1.95–9.80	1.31–10.50	1.85–7.60	–3.38	3.18
Lymphocytes	GI/L	1.09–4.84	1.02–4–54	1.06–6.01	0.84–4.37	1.22–4.71	1.06–4.75	1.22–4.87	0.43–4.72	–1.57	1.54
Monocytes	GI/L	0–0.93	0–0.92	0–1.04	0–0.85	0–0.85	0–0.87	0–0.93	0–0.75	–0.44	0.46
Eosinophils	GI/L	0–0.78	0–0.71	0–0.86	0–0.60	0–0.84	0–0.59	0–0.63	0–0.58	–0.38	0.37
Basophils	GI/L	0–0.23	0–0.19	0–0.14	0–0.14	0–0.19	0–0.21	0–0.15	0–0.15	–0.13	0.13
Platelets	GI/L	136–457	149–586	116–394	158–486	154–499	127–476	184–550	101–480	–93	107
<i>Urine</i>											
Specific gravity		1.006–1.036	1.006–1.034	1.005–1.036	1.006–1.036	1.005–1.036	1.005–1.043	1.005–1.039	1.005–1.032	–0.018	0.017
pH		5–7	5–9	5–8	5–7	5–7	5–8	5–7	5–7	–2.0	2.0

^aThe number of adult patients with at least one analyte result is 2037.

Appendix D Eli Lilly and Company Adult Reference Ranges in Clinical Trials for Routine Laboratory Test Results Smoker/Imbiber ($n = 4265$)^a

Analyte	Age	Male				Female				Delta limits	
		Caucasian		Non-Caucasian		Caucasian		Non-Caucasian			
		<50	≥50	<50	≥50	<50	≥50	<50	≥50	Low	High
Liver											
ALT	U/L	9–102	8–100	6–125	7–101	6–81	7–90	5–52	2–114	–27	28
AST	U/L	11–67	11–83	12–131	12–132	10–70	11–87	8–52	12–106	–25	22
GGT	U/L	6–160	6–200	7–207	9–152	5–117	5–201	5–121	5–113	–30	29
Alk. phosphatase	U/L	25–125	22–124	33–119	22–132	25–115	31–158	29–131	32–192	–44	28
Bilirubin	μmol/L	3–36	3–31	3–29	3–27	3–22	3–22	3–22	3–15	–10	10
Muscle											
Creatine kinase	U/L	20–563	0–427	35–1070	17–2124	25–273	11–275	34–420	10–361	–277	236
Kidney											
Creatinine	μmol/L	71–141	71–159	71–177	71–150	53–124	53–133	53–133	44–133	–35	35
Urea	mmol/L	2.5–9.6	3.2–10.4	2.5–9.3	2.1–10.4	1.8–7.5	2.5–10.4	1.8–7.9	2.9–10.4	–3.2	3.2
Uric acid	μmol/L	178–583	178–583	202–607	190–613	89–446	113–547	131–541	83–613	–125	119
Phosphate(i)	mmol/L	0.71–1.52	0.68–1.49	0.77–1.58	0.65–1.45	0.74–1.58	0.74–1.61	0.74–1.55	0.90–1.68	–0.45	0.42
Calcium	mmol/L	2.10–2.62	2.05–2.62	2.10–2.69	2.10–2.69	2.02–2.62	2.00–2.59	2.05–2.59	1.90–2.64	–0.30	0.30
Electrolytes											
Sodium	mmol/L	134–147	135–148	135–150	133–148	135–146	134–147	136–145	137–153	8	8
Potassium	mmol/L	3.5–5.5	3.3–5.4	3.0–5.5	3.0–6.0	3.2–5.2	3.3–5.3	3.3–5.9	3.1–5.1	–1.1	1.0
Chloride	mmol/L	96–112	97–111	98–110	97–110	95–114	95–113	90–110	90–115	–9	8
Bicarbonate	mmol/L	22–35	22–35	23–36	20–44	21–35	22–36	22–33	22–35	–7	8
Nutritional											
Glucose, fasting	mmol/L	3.7–8.8	4.2–12.5	4.1–8.7	2.9–12.4	3.8–8.1	3.7–18.0	0.9–16.3	2.6–16.6	–3.1	3.2
Glucose, nonfasting	mmol/L	3.4–7.7	1.7–12.1	3.5–13.7	4.8–10.5	4.7–7.7	4.6–13.7	4.8–6.3	4.7–12.0	–4.8	4.2
Albumin	g/L	37–53	36–50	37–53	32–50	34–50	34–50	34–48	33–49	–7	6
Protein	g/L	62–83	61–82	64–86	61–83	62–82	62–83	63–88	59–82	–11	10
Cholesterol	mmol/L	2.95–8.48	3.49–8.43	3.23–8.46	3.41–8.87	3.21–7.47	3.93–9.13	3.59–7.37	3.67–9.36	–1.78	1.73
HDL cholesterol	mmol/L	0.47–2–17	0.47–2.22	0.44–3.03	0.28–2.53	0.72–2.95	0.65–2.74	0.75–2.87	0.78–2.61	–0.85	0.72
LDL cholesterol	mmol/L	1.89–5.96	1.68–6.08	1.55–6.26	1.50–6.49	1.37–5.43	1.84–6.54	1.68–4.76	1.24–7.68	–1.97	1.78
Triglycerides	mmol/L	0.47–10.21	0.53–7.81	0.65–8.99	0.54–8.79	0.29–4.44	0.53–4.90	0.38–6.42	0.52–6.51	–3.01	2.88

(Continued)

Appendix D Eli Lilly and Company Adult Reference Ranges in Clinical Trials for Routine Laboratory Test Results Smoker/Imbiber (*n* = 4265)^a (Continued)

Analyte	Age	Male				Female				Delta limits	
		Caucasian		Non-Caucasian		Caucasian		Non-Caucasian			
		<50	≥50	<50	≥50	<50	≥50	<50	≥50	Low	High
<i>Erythrocytes</i>											
Hemoglobin (Fe)	mmol/L	7.88–10.98	7.32–11.23	7.50–11.11	5.83–10.67	6.83–9.70	7.00–10.18	6.30–9.62	6.33–10.05	–1.20	1.20
Erythrocytes	TI/L	4.2–6.0	4.0–6.1	4.4–6.4	3.2–6.4	3.7–5.4	3.7–5.6	3.8–5.6	3.5–5.7	–0.7	0.7
Hematocrit	l	0.38–0.53	0.36–0.54	0.38–0.54	0.29–0.52	0.33–0.48	0.34–0.50	0.32–0.47	0.33–0.49	–0.07	–0.06
MCV	fL	77–102	77–105	74–104	69–102	78–101	78–103	65–100	71–103	–9	8
MCH (Fe)	fmol	1.60–2.11	1.60–2.20	1.40–2.11	1.31–2.05	1.60–2.10	1.60–2.11	1.20–2.00	1.43–2.17	–0.19	0.20
MCHC (Fe)	mmol/L	19–23	19–23	19–23	10–22	19–23	19–23	18–22	18–24	–2	2
<i>Leukocytes</i>											
Leukoeycles	GI/L	3.4–10.8	3.3–11.8	2.9–9.4	2.9–10.2	3.4–11.8	3.3–10.8	3.0–12.5	3.0–9.4	–4.0	3.7
Bands	GI/L	0–0.29	0–0.26	0–0.16	0–0.20	0–0.34	0–0.21	0–0.08	0–0.30	–0.29	0.26
Neutrophils	GI/L	1.72–7.67	1.63–7.79	1.08–6.12	0.75–6.28	1.67–8.38	1.39–8.36	1.09–8.54	0.87–5.45	–3.38	3.18
Lymphocytes	GI/L	0.87–3.76	0.75–4.42	0.96–4.10	0.82–4.17	0.91–3.65	0.85–3.64	0.94–3.95	0.95–3.96	–1.57	1.54
Monocytes	GI/L	0–0.79	0–0.96	0–0.75	0–0.72	0–0.76	0–0.78	0–0.78	0–1.00	–0.44	0.46
Eosinophils	GI/L	0–0.53	0–0.68	0–0.59	0–0.50	0–0.50	0–0.59	0–0.48	0–0.54	–0.38	0.37
Basophils	GI/L	0–0.15	0–0.17	0–0.15	0–0.19	0–0.17	0–0.17	0–0.20	0–0.19	–0.13	0.13
Platelets	GI/L	147–399	131–429	152–508	123–397	168–474	155–514	165–448	145–409	–93	107
<i>Urine</i>											
Specific gravity		1.006–1.037	1.005–1.036	1.007–1.035	1.005–1.036	1.004–1.036	1.005–1.035	1.006–1.036	1.006–1.034	–0.018	0.017
pH		5–8	5–8	5–8	5–9	5–8	5–8	5–7	5–7	–2.0	2.0

^aThe number of adult patients with at least one analyte result is 4265.

Appendix E Eli Lilly and Company Adult Reference Ranges in Clinical Trials for Routine Laboratory Test Results Nonsmoker/Nonimiber ($n = 4801$)^a

Analyte	Age	Male				Female				Delta limits	
		Caucasian		Non-Caucasian		Caucasian		Non-Caucasian			
		<50	≥50	<50	≥50	<50	≥50	<50	≥50	Low	High
Liver											
ALT	U/L	9-97	6-79	4-109	6-68	5-94	5-77	4-89	4-82	-27	28
AST	U/L	11-67	10-55	11-88	9-60	10-64	9-62	8-64	10-65	-25	22
GGT	U/L	6-117	5-190	7-148	6-135	8-137	5-148	6-130	5-123	-30	29
Alk. phosphatase	U/L	29-136	26-162	29-149	29-146	29-136	23-156	24-138	34-171	-44	28
Bilirubin	μmol/L	5-32	3-36	3-29	3-24	3-22	3-19	2-19	2-21	-10	10
Muscle											
Creatine kinase	U/L	17-667	15-419	15-717	14-1043	10-269	0-238	13-407	11-610	-277	236
Kidney											
Creatinine	μmol/L	62-141	62-168	71-159	71-194	53-124	53-150	53-133	44-150	-35	35
Urea	mmol/L	2.5-9.3	2.9-10.7	1.8-8.6	2.5-11.4	1.8-8.6	2.5-11.1	1.8-7.5	2.9-10.7	-3.2	3.2
Uric acid	μmol/L	155-547	178-601	184-589	244-601	95-500	113-517	125-476	155-619	-125	119
Phosphate(i)	mmol/L	0.81-1.52	0.68-1.45	0.77-1.68	0.68-1.55	0.77-1.55	0.74-1.58	0.77-1.52	0.84-1.58	-0.45	0.42
Calcium	mmol/L	2.07-2.64	2.05-2.59	2.05-2.64	2.05-2.59	2.02-2.59	2.05-2.64	2.02-2.64	2.10-2.69	-0.30	0.30
Electrolytes											
Sodium	mmol/L	132-148	134-148	135-157	135-150	129-145	134-150	135-150	135-152	-8	8
Potassium	mmol/L	3.5-5.5	3.3-5.5	2.8-7.4	2.8-5.4	3.3-5.3	3.4-5.5	3.0-5.0	3-5.2	-1.1	1.0
Chloride	mmol/L	96-112	96-111	93-112	96-111	96-110	95-112	97-110	95-111	-9	8
Bicarbonate	mmol/L	22-36	22-35	21-36	21-36	20-34	21-35	19-36	21-36	-7	8
Nutritional											
Glucose, fasting	mmol/L	3.6-13.2	3.8-16.6	3.9-9.5	3.6-17.5	3.8-12.1	3.8-15.2	3.8-16.0	2.7-21.7	-3.1	3.2
Glucose, nonfasting	mmol/L	4.3-12.6	4.7-14.4	3.4-8.3	4.5-17.3	4.3-9.8	3.5-29.2	4.1-15.8	3.9-22.8	-4.8	4.2
Albumin	g/L	37-51	34-49	34-53	30-48	34-50	33-49	32-49	33-49	-7	6
Protein	g/L	63-84	60-82	61-88	59-86	61-83	62-83	64-85	63-86	-11	10
Cholesterol	mmol/L	3.08-7.81	3.44-8.51	2.79-8.24	3.21-9-05	3.18-8.33	3.70-8.82	3.36-8.51	3.80-9.31	-1.78	1.73
HDL cholesterol	mmol/L	0.42-2.07	0.55-2.12	0.28-2.43	0.36-2.16	0.49-2.22	0.54-2.53	0.47-2.25	0.52-2.61	-0.85	0.72
LDL cholesterol	mmol/L	1.48-5.77	1.84-5.48	1.42-6.30	1.32-6.43	0.91-5.28	1.63-6.23	1.66-5.82	2.20-6.57	-1.97	1.78
Triglycerides	mmol/L	0.54-11.02	0.52-8.85	0.44-16.81	0.50-4.19	0.40-9.08	0.67-7.77	0.55-5.59	0.52-4.44	-3.01	2.88
(Continued)											

(Continued)

Appendix E Eli Lilly and Company Adult Reference Ranges in Clinical Trials for Routine Laboratory Test Results Nonsmoker/Nonimiber (n = 4801)^a (Continued)

Analyte	Age	Male				Female				Delta limits	
		Caucasian		Non-Caucasian		Caucasian		Non-Caucasian			
		<50	≥50	<50	≥50	<50	≥50	<50	≥50	Low	High
Erythrocytes											
Hemoglobin (Fe)	mmol/L	7.80–11–30	7.50–11.20	6.30–11.00	6.80–10.92	6.64–9.90	6.90–10.24	5.00–10.43	6.30–9.99	–1.20	1.20
Erythrocytes	TI/L	4.2–6.3	4.0–6.2	4.0–6.4	3.7–6.2	3.7–5.6	3.7–5.6	3.6–5.9	3.6–5.8	–0.7	0.7
Hematocrit	l	0.37–0.54	0.36–0.54	0.31–0.54	0.33–0.54	0.33–0.48	0.34–0.51	0.27–0.51	0.33–0.49	–0.07	0.06
MCV	fL	77–99	78–103	68–100	74–103	75–101	76–101	66–100	72–100	–9	8
MCH (Fe)	fmol	1.60–2.05	1.60–2.11	1.40–2.00	1.40–2.10	1.49–2.05	1.55–2.10	1.24–2.05	1.40–2.05	–0.19	0.20
MCHC (Fe)	mmol/L	19–23	10–23	18–23	18–22	19–23	19–23	18–22	18–22	–2	2
Leukocytes											
Leukoocytes	GI/L	3.6–11.5	3.6–12.6	3.1–12.8	2.8–10.9	3.4–13.0	3.7–11.8	3.0–11.5	3.0–10.6	–4.0	3.7
Bands	GI/L	0–0.32	0–0.38	0–0.15	0–0.28	0–0.57	0–0.40	0–0.20	0–0.20	–0.29	0.26
Neutrophils	GI/L	1.86–7.31	1.89–8.61	1.19–9.06	1.33–7.50	1.63–8.78	1.88–7.83	1.13–8.24	1.33–7.31	–3.38	3.18
Lymphocytes	GI/L	0.80–3.92	0.88–4.29	0.84–4.26	0.75–4.15	1.03–4.12	0.96–4.17	0.84–4.03	1.02–4.14	–1.57	1.54
Monocytes	GI/L	0–0.85	0–0.90	0–0.78	0–0.82	0–0.80	0–0.88	0–0.83	0–0.71	–0.44	0.46
Eosinophils	GI/L	0–0.48	0–0.64	0–0.68	0–0.54	0–0.62	0–0.59	0–0.81	0–0.55	–0.38	0.37
Basophils	GI/L	0–0.19	0–0.16	0–0.14	0–0.13	0–0.16	0–0.16	0–0.16	0–0.14	–0.13	0.13
Platelets	GI/L	152–420	127–483	156–476	150–443	160–530	157–486	164–541	149–559	–93	107
Urine											
Specific gravity		1.006–1.037	1.006–1.036	1.011–1.040	1.006–1.036	1.005–1.037	1.005–1.036	1.006–1.040	1.006–1.038	–0.018	0.017
pH		5–8	5–8	5–7	5–8	5–8	5–8	5–8	5–8	–2.0	2.0

^aThe number of adult patients with at least one analyte result is 4801.

RESULTS

The Lilly Reference Ranges are based on a sample of 20,102 patients with at least one analyte result recorded. These patients had blood and urine tests during qualification for entry into 50 clinical trials. A total of 1121 investigators studied 11 indications throughout the United States and Europe. Each of these patients had nonmissing values for gender, race, and age.

This reference population appeared comparable to the U.S. population based as reported in the national almanac on gender, race, age, smoking, and ethanol use. A smoker and imbibor are defined to be those patients who at the screening visit report that they currently smoke or take alcohol. Patients whose baseline value for smoking or ethanol use was missing were deleted from the partitioned interval calculation. Table 1 contains the final sample sizes

for each of the five categorized groups. The NCCLS subcommittee recommends a sample size of 120 reference subjects per analyte.

Each of the five categorized groups were further partitioned into eight groups defined by gender, race, and age. The 98% reference intervals for 39 are shown in Appendix A for the total sample. The 98% intervals for 39

Table 1 Sample Sizes by Category

Smoker	Imbibor	Sample size	Appendix
Total	Total	20,102	A
Yes	Yes	2974	B
Yes	No	2037	C
No	Yes	4265	D
No	No	4801	E

Appendix F Eli Lilly and Company the Effect of Demographic Characteristics on Reference Ranges *p* Values

Analyte	Gender		Race		Age		Smoking		Drinking	
	Low	High	Low	High	Low	High	Low	High	Low	High
<i>Liver</i>										
ALT	0.099	<0.001	<0.001	0.580	0.289	0.002	0.035	0.886	<0.001	0.003
AST	0.021	<0.001	0.254	0.011	0.777	0.824	0.165	0.010	0.053	<0.001
GGT	<0.001	<0.001	0.647	<0.001	0.122	0.098	0.525	<0.001	0.140	0.012
Alk. phosphatase	0.961	0.078	0.098	0.542	0.893	<0.001	0.006	0.308	0.141	<0.001
Bilirubin	<0.001	<0.001	0.004	0.197	0.258	0.036	0.006	<0.001	0.377	0.020
<i>Muscle</i>										
Creatine kinase	0.031	<0.001	0.164	<0.001	<0.001	0.004	0.375	0.744	0.003	0.741
<i>Kidney</i>										
Creatinine	<0.001	<0.001	0.706	<0.001	<0.001	<0.001	0.462	0.110	0.607	0.018
Urea	<0.001	0.005	0.281	0.946	<0.001	<0.001	<0.001	0.018	0.476	0.001
Uric acid	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.435	0.061	0.846	0.180
Phosphate (i)	0.015	0.109	0.275	0.587	0.55	<0.001	0.029	0.196	0.808	0.465
Calcium	0.179	0.360	0.129	0.004	0.268	0.135	0.502	0.815	0.999	0.040
<i>Electrolytes</i>										
Sodium	0.495	0.013	0.060	0.032	0.289	0.424	0.143	0.527	0.939	0.200
Potassium	0.465	0.265	<0.001	0.626	0.002	0.826	0.574	0.710	<0.001	0.178
Chloride	0.117	<0.001	0.639	0.004	0.006	0.004	0.727	<0.001	0.201	<0.001
Bicarbonate	0.304	0.623	0.380	0.103	0.265	0.573	0.089	0.006	0.755	0.980
<i>Nutritional</i>										
Glucose, fasting	0.335	0.350	0.355	<0.001	0.372	<0.001	0.399	0.019	0.636	<0.001
Glucose, nonfasting	0.194	0.578	0.610	0.031	0.046	0.002	0.975	0.850	0.068	0.017
Albumin	0.516	<0.001	0.142	0.664	<0.001	<0.001	0.372	0.402	0.030	0.013
Protein	0.944	0.124	0.025	<0.001	<0.001	0.034	0.028	0.047	0.363	0.141
Cholesterol	<0.001	0.832	0.030	0.221	<0.001	<0.001	<0.001	0.171	0.650	0.692
HDL cholesterol	0.006	<0.001	0.676	0.007	0.963	0.948	0.381	0.087	0.040	0.030

(Continued)

Appendix F Eli Lilly and Company the Effect of Demographic Characteristics on Reference Ranges *p* Values (Continued)

Analyte	Gender		Race		Age		Smoking		Drinking	
	Low	High	Low	High	Low	High	Low	High	Low	High
LDL cholesterol	0.670	0.881	0.071	0.009	<0.001	0.059	0.074	0.695	0.279	0.677
Triglycerides	<0.001	<0.001	0.520	0.261	<0.001	0.592	0.624	0.532	0.004	0.464
<i>Erythrocytes</i>										
Hemoglobin	<0.001	<0.001	<0.001	0.092	0.208	0.443	0.017	<0.001	0.002	0.035
Erythrocytes	<0.001	<0.001	0.023	<0.001	<0.001	0.187	0.719	0.021	0.550	0.037
Hematocrit	<0.001	<0.001	<0.001	0.343	0.871	0.011	0.019	<0.001	0.028	0.014
MCV	0.086	<0.001	<0.001	0.012	0.057	<0.001	0.176	<0.001	0.002	<0.001
MCH	0.003	<0.001	<0.001	0.082	0.601	0.008	0.483	<0.001	<0.001	<0.001
MCHC	0.076	0.698	<0.001	0.127	0.019	0.373	0.167	0.792	<0.001	0.200
<i>Leukocytes</i>										
Leukocytes	0.092	0.781	<0.001	0.095	0.929	<0.001	0.017	<0.001	0.912	0.439
Neutrophils	0.083	0.083	<0.001	0.205	0.112	0.015	<0.001	<0.001	0.605	0.491
Lymphocytes	0.002	0.162	0.938	0.012	0.014	0.681	0.275	<0.001	0.188	0.070
Monocytes	–	<0.001	–	0.009	–	0.110	–	<0.001	–	0.664
Eosinophils	–	0.162		0.783		0.079		<0.001		0.230
Basophils	–	0.391		<0.001		0.963		<0.001		0.295
<i>Urine</i>										
Specific gravity	<0.001	0.149	0.007	0.904	0.104	0.303	0.911	0.103	0.067	0.036
pH	–	<0.001	–	0.239	–	<0.001	–	0.011	–	0.167

analytes for the remaining 32 demographic groups are contained in Appendices B–E. Electronic copies of Appendices A–E in ASCII format are available from the author by sending a PC formatted diskette with a self-addressed, stamped envelope.

In addition, the results of two blood samples drawn 1 to 4 weeks apart was examined in 5802 patients who were taking placebo at the onset of a clinical trial to calculate delta limits. Not all 20,102 patients had a second blood draw because some failed to meet inclusion criteria, did not take placebo as directed, or did not return for their study. For the patients who did return, their delta values, defined to be the difference between the two blood draws, were included in the reference sample group. The reference distribution was found to be symmetric about zero. Delta limits were not as influenced by demographic factors as were reference ranges so one pair of delta limits was used to describe an analyte. These delta limits are contained in [Appendix A](#).

For the Lilly Reference Range data of 20,102 patients, the Brown–Mood multisample procedure to analytically compare the impact of these five demographic characteristics on the extreme tails. The Brown–Mood procedure is a non-parametric procedure that answers the question: Do patient demographics influence the reference distribution in the tails? The answer “yes” supports the need for partitioning. This procedure was also applied to the delta limits.

The observed significance levels, *p* values, of the Brown–Mood test for the reference ranges are given in [Appendix F](#). These *p* values are derived from the test of the hypothesis for equality between a specific demographic category in tail proportions. This tests whether the proportion of patients with values below the pooled low reference limit is the same for both strata in a demographic category and whether the proportion of patients with values above the pooled high reference limit is the same for both strata in a demographic category defined by each of the five demographic variables. Missing *p* values occur when there are no observations below (above) the reference limit. The observed significance levels, *p* values, of the Brown–Mood test for the delta limits are not reported.

Contained in [Appendix G](#) are the estimates of the difference in the low (high) reference limit between the two strata defined by each of the five demographic variables relative to the low (high) reference limit calculated from the pooled sampled group. For example, the low AST values for females and males are 9 and 10 U/L, respectively. The low AST value for the entire population is 9 U/L. The entry for the low value of AST for gender is $(9 - 10)/9 \times 100 = -11.1\%$.

The sign of each table entry indicates the direction of the difference between reference limits. Missing values are shown for those estimates where the Brown–Mood test was not statistically significant at the 0.05 level.

Appendix G Eli Lilly and Company Percent Difference Between Reference Values for Each Demographic Group Relative to the Pooled Reference Values

Analyte	Gender (female–male)		Race (Caucasian– non-Caucasian)		Age (young–mature)		Smoking (nonsmoker– smoker)		Drinking (imbiber– nondrinker)	
	Low	High	Low	High	Low	High	Low	High	Low	High
<i>Liver</i>										
ALT	–	–27.1	20.0	–	–	13.5	–12.3	–	34.0	13.7
AST	–11.1 ^a	–31.5	–	–28.2	–	–	–	17.6	–	34.8
GGT	–18.7	–25.0	–	–43.1	–	–	–	39.6	–	22.0
Alk. phosphatase	–	–	–	–	–	–17.6	17.9	–	–	–12.4
Bilirubin	0.0 ^b	–31.0	0.0	–	–	5.2	0.0	–13.2	–	11.5
<i>Muscle</i>										
Creatine kinase	–23.1	–70.0	–	–73.3	92.3	26.7	–	–	33.3	–
<i>Kidney</i>										
Creatinine	–17.0	–23.3	–	–12.0	–0.1	–17.9	–	–	–	–5.9
Urea	–35.0	–6.7	–	–	–35.0	–27.2	–14.3	–10.7	–	–8.0
Uric acid	–58.4	–11.8	–26.5	–10.8	–26.5	–9.7	–	–3.3	–	–
Phosphate (i)	4.2	–	–	–	–	3.8	4.1	–	–	–
Calcium	–	–	–	–0.2	–	–	–	–	–	–0.9
<i>Electrolytes</i>										
Sodium	–	1.3	–	–1.3	0.7	–	–	–	–	–
Potassium	–	–	9.4	–	6.3	–	–	–	6.3	–
Chloride	–	1.8	–	1.8	1.0	1.8	–	1.8	–	2.7
Bicarbonate	–	–	–	–	–	–	–	–2.9	–	–
<i>Nutritional</i>										
Glucose, fasting	–	–	–	–29.6	–	–45.1	–	–21.9	–	–37.3
Glucose, nonfasting	–	–	–	–20.4	–13.2	–50.4	–	–	–	–27.6
Albumin	–	–2.0	–	–	6.1	2.0	–	–	2.9	2.0
Protein	–	–	–1.6	–4.8	1.6	0.0	–1.6	0.0	–	–
Cholesterol	5.9	–	3.3	–	–11.0	–9.1	–8.3	–	–	–
HDL cholesterol	21.0	16.0	–	–6.3	–	–	–	–	10.5	8.5
LDL cholesterol	–	–	15.4	–6.7	–21.5	–	–	–	–	–
Triglycerides	–19.6	–34.7	–	–	–21.9	–	–	–	–17.3	–
<i>Erythrocytes</i>										
Hemoglobin	–10.2	–9.5	9.7	–	–	–	4.6	3.4	5.3	1.0
Erythrocytes	–7.0	–9.5	0.0	–3.3	3.0	–	–	1.7	–	–1.7
Hematocrit	–8.8	–9.3	5.9	–	–	0.0	0.0	3.7	0.0	1.9
MCV	–	–1.9	8.0	1.9	–	–1.9	–	4.8	2.7	2.9
MCH	–6.6	–4.7	13.3	–	–	–2.9	–	4.1	8.3	4.6
MCHC	–	–	5.3	–	0.0	–	–	–	2.0	–
<i>Leukocytes</i>										
Leukocytes	–	–	–15.2	–	–	–7.9	–7.3	–16.6	–	–
Neutrophils	–5.7	–	–38.3	–	–	–6.1	–13.3	–17.9	–	–

(Continued)

Appendix G Eli Lilly and Company Percent Difference Between Reference Values for Each Demographic Group Relative to the Pooled Reference Values (*Continued*)

Analyte	Gender (female–male)		Race (Caucasian– non-Caucasian)		Age (young–mature)		Smoking (nonsmoker– smoker)		Drinking (imbiber– nondrinker)	
	Low	High	Low	High	Low	High	Low	High	Low	High
Lymphocytes	0.002	0.162	0.938	0.012	0.014	0.681	0.275	<0.001	0.188	0.070
Monocytes	–	–9.3	–	7.4	–	–	–	10.5	–	–
Eosinophils	–	–	–	–	–	–	–	19.7	–	–
Basophils	–	–	–	17.6	–	–	–	11.8	–	–
<i>Urine</i>										
Specific gravity	–0.1	–	–0.1	–	–	–	–	–	–	0.0
pH	–	0.0	–	–	–	0.0	–	0.0	–	–

^aExample: $[(9 - 10)/9 \times 100] = -11.1$.

^bExample: $[(3 - 3)/3 \times 100] = 0.0$.

Some have suggested that for the purposes of clinical trials, disease-specific reference intervals would be helpful. Oesterling et al. have constructed these reference intervals for use in clinical trials specific to detecting early prostate cancer.^[12]

CONCLUSION

We offer an alternative to the use of “health-associated” reference ranges in clinical trials. Unlike Lilly Reference Ranges, many “health-associated” reference ranges do not account for the appropriate population, influential demographic characteristics, adequate sample size, non-normal analyte distributions and within patient variability. Proper use of the Lilly Reference Ranges can lead to more precise safety and efficacy conclusions than the use of such “health-associated” reference ranges in clinical trials.

REFERENCES

1. Wilson, M.G.; Enas, G.G. In *Reference Values and Analysis of Routine Clinical Laboratory Data in Clinical Trials*, Joint Meeting of the Society of Clinical Trials and the International Society for Clinical Biostatistics, Brussels, Belgium, July 8–12, 1991.
2. Thompson, W.L.; Brunelle, R.L.; Wilson, M.G. Performance and interpretation of laboratory tests. In *Clinical Drug Trials and Tribulations*, 2nd Ed.; Cato, A.E., Ed.; Dekker: New York, 2001.
3. Wilson, M.G.; Enas, G.G. In *What Analyses Should be Performed on Test Results*, Drug Information Association Workshop On the Integrated Safety Summary, Bethesda, Maryland, April 27, 1990.
4. Thompson, W.L.; Brunelle, R.L.; Enas, G.G.; Simpson, P.J. Routine laboratory tests in clinical trials. *J. Clin. Res. Drug Dev.* **1987**, *1*, 95–119.
5. Solberg, H.E. Approved recommendation (1986) on the theory of reference values. Part 1. The concept of reference values. *Clin. Chim. Acta* **1987**, *167*, 111–118.
6. PetitClerc, C.; Solberg, H.E. Approved recommendation (1987) on the theory of reference values. Part 2. Selection of individuals for the production of reference values. *J. Clin. Chem. Clin. Biochem* **1987**, *25*, 639–644.
7. Solberg, H.E.; PetitClerc, C. Approved recommendation (1988) on the theory of reference values. Part 3. Preparation individuals and collection of specimens for the production of reference values. *Clin. Chim. Acta* **1988**, *177*, S1–S12.
8. Solberg, H.E.; Stamm, D. Approved recommendation on the theory of reference values. Part 4. Control of analytical variation in the production, transfer, and application of reference values. *Eur. J. Clin. Chem. Clin. Biochem.* **1991**, *29*, 531–535.
9. Solberg, H.E. Approved recommendations (1987) on the theory of reference values. Part 5. Statistical treatment of collected reference values. Determination of reference limits. *J. Clin. Chem. Clin. Biochem.* **1987**, *25*, 645–656.
10. Dybkaer, R.; Solberg, H.E. Approved recommendation (1987) on the theory of reference values. Part 6. Presentation of observed values related to reference values. *J. Clin. Chem. Clin. Biochem.* **1987**, *25*, 657–662.
11. Sasse, E.A.; Aziz, K.J.; Harris, E.K.; Krishnamurthy, S.; Lee, H.T.; Ruland, A.; Seamonds, B. *How to Define and Determine Reference Intervals in the Clinical Laboratory; Approved Guideline*; 1995. The National Committee for Clinical Laboratory Standards. Publication C28-A. 771 East Lancaster Avenue, Villanova, Pennsylvania 19085, USA.
12. Oesterling, J.E.; Jacoben, S.J.; Cooner, W.H. The use of age-specific reference ranges for serum prostate specific antigen in men 60 years old or older. *J. Urol.* **1995** Apr, *153* (4), 1160–1163.

Local Influence Analysis

Nicola Sartori

Dipartimento di Statistica, Università Cà Foscari di Venezia, Venezia, Italy

Abstract

Local influence analysis is a method for assessing the influence of small perturbations of a statistical model. This entry formalizes the idea of perturbation schemes, addresses likelihood displacement as a measure of sensitivity given to perturbation, and provides an example application in normal linear regression.

INTRODUCTION

Statistical analyses are usually based on models, which are probabilistic descriptions of the process that generated the data. Very often, models involve some unknown parameters of intrinsic interest. These parameters are estimated using the data and lead to the estimated model, which forms the basis for statistical inference. In this sense, a statistical model is a useful device for extracting and understanding the essential features of a set of data. However, the model is almost always only an approximate description of more complicated processes, and hence it is always wrong. Given this inexactness, it is important to study the variation in inference results under modifications of model formulation. In particular, one important concern is that isolated features of the data or specific assumptions of the model should not seriously influence the key results of the analysis.

In general, we may think of influence analysis as a set of methods for assessing the sensitivity of inferences to perturbations in the data or the model. Inferences that are relatively stable across these perturbations should be preferred, even at the cost of some loss in model exactness. On the other hand, inferences that are heavily affected by minor perturbations should be cause of concern. Aspects of the data or the model, which strongly change the results of the analysis, are said to be influential. Case deletion in regression models is an example of data influence diagnostic.^[1] Local influence analysis, introduced by Cook,^[2] is a method for assessing the influence of small perturbations of a statistical model. The method considers changes in the log-likelihood scale as measures to evaluate influential perturbations and, in particular, the analysis is local because such changes are examined in a small neighborhood of the assumed, unperturbed, model.

In the following section on “Perturbation Schemes,” we formalize the idea of perturbation schemes. Then, in section “Likelihood Displacement,” the likelihood displacement is introduced as a measure of sensitivity to a given

perturbation. The local study of the likelihood displacement in a neighborhood of the assumed model is described in section “Local Influence.” As an example an application in normal linear regression is considered. Finally, we end with a brief description of applications and extensions.

PERTURBATION SCHEMES

Let $L(\theta)$ be the log-likelihood of the postulated statistical model, where θ is a p dimensional parameter. Let us assume that we introduce a perturbation in the model formulation through a q dimensional vector ω confined in an open subset Ω of $\mathcal{R}^q$. Let $L(\theta|\omega)$ be the log-likelihood corresponding to the perturbed model for a given ω . We assume that there exists an ω_0 such that $L(\theta) = L(\theta|\omega_0)$, for all θ . We will use $\hat{\theta}$ and $\hat{\theta}_\omega$ to denote the maximum likelihood estimates obtained from $L(\theta)$ and $L(\theta|\omega)$, respectively.

In general, ω can reflect any well-defined perturbation scheme that can be related to data or to model assumptions. Although local influence can be applied in wide generality, as an illustrative example, we will consider the normal linear regression model as the postulated model

$$Y = X\beta + \varepsilon \quad (1)$$

where Y is the $n \times 1$ response variable, X is the $n \times p$ matrix of explanatory variables, β is the $p \times 1$ vector of regression coefficients, and $\varepsilon \sim N_n(0, \sigma^2 I_n)$.

The parameter is then $\theta = (\beta, \sigma^2)$, and the log-likelihood is

$$L(\theta) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (Y - X\beta)^T (Y - X\beta)$$

Examples of perturbation schemes are as follows:

1. Response perturbations: in model (1), Y is replaced by $Y + \omega$. Hence, ω has dimension $q = n$ and $\omega_0 = 0$.

2. Predictor perturbations: we may perturb, for instance, the i th column of X , X_i , with $X_i + \omega$. Even in this case, $q = n$ and $\omega_0 = 0$.
3. Variances perturbations: we can change the assumption of homoscedasticity replacing the variance of ε in Eq. (1), with $\sigma^2 W^{-1}$, where $W = \text{diag}(\omega_1, \omega_2, \dots, \omega_n) = \text{diag}(\omega)$. Hence $q = n$ and $\omega_0 = 1$. We note that, when a particular ω_j approaches 0, the analysis corresponds to the deletion of the j th observation. For this reason, this is a generalization of deletion diagnostics, such as Cook's distance,^[1] and is also addressed as case-weight perturbation.^[3]

The first two schemes are perturbations of the data, while the third one is a perturbation of an assumption of the model. Other examples of perturbation schemes in the linear model may be found in Refs. [3,4].

LIKELIHOOD DISPLACEMENT

A comparison of $\hat{\theta}$ and $\hat{\theta}_\omega$ is one way of assessing the influence of a particular perturbation ω . If the difference between $\hat{\theta}$ and $\hat{\theta}_\omega$ is not large, when ω varies in Ω , then inference is stable. However, it is not easy to compare these estimates when the dimension p is large. Moreover, different elements of $\hat{\theta}$ may be in different units and estimated with different precisions. In fact, it is more convenient to judge the distance between $\hat{\theta}$ and $\hat{\theta}_\omega$ with a scalar real-valued function based on the likelihood, even because in likelihood analyses, it is the likelihood itself that summarizes the essential inference information.

Cook^[2] introduces the likelihood displacement

$$LD(\omega) = L(\hat{\theta}) - L(\hat{\theta}_\omega) \quad (2)$$

as a measure for assessing the difference between $\hat{\theta}$ and $\hat{\theta}_\omega$. The likelihood displacement is a non-negative function, which is zero in the postulated model, $LD(\omega_0) = 0$, and that is large when influence is present. We note that $L(\theta|\omega)$ is used only to obtain the estimate $\hat{\theta}_\omega$.

LOCAL INFLUENCE

The plot of $LD(\omega)$ as ω varies in Ω is called the influence graph, and it can be a useful tool to explore the presence of influence. However, this is rarely possible in practical situation, say when $q \leq 2$. Moreover, in general, an influence graph requires the evaluation of $\hat{\theta}_\omega$ for each value of ω which, in some models, may be a computationally intensive task.

One possible solution is to study the local behavior of Eq. (2) in a neighborhood of the postulated model, that means, in a neighborhood of ω_0 . Given a direction d in Ω

of unit length, we can study the behavior of $LD(\omega_0 + ad)$ as a function of the scalar a . As the likelihood displacement achieves a minimum in ω_0 , the graph of $LD(\omega_0 + ad)$ has a minimum at $a = 0$. For this reason, Cook^[2] proposed to use the normal curvature $C(d)$ of $LD(\omega + ad)$ around $a = 0$. A large curvature means that $LD(\omega + ad)$ is changing rapidly at $a = 0$ in the direction d . Therefore, the direction of maximum curvature, $d_{\max} = \text{argmax } C(d)$, gives the direction in which $LD(\omega)$ is changing most rapidly around ω_0 .

The normal curvature in the direction d is defined as

$$C(d) = 2|d^T F d|$$

where $F = \partial^2 L(\hat{\theta}_\omega) / \partial \omega \partial \omega^T$, evaluated at $\theta = \hat{\theta}$ and $\omega = \omega_0$. Using chain rule for differentiation, we can write

$$F = \left[\frac{\partial^2 L(\hat{\theta}_\omega)}{\partial \omega \partial \omega^T} \right]_{\hat{\theta}, \omega_0} = \Delta(\hat{\theta}, \omega_0)^T J(\hat{\theta})^{-1} \Delta(\hat{\theta}, \omega_0)$$

where $\Delta(\theta, \omega) = \partial^2 L(\theta|\omega) / \partial \theta \partial \omega^T$ and $J(\theta) = -\partial^2 L(\theta) / \partial \theta \partial \theta^T$ is the observed information matrix. Hence, it follows that

$$C(d) = 2 \left| d^T \Delta(\hat{\theta}, \omega_0)^T J(\hat{\theta})^{-1} \Delta(\hat{\theta}, \omega_0) d \right| \quad (3)$$

Note that all quantities in Eq. (3) are evaluated at $\theta = \hat{\theta}$ and $\omega = \omega_0$, which are quantities of the unperturbed model.

The maximum curvature $C_{\max} = \max C(d)$ is equal to twice the largest eigenvalue of F . The corresponding eigenvector is then the direction of maximum curvature $d_{\max}$. The direction of maximum curvature is the main diagnostic tool that comes with this method. A plot of the elements of $d_{\max}$ may help to identify which components of ω are relatively most influential. Moreover, to assess the presence of local influence in the direction $d_{\max}$, it is possible to plot relevant statistics against the perturbation $\omega(a) = \omega_0 + a d_{\max}$. For instance, we may plot $LD(\omega(a))$ against a or plot elements of $\hat{\theta}_\omega(a)$ against a . We note that the direct interpretation of $C_{\max}$ itself is not recommended, as it depends on the chosen scale for ω , while $d_{\max}$ does not.

When the interest is in a subset of the whole parameter, we consider a different likelihood displacement. Indeed, let us assume that $\theta^T = (\theta_1^T, \theta_2^T)$ and that only θ_1 is of direct interest. Hence, we define the likelihood displacement for the subset θ_1 as

$$LD_1(\omega) = 2 \left\{ L(\hat{\theta}) - L(\theta_{1\omega}, c(\hat{\theta}_{1\omega})) \right\}$$

where $c(\theta_1)$ is the function that maximizes $L(\theta_1, \theta_2)$ for each fixed θ_1 , and $\hat{\theta}_{1\omega}$ is determined by the partition $\hat{\theta}_\omega^T = (\hat{\theta}_{1\omega}^T, \hat{\theta}_{2\omega}^T)$. In Ref. [2], section 3.2, it is shown that the normal curvature for $LD_1(\omega)$ has the form

$$C_1(d) = 2 \left| d^T \Delta(\hat{\theta}, \omega_0)^T \left\{ J(\hat{\theta})^{-1} - B_{22} \right\} \Delta(\hat{\theta}, \omega_0) d \right| \quad (4)$$

where

$$B_{22} = \begin{pmatrix} 0 & 0 \\ 0 & J_{22}(\hat{\theta})^{-1} \end{pmatrix}$$

and $J_{22}(\theta)$ is the block of the information matrix $J(\theta)$ corresponding to θ_2 . Hence, in this case $d_{\max}$ is the eigenvector corresponding to the maximum eigenvalue of $\Delta(\hat{\theta}, \omega_0)^T \{J(\hat{\theta})^{-1} - B_{22}\} \Delta(\hat{\theta}, \omega_0)$.

AN EXAMPLE

As a numerical illustration, let us consider the geese data for observer 1.^[5] The data consist of observations on true flock size obtained by aerial photographs (y) and a visual estimation of the flock size made by the observer (x) on 45 flocks of snow geese. We use a simple linear regression as given by Eq. (1) to illustrate the behavior of $d_{\max}$ in the presence of heteroscedasticity. In particular, as in Ref. [2], section 4.4, we use a case-weights perturbation scheme, as illustrated in point 3 at the end of section “Perturbation Schemes.”

The log-likelihood is given by

$$L(\theta | \omega) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (Y - X\beta)^T W (Y - X\beta)$$

and, assuming that only coefficients β are of interest, the normal curvature [Eq. (4)] is given by

$$C_1(d) = 2d^T \text{diag}(e) X (X^T X)^{-1} X^T \text{diag}(e) d / \hat{\sigma}^2$$

where e is the vector of observed residuals of the postulated model.

The maximum curvature $C_{\max}$ is the largest eigenvalue of the matrix $\text{diag}(e) X (X^T X)^{-1} X^T \text{diag}(e)$ and is equal to 14.39. The corresponding eigenvector $d_{\max}$ is the direction of maximum local influence. A plot of the absolute values of the elements of $d_{\max}$ is given in Fig. 1, showing that observations 29, 28, and 41 are mostly locally influential (relatively to the others). Fig. 2 shows a plot of the elements of $d_{\max}$ vs. the observer counts (x) and gives a clear indication of heteroscedasticity. In this example, $d_{\max}$ is a useful tool to direct attention to cases where specification of weights can be most important.

APPLICATIONS AND EXTENSIONS

Local influence analysis is a diagnostic tool that allows to check if the results of inference are affected by local perturbations of model assumptions. Because it is based on changes in the log-likelihood scale, its use can be quite general, and several applications have been proposed in the literature. Some are listed below.

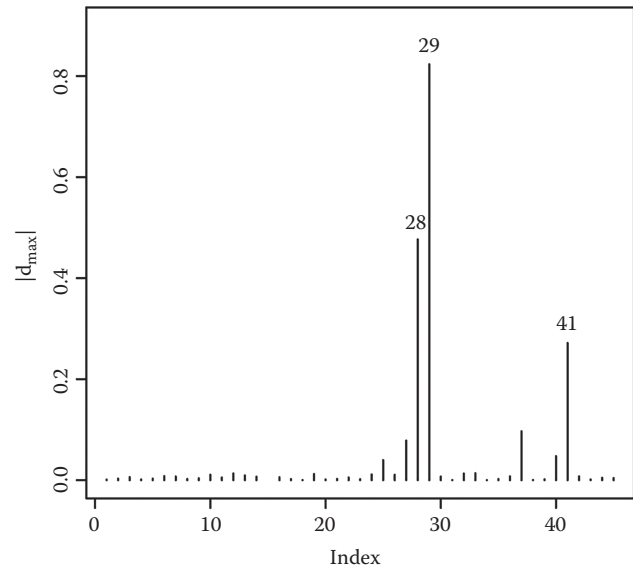


Fig. 1 Geese data. Absolute values of the elements of $d_{\max}$

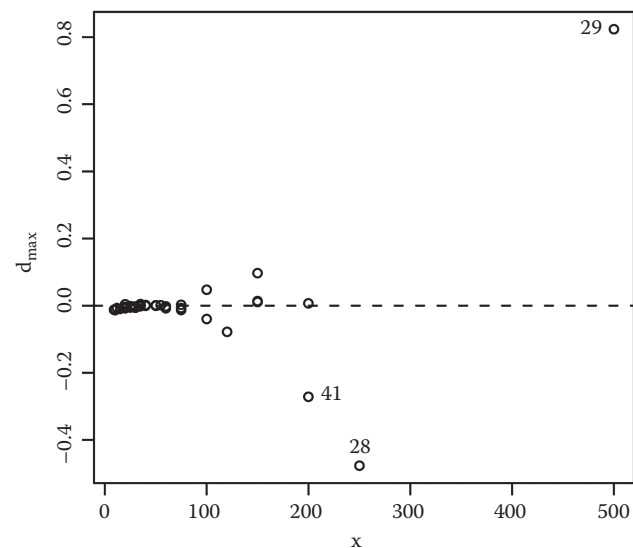


Fig. 2 Geese data. Elements of $d_{\max}$ vs. observer counts (x)

When Cook^[2] introduced local influence, he considered several perturbation schemes but mainly restricted to normal linear regression. A recent review is also given in Ref. [4]. An extension to linear mixed models is given in Refs. [6,7]. Normal non-linear models are considered in Refs. [8,9], while generalized linear models are considered in Refs. [10–12] and, with the addition of random effects, in Ref. [13]. Local influence related to different aspects of the proportional hazards model is studied in Refs. [14–18]. An adaptation of local influence methods to regression analyses with censoring is given in Ref. [19], while problems with incomplete data are investigated in Refs. [20,21].

Finally, we note that the lack of invariance of the normal curvature $C_{\max}$ with respect to the scale of the perturbation scheme has sometimes raised some concern, and variants and extensions of Cook's approach have been proposed in the literature.^[22–27] However, local influence should be viewed mainly as a diagnostic tool and the main tool that comes with this method, which is the direction of maximum curvature $d_{\max}$, is invariant with respect to reparameterizations of the perturbation scheme.

ARTICLES OF FURTHER INTEREST

Confidence Interval and Hypotheses Testing, p. 231.
Dose Response Analysis in Clinical Trials, p. 298.
Failure-Time Model, p. 001.
Logistic Regression, p. 548.
Multicollinearity, p. 001.
Outlier Detection in Clinical Research, p. 001.
Proportional Hazards Regression Model, p. 816.
Survival Analysis, p. 972.

REFERENCES

1. Cook, R.D. Detection of influential observation in linear regression. *Technometrics* **1977**, *19*, 15–18.
2. Cook, R.D. Assessment of local influence. *J. Royal Stat. Soc., Series B: Methodological* **1986**, *48*, 133–155.
3. Cook, R.D. Influence assessment. *J. Appl. Stat* **1987**, *14*, 117–131.
4. Suárez Rancel, M.M.; Gonzalez Sierra, M.A. Regression diagnostic using local influence: a review. *Communic. Stat. Theory Met* **2001**, *30*, 799–813.
5. Weisberg, S. *Applied Linear Regression*, 3rd Ed.; John Wiley & Sons: Hoboken, NJ, 2005.
6. Lesaffre, E.; Verbeke, G. Local influence in linear mixed models. *Biometrics* **1998**, *54*, 570–582.
7. Demidenko, E.; Stukel, T.A. Influence analysis for linear mixed-effects models. *Stat. Med* **2004**, *24*, 893–909.
8. Schwarzmann, B. A connection between local-influence analysis and residual analysis. *Technometrics* **1991**, *33*, 103–104.
9. St. Laurent, R.T.; Cook, R.D. Leverage, local influence and curvature in nonlinear regression. *Biometrika* **1993**, *80*, 99–106.
10. Thomas, W.; Cook, R.D. Assessing influence on regression coefficients in generalized linear models. *Biometrika* **1989**, *76*, 741–750.
11. Thomas, W.; Cook, R.D. Assessing influence on predictions from generalized linear models. *Technometrics* **1990**, *32*, 59–65.
12. Thomas, W. Influence on confidence regions for regression coefficients in generalized linear models. *J. Am. Statist. Assoc.* **1990**, *85*, 393–397.
13. Zhu, H.-T.; Lee, S.-Y. Local influence for generalized linear mixed models. *Can. J. Stat.* **2003**, *31*, 293–309.
14. Pettitt, A.N.; Daud, I.B. Case-weighted measures of influence for proportional hazards regression. *App. Stat* **1989**, *38*, 51–67.
15. Weissfeld, L.A. Influence diagnostics for the proportional hazards model. *Stat. Probabil. Lett.* **1990**, *10*, 411–417.
16. Barlow, W.E. Global measures of local influence for proportional hazards regression models. *Biometrics* **1997**, *53*, 1157–1162.
17. Verbeke, G.; Molenberghs, G.; Thijs, H.; Lesaffre, E.; Kenward, M.G. Sensitivity analysis for nonrandom dropout: a local influence approach. *Biometrics* **2001**, *57*, 7–14.
18. Sartori, N.; Thomaseth, K.; Salvan, A. Local influence analysis when interfacing toxicokinetic and proportional hazard models. *Stat. Med.* **2004**, *23*, 2399–2412.
19. Escobar, L.A.; Meeker J., William Q. Assessing local influence in regression analysis with censored data. *Biometrics* **1992**, *48*, 507–528.
20. Zhu, H.-T.; Lee, S.-Y. Local influence for incomplete-data models. *J. Royal Stat. Soc., Series B: Stat. Methodol.* **2001**, *63*, 111–126.
21. van Steen, K.; Molenberghs, G.; Verbeke, G.; Thijs, H. A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. *Stat. Modell.* **2001**, *1*, 125–142.
22. Schall, R.; Dunne, T.T. A note on the relationship between parameter collinearity and local influence. *Biometrika* **1992**, *79*, 399–404.
23. Billor, N.; Loynes, R.M. Local influence: a new approach. *Communic. Stat.: Theory Meth* **1993**, *22*, 1595–1611.
24. Poon, W.-Y.; Poon, Y.S. Conformal normal curvature and assessment of local influence. *J. Royal Stat. Soc., Series B: Stat. Meth* **1999**, *61*, 51–61.
25. Loynes, R.M. A new measure in local influence. *J. Stat. Planning Inference* **2001**, *92*, 47–53.
26. Critchley, F.; Marriott, P. Data-informed influence analysis. *Biometrika* **2004**, *91*, 125–140.
27. Zhu, H.; Zhang, H. A diagnostic procedure based on local influence. *Biometrika* **2004**, *91*, 579–589.

LOCF: Validity

Bin Cheng

University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

This entry is focused on the analysis of the last observations, which are defined as observations from the last time point for patients who completed the study and the last observations prior to the dropout for patients who did not complete the study.

INTRODUCTION

In clinical trials, data are often collected over a period of time from participating patients. In many situations, however, analyses are only based on data from the last time point (the end of the study) or change from the baseline to the last time point. It is often the case that patients drop out before the completion of the study. Heyting et al.^[1] summarized six common reasons for patient dropout: (1) recovery; (2) lack of improvement; (3) side effect; (4) unpleasantness; (5) intercurrent health problems; and (6) external factors unrelated to trials. These reasons indicate that the dropout rate is often informative in the sense that the population of patients who completed the study is different from the population of patients who dropped out.

In this entry, we focus on the analysis of the last observations, which are defined as observations from the last time point for patients who completed the study and the last observations prior to the dropout for patients who did not complete the study. An analysis based only on data from patients who completed the study is called a completers analysis. Although a *completers analysis* is sufficient in some situations, it is often more desirable to perform an *all randomized subjects* analysis. The analysis based on all randomized subjects is usually referred to as an intention-to-treat (ITT) analysis. Regulatory agencies generally consider the ITT analysis as the primary analysis for evaluation of efficacy and safety in clinical trials with informative dropout (see, e.g., Refs. [2–4]). When the dropout is informative, the target populations of a completers analysis and an ITT analysis are different. Suppose that the population of patients under certain treatment is stratified according to the time of the last observations; then the target population of a completers analysis is only the subpopulation of patients who completed the study under different treatments while the target populations of an ITT analysis include all the subpopulations under different treatments.

For the last observation carry-forward (LOCF) analysis based on ITT population, the last observations are carried forward to the last time point for patients who dropped out. The LOCF analysis treats the carried-forward data as observed data at the last time point. Therefore when the dropout is informative, the LOCF analysis may introduce biases to the statistical inference, which has been speculated upon by the Food and Drug Administration (FDA) in its website (see also Ref. [5]). Although there are discussions, examples, and limited empirical results available in the literature (e.g., Refs. [1,6–8]) regarding this subject, it is still unknown whether or not the LOCF test is asymptotically correct (i.e., its asymptotic size is equal to the nominal size). In this entry, we outline results from Shao and Zhong^[5,9] and Cheng and Shao,^[10] which addressed the above question under the one-way and two-way analysis of variance (ANOVA) models. It is shown that, under certain circumstances, the LOCF test is still asymptotically correct. When the LOCF tests are biased, alternatively, we describe simple but asymptotically correct tests. The proposed tests do not require model assumptions that are nonverifiable by the existing data as pointed out by the FDA.^[2]

Overview

In the section “Validity of the LOCF Test Under a One-Way ANOVA Model,” the LOCF test under a one-way ANOVA model is studied. In the section “Validity of the LOCF ANOVA Test Under a Two-Way ANOVA Model,” we examine the LOCF test under the two-way ANOVA model with and without an interaction for assessing treatment effects, respectively. When the LOCF test is invalid, an alternative asymptotically correct test is proposed in the section “Asymptotically Correct Tests.” The proposed test is illustrated through an example concerning an ITT analysis under a one-way ANOVA model. A brief discussion is given in “Conclusion.”

VALIDITY OF THE LOCF TEST UNDER A ONE-WAY ANOVA MODEL

Consider a parallel-group design for comparing $I \geq 2$ treatments, one of which could be a placebo or an active control. Suppose a total of n patients are randomly assigned to the I treatment groups, where the i th group has n_i patients, $i = 1, \dots, I$. Suppose that patients are evaluated at S time points. Let μ_{is} be the mean of the response under the i th treatment at the last visit. If there is no dropout, then the usual one-way ANOVA test is useful for testing the following null hypothesis

$$H_0: \mu_{1S} = \dots = \mu_{IS} \quad (1)$$

based on the last observations from all patients.

Suppose that, in the i th group, n_{is} patients drop out after the s th visit. Note that n_{is} 's are random but satisfy the constraint $\sum_{s=1}^S n_{is} = n_i$. When there are dropouts in the trial, it is not appropriate to test the null hypothesis (Eq. 1). Under an ITT analysis, the patient population under the treatment i consists of S subpopulations, where the s th subpopulation is the population of patients who are on therapy and drop out after the s th visit (the subpopulation with $s = S$ corresponds to the population of completers). Let μ_{is} be the mean of the s th subpopulation under treatment i , the global mean of the i th treatment group is then

$$\mu_i = \sum_{s=1}^S p_{is} \mu_{is}$$

where p_{is} is the true proportion of patients in the s th subpopulation under treatment i . If we use the global means as the measure of the treatment effect, then the null hypothesis of no treatment effect should be formulated as

$$H_0: \mu_1 = \dots = \mu_I \quad (2)$$

McHugh and Matts^[11] discussed a similar measure for the treatment effect. Note that the two null hypotheses (Eqs. 1 and 2) are different when the dropout is informative.

Let y_{isl} be the last observed response from the l th subject in the group of patients who drop out after the s th visit. The LOCF analysis treats y_{isl} as the unobserved y_{iSl} when $s < S$ and applies the usual one-way ANOVA test. That is, the LOCF test concludes that there is a treatment effect if and only if

$$\frac{\text{MSTR}}{\text{MSE}} > F_{I-1, n-I, \alpha}$$

where $F_{I-1, n-I, \alpha}$ is the $(1 - \alpha)$ th quantile of the F -distribution with $I - 1$ and $n - I$ degrees of freedom and α is the level of significance,

$$\text{MSTR} = \frac{1}{I-1} \sum_{i=1}^I n_i (\bar{y}_{i..} - \bar{y}_{...})^2$$

$$\text{MSE} = \frac{1}{n-I} \sum_{i=1}^I \sum_{s=1}^S \sum_{l=1}^{n_{is}} (y_{isl} - \bar{y}_{i..})^2$$

where

$$\bar{y}_{i..} = \frac{1}{n_i} \sum_{s=1}^S \sum_{l=1}^{n_{is}} y_{isl}$$

and

$$\bar{y}_{...} = \frac{1}{n} \sum_{i=1}^I \sum_{s=1}^S \sum_{l=1}^{n_{is}} y_{isl}$$

In practice, an interesting question arises. When the dropout is informative, what is the null hypothesis dose the resultant LOCF test for, especially, when the observed MSTR/MSE value is large? Considering that MSTR will be large when the difference among $\bar{y}_{i..}$, $i = 1, \dots, I$, is large, it is reasonable to say that the null hypothesis for the LOCF test is the null hypothesis H_0 given in Eq. 2. However, treating y_{isl} as the unobserved y_{iSl} makes one think that the LOCF test preassumes that $\mu_{is} = \mu_{iS} = \mu_i$ for all s (i.e., dropout is noninformative). When the dropout is noninformative, the size of the LOCF test is exactly equal to the nominal size α when y_{isl} 's are independent and normally distributed with a constant variance σ^2 and approximately equal to α when the normality assumptions is removed but n_i 's are large for all i . But what is the behavior of the LOCF test when the dropout is informative, i.e., when $\mu_{is} \equiv \mu_i$ is not true?

Intuitively, the LOCF test cannot be correct or asymptotically correct when the dropout is informative. Surprisingly, Shao and Zhong^[5] showed that, in some cases, the LOCF test is still valid even if the dropout is informative. The following theorem, proved in Shao and Zhong,^[5] fully characterizes the validity of the LOCF test under a one-way ANOVA model.

Theorem 1

Suppose that y_{isl} 's are independent with $E(y_{isl}) = \mu_{is}$ and $\text{Var}(y_{isl}) = \sigma^2$ and that $p_{is} > 0$ for all i, s .

- As $n_i \rightarrow \infty$ for all i ,

$$\text{MSE} \rightarrow_p \sigma^2 + b \quad (3)$$

where $\rightarrow_p$ means convergence in probability and

$$b = \lim \frac{1}{n-I} \sum_{i=1}^I (n_i - 1) \tau_i^2$$

where

$$\tau_i^2 = \sum_{s=1}^S p_{is} (\mu_{is} - \mu_i)^2 \quad (4)$$

- ii. When $I = 2$, under H_0 in Eq. 2,

$$MSTR \rightarrow_d (\sigma^2 + a)\chi_1^2 \quad (5)$$

where $\rightarrow_d$ means convergence in distribution and

$$a = \lim_{I \rightarrow \infty} \frac{1}{I-1} \sum_{i=1}^I \left(1 - \frac{n_i}{n}\right) \tau_i^2$$

Furthermore, if μ_i 's are related so that

$$c = \lim_{I \rightarrow \infty} \frac{1}{I-1} \sum_{i=1}^I n_i \left(\mu_i - \frac{1}{n} \sum_{i=1}^I n_i \mu_i \right)^2$$

is finite, then

$$MSTR \rightarrow_d (\sigma^2 + a)\chi_1^2 \left(\frac{c}{\sigma^2 + a} \right) \quad (6)$$

where $\chi_q^2(\delta)$ denotes a noncentral chi-square random variable with q degrees of freedom and the noncentrality parameter δ . We adopt this notation throughout this paper and when $\delta = 0$, $\chi_q^2(0)$ is abbreviated as χ_q^2 , the central chi-square random variable.

- iii. When $I \geq 3$, under H_0 in Eq. 2, MSTR converges in distribution to a linear combination of $I - 1$ independent central chi-square random variables with 1 degree of freedom.

As can be observed in Theorem 1, in the special case that $I = 2$ treatments, the asymptotic size (when $n_i \rightarrow \infty$) of the LOCF test, when H_0 in Eq. 2 is used as the null hypothesis, is less than or equal to the nominal level α if and only if $a \leq b$, or, equivalently,

$$\lim \left(\frac{n_2 \tau_1^2}{n} + \frac{n_1 \tau_2^2}{n} \right) \leq \lim \left(\frac{n_1 \tau_1^2}{n} + \frac{n_2 \tau_2^2}{n} \right) \quad (7)$$

where τ_i^2 is defined in Eq. 4. If n_1 and n_2 are almost the same, i.e., $\lim \frac{n_1}{n} = \lim \frac{n_2}{n}$, then the equality holds and the LOCF test is asymptotically correct in the sense that the asymptotic size of the LOCF test is α regardless of whether or not the dropout is informative. When $\lim \frac{n_1}{n} \neq \lim \frac{n_2}{n}$, the asymptotic size of the LOCF test is not α unless $\tau_1^2 = \tau_2^2$. However, in practice, this occurs only when $\tau_1^2 = \tau_2^2 = 0$, i.e., $\mu_{is} \equiv \mu_i$ for all s , that means the dropout is noninformative.

When $I = 2$ but $\tau_1^2 \neq \tau_2^2$ and $n_1 \neq n_2$, the LOCF test has an asymptotic size smaller than α if

$$(n_2 - n_1)\tau_1^2 < (n_2 - n_1)\tau_2^2 \quad (8)$$

or larger than α if $<$ in Eq. 8 is replaced by $>$. If treatment 1 is the placebo, then often $n_1 < n_2$ and $\tau_1^2 < \tau_2^2$ (i.e., the

variation among μ_{2s} 's is larger than that among μ_{1s} 's) and therefore Eq. 8 holds and the LOCF test is too conservative in the sense that its size is too small. Note that if the size of a test is unnecessarily small, the power of detecting treatment effects is unnecessarily low, which is a disadvantage for the drug companies.

When $I \geq 3$, the asymptotic size of the LOCF test when H_0 is used as the null hypothesis is generally not α .

VALIDITY OF THE LOCF ANOVA TEST UNDER A TWO-WAY ANOVA MODEL

Consider a parallel-group design for comparing I treatments, one of which could be a placebo or an active control. Suppose a total of n patients are randomly assigned to I treatment groups and J treatment centers, where the i th treatment group has n_i patients ($i = 1, \dots, I$) and of which n_{ij} patients at center j ($j = 1, \dots, J$). The proportion of patients allocated to treatment i at center j is fixed as $p_{ij} = (n_{ij}/n_i > 0)$. Suppose that patients are scheduled to be evaluated at S postbaseline time points (visits) to complete the study. However, owing to some reasons, in the i th treatment and j th center, at the s th visit ($s = 1, \dots, S$), n_{ijs} patients have their evaluations valued as y_{ijsl} , $l = 1, \dots, n_{ijs}$ and then drop out after the visit. Note that n_{ijs} 's are random but they satisfy the constraint $\sum_{s=1}^S n_{ijs} = n_{ij}$.

Similar to the case of one-way ANOVA, under the ITT analysis, the patient population under the treatment i at the center j consists of S subpopulations, where the (i, j, s) th population is the population of patients who are on therapy i at the center j and who drop out after the s th visit. Let μ_{ijs} be the mean of the (i, j, s) th subpopulation. Then global mean of the patient population under the treatment i is

$$\mu_i = \sum_{j=1}^J p_{ij} \sum_{s=1}^S p_{ijs} \mu_{ijs}$$

where p_{ijs} is the true proportion of patients in the (i, j, s) th subpopulation under the treatment i at center j . If we use the global mean as the measure of treatment effect, then the null hypothesis of no treatment effect should be

$$H_0 : \mu_1 = \dots = \mu_I \quad (9)$$

Two-Way ANOVA Model Without Interaction

In some cases, it may be reasonable to assume that there is no interaction between the treatments and centers in the sense that the change of the treatments is not affected across the centers. Under the additive two-way ANOVA model, the ITT LOCF test assumes treatment and center

as fixed factors and concludes that a treatment effect exists if and only if

$$\frac{\text{MSTR}}{\text{MSE}} > F_{I-1, n-I, \alpha}$$

where

$$\text{MSTR} = \frac{1}{I-1} \sum_{i=1}^I n_i (\bar{y}_{i...} - \bar{y}_{...})^2$$

$$\text{MSE} = \frac{1}{n-IJ} \sum_{i=1}^I \sum_{j=1}^J \sum_{s=1}^S \sum_{l=1}^{n_{ijs}} (y_{ijsl} - \bar{y}_{ij..})^2$$

$$\bar{y}_{i...} = \frac{1}{n_i} \sum_{j=1}^J \sum_{s=1}^S \sum_{l=1}^{n_{ijs}} y_{ijsl}, \quad \bar{y}_{...} = \frac{1}{n} \sum_{i=1}^I \sum_{j=1}^J \sum_{s=1}^S \sum_{l=1}^{n_{ijs}} y_{ijsl}$$

and

$$\bar{y}_{ij..} = \frac{1}{n_{ij}} \sum_{s=1}^S \sum_{l=1}^{n_{ijs}} y_{ijsl}$$

The following theorem, given in Shao and Zhong,^[9] summarizes the properties of the LOCF test under the additive model.

Theorem 2

Suppose that y_{ijsl} 's are independent with $\mu_{ijs} = E(y_{ijsl})$ and $\text{Var}(y_{ijsl}) = \sigma^2$ and that $p_{ijs} > 0$ for all i, j, s . Denote

$$\mu_{ij} = \sum_{s=1}^S p_{ijs} \mu_{ijs}, \quad \mu_i = \sum_{j=1}^J p_{ij} \mu_{ij}$$

$$\tau_{ij}^2 = \sum_{s=1}^S p_{ijs} (\mu_{ijs} - \mu_{ij})^2, \quad \tau_i^2 = \sum_{j=1}^J p_{ij} \tau_{ij}^2$$

and

$$\bar{y}_{ij..} = \frac{1}{n_{ij}} \sum_{s=1}^S \sum_{l=1}^{n_{ijs}} y_{ijsl}$$

i. As $n \rightarrow \infty$,

$$\text{MSE} \rightarrow_p \sigma^2 + b \quad (10)$$

where

$$b = \lim \frac{1}{n-IJ} \left(\sum_{i=1}^I n_i \tau_i^2 - \sum_{i=1}^I \frac{1}{n_i} \sum_{j=1}^J \tau_{ij}^2 \right) \quad (11)$$

ii. Under the additive two-way ANOVA model, when $I = 2$, if μ_i 's are related so that

$$c = \lim \sum_{i=1}^2 n_i \left(\mu_i - \frac{1}{n} \sum_{i=1}^2 n_i \mu_i \right)^2$$

is finite, then

$$\frac{\text{MSTR}}{\text{MSE}} \rightarrow_d \frac{\sigma^2 + a}{\sigma^2 + b} \chi_1^2 \left(\frac{c}{\sigma^2 + a} \right) \quad (12)$$

where

$$a = \lim \sum_{i=1}^2 \left(1 - \frac{n_i}{n} \right) \tau_i^2 \quad (13)$$

Furthermore, under H_0 in Eq. 9,

$$\frac{\text{MSTR}}{\text{MSE}} \rightarrow_d \frac{\sigma^2 + a}{\sigma^2 + b} \chi_1^2 \quad (14)$$

iii. Under the additive two-way ANOVA model, when $I \geq 3$, under H_0 in Eq. 9, MSTR converges in distribution to a linear combination of independent central chi-square random variables with 1 degree of freedom.

When there are only two treatment groups (i.e., $I = 2$), Theorem 2 implies that the LOCF test is asymptotically correct if $\lim \frac{n_1}{n} = \lim \frac{n_2}{n}$. That is, when n_1 and n_2 are nearly similar, regardless of whether the dropout is informative or not. And when $\lim \frac{n_1}{n} \neq \lim \frac{n_2}{n}$, the asymptotic size of LOCF test is not α unless the dropout is noninformative. When $\tau_1^2 \neq \tau_2^2$ and $n_1 \neq n_2$, the LOCF test has an asymptotic size smaller than α if

$$(n_2 - n_1) \tau_1^2 < (n_2 - n_1) \tau_2^2 \quad (15)$$

or larger than α if $<$ in Eq. 8 is replaced by $>$. If treatment 1 represents the placebo, then it is often the case that $n_1 < n_2$ and $\tau_1^2 < \tau_2^2$ (i.e., the variation among μ_{2js} 's is larger than that among μ_{1js} 's) and therefore Eq. 8 holds and the LOCF test is too conservative in the sense that its size is too small. Note that if the size of a test is unnecessarily small, the power of detecting treatment effects is unnecessarily low, which is a disadvantage for the drug companies.

When $I \geq 3$, the asymptotic size of the LOCF test when H_0 in Eq. 9 is used as the null hypothesis is generally not α .

Two-Way ANOVA Model with Interaction

In many situations, it may not be appropriate to assume an additive model in the first place. If that is the case, we have to consider a two-way ANOVA model with interaction between the treatments and centers. To test the treatment effect for the two-way ANOVA model with interaction, the interaction effect has to be tested first.

Otherwise, the result for testing the treatment effect cannot be interpreted in a statistically meaningful manner. It is widely recognized that in testing the interaction, when the design is unbalanced, there exists no closed form based on the usual ANOVA method (i.e., the method of sum of squares). To give a clear description of the LOCF test under the interaction model, we consider the balanced case when $n_{ij} \equiv m > 1$, and denote $n_{ijs} = m_s$. When testing for the interaction, the LOCF test concludes that there is an interaction effect if

$$\frac{MSAB}{MSE} > F_{(I-1)(J-1), IJ(m-1), \alpha}$$

where

$$MSAB = \frac{m}{(I-1)(J-1)} \sum_{i=1}^I \sum_{j=1}^J \times (\bar{y}_{ij..} - \bar{y}_{i...} - \bar{y}_{.j..} + \bar{y}_{....})^2$$

and

$$\begin{aligned} \bar{y}_{ij..} &= \frac{1}{m} \sum_{s=1}^S \sum_{l=1}^{m_s} y_{ijsl} \\ \bar{y}_{i...} &= \frac{1}{mJ} \sum_{j=1}^J \sum_{s=1}^S \sum_{l=1}^{m_s} y_{ijsl} \\ \bar{y}_{.j..} &= \frac{1}{mI} \sum_{i=1}^I \sum_{s=1}^S \sum_{l=1}^{m_s} y_{ijsl} \\ \bar{y}_{....} &= \frac{1}{mIJ} \sum_{i=1}^I \sum_{j=1}^J \sum_{s=1}^S \sum_{l=1}^{m_s} y_{ijsl} \end{aligned}$$

The following result addresses the question regarding the validity of the LOCF test in the balanced case (see, e.g., Ref. [10]).

Theorem 3

Under the balanced two-way ANOVA model with interaction,

- i. *when $I = J = 2$, as $m \rightarrow \infty$, for testing of interaction, the LOCF test is asymptotically correct;*
- ii. *when $I = 2$ and $J \geq 3$, for testing Eq. 2, a sufficient condition for LOCF to be valid is τ_{ij}^2 are the same for each fixed i , where $\tau_{ij}^2 = \sum_{s=1}^S p_{ijs} (\mu_{ijs} - \mu_{ij})^2$.*

In the unbalanced case, the LOCF test is much more complicated. Cheng and Shao^[10] proved the following result:

Theorem 4

Assume the unbalanced two-way ANOVA model with interaction. When $I = J = 2$, as $n_{ij} \rightarrow \infty$, the LOCF test is asymptotically valid for testing Eq. 2 if

$$\lim \frac{n_{11}}{n} = \lim \frac{n_{12}}{n} = \lim \frac{n_{21}}{n} = \lim \frac{n_{22}}{n}$$

that is, the cell frequencies n_{11} , n_{12} , n_{21} , and n_{22} are nearly the same, or the design is almost balanced. When $I = J = 2$ is not true, the asymptotic size of the LOCF test for interaction is generally not α .

The above theorem indicates that, under the two-way ANOVA model with interaction, the LOCF test for treatment effect is valid only for the very special case consisting of two treatments and two centers—which is not practically possible. For all the other cases, the asymptotic size of the LOCF test for treatment effect described in the section “Two-way ANOVA Model Without Interaction” is generally not equal to the nominal size α .

ASYMPTOTICALLY CORRECT TESTS

We have shown that the LOCF test is asymptotically correct in some special situations. In those situations, the LOCF method may be used because of its simplicity. However, when the LOCF test for treatment effect is invalid (in the sense that its asymptotic size is not equal to the nominal size α), asymptotically correct tests should be used. Under the one-way ANOVA model, it is recommended that the test proposed by Shao and Zhong,^[5] which is always asymptotically valid, be used.

Not that the poststratified sample mean under treatment i can be written as

$$\hat{\mu}_i = \sum_{s=1}^S \frac{n_{is}}{n_i} \bar{y}_{is.} = \frac{1}{n_i} \sum_{s=1}^S \sum_{l=1}^{n_{is}} y_{isl} = \bar{y}_{i..}$$

It can be verified that $\hat{\mu}_i$ is an unbiased estimator of μ_i and its variance is

$$V_i = \frac{1}{n_i} \left(\sum_{s=1}^S p_{is} \sigma_{is}^2 + \tau_i^2 \right)$$

where $\sigma_{is}^2 = \text{Var}(y_{isl})$. Note that σ_{is}^2 's are allowed to be different. An approximately unbiased estimator of V_i is

$$\hat{V}_i = \frac{1}{n_i(n_i - 1)} \sum_{s=1}^S \sum_{l=1}^{n_{is}} (y_{isl} - \bar{y}_{i..})^2$$

Define

$$T = \sum_{i=1}^I \frac{1}{\hat{V}_i} \left(\bar{y}_{i..} - \frac{\sum_{i=1}^I \bar{y}_{i..} / \hat{V}_i}{\sum_{i=1}^I 1 / \hat{V}_i} \right)^2$$

Thus we have the following result (see Ref. [5]).

Table 1 Subpopulation Parameters

Parameter	Treatment $i = 1$ (placebo)				Treatment $i = 2$ (drug)			
	Visit 1	Visit 2	Visit 3	Global	Visit 1	Visit 2	Visit 3	Global
p_{ij}	0.20	0.20	0.60		0.15	0.15	0.70	
μ_{ij}	25	24	21		40	32	16.6	
σ_{ij}^2	100	100	100		100	100	100	
μ_I				22.4				22.4
τ_i^2				3.04				84.1

Table 2 Simulation Result (2000 Simulations)

$n_1 + n_2$	$p_1 = 30\%$		$p_1 = 50\%$		$p_1 = 70\%$	
	LOCF test	SZ test	LOCF test	SZ test	LOCF test	SZ test
40	0.0275	0.0600	0.0465	0.0535	0.0690	0.0555
80	0.0295	0.0540	0.0485	0.0520	0.0740	0.0520
160	0.0285	0.0540	0.0460	0.0460	0.0755	0.0480

Theorem 5

Under the hypothesis H_0 in Eq. 2, as $n_i \rightarrow \infty$ for all i ,

$$T \rightarrow_d \chi_{I-1}^2$$

Furthermore, if μ_i 's are related so that

$$\tilde{c} = \lim \sum_{i=1}^I \frac{1}{V_i} \left(\mu_i - \frac{\sum_{i=1}^I \mu_i / V_i}{\sum_{i=1}^I 1/V_i} \right)^2$$

is finite, then

$$T \rightarrow_d \chi_{I-1}^2(\tilde{c})$$

An Example

To illustrate the performance of the LOCF test and the test suggested by Shao and Zhong^[5] (SZ test), we present part of the simulation result in Shao and Zhong.^[5] We consider the case when there are two treatments and the total number of complete visits is three for each treatment (i.e., $I = 2$, $S = 3$). The subpopulations are normally distributed with the following parameters (Table 1).

The parameters for the subpopulations are chosen such that the global means are similar for the two treatments while the means for the subpopulations are so dispersed that τ_1^2 and τ_2^2 are very different.

For the fixed nominal level of significance $\alpha = 0.05$, the simulated (approximated) probability of type 1 error (based on 2000 simulations) are presented for the total sample size $n = n_1 + n_2 = 40, 80, 160$, respectively. And

for each fixed sample size, three different percentage of n_1 in $n = n_1 + n_2$, namely, $p_1 = n_1/n = 30\%, 50\%, 70\%$ are considered. The simulation result is given in the following table.

Table 2 indicates that when $p_1 = 50\%$, i.e., when $n_1 = n_2$, the probability of the type 1 error of the LOCF test is close to $\alpha = 0.05$, but is considerably small when $p_1 = 30\%$ while considerably large when $p_1 = 70\%$. The probability of the type 1 error of Shao and Zhong^[5] test is close to α regardless of whether n_i 's are equal or not.

Note that under the two-way ANOVA model, the construction of asymptotically correct tests is more complicated. More details can be found in Cheng and Shao^[10] and Shao and Zhong.^[9]

CONCLUSION

We have shown that the LOCF test is asymptotically correct in some special situations, where the LOCF method may be used because of its simplicity. However, when the LOCF test for treatment effect is invalid (in the sense that its asymptotic size is not equal to the nominal size α), asymptotically correct tests should be used. Under the one-way ANOVA model, it is recommended that the asymptotically correct test proposed by Shao and Zhong^[5,9] or Cheng and Shao^[10] be used.

The validity of LOCF examined in this entry does not incorporate the influence of clinical relevant variables such as demographics and patient characteristics (e.g., disease status and/or medical history) at baseline. In practice, it is suggested that those variables be included in the model for evaluation of the validity of LOCF. This, however, requires further research.

REFERENCES

1. Heyting, A.; Tolboom, J.T.B.M.; Essers, J.G.A. Statistical handling of drop-outs in longitudinal clinical trials. *Stat. Med.* **1992**, *11*, 2043–2061.
2. FDA. *Guideline for Format and Content of the Clinical and Statistical Sections of New Drug Applications*; Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, MD, 1988.
3. ICH. International Conference on Harmonization. Guidelines on Statistical Principles for Clinical Trials (E9), 1997.
4. Knickerbocker, R.K. Intention-to-Treat Analysis. In *Encyclopedia of Biopharmaceutical Statistics*; Chow, S., Ed.; Marcel Dekker, Inc: New York, 2000.
5. Shao, J.; Zhong, B. Last observation carry-forward and last observation analysis. *Stat. Med.* **2002**. in press.
6. Lavori, P.W. Clinical trials in psychiatry: Should protocol deviation censor patient data? *Neuropsychopharmacology* **1992**, *6*, 39–48.
7. Dawson, J.D. Comparing treatment groups on the basis of slopes, areas-under-the-curve, and other summary measures. *Drug Inf. J* **1994**, *28*, 723–732.
8. Ting, N. Carry-Forward Analysis. In *Encyclopedia of Biopharmaceutical Statistics*; Chow, S., Ed.; Marcel Dekker, Inc: New York, 2000.
9. Shao, J.; Zhong, B. *Analysis of global means in clinical trials*; 2002. manuscript in preparation.
10. Cheng, B.; Shao, J. *Some Results for Testing Interactions in Two-Way ANOVA*; 2002. Manuscript in preparation.
11. McHugh, R.; Matts, J. Post-stratification in the randomized clinical trial. *Biometrics* **1983**, *39*, 217–225.

FURTHER READING

Dawson, J.D.; Lagakos, S.W. Size and power of two-sample tests of repeated measures data. *Biometrics* **1993**, *49*, 1022–1032.

Logistic Regression

Rick Chappell and Jason Fine

University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.

Abstract

Logistic regression is a model intended for situations in which the dependent variable Y consists of a binary (yes/no) outcome and the vector of independent variables or covariates $\mathbf{x}$, as in ordinary linear models, may either be continuous, dichotomous, or factors. The aim of this entry is to provide examples of this model to illustrate its application.

INTRODUCTION

Logistic regression is a model intended for situations in which the dependent variable Y consists of a binary (yes/no) outcome and the vector of independent variables or covariates $\mathbf{x}$, as in ordinary linear models, may either be continuous, dichotomous, or factors. Examples of binary outcomes include death/survival, toxicity/lack of toxicity, efficacy/inefficacy, and, in general, the occurrence of an event vs. otherwise. Y is typically coded as 1 if the event occurs and 0 if the event does not occur so that its mean $E(Y|\mathbf{x})$ is the probability of occurrence given covariates.

OVERVIEW

There are two reasons why the standard linear model is, in general, unsatisfactory for binary outcomes. The first is that linearity on the raw scale is unrealistic: The conditional mean of Y , $E(Y|\mathbf{x}) = \pi_{\mathbf{x}}$, is not constrained to lie between 0 and 1, whereas proportions must be so. The second is that for a binary outcome Y , $\text{Var}(Y|\mathbf{x}) = \pi_{\mathbf{x}}(1 - \pi_{\mathbf{x}})$ and so violates the usual linear model assumption of variance homogeneity across covariates.

As an example of the former, consider the example of menarche in Warsaw girls (Ref. 1, Section 7.2). Table 1 shows these data in grouped form giving for each group the total number of girls, the number who have achieved menarche, and the median age.

The proportions are plotted as dots in Fig. 1, in which dashed and solid curves give the fits of linear and logistic regressions, respectively. Note that the former exceeds 1 for ages over 16 years and becomes negative for those under 10 years, whereas the latter is mathematically prohibited from doing either. The logit function is sigmoidal or s-shaped so that the slope is tempered toward each extreme. Formally, if $\mathbf{x}_i$ is the known covariate vector for the i th subject, which in this example includes the i th girl's age, then

$$\begin{aligned}\text{logit}[E(\gamma_i | \mathbf{x}_i)] \\ &= \text{logit}[\text{Pr}(\text{menarche for subject } i \text{ given her age})] \quad (1) \\ &= \text{logit}(\pi_i) = \log[\pi_i/(1 - \pi_i)] \\ &= \beta_0 + \beta_1 \times \text{age}_i\end{aligned}$$

where β_0 and β_1 are unknown. In general, there is a vector of unknown parameters β and

$$\begin{aligned}\text{logit}[\text{Pr}(\text{event for subject } i | \mathbf{x}_i)] &= \log[\pi_i/(1 - \pi_i)] \\ &= \mathbf{x}_i^T \beta\end{aligned} \quad (2)$$

A constant term is usually included whose element of $\mathbf{x}$ is always 1. It corresponds to the intercept in a linear model and is denoted β_0 . As $\pi/(1 - \pi)$ is the odds of an event, logistic regression is a linear model on the logarithm of the odds. For two different subjects denoted i and j , their odds ratio of the event is defined as

$$\begin{aligned}\text{OR}(i, j) &= \left\{ \pi_i / (1 - \pi_i) \right\} / \left\{ \pi_j / (1 - \pi_j) \right\} \\ &= \exp\left\{ (\mathbf{x}_i - \mathbf{x}_j)^T \beta \right\}\end{aligned} \quad (3)$$

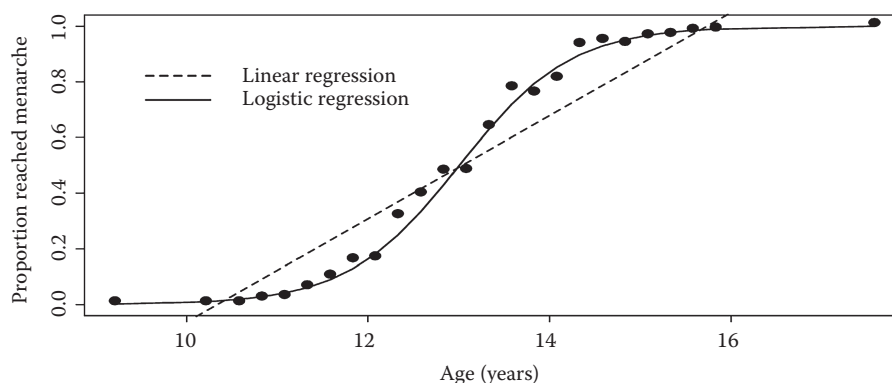
Note that because the logit function is symmetric, logistic regression is invariant to the definition of $Y = 1$ as an event or nonevent. That is, for example, one achieves the same predicted values whether “death” or “survival” is considered to be an event; odds ratios are merely inverted, parameter estimates are the same magnitude but negated, and standard errors and confidence intervals are identical under the two formulations.

Because, as mentioned above, $\pi_{\mathbf{x}}$ can be considered the conditional mean of the random variable Y , logistic regression is a member of the class of general linear models in which a function of the mean is linear in the covariates. Although we use the logit function here, others are practical and sometimes preferable for binary data. These are

Table 1 Menarche in Polish Girls

Age	9.2	10.2	10.6	10.8	11.1	11.3	11.6	11.8	12.1	12.3	12.7	12.8
Total	376	200	93	120	90	88	105	111	100	93	100	108
Menarche	0	0	0	2	2	5	10	17	16	29	39	51

(Continued)

**Fig. 1** Proportion of Warsaw girls who have reached menarche vs. age

usually also sigmoidal and include the probit (inverse normal probability function), $-\log[-\log(\cdot)]$, and $-\log[1 - \log(\cdot)]$. See McCullagh and Nelder^[2] (Section 4.3.1) for more details.

MAXIMUM LIKELIHOOD ESTIMATION

By far, the most commonly used method of obtaining estimates for the coefficients β is that of maximum likelihood estimation. In this paradigm, a likelihood function of the parameters is generated for the sample in which higher values are supported by the data, and the value of β which produces the highest likelihood is the maximum likelihood estimate (MLE), often denoted as $\hat{\beta}$. Eq. 2 can be inverted with $\hat{\beta}$ substituted for β in order to produce a predicted or fitted probability for a vector of covariates $\mathbf{x}_i$:

$$\hat{\pi}_i = \frac{1}{1 + \exp(\mathbf{x}_i^T \hat{\beta})} \quad (4)$$

In addition, $\hat{\beta}$ can be substituted into Eq. 3 to produce estimated odds ratios. Various methods of inference on MLEs exist which generate tests, estimated standard errors, and confidence intervals using $\hat{\beta}$ and functions of it; see Appendix I of Cox and Snell^[3] for a detailed treatment. Almost all statistical packages give asymptotic standard errors for $\hat{\beta}$ and the resultant Wald tests for the null hypotheses that the components of β are 0. These can be unreliable when any component of $\hat{\beta}$ is large, in which case inference based on likelihood ratio statistics is preferred.^[4] A likelihood ratio test of the null hypothesis that one or more elements of β is zero is achieved by fitting both the

full model and one without the corresponding elements of $\mathbf{x}$. Twice the difference in the two log likelihoods is compared to a Chi-squared distribution with degrees of freedom equal to the number of parameters tested.

The method of iteratively reweighted least squares (IRLS) is often used to calculate MLEs (Ref. [2], Section 2.5). In IRLS, Eq. 2 is approximated by a linear model and then fitted. However, this approximation requires knowledge of β ; thus, initial values must be assumed. The fit then produces updated estimates of β which are reinserted, etc., until convergence, with the final estimate being $\hat{\beta}$. IRLS has the advantage of using well-known algorithms for least-square regression; in addition, useful diagnostics can be computed by running only one iteration of IRLS.

A limitation of inference based on unconditional maximum likelihood is its reliance on large-sample approximations. Alternative approaches based on conditional likelihoods may provide more accurate results. The idea is to condition on the sufficient statistic for the intercept in the logistic regression, which is the total number of successes; that is, the number of Y_i equaling one. Inferences about the remaining components of β may then be based on the likelihood for Y_i conditional on $\sum_i Y_i$ ^[5] and do not involve the nuisance parameter, which may cause instability with small to moderate sample sizes. An asymptotic analysis of this conditional likelihood gives distributional results for the Wald's score and likelihood ratio tests which are similar to those for the unconditional likelihood. More recently, so-called exact methods have been developed based on certain permutation distributions which give p values which are valid for all sample sizes.^[6] The methodology has been popularized in the widely used software LogXact.^[7]

Table 1 Menarche in Polish Girls (*Continued*)

Age	13.1	13.3	13.6	13.8	14.1	14.3	14.6	14.8	15.2	15.3	15.6	15.8	17.6
Total	99	106	105	117	98	97	120	102	122	111	94	114	1049
Menarche	47	67	81	88	79	90	113	95	117	107	92	112	1049

INTERPRETATION

The major disadvantage of the nonlinear form of Eq. 2 is that the resultant parameter estimates are not immediately interpretable. Returning to the menarche data of Table 1, consider the linear regression whose fit is represented by the straight line in Fig. 1. Its slope is 0.19/yr; thus, the corresponding estimated chance of reaching menarche increases by 19%/yr, an easily explainable (if obviously incorrect) figure. Logistic regression produces the estimated coefficient $\hat{\beta} = 1.63/\text{yr}$, the amount by which the logarithm of the estimated odds of menarche increases for each year a girl's age advances. This is an unintuitive result. One way in which to make it slightly less is by exponentiating it to form an odds ratio: $\exp(1.63) = 5.11$ so that we can say that adding a year multiplies the odds of menarche by this factor. We would say that 5.11 is the "estimated odds ratio of menarche with respect to age, in years."

If there were other covariates in this example besides age, we would call its exponentiated fitted coefficient the "estimated odds ratio of menarche with respect to age, in years, controlling [or adjusting] for all other factors." This is equivalent to considering two subjects in Eq. 3 for whom $\text{age}_i - \text{age}_j = 1$, but whose other elements of $\mathbf{x}$ are the same.

Finally, one can be dissatisfied with odds ratios because they are not otherwise commonly used in medical discourse in preference for percentages.^[8] Although an exact conversion cannot be made because the logistic model is not linear in percentages, useful approximations do exist. For example, the steepest slope always occurs at the median (about 13 years in Fig. 1) and is computed by multiplying the estimated coefficient by a factor of 25. Thus, for 13 year olds, the percentage of girls reaching menarche increases by $25 \times 1.63 = 41\%/\text{yr}$. The general factor for approximate conversion to linearity is

$$\text{conversion factor} = 100\% \times p \times (1 - p) \quad (5)$$

where p is some relevant proportion (here 0.5, but often, the overall sample proportion of events).

GOODNESS-OF-FIT DIAGNOSTICS

The practical utility of the logistic model is its ability to estimate summary risks for covariates taking a wide range of values. The most appropriate interpretation of the results involves the interpolation of the model over the range of the

covariates used in the model fitting. The common wisdom is that extrapolation may give misleading results, which are driven by modeling assumptions which cannot be verified with the observed data. Of course, model misspecification is also important when interpolating and is often ignored. For this scenario, a variety of techniques are available for model assessment and should be used whenever possible.

In linear regression, the building block for model diagnosis is the residual, i.e., the difference between the observed response and its expected value under the assumed model. Plotting the fitted residuals vs. the covariates may indicate lack of fit when the residuals show systematic deviation from zero. In certain cases, the functional forms of covariates may be identifiable from the plots. Because logistic regression is inherently nonlinear and the binary response is highly discrete with an asymmetric, skewed distribution, these nice properties do not carry over to an analogously defined residual from the logistic regression model. Such scatterplots are notoriously uninformative. However, displaying the results of a nonparametric regression of the residuals on a covariate overlaid on the plot may reveal nonlinearities, which suggest alternative forms of the covariate.

The so-called Pearson residuals, the raw residuals divided by standard error estimates, may be used to construct Pearson-type tests^[9,10] which may be helpful in assessing overall goodness of fit. Unfortunately, such tests may require an arbitrary partitioning of the covariate space which may be unattractive with continuous covariates. In addition, the tests may not be helpful for evaluating individual covariates. A simple approach to this problem is to include other forms of the covariate to check the fit of the form under consideration. For example, with scalar $\mathbf{x}_i$, one may diagnose the appropriateness of the covariate form by fitting a model with $\mathbf{x}_i$ and $\mathbf{x}_i^2$ and evaluating the significance of $\mathbf{x}_i^2$.

To further explore misspecification of the covariates, one may employ the generalized additive model.^[11] In this formulation, $\mathbf{x}_i^T \boldsymbol{\beta}$ is replaced by a sum of unknown functions of $\mathbf{x}_i$ in the logistic model. Here the functions denote the individual covariate effects. If the simple logistic model is correctly specified, then for component j of $\mathbf{x}_i$, say, x_{ij} , the corresponding function is $x_{ij}\beta_j$, where β_j is component j of $\boldsymbol{\beta}$. The estimated functions may be plotted and evaluated visually for departures from linearity. These plots are generally quite useful and may suggest other forms of the covariates. Formal tests for the functional forms of the covariates have not been well studied using this technique.

In some cases, the logistic function relating the binary probability to the covariates may be invalid. To determine whether this is the case, specialized tests^[12] may be used. Alternatively, one may fit a model with a flexible link function involving additional parameters which are simultaneously estimated with the covariate effects using maximum likelihood. This is the approach advocated by Stukel,^[13] where a number of link classes are compared and contrasted, each containing the logistic model as a special case. In practice, it may be difficult to estimate the link parameter well, except in large samples. The menarche example of Table 1 may be such a case. Taylor^[14] has examined the variance inflation for the covariate effects associated with estimating the link parameter when it is known. In general, the cost decreases as the number of covariates in the model increases.

Outlying observations and influential points are common problems in linear regression. Pearson residuals are helpful in identifying outliers, and leverages derived from the hat matrix may provide insight into the influence of individual observations. These techniques are not easily adapted to logistic regression. Alternative procedures based on score and deviance residuals are discussed in Williams^[15] and Davison and Tsai;^[16] they are broadly applicable to assessing generalized linear models.

EXAMPLE: THE BRITISH INSTITUTE OF RADIOLOGY TRIALS OF RADIOTHERAPY IN CANCER OF THE LARYNX AND THE PHARYNX

The British Institute of Radiology (BIR)^[17,18] conducted two large-scale randomized clinical trials to assess the effectiveness of different treatment radiotherapeutic schedules for cancer of the larynx and the pharynx. For the purposes of this analysis, the studies' results are merged; unlike the primary analyses, which were simple comparisons based upon the randomization, we will here model patient outcome as a function of the treatment characteristics in order to gain insight into the underlying radiobiological mechanisms. The event here is a 3-year

local control. There are four components of the covariate vector $\mathbf{x}$: D , the total dose in grays (Gy); DD/f , the “fractionation factor” in square Gy which is large when treatment is administered in many separate increments and small otherwise; t , the total time of treatment in days; and stage, a three-level covariate which represents increasing invasiveness of the tumor (which, as a factor, is coded as two 0/1 elements in $\mathbf{x}$ using stage I as the baseline for comparison). The elements of β corresponding to these predictive factors are β_D , $\beta_{DD/f}$, β_t , and β_{stage} along with the usual constant term β_0 . We fit this model using information from the 858 usable laryngeal cancer patients with negative nodal and metastatic status in both studies, of whom 590 had 3-year local control. See the references cited above for further details, theoretical justification of the model, and analyses.

Table 2 gives MLEs for the BIR data using the logistic model and covariates described above. Although these results do not yield any surprises, they bear commentary and interpretation (as always in any scientific presentation). We compute and display $\hat{\beta}_0$ because of its use for computing fitted values in Eq. 4; however, a p value for the constant term is almost never relevant and should not be stated. For the other covariates, the “Percentage scale” column shows the approximate effects on the 100% $\times$ probability scale. These were obtained by substituting the proportion of events $p = 590/858 = 0.69$ into Eq. 5 to get a conversion factor of 21.5%. Thus, $\hat{\beta}_D = 0.05912$ was multiplied by $100\% \times 0.215$ to get an approximate change per Gy on the raw scale of 1.3%. This implies that each Gy of total dose increases local tumor control by 1.3%; an extra 10 Gy increases control rates by 13%; etc. (here and below, the calculations assume that all other covariates are held constant). A delay of 10 days would decrease local control by nearly 9%. The fractionation factor DD/f is not very important either in terms of the clinical significance of $\hat{\beta}_{DD/f}$ or its p value. The effects of tumor stage are large and in the expected direction; thus, the estimated local control among patients with stage II and stage III is 15% and 22% lower than in patients with stage I invasive tumors, respectively. All these effects are plotted in Fig. 2.

Table 2 Estimated Logistic Regression Model for the BIR Data

Variable	Coefficient estimate	Standard error	p Value	Percentage scale
Constant	−0.937	0.979		
D	0.0591/Gy	0.0186	0.002	1.4%/Gy
DD/f	0.00221/Gy ²	0.00285	0.4	0.047%/Gy ²
t	−0.0419/day	0.0124	0.0007	−0.89%/day
Stage II	−0.711	0.188	<0.001	−15%
Stage III	−1.044	0.187	<0.001	−22%

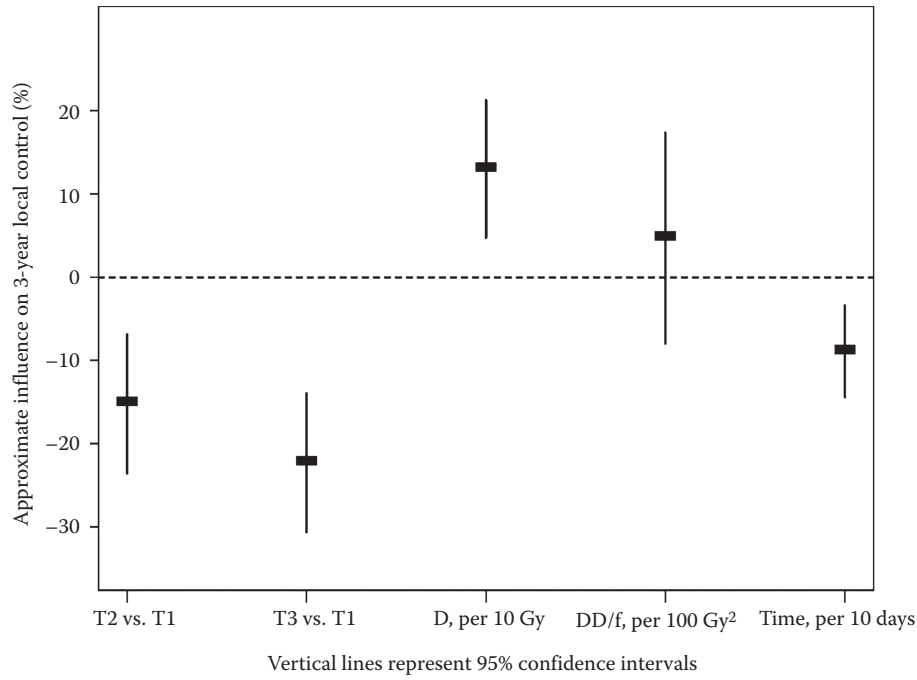


Fig. 2 A graphical display of the coefficients in Table 2

Alternatively, and more commonly, the coefficients can be exponentiated to form odds ratios. The odds ratio of local control with respect to total dose is $\exp^{0.0591} = 1.061/\text{Gy}$; the odds ratio of control with respect to time is 0.96/day or 0.66 per 10 days; the odds ratio of control with respect to stage II vs. I is 0.49; and with respect to stage III vs. I, it is down to 0.35. As before, these should be interpreted as the estimated effects after controlling for all other predictors in the model.

These results also directly address other interesting clinical questions. For example, some patients in the second study received 50 Gy in 16 fractions over 3 weeks and others received 63 Gy in 31 fractions over 6 weeks. If both such patients had stage I tumors, we can substitute ($D = 50$ Gy, $DD/f = 156.3$ Gy², $t = 21$ days) and ($D = 63$ Gy, $DD/f = 128$ Gy², $t = 42$ days) into Eq. 4 using the estimated coefficients in Table 2 to yield the estimated local control percentages of 82% and 79%. These give an estimated odds ratio of 1.19 although we can use Eq. 3 as a shortcut

$$\begin{aligned} & \log[\widehat{\text{OR}}(50\text{Gy}/156.3\text{Gy}^2/21\text{days}, \\ & \quad 63\text{Gy}/128\text{Gy}^2/42\text{days})] \\ &= \hat{\beta}_D(50 - 63) + \hat{\beta}_{DD/f}(156.3 - 128) + \hat{\beta}_t(21 - 42) \end{aligned} \quad (6)$$

resulting in the same answer. This formulation has the advantage of making it simple to compute an estimated variance of the log odds ratio:

$$\begin{aligned} & \widehat{\text{Var}}[\log[\widehat{\text{OR}}(50\text{Gy}/156.3\text{Gy}^2/21\text{days}, \\ & \quad 63\text{Gy}/128\text{Gy}^2/42\text{days})]] \\ &= \widehat{\text{Var}}(\hat{\beta}_D) \times (50 - 63)^2 + \widehat{\text{Var}}(\hat{\beta}_{DD/f}) \\ & \quad \times (156.3 - 128)^2 + \widehat{\text{Var}}(\hat{\beta}_t) \times (21 - 42)^2 \\ & \quad + \widehat{\text{Cov}}(\hat{\beta}_D, \hat{\beta}_{DD/f}) \times (50 - 63) \times (156.3 - 128) \\ & \quad + \widehat{\text{Cov}}(\hat{\beta}_D, \hat{\beta}_t) \times (50 - 63) \times (21 - 42) \\ & \quad + \widehat{\text{Cov}}(\hat{\beta}_{DD/f}, \hat{\beta}_t) \times (156.3 - 128) \times (21 - 42) \end{aligned} \quad (7)$$

where the estimated variances and covariances are the relevant elements of the asymptotic covariance matrix (inverted Fisher information matrix) calculated by most logistic regression programs. One uses this to compute a confidence interval for the log OR which is then exponentiated to create a confidence interval for the OR of (0.83, 1.71). Therefore, the evidence cannot discriminate between the two regimens.

Finally, one can also depict these results by computing the fitted probabilities using Eq. 4 and plotting them. This is done in one dimension by varying a single predictor while holding the others fixed, or in two dimensions by varying two predictors. Fig. 3 shows the estimated probabilities of local control at 3 years for a range of total doses and times. For this purpose, tumor stage was set at I and the number of fractions f was fixed at 19 (this was thought to yield a more interpretable plot in this case

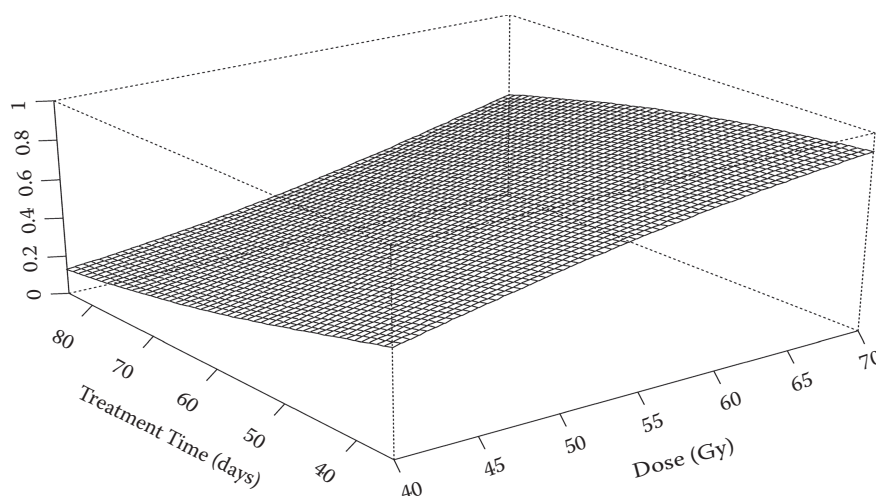


Fig. 3 Probability of local control at 3 years vs. dose and treatment time; 19 fractions, stage I

than just setting $DDIf$ to a constant value as could also be done). Fig. 3 shows how the estimated response surface of tumor control probability changes dramatically with time and dose.

REFERENCES

1. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S-PLUS*, 3rd Ed.; Springer-Verlag: New York, 1999.
2. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd Ed.; Chapman and Hall: New York, 1989.
3. Cox, D.R.; Snell, E.J. *Analysis of Binary Data*, 2nd Ed.; Chapman and Hall: New York, 1989.
4. Hauck, W.W.; Donner, A. Wald's test as applied to hypotheses in logit analysis. *J. Am. Stat. Assoc.* **1977**, *72*, 851–853.
5. Breslow, N.E.; Day, N.E. *Statistical Methods in Cancer Research. The Analysis of Case-Control Studies*; International Agency on Cancer: Lyon, 1980, Vol. 1.
6. Mehta, C.R.; Patel, N.R. Exact logistic regression: Theory and examples. *Stat. Med.* **1995**, *14*, 2143–2160.
7. Cytel Software Corporation, *LogXact-4 for Windows, a Software Package for Logistic Regression Featuring Exact Methods*; Cytel Software Corporation: Cambridge, MA, 2000.
8. Chappell, R. Presenting the coefficients of the linear quadratic formula for clinical use. *Int. J. Radiat. Oncol. Biol. Phys.* **1994**, *29*, 191–193.
9. Hosmer, D.W.; Lemeshow, S. *Applied Logistic Regression*, 2nd Ed.; John Wiley and Sons: New York, 2000.
10. Tsiatis, A.A. A note on a goodness-of-fit test for the logit regression without replication. *Biometrika* **1980**, *67*, 250–251.
11. Hastie, T.J.; Tibshirani, R.J. *Generalized Additive Models*; Chapman and Hall: New York, 1990.
12. Pregibon, D. Link Tests. In *Encyclopedia of Statistical Sciences*; Kotz, S., Johnson, N.L., Eds.; Wiley: New York, 1985, Vol. 5, 82–85.
13. Stukel, T.A. Generalized logistic models. *J. Am. Stat. Assoc.* **1988**, *83*, 426–431.
14. Taylor, J.M.G. The cost of generalizing logistic regression. *J. Am. Stat. Assoc.* **1988**, *83*, 1078–1083.
15. Williams, D.A. Generalized linear model diagnostics using the deviance and single case deletions. *Appl. Stat.* **1987**, *36*, 181–191.
16. Davison, A.C.; Tsai, C.-L. Regression model diagnostics. *Int. Stat. Rev.* **1992**, *60*, 337–353.
17. Chappell, R.; Nondahl, D.; Fowler, J.F. Modeling dose and local control in radiotherapy. *J. Am. Stat. Assoc.* **1995**, *90*, 829–838.
18. Chappell, R.; Nondahl, D.; Rezvani, M.; Fowler, J.F. Further analysis of the radiobiological parameters from the first and second British Institute of Radiology randomized studies of larynx/pharynx radiotherapy. *Int. J. Radiat. Oncol. Biol. Phys.* **1995**, *33*, 509–518.

Logistic Regression Process: Predictive Model Building in Clinical Research

Mo Liu

*National Clinical Research Center for Digestive Diseases, Beijing Friendship Hospital,
Capital Medical University, Beijing, China*

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

This article focuses on logistic regression process for model building, its validation, and its generalizability in case-control studies. The process includes determination of associations between potential risk factors and clinical outcomes, identification of risk factors for medical predictive model building, and internal/external validation of the established predictive model. The process starts with propensity score matching between the case and control groups followed by (i) descriptive analysis to have better understanding of the data, (ii) univariate analysis to test associations between the risk factors (or predictors) and the clinical outcomes (e.g., safety and efficacy), (iii) collinearity analysis to test associations/correlations between predictors (explanatory variables or risk factors), (iv) multivariate analysis to test the association of a predictor after adjusting for other predictors or potential confounders, and (v) model diagnostic/validation to determine the goodness of fit of the final model.

Keywords: Clinical research; Generalizability; Logistic regression process; Model validation; Predictive model building.

INTRODUCTION

In clinical research, case-control studies utilizing the existing database are commonly considered. A case-control study typically consists of two existing groups of patients, i.e., the group of cases and the group of control, which are identified and compared on the basis of some predetermined causal attributes. In practice, case-control studies are often conducted to identify the factors that may contribute to a medical condition by comparing subjects who have that condition and/or disease. These subjects are referred to as *cases*. Those subjects who do not have the condition and/or disease but have similar characteristics are considered as the *controls*. One of the most significant examples of the case-control study was the demonstration of the link between tobacco smoking and lung cancer.^[1] They found a statistically significant association between smoking and lung cancer in a large case-control study. The results, which were obtained based on nonrandomized, epidemiological observational studies rather than randomized trials, have been criticized in that this type of study cannot prove causation, but do confirm the causal link between smoking and lung cancer.

The primary objective of case-control studies is not only to study the relationship between clinical outcomes (e.g., safety and efficacy) and possible risk factors (e.g., demographics and patient characteristics), but also to develop a medical predictive model based on the identified risk factors. In recent years, it is also of particular interest to examine the generalizability of the established predictive model (e.g., from one patient population such as adults to another similar but different patient population such as pediatrics or from one medical center to another). For this purpose, multivariate (logistic) regression process is probably the most commonly used method. The use of logistic regression analysis for identifying potential risk factors (or predictors) in order to build a predictive model has become very popular in clinical research.^[2]

In practice, for a given database, the process of logistics regression analysis for model building in case-control studies starts with propensity score matching between the case and control groups followed by (i) descriptive analysis to have better understanding of the data, (ii) univariate analysis to test associations between the risk factors (or predictors) and the clinical outcomes (e.g., safety and efficacy), (iii) collinearity analysis to test associations/

correlations between predictors (explanatory variables or risk factors), (iv) multivariate analysis to test the association of a predictor after adjusting for other predictors or potential confounders, and (v) model diagnostic/validation to determine the goodness of fit of the final model. This step is usually considered as an internal validation. In addition, one may consider testing generalizability of the established medical predictive model from either one patient population (e.g., adults or elderly) to another similar but different patient population (e.g., pediatrics) or one medical center to another in different geographical location. Steps (i)–(iii) are usually implemented by the univariate logistic regression and steps (iv)–(v) are by the multiple logistic regression analysis. In this entry, the process of logistics regression analysis for model building including its generalizability is described.

PROPENSITY SCORE MATCHING

In clinical research, one of the major concerns of a case–control study is selection bias, which is often caused by a significant difference or imbalance between the case and control groups, especially for large observational studies.^[3–5] In this case, the target patient population under study in the control group may not be comparable to that in the case group. This selection bias could alter the conclusion of the treatment effect due to possible confounding effects. Consequently, the conclusion may be biased and hence misleading.

An example regarding the impact of imbalance between the case and control groups (e.g., in ethnic factors on the responses to therapeutics) is the epidermal growth factor receptor (EGFR) tyrosine kinase inhibitor gefitinib (Iressa). Recently, Iressa was approved in Japan and the United States for the treatment of non-small-cell lung cancer (NSCLC). The EGFR is a promising target anticancer therapy because it is more abundantly expressed in lung carcinoma tissue than in adjacent normal lung tissue. However, clinical trials have revealed a significant variability in the response to gefitinib with higher responses observed in Japanese patients than in a predominantly European-derived population (27.5% vs. 10.4%, in a multi-institutional phase II trial^[6]). Paez et al.^[7] also show that somatic mutations of the EGFR were found in 15 of 58 unselected tumors from Japan and 1 of 61 from the United States. Treatment with Iressa causes tumor regression in some patients with NSCLC, more frequently in Japan. Finally, the striking differences in the frequency of EGFR mutation and the response to Iressa between Japanese and American patients raise general questions regarding variations in the molecular pathogenesis of cancer in different ethnic, cultural, and geographic groups. Recently, generalibility of a new treatment has attracted more and more attention from sponsors as well as regulatory authorities; the key issues

such as when and how to address the geographic variations of efficacy and safety for the statistical inference are still some challenges.

To overcome this problem, Rosenbaum and Rubin^[3] proposed the concept of *propensity score* as a method to reduce the selection bias in observational studies. Propensity score is a conditional probability (or score) of the subject being in a particular group when given chosen characteristics. The use of propensity score helps to match the case and control groups based on baseline characteristics as in a randomized experiment for reducing bias. In other words, using propensity score enables the assessment of the treatment effects more reliably. Consider the case group as those who received a certain treatment ($T = 1$) and the control group as those who did not receive this treatment ($T = 0$). Let X be a vector that represents baseline demographics and/or patient characteristics that are important (e.g., possible confounding factors) for matching the case and control populations for reducing the selection bias. The propensity score is then given by

$$p(X) = \Pr[T = 1 | X] = E(T | X)$$

where $0 < p(X) < 1$. The exclusion of $p(X)$ extreme values such as 0 and 1 indicates that every subject has a nonzero change of being in the case or the control group. This condition is part of the assumption that treatment assignment should be strongly ignorable. Propensity score matching is a powerful tool reducing the bias due to possible confounding effects. Sometimes, propensity score matching is also considered as *post-study* randomization compared to a randomized clinical trial for reducing the bias.

In practice, several methods can be applied to build a propensity score model. One of the commonly used approaches is probably the *logistic regression* for binary outcomes, i.e., $T = 1$ for the case group and $T = 0$ for the control group. The propensity score $p(X)$ can be estimated based on the logistic regression model, where $p(X)$ is the conditional probability of being assigned to the case group given covariates X : $p(X) = \Pr[T = 1 | X]$. The natural log odds ratio of $p(X)$ equals a linear function of covariates X , and $p(X)$ is assumed to be between 0 and 1. This linear combination of X indicates that every increment of a component of x would add or subtract so much to the final conditional probability $p(X)$.

$$\begin{aligned} \text{logit}(p(X)) &= \log\left(\frac{p(X)}{1 - p(X)}\right) \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n = X^T \beta \end{aligned} \quad (1)$$

where X is the covariate matrix: $X = [1, X_1, X_2, \dots, X_n]^T$, and β is the coefficient matrix, which is given by $\beta = [\beta_0, \beta_1, \beta_2, \dots, \beta_n]^T$. Then, the propensity score $p(X)$ can be solved as follows:

$$p(X) = \frac{e^{x^T \beta}}{1 + e^{x^T \beta}} = \frac{1}{1 + e^{-x^T \beta}} \quad (2)$$

It should be noted that for propensity score matching, data available for matching will decrease as the number of matching factors (potential confounding factors) increases.

MODEL BUILDING

Model building involves univariate analysis for identifying risk factors (or predictors) and multivariate analysis or possible collinearity, which are critical for establishing an accurate and reliable predictive model and will be briefly described subsequently.

Univariate Analysis

Let Y be the outcome variable, which could be a discrete response variable, e.g., a binary response variable where $Y = 1$ if it is a success and $Y = 0$ if it is a failure or a categorical variable where $Y = 1$ if there is an improvement, $Y = 0$ where there is no change, or $Y = -1$ when the symptom has worsened. Also, let X be any type of covariate (e.g., continuous or dichotomous). For simplicity, consider a cirrhotic study. The 6-month survival of cirrhotic patients, X could be TCR functional status, TCR marked muscle wasting and serum creatinine, or demographics characteristics that could be potential risk factors (predictors) such as gender, age, or weight. Considering univariate logistic regression, the general model with on covariate is given by

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x \quad (3)$$

where π is the probability of success at a covariate level x . The logistic regression model can be rewritten as

$$\frac{\pi}{1-\pi} = e^{\alpha + \beta x} = e^{\alpha} \{e^{\beta}\}^x$$

where e^{β} represents the change in the log odds of the outcome by increasing x by one unit. In other words, every one unit increase in x increases the odds by a factor of e^{β} . More specifically, if $\beta = 0$ (i.e., $e^{\beta} = 1$), then the probability of success is the same at each level of x . When $\beta > 0$ (i.e., $e^{\beta} > 1$), the probability of success increases as x increases. Similarly, when $\beta < 0$ (i.e., $e^{\beta} < 1$), the probability of success decreases as x increases. To test the significance of the single variable model for the null hypothesis, $\beta = 0$, one can use the Wald test statistic^[8]:

$$Z = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

where $\hat{\beta}$ is the coefficient estimate of x and $SE(\hat{\beta})$ is the standard error of $\hat{\beta}$. Z follows the standard normal distribution with mean 0 and standard deviation 1. The null hypothesis will be rejected when $Z > Z_{1-\alpha/2}$ for a two-sided test at α level. Similarly, the log-likelihood test can also be used to determine whether x is a significant predictor of Y . Under the null hypothesis that $H_0: \beta = 0$, the log-likelihood test is given by

$$G = -2 \ln \left(\frac{\text{Likelihood without the variable}}{\text{Likelihood with the variable}} \right)$$

where G follows a chi-square distribution with 1 degree of freedom.^[8] The null hypothesis will be rejected when $G > \chi^2(1)$ for a two-sided test at α level.

Test for Collinearity

As indicated by Belsley et al.,^[9] two varieties are collinear if the data vectors representing them lie on the same line. Generally, k varieties are collinear when the vectors that represent them lie in a subspace of dimension less than k , or if one of the vectors is a linear combination of the others. Since collinear varieties do not provide information that is significantly different from the information provided by the other variables included, it would be difficult to make inference on the relationship between the covariates and the outcome. Therefore, a collinearity analysis between each pair of the predictors is necessary before building a multivariate model.

To test collinearity between quantitative variables, the Pearson correlation coefficient between independent predictors can be checked. The correlation coefficient between two variables (X_1, X_2) can be calculated as follows:

$$\rho_{X_1, X_2} = \frac{\text{cov}(X_1, X_2)}{\sigma_{X_1} \sigma_{X_2}}$$

where $\text{cov}(X_1, X_2)$ is the covariance between X_1 and X_2 , σ_{X_1} is the standard deviation of X_1 , and σ_{X_2} is the standard deviation of X_2 . For binomial or multinomial variables, the chi-square test can be used to check collinearity issues. The variance inflation factors (VIFs) can also be used for further measuring the collinearity in a multiple regression model, and it is defined as follows:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}, \quad j = 1, 2, 3, \dots, k$$

where R_j^2 is the squared multiple correlation based on regression X_j on the remaining $k - 1$ predictors.^[9] That is, $R_j^2 = R_{X_j|X_1, X_2, \dots, X_{j-1}, X_{j+1}, \dots, X_k}^2$. The larger the value of VIF, the more signal of collinearity will be between the independent predictors.

Once the highly correlated covariates are identified, decisions will be made regarding the variables that should be excluded from the next modeling building step. The variables to be included in the modeling building step should also have clinical significance. The reason is that removing some highly correlated variables will not only simplify the model built but also make the estimates of covariates more precise.

Multivariate Analysis

Similarly, the general logistic regression model with multiple covariates can be written as follows:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (4)$$

where log odds are a linear function of the covariates. To assess the significance of the model, the likelihood ratio test can be performed in the same way as described for the univariate testing step. The null hypothesis for the multivariable model is $H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$, and the alternative hypothesis is that at least one of the coefficients is not equal to zero. The test is given by

$$G = -2\ln\left(\frac{\text{Likelihood of the null model}}{\text{Likelihood of the full model}}\right)$$

The null hypothesis would be rejected for the two-sided test when $G > \chi^2_{\alpha/2}(k)$ at α level or when the p -value of the test is smaller than α (Hosmer and Lemeshow, 2013). To identify the significant risk factors, the Wald test should be performed for each predictor. The null hypothesis is $\beta_j = 0$, where $j \in 1, 2, \dots, k$, and the Wald test statistic is calculated as follows:

$$Z_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

where $\hat{\beta}_j$ is the coefficient estimate of x_j and $SE(\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$. Z follows a standard normal distribution with mean 0 and standard deviation 1. The null hypothesis will be rejected when $Z > Z_{1-\alpha/2}$ for a two-sided test at α level.

Model Selection

For model building, selection of an efficient subset of the explanatory variables or predictors that account for maximum variability in the outcome is the ultimate goal of the multiple logistic analysis. In practice, three model selection procedures commonly used are forward, backward, and stepwise processes. Forward selection procedure starts with an empty model and iteratively adds variables into the model one at a time. The variables added to the

model are those that can minimize the criterion mentioned previously. On the other hand, backward selection procedure starts with a full model and removes the least significant variable step by step. After fitting the model with all variables, the best model can be selected based on either the Akaike information criteria (AIC) or the Bayesian information criteria (BIC) given as follows:

$$\begin{aligned} \text{AIC} &= -2\log L(\hat{\pi}, y) + 2p \\ \text{BIC} &= -2\log L(\hat{\pi}, y) + 2p\ln(n) \end{aligned}$$

where p is the number of variables in the model of interest and n is the number of observations in the data.^[10] Both AIC and BIC add penalty to the criteria, and we can see from the definition that BIC has more penalty added compared to AIC. As a result, BIC tends to select a model with fewer variables compared to the model selected using AIC.

Model Diagnosis and Internal Validation

The developed model can then be tested for goodness of fit, which can serve as the diagnostic tool for model selection. Commonly considered goodness-of-fit tests include (i) Pearson goodness-of-fit test, (ii) deviance goodness-of-fit test, (iii) Hosmer and Lemeshow goodness-of-fit test (only for binomial outcome), (iv) pseudo R -square, (v) pseudo adjusted R -square, and (vi) score test for assumption of proportional odds (only for ordinal outcome). Detailed information regarding these tests can be found in the work of Hosmer and Lemeshow.^[2]

For the validation of the final model, an internal validation of the developed model is usually performed, whereas an external validation for assessment of its generalizability is often performed. In practice, a commonly considered procedure for model validation is to randomly split the data set into two sub-data sets: one contains 90% of the data and the other one contains 10% of the data. The data set that contains 90% of the data is usually referred to as *training* data set, which is used to build up the predictive model based on the identified possible risk factors (predictors). As indicated earlier, the commonly considered criterion for model selection is the so-called information criterion such as AIC, BIC, or Schwartz criteria. The second data set that contains 10% of the data is known as the *validation* data set, which is used to validate the selected final model. If the closeness of the predictive values and the observed responses is within a prespecified range, we claim that the model is validated.

MODEL GENERALIZABILITY

In clinical research, it is often of interest to generalize clinical results obtained from a given target patient population at a medical center to a similar patient population at another medical center. Thus, model generalizability is

also referred to as external validation as the established model is validated based on a similar but different database obtained either from a similar but different patient population (e.g., elderly patient population) or from a different medical research center. Denote the original target patient population by (μ_0, σ_0) and the similar but different patient population by (μ_1, σ_1) . Since the two populations are similar but different, it is reasonable to assume that $\mu_1 = \mu_0 + \varepsilon$ and $\sigma_1 = C\sigma_0$ ($C > 0$), where ε is referred to as the shift in location parameter (population mean) and C is the inflation factor of the scale parameter (population standard deviation). Thus, the (treatment) effect size adjusted for the standard deviation of population (μ_1, σ_1) can be expressed as follows:

$$E_1 = \frac{\left| \frac{\mu_1}{\sigma_1} \right|}{\left| \frac{\mu_0}{\sigma_0} \right|} = \frac{\left| \frac{\mu_0 + \varepsilon}{C\sigma_0} \right|}{\left| \frac{\mu_0}{\sigma_0} \right|} = |\Delta| \frac{\left| \frac{\mu_0}{\sigma_0} \right|}{\left| \frac{\mu_0}{\sigma_0} \right|} = |\Delta| E_0 \quad (5)$$

where $\Delta = (1 + \varepsilon/(\mu_0)/C)$, and E_0 and E_1 are the effect sizes (of clinically meaningful importance) of the original target patient population and the similar but different patient population, respectively. Chow et al.^[11] and Chow and Chang^[12] refer Δ as a sensitivity index measuring the change in the effect size between patient populations. As it can be seen from the above, if $\varepsilon = 0$ and $C = 1$ (there is no shift in the target patient population; the two patient populations are identical), $E_1 = E_0$. That is, the effect sizes of the two populations are identical. In this case, we claim that the results observed from the original target patient population (e.g., adults) can be generalized to the similar but different patient populations. Thus, if we can show that the sensitivity index Δ is within an acceptable range, say 80% and 120%, we may claim that the results observed at the original medical center can be generalized to another medical center with the similar but different patient populations.

As indicated in Chow et al.,^[13] the effect sizes of the two populations could be linked by baseline demographics or patient characteristics if there is a relationship between the effect sizes and the baseline demographics and/or patient characteristics (a covariate vector). However, such covariates may not exist or may not be observable in practice. In this case, Chow et al.^[13] suggested assessing the sensitivity index by simply replacing ε and C with their corresponding estimates. Intuitively, ε and C can be estimated by

$$\hat{\varepsilon} = \hat{\mu}_1 - \hat{\mu}_0 \text{ and } C = \hat{\sigma}_1 / \hat{\sigma}_0$$

where $(\hat{\mu}_0, \hat{\sigma}_0)$ and $(\hat{\mu}_1, \hat{\sigma}_1)$ are some estimates of (μ_0, σ_0) and (μ_1, σ_1) , respectively. Thus, the sensitivity index can be estimated by

$$\hat{\Delta} = \frac{1 + \hat{\varepsilon} / \hat{\mu}_0}{\hat{C}} \quad (6)$$

Note that in practice, the shift in location parameter (i.e., ε) and/or the change in scale parameter (i.e., C) could be random. If both ε and C are fixed, the sensitivity can be assessed based on the sample means and sample variances obtained from the two populations. In real-world problems, however, ε and C could be either a fixed or a random variable. In other words, there are three possible scenarios: (i) the case where ε is random and C is fixed, (ii) the case where ε is fixed and C is random, and (iii) the case where both ε and C are random. These possible scenarios have been studied by Luet al.^[14]

CONCLUDING REMARKS

Logistic regression analysis is useful in determining/identifying the risk factors (predictors) for building the medical predictive model in case-control studies. Goodness of fit and internal validation are necessarily performed to ensure the predictability of the established medical predictive model. In addition, the assessment of generalizability of the established medical predictive model from one patient population to another and/or from one medical center to another is encouraged. In case-control studies, however, it is always a major challenge to the principal investigator when there are significant differences (or imbalances) in demographics and/or patient characteristics, which may cause confounding effects. Consequently, the results may be biased and hence misleading. In addition, the imbalances could have a negative impact on the generalizability of the clinical results. To avoid possible confounding effects due to imbalances between the case and control groups, the use of propensity score is recommended.

REFERENCES

1. Doll, R., Hill, A.B. Smoking and carcinoma of the lung; preliminary report. *Br. Med. J.* **1950**, 2 (4682), 739–748.
2. Hosmer, D.W.; Lemeshow, S.L. *Applied Logistic Regression*, 2nd Ed.; Wiley-Interscience: New York, 2000.
3. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, 70 (1), 41–55.
4. Rosenbaum, P.R.; Rubin, D.B. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* **1984**, 79 (387), 516–524.
5. Austin, P.C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multiv. Behav. Res.* **2011**, 46 (3), 399–424.
6. Fukuoka, M.; Yano, S.; Giaccone, G.; Tamura, T.; Nakagawa, K.; Douillard, J.Y.; Nishiaki, Y.; Vansteenkiste, J.; Kudoh, S.; Rischin, D.; Eek, R.; Horai, T.; Noda, K.; Takata, I.; Smit, E.; Averbuch, S.; Macleod, A.; Feyereislova, A.; Dong, R.P.; Baselga, J. Multi-institutional randomized phase II trial of gefitinib for previously treated

- patients with advanced non-small-cell lung cancer. *J. Clin. Oncol.* **2003**, *21* (12), 2237–2246.
7. Paez, J.G.; Janne, P.A.; Lee, J.C.; Tracy, S.; Greulich, H.; Gabriel, S.; Herman, P.; Kaye, F.J.; Lindeman, N.; Boggon, T.J.; Naoki, K.; Sasaki, H.; Fujii, Y.; Eck, M.J.; Sellers, W.R.; Johnson, B.E.; Meyerson, M. EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy. *Science* **2004**, *304* (5676), 1497–1500.
 8. Hosmer, D.W.; Lemeshow, S.L.; Sturdivant, R.X. *Applied Logistic Regression*, 3rd Ed.; Wiley-Interscience: New York, 2013.
 9. Belsley, D.A.; Kuh, E.; Welsch, R.E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; John Wiley & Sons: New York, 1980.
 10. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*; Springer: New York, 2013.
 11. Chow, S.C.; Shao, J.; Hu, O.Y.P. Assessing sensitivity and similarity in bridging studies. *J. Biopharm. Stat.* **2002**, *12* (3), 385–400.
 12. Chow, S.C.; Chang, M. *Adaptive Design Methods in Clinical Trials*; Taylor & Francis: New York, 2006.
 13. Chow, S.C.; Chang, M.; Pong, A. Statistical consideration of adaptive methods in clinical development. *J. Biopharm. Stat.* **2005**, *15* (4), 575–591.
 14. Lu, Y.; Kong, Y.Y.; Chow, S.C. Analysis of sensitivity index for assessing generalizability in clinical research. *J. Biostat.* **2017**, *2* (1), 009.

Logistic Regression: Three-Point Designs

Mårten Vågerö

AstraZeneca R&D Södertälje, Södertälje, Sweden

Rolf Sundberg

Stockholm University, Stockholm, Sweden

Abstract

This entry discusses the problem and practical implications of infinite parameter estimates in a three-point univariate logistic regression.

INTRODUCTION

Three-point designs are frequently used in phase II clinical studies to establish a dose–response relationship before proceeding to larger confirmatory studies. In cases with a binary outcome, a logistic regression model is often adopted. As these studies are most often of small or moderate size with little prior knowledge about the effective dose interval, there is a substantial risk for an outcome with nonunique or infinite parameter estimates.

Morgan's book^[1] discusses both sequential and fixed-point designs for estimation of a particular characteristic or the entire distribution of a dose–tolerance threshold model for quantal response data. In a threshold model, each individual in the target population has a dose tolerance (threshold) to the exposure of a drug. If a subject (patient, healthy volunteer, or animal) is administered a dose above its threshold, it will respond. Otherwise, it will be a nonresponder. If subjects are assumed to be randomly sampled from a logistic tolerance threshold distribution, the probability of responding when administered a dose x can be written as

$$P(x; \mu, \beta) = \frac{e^{\beta(x-\mu)}}{1 + e^{\beta(x-\mu)}} \quad (1)$$

where μ is the median (ED_{50}) and β represents the steepness of the dose–response curve. Parameter estimates that are asymptotically unbiased, and asymptotic standard errors for these estimates, can be determined by the method of maximum likelihood through many available statistical computer packages.

However, in small sample studies, problems with infinite parameter estimates as a result of a complete or quasi-complete separation of the data may occur (see below).^[2] These situations also cause problems when working with some statistical computer packages. Certainly, there are textbooks discussing this problem,^[3] but many modern textbooks on logistic regression do not mention it at all.^[4]

A published example where this was not properly addressed is discussed in Ref. [5]. In simple studies, it is easily seen if there is a complete or quasi-complete separation of the data leading to an infinite parameter estimate. However, in more complex models with several explanatory variables, it need not be obvious from inspection of the outcome whether some parameter estimate is infinite or not. This paper will discuss the problem and practical implications of infinite parameter estimates in a three-point univariate logistic regression.

MAXIMUM LIKELIHOOD ANALYSIS

With the notation

n_i = the number of subjects who are administered the dose x_i , and
 y_i = the number of subjects responding to dose x_i , the log-likelihood derived from Eq. 1 is

$$L(\mu, \beta) = \sum_i y_i \beta (x_i - \mu) + \sum_i n_i \log \{1 - P(x_i; \mu, \beta)\} + \text{constant} \quad (2)$$

The maximum likelihood estimator (MLE) $(\hat{\mu}, \hat{\beta})$ is determined by maximizing Eq. 2. The second derivatives form the observed information matrix, here denoted $J(\mu, \beta)$. We have

$$J(\hat{\mu}, \hat{\beta}) = \begin{bmatrix} \hat{\beta}^2 \sum_i n_i \hat{P}_i (1 - \hat{P}_i) & -\hat{\beta} \sum_i (x_i - \hat{\mu}) n_i \hat{P}_i (1 - \hat{P}_i) \\ -\hat{\beta} \sum_i (x_i - \hat{\mu}) n_i \hat{P}_i (1 - \hat{P}_i) & \sum_i (x_i - \hat{\mu})^2 n_i \hat{P}_i (1 - \hat{P}_i) \end{bmatrix} \quad (3)$$

where $\hat{P}_i = P(x_i; \hat{\mu}, \hat{\beta})$.

The variance–covariance matrix of the maximum likelihood estimator (MLE) is estimated by the inverse $J^{-1}(\hat{\mu}, \hat{\beta})$ of the observed information matrix in the maximum likelihood point. We will use superscripts to denote the elements of J^{-1} , e.g., j^{11} for its upper left corner. The expected Fisher information, i.e., the expected value of J over the possible outcomes of data, will be denoted as $I(\mu, \beta)$.

CHOICE OF DESIGN POINTS

Fixed-point designs usually divide the total number of subjects equally between all dose levels. The dose levels are ideally chosen symmetrically around the center μ of the tolerance distribution and with an equal distance between the dose levels. See, for example, Ref. [6] or [1] for more details on design considerations for quantal response data. Consider now a three-point design, with $3n$ subjects equally divided between the three doses, $x_1 < x_2 < x_3$. The middle dose, x_2 , is assumed located in the center μ of the tolerance threshold distribution, and the other two doses, x_1 and x_3 , symmetrically around x_2 , say at $x_2 \pm \delta$. This makes $P_2 = 1/2$ and $P_3 = 1 - P_1$, and the expected Fisher information, $I(\mu, \beta)$, is diagonal; thus $\hat{\mu}$ and $\hat{\beta}$ are approximately uncorrelated.

$$I(\mu, \beta) = n \begin{bmatrix} \beta^2(2P_1(1-P_1)+1/4) & 0 \\ 0 & 2P_1(1-P_1)\delta^2 \end{bmatrix} \quad (4)$$

Many different optimality criteria are found in the literature. Most criteria are based on the information matrix and aim at minimizing elements or functions of the elements in the inverse information. To achieve a μ -optimal design, I^{11} should be minimized, or the equivalent $P_1(1-P_1)$ maximized; thus the design points should be chosen close to μ . An optimal design for estimating β should instead minimize I^{22} ; that is, maximize $P_1(1-P_1)\delta^2$. Using this optimality criteria, the spacing δ should be chosen so that $\delta = 2.4\beta$ or equivalent so that $P_1 = 1 - P_3 = 0.083$. A widely used optimality criterion is D-optimality, where the determinant of I is maximized.^[7] This is a design with a good balance between a low risk of a separation in data and providing an informative dose–response relation. Using the D-optimality criteria, the design points are chosen so that $P_1 = 1 - P_3 = 0.136$ or equivalently so that $\delta = 1.85\beta$.

DATA CONFIGURATIONS

In a binary dose–response model (with a single dose or a dose vector), some degenerate types of data configurations yield infinite and/or nonunique parameter estimates. Essentially, following Albert and Anderson,^[2] we characterize the sample outcomes as follows:

1. Only one response type represented in data.
2. A complete separation of responders from nonresponders.
3. A quasi-complete separation of responders from nonresponders.
4. An overlap of the two response types.

These authors^[2] discuss the problem of existence, finiteness, and uniqueness of maximum likelihood estimates in the case of several explanatory variables. A further discussion of how to interpret and handle infinite parameter estimates in a general logistic regression model is found in Ref. [8]. Next we describe what Albert and Anderson's results imply in our case with only one explanatory variable; that is, the dose.

Only One Response Type Represented in Data

The first situation is when the outcome consists of only responders or only nonresponders. The only information provided by data in this trivial case is that the design points were chosen outside the effective dose range for the specific drug under investigation. The experiment must be supplemented with more design points in the direction of the effective dose range.

A Complete Separation of Responders from Nonresponders

A complete separation means that there is a specific (but nonunique) dose level so that all responders are found at higher doses than the nonresponders (or vice versa).

Thus

$$\min_i(x_i|y_i > 0) > \max_i(x_i|y_i < n_i)$$

The probability is particularly high for obtaining such data when the dose–response curve steeply rises from ≈ 0 to ≈ 1 somewhere between $\max(x_i|y_i < n_i)$ and $\min(x_i|y_i > 0)$. It is easily verified that the maximum likelihood estimates are degenerated, being

$$\hat{\beta} = +\infty$$

and

$$\max_i(x_i|y_i < n_i) < \hat{\mu} < \min_i(x_i|y_i > 0)$$

A Quasi-Complete Separation of Responders from Nonresponders

A quasi-complete separation of responders from nonresponders means that there is a unique dose level x^* in the design such that all responders occur at $x_i \geq x^*$, and all nonresponders occur at dose levels $x_i \leq x^*$ (or vice versa), with both responders and nonresponders at x^* .

In this situation, the highest probability for the observed data is obtained if the dose–response curve steeply rises at the particular dose x^* . Hence the MLE in this case is

$$(\hat{\beta}, \hat{\mu}) = (+\infty, x^*)$$

An Overlap of the Two Response Types

There is an overlap of the two response types if the smallest dose level with responders is smaller than the largest dose level with nonresponders.

Under the parameterization with μ and β , a degenerate MLE will occur in the extreme case when data suggest that there is no dose–response relation; that is, $\hat{\beta} = 0$. In a three-point design, this means that the proportion of responders is the same for dose x_1 as for dose x_3 . In this case, $\hat{\beta} = 0$ and $\hat{\mu}$ is nonunique or infinite.

Excluding this rare outcome, an overlap implies that the MLE $(\hat{\mu}, \hat{\beta})$ is unique and finite and can be determined from the likelihood equations.

Properties of the Maximum Likelihood Estimator Illustrated

From a practical point of view, after having collected data from all patients in the experiment, we would like to base our inference on standard methods no matter what the underlying risk was for a separation of responders from nonresponders.

If data suggest that there is a separation of responders from nonresponders, there is not much information to gain from the experiment so you would need to supplement your experiment in some way. However, if all parameter estimates exist and are finite, you would like to determine the parameter estimates and make conclusions about the dose–response relation without taking into account the underlying risk of a separation of responders from nonresponders.

Here we demonstrate that the asymptotic inference is valid as soon as $\hat{\beta}$ is finite. We illustrate this in a three-point design with a somewhat larger spacing than in a D-optimal design, design points at quantiles 10%, 50%, and 90%; that is, we chose x_i such that

$$\begin{aligned} P_1 &= 1 - P_3 = 0.10 \\ P_2 &= 1/2 \end{aligned}$$

This naturally represents an idealized situation, because in reality we do not know the position of the center nor the steepness. Placing the middle dose away from the center would increase the risk of a complete or quasi-complete separation, further stressing the importance of a good choice of design points.

The exact distribution for the MLE $(\hat{\mu}, \hat{\beta})$ together with its observed information matrix $J(\hat{\mu}, \hat{\beta})$ was found by

going through the $(n + 1)^3$ possible distinct outcomes in a three-point design for a binary response variable with n observations per dose level.

RESULTS

In a three-point design, $\hat{\beta} = +\infty$ if there are no responders at the lowest dose and no nonresponders at the highest dose. The probability for this event is

$$P(\hat{\beta} = +\infty) \approx (1 - P_1)^n P_3^n \quad (5)$$

The corresponding probability for $\hat{\beta} = -\infty$ is negligible.

In the present investigation, this formula yields the probabilities 0.122, 0.015, and 0.00003 for $n = 10$, $n = 20$, and $n = 50$ observations per dose group, respectively.

In Fig. 1, the exact distribution of $\hat{\beta} - \beta$ standardized by j^{22} is presented for $n = 10$ and $n = 20$.

The j^{22} -standardized distribution curve of $\hat{\beta} - \beta$ follow a standard normal distribution quite well, indicating that the asymptotic inference about β is adequate, as soon as $\hat{\beta}$ is finite. In particular, the median of $\hat{\beta}$ is close to the true β value. Hence if centrality is measured by the median, there is no substantial systematic error in the MLE. Furthermore, as seen in Table 1, the coverage probabilities for one-sided confidence intervals $(-\infty, \hat{\beta} + z_\alpha \sqrt{j^{22}})$ based on a normal approximation are close to the nominal confidence levels for $n = 10$, $n = 20$, and $n = 50$ even when $P(\hat{\beta} = +\infty)$ is substantial. Similar results were found for a D-optimal design and for designs deviating in location and spacing from the D-optimal, except that the probability for a degenerated MLE considerably differed. However, it should be noted that the distribution of $(\hat{\beta} - \beta)/j^{22}$ has no or little probability mass in some intervals but takes large jumps in other points/intervals (Fig. 1). As the location of these intervals is different for different n values, this causes an irregular pattern in the coverage probabilities seen in Table 2.

Fig. 2 shows the $\sqrt{j^{11}}$ -standardized error in $\hat{\mu}$ for $n = 10$, given that $\hat{\beta}$ is finite. The figure illustrates that the $\sqrt{j^{11}}$ -standardized error in $\hat{\mu}$ follows a standard normal distribution fairly well even with only $n = 10$ observations per design point. For the case where $n = 20$ observations per design point, the distribution of the error is even closer to the standard normal distribution. This case has been omitted in Fig. 2 to increase readability. When $\hat{\beta} = +\infty$, use of the observed information from Eq. 3 would indicate a confidence interval for μ of length zero. However, in this case, the likelihood is nonregular, having its maximum on the boundary of the parameter space, so the observed information in Eq. 3 is not justified and should not be used. The results from the exact calculations presented in Table 2 demonstrate that the conditional coverage probabilities are fairly close to the nominal level. It must be stressed that these coverage probabilities are conditional on a finite $\hat{\beta}$, indicating that as soon as we have an

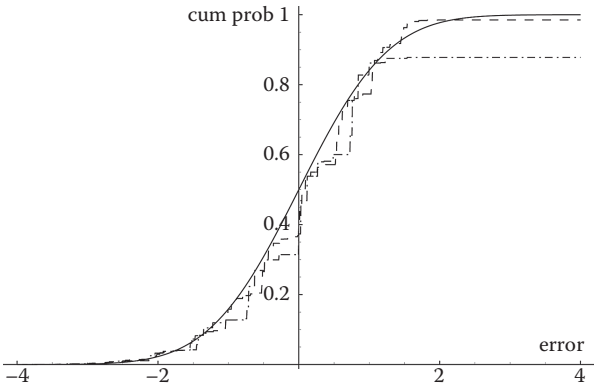


Fig. 1 Empirical distribution of the $\sqrt{j^{22}}$ -standardized error in $\hat{\beta}$ with $n = 10$ (dot-dashed line) and $n = 20$ (short broken) observations per dose level. A $N(0,1)$ is included for reference (solid line)

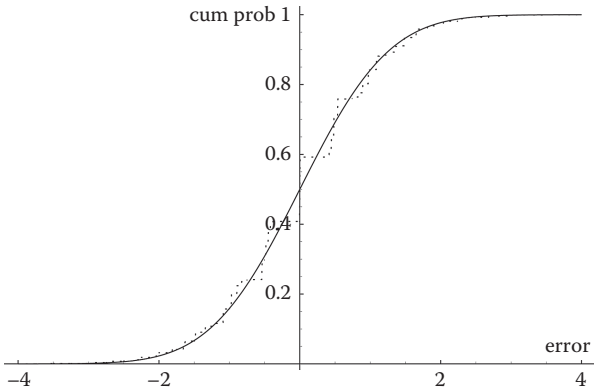


Fig. 2 Empirical distribution of the $\sqrt{j^{11}}$ -standardized error in $\hat{\mu}$ for the outcomes where $\beta < +\infty$ with $n = 10$ (dot-dashed line). A $N(0,1)$ is included for reference (solid line)

Table 1 Coverage Probabilities for $-\infty, \hat{\beta} + z_{\alpha}\sqrt{j^{22}}$ Intervals

n	97.5% CI	95% CI	90% CI
10	0.9674	0.9592	0.8953
20	0.9663	0.9596	0.9061
50	0.9698	0.9530	0.8953

overlap of the two response types, asymptotic normality of $\hat{\mu}$ seems to be appropriate. The same results have been qualitatively established not only for a D-optimal design, but also for designs deviating in location and spacing from the D-optimal; for example, a design in Ref. [5] with an extremely large spacing yielding a finite $\hat{\beta}$ with probability of only 0.3. A partial explanation of this remarkable fact is that $\hat{\mu}$ and $\hat{\beta}$ are approximately uncorrelated [note that the expected Fisher information in this case is diagonal (Eq. 4)]

DISCUSSION

The overall conclusion of this work is that three-point designs do not produce systematic errors in the distribution of the maximum likelihood estimator of the steepness parameter, β , but if the design points are too widely spaced, the probability of an infinite $\hat{\beta}$ is unreasonably high. If one succeeds in choosing the spacing according to a D-optimal design, $\hat{\beta}$ seems to be an effective estimator with a reasonably low probability of being infinite, even with only 10 subjects per dose level. However, from a practical point of view, it is difficult to determine the doses D-optimally as it must be based on prior beliefs about the dose-response relationship.

Prior beliefs are required also for a Bayesian analysis of data. Using a data augmentation prior (DAP) (see Ch. 13 in Ref. [4]) would be equivalent to adding some extra observations to the experiment. This would reduce/eliminate

Table 2 Coverage Probabilities for $\hat{\mu} \pm z_{\alpha/2}\sqrt{j^{11}}$ Intervals, Conditional on $\beta < \infty$

n	95% CI	90% CI	80% CI
10	0.9364	0.8782	0.7847
20	0.9474	0.8921	0.7920
50	0.9496	0.8982	0.7972

the risk for infinite or nonunique parameter estimates, but it might then hide the lack of information in the data themselves.

An important practical conclusion of the examples investigated is that once we have an outcome yielding a finite $\hat{\beta}$, asymptotic normality-based confidence intervals are justified for inference about μ given the finite $\hat{\beta}$ and with some extra care also for inference about β .

Therefore the main problem is outcomes with infinite $\hat{\beta}$. Most commonly used statistical computer packages provide warnings that we might have the case of a complete or quasi-complete separation, but some of them nevertheless return finite estimates for all parameters in these situations. It is also important when teaching logistic regression to users of these statistical tools, to discuss the interpretation of a complete or quasi-complete separation of the data.

In practice, what should be done with an experiment that gives an infinite $\hat{\beta}$? First of all, its data indicate that the steepness is greater than expected in advance. The next step would either be to stop the trial here, or more likely, to continue the experiment by including more subjects and more design points in the trial. Different sequential types of procedures have been suggested.^[1]

CONCLUSION

It is concluded that the main problem in three-point designs with a binary outcome is the risk of observing an infinite $\hat{\beta}$.

We need to take this risk into account when designing the experiment. A D-optimal design provides an informative dose–response relation and still has a low risk of observing an infinite $\hat{\beta}$. However, irrespectively of the design, once we have an outcome where all parameter estimates exist and are finite, asymptotic inference is remarkably adequate even in small-sample studies.

REFERENCES

1. Morgan, B.J.T. *Analysis of Quantal Response Data*; Chapman & Hall: London, 1992.
2. Albert, A.; Anderson, J.A. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **1984**, *71*, 1–10.
3. Hosmer, D.; Lemeshow, S. *Applied Logistic Regression*; Wiley: New York, 1989.
4. Christensen, R. *Log-Linear Models and Logistic Regression*, 2nd Ed.; Springer-Verlag: New York, 1997.
5. Vågerö, M. A remark on small-sample properties of logistic regression in three-point designs. *J. Biopharm. Stat.* **2000**, *10*, 573–587.
6. Finney, D.J. *Statistical Methods in Biological Assay*, 3rd Ed.; Griffin: London, 1978.
7. Kalish, L.A. Efficient design for estimation of median lethal dose and quantal dose–response curves. *Biometrics* **1990**, *46*, 737–748.
8. Kolassa, J.E. Infinite parameter estimates in logistic regression, with application to approximate conditional inference. *Scand. J. Stat.* **1997**, *24*, 520–530.

McNemar's Test

Robert G. Lehr

Bayer Healthcare Pharmaceuticals, Inc., Montville, New Jersey, U.S.A.

Abstract

McNemar's test is a statistical procedure used to compare two proportions which are dependent or correlated. This entry provides details of the test calculations, variations of the test, sample size calculation methods, discussion of the test characteristics, related confidence intervals, alternative tests, and extensions of the test.

Keywords: Correlated proportions; Discordant pairs; Matched pairs; McNemar's statistic; phi-Correlation; Symmetry.

INTRODUCTION

McNemar's test is a statistical procedure used to compare two proportions which are dependent or correlated. Percentages or proportions of events resulting from 2 observations made on the same or matched experimental units under 2 different conditions may be tested for equality using this procedure. The test may have a one-tailed or two-tailed alternate hypothesis. Details of the test first appeared in a 1947 *Psychometrika* article written by Quinn McNemar (1900–1986), a Stanford University Professor of Psychology and Statistics who made many contributions to psychometric theory. This discourse provides details of the test calculations, variations of the test, sample size calculation methods, discussion of the test characteristics, related confidence intervals, alternative tests, and extensions of the test.

CALCULATION OF THE TEST STATISTIC

The null hypothesis for McNemar's test is the equality of π_1 and π_2 , the two proportions of interest. It may be stated in several equivalent ways, including: $H_0: \pi_1 = \pi_2$, $H_0: \pi_1 - \pi_2 = 0$, and $H_0: OR = 1$, where $OR = (\pi_1/(1 - \pi_1))/(\pi_2/(1 - \pi_2))$. The corresponding alternate hypotheses for a one-tailed test may be $H_0: \pi_1 - \pi_2 > 0$ and $H_0: OR > 1$. A two-tailed alternate hypothesis would be stated using the non-directional “ $\neq$.”

Table 1 below uses a pretreatment-posttreatment scenario to illustrate summarized data for comparing proportions resulting from assessments made on the same or matched experimental units.

In Table 1, $n = a + b + c + d$, b = the number of subjects classified a success post but not pre, c = the number of subjects classified a success pre but not post,

$n_1 = a + b$ = the number of observed pretreatment successes, $n_s = a + c$ = the number of observed posttreatment successes, $p_1 = n_1/n = (a+b)/n$ = the observed posttreatment proportion of success, and $p_2 = n_2/n = (a + c)/n$ = the observed pretreatment proportion of success. The frequencies b and c are usually referred to as mismatch-cell or discordant-pair frequencies, as the pair of observations represents different responses.

A statistic which may be used to quantify the treatment effect and estimate the population difference is $(p_1 - p_2) = (b - c)/n$, the difference in proportions of success which is an estimate of $\pi_1 - \pi_2$. When p_1 and p_2 are likely to be near 0 or 1 and their difference is likely to be small, the odds ratio, $(p_1/(1 - p_1))/(p_2/(1 - p_2)) = ad/bc$, may be preferred to compare the two.

The test statistic most commonly used for McNemar's test is

$$T_u = \frac{(b - c)^2}{(b + c)} \quad (1)$$

using the notation of Table 1. It is approximately distributed as chi-square with one degree of freedom.

The test statistic in Eq. 1 can be derived in several ways, two of which were provided in the original entry.^[1] As $(p_1 - p_2) = (b - c)/n$ is not dependent on a or d , the numbers of concordant pairs, a statistic may be formed using only the expected cell frequencies for the discordant cells under the null hypothesis. Substituting these two expected values, $(b + c)/2$, into the usual chi-square (observed – expected)²/(expected) expression yields Eq. 1 directly with some algebraic simplification. Alternately, creating a Wald statistic by forming a ratio of the difference of the two correlated proportions divided by the standard deviation of the difference results in an equivalent test using a statistic which is approximately distributed as Normal (0,1).

Table 1 Response Frequencies

		Pretreatment		
		Success	Failure	
Posttreatment	Success	<i>a</i>	<i>b</i>	<i>n</i> ₁
	Failure	<i>c</i>	<i>d</i>	
		<i>n</i> ₂		<i>n</i>

The continuity-corrected McNemar chi-square test uses the statistic

$$T_c = \begin{cases} (b - c - 1)^2 / (b + c) & \text{if } b > c, \\ (b - c + 1)^2 / (b + c) & \text{if } b < c, \\ \text{and } 0, & \text{if } b = c \end{cases} \quad (2)$$

This given variation^[2] is analogous to the Yates correction for comparing independent proportions.

An exact version (T_c) of McNemar's test, described by Mosteller,^[3] may be performed as follows: If $c > b$, the p -value = the probability of getting (c or more) or (b or less) successes in a binomial experiment of $b + c$ trials each with probability of success $p = 1/2$, that is $B(\geq c, b + c; 1/2) + B(\leq b, b + c; 1/2)$, where

$$B(\leq x, n; p) = \sum_x^n \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

Although McNemar only gives the first of these three versions of his test in the seminal 1947 paper, he mentions the latter two in his textbook *Psychological Statistics*.^[4] It is likely that several statisticians independently derived comparable test statistics at approximately the same time, but McNemar's version was apparently the first to reach print. Siegel's work may be the first textbook^[5] in which the method was called "McNemar's test."

An interesting relationship between McNemar's test and the Cochran Mantel-Haenszel (CMH) test is described by Agresti^[6] in his *Categorical Data Analysis* book. Each pair of binary responses may be represented as a 2×2 table, and then the CMH test for conditional independence for the set of 2×2 tables is a test for marginal homogeneity in the collapsed table.

AN EXAMPLE

A researcher instructs 50 patients with a mild rash on both arms to apply a steroid cream to one arm for 14 days, and a pink drying lotion to the other arm. The condition of interest is complete healing of all lesions. Table 2 gives the results.

The steroid cream had a better healing rate with 66% of the arms healed compared to 46% for the pink drying lotion. Using Eq. 1, the McNemar test statistic is

Table 2 Fourteen-day Complete Healing Rates

		Pink drying lotion	
		Healed	Not healed
Steroid cream	Healed	8	25
	Not healed	15	2

$X^2 = (25 - 15)^2 / (25 + 15) = 2.5$. The corresponding two-tailed probability from a chi-square distribution with one degree of freedom is $p = 0.113$. Using the 0.05 significance level, the data do not provide sufficient evidence to conclude that the two treatments differ in healing effect.

APPLICATIONS OF McNEMAR'S TEST

As is evidenced by the earliest references, McNemar's test has seen extensive use in the behavioral sciences. It is the simplest test to use for comparing binary data for short- or long-term studies involving matched subjects. Changes for identical twins in a given characteristic, increase or decrease in a given score for the same test subject, or success or failure on a given test for a matched pair of subjects are all reasonable applications of this test.

Any study using a design that involves two observations on a binary variable on the same subject may be analyzed with McNemar's test. Preintervention and postintervention results, two different interventions on different locations, or two observations taken at successive time points are all examples of where the test may be used. Diagnostic imaging studies often are designed to obtain two scans on the same area of the same subject and compare the diagnostic accuracies of readers using the scans to determine presence or absence of a given condition.

Studies in which a binary or dichotomous variable is evaluated on matched pairs of patients may also be analyzed using McNemar's test. Long-term epidemiology studies on twins, other siblings or family relations, or subjects matched on a set of demographic and background factors are among the most popular designs.

ALTERNATIVE TESTS

Several tests have been proposed as alternatives to McNemar's test for the equality of two correlated proportions. None exhibit the simplicity or frequency of use.

Bennett and Underwood^[7] discuss a statistic formed by adding 1/2 to each cell frequency, which leads to the statistic $T_{u*} = (b - c)^2 / (b + c + 1)$. They observe that this statistic seems to perform better than the previously mentioned corrected McNemar statistic based on empirical investigations.

A Wald statistic that results from the application of a general method proposed by Grizzle, Starmer, and Koch^[8]

uses a variance estimator not restricted by the null hypothesis. The statistic was put forth as an alternative to the more general Q statistic of Cochran,^[9] which is in a sense a more general version of McNemar's test. No claims are made regarding the performance of the statistic relative to that of the prior tests.

Suissa and Shuster^[10] proposed a method of calculating the exact significance level for the statistic formed by the square root of the McNemar statistic. The procedure uses the complete domain of the nuisance parameter representing the probability of a discordant pair under the null hypothesis.

May and Johnson^[11] compare the size and power of the uncorrected, corrected, and exact versions of McNemar's test already mentioned and several other tests. They found that the Wald statistic tends to be anticonservative while the continuity-corrected version of McNemar's test tends to be overly conservative.

Irony, Pereira, and Tiwari^[12] suggest a Bayesian hypothesis test for the comparison of two correlated proportions. The test consists of evaluating the posterior odds ratio, the ratio of the posterior probabilities, that is, the population proportion of one of the discordant-pair cells to the total of the two discordant-pair cells.

Lloyd^[13] has proposed an exact unconditional test of unequal effects from binary matched pairs. Starting with an estimate of the nuisance parameter from an inexact test, he eliminates the residual dependence by maximization. He calls the resulting probabilities E + M *p*-values, presumably for "estimate and maximize."

SAMPLE SIZE DETERMINATION

Different approaches to calculating sample size requirements for the McNemar test have been given by several authors. The methods vary with respect to approach (conditional or unconditional), parameters that need to be specified, and whether they are asymptotic or exact.

One of the first methods proposed for estimating sample size for McNemar's test was proposed by Miettinen.^[14] Miettinen's equation for a one-sided test states that the required sample size is

$$n = \left[z_{\alpha} \Psi + z_{\beta} (\Psi^2 - 1/4 \delta^2) + \Psi^{1/2} \right]^2 / \Psi \delta^2 \quad (3)$$

where δ is the detectable difference, Ψ is the proportion of mismatches, and z_{α} and z_{β} are the usual normal values.

Schlesselman^[15] gives a two-step method requiring specification of the estimated proportion of discordant pairs and odds ratio to be detected. The required number of experimental units resulting in discordant pairs is then determined, leading then to an estimate of the total units required.

Connor^[16] used an unconditional multinomial-based approach and derived an equation similar to that of

Miettinen's. He suggested the use of an independence assumption if no information regarding discordant-pair frequency was available.

Dupont^[17] developed a method of estimating the sample size for McNemar's test which required the specification of the probability of exposure for control patients, an odds ratio to be detected, and a correlation coefficient for exposure in matched pairs of case-control studies.

Connett, Smith, and McHugh^[18] derive a formula for McNemar's tests with a completely unconditional approach. They state that the derivation is analogous to that used by Fleiss^[19] for the unmatched design. They provide a proof that their unconditional method will always produce a result equal to or greater than Schlesselman's conditional formula.

Using Connett's equation, Machin et al.^[20] have constructed tables of sample sizes for McNemar's test. Various values for the alternate hypothesis odds ratio and the proportion of discordant pairs cover 17 pages.

Parker and Bregman^[21] developed a sample size formula that allows adjustment for varying probability of discordant pairs among matched pairs. They demonstrate that allowing for this heterogeneity increases the calculated sample size requirements.

Lachenbruch^[22] proposed a method that provided a range of possible sample sizes based on the smallest and largest possible numbers of discordant pairs. The method is proposed for situations where estimates of discordant-pair frequencies are not readily available.

Several authors have evaluated and/or compared one or more sample size method for McNemar's test. Duffy^[23] compared Miettinen's asymptotic formula against exact powers, and concluded the asymptotic formula was generally adequate. Lachin^[24] thoroughly analyzed the methods previously mentioned among others and discussed which tend to be conservative and anticonservative. He demonstrated that the differences among the methods can be traced to the variance estimates under the alternate hypotheses. Royston^[25] compared several conditional and unconditional approaches and suggested a strategy for sample size choice based on a binomial unconditional estimate.

A quick formula that provides an easy method for obtaining rough sample size estimates for McNemar's test requires the specification of a pair of proportions and estimate of the correlation between them.^[26] For a two-tailed McNemar test with alpha level 0.05, the following relation specifies a sample size, which provides an approximate power of 0.80 to detect a difference d between proportions p_1 and p_2

$$n = \frac{16pq(1-r)}{d^2} \quad (4)$$

where n is the size of the sample, $p = (p_1 + p_2)/2$, $q = 1 - p$, $d = |p_2 - p_1|$, and r is the phi correlation coefficient. The possible values for r are dependent on p_1 and p_2 , and do not in general span the entire range of -1 to 1 .

Other authors have developed sample size estimation methods for extensions of McNemar's test. Woolson, Bean, and Rojas^[27] proposed a method for Cochran's test. Feuer and Kessler^[28] derive a test statistic and sample size method for testing whether the marginal changes in two independently sampled tables are equal. Lu and Bean^[29] discuss the problem of estimating sample size for one-sided equivalence based on McNemar's test.

DISCUSSION OF THE CHARACTERISTICS OF McNemar's TEST

Some users are uncomfortable with McNemar's test because the numbers the statistic used do not reflect the entire sample size; so at first glance the test does not seem to be a comparison of the two paired proportions.^[30] These features are, however, just a consequence of the pairing. Although only the frequencies or proportions from the mismatch cells are used in the calculation, all cases contribute to the inference about whether the marginal proportions differ.^[6]

The relationship involving the paired values can be expressed in several ways, two of which have been mentioned previously. The quantities Ψ and r affect power, required sample size, and operating characteristics of McNemar's test. The symbol Ψ , used here to represent the proportion of discordant pairs, unfortunately, is also used to denote odds ratios. The correlation coefficient r is the familiar Pearson r calculated on zeros and ones, and has a simpler form expressed in terms of the four cell frequencies from the 2×2 table, for which case it is more commonly denoted ϕ (phi). In terms of the quantities in Table 1, this simpler form is

$$r = \phi = \frac{(ad - bc)}{\sqrt{(a+B)(c+d)(a+c)(b+d)}} \quad (5)$$

The two quantities r and Ψ , which both characterize the degree of relationship between the paired variables, are related. For a fixed pair of proportions, the correlation r and the proportion of discordant pairs Ψ are functions of one another, i.e., r , generally, decreases as Ψ increases. The relationship between Ψ and r for a given p_1 and p_2 can be specified by^[26]

$$\Psi = p_1 + p_2 - 2p_1p_2 - 2r \times \sqrt{p_1(1-p_1)p_2(1-p_2)} \quad (6)$$

The range of possible values which Ψ may take is from $|p_2 - p_1|$ to $p_1 + p_2$ or $2 - (p_1 + p_2)$, whichever is smaller. A practical lower bound of $p_1(1 - p_2) + p_2(1 - p_1)$, the proportion of discordant cells which would result from independence, is also useful for some situations. As previously stated, r is a function of Ψ , p_1 , and p_2 , and may thus be obtained using

$$r = \frac{(p_1 + p_2 - \Psi - 2p_1p_2)}{2\sqrt{p_1(1-p_1)p_2(1-p_2)}} \quad (7)$$

a direct inversion of Eq. 6. Substituting the value of Ψ corresponding to independence will produce an r value of 0. Substituting the minimum Ψ of $p_2 - p_1$ into Eq. 7 produces

$$r = \sqrt{\frac{p_1(1-p_2)}{p_2(1-p_1)}} \quad (8)$$

The right-hand side of Eq. 8 is the square root of the odds ratio where $p_1 < p_2$. This gives an easy way of remembering the maximum possible r given proportions p_1 and p_2 . Although r may be less than 0, for most practical situations values ranging from 0 to the square root of the odds ratio may be assumed.^[26]

If only a difference δ is specified without p_1 and p_2 , the value of r has a maximum of

$$r_{\max} = \frac{(1 - \delta)}{(1 + \delta)} \quad (9)$$

which is reached when $(p_1 + p_2)/2 = 0.5$, i.e., $p_1 + p_2 = 1$.

RELATED CONFIDENCE INTERVALS

Asymptotic confidence intervals (CIs) for the difference of correlated proportions ($\pi_1 - \pi_2$) are referenced in some statistics textbooks discussing categorical data analysis. Using the notation of Table 1 and the ensuing discussion, a formula for an asymptotic CI given by Fleiss^[19] is $((p_2 - p_1) + (1.96\text{se}_{p_2-p_1}) + 1/n)$, where

$$\text{se}_{p_2-p_1} = \sqrt{(a+d)(b+c) + 4bc/n} \sqrt{n}$$

Breslow and Day^[31] suggested a CI that can be obtained by generating an exact CI for the proportion $b/(b+c)$ from Table 1 and then translating these limits back into odds ratios.

Armitage and Berry^[32] suggested a CI which can be obtained in a similar way to that used by Breslow and Day for the differences $(p_2 - p_1)$ which would result given the b and c cells. This interval, in terms of the original cell frequencies, is

$$\left(\frac{b+c}{n} \right) \left(\frac{2b}{b+(c+1)F_{[2(c+1), 2b]}} - 1 \right), \dots, \left(\frac{b+c}{n} \right) \left(\frac{(2b+2)F_{[2(b+2), 2c]}}{c+(b+1)F_{[2(b+2), 2c]}} - 1 \right)$$

Exhaustive empirical investigations suggest that this interval is perfectly consistent with the exact version of the McNemar test in the sense that the test rejects H_0 if and

only if the CI excludes 0.^[33] Lloyd^[34] points out a coverage problem with these two intervals under certain conditions, and suggests an ad hoc adjustment for these situations.

Several other authors have proposed different CI methods for differences in correlated proportions. Tango^[35] proposed a score-based interval for the difference of two matched proportions, which he found to have better coverage than several other intervals used for comparison. Agresti and Min^[36] propose a simple adjustment for the usual Wald CI for the difference of two proportions, which consists of adding 2 to each observed frequency. They find that this adjustment results in an interval that provides coverage much closer to the nominal value. May and Johnson^[37] compare several CIs for correlated binary proportions. Newcombe^[38] also compared several different CI methods and proposed another.

GENERALIZATIONS AND EXTENSIONS OF McNEMAR'S TEST

Statistical methods that generalize McNemar's test to closely related problems have been developed by several statisticians. McNemar's test compares two proportions resulting from the same or matched units, whereas Cochran's Q test^[9] compares n such proportions. The statistic is distributed approximately as chi-square with $n - 1$ degrees of freedom. Cochran's Q statistic is identical to the McNemar statistic when $n = 2$.

The McNemar test is sometimes referred to as a test of symmetry across a diagonal in a 2×2 table. Bowker's test is used to assess symmetry across a diagonal in a square table of higher dimension.^[39]

Another generalization of the McNemar test applies to paired ordinal categorical data, a natural extension of the binary case to three or more categories. Agresti^[40] describes a test that is an adaptation of the Mann-Whitney test, which takes into account the ordering of the variables.

Hamdan, Pirie, and Arnold^[41] proposed methods for testing the simultaneous equality of proportions of unlike pairs in several independent populations.

Paired-binary methods that extend McNemar's test to clustered data have been developed by Eliasziw and Donner,^[42] Obuchowski,^[43] and Durkowski et al.^[44] These methods may be applied when the clustering involves sampled individuals in different geographical regions. Another application is the consideration of multiple measured sites or units on the same individual to comprise a cluster.

Other specialized situations leading to extensions of McNemar's test have drawn the attention of several authors. Marascuillo and Serlin^[45] proposed a test to compare paired results for two or more independent treatments for identical change. Prescott's test and the Mainland-Gart test allow for a period effect in comparing paired proportions resulting from a crossover trial.^[46] Lachenbruch and

Lynch^[47] propose several modifications of McNemar's procedure to assess equivalence using a compound null hypothesis. Klingenberg and Agresti^[48] have proposed a multivariate extension of McNemar's test and show that it is a generalized score test under a GEE approach.

CONCLUSIONS

McNemar's test is a simple and widely used procedure for comparing two dependant or matched proportions. Although it was developed in 1947, no alternative has replaced it as the most commonly used method for its purpose. Since its introduction, statisticians have been proposing alternatives, improvements, sample size methods, CIs, and extensions related to the test.

ARTICLES OF FURTHER INTEREST

Case-Control Studies, Inference in, p. 1.
Crossover Design, p. 255.
Sample Size Determination, p. 899.

REFERENCES

1. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157.
2. Edwards, A.L. Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika* **1948**, *13*, 185–187.
3. Mostellar, F. Some statistical problems in measuring the subjective response to drugs. *Biometrics* **1952**, *8*, 220–226.
4. McNemar, Q. *Psychological Statistics*; John Wiley and Sons: New York, 1962, 52–61.
5. Siegel, S. *Nonparametric Statistics for the Behavioral Sciences*; McGraw Hill: New York, 1956, 63–67.
6. Agresti, A. *Categorical Data Analysis*, 2nd Ed.; Wiley: New York, 2002, 409–414.
7. Bennett, B.M.; Underwood, R.E. On McNemar's test for the 2×2 table and its power function. *Biometrics* **1970**, *26*, 339–343.
8. Grizzle, J.E.; Starmer, C.F.; Koch, G.G. Analysis of categorical data by linear models. *Biometrics* **1969**, *25*, 489–504.
9. Cochran, W.G. The comparison of percentages in matched samples. *Biometrika* **1950**, *37*, 256–266.
10. Suissa, S.; Shuster, J.J. The 2×2 matched pairs trial: exact unconditional design and analysis. *Biometrics* **1991**, *47*, 361–372.
11. May, W.L.; Johnson, W.D. The validity and power of tests for equality of two correlated proportions. *Stat. Med.* **1997**, *16*, 1081–1096.
12. Irony, T.; Pereira, C.; Tiwari, R. Analysis of opinion swing: comparison of two correlated proportions. *Am. Stat.* **2000**, *54* (1), 57–62.

13. Lloyd, Chris J. A new exact and more powerful unconditional test of no treatment effect from binary matched pairs. *Biometrics* **2008**, *64*, 716–723.
14. Miettinen, O. The matched pairs design in the case of all-or-none responses. *Biometrics* **1968**, *24*, 339–352.
15. Schlesselman, J.J. *Case-Control Studies—Design, Conduct, Analysis*; Oxford University Press: New York, 1982.
16. Connor, R. Sample size for testing differences in proportions for the paired-sample design. *Biometrics* **1987**, *43*, 207–211.
17. Dupont, W. Power calculations for matched case-control studies. *Biometrics* **1988**, *44*, 1157–1168.
18. Connett, J.; Smith, J.; McHugh, R. Sample size and power for pair-matched case-control studies. *Stat. Med.* **1987**, *6*, 53–59.
19. Fleiss, J.L. *Statistical Methods for Rates and Proportions*, 2nd Ed.; Wiley: New York, 1981.
20. Machin, D.; Campbell, M.; Fayers, P.; Pinol, A. *Sample Size Tables for Clinical Studies*; Blackwell Science Inc.: Malden, MA, 1997, 79–98.
21. Parker, R.A.; Bregman, D.J. Sample size for individually matched case-control studies. *Biometrics* **1986**, *42*, 919–926.
22. Lachenbruch, P. On the sample size for studies based upon McNemar's test. *Stat. Med.* **1992**, *11*, 1521–1525.
23. Duffy, S. Asymptotic and exact power for the McNemar test and its analogue with r controls per case. *Biometrics* **1984**, *40*, 1005–1015.
24. Lachin, J. Power and sample size evaluation for the McNemar test with application to matched case-control studies. *Stat. Med.* **1992**, *11*, 1239–1251.
25. Royston, P. Exact conditional and unconditional sample size for pair-matched studies with binary outcome: a practical guide. *Stat. Med.* **1993**, *12*, 699–712.
26. Lehr, R. Some practical considerations and a crude formula for estimating sample size for McNemar's test. *Drug Inform. J.* **2001**, *35*, 1227–1233.
27. Woolson, R.; Bean, J.A.; Rojas, B. Sample size for case-controlled studies using Cochran's statistic. *Biometrics* **1986**, *42*, 927–932.
28. Feuer, E.J.; Kessler, L.G. Test statistic and sample size for a two-sample McNemar test. *Biometrics* **1989**, *45*, 629–636.
29. Lu, Y.; Bean, J.A. On the sample size for one-sided equivalence of sensitivities based upon McNemar's test. *Stat. Med.* **1995**, *14*, 1831–1839.
30. Snedecor, G.W.; Cochran, W.G. *Statistical Methods*; Iowa University Press, 1980, 121–124.
31. Breslow, N.; Day, N. *Statistical Methods in Cancer Research. Vol. 1. The Analysis of Case-Control Studies*; International Agency for Research in Cancer: Lyon, 1980; 165–166.
32. Armitage, P.; Berry, G. *Statistical Methods in Medical Research*; Blackwell Scientific Publications: London, 1987, 1–123.
33. Lehr, R.G. Letter to the editor. *Am. Stat.* **2000**, *54* (4), 325.
34. Lloyd, C.J. Confidence intervals from the difference between two correlated proportions. *JASA* **1990**, *85*, 1154–1158.
35. Tango, T. Equivalence test and confidence interval for the difference in proportions for the paired sample design. *Stat. Med.* **1998**, *17*, 891–908.
36. Agresti, A.; Min, Y. Simple improved confidence intervals for comparing matched proportions. *Stat. Med.* **2005**, *24*, 729–740.
37. May, W.L.; Johnson, W.D. Confidence intervals for differences in correlated binary proportions. *Stat. Med.* **1997**, *16*, 2127–2136.
38. Newcombe, R.G. Improved confidence intervals for the difference between binomial proportions based on paired data. *Stat. Med.* **1998**, *17*, 2635–2650.
39. Sachs, L. *Applied Statistics, A Handbook of Techniques*; Springer Verlag: New York, 1984.
40. Agresti, A. *Analysis of Ordinal Categorical Data*; John Wiley & Sons: New York, 1984, 1–208.
41. Hamdan, M.A.; Pirie, W.R.; Arnold, J.C. Simultaneous testing of McNemar's problem for several populations. *Psychometrika* **1975**, *40*, 153–161.
42. Eliasziw, M.; Donner, A. Application of the McNemar test to non-independent matched pair data. *Stat. Med.* **1991**, *10*, 1981–1991.
43. Obuchowski, N. On the comparison of correlated proportions for clustered data. *Stat. Med.* **1998**, *17*, 1495–1507.
44. Durkowsky, V.; Palesch, Y.; Lipsitz, S.; Rust, P. Analysis of clustered matched-pair data. *Stat. Med.* **2003**, *22*, 2417–2428.
45. Marascuillo, L.A.; Serlin, R.C. Tests and contrasts for comparing change parameters for a multiple sample McNemar data model. *Br. J. Math. Stat. Psychol.* **1979**, *32*, 105–112.
46. Senn, S. *Cross-over Trials in Clinical Research*; Wiley: Chichester 1993, 106–108.
47. Lachenbruch, P.A.; Lynch, J. Assessing screening tests: extensions of McNemar's test. *Stat. Med.* **1998**, *17*, 2207–2217.
48. Klingenberg, B.; Agresti, A. Multivariate extensions of McNemar's test. *Biometrics* **2006**, *62*, 921–928.

MCP-Mod: Dose–Response Testing and Estimation

Björn Bornkamp

Clinical Development & Analytics, Novartis Pharma AG, Basel, Switzerland

Naitee Ting

Boehringer-Ingelheim Pharmaceuticals, Inc., Ridgefield, Connecticut, U.S.A.

Abstract

Multiple comparison procedures and modeling (MCP-Mod) is a method for dose–response testing and dose–response estimation under model uncertainty. The dose–response modeling is based on a set of candidate dose–response shapes that is predefined at the design stage of the trial. The methodology consists of two main steps. The first step is a statistical test for a dose–response signal, assessing whether there is any effect of the test doses over placebo (the MCP step). This is performed by utilizing a multiple contrast test procedure, where each of the candidate shape is represented by a contrast. In the second step, the dose–response curve is estimated based on model selection or model averaging techniques, utilizing the predefined candidate dose–response models.

Keywords: Dose finding; Dose–response; Model uncertainty; Multiple contrast test; Phase II study.

INTRODUCTION

Multiple comparison procedures and modeling (MCP-Mod) is a method for dose–response testing and dose–response estimation under model uncertainty. The dose–response modeling is based on a set of candidate dose–response shapes that is predefined at the design stage of the trial. The methodology consists of two main steps. The first step is a statistical test for a dose–response signal, assessing whether there is any effect of the test doses over placebo (the MCP step). This is performed by utilizing a multiple contrast test procedure, where each of the candidate shapes is represented by a contrast. In the second step, the dose–response curve is estimated based on model selection or model averaging techniques, utilizing the predefined candidate dose–response models.

The goal of this entry is to provide an introduction and review of the MCP-Mod methodology. The section “Dose finding and MCP-Mod” contains a high-level introduction to the procedure, and then section “MCP-Mod: Literature Review and Historical Overview” reviews the historical development of MCP-Mod and references related to MCP-Mod. The sections “MCP-Mod for Normally Distributed Data” and “MCP-Mod for General Data” provide a more detailed technical overview of the method, first for normally distributed data and then for general parametric models. The section “Design Aspects” considers study design aspects related to MCP-Mod. The section “Scope of MCP-Mod” presents more details on which

dose-finding problems MCP-Mod applies to and to which it does not apply, and what alternative methods can be useful in other situations. The final section reviews software implementations available for MCP-Mod.

DOSE-FINDING AND MCP-Mod

A general overview of dose-finding in the whole clinical drug development is given in the book by Ting.^[1] The MCP-Mod methodology applies primarily to phase II dose-finding studies. These studies play an important role in drug development—they are the last study/studies done before the start of the confirmatory studies in phase III. Because of this, they play an important role both in supporting the go/no-go decision to phase III and also in terms of dose-finding and dose- selection for phase III clinical trials.

The primary regulatory guidance for dose finding is the ICH E4 document on Dose–Response Information to Support Drug Registration. Despite the fact that this document is now years old, dose-finding and dose selection are often inadequately performed. For example, Cross et al.^[2] examined post approval changes in dosages during the time period 1980–1999. They concluded that around 20% of the drugs had their dosages changed. In 80% of those cases, the dose was reduced. Sacks et al.^[3] provided an overview on the scientific reasons for delay or denial of approval of a drug by the Food and Drug Administration (FDA) during the time period of 2000–2012. Of those submissions that

were not approved the first-time, uncertainties related to dose selection was one of the most common deficiencies.

A more prominent role of dose–response modeling in phase II studies was identified as a way of improving the design and analysis of dose-finding in drug development. Traditionally, phase II dose–response studies are designed for pair-wise testing approaches. These methods do not directly focus on selection of a specific dose or appropriate determination of the full dose–response curve. In addition, if multiple test strategies are used, there is a further penalty for utilizing multiple dose levels. When analyzing a study for dose–response (trend) testing and estimation, more emphasis will be put on estimation of the dose–response curve and so more informative designs should result, in the sense of being able to investigate more doses and more informatively selected doses.

An extensive overview and simulation-based comparison of different methods is provided in the white papers of the Pharmaceutical Research and Manufacturers of America (PhRMA) group on adaptive-dose ranging studies^[4,5] studying parametric and non-parametric as well as Bayesian and non-Bayesian methods. Parametric methods to estimate the dose–response curve using the sigmoid Emax dose–response model have been proposed, for example, in the article by Thomas^[6] in a Bayesian setting and in the article by Dragalin, Hsuan, and Padmanabhan^[7] in a non-Bayesian setting. Empirical evidence of the appropriateness of the sigmoid Emax model using a large-scale meta-analysis is provided in the article by Thomas, Sweeney, and Somayaji.^[8] Non-parametric Bayesian methods for estimating the dose–response curve based on dynamic linear models have been considered,

for example, in the articles by Grieve and Krams^[9] and Berry et al.,^[10] often coupled with the use of adaptive designs. Non-Bayesian flexible methods to estimate the dose–response curve based on splines are, for example, discussed in the articles by Desquilbet and Mariotti^[11] and Helms et al.^[12] The MCP-Mod method discussed in detail here is an approach somewhere between the parametric and non-parametric modeling, because a candidate set of parametric dose–response curves is used to model the dose–response curve. Depending on the diversity of the candidate models, this approach is rather flexible in terms of estimating the dose–response curves.

Fig. 1 outlines the steps it takes to implement the MCP-Mod procedure in a dose-finding trial. At the trial design stage, one needs to include design considerations as in any other clinical trial, for example, selection of study population and study end point. What is special about MCP-Mod is that a preselection of candidate dose–response models is performed, which will be utilized as part of the methodology. Ideally, this candidate set of dose–response shapes should reflect the available clinical information on the shape of the dose–response curve. That is when the uncertainty on the shape is large, a broad set of candidate shapes and models are selected, and when the uncertainty is less, a more narrow set is selected. Together with an assumption on the maximum effect size of the drug, the chosen candidate set of models can not only be used to derive the total sample size of the trial, but also to evaluate the necessary doses needed for efficient estimation of the dose–response curve or target doses of interest. Section “Design Aspects” will elaborate more on the selection of the candidate set and design considerations.

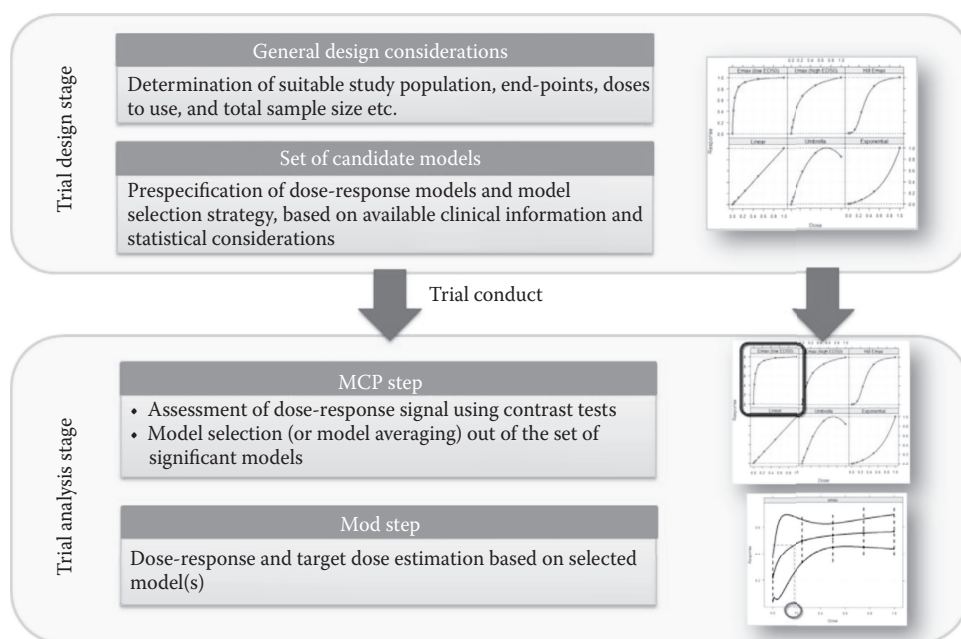


Fig. 1 Overview of the different steps involved, when implementing the MCP-Mod procedure in a dose-finding trial

At the analysis stage, the MCP-Mod method consists of two main steps. The first step is a statistical test for a dose–response signal, assessing whether there is any effect of the test doses over placebo (the MCP step). This is performed by utilizing a multiple contrast test that is optimized for the candidate shapes included. In the second step, the dose–response curve is estimated based on model selection or model averaging techniques, utilizing the predefined candidate dose–response models. The analysis methodology will be discussed in more detail in the sections “MCP-Mod for Normally Distributed Data” and “MCP-Mod for General Data.”

MCP-Mod: LITERATURE REVIEW AND HISTORICAL OVERVIEW

The original MCP-Mod procedure of Bretz, Pinheiro, and Branson^[13] was motivated by the work of Tukey, Ciminera, and Heyse,^[14] who proposed to use several trend tests simultaneously based on different functional dose–response descriptions and to subsequently adjust the resulting *p* values for multiplicity. Since then, the methodology has been subject to several investigations. Pinheiro, Bornkamp, and Bretz^[15] discussed practical considerations regarding the implementation of this methodology, including sample size calculations for the MCP part. Dette et al.^[16] constructed optimal designs for target dose estimation to determine the optimum location of doses and allocation of patients to the dose levels, taking model uncertainty into account. Klingenberg^[17] applied the MCP-Mod approach to proof-of-concept studies with binary responses. Tao et al.^[18] incorporated both efficacy and safety end points through joint modeling. Dette et al.,^[19] Baayen, Hougaard, and Phipper,^[20] and Gutjahr and Bornkamp^[21] investigated the use of likelihood ratio tests instead of multiple contrast tests. Miller^[22] and Franchetti, Anderson, and Sampson^[23] extended the MCP-Mod approach to flexible two-stage designs that include the possibility to perform adaptations (e.g., adding or dropping doses) under a strict family-wise error rate (FWER) control. Bornkamp et al.,^[24] Vandemeulebroecke et al.,^[25] and Mercier et al.^[26] investigated the MCP-Mod procedure in response-adaptive designs, addressing two major challenges in dose-finding studies—uncertainty about the dose–response models and large variability in parameter estimates. Schorning et al.^[27] investigated the Mod step of MCP-Mod in more detail and compared different model selection criteria and model selection versus model averaging, using a simulation study. While no uniformly best model selection criteria can be established, model averaging approaches outperform the model selection approaches in the simulation scenarios covered. König et al.^[28] extended the MCP part of the MCP-Mod procedure using the closed test principle so that pair-wise comparisons between test treatment and control can be performed, controlling the FWER for

these comparisons, so that the MCP part of MCP-Mod could also be utilized in confirmatory studies. A general overview of MCP-Mod and many practical implementation aspects were provided by Xun and Bretz.^[29] Design aspects such as ways of determining the doses to utilize in a dose-finding trial and the total sample size are reviewed in the article by Pinheiro and Bornkamp.^[30]

In 2014, the MCP-Mod procedure received a qualification opinion by the European Medicines Agency, [Committee for Medicinal Products for Human Use (CHMP)] as an efficient statistical methodology for model-based design and analysis of phase II dose-finding studies under model uncertainty.^[31] In 2016, the methodology received a fit-for-purpose determination by the FDA.^[32]

MCP-Mod FOR NORMALLY DISTRIBUTED DATA

Assume a primary response variable *y* for a given set of parallel groups of patients corresponding to test doses *d*₂, *d*₃, ..., *d*_{*k*} plus placebo *d*₁, for a total of *k* arms. For the purpose of dose–response signal testing and estimating target doses, we consider the model specification:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2), j = 1, \dots, n_i, i = 1, \dots, k \quad (1)$$

where the mean response at dose *d*_{*i*} can be represented as $\mu_i = \mu(d_i, \theta)$ for some dose–response model $\mu(\cdot)$, parameterized by a vector of parameters θ , *n*_{*i*} denotes the number of patients allocated to dose *d*_{*i*}, and ε_{ij} is the error term for patient *j* within dose group *i* (assuming independent residuals). We note that most dose–response models used in practice can be written as $\mu(d, \theta) = \theta_0 + \theta_1 f^0(d, \theta_2)$, and $f^0(d, \theta_2)$ is typically a non-linear transformation of the dose levels depending on θ_2 . The parameters θ_0 and θ_1 determine location and scaling of the function *f*. An example for f^0 is $f^0(d, \theta_2) = d/(d + \theta_2)$, resulting in the (hyperbolic) Emax model and where θ_2 is the ED₅₀ parameter of the Emax model. More dose–response models are given in

Table 1 Example dose–response models often used in clinical trials^a

Model name	$f(d, \theta)$	(*)
Linear	$E_0 + \delta d$	
Emax	$E_0 + E_{\max} d / (ED_{50} + d)$	ED ₅₀
exponential	$E_0 + E_1 (\exp(d/\delta) - 1)$	δ
Sigmoid	$E_0 + E_{\max} d^h / (ED_{50}^h + d^h)$	$(ED_{50}, h)'$
Emax		
Beta model	$E_0 + E_{\max} B(\delta_1, \delta_2) (d/D)^{\delta_1} (1 - d/D)^{\delta_2}$	$(\delta_1, \delta_2)'$

^aColumn (*) lists for each model the parameters that determine the shape of the model. For the beta model, $B(\delta_1, \delta_2) = (\delta_1 + \delta_2)^{\delta_1 + \delta_2} / (\delta_1^{\delta_1} \delta_2^{\delta_2})$ and for the quadratic model $\delta = \frac{\beta_2}{|\beta_1|}$.

Table 1, with example shapes for each model presented in Fig. 2. More details on these dose–response models can be found in the article by Pinheiro, Bretz, and Branson^[33] and a guide on how to select candidate dose–response models for a specific trial is given in the section “Design Aspects.”

The first step of MCP-Mod is to identify a set of M parameterized candidate models together with parameter values for their standardized versions. When used with the doses planned for the trial, $d_1, \dots, d_k$, these candidate models produce mean response vectors $\boldsymbol{\mu}_m = (\mu_{m1}, \dots, \mu_{mk})$, where $\mu_{mi} = \mu_m(d_i, \theta_{2m})$ with a specified parameter θ_{2m} and μ_m is the m^{th} dose–response function, $m = 1, \dots, M$.

The second step is to test each of the dose–response shapes in the candidate set using a single contrast test, with coefficients chosen to maximize the power of the test when the true underlying mean response equals $\boldsymbol{\mu}_m$. This approach formulates the test for a dose–response trend in terms of a linear contrast hypothesis $\mathbf{c}'\boldsymbol{\mu}$, where $\mathbf{c}$ is a contrast vector subject to $\sum_i c_i = 0$ and $\boldsymbol{\mu}$ is the true mean at the dose levels. For a given vector $\boldsymbol{\mu}$, optimum contrast coefficients for model testing are proportional to $n_i(\mu_i - \mu^*)$, $i = 1, \dots, k$, where $\mu^* = \boldsymbol{\mu}'\mathbf{1}/k$. After

normalization, this leads to the unique solution (up to the sign) $\mathbf{c}/\|\mathbf{c}\|$. Each of M possible dose–response shape gets represented by one contrast $\mathbf{c}_m = (c_{m1}, \dots, c_{mk})$. For example, a linear contrast test for equally spaced doses and balanced patient allocation is defined such that the difference of any two adjacent contrast coefficients is a constant. Assuming that the linear model has been included in the candidate model set, the linear contrast test is then a powerful test to detect the linear trend. Similarly, any dose–response relationship characterized through $\boldsymbol{\mu}_m$ can be tested equally powerfully by selecting an appropriate contrast test, whose coefficients are defined in dependence of the assumed $\boldsymbol{\mu}_m$. Note that contrast tests are shift and scale invariant, and thus it is sufficient to work with the standardized modeling functions f^0 .

The third step is to test for an overall dose–response signal. The single contrast test is defined as:

$$T_m = \frac{\sum_{i=1}^k c_{mi} \bar{Y}_i}{S \sqrt{\sum_{i=1}^k c_{mi}^2 / n_i}}, \quad m = 1, \dots, M, \quad (2)$$

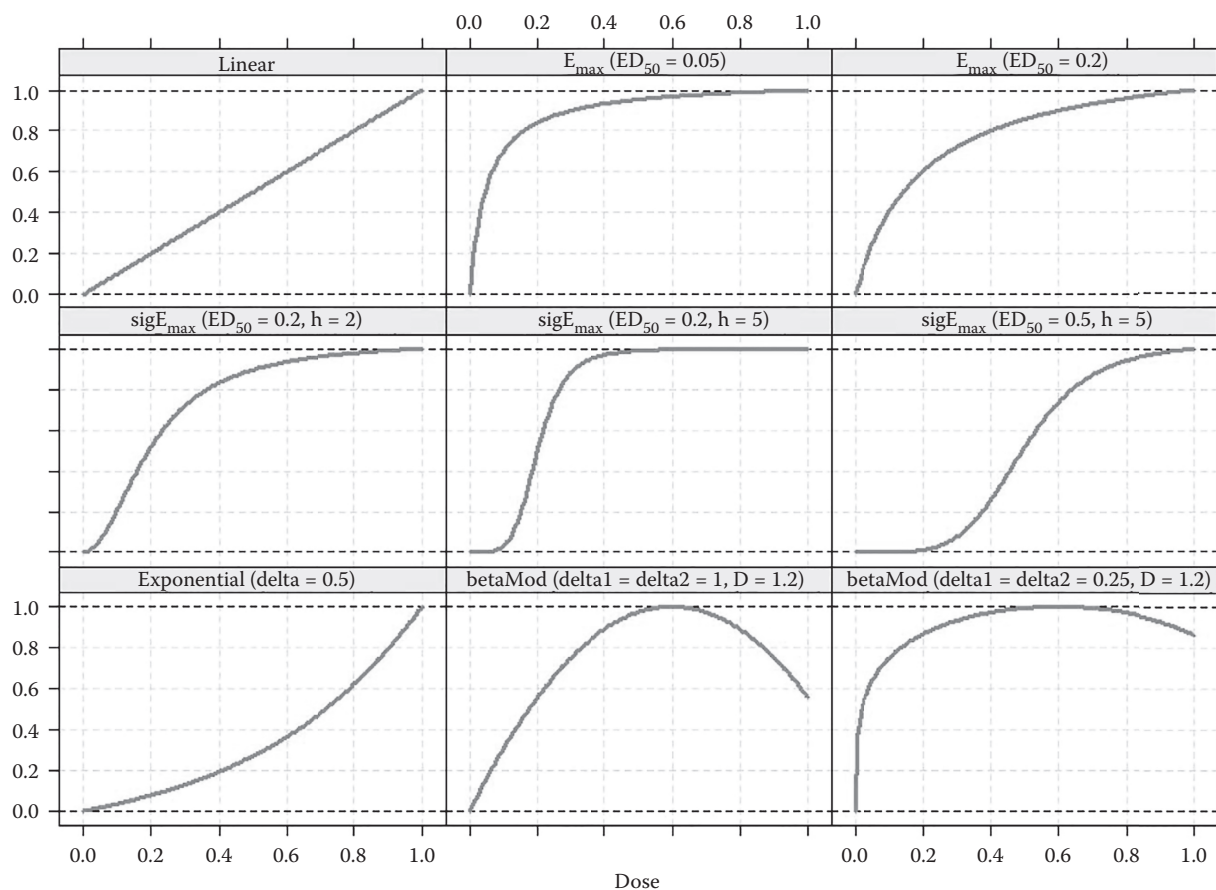


Fig. 2 Example dose–response shapes for the models in Table 1. For simplicity the dose range is from $[0,1]$, the placebo effect is assumed to be 0, and the maximum effect is assumed to be 1

Where,

$$S^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 / (N - k) \quad (3)$$

denotes the variance estimate and $N = \sum_i n_i$, the total sample size. Under the null hypothesis $H: \mathbf{c}'\boldsymbol{\mu} = 0$ of no dose-response effect, i.e., $\mu_1 = \dots = \mu_k$, and under the distributional assumptions from above, $T_1, \dots, T_M$ jointly follow a central multivariate t distribution with $N - k$ degrees of freedom and the correlation matrix only depends on the sample sizes and the correlations between the different contrasts. Note that for the data analysis step, the critical value, contrast weights, etc. might need to be recalculated at the analysis stage due to dropouts, covariate effects, etc.

The final test statistic $T_{\max}$ is based on the maximum contrast test, i.e., $T_{\max} = \max_m T_m$. A dose-response signal is verified when this maximum statistic $T_{\max}$, and thus at least one single contrast test is statistically significant while controlling the FWER at prespecified level α . Let $q_{1-\alpha}$ denotes the multiplicity adjusted critical value. A dose-response signal is then established if $T_{\max} \geq q_{1-\alpha}$. In fact, any dose-response model with a test statistic larger than $q_{1-\alpha}$ can be declared statistically significant at level α . If no candidate model is statistically significant, the MCP-Mod procedure stops indicating that a dose-response signal cannot be established from the observed data.

The fourth step is to select one model out of the reference set or to perform model averaging. Model selection criteria such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC) can be used for model selection or model averaging. An overview of different approaches is presented in the study by Schorning et al.^[27]

The fifth and final step consists of estimating the dose-response curve and target doses of interest using the dose-response model(s) selected in the previous step.

MCP-Mod FOR GENERAL DATA

In this section, we describe an extension of the MCP-Mod procedure to general parametric end points and more complex situations than discussed previously. For example, the response variable might be a count, binary or time-to-event variable, or in a normally distributed setting, the variance might be different at different dose groups (e.g., proportional to the dose or the mean response). In addition, the final analysis has to be usually adjusted for relevant covariates (e.g., region, age, etc.). Patient measurements are often recorded over time (necessitating the use of longitudinal models), and patients might receive more than one treatment (such as in crossover or incomplete block designs). The core ideas of MCP-Mod remain applicable in

these situations, as sketched in the following and laid out in the article by Pinheiro et al.^[34] in more detail.

Let $\mathbf{y}$ denote the response vector of an experimental unit in the trial (e.g., a patient) which has been assigned a dose d (the formulation can easily be extended to the case of multiple doses). We assume that the residual distribution function is given by $\mathbf{y} \sim F(\mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\mu}(x))$, where $\boldsymbol{\mu}(d)$ denotes the dose-response parameter, $\boldsymbol{\eta}$ the nuisance parameters (fixed effects and random effects), and $\mathbf{z}$ the possible covariates. The key features of MCP-Mod can then be formulated with respect to $\boldsymbol{\mu}(d)$, including:

- accounting for uncertainty in the dose-response model via a set of candidate dose-response models
- testing for a dose-response signal via optimal contrasts based on plausible dose-response shapes
- model selection or model averaging to combine different models
- dose-response estimation and dose selection using non-linear regression.

Because all dose-response information is assumed to be represented by $\boldsymbol{\mu}(d)$, the interpretability of this parameter is critical for communicating with clinical teams, choosing candidate dose-response shapes, specifying clinically relevant effects, and etc.. As an example, consider the Weibull distribution, which is typically parameterized by a scale parameter λ and shape parameter α , neither of which is easily interpretable. For the purpose of interpretability, the model could be reparameterized in terms of the median time-to-event $\mu = \log(2)^{1/\alpha} / \lambda$, and α , and then use μ as an interpretable dose-response parameter.

Let $\hat{\boldsymbol{\mu}}$ denotes the vector of estimated dose-response parameters under an analysis of variance (ANOVA) parameterization, obtained using the appropriate estimation method for the general parametric model above via maximum likelihood (ML), generalized estimating equations, partial likelihood, etc. The key assumption needed for this general version of MCP-Mod to work is that the response estimates $\hat{\boldsymbol{\mu}}$ at the doses are asymptotically multivariate normal distribution with covariance matrix $\mathbf{S}$. This approximation can be shown to hold for most parametric estimation problems, such as, generalized linear models, parametric time-to-event models, mixed-effects models, etc. The MCP step consists of specifying a set of candidate models for the dose-response relationship $\boldsymbol{\mu}(d)$. Similar to the original MCP-Mod approach, each of the candidate model shape determines an optimal contrast for a trend test to evaluate the associated dose-response model signal. The optimal contrasts are applied to the previously described ANOVA-type estimates $\hat{\boldsymbol{\mu}}$, with the associated asymptotic distribution used for implementing the corresponding tests (i.e., critical values and p values). It can be shown that the (optimal) contrast for testing the hypothesis that $\mathbf{c}'\boldsymbol{\mu} = 0$ with maximal power for a single given candidate model shape $\boldsymbol{\mu}_m$ is given by $\mathbf{S}^{-1}(\boldsymbol{\mu}_m - (\boldsymbol{\mu}_m'\mathbf{S}^{-1})/(\mathbf{1}'\mathbf{S}^{-1}\mathbf{1}))$, which

generalizes the formula given in the previous section. Once a dose–response signal is established, one proceeds to the Mod step, fitting the dose–response profile and estimating target doses based on the models identified in the MCP step. There are several ways to fit the dose–response models to the observed data, including approaches based on maximizing the likelihood or the restricted likelihood. Pinheiro et al.^[34] suggested an alternative two-stage approach to the dose–response model fitting based on generalized least squares, where the dose–response model is directly fitted to $\hat{\mu}$ utilizing the inverse of $\hat{S}$ as a weighting matrix. This approach has some computational advantages compared to a full likelihood approach. Although this approach relies on asymptotic results, it has the appeal of being a general purpose application, as it depends only on $\hat{\mu}$ and $\hat{S}$.

DESIGN ASPECTS

There are at least three important design considerations regarding the use of MCP-Mod: 1) selection of the candidate set of models at the design stage; 2) selecting doses to utilize in the trial; and 3) selecting the total sample size of the trial. The material in this section is based on the article by Xun and Bretz^[29] as well as the article by Pinheiro and Bornkamp.^[30]

Selection of the Candidate Set of Models

As a rule of thumb, in many cases, a set of three to seven dose–response shapes corresponding to two to four dose–response models can be appropriate. Note that multiple dose–response shapes might be used for each dose–response model, where the shapes define the contrasts to be used in the MCP step and the models are those to be fitted in the Mod step. There are at least three considerations that play a role in selection of the candidate set of models.

1. *Previous information:* There might exist information for similar compounds in the same indication or the same compound in a different indication. Also an early longitudinal dose–exposure–response model might have been developed on earlier data, which might be used to predict the dose–response curve at the time point of interest.
2. *General experience:* An Emax model should always be included in the candidate model set. A meta-analysis of the shape of dose–response curve, conducted by Thomas et al.,^[8] showed that the sigmoid Emax model can adequately describe the dose–response shape observed in a variety of cases.
3. *Statistical considerations:* If only few test doses are available, it is difficult to fit complex models, such as the sigmoid Emax with four parameters in a trial with three test doses. Such models would thus have to be excluded from the candidate model set and one

would rather rely on more dose–response models with less parameters to obtain an adequate breadth of the candidate set. Another consideration is to evaluate the performance of MCP-Mod under model misspecification. For example, the impact of omitting the true dose–response model from the candidate model set can be negligible if other models in the candidate set can pick up the omitted model. This can be evaluated for the MCP part (impact on power) using explicit calculations and for the Mod part (impact on precision for dose–response and dose estimation) using simulations. Using a broad candidate model set may have an impact on the degree of multiplicity adjustment for the MCP part and increase the critical value. The required sample size to achieve a designated power tends to increase with a broader candidate model set. A similar trade-off exists also in terms of dose–response model fitting, because a broader candidate set will decrease potential bias (in case of model misspecification) but increase the variance of the estimates.

Selection of the Doses to Utilize in the Trial

In many clinical projects, the doses that can be chosen are limited, because only a specific dose strength might be available for administration. Nevertheless, by combining different dose strengths, a variety of doses might be available to be administered. How are specific doses selected to be utilized? As a rule of thumb, four to seven doses and a greater than tenfold dose range (ratio of highest to lowest dose) should be considered (see European Medicines Agency^[35]). Regarding the dose spacing, often log-dose spacing is utilized. The heuristic justification for this is that the variance of drug exposure metrics inside the body increases with the dose (often following a log-normal distribution), which means that the lower dose range needs to be investigated more densely to cover the exposure ranges uniformly. On the other hand, if doses are spaced too close in the upper dose range, they will be too similar in terms of exposure (due to the increase in variance). Motivated by considerations related to safety and also to ensure proper estimation of the steeply increasing part of the dose–response curve, Hamlett et al.^[36] introduce the binary dose spacing design. In this proposal, the designed doses have a twofold, or more generally a p-fold difference, so that, in this design, the lower dose range is also sampled more intensely. An alternative and more formal approach to evaluate different designs is to use optimal design theory based on the candidate dose–response models, as discussed from a practical viewpoint by Pinheiro and Bornkamp.^[30]

Determination of the Total Sample Size

Calculation of the total sample size for the trial can be based on two aspects of MCP-Mod—the MCP-part and

the Mod-part. When calculating the sample size for the MCP-part, the sample size for the study is calculated so that a specific summary power across the different candidate models considered is achieved (e.g., average power or minimum power across all candidate models larger than a specific value). This approach is discussed in detail in the article by Pinheiro et al.^[15] An alternative approach is to base sample size calculations on estimation aspects, for example, estimating the dose–response curve with a given precision or target doses of interest with a given precision. Pinheiro and Bornkamp^[30] discuss different approaches to do this and compare this to the sample sizes obtained, when calculating the sample size based on the MCP part. In general, sample sizes obtained based on modeling considerations tend to be considerably higher than when considering the MCP part alone.

SCOPE OF MCP-Mod

It should be noted that in clinical development, the term “dose” in “dose-finding” refers more generally to a treatment regimen, i.e., the combination of dose and dosing interval. Even more broadly speaking, aspects like the duration of therapy, the route of administration and dosage form (oral, injection, etc.), the time of dosing (after a meal, in the morning, etc.), finding partner drugs or drug–drug interactions, and the population are closely related to the question of dose-finding.

Certainly the MCP-Mod methodology is not applicable to all of these questions. MCP-Mod applies, when the primary focus of the study is the investigation of the population dose–response curve at a specific time point in the context of a phase II study. The method is most informative if it is applied at a time point when the dose–response has stabilized over time and reached its full effect. It applies to parallel group or crossover dose–response studies, and when at least three test doses are utilized, and a placebo (or placebo-like) comparator.

The method does not apply, when the major focus of the study is the determination of the administration frequency and not the dose within given regimens. It is possible to extend MCP-Mod to allow modeling of more than one administration frequency in certain situations. Then additional modeling assumptions would need to be made (e.g., assuming some of the model parameters are shared between regimen and others are different between regimens, see Feller et al.^[37]). MCP-Mod also does not apply, when very different irregular treatment regimens are compared (e.g., some regimens with loading dose versus others without loading dose) or different administration forms are used in the same study e.g., intravenous, subcutaneous injection, etc. When the main study objective is to model the evolvement of the treatment effect over time, the method does not apply, although extensions of MCP-Mod exist that accommodate this when additional modeling

assumptions are being made (e.g., see Pinheiro et al.^[34]). The method should not be used when patients are titrated based on efficacy and/or safety outcomes in a dose titration design. Then, a simple population dose–response analysis will be misleading, as other covariates will not be balanced between the dose groups (due to the lack of randomization). The method also does not apply to dose escalation studies with the purpose of determining the maximum tolerated dose.

Adequate tools to approach many of the questions that are not covered by MCP-Mod (e.g., comparing different irregular treatment regimen, modeling different administration forms, etc.) are more pharmacologically motivated models. In these approaches, typically the relationship of administered dose, blood concentrations, and effect are modeled over time. A great introduction to these models is provided by Källén.^[38]

SOFTWARE IMPLEMENTATION

The most flexible way to implement MCP-Mod is via the DoseFinding R package (<http://CRAN.R-project.org/package=DoseFinding>), which contains extensive functionality both for the design and the analysis of dose-finding studies and examples. Worked examples for design and analysis showing extensive code based on the DoseFinding package are given in the articles by Xun and Bretz^[29] and Pinheiro and Bornkamp.^[30] The commercial software package ADDPLAN-DF contains extensive tools for designing dose-finding studies using MCP-Mod, including complex adaptive dose-finding designs, which are not easily available in the DoseFinding package. The company Cytel has developed PROC MCPMOD as an add-on to SAS that focusses on the analysis of dose-finding studies using MCP-Mod. SAS code to implement MCP-Mod analyses is also available in the article by Bornkamp, Bezlyak, and Bretz^[39].

CONCLUSION

This entry provides an introduction to and a general review of the MCP-Mod methodology. The methodology is described and practical implementation considerations are provided as well as pointers to references containing further information are given.

REFERENCES

1. Ting, N. *Dose Finding in Drug Development*; Springer: New York, 2006.
2. Cross, J.; Lee, H.; Westelinck, A.; Nelson, J.; Grudzikas, C.; Peck, C. Postmarketing drug dosage changes of 499 FDA-approved new molecular entities, 1980–1999. *Pharmacoepidem. Dr. S.* **2002**, *11*, 439–446.

3. Sacks, L.; Shamsuddin, H.; Yasinskaya, Y.; Bouri, K.; Lanthier, M.; Sherman, R. Scientific and regulatory reasons for delay and denial of FDA approval of initial applications for new drugs, 2000–2012. *J. Am. Med. Assoc.* **2014**, *311*, 378–384.
4. Bornkamp, B.; Bretz, F.; Dmitrienko, A.; Enas, G.; Gaydos, B.; Hsu, C.H.; Koenig, F.; Krams, M.; Liu, Q.; Neuenchwander, B.; Parke, T.; Pinheiro, J.; Roy, A.; Sax, R.; Shen, F. Innovative approaches for designing and analyzing adaptive dose-ranging trials (with discussion). *J. Biopharm. Stat.* **2007**, *17*, 965–995.
5. Dragalin, V.; Bornkamp, B.; Bretz, F.; Miller, F.; Padmanabhan, S.K.; Patel, N.; Perevozskaya, I.; Pinheiro, J.; Smith, J.R. A simulation study to compare new adaptive dose-ranging designs. *Stat. Biopharm. Res.* **2010**, *2*, 487–512.
6. Thomas, N. Hypothesis testing and Bayesian estimation using a sigmoid Emax model applied to sparse dose response designs. *J. Biopharm. Stat.* **2006**, *16*, 657–677.
7. Dragalin, V.; Hsuan, F.; Padmanabhan, S.K. Adaptive designs for dose-finding studies based on the sigmoid Emax model. *J. Biopharm. Stat.* **2007**, *17*, 1051–1070.
8. Thomas, N.; Sweeney, K.; Somayaji, V. Meta-analysis of clinical dose-response in a large drug development portfolio. *Stat. Biopharm. Res.* **2014**, *6*, 302–317.
9. Grieve, A.; Krams, M. ASTIN: A Bayesian adaptive dose-response trial in acute stroke. *Clin. Trials.* **2005**, *2*, 340–351.
10. Berry, S.; Carlin, B.; Lee, J.; Müller, P. *Bayesian Adaptive Methods for Clinical Trials*; CRC Press: Boca Raton, FL, 2011.
11. Desquilbet, L.; Mariotti, F. Dose-response analyses using restricted cubic spline functions in public health research. *Stat. Med.* **2010**, *29*, 1037–1057.
12. Helms, H.-J.; Benda, N.; Zinserling, J.; Kneib, T.; Friede, T. Spline-based procedures for dose-finding studies with active control. *Stat. Med.* **2014**, *34*, 232–248.
13. Bretz, F.; Pinheiro, J.; Branson, M. Combining multiple comparisons and modeling techniques in dose response studies. *Biometrics.* **2005**, *61*, 738–748.
14. Tukey, J.W.; Ciminera, J.L.; Heyse, J.F. Testing the statistical certainty of a response to increasing doses of a drug. *Biometrics.* **1985**, *41*, 295–301.
15. Pinheiro, J.; Bornkamp, B.; Bretz, F. Design and analysis of dose finding studies combining multiple comparisons and modeling procedures. *J. Biopharm. Stat.* **2006**, *16*, 639–656.
16. Dette, H.; Bretz, F.; Pepelyshev, A.; Pinheiro, J.C. Optimal designs for dose finding studies. *J. Am. Stat. Assoc.* **2008**, *103*, 1225–1237.
17. Klingenberg, B. Proof of concept and dose estimation with binary responses under model uncertainty. *Stat. Med.* **2009**, *28*, 274–292.
18. Tao, A.; Lin, Y.; Pinheiro, J.; Shih, W.S. Dose finding method in joint modeling of efficacy and safety endpoints in phase II studies. *Int. J. Stat. Probab.* **2015**, *4*, 33–45.
19. Dette, H.; Titoff, S.; Volgushev, S.; Bretz, F. Dose response signal detection under model uncertainty. *Biometrics.* **2015**, *71*, 996–1008.
20. Baayen, C.; Hougaard, P.; Pipper, B. Testing effect of a drug using multiple nested models for the dose-response. *Biometrics.* **2015**, *71*, 417–427.
21. Gutjahr, G.; Bornkamp, B. Likelihood ratio tests for a dose-response effect using multiple nonlinear regression models. *Biometrics.* **2016**, DOI: 10.1111/biom.12563.
22. Miller, F. Adaptive dose-finding: Proof of concept with type I error control. *Biometrical J.* **2010**, *52*, 577–589.
23. Franchetti, Y.; Anderson, S.; Sampson, A. An adaptive two-stage dose-response design method for establishing proof-of-concept. *J. Biopharm. Stat.* **2013**, *23*, 1124–1154.
24. Bornkamp, B.; Bretz, F.; Dette, H.; Pinheiro, J. Response-adaptive dose-finding under model uncertainty. *Ann. Appl. Stat.* **2011**, *5* (2B), 1611–1631.
25. Vandemeulebroecke, M.; Bretz, F.; Pinheiro, J.; Bornkamp, B. Adaptive dose-ranging studies. In *Handbook of Adaptive Designs in Pharmaceutical and Clinical Development*; Pong, A., Chow, S.-C., Eds.; CRC Press: Boca Raton, FL, 2011; 11–1–11–16.
26. Mercier, F.; Bornkamp, B.; Ohlssen, D.; Wallstroem, E. Characterization of dose-response for count data using a generalized MCP-Mod approach in an adaptive dose-ranging trial. *Pharma. Stat.* **2015**, *14*, 359–367.
27. Schorning, S.; Bornkamp, B.; Bretz, F.; Dette, H. Model selection versus model averaging in dose finding studies. *Stat. Med.* **2016**, *35*, 4021–4040.
28. König, F.; Krasnozhan, S.; Bornkamp, B.; Bretz, F.; Glimm, E. *Closed MCP-mod for Pairwise Comparisons. Technical Report*; 2016.
29. Xun, X.; Bretz, F. The MCP-mod methodology: Practical considerations and the Dosefinding R package. *Handbook of Methods for Designing, Monitoring and Analyzing Dose Finding Trials* O’Quigley, J.; Iasonos, A.; Bornkamp, B., Eds., Chapman and Hall/CRC: Boca Raton, FL, 2017.
30. Pinheiro, J.; Bornkamp, B. Designing phase II dose-finding studies: Sample size, doses and dose allocation weights. *Handbook of Methods for Designing, Monitoring and Analyzing Dose Finding Trials*; O’Quigley, J., Iasonos, A.; Bornkamp, B., Eds.; Chapman and Hall/CRC: Boca Raton, FL, 2017.
31. European Medicines Agency. Qualification opinion of MCP-Mod as an efficient statistical methodology for model-based design and analysis of phase II dose-finding studies under model-uncertainty. Accessed at: <http://googl/imT7IT2014> (accessed March 1, 2017).
32. Food and Drug Administration. 2016, Determination Letter. Accessed at: <http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/UCM508700.pdf> (accessed March 1, 2017).
33. Pinheiro, J.; Bretz, F.; Branson, M. Analysis of dose-response studies—modeling approaches. In *Dose Finding in Drug Development*; Ting, Naitee, Ed.; Springer: New York, 2006; 146–171.
34. Pinheiro, J.; Bornkamp, B.; Glimm, E.; Bretz, F. Multiple comparisons and modeling for dose finding using general parametric models. *Stat. Med.* **2014**, *33*, 1646–1661.
35. European Medicines Agency. 2015, Report from dose finding workshop. Accessed at: http://www.ema.europa.eu/docs/en_GB/document_library/Report/2015/04/WC500185864.pdf (accessed March 1, 2017).
36. Hamlett, A.; Ting, N.; Choudary Hanumara, R.; Finman, J.S. Dose spacing in early dose response clinical trial designs. *Drug Inf. J.* **2002**, *36*, 855–864.

37. Feller, C.; Schorning, K.; Dette, H.; Bermann, G.; Bornkamp, B. Optimal designs for dose response curves with common parameters. *Ann. Stat.* **2016**, *45* (5), 2102–32.
38. Källén, A. *Computational Pharmacokinetics*; Chapman and Hall: Boca Raton, FL, 2007.
39. Bornkamp, B.; Bezlyak, V.; Bretz, F. Implementing the MCPMod procedure for dose-response testing and estimation. In *Modern Approaches to Clinical Trials Using SAS: Classical, Adaptive, and Bayesian Methods*; Menon, S.M., Zink, R.C., Eds.; SAS Institute: Cary, NC, 2015, 193–224.

Measuring Agreement

Lawrence I-Kuei Lin

Baxter Healthcare Corporation, Round Lake, Illinois, U.S.A.

Abstract

The common theme when assessing the acceptability of a new or generic process, methodology, or formulation is the assessment of the agreement between observations and their target values. This entry reviews methods of performing this type of assessment and provides example applications for further illustration of the main concepts.

INTRODUCTION

Consider the problems of assessing the acceptability of a new or generic process, methodology, and/or formulation in areas of laboratory performance, instrument/assay validation or method comparisons, statistical process control, goodness of fit, and individual bioequivalence. The common theme is to assess the agreement between observations (Y) and their target values (X). Target values may be considered random or fixed. Random target values are measured with error. Common random target values are the gold standard measurements, which are the proven and widely acceptable methods. Sometimes we may also be interested in comparing two methods without a designated gold standard method, or in comparing two technicians (times, reagents, etc.) by the same method. Common fixed target values are the expected or known values.

OVERVIEW

Fig. 1 presents a typical situation for assessing agreement. When we plot the observed values on the y axis vs. the target values over a desirable range on the x axis, we would like to see an agreement in that the observations fall closely along the identity line (zero intercept and unit slope). When there is an evidence of disagreement, it is important to address the sources of the disagreement.

Precision and Accuracy

Common basic sources of disagreement would come from a large within-sample variation (imprecision) and/or a shift in the marginal distributions (inaccuracy). Fixing imprecision is a variance reduction exercise, which typically is more cumbersome than fixing inaccuracy. The cause of inaccuracy is most likely a calibration problem. The concepts of precision and accuracy can be seen in Fig. 2.

Traditional Approaches

Traditionally, the assessment of agreement between observed and target values has been by a paired t -test analysis of the data, which can be misleading. Another slightly better approach to the assessment of agreement has been to test the least squares estimates of the linear line against the identity line. We will denote this method as LS01. These two approaches capture only the accuracy information relative to the precision. The potential for obtaining misleading results using these two approaches can be seen in the situations in Fig. 3.^[1]

Another popular traditional method for assessing agreement of observed and target values has been the Pearson correlation coefficient. However, this is a measure of precision only. The potential for obtaining misleading results using this approach can be seen in the situations in Fig. 4.^[1]

Other traditionally used approaches for assessing the agreement of observed and target values have included the coefficient of variation (measure of precision only), the mean square error of the regression analysis (measure of precision only), and the estimate of intercept and slope (measure of accuracy only). Typically, the target values have been assumed fixed in the regression analysis, even when the target values are obviously random.

The best traditional approach for assessing the agreement of observed and target values has been the intraclass correlation coefficient (ICC). The ICC, in its original form,^[2] is the ratio of between the sample variance and the total (within + between) variance under the model of equal marginal distributions. The ICC is intended to only measure precision. However, when the marginal distributions are different (inaccuracy), the ICC is able to capture this deviation information and consider it as imprecision. Several forms of ICC have evolved, but none has distinct and meaningful components of precision and accuracy.

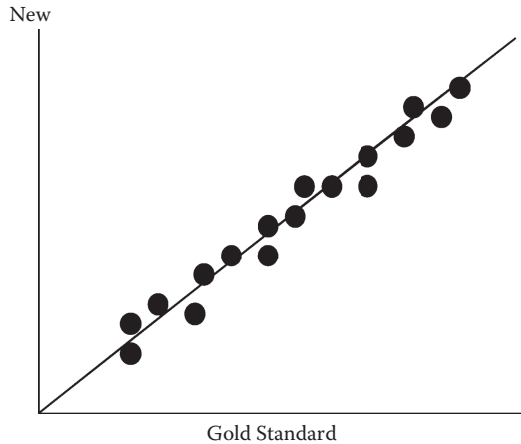


Fig. 1 Assessing the agreement of observed values (new) and target values (gold standard method)

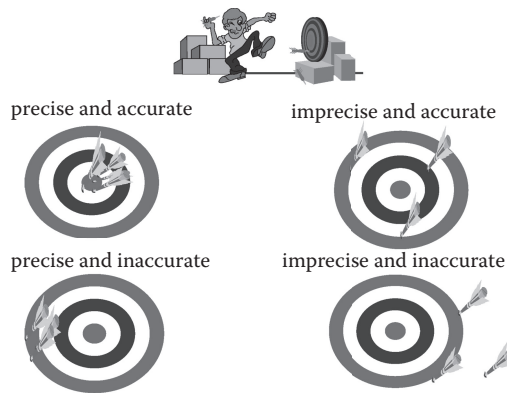


Fig. 2 Precision and accuracy

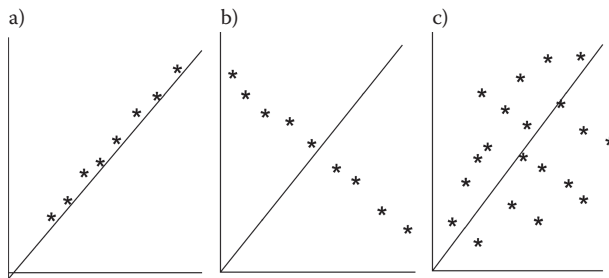


Fig. 3 Situations when paired t -test or least square test against the identity line (LS01) can be misleading. (a) Rejected by paired t -test and by LS01. (b) Accepted by paired t -test and rejected by LS01. (c) Accepted by paired t -test and by LS01

In addition, the traditional hypothesis testing setting is not appropriate for assessing the agreement of observed and target values. In traditional hypothesis testing, the rejection region (alternative hypothesis) is the region for declaring a difference based on strong evidence presented in the data. Failing to reject the null hypothesis does not

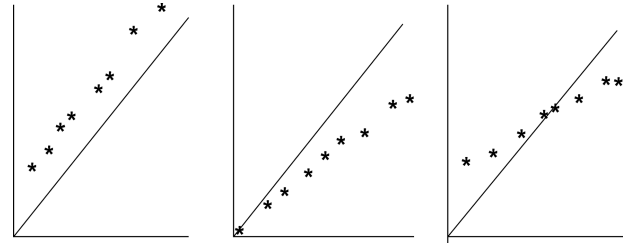


Fig. 4 Situations when Pearson correlation coefficient can be misleading

imply accepting agreement, but implies a lack of evidence for declaring a difference. The proper setting is to reverse the null and the alternative hypotheses, so that the conventional rejection region actually is the region for declaring agreement. Therefore we would reject the null hypothesis of a difference and accept the alternative hypothesis of agreement based on strong evidence presented in the data.^[3–6] Here, a meaningful criterion of an acceptable difference in agreement must be prespecified, and the hypothesis testing is usually one-sided. Actually, the proper hypothesis testing approach is equivalent to computing the one-sided confidence limit for an agreement statistic. If this computed one-sided limit is better than the prespecified criterion, we would accept the alternative hypothesis of agreement. The confidence limit approach is much more straightforward and less confusing than the hypothesis testing approach.

We will now introduce new approaches that evolved in the last 15 years. These new approaches were summarized, studied, and compared by Lin et al.^[7] Most of the following materials have been extracted from the above manuscript.

MODELS

We begin with the most basic model when paired observations (Y and X) are collected.

When Target Values are Random

The joint distribution of Y and X is assumed to have finite second moments with means μ_y and μ_x , variances σ_y^2 and σ_x^2 , and covariance σ_{yx} .

When Target Values are Fixed

$\{y_i|x_i, i = 1, \dots, n\}$ are assumed to be observations in a random sample from the regression model:

$$Y = \beta_0 + \beta_1 X + e_r$$

Here, e_r is the residual error with a mean of zero and a variance σ_e^2 .

METHODS

Mean Square Deviation

This measure evaluates the deviation from the identity line, mean square deviation (MSD) = $E(Y - X)^2$. This is equal to:

1. $\epsilon^2 = (\mu_y - \mu_x)^2 + (\sigma_y^2 + \sigma_x^2 - 2\sigma_{yx})$ when target values are random, or
2. $\epsilon_{|X}^2 = (\mu_{y|\bar{x}} - \bar{x})^2 + s_x^2(1 - \beta_1)^2 + \sigma_e^2$ when target values are fixed

where $\bar{x}$ is the arithmetic average of the X values.

Estimated by sample counterparts (e^2) with a log transformation, $W = \ln(e^2)$ has an asymptotic normal distribution with mean $\ln(e^2)$ and variance:

1. $\sigma_W^2 = 2[1 - (\mu_y - \mu_x)^4/\epsilon^4]/(n - 2)$ when target values are random, or
2. $\sigma_{W|X}^2 = 2[1 - (\epsilon_{|X}^2 - \sigma_e^2)/\epsilon_{|X}^4]/(n - 2)$ when target values are fixed.

In the following, the variance under random ($\sigma_{(.)}^2$) or fixed ($\sigma_{(.)|X}^2$) target values will be presented in similar fashion.

The MSD is difficult to interpret. The intention of the following methods is to put some meaningful interpretation on this basic MSD measurement.

Concordance Correlation Coefficient

The MSD can be standardized such that: 1 = *perfect agreement* ($Y = X$), 0 = *no agreement*, and -1 = *reversed agreement* ($Y = -X$). According to Lin,¹ the concordance correlation coefficient (CCC) is:

$$\rho_c = 1 - \epsilon^2/(\epsilon^2|_{p=0})$$

For ordinal categorical data, CCC degenerates into the weighted kappa suggested by Cohen.^[8,9] The CCC is closely related to the intraclass correlation. The CCC has a meaningful geometric interpretation. It is inversely related to the mean square of the ratio of “within sample total deviation” and “total deviation.” Total deviation is the largest possible deviation among non-negatively correlated samples. Both total deviations in the numerator and the denominator contain the absolute bias. For example, if the within-sample total deviation is 10%, 32%, or 45% of the total deviation, the CCC is 0.99, 0.9, or 0.8, respectively.

The CCC is the product of accuracy and precision. The accuracy measures the closeness of the marginal distributions of Y and X , where 1 = *equal means and variances* and 0 = *no agreement* (far apart). The accuracy can be broken down to measures of location and/or scale shifts, where the

location shift is $v = (\mu_y - \mu_x)/(\sigma_y \sigma_x)^{1/2}$ and the scale shift is $\varpi = \sigma_y/\sigma_x$. Here, accuracy is:

$$\chi_a = 2/(\varpi + 1/\varpi + v^2)$$

Precision measures how close the observations deviate from the best-fit linear line. It is the Pearson correlation coefficient (ρ). ρ^2 has the same scale as accuracy, from 0 (*no agreement*) to 1 (*perfect*).

The CCC estimate (r_c) using sample counterparts by the Z transformation has an asymptotic normal distribution with a mean of $\ln[(1 + \rho_c)/(1 - \rho_c)]/2$ and a variance of:

1. $\sigma_Z^2 = \{(1 - \rho^2)\rho_c^2/[(1 - \rho_c^2)\rho^2] + 2\rho_c^3(1 - \rho_c)v^2/[\rho(1 - \rho_c^2)^2] - \rho_c^4v^4/[2\rho^2(1 - \rho_c^2)^2]\}/(n - 2)$, or
2. $\sigma_{Z|X}^2 = (1 - \rho^2)\rho_c^2/[(n - 2)(1 - \rho_c^2)^2\rho^2] \times [\varpi v^2\rho_c^2 + (1 - \rho_c\rho\varpi)^2 + \varpi^2\rho_c^2(1 - \rho^2)/2]$.

Note that there were typographical errors in Lin.^[1] The above formula for σ_Z^2 is correct.

The accuracy estimate (c_a) using sample counterparts by the logit transformation has an asymptotic normal distribution with a mean of $\ln[\chi_a/(1 - \chi_a)]$ and a variance of:

1. $\sigma_L^2 = [\chi_a^2v^2(\varpi + 1/\varpi - 2\rho) + \chi_a^2(\varpi^2 + 1/\varpi^2 + 2\rho^2)/2 + (1 + \rho^2)(\chi_a v^2 - 1)]/[(n - 2)(1 - \chi_a)^2]$, or
2. $\sigma_{L|X}^2 = [v^2\varpi\chi_a^2(1 - \rho^2) + (1 - \varpi\chi_a)^2(1 - \rho^4)/2]/[(n - 2)(1 - \chi_a)^2]$.

The precision estimate using sample counterparts (r) by the Z transformation has an asymptotic normal distribution with a mean of $\ln[(1 + \rho)/(1 - \rho)]/2$ and a variance of $1/(n - 3)$ when target values are random, or $(1 - \rho^2/2)/(n - 3)$ when target values are fixed.

Total Deviation Index

To examine the agreement from a different perspective, a measure that captures a large proportion (π) of data within a boundary (κ) from target values was considered.^[7,10] Assume that $D = Y - X$ has a normal distribution with a mean of $\mu_d = \mu_y - \mu_x$ and a variance of σ_d^2 . We may find π for a given κ , which is:

$$\pi_\kappa = P(D^2 < \kappa^2) = \chi^2[\kappa^2, 1, \mu_d^2/\sigma_d^2]$$

This measure will be presented later. We may also find κ for a given π , which is:

$$\kappa_\pi = \left\{ \chi^{2(-1)}[\pi, 1, \mu_d^2/\sigma_d^2] \right\}^{1/2}$$

The estimate of this index has intractable asymptotic properties; therefore Lin^[10] suggested the following approximation:

$$\text{TDI}_\pi = \kappa_{\pi^-} = \Phi^{-1}[1 - (1 - \pi)/2]|\epsilon|$$

The approximation is good when $\pi = 0.75$ and $\mu_d^2/\sigma_d^2 \leq 1/2$; $\pi = 0.8$ and $\mu_d^2/\sigma_d^2 \leq 8$; $\pi = 0.85$ and $\mu_d^2/\sigma_d^2 \leq 2$; and $\pi = 0.9$ and $\mu_d^2/\sigma_d^2 \leq 1$, as well as $\pi = 0.95$ and $\mu_d^2/\sigma_d^2 \leq 1/2$. The interpretation of this measure is that approximately $100\pi\%$ of observations are within κ_π from the target values. The TDI_π^2 is proportional to MSD and therefore we may perform inference based on the asymptotic normality of $W = \ln(e^2)$.

Coverage Probability

Now we consider finding the coverage probability (CP) π for a given κ . This is:

1. $\pi_\kappa = P(D^2 < \kappa^2) = \chi^2[\kappa^2, 1, \mu_d^2/\sigma_d^2]$ when target values are random, or
2. $\pi_{\kappa|X} = \sum \pi_{\kappa i}/n$, where $\pi_{\kappa i} = \chi^2[\kappa^2, 1, [\beta_0 + (\beta_1 - 1)x_i/\sigma_e]^2]$ when targeted values are fixed.

The estimate using sample counterparts (p_κ) by the logit transformation has an asymptotic normal distribution with a mean of $\ln[\pi_\kappa/(1 - \pi_\kappa)]$ and a variance of:

1. $\sigma_\pi^2 = \{[\kappa_{+\mu\sigma}\phi(-\kappa_{+\mu\sigma}) + \kappa_{-\mu\sigma}\phi(\kappa_{-\mu\sigma})]^2/(2n) + [\phi(-\kappa_{+\mu\sigma}) - \phi(\kappa_{-\mu\sigma})]^2/[(n-3)(1 - \pi_\kappa)^2\pi_\kappa^2]\}$, or
2. $\sigma_{\pi|X}^2 = [C_0^2/n^2 + (C_0\bar{x} - C_1)^2/(n^2s_x^2) + C_2^2/(2n^2)]/[(n-3)(1 - \pi_\kappa)^2\pi_\kappa^2]$

where

$$\begin{aligned}\kappa_{+\mu\sigma} &= (\kappa + \mu_d)/\sigma_d, \kappa_{-\mu\sigma} = (\kappa - \mu_d)/\sigma_d. \\ \kappa_{+\mu\beta\sigma i} &= [\kappa + \beta_0 + (\beta_1 - 1)x_i]/\sigma_e. \\ \kappa_{-\mu\beta\sigma i} &= [\kappa - \beta_0 - (\beta_1 - 1)x_i]/\sigma_e. \\ C_0 &= \sum[\phi(-\kappa_{+\mu\beta\sigma i}) - \phi(\kappa_{-\mu\beta\sigma i})]. \\ C_1 &= \sum[\phi(-\kappa_{+\mu\beta\sigma i}) - \phi(\kappa_{-\mu\beta\sigma i})]x_i. \\ C_2 &= \sum[\kappa_{+\mu\beta\sigma i}\phi(-\kappa_{+\mu\beta\sigma i}) + \kappa_{-\mu\beta\sigma i}\phi(\kappa_{-\mu\beta\sigma i})].\end{aligned}$$

Proportional Error Case

When Y and X are positively valued variables and the errors are proportional to X , we assume that $\ln(Y)$ and $\ln(X)$ have a bivariate normal distribution. Let $100\theta\%$ be the percent change in Y and X , then:

$$\begin{aligned}\pi_\theta &= P[1/(1 + \theta) < Y/X < (1 + \theta)] \\ &= P[|\ln(Y) - \ln(X)| < \ln(1 + \theta)]\end{aligned}$$

Let $D = \ln(Y) - \ln(X)$ and $\kappa = \ln(1 + \theta)$ then $\theta = 100[\exp(\kappa) - 1]\%$. This $100\theta\%$ is called the $\text{TDI}_\pi\%$. In the case of proportional errors, the CCC, precision, and accuracy should be computed from log-transformed data. The same is applicable when target values are fixed.

Simulation Results

The parameters discussed above can be replaced with their sample estimates that are consistent estimators, usually the moments estimates, in order to perform inferences. A comprehensive simulation study was conducted^[7] to study the small sample properties of CCC, precision, accuracy, total deviation index (TDI), and CP. The results showed excellent agreement with theoretical values from normal samples even when $n = 15$. However, these estimates are not expected to be robust against outliers and/or large deviations from normality (or log normal). The robustness issues of the CCC have been addressed in King and Chinchilli.^[11]

Asymptotic Power

In assessing agreement, the null and the alternative hypotheses should be reversed. The conventional rejection region actually is the region of declaring agreement (one-sided). Asymptotic power and sample size calculation should proceed by the above principle. These powers of CCC, TDI, and CP were compared.⁷ The results showed that the TDI and CP estimates have similar power, and are superior to CCC. Therefore for inference, TDI and CP are superior to CCC. The CCC, precision, and accuracy remain as very useful descriptive tools.

EXAMPLES

Methods Comparison Example (Random Target Values, Constant Error)

The DCLHb is a treatment solution containing oxygen-carrying hemoglobin. The DCLHb level in a patient's serum is routinely measured by the Sigma method. The HemoCue method was modified to reproduce the DCLHb values of the Sigma method. Serum samples from 299 patients over a 50–2000 mg/dL range were collected. The DCLHb values of each sample were measured by both methods. The client wanted to be able to state with 95% confidence that the within-sample total deviation would be less than 15% of the total deviation. This means that the allowable CCC is $1 - 0.15^2 = 0.9775$. The client also wanted to state that 90% of the HemoCue observations would be within 150 mg/dL from the targeted Sigma values. This means that the allowable $\text{TDI}_{0.9}$ is 150 mg/dL, or that the allowable CP_{150} is 0.9.

The results are presented in Fig. 5. The plot indicates that the within-sample error is relatively constant across the clinical range. The plot also indicates that the HemoCue accuracy is excellent and that the precision is adequate.

The CCC estimate is 0.987, which means that the within-sample total deviation is about 11.6% of the total

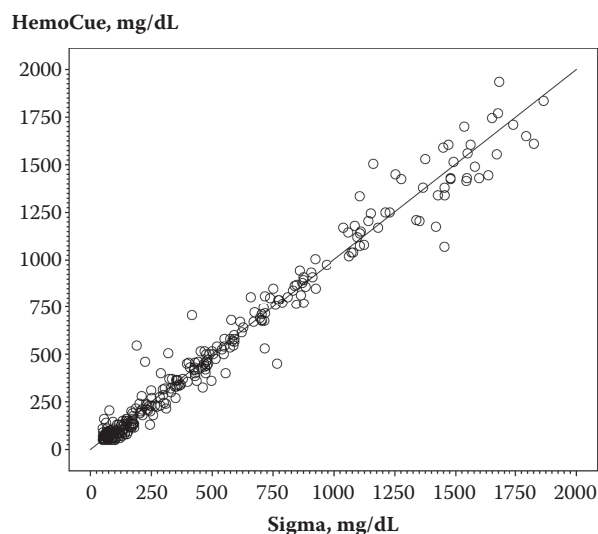


Fig. 5 HemoCue (vertical) vs. Sigma (horizontal) [mg/dL]. $CC = 0.987$ ($0.984 > 0.9775$); $r^2 = 0.974$ (0.968); $c_a = 0.9999$ (0.9989); $TDI_{0.9} = 127.5$ ($136.4 < 150$); $CP_{150} = 0.946$ ($0.928 > 0.9$)

deviation. The CCC one-sided lower confidence limit is 0.984, which is greater than 0.9775. The precision squared estimate is 0.974 with a one-sided lower confidence limit of 0.968. The accuracy estimate is 0.9999 with a one-sided lower confidence limit of 0.9989. The $TDI_{0.9}$ estimate is 127.5 mg/dL, which means that 90% of HemoCue observations were within 127.5 mg/dL of their target values. The one-sided upper confidence limit for $TDI_{0.9}$ is 136.4 mg/dL, which is less than 150 mg/dL. Finally, the CP_{150} estimate is 0.946, which means that 94.6% of HemoCue observations are within 150 mg/dL of their target values. The one-sided lower confidence limit for CP_{150} is 0.928, which is greater than 0.9. The agreement between HemoCue and Sigma is acceptable with excellent accuracy and adequate precision. The relative bias squared is estimated to be 0.003, so the approximation of TDI should be excellent.

Assay Validation Example (Fixed Target Values, Proportional Error)

FVIII is a clotting agent in plasma. The FVIII assay uses a marker with varying dilutions of known FVIII activities to form a standard curve. The standard curve started at 1:5 or 1:10, and serial dilutions were prepared until they reached the target values. Target values were 3%, 8%, 38%, 91%, and 108%. Six samples were assayed per target value. The error was expected to be proportional due largely to dilutions. The client wanted to be able to state with 95% confidence that the within-sample total deviation would be less than 15% of the total deviation. This means that the allowable CCC is $1 - 0.15^2 = 0.9775$. The client also wanted to be able to state that 80% of FVIII observations (%) would be within 50% from target values percent (percentage of percentage). This means that the allowable $TDI_{0.8}$ is 50%, or that the allowable $CP_{50\%}$ is 0.8.

Fig. 6 presents the results started at 1:5 and at 1:10 for the plots of observed FVIII assay results vs. targeted values in $\log_2$ scale. Note that in Fig. 6a, four 3% and two 2% were observed at the target value of 3%, and circles at the target value of 8% represent duplicate readings of 8%, 9%, and 10%. Duplicate readings of 45% were observed at target values of 38%. Also note that in Fig. 6b, four 5% and two 4% were observed at the target value of 3%. Three 11% and two 12% were observed at the target value of 8%. Duplicate readings of 49% were observed at target values of 38%. Duplicate readings of 124% were observed at target values of 91%. The plots indicate that the within-sample error was relatively constant across the target values in log scale. The precision was good for both assays started at 1:5 and at 1:10, but the accuracy was not as good for the assay started at 1:10.

For the results started at 1:5, the CCC was estimated to be 0.992, which means that the within-sample total deviation is about 9.1% of the total deviation. The one-sided lower confidence limit was 0.987, which was greater than 0.9775. The precision squared was estimated to be 0.988 with a one-sided lower confidence limit of 0.982. The accuracy was estimated to be 0.998 with a one-sided lower confidence limit of 0.994. The $TDI_{0.8}$ was estimated to be 27.3%, which means that 80% of observations were within 27.3% change from target values (percentage of percentage values). The one-sided upper confidence limit was 35.0%, which is less than 50%. Finally, the $CP_{50\%}$ was estimated to be 0.965, which means that 96.5% of observations were within 50% change from target values. The one-sided lower confidence limit was 0.892, which was greater than 0.8. The agreement between the FVIII assay and the actual concentration was acceptable with good precision and accuracy. The relative bias squared was estimated to be 0.009, so that the approximation of TDI should be excellent.

For the results started at 1:10, the CCC was estimated to be 0.967, which means that the within-sample total deviation is about 18.2% of the total deviation. The one-sided lower confidence limit was 0.958, which is less than 0.9775. The precision squared was estimated to be 0.989 with a one-sided lower confidence limit of 0.983. The accuracy was estimated to be 0.972 with a one-sided lower confidence limit of 0.964. The $TDI_{0.8}$ was estimated to be 58.9%, which means that 80% of observations were within 58.9% change from target percentage values. The one-sided upper confidence limit was 69.0%, which is greater than 50%. Finally, the $CP_{50\%}$ was estimated to be 0.702, which means that 70.2% of observations were within 50% change from target values. The one-sided lower confidence limit was 0.590, which was less than 0.8. The agreement between the FVIII assay and the actual concentration had good precision but was not acceptable because of mediocre accuracy. The relative bias squared was estimated to be 3.747, so that the approximation of TDI should be acceptable.

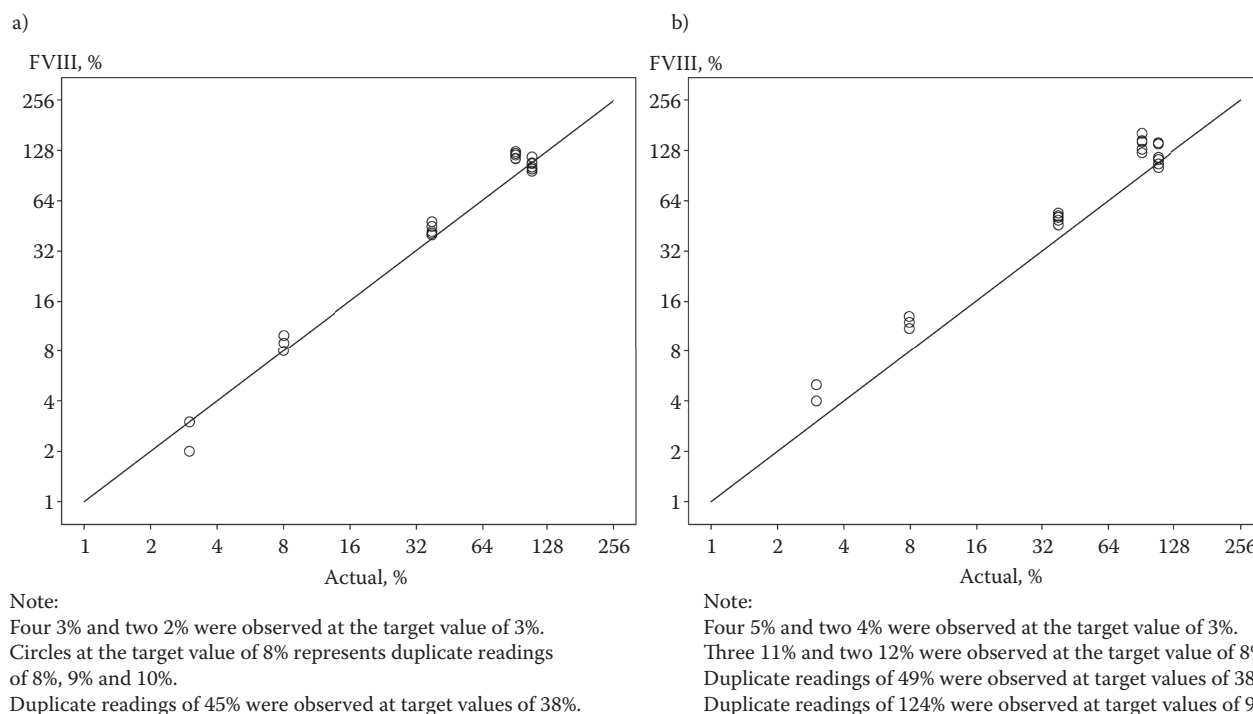


Fig. 6 FVIII assay (vertical) vs. actual values (horizontal) [%] (in $\log_2$ scale). (a) Started at 1:5: CCC = 0.992 (0.987 > 0.9775); $r^2 = 0.988$ (0.982); $c_a = 0.998$ (0.994); $\text{TDI}_{0.8}\%$ = 27.3% (35.0% < 50%); $\text{CP}_{50\%} = 0.965$ (0.892 > 0.8). (b) Started at 1:10: CCC = 0.967 (0.958 < 0.9775); $r^2 = 0.989$ (0.983); $c_a = 0.972$ (0.964); $\text{TDI}_{0.8}\%$ = 58.9% (69.0% > 50%); $\text{CP}_{50\%} = 0.702$ (0.590 < 0.8)

CONCLUSION

Under normality assumptions, the CCC, TDI (MSD), and CP basically measure the same information from different perspectives. The CCC, precision, accuracy, and ICC largely depend on the study range (intersample variation). Comparisons among any of these are valid only with similar study ranges. When the study range is fixed and meaningful, the CCC, precision, and accuracy offer meaningful geometric interpretations. The TDI and CP offer the most intuitively clear interpretation and have better power for inference, but they do not have precision and accuracy components.

For computing the above agreement statistics and generating an agreement plot, a SAS macro is available (<http://www.uic.edu/~hedayat/>), which can be downloaded by users.

RECENT DEVELOPMENTS AND FUTURE STUDIES

Chinchilli et al.^[12] addressed the topic regarding repeated measures CCC. Vonesh et al.^[13] and Vonesh and Chinchilli^[14] addressed the topic regarding the goodness of fit using CCC. King and Chinchilli^[11] addressed the CCC among more than two methods and robust CCC. Barnhart and Williamson^[15] used a very general approach to model

the CCC using generalized estimation equations (GEE). Our ultimate goal is to work on robust CCC, accuracy, precision, CP, and TDI via GEE. This will then be applicable to more general cases involving repeated measures of Y and X with covariates. Therefore it will be applicable to the individual bioequivalence problem (target values are always random), where repeated measures of test and reference drugs from a patient with covariates are taken.

REFERENCES

1. Lin, L.I. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **1989**, *45*, 255–268.
2. Fisher, R.A. *Statistical Methods for Research Workers*; Oliver & Boyd: Edinburgh, 1925.
3. Lin, L.I. Assay validation using the concordance correlation coefficient. *Biometrics* **1992**, *48*, 599–604.
4. Dunnett, C.W.; Gent, M. Significance testing to establish equivalence between treatments, with special reference to data in the form of 2×2 tables. *Biometrics* **1977**, *33*, 593–602.
5. Bross, I.D. Reader viewpoint: Why proof of safety is much more difficult than proof of hazard. *Biometrics* **1985**, *41*, 785–793.
6. Rodary, C.; Com-Noughe, C.; Tournade, M.F. How to establish equivalence between treatments: A one-sided clinical trial in pediatric oncology. *Stat. Med.* **1989**, *8*, 593–598.

7. Lin, L.I.; Hedayat, A.S.; Sinha, B.; Yang, M. Statistical methods in assessing agreement: Models, issues, and tools. *J. Am. Stat. Assoc.* **2002**, *97*, 257–270, 457.
8. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46.
9. Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213–220.
10. Lin, L.I. Total deviation index for measuring individual agreement: With application in lab performance and bioequivalence. *Stat. Med.* **2000**, *19-2*, 255–270.
11. King, T.S.; Chinchilli, V.M. A generalized concordance correlation coefficient for continuous and categorical data. *Stat. Med.* **2001**, *20*, 2131–2147.
12. Chinchilli, V.M.; Martel, J.K.; Kumanyka, S.; Lloyd, T. A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics* **1996**, *52*, 341–353.
13. Vonesh, E.F.; Chinchilli, V.M.; Pu, K. Goodness-of-fit in generalized nonlinear mixed-effect models. *Biometrics* **1996**, *52*, 572–587.
14. Vonesh, E.F.; Chinchilli, V.M. *Linear and Nonlinear Models for the Analysis of Repeated Measurements*; Marcel Dekker, Inc.: New York, 1997.
15. Barnhart, H.X.; Williamson, J.M. Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics* **2001**, *57*, 931–940.

MedDRA: Impact on Pharmaceutical Development

Keya T. Pitts

Bristol-Myers Squibb Company, Princeton, New Jersey, U.S.A.

Abstract

The Medical Dictionary for Regulatory Activities (MedDRA) was selected in 1999 by the International Conference on Harmonization (ICH) as the international standard for the encoding of medical terms captured in both the pre- and post-approval phases of drug development. This entry discusses the creation of the MedDRA as well as its reception and use in the pharmaceutical field.

Keywords: Adverse events; Clinical trials; Coding MedDRA; Postmarketing; Versioning.

INTRODUCTION

The Medical Dictionary for Regulatory Activities (MedDRA)^a was selected in 1999 by the International Conference on Harmonization (ICH) as the international standard for the encoding of medical terms captured in both the pre- and post-approval phases of drug development. This new coding terminology was not like the coding dictionaries of the past; with this new dictionary came a previously unobserved set of both advantages and challenges. The MedDRA Maintenance and Support Services Organization (MSSO), tendered by the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA), is charged with the administration of the terminology, and provision of guidance to its subscribers in its appropriate and intended application to patient data for the purposes of standardized regulatory reporting to health authorities.

Current State of the MedDRA Structure

Since its formal commercial introduction in 1999, MedDRA remains granular, containing more than 85,000 terms across the encoding levels (Lowest Level Terms (LLTs) and Preferred Terms (PTs)). The term level used for encoding may vary with the sponsor or manufacturer's coding philosophy. The Lowest Level Term consists of synonyms, culturally unique terms, and lexical variants of the corresponding Preferred Term. The Preferred Term is the level at which MedDRA establishes individual unique medical concepts and specific medical conditions. The High Level Term (HLT) is somewhat broader, and usually groups terms related to a general condition. Above the HLT is the High Level Group Term (HLGT), which further collectively

presents associated concepts into broadly presented groups. Finally, at the peak of the hierarchy is the System Organ Class (SOC). The scope of the SOC in MedDRA have not essentially changed since its commercial release, remaining 26 in the number of groupings, which has been important in maintaining consistency over time for sponsors that have utilized or referenced MedDRA in adverse event presentation within package labeling.

Although the number of term selections available at the Lowest Level and Preferred Term levels has grown considerably, the uppermost levels have been somewhat more stable. These intermediate levels may also be used by sponsors in meaningful data presentation and analyses, although the Preferred Term may be the more common level used in presentation of adverse event concepts in the labeling.

Some of the issues that were observed with the use of past terminologies continue to exist with this one dictionary, but now involve the use of different versions, as MedDRA is released twice-yearly (at the time of the writing this article). Using different terminologies at various stages in a product's life complicates data retrieval and analysis, making it difficult to cross-reference data. In the past, safety data had frequently been classified for pre-registration clinical trials using ICD terminology and for post-marketing surveillance using J-ART, WHO-ART, or COSTART. Furthermore, using different terminologies in separate geographic regions impaired international communication and necessitated the conversion of data from one terminology to another. This data conversion had the potential to cause time delays, and loss or distortion of data. In particular, these problems affected multinational pharmaceutical companies whose subsidiaries used multiple terminologies to fulfill the different local data submission requirements of regulators. The use of multiple terminologies also affected communication between companies and clinical research organizations.^[1]

MedDRA terms each have a unique numeric identifier; however, the field is non-expressive, meaning that no

^aMedDRA® is a registered trademark of the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA).

information can be further derived from it. Because the terms at the preferred level are repeated again at the lowest level for data entry purposes, these terms will have the same code or identifier for each level, thus identifying them as preferred terms, but providing no information such as related body system(s) or other groupings. Use of the numeric code became increasingly important as sponsors had to manage the process of versioning in the creation of the Integrated Summary of Safety for marketing authorizations. Versioning approaches will be discussed in the next section.

MedDRA-Encoded Data and the Concept of Versioning

As companies began to utilize MedDRA in the conduct of clinical trials to support Investigational New Drug (IND) or Clinical Trial Applications (CTAs), and New Drug Applications (NDAs) or Marketing Authorizations (MAs) for their products, a learning curve had to be mastered. With MedDRA commercially released twice per year, the issue of versioning in clinical trials had to be addressed. The MedDRA Maintenance and Support Services Organization (MSSO), responsible for maintaining the terminology for its subscribers, published recommendations for industry on how to manage versioning for clinical trials and the investigational phases of product development.^[2] These approaches served to provide companies some basic guidance in determining which versioning strategy best suited their processes, while regulators sought to also navigate the same waters and figure out the regulatory perspectives on clinical data presentation with MedDRA.

The six recommendations comprised every possible approach that could reasonably be applied for clinical trial data, given that trials are conducted over time and many different dictionary versions would be impacted.

The MSSO's **Option 1**—to select one version of MedDRA at the beginning of a project and continue to utilize for the life of the project through application submission—represents a seemingly uncomplicated approach to versioning. Benefits include ease in coding review across studies, report generation, and ease in development of data management, coding and statistical plans across studies. Some of the challenges with applying this strategy include not gaining the advantages of the evolving terminology and structural changes as well as data reconciliation between the clinical and safety databases (when companies opt to have their post-marketing pharmacovigilance units manage serious adverse event reporting to health authorities).

Option 2 presented a slight variant to option 1, the difference mainly being that users would begin with one version and use throughout, but once all submission studies have been completed in that version, all data would be upversioned to the version of MedDRA current at that point in time, to analyze and report. This brings a current perspective to the data, and will benefit companies with

detection of any potential issues that may have obscured through earlier version encoding. However, this does also potentially lengthen the time to submission because the upversioning work needs to be accounted for at the end of the process but prior to analyses, and addressing the changes from individual clinical study reports to the Integrated Safety Summary (ISS) data presentation.

Option 3 was another variant, allowing sponsors to begin each study with the version of MedDRA current at that point in time, and again upversion all study data at the time prior to summary analysis. This option would slowly introduce any changes in the observed safety profile through changes in coding from study to study. As the dictionary itself improves over time, these advantages would now be made available to users for more precise review. When the time comes to upversion all the data for the ISS, there is a small decrease in time needed, and there should be fewer unforeseen issues, previously undetected through coding.

The next option is not known by the author to be widely implemented—**Option 4** describes an approach that does not require the sponsor to encode individual clinical trial data as it is received but rather to hold all coding to be performed at the end of a study. Gains would be made in coding consistency with having it done at one time. Although this may seem like an approach to minimize the impact of versioning, it loads the work to be done at the back end of a study, and would not support accurate signal detection and safety risk management during the course of study conduct.

Option 5 is also a variant of Option 1, in that sponsors may initiate a study with a given version of MedDRA and *optionally* recode in a more recent version of MedDRA at the end of the study, but they must definitely upversion all data at the time prior to summary analysis. This would take advantage of the use of supplemental terms, or terms that may have been requested by subscribers for an upcoming release but are not yet available at the time of coding. Companies may choose to record the approved supplemental term in their database, but until the next version is released, the term technically is not associated with a version of MedDRA; or, they can enter a best-available choice from the version they are using and in either situation, exercise the option to recode at the end of the study before generation of the final clinical study report. Longer studies would gain from this approach and provide more beneficial analyses of data.

The final option is **Option 6**, which involves versioning of the trial data within a program with each commercial release of MedDRA. All trial data would be updated with each version and would provide consistency for reporting. Using this approach removes reconciliation issues with post-market safety data since post-market data will usually tend to be captured, coded, and reported with each new current version of MedDRA. This strategy does require some additional resourcing with respect to coding activity and more frequent maintenance of coding guidelines and

conventions, but provides a consistently current perspective of the data.

Pharmacovigilance functions in the post-marketing arena were predominantly first to utilize MedDRA for data capture and reporting expedited or periodic reports. Within industry and regulatory health authorities, post-marketing drug safety and pharmacovigilance were heavily represented on the international working groups that championed its evolution and implementation. With the regards to the versioning approach for safety and pharmacovigilance, those regulators that require reporting in MedDRA facilitated the versioning option for implementation with mandates of timeframes and version requirements in certain regions. As of the writing of this article, the European Commission and the Ministry of Health, Labour and Welfare (MHLW) for Japan both require reporting of spontaneous adverse event reporting to be done in MedDRA.^[3,4] The U.S. Food and Drug Administration (FDA) recommended MedDRA's use via draft and final guidances in both post-marketing reporting and NDA submissions; MedDRA was indicated in a Proposed Rule in 2003, but at this time, it is not mandated.^[5] The regulators mentioned have converted their respective databases to MedDRA and do have varying strategies and approaches to version control.

Aggregate Data Analysis and Presentation Using MedDRA

As the industry began to determine strategies for coding and versioning with MedDRA, the issue of data presentation also had to be addressed. MedDRA, being a much granular dictionary, required sponsors to really understand the terminology's structure and how that would impact data analysis and presentation. The quality of the "Preferred Term" from legacy terminology and previously used in analyses and listings would change significantly with the use of MedDRA. The MedDRA Preferred Term reflects a unique medical concept, and with over 18,400 preferred terms in version 12.0, the listings of coded data presented at the same level in MedDRA and, for a single study, will increase dramatically. As the System Organ Class (SOC) level is most like a "Body System" used in legacy dictionaries, and the MedDRA Preferred Term is the level of the medical concept, they currently remain the typical levels reflected within summary data tabulations and package labeling. The next sections will address how data presentation has been formally addressed by international working groups, industry, and the MedDRA maintenance organization.

Investigational Data Presentation

Requirements for data presentation for products under investigation (prior to market authorization will vary based upon the regional regulations. There will be requirements

for which cases will be presented (e.g., serious and associated adverse events), and grouping by SOC with the PTs observed has essentially continued to remain the industry standard, as recommended by internationally recognized pharmaceutical working groups.^[6]

Where there are many terms to be listed, inclusion of the intermediary MedDRA hierarchy grouping terms may prove clinically beneficial. As the MedDRA Preferred Term has been somewhat diluted compared to the "Preferred Term" of legacy terminologies, the High Level Term or High Level Group Term may aid to articulate the adverse event observations from a study as more meaningful medical concepts.

Presentation of the MedDRA concepts at any level of the hierarchy would not normally impact any of the typical summary tabulations that are generated, with the exception of a summary tabulation by age group. For pediatric studies, perhaps done as a post-marketing commitment that may be required in some regions, studies may comprise pediatric subjects across age groups. In this situation, utilizing the intermediate levels for data presentation as well as the Preferred Term level, which include concepts like HLGT Neonatal respiratory disorders and HLT Neonatal gastrointestinal disorders, will group terms specific to the neonatal period of life, away from concepts, to enable identification of such events. Of course, proper encoding of the reported adverse event term is required to ensure that it is grouped with its accurate grouping term. This provides additional benefit when the original reported information or demographic variables are not included in the specific presentation.

MedDRA aligns a medical concept under its most relevant SOC, but secondarily links concepts to other SOC's that are also related. Thus, industry initially struggled with how to apply the secondary SOC's in analysis and presentation for premarketing applications. The MedDRA MSSO subsequently developed a white paper addressing the issue, in the MedDRA Data Retrieval and Presentation: Points to Consider Guidance has been provided that recommends sponsors present the data by primary SOC as a standard in both marketing authorizations and the proposed package labeling for consistency across drug classes; then, if desired or requested, data may be displayed in focused searches or analyses utilizing secondary term assignments.^[7]

In the realm of product labeling, MedDRA is only known to be mandated for use in the presentation (i.e., labeling) of adverse drug reactions/events by the European Commission. The intermediate levels are recommended for inclusion in the Summary of Product Characteristics, if they further assist to present adverse drug reactions in a more meaningful way to the prescribing professional and patient.

Post-marketing Data Presentation

The use of MedDRA in post-marketing data presentation has been addressed within guidance documents proposed and finalized by the International Conference on Harmonization

(ICH). Although at the time of finalization of the E2C guidance on Periodic Safety Update Reporting, MedDRA had not yet been agreed, but a standardized coding terminology was recommended by the ICH for the presentation of adverse drug reaction data within a product's Core Safety Information section of the Company Core Data Sheet (CCDS), and for data presentation in the line listings within the Periodic Safety Update Report (PSUR).^[8] MedDRA was, of course, later agreed in the ICH E2B R3 Electronic Transmission of Individual Case Safety Reports guidance.^[9] The PSUR requires data to be presented in increments of time; and as MedDRA is upversioned with each new release twice a year, stating the version number is essential for regulators to fully evaluate the data in its context.

Within the PSUR, tabulations of adverse events/adverse drug reactions are requested to be presented by the body system using the standardized coded term rather than the verbatim. Given that product labeling may reflect intermediate level MedDRA terms, the periodic data may be further summarized in the PSUR using the same concepts for consistency but always reflecting the version to help interpret, as the regulators who may receive the report will be working with varying MedDRA versions.

Primarily in support of pharmacovigilance, the original concept behind Special Search Categories (SSCs) in the use of MedDRA was to aid MedDRA subscribers in reviewing collection of MedDRA-encoded data for clusters of specific syndrome terms across system organ classes that may not be readily observed with manual review of the data. As the SSCs initially developed were determined to not be of much benefit to industry and regulators alike, the Council for International Organizations of Medical Sciences (CIOMS) formed a working group comprised of regulators, prominent industry safety leaders, representatives of leading health organizations, such as the World Health Organization (WHO), and the MedDRA MSSO. The group's purpose was to develop rational use of the MedDRA Terminology for Drug Safety Database Searches and was established in May 2002.^[10] The group formulated what are now known as Standard MedDRA Queries (SMQs) that can be utilized by regulatory authorities, scientific institutions, and pharmaceutical companies worldwide to enable consistent and appropriate use of SMQs in drug surveillance.

As of MedDRA v12.0, there were 74 SMQs in production use that replaced the SSCs. Many others continue under development by the MedDRA MSSO in conjunction with the CIOMS SMQ working group, with recommendations always made possible through subscriber requests. Just as with dictionary content, the SMQs are maintained with respect to currency, relevance, and clinical accuracy.

SMQs are designed to follow an algorithm that, when applied, will look at adverse event data within a case and determine if the report's encoded terms coincide with the criteria within the SMQ to reflect the SMQ's topic, and cases may be weighted or ranked, according to how the

subscriber applies the SMQ and other criteria that may also be relevant. SMQs collate related Preferred Terms across the terminology that may be indicative of a sign and/or symptom of a specific condition or syndrome. The SMQ will include conditions, laboratory test results, and surgical procedures, where relevant. The concepts currently reflected as SMQ topics are of greater relevance to safety issues in pharmaceutical development.

Companies may opt to apply a relevant SMQ in response to a regulator's inquiry on a specific topic or issue, which enables reproducibility of results for both parties, with consideration given to the version and stability of the underlying data.

As subscribers become more comfortable with the application of the SMQ for post-marketing pharmacovigilance, the benefit of their application to assist with detection of potential safety issues, prior to a marketing application's submission, will begin to emerge.

Data Collection Principles Affecting Analyses

Be it within clinical development or the post-marketing environment, data collection techniques and guidelines will affect how the data is interpreted in coding and later analyzed. As was mentioned earlier, not only does demographic information at times affect how data is encoded, grouped, and analyzed, but as well, details that can be collected about the adverse event information will affect it. The route of administration for the investigational or marketed product can pose a factor in adverse event capture and collection. Administration site reactions are grouped separately from the body system events, in the General and Administration Site Conditions SOC (Fig. 1); however, they will have a secondary link, through MedDRA's multi-axiality, to the underlying body system (Fig. 2).

These types of characteristics should be taken into consideration in order to obtain more concise data interpretation for a clinical trial or post-marketing data. If data

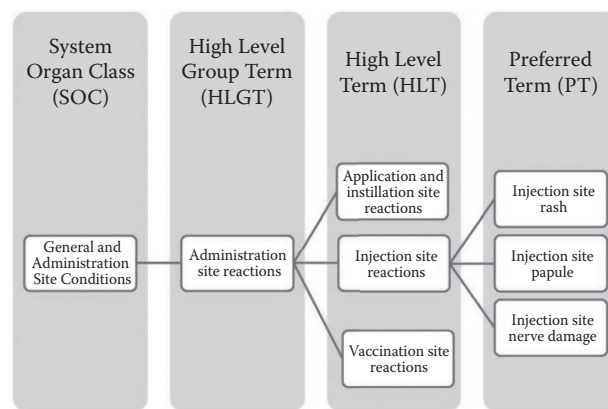


Fig. 1 Display of primary MedDRA hierarchy for administration site concepts (v12.0). Terms shown do not represent the full set of terms assigned to the levels displayed

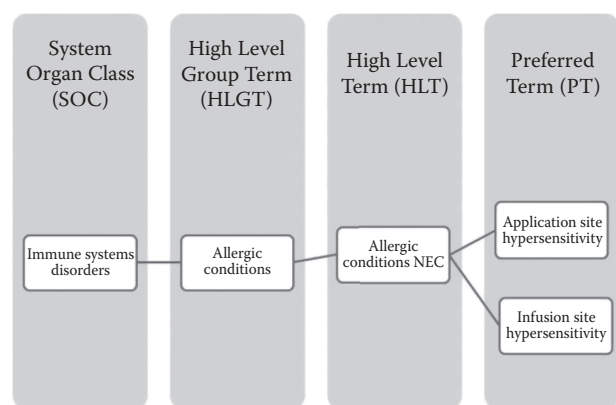


Fig. 2 Display of secondary MedDRA hierarchy for site reaction concepts. PTs Application site hypersensitivity and Infusion site hypersensitivity are linked secondarily to the Immune SOC

collected clearly denotes if an event occurred at the site of administration for the product, it will be more accurately encoded under the proper PT, such as those reflected in Fig. 2. Without this information collected from the subject or patient, it may have been presumed to have a different etiology or causality. For example, with a product that is administered via injection, capturing an event as “injection site-related” when appropriate is ultimately important for accurate display of observed adverse events within the prescribing information and medication guide.

CONCLUSION

As the Medical Dictionary for Regulatory Activities has been in commercial use by industry and regulators alike for eight years now, the benefits and challenges of adapting to MedDRA have surfaced. The MedDRA MSSO, in conjunction with its user groups and industry expert panels, has worked continuously to address those challenges, such as versioning, utility, and validity of MedDRA’s multi-axiality in data analyses.

The issue of versioning proved to be one that affected the pre- and post-marketing arenas in vastly different ways. With divisions of health authorities attaining differing levels of MedDRA readiness, given it was driven predominantly by the post-approval division of the major authorities. Post-marketing data was quickly required to be submitted encoded in MedDRA in some regions, and subsequently requirements were set for the timely use of MedDRA versions for some level of consistency with the regulators receiving the data. However, not all regulators in all regions set requirements for use of MedDRA, let alone a specific version with which data should be reported.

With the Proposed Rule on post-marketing safety data issued by the FDA in 2003, it was evident that the US was planning to require the use of MedDRA, but as part of a

much larger set of proposed regulations, the use of MedDRA still has yet to be mandated in U.S. regulatory reporting.

In the environment of pre-approval data, the European Commission has required the use of MedDRA in submissions, but the United States has not yet finalized a regulation on a position on MedDRA use in investigational and marketing applications and product labeling.^[11]

Subscribers have the MedDRA MSSO and industry expert working groups to look to for guidance in the management of MedDRA with respect to not only coding, but versioning, data analysis, and presentation. It is helpful to know that the health authorities also faced the same issues, and thoughtfully collaborated with industry to develop feasible and effective approaches to these issues. The key overarching benefit to the use of MedDRA include standardization of the patient safety data collected between pharmacovigilance and development functions, from company to company in safety data exchange, and among regulators and industry. To fully realize this benefit, standardized approaches to how versioning of MedDRA should be managed should be consistent across the regions for reporting both investigational and post-marketing data; as well, industry should aim to continue to partner with the regulators in evaluating the most effective and accurate standards for analysis and presentation of MedDRA-encoded data, for the ultimate benefit of delivering safe yet effective medication options to the patient.

REFERENCES

1. International Federation of Pharmaceutical Manufacturers’ and Associations (IFPMA). MedDRA Term Selection: Points to Consider Release 3.12 (Based on MedDRA Version 12.0). 2009: Geneva, Switzerland. Available on MedDRA MSSO website: <http://www.meddramsso.com/MSSOWeb/activities/PTC.htm>.
2. MedDRA MSSO. Recommendations for MedDRA® Implementation and Versioning for Clinical Trials (White Paper). 2001. Available on MedDRA MSSO website: <http://meddramsso.com/MSSOWeb/Docs/clinicaltrialversioning.pdf>.
3. European Commission Enterprise and Industry Directorate. Volume 9A of the Rules Governing Medicinal Products in the European Union – Guidelines on Pharmacovigilance for Medicinal Products for Human Use. Final January 2007: Available on European Commission website: http://ec.europa.eu/enterprise/pharmaceuticals/eudralex/vol-9/pdf/vol9a_09-2008.pdf.
4. MedDRA MSSO unofficial translation of Yakushokuanhatsu Notification No. 0325001 and Yakushokushinsatsu Notification No. 0325032. MHLW Notification on MedDRA Usage. 2004 March 25. Available on MedDRA website: http://www.meddramsso.com/MSSOWeb/document_library/MHLW_Notification_on_MedDRA_20040325.pdf.
5. Department of Health and Human Services – Food and Drug Administration. 21 CFR Parts 310, 312, et al.,

- Safety Reporting Requirements for Human Drug and Biological Products. – Proposed Rule. 14 March 2003. Available on FDA website: <http://www.fda.gov>.
6. Council for International Organizations of Medical Sciences (CIOMS). Development Safety Update Report (DSUR): Harmonizing the Format and Content for Periodic Safety Reporting During Clinical Trials Report of CIOMS Working Group VII. Geneva, Switzerland; CIOMS: 2006. Available on CIOMS website: <http://www.cioms.ch/>.
 7. International Federation of Pharmaceutical Manufacturers' and Associations (IFPMA). MedDRA® Data Retrieval and Presentation: Points to Consider Release 2.0 based on MedDRA Version 12.0. 1 April 2009; p. 18. MedDRA MSSO website: http://meddramssso.com/MSSOWeb/Document_Library/9610-1200_DatRetPTC_R20_mar2009.pdf.
 8. ICH Steering Committee. ICH Harmonised Tripartite Guideline Clinical Safety Data Management: Periodic Safety Update Reports For Marketed Drugs E2C (R1). 06 November 1996. Available on ICH website: <http://www.ich.org/cache/compo/276-254-1.html>.
 9. ICH Steering Committee. Revision of The ICH Guideline On Clinical Safety Data Management Data Elements For Transmission of Individual Case Safety Reports E2B(R3), Current Step 2 version. 12 May 2005; pg 2. Available on ICH website: http://www.ich.org/cache/compo/MediaServer.jsr?@_ID=632&@_TYPE=MULTIMEDIA&@_TEMPLATE=616&@_MODE=GLB.
 10. Council for International Organizations of Medical Sciences (CIOMS). Development and Rational Use of Standardised Med-DRA Queries (SMQs). Geneva, Switzerland: CIOMS; 2004. Available on CIOMS website: <http://www.cioms.ch/>.
 11. European Commission Enterprise and Industry Directorate. General Detailed guidance on the collection, verification and presentation of adverse reaction reports arising from clinical trials on medicinal products for human use – Revision 2. April 2006. Available on European Commission website: <http://ec.europa.eu/>.

Medical Devices

Greg Campbell, Lilly Q. Yue, Gene Pennello, and Mary K. Barrick

CDRH, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

Medical devices comprise a staggering array of products that include everything from diagnostic tests and monitors, to therapeutic machines and instruments and replacements for failing body parts. This entry focuses on statistical issues in medical devices, touching on methodological discussions for implants, statistical challenges of diagnostic devices, and the crucial role of prior information in the design of medical devices.

Keywords: Diagnostic devices; Masking; Medical devices; Placebo; PMA; Surrogates.

INTRODUCTION

Medical devices comprise a staggering array of products that include everything from diagnostic tests and monitors, to therapeutic machines and instruments and replacements for failing body parts. They include artificial skin, hips, knees, and breast implants; apnea and cardiac monitors; glucose and cholesterol tests; magnetic resonance imaging machines and ultrasound bone sonometers; intraocular lenses and ophthalmic lasers; manual and robotic surgical instruments; and gloves. They also include prostate-screening antigen (PSA) tests for prostate cancer detection and monitoring, cardiac defibrillators, left ventricular assist devices, and the coronary stents that are revolutionizing interventional cardiology.

Within the Food and Drug Administration (FDA), the Center for Devices and Radiological Health (CDRH) is charged with the regulation of medical devices in the United States. For regulatory purposes, a *device* is defined by law as “an instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article, including any component, part, accessory, which is (i) recognized in the Official National Formulary, or the United States Pharmacopoeia, or any supplement to them, (ii) intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease, in man or other animals, or (iii) intended to affect the structure, or any function of the body of man or other animals, and which does not achieve its primary intended purposes through chemical action within or on the body of man or other animal and which is not dependent upon being metabolized for the achievement of its primary intended purposes.”^[1] A more practical and somewhat simpler definition for many people is one of exclusions, namely, any medical item that is *not* a drug and *not* a biological product.

This paper will discuss the statistical issues in medical devices in four subsequent sections. The section “Regulation of Medical Devices” concerns the regulation of medical devices. The section “Design Issues in Non-diagnostic Medical Device Studies” concerns non-diagnostic devices and includes methodological discussions for implants (knee, hip, breast prosthesis) as well as therapeutic devices such as external defibrillators or lasers for eye surgery. The section “Diagnostic Devices” will treat the unique statistical challenges that diagnostic devices pose. Diagnostic medical devices can be used in screening, diagnosing, or monitoring of patients; such products include genetic markers for cancer, PSA tests, MRI machines, bone densitometers, mammography machines, glucose meters for home testing as well as the wide variety of clinical laboratory tests such as ones for creatine kinase, chlamydia, and pregnancy. The section “Bayesian Approach in Evaluating Medical Devices” will address the crucial role that prior information can play in the design and analysis of medical device clinical trials. The section “Statistical Issues in the Postmarket” and the “Conclusion” section will address postmarketing issues and the future of medical device evaluation, respectively.

REGULATION OF MEDICAL DEVICES

Regulatory History in the United States

In terms of regulatory history in the United States, 1976 marked the beginning of medical device regulation in its own right. The 1976 Medical Device Amendments to the Food, Drug, and Cosmetic (FD&C) Act gave the FDA explicit authority to evaluate safety and effectiveness of medical devices. These amendments also recognized the special nature of medical devices and defined specific

provisions including the Premarket Approval (PMA) and the Premarket Notification pathways to market. To carry out these provisions, FDA created the Bureau of Medical Devices. In 1982, this fledgling bureau merged with the Bureau of Radiological Health to form CDRH, and in 1990, 1992, and 1997, Congress passed further amendments to refine the regulation of medical devices as well as other FDA-regulated products. Headquartered in Silver Spring, MD, CDRH^[2] now has a staff of about 1200. Statistical programs in the Center are organizationally located in the Division of Biostatistics, which has a staff of 55.

Regulation of Medical Devices vs. Pharmaceutical Drugs in the United States

In comparison to pharmaceutical drugs, there are some similarities and differences in how medical devices are regulated. There are often preclinical studies, sometimes using animal models, for medical devices, but the issue is usually not toxicity as in pharmaceuticals. In many instances, the mechanism of action of even a new device technology is well understood, physical and local as opposed to chemical and systemic. Most device product lines evolve substantially through a series of changes. Whereas changes to a drug formulation can drastically change the safety and efficacy profiles, small changes to a device may have little or no effect on either safety or effectiveness. Unlike the pharmaceutical industry, which consists of relatively few very large companies, the medical device industry has a large number of companies (over 10,000). While a few of these are quite large, the median size of medical device firms is under 50 people. The regulation of medical devices in the United States reflects the above differences. Whereas the standard for pharmaceutical drug regulation is two well-controlled Phase III trials that demonstrate safety and efficacy, the standard for medical devices is different depending on the regulatory pathways described below, but is usually not more than one confirmatory, well-controlled, multicenter study.^[3–5]

Premarket Approval Applications

The PMA process is the most rigorous regulatory pathway to market a new device technology. By law, the decision to approve or disapprove a PMA application must “rely upon valid scientific evidence to determine whether there is reasonable assurance that the device is safe and effective.” Again, by law, “Valid scientific evidence is evidence from well-controlled studies, partially controlled studies and objective trials without matched controls, well-documented case histories conducted by qualified experts that there is a reasonable assurance of safety and effectiveness ...”⁶ These criteria differ from the regulatory standards set by law for pharmaceutical drugs and biological products in the United States and these differences lead to a more varied approach in the nature of studies and data

required to support market approval for devices. As part of the PMA process, the FDA seeks advice from outside experts who serve on 18 Advisory Panels of the Medical Devices Advisory Committee. For more information about the Advisory Committee structure for FDA, see Ref. [7].

Premarket Notifications (510(k))

A pathway to market for medical devices that has no direct analog in drug regulation is Premarket Notification, often referred to as the 510(k) process (from FD&C provision). In general terms, this process allows manufacturers to notify the FDA that they intend to market a product as a “me too” device. The manufacturer must establish that the new device is “substantially equivalent” to a predicate device, i.e., a device marketed at the time of the 1976 Medical Device Amendments. If the FDA agrees that the device is substantially equivalent, the new device is then cleared for marketing in the United States.

Investigational Device Exemption

Usually, to demonstrate that a new device is safe and effective and, sometimes, to show that it is substantially equivalent to another device, the new device must be studied in humans. To protect patients participating in studies where there is significant risk, a company or individual researcher must apply for an Investigational Device Exemption (IDE) prior to beginning the study. FDA then provides statistical, clinical, and engineering reviews of the preclinical data and investigational plan and weighs the potential benefits against the potential risks to determine if the participation of patients in the clinical study is justified.

DIAGNOSTIC DEVICE REGULATION

The FDA regulates diagnostic tests in several ways. One way has been through labeling; for microbial sensitivity diagnostics, this began in 1967 and has expanded to in vitro diagnostic (IVD) tests. As a result of the 1976 Medical Device amendments to the Food, Drug and Cosmetic Act, the same two mechanisms of pre-market regulation, the PreMarket Notification and the PreMarket Approval application, apply to diagnostic as well as therapeutic devices. In addition to the diagnostic device manufacturers which are regulated by the FDA, many in vitro diagnostic tests require laboratories which analyze the specimens and report the results. Laboratories are regulated in the United States by several congressional actions. In 1967, the Clinical Laboratory Improvement Act (CLIA) which regulated what are called high-complexity laboratories was enacted into law and in 1988 the CLIA Amendments expanded the scope from high-complexity laboratories to all laboratories, including physician’s offices. Many diagnostic tests may no pose any significant risk to the patient, in which

case an IDE would not be required; however, CDRH encourages companies to meet with the FDA before any clinical study is undertaken for regulatory consideration. For a discussion of some of the regulatory issues in genetic and genomic tests, see Ref. [8].

GLOBAL REGULATORY EFFORTS

Medical devices are regulated throughout the world. Governmental organizations that utilize statistical review in their evaluation of medical devices include the Medicines and Healthcare Products Regulatory Agency (MHRA) of the United Kingdom,^[9] Japan's Pharmaceuticals and Medical Devices Agency (PMDA),^[10] and Canada's Bureau of Medical Devices in Health Canada.^[11] Efforts are underway to harmonize medical device regulations among many countries throughout the world through Global Harmonization Task Force,^[12] an organization similar to the International Conference on Harmonization (ICH) for pharmaceutical products.

Design Issues in Non-Diagnostic Medical Device Studies

There are a number of important design issues that arise in the planning of medical device studies. A general FDA guidance document that describes many of these issues is "Statistical Guidance for Clinical Trials of Non-Diagnostic Medical Devices."^[13] For sample-size estimation, see the encyclopedia entry on "Sample Size Determination" because the sample size issues involved in medical device studies are similar to those in drug development.

Implants

While medical devices come in a mind-boggling array of shapes, sizes, and intended uses as well as attendant problems in evaluation, devices that are implanted provide special statistical challenges. For example, implants are usually meant to be in the body for a very long time, unlike a drug, which is usually out of the body shortly after the ingestion. Thus, how the implant interacts with the body, and the body with the implant over time, is important to evaluate. This poses some unique statistical challenges, especially regarding survival analysis,^[14] repeated measures (longitudinal data analysis), and missing data. In addition, implants, by definition, are *implanted*, usually in a surgical procedure; the safety and effectiveness of the implant may depend on the skill of the surgeon. Thus, the variability of patient outcome across study centers (or surgeons) may be a critical issue. There may even be (and often is) a *learning curve* where a surgeon becomes more skilled and results improve as a surgeon gains experience with the implant and implantation procedure. On the other hand, sometimes, devices can be predictable. Most medical

devices work through a localized, physical mechanism of action. Thus, bench testing and information from similar products may be useful in predicting how a new device will perform. Because medical devices do not fit nicely into a neat category where a single analytical approach is optimal, or even possible, the study design and analyses employed will vary greatly depending on the particular device and indication involved.

Sham Controls (The Placebo Effect)

One critical element in study design that represents a major departure from the usual pharmaceutical study lies in the choice of an appropriate control. The well-studied and highly successful placebo-control trial design is used less frequently in medical device evaluation. For many devices, there is no placebo arm possible, while for others, a placebo (or sham) would be unethical. For these reasons, approaches employing active or historical controls are much more common although this can introduce the knotty question of statistical bias. The placebo effect can be quite substantial in many studies where the primary outcome variable is pain or function. Kaptchuk considers whether there is, in some cases, an enhanced placebo effect for devices; if there were such an added effect for a new device, it would make its effectiveness more difficult to demonstrate.^[15]

While an active control may appear to be the more appropriate alternative, this type of control can bring major analytical challenges to the device arena. Consider, for example, an implanted nerve stimulator intended for weight reduction. In this case, the most appropriate control can be a drug intended for the same purpose. However, both the drug and the device probably have a placebo effect and the extent of the placebo effects may easily differ.^[16] This placebo-by-treatment interaction may be difficult or, at worst, impossible to estimate at the analysis stage.

Masking (Blinding)

Often in a medical device trial, it may not be possible to mask the health-care provider or the patient to the treatment because it may be quite apparent to all who has which device. It is nonetheless desirable that the third-party evaluator be masked if possible at all. Masking (or blinding) can avoid statistical bias.

Historical Controls

In some cases, a randomized trial with an active control may not even be practical or feasible. For example, in a rapidly advancing technology, it may be difficult and, sometimes, unethical to recruit patients for the control arm. In such cases, a single-arm study is often conducted, with the experimental device being compared to one or more historical controls from previous studies. A problem with this comparison is that the device effect is confounded by the

effect of the type of study as well as the device-by-study interactions. These confounding effects can result from differences between the studies in objectives, endpoints studied, protocol, patient population, investigational sites, physician training, and patient management. In fact, any variable confounds the comparison if it has an effect on the outcome of interest and the variable's distribution differs between the historical studies and the current study. Estimation of the treatment effect in a study with an historical control is considered in Zhang.^[17]

Sometimes, the historical control data are summarized by a single-target value. The comparison of the device to the target value not only can be confounded but also ignores the random variation inherent in the historical study used to create the target value. Bayesian hierarchical modeling could be used to address confounding and historical control random variation in the development of target values for medical device trials.^[18]

Noninferiority Trials

When establishing substantial equivalence, superiority is usually not the issue, and a superiority trial design usually does not make sense. For devices where there are alternative devices already available, knowing that a new device is no worse than a competitor may demonstrate that the new device is substantially equivalent to its comparator, or that it is safe and effective. For these reasons, the noninferiority trial design is an increasingly popular approach in evaluating medical devices. The fundamental principles of such noninferiority trials and statistical methodologies applied in the area of therapeutic devices are the same as in the area of drug development.^[19–24] However, some statistical issues and design questions have been encountered in this area more frequently than others.^[24]

In statistical terms, the null and the alternative hypotheses with respect to population mean in a noninferiority trial are specified as $H_0: \mu_c - \mu_e \geq \delta$ vs. $H_a: \mu_c - \mu_e < \delta$ where $\delta > 0$ is the prespecified noninferiority margin, μ_e and μ_c denote experimental and active control treatment means, respectively, and it is assumed that larger means are more desirable. While choosing the correct formulation of study hypothesis might not be an issue in most drug development applications, in some medical device submissions, the misleading hypotheses of $H_0: \mu_c - \mu_e = 0$ vs. $H_a: \mu_c - \mu_e \neq 0$ are still employed. As Blackwelder has pointed out,^[19] in determining whether a new device is as effective as a standard device, the test of the conventional null hypothesis of equal effects is inappropriate and leads to logical difficulties. In particular, failure to reject the alternative hypothesis does not establish noninferiority.^[25]

Choice of Active Control

An active control could be an FDA-approved device or an accepted standard of care. To demonstrate the effectiveness

of a new device, a noninferiority study should have the ability to distinguish effective from ineffective products. One crucial requirement is the implication of the active control treatment effect; that is, the active control selected should be known to be effective in the planned study because equivalence/noninferiority to an ineffective comparator is meaningless.

Choice of Noninferiority Margin

Another crucial consideration concerns the choice of an acceptable noninferiority margin, δ . This quantity should be sufficiently small so that from a clinical point of view, the new device can be considered to be as good as the control (and, therefore, effective) as long as the new device is no more than δ worse than the active control. In particular, the choice of δ should not lead to a situation where the new device is only slightly better or even worse than no treatment.

When planning a study and sizing it, the specification of the population parameters, μ_e and μ_c , has to be clinically meaningful and reasonable. The treatment effects for the new and active control are frequently set at the same level, meaning the new device is expected to perform as well as the control although somewhat lesser performance may be tolerable. However, the possibility that the active control is more effective than the new device might be more realistic in some noninferiority medical device trials. For example, in a comparison of invasive and noninvasive devices, if it is known that the noninvasive device has somewhat inferior effectiveness performance compared to the invasive active control but has other significant benefits, it may be more realistic to allow that the active control is slightly better.

As a matter of fact, a superiority study and a noninferiority study can be practically planned at the same time, but it is essential that the planning be done in advance of carrying out the study.^[23]

Surrogates

Surrogates can play an extremely important role in the evaluation of medical device effectiveness. In the case of implants, even a premarket study of, say, 3 years may not predict the long-term behavior of the implant. Unlike drugs, where most of the drug is cleared from the system in a moderate amount of time after use is discontinued, an implant is there until it is surgically removed (if that is even possible). Thus, the long-term effects of the device are of interest. The ability of a surrogate to provide insight as the performance of the primary endpoint can streamline the introduction of the medical device into the marketplace. However, it is not sufficient to validate any surrogate merely to show strong correlation between it and the primary endpoint; general conditions for valid surrogates are put forth by Prentice.^[26] DeMets^[27] considers the difficult issue of surrogate endpoints in medical device trials.

Monitoring

There are important statistical issues that arise in the monitoring of medical device studies. This includes patient accountability (missing data), modifications in the protocol or in the device, sample-size re-estimation, and sequential monitoring. Another issue that arises in device trials is called crossover, not the statistical design but the notion of transferring patients, usually for ethical reasons, from one therapy to another during the study. For implants, this can create the inability to do long-term follow-up on both arms.

Analysis

Many of the same issues arise in the analysis of medical device studies as are encountered in the analysis of pharmaceutical trials. A few of the notable statistical issues that may be somewhat different for device studies follow.

Center-by-Treatment Interaction

In device trials, center-to-center variation in the device effect can be large. The main reason is sometimes large variation between centers in physician experience and training in using or implanting the device. Other reasons include variability between centers in patient population, patient management, and reporting practices. When center-by-device interactions are large or are qualitative (i.e., the device effect changes sign depending on the center), then a scientifically credible explanation is required at the minimum. Sometimes, close scrutiny of the centers can reveal explanations for the interactions.

Repeated Measurements

Repeated measurements are generated when a study subject is observed on several successive follow-up times, occasions, or experimental conditions. The use of repeated measurements in medical device studies has become increasingly popular recently. There are generally two types of frequently used repeated measurement design in therapeutic device trials: single-arm studies (studies without a concurrent control group) and two-arm, randomized, possibly masked, controlled trials. In a single-arm study, the hypothesis of interest is usually to compare pre-treatment and post-treatment effects or to evaluate the effect of device over follow-up times. In a two-arm study, there are generally three types of hypothesis of interest: treatment-by-time interaction, treatment effect, and time trend. The equivalence of diagnostic assay methods is often evaluated with matched-paired repeated measurements through regression model, repeatability, and reproducibility studies. There are mixed types of data seen in medical device clinical studies including continuous, binary, and ordinal data. Appropriate statistical analysis incorporating

correlation structure among repeated measures within a subject and important covariates is needed.^[28–31]

Subgroup Analysis

Planned subgroup analysis can test for significant effects within subgroups. For example, for spinal fusion devices, the device effect could be significant for obtaining fusion at one level of the spine, but not for obtaining fusion at two contiguous levels. When the device effect is only significant for some of the subgroups, then the inferences should be adjusted for multiple comparisons.^[32]

Causal Inference

In many medical device studies, the actual device effects may be distorted because of statistical biases introduced by the design and the conduct of the study. For example, bias is introduced by lack of masking of treatments to patients and investigators, nonrandomized comparisons with historical controls, noncompliance, missing data, censoring of quality-of-life variables due to death, etc. When a study has a known or suspected bias, causal inference theory can be used to tease out the actual or causal treatment effect from the bias. As a result, the theory can provide a framework for investigating biases inherent to studies. For example, Frangakis and Rubin^[33] employ the causal inference method of principle stratification to obtain estimates of causal effects in studies complicated by post-treatment variables such as noncompliance. Other causal inference methods include propensity scores and sensitivity analysis as discussed below. Other causal inference topics are counterfactuals, randomization theory, structural equation models, path analysis, and directed acyclic graphs. The causal inference literature is growing and may prove to be very helpful in evaluating medical devices.^[34]

Propensity Scores

Propensity scores are used to adjust nonrandomized comparisons for covariates. For medical devices, they can be useful when an experimental device is compared with a historical control using patient-level information in the two data sets. In this context, the propensity score $e(X)$ for a subject with a vector X of observed covariates is the conditional probability of receiving treatment ($Z = 1$) rather than control ($Z = 0$) given X ; i.e., $e(X) = \Pr(Z = 1 | X)$.^[35–37] The propensity score $e(X)$ is a balancing score in the sense that it is a function of the observed covariates X such that the conditional distribution of X given $e(X)$ is the same for subjects in the treatment group and subjects in the control group (1). If treatment assignment is strongly ignorable, that is, the treatment assignment Z and the potential outcome Y are conditionally independent given the observed covariates, X , i.e., $\Pr(Z | X, Y) = \Pr(Z | X)$, and if $0 < \Pr(e(X)) < 1$, for all X , then the observed treatment effect at

each value of the propensity score is an unbiased estimate of the true treatment effect at that propensity score.^[35] Propensity score matching, stratification, weighting, covariate adjustment, and some combination of these strategies can then be used to produce an unbiased estimate of the average treatment effect (Ref. [35–40] and references therein). In application, the propensity score is estimated by modeling the probability of treatment group membership as a function of the observed baseline covariates via, for example, a multiple logistic regression. The assumption of strongly ignorable treatment assignment requires that all covariates relevant to both treatment assignment and potential outcomes are measured (i.e., there is no hidden bias).

As pointed out by Rubin,^[37] as a balancing score, the propensity score is a one-dimensional summary of the observed covariates such that when the propensity scores are balanced across two treatment groups, the distribution of all the covariates are balanced in expectation across the two groups. Therefore, evaluating treatment group overlap in terms of the distributions of propensity scores provides a straightforward assessment of whether the two treatment groups are sufficiently similar with respect to baseline covariates to allow a sensible treatment comparison. When sufficient overlap is present, the propensity score analysis can provide a straightforward treatment comparison that reflects adjustment for differences in all observed baseline covariates. Otherwise, as indicated by Rubin,^[37] either the propensity score prediction model would need to be reformulated or it would have to be concluded that the covariate distributions did not overlap sufficiently to allow propensity score analysis to adjust for these covariates. Therefore, in a regulatory environment, it is crucial to start with comparable patient populations, since propensity score analysis does not work well (in fact, no statistical method works well) when there is serious imbalance in baseline covariates between the two treatment groups.^[41,42]

Like other covariate adjustment methods, propensity score methods can only adjust for observed covariates and not for unobserved ones. This is always a limitation of non-randomized studies compared with randomized trials where randomization tends to balance both observed and unobserved covariates. Therefore, some sensitivity analysis for hidden bias is often desirable in a medical product regulatory submission based on a non-randomized study. It should also be noted that propensity score methods may not eliminate all the bias resulting from imbalance in the observed covariates, due to the limitation of propensity score modeling, which typically uses a linear combination of covariates.^[37] Braitman and Rosenbaum^[43] state that the propensity score methods work better under three conditions: first, when clinical outcome event is rare, second, when there are a large number of patients in each treatment group, and third, when there are many covariates measured.

Rubin^[44] emphasizes the importance of designing of observational studies prospectively, and advocates

that observational studies can and should be designed to approximate a randomized experiment as closely as possible before any access to any outcome data, thereby assuring the objectivity of the design. Particularly in a regulatory environment, the prospective design of a non-randomized confirmatory study with propensity score analysis specified as primary analysis is essential, in order to provide valid and convincing evidence of safety and effectiveness of a medical product.^[41,42]

Sensitivity Analysis

Sensitivity analyses quantify how sensitive inferences are to statistical biases that might be present in studies. For medical device studies under regulatory review, such analyses can be used to demonstrate that inferences are robust to potential sources of bias such as missing data, placebo-by-treatment interactions, and unmeasured variables that confound comparisons of devices with historical controls. For example, a significant device effect may still be regarded as significant in the presence of an unmeasured confounder if it is large enough. However, the device effect has to be larger than if that bias did not exist. Thus, there is a trade-off between the clinical burden of a well-designed and well-conducted study (e.g., a randomized controlled trial with no missing data) and the statistical burden of less than ideal studies (e.g., a one-armed study comparing an experimental device to a historical control) that introduce statistical bias. An example of a missing data sensitivity analysis for a medical device trial was presented at a recent Orthopedic and Rehabilitation Devices Advisory Panel; the interested reader is referred to the Summary of Safety and Effectiveness for this spinal fixation product.^[45] Sensitivity analysis methodology for an unmeasured confounding variable is given in Rosenbaum^[46] and in Lin et al.^[47]

DIAGNOSTIC DEVICES

Diagnostic Design Issues

Diagnostic devices differ from therapeutic devices in a fundamental way. Rather than treating a patient with a specific condition, diagnostic devices are intended to determine whether the patient has the specific condition. Thus, diagnostic studies pose different statistical challenges in terms of both the design and the analysis. The design of most of these studies involves using the diagnostic device to test patients with and without the condition of interest and determining how often the diagnostic test is correct. While this design may seem simple in concept, it can be very difficult to actually carry out. One major problem is in identifying the patients who truly have the condition. Often, this is done with another diagnostic test. When that test is extremely accurate so that a *truth standard* (sometimes called a “gold standard”) exists, then defining the

accuracy of the new test is relatively straightforward. When the comparison test is less than ideal, the study becomes a comparison of two tests, where, even when they agree, neither may be correct.^[48] These issues are approached from a regulatory perspective with the FDA Guidance document “Statistical Guidance on Reporting the Results from Studies Evaluating Diagnostic Tests.”^[49]

The Randomized Clinical Trial, a design that is so often relied upon in therapeutic trials, is of limited use in many diagnostic evaluations. It is generally much more efficient to use each patient as his/her own control in the diagnostic setting. It is often simple to subject each patient to both diagnostic procedures (although whether the order of the tests is important may depend on the situation); further, if the two tests are correlated, this design can be extremely efficient. However, bias can become an overwhelming issue. In particular, the performance of a diagnostic test can be greatly affected by the selection of the patients. The order in which the two tests are conducted can sometimes also produce bias. Verification bias, the bias introduced when the gold standard is missing for some patients, is also present in many diagnostic studies. For an example of verification bias and a proposed remedy, see Kondratovich.^[50] Thus there has to be exceptional vigilance to ensure that all the different sorts of statistical biases do not threaten the scientific validity of the clinical diagnostic study. Begg^[51] and Pepe^[52] survey many of the biases that can plague diagnostic studies.

Diagnostic test results have little clinical value if they do not confer any additional diagnostic information beyond information already available from standard assessments. For example, a prostate specific antigen (PSA) test should add diagnostic value over the digital rectal exam in detecting prostate cancer. As another example, bone density, as measured by a bone densitometer, should add value as a risk factor for bone fracture after considering a patient's age.

In terms of reporting the performance of a new diagnostic medical device, sensitivity and specificity are most helpful when using a truth standard (“gold standard”). Unfortunately, there is often no truth standard because even if there is some standard for disease classification, it is usually not perfect. For example, a standard may be accurate if the disease is detected, but less than perfect if the disease is not detected. Without a truth standard, it is impossible to report sensitivity and specificity in a direct manner. One approach that some have attempted is to compare the new device to another test (not a truth standard) and then subject the cases that are discrepant to another perfect or imperfect test. Such a practice inevitably results in biased estimates of sensitivity and specificity that may be very misleading.^[53] Recent research into treating the prevalence as an unknown parameter that can be estimated using the expectation-maximization (E-M) algorithm may hold some promise.^[54] Another approach called *method comparison*, which is discussed in the “Measurement

Method Comparison for Diagnostic Tests,” can be used to compare the new test with another diagnostic test. There are a number of interesting statistical problems associated with the demonstration of agreement.^[55] Regardless of the approach, however, without a truth standard, such a comparative agreement approach can only answer questions about equivalence, but not superiority.

Analysis Issues for Diagnostic Devices

The statistical analysis for diagnostic devices is generally quite different from that for therapeutic agents. When truth is known, statistical inference using confidence intervals rather than hypothesis testing is often much more useful in reporting the performance of a dichotomous test. When the threshold or cutoff is on a continuous or ordinal scale, it is particularly helpful to use receiver-operating characteristic (ROC) methodology.^[56] For studies with important covariates, logistic regression is a useful tool. For studies that rely on agreement, kappa and percent agreement positive and negative are helpful although the null value of kappa equal to zero is not of interest. In some instances, especially for diagnostic imaging, the skill of the image reader can be quite important. In such cases, it may be helpful to use variance components and bootstrapping to study the variability due to readers as well as due to cases and to design the study accordingly.^[57] For comparative studies, sample size can be difficult to determine because the correlation between two diagnostic tests can be unknown *a priori*, suggesting that sample size should be determined adaptively.^[58] An overview of statistical analysis for medical imaging is given in Chow and Shao.^[59]

Measurement Method Comparison for Diagnostic Tests

The purpose of a diagnostic device study with quantitative measurements is usually to assess whether the new device is comparable to an established one. Here comparable means equivalent, which can be defined in different ways including *average equivalence*, *population equivalence*, or *individual equivalence*.^[60–63] Evaluating average equivalence could be done using regression analysis and assessing the closeness of the slope to 1 and the intercept to 0. Key considerations in this type of analysis would include:

- Criteria for claiming equivalence
- Errors in measuring both X and Y
- Variance assumptions.

The statistical model could be specified as follows: Y indicates the measurement with the new device, with mean y and random error ε ; X represents the measurement using the established device, with mean x and random error η , with random errors ε and η independent. The unobserved

mean values x and y have a linear relationship with intercept β_0 and slope β_1 :

$$Y = y + \varepsilon, \quad X = x + \eta, \quad y = \beta_0 + \beta_1 x$$

For purposes of illustration, consider the criteria for regression slope equivalence as closeness of slope to 1 and closeness of intercept to 0. In assessing slope, a power approach has often been used, with the statistical null hypothesis of slope equal to 1 and the alternative hypothesis of slope not equal 1:

$$H'_0 : \beta_1 = 1$$

$$H'_a : \beta_1 \neq 1$$

Regression equivalence for the slope could be claimed if the null hypothesis were not rejected or, equivalently, if a 95% confidence interval on slope includes 1, provided there is sufficient power to detect an important difference. Usually, the sample size is calculated at a nominal α level of 0.05% and 80% statistical power to detect a slope that differs from the null slope of 1 by a clinically meaningful value of, say, 0.1. The logic underlying the power approach is that if the slope had actually been 1.1 (or larger), the probability is at least 80% that the null hypothesis of slope equal to 1 would have been rejected at level α . Thus, if the null hypothesis is not rejected, it is concluded that slope was not as large as 1.1; that is, the equivalence is achieved. However, as pointed out by Schuirmann,^[63] this test leads to logical difficulties. For example, insufficient evidence to reject the null hypothesis of equality does not imply sufficient evidence to accept it. If the slope is actually clinically different from 1, but the desired result is the failure to find it,^[1] that result is easier to obtain with a small study than with a large one. There is actually a penalty for conducting an efficient trial with a narrow confidence interval.

By contrast, an interval hypothesis procedure, commonly used in pharmaceutical bioequivalence studies, does not have these disadvantages. Consider the hypotheses

$$H_0 : \beta_1 \leq 1 - \delta \text{ or } \beta_1 \geq 1 + \delta$$

$$H_a : 1 - \delta < \beta_1 < 1 + \delta$$

where δ is a clinically meaningful equivalence margin. This procedure is a familiar one in assessing equivalence of average bioavailability/bioequivalence.^[64,65] Regression equivalence of the slope could be claimed if the null hypothesis is rejected or, equivalently, if a two-sided 90% confidence interval on slope lies entirely within the equivalence interval $(1 - \delta, 1 + \delta)$. Average equivalence would also require inference about the intercept as well; and individual equivalence additionally requires some assurance in

this slope-intercept regression approach that the variance about the fitted line is small.

In terms of analysis, errors in measurements and underlying variance assumptions are critical issues. Because the measurements obtained from both the new and the established devices are subject to random measurement error, using a simple model that ignores the errors in measurements from *both* devices will lead to bias in parameter estimates and inferences. Hence, more complex methodologies, such as measurement error models, e.g., Deming regression^[66] or Passing-Bablok regression^[67] should be used. Assumptions about variance are also critical in these assessments. When, as is often the case, the variance of these measurements quite commonly increases as the true reference value increases, this heteroscedasticity in variance must be taken into account, otherwise the subsequent test for measurement method equivalence will be invalid.

INDIVIDUAL EQUIVALENCE

To evaluate individual equivalence, one could, for example, characterize the distribution of the differences between paired measurements, Y and X , from the two devices across the measurement range, and evaluate whether $P(|Y - X| < d)$ is close to 1 for a given $d > 0$. A graphical presentation of individual agreement is the Bland-Altman scatter plot of the differences $(Y - X)$ against the mean of the measurements $((X + Y)/2)$ to enable assessment of the scatter for uniformity and symmetry about zero.^[68] The 95% limits of agreement, the average difference plus or minus two standard deviations, is also useful to compare against d .^[69] Another measure of individual agreement is the individual equivalence criterion (IEC)^[70] which compares the variability between modalities to the variability within a modality, which for 510(k)s is naturally the predicate device. Coefficients of accuracy and agreement have also been developed.^[71]

Diagnostic Imaging

Imaging systems are medical devices which are becoming an increasingly important in diagnostic medicine. Examples of diagnostic imaging include digital mammography, ultrasound, chest x-ray, computed tomography (CT) scans, and magnetic resonance imaging (MRI). Diagnostic imaging usually involves interpretation of the image by a reader, e.g., a radiologist reading a mammogram for suspicious lesions. Computer assisted diagnostic devices (CAD) provide information to aid the reader's interpretation of the image. CADs have been developed for breast, lung, and colon imaging to detect abnormalities, such as malignant lesions. The evaluation of imaging systems and CADs can pose unique design and analysis challenges, including estimating intra- and inter-reader variability.^[72,73]

Monitoring Devices

Many diagnostic devices monitor patients for conditions of interest. Examples include bone densitometers, fetal heart rate monitors, glucose monitors, and pulse oximeters. Sample size determination for studies of monitoring devices that involve repeated measurements is discussed in Pennello.^[74]

Analytical Studies

For *in vitro* diagnostic tests, analytical performance should be characterized. Analytical performance refers to accuracy of the test in detecting an analyte or quantifying its concentration. Analytical performance also can include bias relative a predicate or reference test, test precision (repeatability), analytical specificity, and when appropriate limits of detection or measurement. Determining the limit of detection is discussed in Linnet and Kondratovich.^[75]

Biomarkers, Diagnostic Tests, and Microarrays

Biomarkers play an important role in many clinical studies, whether they involve other medical devices or not. A large number of biomarkers are obtained from medical devices, such as blood pressure, medical imaging, and genetic and genomic tests including microarrays. A question in some cases is whether the biomarker can serve as a surrogate or a partial predictor of disease or of response to treatment. Biomarkers can help to predict outcome in the presence of the standard treatment or no treatment (prognostic biomarker) or to predict differential efficacy or safety associated with a particular treatment (predictive biomarker).^[76] Biomarkers may prove to be indispensable in selecting the right therapies for the right patients, in short, in personalizing medicine. Technologies such as microarrays, which measure gene expression for a large number of genetic segments, are fueling some of the excitement associated with the discovery of new biomarkers. For example, MammaPrint®, by Agendia, is an FDA-approved microarray-based prognostic biomarker for risk of distant metastasis of breast cancer.^[77] Microarray-based classifiers pose interesting challenges in development, validation, and FDA regulation^[78,79] and are a subset of In Vitro Multivariate Index Assays, the subject of FDA draft guidance.^[80,81]

Bayesian Approach in Evaluating Medical Devices

Bayesian statistics is playing an increasing role in pre-market medical device studies. In several well-planned successful submissions to the FDA's CDRH, a Bayesian approach has provided the primary basis for the statistical analysis. Bayesian statistical analyses take advantage of prior information on safety and effectiveness parameters by formally combining this information with clinical data via Bayes Theorem.

At the FDA's CDRH, the ability to take advantage of prior information has very attractive regulatory implications. In device development, improvements tend to be incremental; thus, there is usually an older version of the device, often with associated clinical testing. These clinical data are often relevant. For example, implantable and therapeutic devices tend to have a physical mechanism of action. Thus, rather than dealing with potential major systemic reactions, small changes in such a device tend to result in only a small change in a localized effect. Using Bayesian methods to take advantage of prior information often leads to estimates of effectiveness and safety parameters that are more precise than if the prior information were ignored. Therefore, by utilizing prior information, CDRH can, in principle, reach the same decision on a device with smaller and perhaps shorter trials.

When prior information is not available, the Bayesian approach can still be helpful. For instance, the Bayesian approach can be more flexible than the frequentist approach with regard to interim looks,^[82] missing data,^[83] multiplicity,^[82] and adaptive modifications of trials such as changes to the sample size and dropping of a treatment arm.^[84] In addition, Bayesian inferences provide exact posterior distributions of parameters, while many frequentist methods appeal to asymptotic theory, which may not be appropriate for small-sized studies or for rare events.

Practical applications of Bayesian analyses are discussed in the excellent texts by Spiegelhalter et al. and others^[85–88] A tutorial of Bayesian methods for medical device development is given in Berry.^[89] Pennello and Thompson^[90] discuss a variety of issues concerning Bayesian medical device trials under regulatory review.

Bayesian Statistics in Brief

Most of the familiar techniques in statistics are referred to as frequentist. The frequentist approach considers parameters as fixed unknown quantities. In contrast, the Bayesian approach considers parameters to be random quantities with a probability distribution. In the Bayesian approach, probabilities are assigned to parameter values prior to the data being observed. These probabilities are called the prior distribution. When study data are later observed, the prior distribution is mathematically updated to the posterior distribution, according to Bayes Theorem. The posterior distribution is then used to compute posterior probabilities of important hypotheses. If the posterior probability of a hypothesis is large enough, then the hypothesis is accepted. This approach provides a scientifically valid way of combining previous information with current data.^[91]

FDA's CDRH has released a draft guidance document on the use of Bayesian statistics in medical device clinical trials.^[92]

Valid Prior Information

Sources and validation of prior information for medical devices under regulatory review are discussed in Irony and Pennello^[93] and Pennello and Thompson.^[90] The prior information is usually based on data from other studies. These prior studies should be similar to the current study in protocol (endpoints, target population, etc.) and timeframe, e.g., to ensure the practice of medicine in the studies is similar. Covariates, such as demographics and prognostic variables, can be extremely helpful in calibrating the previous studies to the current study. Covariate adjustment is often necessary to ensure that the prior studies are exchangeable with the current study, an assumption commonly invoked in Bayesian hierarchical models for combining the studies. The previous studies should also be representative of the possible studies that could have been selected.

Bayesian Sample Size

The Bayesian approach does not require the sample size to be fixed at the study outset because inferences are based not on the sample space, as in the frequentist approach, but on the parameter space. That is, the posterior distribution represents probabilities for parameters, not data. In practice, however, a minimum sample size may be required to evaluate safety parameters (which are often rare adverse events), to verify model assumptions or to ensure that the prior does not overwhelm the clinical data. A maximum sample size may be needed for economical or ethical reasons or for planning the logistics of the trial. Bayesian approaches to sample-size determination include those based on power to make decisions^[94,95] interval length and coverage probability,^[96–98] and decision theory.^[99] Because the sample size is not fixed, it can, in principle, be revised at any point in the study by considering the current posterior distribution instead of the prior distribution.

Bayesian Interim Looks

Bayesian interim analyses do not depend on the stopping rule, provided it is non-informative. This is the Stopping Rule Principle, which is a consequence of the Likelihood Principle,^[100] adhered to in Bayesian analyses. Consequently, Bayesian interim analyses do not depend on the number of interim looks. Nevertheless, in the regulatory setting, the Type 1 error that a medical device is approved when it is not safe and effective is a major concern, and the Type 1 error rate is inflated if interim analyses are made. Therefore, the FDA's CDRH usually recommends that the sponsor computes the Type 1 error rate for the design they are considering.

One Bayesian interim monitoring strategy stops patient accrual when the posterior probability of the primary hypothesis of interest exceeds a pre-specified threshold, e.g., 0.975. However, if outcomes are not yet available for

patients just accrued to the trial, a better strategy is to monitor the predictive probability that when these outcomes are observed, the posterior probability threshold will be met. One could also monitor the predictive probability of success at the end of the study were it not stopped. Stopping criteria can also be based on formal decision analysis.^[101] For more information on Bayesian interim analysis for medical devices.^[90,102]

Bayesian Adaptive Design

Adaptive designs use accumulating data to decide on how to modify certain aspects of a trial according to a pre-specified plan. Because of the Likelihood principle, a Bayesian approach can offer flexibility in the design and analysis of adaptive trials. Some examples of adaptive designs for medical device trials are sample size modification, adaptive randomization,^[103] and the dropping or adding of an arm because of safety concerns or lack of enrollment. Another example is a hybrid trial, during which an inactive control, e.g., no device, is replaced with an active control, e.g., a recently approved new device, when investigational sites begin to use the active control in their practice, in anticipation that once the investigational sites begin using the recently approved device, randomization to the no device control arm would become unethical. In turn, the primary hypothesis changes from one of superiority to the inactive control to one of non-inferiority with the active control. A specific instance is hybrid trials for embolic protection devices, which collect potentially harmful emboli caused by placement of coronary stents.^[104]

Historical Controls as Prior Information

A frequent application of Bayesian statistics uses historical control data as prior information for the control arm in a randomized controlled study.^[105] This prior information translates to additional sample size for the randomized control;^[106] thus, fewer patients need to be randomized to it. Hierarchical modeling (see the section “Bayesian Hierarchical Modeling”) of the studies can be used to account for between-study and within-study variation of the controls. For a single-armed study without a randomized control, historical controls can still be used as prior information to derive a synthetic control arm.^[107]

Bayesian Noninferiority Testing

Some commonly used prior distributions are not appropriate for noninferiority trials. “Non-informative” priors center the experimental device effect at zero, implying a prior probability of noninferiority that is greater than 50%. If it were really the case that no prior information is available, then the prior probability should be exactly 50%. Similarly, hierarchical models (see the section “Bayesian Hierarchical Modeling”) on the experimental device and

control arms center the device effect at zero. Additionally, they use data on one arm as prior information for the other, which pulls the device effect toward zero, i.e., toward the region of noninferiority. This prior can be difficult to justify because it assumes what is sought to be proved by the study, namely, that the experimental device is similar (non-inferior) to the control.

Bayesian Hierarchical Modeling

Bayesian hierarchical modeling^[108] is a flexible approach for combining previous device studies with a current study. These models separate variation between studies from variation within studies by assuming that study-specific parameters are random. These random effects enable the current study to borrow strength from the previous studies, i.e., use them as prior information. This usually results in more precise estimates of the current study parameters; however, that is not always the case.^[109] Hierarchical models are flexible in that the current study borrows less strength from the previous studies as the variation between studies increases. Thus, hierarchical models protect against overreliance on bad prior information, which is an attractive property, especially in the regulatory setting.

Hierarchical models can also be used to combine data across centers in a multicenter device trial. Between-center variability, which can be large in device trials, is accounted for by assuming that center-specific device effects are random.

Hierarchical models can also be used to adjust subgroup analyses for multiplicity effects.^[110] Briefly, subgroup effects that are considered related are assumed random. Thus, the estimate of each subgroup effect borrows strength from related subgroups. A subgroup that has a large observed effect will have that effect adjusted downward if the related subgroups have mostly small observed effects.

Bayesian Calculations

Bayesian calculations require integration, usually in high dimensions. For example, the posterior probability of a hypothesis is expressed as an integral involving the posterior distribution. Until recently, the high dimensionality of these integrals made Bayesian calculations prohibitive. Advances in Markov Chain Monte Carlo (MCMC) sampling methods^[111] and computing speed have enabled calculation of these integrals and are the main reasons that the field of Bayesian statistics has become popular in recent years.

A popular MCMC method is Gibbs sampling. Gibbs sampling avoids the problem of sampling from the posterior distribution, which is often infeasible. The Gibbs sampling method creates a Markov Chain by sampling sequentially from the posterior distribution for each parameter conditional on the last sampled values of the other parameters and the data. These distributions are called full conditional

distributions. Often, the full conditionals can be written in closed form, making sampling from them straightforward. Under mild regularity conditions, the sampling distribution converges to the posterior distribution. Once the sampling distribution has converged, subsequent samples can be used to compute desired Bayesian integrals. Because the samples form a Markov Chain, they are dependent samples. Therefore, enough samples need to be taken such that the support of the posterior distribution has been completely explored. A Bayesian statistics program that uses Gibbs sampling is Bayesian inference Using Gibbs Sampling (BUGS).^[112]

STATISTICAL ISSUES IN THE POSTMARKET

Medical devices are also studied in the postmarket phase; the longer-term safety of new medical devices is often monitored after they have been approved for marketing. The ability to decide on the seriousness of a medical device adverse event depends on exposure. Information concerning the number of devices in the hands of the intended users is often not well known; hence, exposure may be most difficult to estimate. In addition, there is a well-recognized, serious underreporting of adverse events that may be associated with devices; this can create a severe bias. Furthermore, even if there is an adverse event reported in conjunction with the use of a medical device, it may be difficult to ascertain whether the device caused the adverse event. Different statistical methodologies have been proposed to address these difficult problems in postmarket surveillance of both medical devices^[113,114] and drugs,^[115] e.g., a Bayesian approach using a gamma-Poisson model for the latter.

CONCLUSION

The future of medical devices poses no end of exciting and challenging issues for the development of new statistical methodology. Indeed, a special issue in *J Biopharmaceutical Statistics* was recently devoted to medical devices.^[116] It is predicted that the pace of medical device revolution will continue to accelerate. The speed and the variety of the new technology will encourage new statistical methods to meet these challenges. The globalization of the medical device industry will pose challenges to the regulators as efforts are made to harmonize requirements for demonstrating safety and effectiveness.

REFERENCES

1. U.S. Code 21 (h) Food, Drug, and Cosmetic Act; <http://www.fda.gov/opacom/laws/fdcact/fdcact1.htm> (Accessed August, 2009).

2. Web address <http://www.fda.gov/MedicalDevices/default.htm> (accessed August, 2009).
3. Wittes, J. Clinical trials of the effectiveness of devices: An analogy with drugs. *Surgery* **2001**, *129* (5), 517–523.
4. Campbell, G. Statistical Issues in Medical Devices: A Regulatory Perspective. In *American Statistical Association 1996 Proceedings of the Biopharmaceutical Section*, Joint Statistical Meetings, American Statistical Association: Chicago, IL and Alexandria, VA, 1996, 224–229.
5. Campbell, G. Statistics in the world of medical devices: The contrast with pharmaceuticals. *J. Biopharm. Stat.* **2008**, *18*, 4–19.
6. U.S. Code of Federal Regulations, Title 21 (Food and Drugs), U.S. Government Printing Office, Washington, DC, 2001, Part 860.7. http://www.access.gpo.gov/nara/cfr/waisidx_01/21cfr860_01.html (accessed August, 2009).
7. Web address <http://www.fda.gov/AdvisoryCommittees/default.htm> (accessed August, 2009).
8. Campbell, G. Some issues in the statistical evaluation of genetic and genomic tests. *J. Biopharm. Stat.* **2004**, *14*, 539–552.
9. UK's Medicines and Healthcare Products Regulatory Agency (MHRA) web address <http://www.mhra.gov.uk/Howweregulate/Devices/index.htm> (accessed August 2009).
10. Japan's Pharmaceuticals and Medical Devices Agency (PMDA) web address <http://www.pmda.go.jp/english/about/index.html> (accessed August 2009).
11. Canada's Bureau of Medical Devices in Health Canada Web address <http://www.hc-sc.gc.ca/dhp-mps/md-im/index-eng.php> (accessed August 2009).
12. Global Harmonization Task Force web address <http://www.ghftf.org> (accessed August 2009).
13. U.S. Department of Health and Human Services, Food and Drug Administration, Division of Small Manufacturers. *Assistance Statistical Guidance for Clinical Trials of Non-Diagnostic Medical Devices*. U.S. Government Printing Office: Washington, DC 1996-32 pp. (Also available at <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm106757.htm>).
14. Lao, C.S. Statistical considerations for survival analysis from medical device clinical studies. *J. Biopharm. Stat.* **1995**, *5*, 159–170.
15. Kaptchuk, T.J.; Goldman, P.; Stone, D.A.; Stason, W.B. Do medical devices have enhanced placebo effects?. *J. Clin. Epidemiol.* **2000**, *53* (8), 786–792.
16. Kaptchuk, T.J.; Stason, W.B.; Davis, R.B.; Legedza, A.T.R.; Schnyer, R.N.; Kerr, C.E.; Stone, D.A.; Nam, B.H.; Kirsch, I.; Goldman, R.H. Sham device v inert pill: randomised controlled trial of two placebo treatments. *Br. Med. J.* **2006**, *332*, 391–397.
17. Zhang, Z. Estimating the current treatment effect with historical control data. *JP J. Biostat.* **2007**, *1*, 217–247.
18. O'Malley, A.; Normand, S.; Kuntz, R. Application of models for multivariate mixed outcomes to medical device trials: coronary artery stenting. *Stat. Med.* **2003**, *22*, 313–336.
19. Blackwelder, W.C. "Proving the null hypothesis" in clinical trials. *Control. Clin. Trials* **1982**, *3*, 345–353.
20. Farrington, C.P.; Manning, G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Stat. Med.* **1990**, *9*, 1447–1454.
21. Hwang, I.K.; Morikawa, T. Design issues in non-inferiority/equivalence trials. *Drug Inf. J.* **1999**, *33*, 1205–1218.
22. Simon, R. Bayesian design and analysis of active control clinical trials. *Biometrics* **1999**, *55*, 484–487.
23. Morikawa, T.; Yoshida, M. A useful testing strategy in Phase III trials: Combined test of superiority and test of equivalence. *J. Biopharm. Stat.* **1995**, *5*, 297–306.
24. Yue, L.Q. Design Issues in Non-inferiority Medical Device Clinical Trials. In *American Statistical Association 2001 Proceedings of the Biopharmaceutical Section*, Joint Statistical Meetings, American Statistical Association: Atlanta, GA, Alexandria, VA, 2001; M, #451.
25. Altman, D.G.; Bland, J.M. Absence of evidence is not evidence of absence. *Br. Med. J.* **1995**, *311*, 485.
26. Prentice, R.L. Surrogate endpoints in clinical trials: Definitions and operational criteria. *Stat. Med.* **1989**, *8*, 431–440.
27. DeMets, D.L. The role of surrogate outcome measures in evaluating medical devices. *Surgery* **2000**, *128* (3), 379–385.
28. Lao, C.S. Application and Analysis of Repeated-Measure Design in Medical Device Clinical Trials. *American Statistical Association 2001 Proceedings of the Biopharmaceutical Section*; Joint Statistical Meetings, American Statistical Association: Atlanta, GA and Alexandria, VA, 2001.
29. Lao, C.S. Equivalence in test assay method comparisons for the repeated-measure, matched-pair design in medical device studies: Statistical considerations. *J. Biopharm. Stat.* **2000**, *10* (3), 433–445.
30. Wang, Z.J. Design effects for highly clustered count data with varying length of follow-up – applying to the analysis of ICD shocks. *J. Biopharm. Stat.* **2008**, *18* (1), 31–43.
31. Lao, C.S.; Bushar, H.F. Longitudinal data analysis in medical device clinical studies. *J. Biopharm. Stat.* **2008**, *18* (1), 44–53.
32. Scott, P.E.; Campbell, G. Interpretation of subgroup analyses in medical device clinical trials. *Drug Inf. J.* **1997**, *32*, 213–220.
33. Frangakis, C.E.; Rubin, D.B. Principal stratification in causal inference. *Biometrics* **2002**, *58*, 21–29.
34. Van der Laan, M.; Robins, J. *Unified Methods for Censored Longitudinal Data and Causality*; Springer Verlag: New York, 2003.
35. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, *70*, 41–55.
36. Rosenbaum, P.R.; Rubin, D.B. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* **1984**, *79*, 516–524.
37. Rubin, D.B. Estimating casual effects from large data sets using propensity scores. *Ann. Intern. Med.* **1997**, *127*, 757–763.
38. D'Agostino, R.B., Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat. Med.* **1998**, *17*, 2265–2281.
39. Lunceford, J.K.; Davidian, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat. Med.* **2004**, *23*, 2937–2960.
40. Hill, J.; Reiter, J. Interval estimation for treatment effects using propensity score matching. *Stat. Med.* **2006**, *25*, 2230–2256.

41. Yue, L.Q. Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies. *J. Biopharm. Stat.* **2007**, *17*, 1–13.
42. Li, H.; Yue, L.Q. Statistical and regulatory issues in non-randomized medical device clinical studies. *J. Biopharm. Stat.* **2008**, *18*, 20–30.
43. Braitman, L.; Rosenbaum, P.R. Rare outcomes, common treatments: Analytic strategies using propensity scores. *Ann. Intern. Med.* **2002**, *137*, 693–696.
44. Rubin, D.B. The design versus the analysis of observational studies for causal effects: Parallel with the design of randomized trials. *Stat. Med.* **2007**, *26*, 20–36.
45. U.S. Department of Health and Human Services, Food and Drug Administration, CDRH Advisory Meeting Materials Archive, Orthopaedic and Rehabilitation Devices Panel, <http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfAdvisory/details.cfm?mtg=128> (accessed August, 2009).
46. Rosenbaum, P.R. *Observational Studies*. Springer-Verlag: New York, 1995.
47. Lin, D.Y.; Psaty, B.M.; Kronmal, R.A. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **1998**, *54*, 948–963.
48. Meier, K. Study Design Requirements for Evaluating the Performance of a Diagnostic Test. In *American Statistical Association 1998 Proceedings of the Biopharmaceutical Section*, Joint Statistical Meetings, American Statistical Association: Dallas, TX and Alexandria, VA, 1998, 84–87.
49. U.S. Department of Health and Human Services, Food and Drug Administration. *Statistical Guidance on Reporting the Results from Studies Evaluating Diagnostic Tests (released March 13, 2007)* <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071148.htm> (accessed August 2009).
50. Kondratovich, M.V. Comparing two medical tests when results of reference standard are unavailable for those negative via both tests, *J. Biopharm. Stat.* **2008**, *18* (1), 145–166.
51. Begg, C.G. Biases in the assessment of diagnostic tests. *Stat. Med.* **1987**, *6*, 411–423.
52. Pepe, M.S. *The Statistical Evaluation of Medical Tests for Classification and Prediction*; Oxford Press: New York, 2003.
53. Meier, K. Reporting results from studies evaluating diagnostic tests. *Clin. Microbiol. Newsl.* **2002**, *24*, 60–63.
54. Beiden, S.V.; Campbell, G.; Meier, K.L.; Wagner, R.F. On the Problem of ROC Analysis Without Truth: The E.M. Algorithm and the Information Matrix, Medical Imaging 2000: Image Perception and Performance. *Proc. SPIE* **2000**, *3981*, 126–134.
55. Weng, T.S. Evaluation of diagnostic tests: Measuring degree of agreement and beyond. *Drug Inf. J.* **2001**, *35* (2), 577–588.
56. Zweig, M.H.; Campbell, G. Receiver operating characteristic plots: A fundamental evaluation tool in clinical medicine. *Clin. Chem.* **1993**, *39*, 561–577.
57. Beiden, S.V.; Wagner, R.F.; Campbell, G. Components-of-variance models and multiple bootstrap experiments: An alternative methodology for random-effects, receiver operating characteristic analysis. *Acad. Radiol.* **2000**, *7*, 341–349.
58. Wu, C.; Liu, A.; Yu, K.F. An adaptive approach to designing comparative diagnostic accuracy studies. *J. Biopharm. Stat.* **2008**, *18* (1), 116–125.
59. Chow, S.C.; Shao, J. *Statistics in Drug Research: Methodologies and Recent Developments*; Marcel Dekker: New York, 2002.
60. Ponnappalli, R.M.; Vishnuvajjala, R.L.; Campbell, G. On Equivalence Trials with Medical Devices. In *American Statistical Association 1999 Proceedings of the Biopharmaceutical Section*, Joint Statistical Meetings, American Statistical Association: Baltimore, MD and Alexandria, VA, 1999, 224–226.
61. Liu, J.P. Equivalence Trials. *Encyclopedia of Biopharmaceutical Statistics*, 1st Ed.; 2000.
62. Chow, S.C.; Liu, J.P. Individual equivalence. *Encyclopedia of Biopharmaceutical Statistics*, 1st Ed.; 2000.
63. Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* **1987**, *15* (6), 657–680.
64. Tan, C.Y.; Iglewicz, B. Measurement—methods comparison and linear statistical relationship. *Technometrics* **1999**, *41* (3), 192–201.
65. Kondratovich, M.; Yue, L.Q.; Dawson, J.M. Power Approach Versus Two One-Sided Tests Procedure in Equivalence Evaluation with Diagnostic Medical Devices. In *American Statistical Association 2001 Proceedings of the Biopharmaceutical Section*, Joint Statistical Meetings, American Statistical Association: Atlanta, GA and Alexandria, VA, 2001.
66. Deming, W.E. *Statistical Adjustment of Data*; Wiley: New York, 1943.
67. Passing, H.; Bablock, W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. *J. Clin. Chem. Clin. Biochem.* **1983**, *21*, 709–720.
68. Bland, J.M.; Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1986**, *327* (8476), 307–310.
69. Bland, J.M.; Altman, D.G. Measuring agreement in method comparisons studies. *Stat. Methods Med. Res.* **1999**, *8*, 135–160.
70. Barnhart, H.X.; Kosinski, A.S.; Haber, M.J. Assessing individual agreement, *J. Biopharm. Stat.* **2007**, *17*, 697–719.
71. Lin, L. Overview of agreement statistics for medical devices. *J. Biopharm. Stat.* **2008**, *18*, 126–144.
72. Wagner, R.F.; Beiden, S.V.; Campbell, G.; Metz, C.; Sacks, W.M. Assessment of medical imaging and computer-assist systems. *Acad. Radiol.* **2002**, *9*, 1264–1277.
73. Zhou, X.-H.; McClish, D.K.; Obuchowski, N.A. *Statistical Methods in Diagnostic Medicine*; Wiley: Hoboken, NJ, 2002.
74. Pennello, G.A. Sample size for paired repeated measures designs of method comparison studies: application to pulse oximetry. In *American Stat. Assoc. 2002 Proc. of the Biopharm. Section*, Volume M(85); Joint Stat. Meetings, New York, New York, American Stat. Assoc. Alexandria, VA, 2002.
75. Linnet, K.; Kondratovich, M. Partly nonparametric approach of determining the limit of detection. *Clin. Chem.*, **2004**, *50*, 732–740.
76. Sargent, D.J.; Conley, B.A.; Allegra, C.; Collette, L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J. Clin. Oncol.* **2005**, *23* (9), 2020–2027.
77. Buyse, M.; Loi, S.; van't Veer, L.; Viale, G.; Delorenzi, M.; Glas, A.M.; et al., Validation and clinical utility of a 70-gene

- prognostic signature for women with node-negative breast cancer. *J. Natl. Cancer Inst.* **2006**, *98*, 1183–1192.
78. Simon, R. Roadmap for developing and validating therapeutically relevant genomic classifiers. *J. Clin. Oncol.* **2005**, *23*, 7332–7341.
 79. Lababidi, S. Challenges in DNA microarray studies from the regulatory perspective. *J. Biopharm. Stat.* **2008**, *18*, 183–202.
 80. U.S. Department of Health and Human Services, Food and Drug Administration. *Draft Guidance for Industry, Clinical Laboratories, and FDA Staff – In Vitro Diagnostic Multivariate Index Assays*. (released July 26, 2007), <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm079148.htm> (accessed August 2009).
 81. Liu, J.-P.; Chow, S.-C. Statistical issues on the diagnostic multivariate index assay for targeted clinical trials. *J. Biopharm. Stat.* **2008**, *18*, 167–182.
 82. Berry, D.A., Multiple Comparisons, Multiple Tests, and Data Dredging: A Bayesian Perspective, In *Bayesian Statistics 3*, Bernardo, DeGroot, Lindley, Smith, Eds., 1988, 79–94.
 83. Little, R.; Rubin, D. *Statistical Analysis with Missing Data*, 2nd Ed.; Wiley: New York, 2002.
 84. Berry, D.A. Bayesian statistics and the efficiency and ethics of clinical trials. *Stat. Sci.* **2004**, *19* (1), 175–187.
 85. Spiegelhalter, D.J.; Abrams, K.R.; Myles, J.P. *Bayesian Approaches to Clinical Trials and Health-care Evaluation*; Wiley: New York, 2004.
 86. Congdon, P. *Bayesian Statistical Modelling*, 2nd Ed.; Wiley: New York, 2007.
 87. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*, 2nd Ed.; Chapman and Hall/CRC: London, 2004.
 88. Carlin, B.P.; Louis, T. *Bayesian Methods for Data Analysis*, 3rd Ed.; Chapman and Hall/CRC: London, 2008.
 89. Berry, D. *Using a Bayesian Approach in Medical Device Development. Technical Report 97-21*. Duke University Institute of Statistics and Decision Sciences: Durham, NC, 1997, Web address <http://citeseer.ist.psu.edu/old/401145.html> (accessed August 2009).
 90. Pennello, G.; Thompson, L. Experience with reviewing Bayesian medical device trials. *J. Biopharm. Stat.* **2008**, *18* (1), 81–115.
 91. Campbell, G. The experience in the center for devices and radiological health with bayesian strategies. *Clin. Trials* **2005**, *2*, 359–363.
 92. U.S. Department of Health and Human Services, Food and Drug Administration, *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials; Draft Guidance*. (released May 23, 2006) <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm> (accessed August 2009).
 93. Irony, T.Z.; Pennello, G.A. Choosing an Appropriate Prior for Bayesian Medical Device Trials in the Regulatory Setting. In *American Statistical Association 2001 Proceedings of the Biopharmaceutical Section*, Joint Statistical Meetings, American Statistical Association: Atlanta, GA, Alexandria, VA, 2001; *M*, #85.
 94. Rubin, D.B.; Stern, H.S. Sample size determination using posterior predictive distributions. *Sankhya* **1998**, *60*, 161–175.
 95. Katsis, A.; Toman, B. Bayesian sample size calculations for binomial experiments. *J. Stat. Plan. Inference* **1999**, *81*, 349–362.
 96. Joseph, L.; Wolfson, D.B.; Berger, R. Sample size calculations for binomial proportions via highest posterior density intervals. *Statistician* **1995**, *46*, 139–144.
 97. Joseph, L.; Wolfson, D.B. and Berger, R. Some comments on Bayesian sample size determination. *Statistician* **1995**, *44*, 167–171.
 98. Zou, K.H.; Normand, S.T. On determination of sample size in hierarchical binomial models. *Stat. Med.* **2001**, *20*, 2163–2182.
 99. Lindley, D.V. The choice of sample size. *Statistician* **1997**, *46* (2), 129–138.
 100. Berger, J.O.; Wolpert, R.L. *The Likelihood Principle*, 2nd Ed.; IMS: Hayward, CA, 1988.
 101. Carlin, B.P.; Kadane, J.; Gelfand, A. Approaches for optimal sequential decision analysis in clinical trials. *Biometrics*, **1998**, *54*, 964–975.
 102. Hobbs, B.P.; Carlin, B.P. Practical bayesian design and analysis for drug and device clinical trials. *J. Biopharm. Stat.* **2008**, *18* (1), 54–80.
 103. Kadane, J.B. *Bayesian Methods and Ethics in a Clinical Trial Design*; Wiley: New York, 1996.
 104. U.S. Department of Health and Human Services, Food and Drug Administration. 510(k) Summary; (released June 4, 2003) http://www.accessdata.fda.gov/cdrh_docs/pdf2/K023691.pdf (accessed August 2009).
 105. Tarone, R.E. The use of historical control information in testing for a trend in proportions. *Biometrics* **1982**, *38*, 215–220.
 106. Pocock, S. The combination of randomized and historical controls in clinical trials. *J. Chron. Dis.* **1976**, *29*, 175–188.
 107. Gould, A.L. Using prior findings to augment active-controlled trials and trials with small placebo groups. *Drug Inf. J.* **1991**, *25*, 369–380.
 108. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*; Chapman and Hall: London, 1995.
 109. Malec, D. A closer look at combining data among a small number of binomial experiments. *Stat. Med.* **2001**, *20* (12), 1811–1824.
 110. Dixon, D.O.; Simon, R. Bayesian subset analysis in a colorectal cancer clinical trial. *Stat. Med.* **1992**, *11*, 13–22.
 111. Gilks, W.R.; Richardson, S.; Spiegelhalter, D.J. *Markov Chain Monte Carlo in Practice*; Chapman and Hall: London, 1996.
 112. Spiegelhalter, D.J.; Thomas, A.; Best, N.G.; Gilks, W.R. *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.5 (version ii)*; MRC Biostatistics Unit: Cambridge, 1996, Available at the website <http://www.mrc-bsu.cam.ac.uk> (accessed February 2002).
 113. Lao, C.S.; Kessler, L.G.; Gross, T. Proposed statistical methods for signal detection of adverse medical device events. *Drug Inf. J.* **1998**, *32*, 183–191.
 114. Lao, C.S. Statistical issues involved in medical device post-marketing surveillance. *Drug Inf. J.* **2000**, *34*, 483–493.
 115. DuMouchel, W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am. Stat.* **1999**, *53* (3), 177–202.
 116. Yue, L.Q. Special issue on medical device clinical studies – Guest editor's note, *J. Biopharm. Stat.* **2008**, *18* (1), 1–3.

Meta-Analysis of Therapeutic Trials

Joseph C. Cappelleri

Pfizer Inc., New London, Connecticut, U.S.A.

John P. A. Ioannidis

Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece

Joseph Lau

Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, Massachusetts, U.S.A.

Abstract

In this entry, the focus is primarily on the meta-analysis of randomized therapeutic trials in which aggregated, or summary, estimates of treatment effect are combined.

INTRODUCTION

How can a large amount of independent quantitative information on the same question come together in a coherent and meaningful manner? Many researchers have relied on meta-analysis to achieve such synthesis of evidence. Traditional journal review articles and textbook chapters, upon which most clinicians have traditionally relied for therapeutic guidance, generally tend to make selective and subjective appraisals of the evidence and usually do not provide a quantitative synthesis of the data.^[1] Although it can be applied to data of any sort, meta-analysis has been commonly used to combine results from different studies to draw conclusions about treatments.

“Meta-analysis” may be defined as the statistical analysis of data from multiple studies. A meta-analysis typically identifies data systematically, summarizes results, and evaluates quantitatively sources of heterogeneity and bias. A systematic literature review encompasses an explicit and detailed description of how a review was conducted. Some have characterized meta-analysis broadly to include a literature systematic literature review, as well as the statistical analysis of data. In this sense, meta-analysis has been also referred to as a quantitative systematic review or an overview.

The advent of the meta-analytic research method in the biopharmaceutical sciences has paralleled the explosion in the number of randomized trials being conducted and the increasing need to provide an unbiased quantitative synthesis in the era of evidence-based medicine and comparative effectiveness. Meta-analysis thus far has used mostly data from completed trials, but there is an increasing momentum for using the same principles of data synthesis in the prospective design of sets of clinical trials addressing a similar question.^[2,3] Meta-analysis has also been applied to

epidemiological studies to derive better risk estimates and to evaluations of diagnostic tests to obtain more reliable estimates of test performance. The basic statistical principles of meta-analysis can be expanded to apply to diverse types of data. In this chapter, however, we focus primarily on the meta-analysis of randomized therapeutic trials in which aggregated, or summary, estimates of treatment effect are combined.

We provide a review on the application of meta-analysis as a reference to aid pharmaceutical scientists, regulatory reviewers, and biostatisticians who are engaged in pharmaceutical research and development. The section “Benefits and Limitations” highlights the benefits and limitations of meta-analysis. “Statistical Analysis for Discrete Data” discusses common statistical methods for pooling discrete data from two-by-two contingency tables. “Statistical Analysis for Continuous Data” focuses on a method of pooling continuous data from multiple studies. The section “Basic Steps in Performing and Evaluating a Meta-Analysis” covers the basic steps to perform and evaluate a meta-analysis of randomized therapeutic trials. “Applications” illustrates several applications on how meta-analysis has helped to determine optimal therapeutic use and to enhance the understanding of the efficacy, safety, and cost-effectiveness of different pharmaceuticals. Finally, the section “Some Key Developments” highlights some of the major key methodological developments and advancements.

BENEFITS AND LIMITATIONS

Benefits

Meta-analysis offers several benefits. It may be used to address uncertainty and heterogeneity when results of

studies disagree, to increase statistical power for primary outcomes and subgroups, to improve estimates of treatment effect, and to lead to new knowledge and formulation of new questions.^[4] It is also a powerful tool for exploring sources of heterogeneity and bias in clinical research. The rapidly increasing volume of research, often with discrepant findings, has led to an increased need for meta-analysis. We identified 225 published meta-analyses of randomized controlled trials in 1993, compared with 7 in 1980. At least 500 meta-analyses of randomized trials were published in 1996 and 1997 alone. Several thousands of meta-analyses have been published to date, and thousands more are on the way.

Pooling results from studies with inadequate sample sizes may lead to a more precise and statistically significant estimate of a treatment effect. Applying regression methods to combined (pooled) results from studies may help to explain heterogeneity of treatment effects and differences across studies; as such, enhanced understanding may be gained and research hypotheses may be generated and subsequently evaluated. A comprehensive, well-done meta-analysis may help the researcher to know the strengths and weaknesses of the studies and therefore may aid in the design and analysis of future studies, which may bring about improvements in the quality of research. Finally, when performed prospectively, a meta-analysis may serve as a tool for planning the collection and analysis of large-scale evidence.

Limitations

Several criticisms have been leveled at meta-analysis.^[5,6] The major limitation of retrospective meta-analysis is that it is based only on what studies are available. An offshoot of this is the “apples and oranges” phenomenon pertaining to the external validity (generalizability) of the process: Combining different studies may lead to uncertainty as to which study or to what specific combination of studies and populations the results apply to. Furthermore, the studies available for pooling may vary in design, quality, outcome measures, or populations studied, calling into question whether it is logical and useful to combine their results. To address this issue, the researcher should prepare a protocol that includes well-defined criteria and objectives for including the studies in a meta-analysis as well as plans for subgroup analyses and regression analyses that may examine differences (heterogeneity) among studies. Another way to address this is to adhere to guidelines, whenever feasible, on the quality of reporting of meta-analyses.^[7]

Publication bias is a second limitation of using only available studies. Publication bias is induced when the decision to publish is influenced by whether the study result of an experimental treatment is “positive” (i.e., favorable and statistically significant) or “negative” (i.e., statistically unfavorable or not statistically significant). Theoretical and empirical evidence suggests that, if the treatment effects

of the studies included in the meta-analysis are found to be related to the sample size or the variance of the treatment effects, this association in some occasions may suggest publication bias,^[8] as small negative studies are more likely to be unpublished than large ones.

On the other hand, it can be argued that unpublished data have not been evaluated by peer review and therefore may be of dubious quality and should not be trusted. Although a source of concern, if there are only a few unpublished studies and each has a small sample size, their omission may be less likely to cause a substantial problem because they may then have minimal impact on the overall result. Nevertheless, there is evidence that even large efficacy trials may suffer from “publication lag”,^[9] a term coined to reflect the situation when negative efficacy trials appear in the literature with a time lag compared with their positive counterparts which are completed, submitted, and published more promptly. This may result in larger treatment effects in meta-analyses of the early evidence.

Several tests have been suggested in the literature to detect the possibility of publication bias, and some of them such as the “funnel plot” in which a measure of precision (e.g., sample size or the reciprocal of the variance) is plotted against the treatment effect are popular and commonly applied.^[8] Non-parametric^[10] and parametric methods^[11] to formally test for such funnel asymmetry have become accepted. When selection bias is present, the plot becomes asymmetrical, and the overall effect of the meta-analysis is biased. Empirical research, though, indicates that the shape of a funnel plot may be largely determined by the arbitrary choice of the method to construct the plot (i.e., by the definition of precision and the effect measure).^[12] Asymmetry in a funnel plot may also stem from the heterogeneity of treatment effect across studies as well as from chance and subjectivity.^[13] Overall, visual assessment of funnel plot asymmetry is an unreliable method, and even formal statistical methods for assessing asymmetry can only tell at best whether small studies tend to have different results from large studies (small-study bias), which may have nothing to do with publication bias.^[14]

In mitigating publication bias, research registries^[15,16] and other reasonable means (e.g., following up on published abstracts) can be used to track down unpublished studies. Prospective trial registries is probably the best way to circumvent publication bias and publication lag because, as new trials become registered before they are started, the decision to register a trial cannot be affected by the result of the study, which is not yet known.

A well-rounded, thorough review on publication bias has been published elsewhere.^[17] More recently, an up-to-date volume on publication bias in meta-analysis provides comprehensive coverage that includes the following: different types of publication bias, mechanisms that may induce them, empirical evidence for their existence, statistical methods to address them, and ways they can be avoided.^[18] This edited volume features worked examples

with common data sets and explains and compares available software used for analyzing and reducing publication bias.

A final criticism against meta-analysis centers on its internal validity in that it cannot correct the qualitative flaws of the studies that it includes and thus it carries along the biases inherent in each included trial. As a careful meta-analysis identifies the quality defects of the individual trials; however, the identification of bias becomes one of the prime benefits of a systematic meta-analysis. Several scales on assessing the quality of randomized trials in meta-analyses have gained currency, such as the ones developed by Chalmers et al.^[19] and Jadad et al.^[20] A comprehensive bibliography of quality assessment methods has been published.^[21] So far, there is little evidence that the magnitude of the treatment effect corresponds to the quality of the studies.

An investigation of 25 scales for quality assessment found that the type of scale used to assess trial quality could result in different conclusions.^[22] Because the use of summary scores to identify trials of high quality can be problematic, it has been suggested that relevant methodological aspects should be assessed individually and their influence on effect sizes explored.^[22] By focusing attention on the quality of clinical trials, meta-analysis has probably provided a powerful stimulus for improving their conduct and their reporting.^[23]

STATISTICAL ANALYSIS FOR DISCRETE DATA

This section emphasizes the main statistical methods of meta-analysis that have been used when a study's results come from a 2×2 contingency table with two treatment groups and a binary outcome, a situation common in randomized therapeutic trials. A compendium of methods for combining discrete data can be found in *The Handbook of Research Synthesis*,^[24] one of the most comprehensive books on statistical, as well as non-statistical, matters pertaining to meta-analysis. Examples are also provided there.

It is not advisable to simply lump the results of different studies together as if only one large study had been performed. Not only may the results of larger studies unduly overwhelm the results of smaller studies, but the overall results may be distorted. As shown in Table 1, the risk ratio (i.e., relative risk) for each study is the same (0.50). Yet

simply adding their results gives a misleading value (0.39), often referred to as Simpson's paradox. Correct methods of pooling would, of course, give a pooled risk ratio of 0.50.

Fixed-Effects and Random-Effects Models

Two types of models have been employed to appropriately adjust for the potential confounding effect of study. In the fixed-effects model, which leads to inferences only about the studies assembled, the true treatment effect is assumed to be the same across all studies and the only variability of treatment effect considered is within each study (i.e., within-study variability). In the random-effects model, which leads to inference about all studies from a hypothetical population, each study can have a different effect that is randomly drawn from a normal distribution and is positioned about a central value; in this model, not only within-study variability but also variability of treatment effects across studies (i.e., among-study variability) is considered.

The fixed-effects model weights each study by its sampling variability of treatment effect, which incorporates the number of events and sample size in each treatment group of a study. Along with weighting by this sampling variation, the random-effects model also weights each study by an overall estimate of the variability of true treatment effects across studies. Compared with the fixed-effects model, the random-effects model tends to give wider confidence intervals, which makes its statistical significance more conservative than the fixed-effects models.

In practice, important differences in the results between these two statistical models occur infrequently. The random-effects model should be used when heterogeneity of treatment effect is present—that is, when there is evidence that there is no single effect of treatment across studies (see the next section “Assessing Homogeneity of Treatment Effects”).

For combining discrete data from 2×2 tables on, for instance, survival status and two treatments, fixed-effects models that have been used to combine estimates of treatment effect across studies include (but are not limited to) the following: directly weighted pooled risk differences and pooled risk ratios;^[25] Mantel–Haenszel procedure for pooling odds ratios and pooling risk ratios;^[25] and Peto's method for pooling odds ratio.^[26] Analogous random-effects models include the DerSimonian–Laird method for pooling risk differences,^[27] pooling odds ratios^[28] and

Table 1 Fictional Results of Two Randomized Controlled Trials

Trial	Treatment group			Control group			Risk ratio
	Deaths	Number	Risk	Deaths	Number	Risk	
A	20	100	20%	40	100	40%	20/40 = 0.50
B	50	500	10%	20	100	20%	10/20 = 0.50
Total	70	600	11.7%	60	200	30%	11.7/30 = 0.39
							11.7/30 ≠ 0.50

Table 2 Statistical Methods of Pooling Treatment Effects from 2×2 Contingency Tables

	Risk difference	Odds ratio	Risk ratio
Fixed Effects Model	Directly Weighted	Mantel–Haenszel with RBG Peto	Mantel–Haenszel with GR Directly weighted
Random Effects Model	D & L	D & L	D & L

D & L: DerSimonian and Laird (From Ref. [27])

RBG: Robins-Breslow-Greenland variance estimator (From Ref. [25])

GR: Greenland-Robins variance estimator (From Ref. [25])

pooling risk ratios,^[29] all of which add to the fixed-effects models an extra variance component of between-study variability of treatment effects.

Table 2 compartmentalizes these methods for pooling results from different studies. Programs, publicly available without charge, to analyze such data include Meta-Analyst,^[30] MIX,^[31,32] and Revman.^[33] We recommend that the user of such programs be fully familiar with the limitations of different formulas for calculating treatment effects and variances. For example, the Peto method may be biased if the allocation ratio of patients in the study arms is far from 1.^[34]

If person-time were used instead of number of persons in the denominator of a rate in a study, formulas for computing the incidence rate ratio and incidence rate difference of a study—along with how to pool them with a fixed-effects model—have been documented.^[25] Corresponding methods of pooling for a random-effects model are the same as those for a fixed-effects model but add between-study variability of treatment effects.^[29] For continuous data, pooled effect sizes exist for fixed-effects models^[35] and random-effects models.^[36]

We delineate the steps and formulas to derive the pooled risk ratio, a common measure used to combine the results of prospective trials, in **Table 3** with the directly weighted fixed-effects approach^[25] and in **Table 4** with the DerSimonian–Laird (random-effects) approach.^[27,29] The directly weighted fixed-effects approach is essentially the DerSimonian–Laird method with no interstudy variability of treatment effect (i.e., $\tau^2 = 0$ in **Table 4**).

Assessing Homogeneity of Treatment Effects

A few of the several ways of measuring heterogeneity are highlighted here. A chi-square test of homogeneity (or, some may say, heterogeneity) is used to test whether there is evidence that the treatment effects are different across studies.^[37] Step 1 of **Table 4** shows a chi-square statistic to assess whether the risk ratios are different across studies. The issue is not whether differences exist, for the population treatment effects of different studies are not expected to be identical, but whether they can be reasonably ignored. Given the low power of the chi-square test of homogeneity, as an arbitrary rule of thumb, P -values at or below 0.1 can usually be taken to mean that the differences should not be ignored. However, in most typical applications of

Table 3 Steps in Computing an Estimate and 95% Confidence Interval of a Pooled Fixed-Effects Risk Ratio

1. Compute natural logarithm of the risk ratio for the i th study ($i = 1, \dots, n$ total studies):

$$\ln(RR_i) = \ln(TR_i/CR_i)$$

2. Compute its standard error:

$$se_i = \left[\frac{b_i}{a_i(a_i + b_i)} + \frac{d_i}{c_i(c_i + d_i)} \right]^{1/2}$$

3. Compute the fixed-effects weight for each study:

$$w_i = 1/(se_i)^2$$

4. Compute the pooled natural logarithm of the fixed-effect risk ratio:

$$\ln(RRf) = \frac{\sum w_i \ln(RR_i)}{\sum w_i}$$

5. Compute its antilogarithm to obtain the pooled fixed-effects risk ratio:

$$RRf = e^{\ln(RRf)}$$

6. Compute the corresponding 95% confidence interval:

$$e^{\ln(RRf) \pm 1.96/\sqrt{\sum w_i}}$$

a_i = number of treated patients with the event of interest in the i th study ($i = 1, \dots, n$ studies)

b_i = number of treated patients without the event of interest in the i th study

c_i = number of control patients with the event of interest in the i th study

d_i = number of control patients without the event of interest in the i th study

$TR_i = a_i/(a_i + b_i)$ = the treatment rate in the i th study

$CR_i = c_i/(c_i + d_i)$ = control rate in the i th study

meta-analysis in biomedical fields, the number and size of studies is too limited and the chi-square test is thus underpowered. A non-significant result may not rule out heterogeneity.^[38]

As a function of the chi-square statistic of homogeneity (Q), the I^2 statistic is a descriptive measure that quantifies the extent of total variation in treatment effects across studies that is due to heterogeneity alone rather than chance;^[39,40] it is a kind of signal-to-noise ratio. It is computed as

$$I^2 = \left(\frac{Q - df}{Q} \right) 100\%,$$

where df refers to degrees of freedom, which equals the number of studies minus 1. Values on I^2 can range from 0–100%. Proposed benchmark values on I^2 of 25%, 50%,

Table 4 Steps in Calculating an Estimate and 95% Confidence Interval of a Pooled Random-Effects Risk Ratio

1. Calculate the homogeneity test statistic for the underlying risk ratios:

$$Q = \sum w_i [\ln(RR_i) - \ln(RR_f)]^2$$

Note: The null hypothesis is that the n underlying risk ratios are equal.

For hypothesis testing, this test statistic (Q) is compared with a given percentile (for example, 90th) of a chi-square distribution with $n - 1$ degrees of freedom.

2. Calculate the interstudy variability of treatment effects:

$$D = \frac{Q - (n - 1)}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}}$$

$$I^2 = D \text{ if } D > 0 \text{ or } I^2 = 0 \text{ if } D \leq 0$$

3. Calculate the random-effects weight for each study:

$$w_i^* = [I^2 + (1 / w_i)]^{-1}$$

4. Calculate the pooled natural logarithm of the random effects risk ratio:

$$\ln(RR^*) = \frac{\sum w_i^* \ln(RR_i)}{\sum w_i^*}$$

5. Calculate its antilogarithm to obtain the pooled random effects risk ratio:

$$RR^* = e^{\ln(RR^*)}$$

6. Calculate the corresponding 95% confidence interval:

$$e^{\ln(RR^*) \pm 1.96 / \sqrt{\sum w_i^*}}$$

Definition and calculation of terms not appearing here appear in Table 3.

and 75% might be considered low, medium, and high, respectively. The I^2 statistic is viewed as a measure of inconsistency across the findings of the studies. A caveat about the I^2 is that with a limited number of studies, it tends to have large uncertainty. The 95% confidence intervals of I^2 can be readily estimated and are useful to convey the extent of uncertainty.^[41] One may also formally examine in sensitivity analyses the impact of specific studies on the extent of estimated between-study heterogeneity.^[42]

The Galbraith plot,^[43] a plot of the z-score of treatment effect versus the square root of the weight for each trial, is a convenient graphical technique to examine heterogeneity of effects. The L'Abbe plot,^[44] a plot of the treatment rate against the control rate, is another graphical technique to explore such heterogeneity. An empirical assessment of treatment effect and heterogeneity across 125 meta-analyses revealed that, while summary odds ratios and risk differences tended to agree in statistical significance, risk differences usually displayed more heterogeneity than odds ratios.^[45]

Meta-Regression Analysis

Meta-regression analysis^[46,47] is sometimes used as part of a meta-analysis to describe how a treatment effect may vary

with patient and study characteristics. Meta-regressions have been applied to several clinical areas.^[47–51] In a meta-regression of summary data, the unit of analysis is the individual study in the meta-analysis. Aggregate quantities reported in the studies, such as the mean dosage, mean age, and mean duration of disease, could be extracted and used as independent variables. The treatment effect (e.g., the risk difference) can be the dependent variable.

Taking the natural logarithm of a treatment effect measured on a multiplicative scale, such as the risk ratio and odds ratio, is suggested to help ensure that the dependent variable becomes normally distributed in order to make valid statistical inferences. It is more desirable to weight each study in the meta-regression proportional to its precision (e.g., the inverse for the variance of a study's treatment effect) rather than weighting the studies equally. When output is taken from certain standard statistical packages, like the Statistical Analysis System (SAS; Ref. [52]), is used to obtain the standard error of a regression coefficient, this standard error should be adjusted by dividing it by the square root of the mean squared error, as is appropriate for inverse-variance weighted least squares regression.^[46] Though potentially informative, the results of meta-regression based on covariates with aggregate numbers (e.g., group averages) do not necessarily prove causality and should be interpreted as cautiously as other ecological investigations when making inferences to individual patients.^[53,54] There are many caveats in the application of meta-regressions in practice, and these are discussed more extensively elsewhere.^[55,56]

STATISTICAL ANALYSIS FOR CONTINUOUS DATA

Characterizations based on fixed and random effects, assessment of homogeneity of treatment effects, and meta-regression analysis described above for discrete data also pertain for continuous data. In this section, a common statistical approach is highlighted for pooling summary estimates of continuous data for the difference between two treatment groups. This approach combines mean differences in the original units of the variable of interest. Details and examples on this approach are found elsewhere,^[24] as are combining standardized mean differences and combining correlations.

Fixed-Effects Model

Under the fixed-effects model, all estimates measure a common treatment effect and differences among these estimates are due to chance. Suppose that the data to be combined arise from a series of n independent studies. Let T_i be the mean difference between treatment and control group means in the i th study, and let $S^2(T_i)$ be the sample variance of T_i . Algebraically,

$$T_i = \bar{X}_i^t - \bar{X}_i^c$$

and

$$S^2(T_i) = S_{pi}^2 \left(\frac{1}{n_i^t} + \frac{1}{n_i^c} \right)$$

where, for the i th study, S_{pi}^2 is the estimated pooled common variance found in introductory statistics books, and n_i^t and n_i^c denote the sample sizes in the treatment and control groups, respectively.

The weighted average effect over the n studies becomes

$$\bar{T}_c = \frac{\sum_{i=1}^n w_i T_i}{\sum_{i=1}^n w_i}$$

where

$$w_i = \frac{1}{s^2(T_i)}$$

represents the weights that are inversely proportional to the sample variance in each study and that minimize the sample variance of $\bar{T}_c$, which is

$$S^2(\bar{T}_c) = \frac{1}{\sum_{i=1}^n \frac{1}{s^2(T_i)}}$$

The 95% confidence interval of the true treatment effect then becomes

$$\bar{T}_c \pm 1.96[S^2(\bar{T}_c)]^{1/2}.$$

If the confidence interval does not contain zero, the null hypothesis of a population treatment effect of zero gets rejected at the 0.05 level of significance.

Random-Effects Model

Regardless of whether a fixed-effects or random-effects model is employed, a test should be made about the assumption that the studies share a common treatment effect. This assumption can be tested with the following homogeneity test statistic:

$$Q^* = \sum_{i=1}^n w_i (T_i - \bar{T}_c)^2$$

In general, if Q^* exceeds the upper-tail critical value of the chi-square distribution with n minus 1 degrees of freedom, the observed variance across treatment effects is significantly greater than what would be expected by chance if all studies shared a common effect.

A random-effects model would incorporate an estimate of the between-study variance component:

$$S_T^2 = \frac{Q^* - (n-1)}{\sum_{i=1}^n w_i - \frac{\sum_{i=1}^n w_i^2}{\sum_{i=1}^n w_i}}$$

The computations follow those for the fixed-effects model except that the revised variance estimate

$$S^{2*}(T_i) = S_T^2 + S^2(T_i)$$

replaces $S^2(T_i)$.

BASIC STEPS IN PERFORMING AND EVALUATING A META-ANALYSIS

Though important, statistical analysis is only one facet of a comprehensive meta-analysis. A properly conducted meta-analysis can be most demanding and requires the utmost advanced planning and care in its conduct. Fig. 1 depicts the basic steps to perform a retrospective meta-analysis.

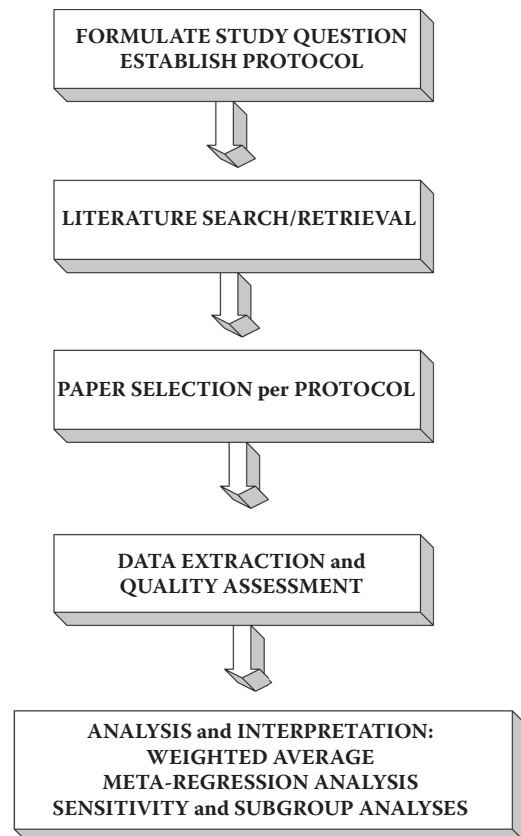


Fig. 1 Basic steps in performing a meta-analysis

Quantitative methods give an aura of credence to any analysis. A meta-analysis, similar to any clinical study, must be read carefully, recognizing its strengths and limitations. Guidelines are available on how to approach the scientific quality of a meta-analysis,^[57–59] although there is no full consensus on how to do this even among experts. There is some wider acceptance of the need for reporting standards, as exemplified by the QUORUM and PRISMA recommendations.^[7,23] In general, to assess quality, the following questions may be posed:

- a. Objectives: Were the objectives clearly stated, clinically useful, and well-formulated? Was the design of the meta-analysis addressing a well-specified population and clinical question?
- b. Identification and Selection of Studies: How were the studies selected? Were the inclusion criteria explicit? Which studies were excluded? Why? Were unpublished papers sought and their impact examined? How many sources of data were screened (e.g., electronic databases, meeting presentations, communications with experts and pharmaceutical sponsors)?
- c. Information about the Studies: Was information given on the specific study design, patient and disease characteristics, each treatment being compared, and study duration? What precautions were taken to avoid bias? What were the results of the individual studies? Were the clinical outcomes of the studies well-defined and useful? What criteria were chosen to appraise the quality of the studies?
- d. Methods of Analysis: What method was employed to combine the results? Was the method technically correct? Were confidence intervals, as well as *P* values, given? Was a fixed-effect model or a random-effect model used? Was there justification to combine the trials? Was a test of homogeneity of treatment effect performed? Were subgroup analyses defined a priori? Was sensitivity analysis used? Were meta-regressions performed to explore whether treatment effect was influenced by potential modifier variables? Was it considered whether the baseline disease risk of the patients affected the magnitude of the treatment effect?
- e. Results and Conclusions: Were the results interpreted cautiously? Were the conclusions justified by the data? What was the generalizability (external validity) of the included studies and the meta-analysis? Were limitations of the meta-analysis mentioned?

APPLICATIONS

Numerous applications appear in the literature on how the results of meta-analyses of pharmaceuticals have helped to determine optimal therapeutic use and to enhance the understanding of pharmaceuticals with respect to their

efficacy, safety, and cost-effectiveness. We highlight some of these applications, classifying them by a methodological framework.

Cumulative Meta-Analysis vs. Expert Opinion

Cumulative meta-analysis is a method of updating previous meta-analyses with the appearance of new studies.^[60] The method allows the monitoring of developing trends of therapeutic efficacy. When performed routinely, the earliest time when statistical significance is reached (by whatever criterion is chosen) can be identified.

The first placebo-controlled randomized study of a thrombolytic drug in the treatment of acute myocardial infarction was reported in 1959. A cumulative meta-analysis on this treatment (Fig. 2) could have found that in 1973, after eight studies included 2,432 patients, a statistically significant reduction in overall mortality of about 25% had been achieved with thrombolytic drugs relative to placebo (Mantel–Haenszel odds ratio = 0.74, 95% confidence interval, 0.59–0.92).^[61] Yet 62 more placebo-controlled trials were reported subsequently, involving an additional 45,000 patients. The Federal Drug Administration (FDA) did not approve a thrombolytic drug for this indication until 1988, 15 years after cumulative meta-analysis indicated that the treatment could have been found effective in saving lives. Compared with the results of the cumulative meta-analysis, expert opinions lagged substantially in recognizing the efficacy of the treatment and did not begin to recommend it until around the time of FDA approval.

Meta-Analysis in Cost-Effectiveness Analysis

Pooled estimates from a meta-analysis could serve as input into a decision analysis or a cost-effectiveness analysis. For example, a cost-effectiveness of enoxaparin, a low-molecular weight heparin derivative, was compared with low-dose warfarin in the prevention of deep-vein thrombosis (DVT) after total hip replacement.^[62] To determine the proportion of DVT with enoxaparin and warfarin prophylaxis, the authors pooled the proportions from the available studies to arrive at the efficacy of the drugs as input into determining that the incremental cost-effectiveness of enoxaparin relative to warfarin was estimated as \$29,120 per life-year gained (in Canadian dollars).

Meta-Analysis for Optimal Dosing

A meta-analysis of 35 randomized placebo-controlled trials was undertaken to study the efficacy and safety of different aspirin dosages for patients at high risk of vascular disease.^[49] Dose-response meta-regressions indicated no evidence that lower doses of aspirin were more efficacious or less efficacious in preventing overall mortality and vascular events. There was evidence, however, that lower doses of aspirin significantly reduced gastrointestinal

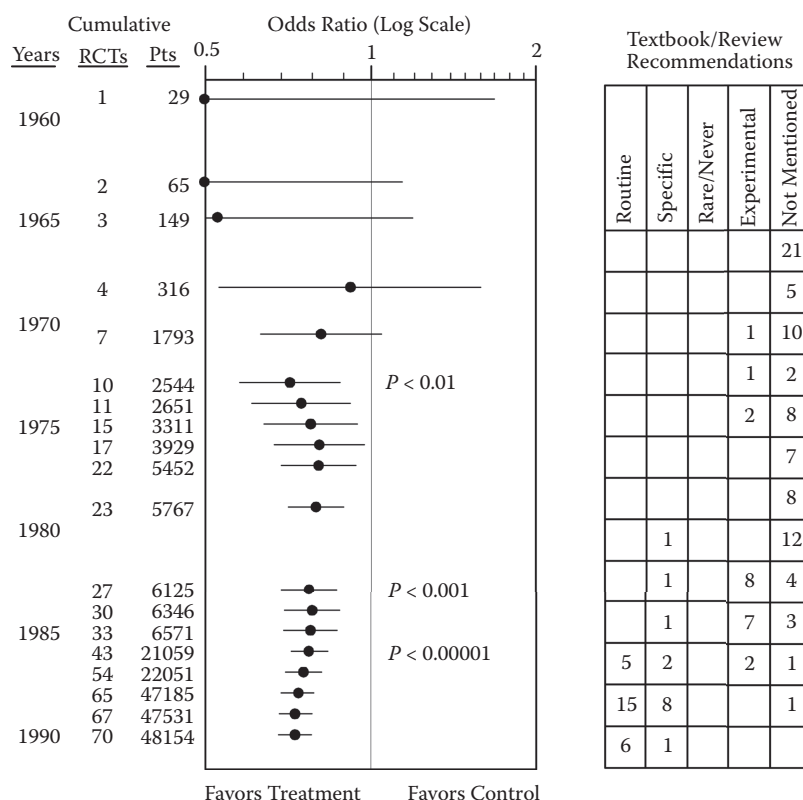


Fig. 2 Cumulative meta-analysis of thrombolytic therapy vs. placebo for acute myocardial infarction. Note: The cumulative meta-analyses by year of publication of randomized controlled trials (RCTs) are presented on the left. The cumulative number of trials and patients (Pts) are also presented. On the right, the recommendations of the clinical expert reviewers are presented in two-year segments, except for the entry in 1966, which represents all previous years (From Ref. [61].)

symptoms. For a daily dosage increase of 100 milligrams of aspirin, the (unadjusted) odds ratio of gastrointestinal symptoms increased by 4.5% on average (99% confidence interval, 1.83–7.25%).

Aminoglycosides are one of the most commonly used antibiotics for serious infections. There are several agents in this class with very similar mechanism of action and toxicity profile. The drugs have traditionally been administered with multiple doses a day, although experimental evidence would suggest that giving the whole daily dose in a single daily administration should be as efficacious and possibly less toxic. Twenty-one trials were used to compare single versus multiple daily doses of aminoglycosides. A meta-analysis of these trials convincingly showed that single daily doses were about 25% less nephrotoxic than and at least as efficacious as multiple daily doses.^[50] These findings are expected to result in better outcomes, substantial cost savings, and increased convenience.

Many randomized controlled trials addressed the efficacy of many different regimens in the prophylaxis of *Pneumocystis carinii* infection among patients with advanced human immunodeficiency virus infection. A meta-analysis of 35 randomized trials showed the marked superiority of trimethoprim-sulfamethoxazole regimens over aerosolized

pentamidine and dapsone.^[51] The meta-analysis predicted a 43% reduction in the odds of discontinuations of prophylaxis due to side effects when trimethoprim-sulfamethoxazole was used at a lower dose (one double strength tablet three times a week) instead of the regular dose (one double strength tablet daily), with no substantial loss of efficacy.

Meta-Analysis for Bioequivalence

Under current FDA regulation, a patient may switch from a brand-name drug to a generic drug if they are shown to be bioequivalent. Bioequivalence between generic copies of the brand-name drug, however, is not required by the FDA. Chow and Liu^[63] have proposed a methodology to provide an overview on the assessment of (average) bioequivalence among approved generic copies of an expired brand-name drug. Their methodology was intended to help guide the FDA and others in evaluating the bioequivalence among the approved generic copies. More work is needed in this area.

In general, meta-analysis has been gaining recognition by FDA. Although it is frequently used in issuing warnings and treatment guidelines, though, large and randomized clinical trials continue to be relied on and necessary for approval.

Drug Planning and Development

Perhaps the most fruitful applications of meta-analysis in the years to come are in the prospective planning of drug and therapeutic strategy development. Current drug development typically is trying to meet the regulatory mandates for showing treatment efficacy and safety, and is also largely dictated by piecemeal planning and clinical opportunities for the drug. Prospective meta-analysis offers a framework for building all the trials in the development of a drug within a single matrix with explicit quantitative plans for the analysis and interpretation of the complete accumulated evidence. This approach will alleviate the threat of publication bias and will instill full transparency and credibility to the results of developmental trials seen in their totality. On an even larger scale, prospective meta-analysis matrices may incorporate different drugs for the same disease or indications. Such prospective meta-analyses are one of the most appealing challenges of current therapeutics. Some current research addresses specifically the possibility to design new trials based on the appraisal of prior evidence to see what types of comparisons are missing^[64] and how big future studies should be.^[65]

SOME KEY DEVELOPMENTS

There have been a multitude of recent developments in meta-analysis and the field has been a fertile area for statistical and epidemiological research. We highlight three major developments here.

Control Rate

For a dichotomous outcome, the control rate is the proportion of patients in the control group with the event of interest. The control rate in a clinical trial has gained acceptance as an available summary measure of patient or study differences that may help to explain the heterogeneity of treatment effects across studies.^[66–68] The control rate is a summary measure that is usually reported. Variation in control rates across studies may reflect different patient populations, underlying baseline risk of patients, length of study follow-up, and treatment delivery. As a readily available simple measure to explore treatment effect heterogeneity, the control rate may help delineate when one needs further research on risk factors.

The size of the treatment effects of studies in a meta-analysis may tend to vary with the size of the control rates in these studies. Several methodological advances have been made in this area.^[69–73] For example, in an empirical investigation that analyzed 112 meta-analyses, Schmid et al.^[73] used a hierarchical model to find that the control rate was significantly associated with the risk difference in 35 meta-analyses (31%), with the risk ratio in 15 instances (13%), and with the odds ratio in 16 (14%). The impact of

differing patient risks on the results of clinical trials, along with the advantages and limitations of using the control rate as a measure of baseline risk, have been discussed.^[74,75]

Large Trials and Meta-Analyses

The extent of concordance between the results of large trials and the results of meta-analyses of smaller trials on the same topic has stirred considerable debate. Three different groups of investigators evaluated and compared these results. Villar et al.^[76] found moderate agreement, beyond chance, between results of large trials and those of smaller trials. Cappelleri et al.^[77] concluded that the results of the two methods usually agree, but discrepancies do occur; however, clinically important discrepancies without an obvious explanation were unusual. LeLorier et al.^[78] concluded that up to 35% of the time a meta-analysis failed to predict accurately the results of a subsequent, large trial.

In their evaluation of the three protocols, Ioannidis et al.^[79] concluded that a comparison of large trials with meta-analyses may reach different conclusions depending on the design used, the studies selected, and the method of analysis employed. When a robust approach was used that accounts for both the magnitude and the uncertainty of treatment effect, the estimated range of disagreements was fairly similar in the three protocols (10–23%). Future clinical trials may benefit in addressing sources of heterogeneity of treatment effect between large trials themselves and even within a large trial. Similarly, future meta-analyses may benefit in addressing sources of heterogeneity rather than try to always fit a common estimate among diverse studies.

Meta-Analyses of Individual Patient Data

Most meta-analyses in the past have been conducted using summary, or aggregated, data on treatments and subgroups from clinical trials. Meta-analyses of individual patient data (MIPD) offers several distinct advantage of performing more detailed analyses using all the data on each individual patient participating in a clinical trial to be included in meta-analysis. This may allow, with the use of regression analyses, for detailed time-to-event analysis, the development of more accurate predictive models for the risk of the studied outcome, and the quantification of different treatment effects for different subgroups. Furthermore, the process of performing MIPD may also improve the quality of trials by revealing additional issues that cannot be gleaned from published reports. Stewart and Clarke^[80] provided a description of the organizational steps required for performing a MIPD. Increasingly, applications using MIPD from randomized trials, as well as from observational studies, have burgeoned in different therapeutic areas.^[81–84]

The results of MIPD against meta-analyses on group data give similar, if not identical, findings for the main effect of treatment when no covariates are in the model. They may

not necessarily give similar results or even similar conclusions with regard to modification and heterogeneity of treatment effect for different levels of a covariate.

For instance, in a survey of group-level and individual patient-level data from five randomized trials on the benefits of anti-lymphocyte antibody induction therapy among renal transplant patients, Berlin et al.^[85] found that the treatment was significantly more efficacious among patients with elevated levels (20% or more) panel reactive antibodies than among patients without such levels based on the individual-patient analysis; the group-level analysis failed to detect this interaction. The authors concluded that individual patient data should be used, when feasible, to study patient characteristics in order to avoid the potential for ecological bias introduced by group-level analyses.

Schmid et al.^[86] conducted two investigations to evaluate Bayesian meta-regression for detecting treatment interactions. The first investigation involved the comparison of aggregate and individual patient data on 1,860 subjects from 11 trials testing angiotensin converting enzyme inhibitors for nondiabetic kidney disease; the second explored meta-regression for detecting treatment interaction 671 covariates, including the baseline risk from 232 meta-analyses of binary outcomes. These authors concluded that meta-regression can detect interaction of treatment with study-level factors when treatment effects are heterogeneous, and that individual patient data are needed for patient-level factors and homogeneous effects.

Discrepancies may arise when different data are used by the MIPD and the meta-analysis of group data (such as updated information beyond the main trial follow-up in the MIPD). Disadvantages of MIPD include the fact that it requires more effort and coordination on an international level, and the potential for retrieval bias when individual patient data can be retrieved only for a selected group of the pertinent trials.

Nevertheless, meta-analysis of individual patient data may present the highest step in the hierarchy of evidence.^[87] The statistical methods for synthesis at the patient level are, to a large extent, similar to those used in the analysis of multicenter clinical trials. The main feature that differentiates MIPD from a single trial is that a covariate can be included at the study level, with the effect of the covariate on outcome being either fixed or random and possibly made to depend on other factors. To learn more about the advantages and disadvantages of systematic reviews using individual patient data, the reader can consult Stewart and Tierney.^[88]

Methodological research on MIPD is well-suited to examine possible heterogeneity of treatment effect with both patient-level and study-level covariates in multi-level models. The use of prognostic scores derived from MIPD, for example, may help to better explain a possible relationship between underlying risk and treatment effect. Moreover, when MIPD is available only from a proportion of known studies, combining MIPD with aggregate data can

lead to more complete information. For more information on using multilevel (hierarchical) models for these purposes, the reader is referred to Goldstein et al.,^[89] Turner et al.,^[90] and Sutton et al.^[91]

Cumulative Meta-Analysis

Cumulative meta-analysis, which has been used to illustrate important gaps between accumulated evidence and the behavior of researchers and practitioners, typically involves performing an updated meta-analysis when a new trial or trials are added to a series of similar trials, which by definition involves multiple inspections. Neither the commonly used random effects model nor the conventional group sequential method can control the type I error for many practical situations.

In addressing this problem, Lan et al.,^[92] for the case of continuous outcomes, and Hu et al.,^[93] for the case of binary outcomes, invoked the Law of Iterated Logarithm (LIL) to “penalize” the Z-value of the test statistics to account for multiple tests and for the estimate of heterogeneity in treatment effects across studies. The LIL method involves an adjustment factor that is directly related to the control of type I error and determined through extensive simulations under various conditions. The LIL method controls the overall type I error for a broad span of practical situations with a binary outcome or continuous outcome; the LIL works properly in controlling the type I error rates as the number of inspections becomes large.

Other researchers have made refinements and extensions to cumulative meta-analysis.^[94–97]

Indirect Comparisons

Although the randomized controlled trial is the most valid design for comparing the efficacy of interventions, many competing interventions have not been compared in randomized controlled trials (or, if compared, scantily so) and indirect methods have been used instead for medical evidence and policy decisions.^[98,99] Suppose, for example, that Treatment A is directly compared with Treatment B, and Treatment B is directly compared with Treatment C, across several studies. Although no direct comparison can be made here between Treatment A versus Treatment C, an indirect comparison can be between them. Such indirect comparisons, though, are generally subject to greater bias than head-to-head comparisons, as the data of interest are not available; therefore, it is essential to promote research that evaluates such bias, which may distort the estimate of treatment effect and, in doing, lead to inappropriate decisions.

These so-called mixed treatment comparisons can be viewed as an extension of traditional meta-analysis to which all possible treatment options can be compared. Consider two simple ways to indirectly compare two treatments A and B: (i) combine all A arms and combine

all B arms and (ii) use A versus C arms and B versus C arms as additional evidence. The latter, but not the former, approach preserves randomization and within-study comparisons. For this latter method, the difference in the treatment effect between respective pairs of treatment (A and C against B and C) gives an (indirect) estimate of the treatment effect between A and B. This indirect comparison, however, involves an assumption of coherence: that the treatment effects from indirect comparisons should be the same as the treatment effects from direct comparisons.

In addition to respecting the randomization of the included trials, Bayesian Markov Chain Monte Carlo (MCMC) methods and network meta-analysis can provide an efficient way to combine data for indirect comparisons, acknowledge their extra variability, and examine the plausibility of assumptions.^[100–107] These methods also yield probability statements on how likely each treatment is to another and, as such, enhance interpretation of results and overcome the difficulty created by multiple comparisons. Time will tell whether the MCMC methodology will become mainstream to assess treatment efficacy.

Other Developments: Methods, Software, Education

Numerous other developments on meta-analysis have been made on methodological, computational, and educational fronts. Some methodological advancements include (but are not restricted to) those on publication bias,^[18] Bayesian approaches,^[108] longitudinal analysis,^[109] survival analysis,^[110] and a general multivariate approach to meta-analysis.^[111] Sutton and Higgins^[112] have compiled a thorough, wide-ranging review on recent development in meta-analysis, which include the following: heterogeneity and random-effects analyses, special considerations in different areas of application, an assessment of bias within and across studies, and an extension of ideas to complex or nonstandard evidence synthesis.

Advances in software programs dedicated to meta-analysis have made statistical computing of pooled data more accessible and powerful.^[113] Among the most highly rated of them are the commercially available Comprehensive Meta-Analysis^[114] and freely available MIX.^[31,32] Among the general-purpose programs, the freely available BUGS for Windows^[115] and the commercially available SAS^[52,109,111,116–118] and STATA^[119] are among the most comprehensive. A spreadsheet application like Microsoft Office EXCEL is well-suited to handle basic calculations.^[120] The freely available MetaEasy is a meta-analysis add-in for Microsoft EXCEL.^[121]

Many instructional articles and textbooks have appeared, especially over the last dozen years, to enrich and expand one's knowledge of meta-analysis.^[18,24,109,111,118,120,122–133]

CONCLUSIONS

Meta-analysis may be defined as the statistical analysis of data from multiple studies. A meta-analysis typically identifies data systematically, summarizes results, and evaluates quantitatively sources of heterogeneity and bias. This chapter has highlighted the benefits and limitations of meta-analysis. Common statistical methods were described for pooling discrete data from 2×2 contingency tables and for pooling continuous data from multiple studies. Basic steps to perform and evaluate a meta-analysis of randomized therapeutic trials were featured. Several applications of meta-analysis were highlighted. Some of the major methodological developments and resources for further reading were noted.

REFERENCES

1. Mulrow, C. The medical review article: state of the science. *Ann. Intern. Med.* **1987**, *106*, 485–488.
2. Laupacis, A.; Connolly, S.J.; Gent, M.; Roberts, R.S.; Cairns, J.; Joyner, C. How should results from completed studies influence ongoing clinical trials? The CAFA Study experience. *Ann. Intern. Med.* **1991**, *115*, 818–822.
3. Henderson, W.G.; Moritz, T.; Goldman, S.; Copeland, J.; Sethi, G. Use of cumulative meta-analysis in the design, monitoring, and final analysis of a clinical trial: a case study. *Control. Clin. Trials* **1995**, *16*, 331–341.
4. Sacks, H.S.; Berrier, J.; Reitman, D.; Ancona-Berk, V.A.; Chalmers, T.C. Meta-analyses of randomized controlled trials. *N. Engl. J. Med.* **1987**, *316*, 450–455.
5. Thacker, S.B. Meta-analysis. A quantitative approach to research integration. *JAMA* **1988**, *259*, 1685–1689.
6. Goodman, S.N. Have you ever seen a meta-analysis you didn't like? *Ann. Intern. Med.* **1991**, *114*, 244–246.
7. Moher, D.; Cook, D.J.; Eastwood, S.; Olkin, I.; Rennie, D.; Stroup, D. For the QUORUM group. Improving the quality of reporting of meta-analysis of randomized controlled trials: the QUORUM statement. *Lancet* **1999**, *354*, 1896–1900.
8. Sutton, A.J. Publication bias. In *The Handbook of Research Synthesis*, 2nd Ed.; Cooper, H.; Hedges, L.V.; Valentine, J.C.; Eds.; Russell Sage Foundation: New York, 2009, 435–452.
9. Ioannidis, J.P. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA* **1998**, *279*, 281–286.
10. Begg, C.B.; Mazumdar, M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* **1994**, *50*, 1088–1101.
11. Egger, M.; Smith, G.D.; Schneider, M.; Minder, C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* **1997**, *315*, 629–634.
12. Tang, J.-L.; Liu, J.L.Y. Misleading funnel plot for detection of bias in meta-analysis. *J. Clin. Epidemiol.* **2000**, *53*, 477–484.

13. Terrin, N.; Schmid, C.H.; Lau, J.; Olkin, I. Adjusting for publication bias in the presence of heterogeneity. *Stat. Med.* **2003**, *22*, 2113–2126.
14. Lau, J.; Ioannidis, J.P.A.; Terrin, N.; Schmid, C.H.; Olkin, I. The case of the misleading funnel plot. *BMJ* **2006**, *333*, 597–600.
15. Easterbrook, P.J. Directory of registries of clinical trials. *Stat. Med.* **1992**, *11*, 345–359.
16. Matt, G.E.; Cook, T.D. Threats to the validity of generalized inferences. In *The Handbook of Research Synthesis*, 2nd Ed.; Cooper, H.; Hedges, L.V.; Valentine, J.C.; Eds.; Russell Sage Foundation: New York, 2009, 537–560.
17. Thornton, A.; Lee, P. Publication bias in meta-analysis: its causes and consequences. *J. Clin. Epidemiol.* **2000**, *53*, 207–216.
18. Rothstein, H.R.; Sutton, A.J.; Borenstein, M.; Eds.; *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*; Wiley: Chichester, 2005.
19. Chalmers, T.C.; Smith, H.; Blackburn, B.; Silverman, B.; Schroeder, B.; Reitman, D.; Ambroz, A. A method for assessing the quality of a randomized control trial. *Control. Clin. Trials* **1981**, *2*, 31–49.
20. Jadad, A.R.; Moore, A.; Carroll, D.; Jenkinson, C.; Reynolds, D.J.; Gavaghan, D.J.; McQuay, H.J. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Control Clin. Trials* **1996**, *17*, 1–12.
21. Moher, D.; Jadad, A.R.; Nichol, G.; Penman, M.; Tugwell, P.; Walsh, S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control. Clin. Trials* **1995**, *16*, 62–73.
22. Jüni, P.; Witschi, A.; Bloch, R.; Egger, M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* **1999**, *282*, 1054–1060.
23. Liberati, A.; Altman, D.G.; Tetzlaff, J.; Mulrow, C.; Gøtzsche, P.C.; Ioannidis, J.P.A.; Clarke, M.; Devereaux, P.J.; Kleijnen, J.; Moher, D.; and the PRISMA Group. The PRISMA Statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Ann. Intern. Med.* **2009**, *151*, W64–W94.
24. Cooper, H.; Hedges, L.V.; Valentine, J.C.; Eds.; *The Handbook of Research Synthesis*, 2nd Ed.; Russell Sage Foundation: New York, 2009.
25. Rothman, K.J.; Greenland, S.; Lash, T.L. *Modern Epidemiology*, 3rd Ed.; Lippincott Williams & Wilkins: Philadelphia, PA, 2008.
26. Yusuf, S.; Peto, R.; Lewis, J.; Collins, R.; Sleight, P. Beta blockade during and after myocardial infarction: an overview of randomized trials. *Prog. Cardiovasc. Dis.* **1985**, *27*, 335–371.
27. DerSimonian, R.; Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* **1986**, *7*, 177–188.
28. Fleiss, J.L.; Gross, A.J. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *J. Clin. Epidemiol.* **1991**, *44*, 127–139.
29. Ioannidis, J.P.; Cappelleri, J.C.; Lau, J.; Skolnik, P.R.; Melville, B.; Chalmers, T.C.; Sacks, H.S. Early or deferred zidovudine therapy in HIV-infected patients without an AIDS-defining illness. *Ann. Intern. Med.* **1995**, *122*, 856–866.
30. Meta-Analyst [computer program]. J. Lau version 0.991. New England Medical Center: Boston, MA, 1997.
31. Bax, L.; Yu, L.M.; Ikeda, N.; Tsuruta, H.; Moons, K.G.M. Development and validation of MIX: comprehensive free software for meta-analysis of causal research. *BMB Med. Res. Methodol.* **2006**, *6*, 50.
32. Bax, L.; Yu, L.M.; Ikeda, N.; Tsuruta, H.; Moons, K.G.M. *MIX: Comprehensive Free Software for Meta-Analysis of Causal Research Data*. Version 1.7. <http://mix-for-meta-analysis.info>. Sagamihara, Japan: Evidence Synthesis Division—MIX, 2008.
33. Review Manager (RevMan) [Computer program]. Version 5.0. The Nordic Cochrane Centre, The Cochrane Collaboration: Copenhagen, Denmark, 2008.
34. Greenland, S.; Salvan, A. Bias in the one-step method for pooling study results. *Stat. Med.* **1990**, *9*, 247–252.
35. Konstantopoulos, S.; Hedges, L.V. Analyzing effect sizes: Fixed-effects models. In *The Handbook of Research Synthesis*, 2nd Ed.; Cooper, H.; Hedges, L.V.; Valentine, J.C.; Eds.; Russell Sage Foundation: New York, 2009, 279–293.
36. Raudenbush, S.W. Analyzing effect size: Random-effects models. In *The Handbook of Research Synthesis*, 2nd Ed.; Cooper, H.; Hedges, L.V.; Valentine, J.C.; Eds.; Russell Sage Foundation: New York, 2009, 295–315.
37. Fleiss, J.L.; Levin, B.; Paik, C.M. *Statistical Methods for Rates and Proportions*, 3rd Ed.; Wiley: Hoboken, NJ, 2003, 234–283.
38. Ioannidis, J.P. Interpretation of tests of heterogeneity and bias in meta-analysis. *J. Eval. Clin. Pract.* **2008**, *14*, 951–957.
39. Higgins, J.P.T.; Thompson, S.G.; Deeks, J.J.; Altman, D.G. Measuring inconsistency in meta-analyses. *BMJ* **2003**, *327*, 557–560.
40. Higgins, J.P.T.; Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **2002**, *21*, 1539–1558.
41. Ioannidis, J.P.; Patsopoulos, N.A.; Evangelou, E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ* **2007**, *335*, 914–916.
42. Patsopoulos, N.A.; Evangelou, E.; Ioannidis, J.P. Sensitivity of between-study heterogeneity in meta-analysis: proposed metrics and empirical evaluation. *Int. J. Epidemiol.* **2008**, *37*, 1148–1157.
43. Galbraith, R.F. A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat. Med.* **1988**, *7*, 889–894.
44. L'Abbe, K.A.; Detsky, A.S.; O'Rourke, K. Meta-analysis in clinical research. *Ann. Intern. Med.* **1987**, *107*, 224–233.
45. Engels, E.A.; Schmid, C.H.; Terrin, N.; Olkin, I.; Lau, J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Stat. Med.* **2000**, *19*, 1707–1728.
46. Berlin, J.A.; Antman, E.M. Advantages and limitations of metaanalytic regressions of clinical trials data. *Online J. Curr. Clin. Trials* **1994**, Doc No 134, June 4.
47. Kasiske, B.L.; Kalil, R.S.N.; Ma, J.Z.; Liao, M.; Keane, W.F. Effect of antihypertensive therapy on the kidney in patients with diabetes: a meta-regression analysis. *Ann. Intern. Med.* **1993**, *118*, 129–138.
48. Thompson, S.G. Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. *Stat. Methods Med. Res.* **1993**, *2*, 173–192.

49. Cappelleri, J.C.; Lau, J.; Kupelnick, B.; Chalmers, T.C. Efficacy and safety of different aspirin dosages on vascular diseases in high-risk patients. A metaregression analysis. *Online J. Curr. Clin. Trials* **1995**, Doc No 174, March, 14.
50. Barza, M.; Ioannidis, J.P.; Cappelleri, J.C.; Lau, J. Single or multiple daily doses of aminoglycosides: a meta-analysis. *BMJ* **1996**, *312*, 338–345.
51. Ioannidis, J.P.; Cappelleri, J.C.; Skolnick, P.R.; Lau, J.; Sacks, H.S. A meta-analysis of the relative efficacy and toxicity of *Pneumocystis carinii* prophylactic regimens. *Arch. Intern. Med.* **1996**, *156*, 177–188.
52. SAS Institute Inc. *SAS/STAT User's Guide*; Version 9.1. SAS Institute: Cary, NC, 2003.
53. Langbein, L.I.; Lichtman, A.J. *Ecological Inference*; Sage Publications: Newbury Park, CA, 1978.
54. Morgenstern, H. Uses of ecologic analysis in epidemiologic research. *Am. J. Public Health* **1982**, *72*, 1336–1344.
55. Thompson, S.G.; Higgins, J.P. How should meta-regression analyses be undertaken and interpreted? *Stat. Med.* **2002**, *21*, 1559–1573.
56. Higgins, J.P.; Thompson, S.G. Controlling the risk of spurious findings from meta-regression. *Stat. Med.* **2004**, *23*, 1663–1682.
57. Abramson, J.H. *Making Sense of Data*, 2nd Ed.; Oxford University Press: New York, 1994, 318–389.
58. Higgins, J.P.T.; Green, S., Eds.; *Cochrane Handbook for Systematic Reviews of Interventions*; The Cochrane Collaboration and John Wiley & Sons, Ltd: Chichester, 2008.
59. Geller, N.L.; Proschan, M. Meta-analysis of clinical trials: a consumer's guide. *J. Biopharm. Stat.* **1996**, *6*, 377–394.
60. Lau, J.; Antman, E.M.; Jimenez-Silva, J.; Kupelnick, B.; Mosteller, F.; Chalmers, T.C. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N. Engl. J. Med.* **1992**, *327*, 248–254.
61. Antman, E.M.; Lau, J.; Kupelnick, B.; Mosteller, F.; Chalmers, T.C. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *JAMA* **1992**, *268*, 240–248.
62. O'Brien, B.J.; Anderson, D.R.; Goeree, R. Cost-effectiveness of enoxaparin versus warfarin prophylaxis against deep-vein thrombosis after total hip replacement. *CMAJ* **1994**, *150*, 1083–1090.
63. Chow, S.C.; Liu, J.P. Meta-analysis for bioequivalence review. *J. Biopharm. Stat.* **1997**, *7*, 97–111.
64. Salanti, G.; Kavvoura, F.K.; Ioannidis, J.P. Exploring the geometry of treatment networks. *Ann. Intern. Med.* **2008**, *148*, 544–553.
65. Sutton, A.J.; Cooper, N.J.; Jones, D.R. Evidence synthesis as the key to more coherent and efficient research. *BMC Med. Res. Methodol.* **2009**, *9*, 29.
66. Smith, G.D.; Song, F.; Sheldon, R.A. Cholesterol lowering and mortality: the importance of considering initial level of risk. *BMJ* **1993**, *306*, 1367–1373.
67. Boissel, J.P.; Collet, J.P.; Lievre, M.; Girard, P. An effect model for the assessment of drug benefit: example of anti-arrhythmic drugs in postmyocardial infarction patients. *J. Cardiovasc. Pharmacol.* **1993**, *22*, 356–363.
68. Brand, R.; Kragt, H. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Stat. Med.* **1992**, *11*, 2077–2082.
69. Sharp, S.J.; Thompson, S.G.; Altman, D.G. The relation between treatment benefit and underlying risk in meta-analysis. *BMJ* **1996**, *313*, 735–738.
70. McIntosh, M.W. The population risk as an explanatory variable in research synthesis of clinical trials. *Stat. Med.* **1996**, *15*, 1713–1728.
71. Thompson, S.G.; Smith, T.C.; Sharp, S.J. Investigation underlying risk as a source of heterogeneity in meta-analysis. *Stat. Med.* **1997**, *16*, 2741–2758.
72. Walter, S.D. Variation in baseline risk as an explanation of heterogeneity in meta-analysis. *Stat. Med.* **1997**, *16*, 2883–2990.
73. Schmid, C.H.; Lau, J.; McIntosh, M.W.; Cappelleri, J.C. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Stat. Med.* **1998**, *17*, 1923–1942.
74. Lau, J.; Ioannidis, J.P.; Schmid, C.H. Summing up evidence: one answer is not always enough. *Lancet* **1997**, *351*, 123–127.
75. Ioannidis, J.P.; Lau, J. The impact of high-risk patients on the results of clinical trials. *J. Clin. Epidemiol.* **1997**, *50*, 1089–1098.
76. Villar, J.; Carroli, G.; Belizan, J.M. Predictive ability of meta-analyses of randomised controlled trials. *Lancet* **1995**, *345*, 772–776.
77. Cappelleri, J.C.; Ioannidis, J.P.; Schmid, C.H.; de Ferranti, S.D.; Aubert, M.; Chalmers, T.C.; Lau, J. Large trials vs meta-analysis of smaller trials: how do their results compare? *JAMA* **1996**, *276*, 1332–1338.
78. LeLorier, J.; Gregoire, G.; Benhaddad, A.; Lapierre, L.; Derderian, F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *N. Engl. J. Med.* **1997**, *337*, 536–542.
79. Ioannidis, J.P.; Cappelleri, J.C.; Lau, J. Issues in comparisons between meta-analyses and large trials. *JAMA* **1998**, *279*, 1089–1093.
80. Stewart, L.A.; Clarke, M.J. Practical methodology of meta-analyses (overviews) using updated individual patient data. Cochrane Working Group. *Stat. Med.* **1995**, *14*, 2057–2079.
81. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiologic studies of 52,705 women with breast cancer and 108,411 women without breast cancer. *Lancet* **1997**, *350*, 1047–1059.
82. Camma, C.; Giunta, M.; Chemello, L.; Alberti, A.; Toyoda, H.; Trepo, C.; Marcellin, P.; Zahm, F.; Schalm, S.; Craxi, A. Chronic hepatitis C: interferon retreatment of relapsers. A meta-analysis of individual patient data. European Concerted Action on Viral Hepatitis (EUROHEP). *Hepatology* **1999**, *30*, 801–807.
83. Trikalinos, T.A.; Ioannidis, J.P. Predictive modeling and heterogeneity of baseline risk in meta-analysis of individual patient data. *J. Clin. Epidemiol.* **2001**, *54*, 245–252.
84. The ACE Inhibitors in Diabetic Nephropathy Trialist Group. Should all patients with type 1 diabetes mellitus and microalbuminuria receive angiotensin-converting enzyme inhibitors? a meta-analysis of individual patient data. *Ann. Intern. Med.* **2001**, *134*, 370–379.
85. Berlin, J.A.; Santanna, J.; Schmid, C.H.; Szczech, L.A.; Feldman, H.I. Individual patient-versus group-level data

- meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Stat. Med.* **2002**, *21*, 371–387.
86. Schmid, C.H.; Stark, P.C.; Berlin, J.A.; Landais, P.; Lau, J. Meta-regression detected association between heterogeneous treatment effects and study-level, but not patient-level factors. *J. Clin. Epidemiol.* **2004**, *57*, 683–697.
 87. Olkin, I. Statistical and theoretical considerations in meta-analysis. *J. Clin. Epidemiol.* **1995**, *48*, 133–146.
 88. Stewart, L.A.; Tierney, J.F. To IPD or not to IPD? Advantages and disadvantages of systematic review using individual patient data. *Eval. Health Prof.* **2002**, *25*, 76–97.
 89. Goldstein, H.; Yang, M.; Omar, R.; Turner, R.; Thompson, S.G. Meta-analysis using multilevel models with an application to the study of class size effects. *Appl. Stat.* **1999**, *49*, 399–412.
 90. Turner, R.M.; Omar, R.Z.; Yang, M.; Goldstein, H.; Thompson, S.G. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Stat. Med.* **2000**, *19*, 3417–3432.
 91. Sutton, A.J.; Kendrick, D.; Coupland, C.A.C. Meta-analysis of individual- and aggregate-level data. *Stat. Med.* **2008**, *27*, 651–669.
 92. Lan, K.K.G.; Hu, M.; Cappelleri, J.C. Applying the law of iterated logarithm to cumulative meta-analysis of a continuous endpoint. *Stat. Sin.* **2003**, *13*, 1135–1145.
 93. Hu, M.; Cappelleri, J.C.; Lan, K.K.G. Applying the law of iterated logarithm to control type 1 error in cumulative meta-analysis of binary outcomes. *Clin. Trials* **2007**, *4*, 329–340.
 94. Whitehead, A. A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. *Stat. Med.* **1997**, *16*, 2901–2913.
 95. Pogue, J.; Yusuf, S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Control. Clin. Trials* **1997**, *18*, 580–593.
 96. Ioannidis, J.P.; Contopoulos-Ioannidis, D.G.; Lau, J. Recursive cumulative meta-analysis: a diagnostic for the evolution of total randomized evidence from group and individual patient data. *J. Clin. Epidemiol.* **1999**, *52*, 281–291.
 97. Ioannidis, J.P.; Lau, J. Evolution of treatment effects over time: empirical insight from recursive cumulative meta-analyses. *Proc. Natl. Acad. Sci.* **2001**, *98*, 831–836.
 98. Bucher, H.C.; Guyatt, G.H.; Griffith, L.E.; Walter, S.D. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J. Clin. Epidemiol.* **1997**, *50*, 683–691.
 99. Sutton, A.; Ades, A.E.; Cooper, N.; Abrams, K. Use of indirect and mixed treatment comparisons for technology assessment. *Pharmacoeconomics* **2008**, *26*, 753–767.
 100. Higgins, J.P.T.; Whitehead, A. Borrowing strength from external trials in meta-analysis. *Stat. Med.* **1996**, *15*, 2733–2749.
 101. Caldwell, D.M.; Ades, A.E.; Higgins, J.P.T. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ* **2005**, *331*, 897–900.
 102. Song, F.; Altman, D.G.; Glenny, M.-A.; Deeks, J.J. Validity of indirect comparison for estimating of competing interventions: empirical evidence from published meta-analyses. *BMJ* **2003**, *326*, 472–476.
 103. Lu, G.; Ades, A.E. Combination of direct and indirect evidence in mixed treatment comparison. *Stat. Med.* **2004**, *23*, 3105–3124.
 104. Lumley, T. Network meta-analysis for indirect treatment comparisons. *Stat. Med.* **2002**, *21*, 2313–2324.
 105. Lu, G.; Ades, A.E. Assessing evidence inconsistency in mixed treatment comparisons. *JAMA* **2006**, *101*, 447–459.
 106. Dominici, F.; Parmigiani, G.; Wolpert, R.L.; Hasselblad, V. Meta-analysis of migraine headache treatments: combining information from heterogeneous designs. *JASA* **1999**, *94*, 16–28.
 107. Salanti, G.; Higgins, J.P.T.; Ades, A.E.; Ioannidis, J.P.A. Evaluation of networks of randomized trials. *Stat. Methods Med. Res.* **2008**, *17*, 279–301.
 108. Stangl, D.K.; Berry, D.A., Eds.; *Meta-Analysis in Medicine and Health Policy*; Marcel Dekker: New York, 2000.
 109. Ishak, K.J.; Platt, R.W.; Joseph, L.; Hanley, J.A.; Caro, I.J. Meta-analysis of longitudinal studies. *Clin. Trials* **2007**, *4*, 525–539.
 110. Arends, L.R.; Myriam Hunink, M.G.; Stijnen, T. Meta-analysis of summary survival curve data. *Stat. Med.* **2008**, *27*, 4381–4396.
 111. van Houwelingen, H.C.; Arends, L.R.; Stijnen, T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat. Med.* **2002**, *21*, 589–624.
 112. Sutton, A.J.; Higgins, J.P.T. Recent developments in meta-analysis. *Stat. Med.* **2008**, *27*, 625–650.
 113. Bax, L.; Yu, L.-M.; Ikeda, N.; Moons, K.G.M. A systematic comparison of software dedicated to meta-analysis causal studies. *BMC Med. Res. Methods* **2007**, *7*, 40.
 114. Borenstein, M.; Hedges, L.; Higgins, J.; Rothstein, H. *Comprehensive Meta Analysis Version 2*; Biostat: Englewood, NJ, 2005.
 115. Spiegelhalter, D.J.; Thomas, A.; Best, N.G.; Lunn, D. *WinBUGS User Manual Version 1.4*; MRC Biostatistics Unit: Cambridge, 2002.
 116. Arthur Jr., W.; Bennett Jr., W.; Huffcutt, A.I. *Conducting Meta-Analysis Using SAS*; Lawrence Erlbaum Associates, Mahwah, NJ, 2001.
 117. Wang, M.C.; Bushman, B.J. *Integrating Results Through Meta-Analytic Review Using SAS® Software*; SAS Institute: Cary, NC, 1999.
 118. Whitehead, A. *Meta-Analysis of Controlled Clinical Trials*; Wiley: Chichester, 2002.
 119. Sterne, J.A.C.; Bradburn, M.J.; Egger, M. Meta-analysis in Stata. In *Systematic Reviews in Health Care*; Egger, M.; Davey Smith, G.; Altman, D.G.; Eds.; BMJ Publishing Group: London, 2001, 347–369.
 120. Lipsey, M.W.; Wilson, D.B. *Practical Meta-Analysis*; Sage Publications: Thousand Oaks, CA, 2001.
 121. Kontopantelis, E.; Reeves, D. MetaEasy: A meta-analysis add-in for Microsoft Excel. *J. Stat. Softw.* **2009**, *30*, 7.
 122. Lau, J.; Ioannidis, J.P.; Schmid, C.H. Quantitative synthesis in systematic reviews. *Ann. Intern. Med.* **1997**, *127*, 820–826.
 123. Normand, S.L. Meta-analysis: formulating, evaluating, combining, and reporting. *Stat. Med.* **1999**, *18*, 321–359.

124. Petitti, D.B. *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine*, 2nd Ed.; Oxford University Press, New York, 2000.
125. Stevens, A.; Abrams, K.; Brazier, J.; Fitzpatrick, R.; Lilford, R.; Eds.; *The Advanced Handbook of Methods in Evidence Based Healthcare*; Sage Publications: Thousand Oaks, CA, 2001.
126. Sutton, A.J.; Abrams, K.R.; Jones, D.R.; Sheldon, T.A.; Song, F. *Methods for Meta-Analysis in Medical Research*; John Wiley & Sons: Chichester, 2000.
127. Borenstein, M.; Hedges, L.V.; Higgins, J.P.T.; Rothstein, H.R. *Introduction to Meta-Analysis*; Wiley: Chichester, 2009.
128. Littell, J.H.; Corcoran, J.; Pillai, V. *Systematic Reviews and Meta-Analysis*; Oxford University Press: New York, 2008.
129. Egger, M.; Smith, D.G.; Altman, D.G., Eds.; *Systematic Reviews in Health Care: Meta-Analysis in Context*, 2nd Ed.; BMJ Books: London, 2001.
130. Hartung, J.; Knapp, G.; Sinha, B.K. *Statistical Meta-Analysis with Applications*; John Wiley & Sons: Hoboken, NJ, 2008.
131. Kulinskaya, E.; Morgenthaler, S.; Staudte, R.G. *Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence*; John Wiley & Sons, Ltd: Chichester, 2008.
132. Hedges, L.V.; Olkin, I. *Statistical Methods for Meta-Analysis*; Academic Press: Orlando, FL, 1985.
133. Cooper, H. *Research Synthesis and Meta-Analysis: A Step-by-Step Approach*, 4th Ed.; Sage Publications: Thousands Oaks, CA, 2009.

Microarray Gene Expression

James J. Chen

National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, Arkansas, U.S.A.

Chun-Houh Chen

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

Abstract

Microarray is a device for measuring transcription abundance of a large number of genes simultaneously under experimental conditions. This entry discusses microarray technology and experimental designs, and also touches on sample size estimation.

Keywords: Classification and prediction; Clustering analysis; Experimental design; Gene Selection; Gene set analysis; Matrix map, Sample size; Visualization.

INTRODUCTION

A gene consists of a segment of deoxyribonucleic acid (DNA). A DNA molecule is a double-stranded polymer composed of four basic molecular units called nucleotides. The expression of genetic information stored in the DNA molecule occurs in two stages: (i) transcription, during which DNA is transcribed into mRNA (messenger ribonucleic acid), a single-stranded complementary copy of the base sequence in the DNA molecule; and (ii) translation, during which mRNA is translated into protein, which are action molecules of the cell responsible for nearly all cellular processes. The majority of genes are expressed as the protein they encode. Although regulation of protein synthesis is not solely controlled by mRNA levels, the changes in protein abundance are determined in part by changes in the levels of mRNA. By measuring transcription levels of genes in an organism under various conditions, at different developmental stages and different tissues, one can characterize the dynamic function of each gene in the genome. DNA microarray is a device for measuring transcription abundance of a large number of genes simultaneously under experimental conditions.

DNA microarray technology provides tools for studying the expression levels of a large number of distinct genes simultaneously.^[1] The technology is based on the hybridization of known DNA segments on the array with an unknown labeled DNA or RNA sample based on base-pairing rules. Gene expression levels are quantified from the image of a hybridized microarray excited by a

laser scanner. Signal intensities reflect the amount of transcript present for the gene in the mRNA sample. A typical microarray experiment data set includes expression levels of thousands of genes in a number of experimental samples (conditions). These samples may correspond to different toxins or serial time points taken during a biological process. The expression data can be summarized by a matrix with rows representing genes and columns representing samples. This matrix is referred to as a gene expression matrix. A gene expression matrix can be analyzed by comparing expression profiles of an individual gene (row) under various conditions or comparing expression profiles of a sample (column) to understand their similarities and differences and to find the relationships among genes and sample conditions.

DNA MICROARRAY EXPERIMENT

A microarray experiment is a multi-step process, where each step is potentially a source of variation. Usually, raw data generated from image analysis cannot be directly compared. Data need to be normalized in order to minimize systematic biases so that biological differences can be distinguished. Statistical analyses generally are of two types. The first type of analysis involves identification of the genes that are differentially expressed among different sample groups. The second type seeks to determine the relationships between genes or gene clusters to identify biological functions, or to predict specific biological outcomes (or diseases) from the analysis of expression patterns. Many statistical methods are used in microarray data processing and analysis. Microarray technology, experimental design, data preprocessing and normalization, significance

* The views presented in this paper are those of the authors and do not necessarily represent those of the U.S. Food and Drug Administration.

testing, classification and prediction, and cluster analysis and pattern recognition are discussed in the following sections.

MICROARRAY TECHNOLOGY

cDNA Microarray

A cDNA microarray consists of thousands of single-stranded known DNA fragments attached at fixed spots by a robotic arrayer. The known DNA fragments are selected from a cDNA library prepared from reverse transcription of mRNA from various tissues and organisms. Ideally, each gene fragment should represent a unique gene or alternative splice variant. There are rat, mouse, and human arrays commercially available. Additionally, tissue-specific arrays can be constructed from a cDNA library extracted from the tissue.

In the experiment, the mRNA extracted from the tissue cell sample under study is purified, reverse transcribed into cDNA target, and labeled with green or red fluorescent dyes for the treatment or reference sample, respectively. Labeled cDNA hybridizes to the spots containing complementary sequence probes on the array based on base-pairing rules. After hybridization, the fluorescent signal intensities are imaged using a scanner. Two intensities, one for each dye, are measured on each spot. The intensities are considered as surrogates for the expression levels of genes in the sample under study. Some experiments may use only a single fluorescent dye approach; therefore, one intensity is measured on each spot.

The process of transforming the fluorescence intensity of each spot into a measure of transcript abundance is referred to as image analysis. In addition, image analysis includes evaluation of the quality of the quantified spot intensities and flagging unreliable spots or arrays. Image analysis can be separated into three steps: (i) image acquisition, (ii) spot location, and (iii) intensity extraction. First, the microarray is scanned to produce two images, one for each color. These images are stored as a 16-bit tag image file format (TIFF) file (a digital record of fluorescence intensities of pixels for the array), representing a number between 0 and $2^{16}-1$. As soon as the image has been captured, the next step is to locate each spot on the array. This can be performed automatically by the image analysis software. The image analysis software must segment the pixels as foreground (feature) or background. Once the foreground pixels and background pixels have been identified, the spot intensities are calculated using the mean (or median) intensity of all foreground pixels and, subtracting from them the mean (or median) intensity of the background pixels. The adjusted intensities are the primary data for subsequent analyses. The background-corrected intensity may be negative for some spots. Fig. 1 depicts the scheme of a two-color microarray experiment.

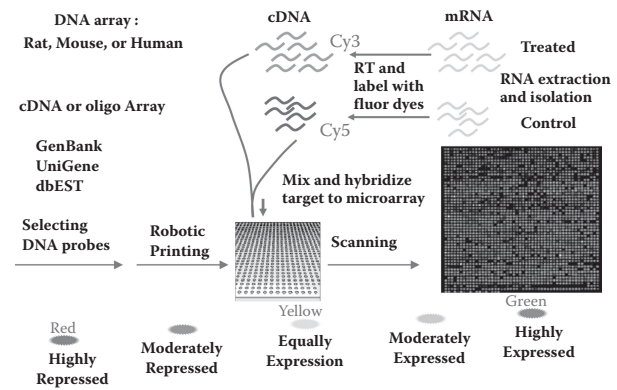


Fig. 1 Two-color DNA microarray experiment

Oligonucleotide Microarray

The oligonucleotide array technology was introduced by Lockhart et al.^[2] and Lipschutz et al.,^[3] and is commonly known as Affymetrix GeneChip™ (Santa Clara, CA). This technology does not use gene fragments extracted from a sequenced cDNA library as in cDNA microarray. Typically, each gene will be represented by 16 to 20 pairs of oligonucleotides referred to as probe sets. Each probe pair consists of 25 base sequences. One is perfectly complementary to a portion of a given transcript referred to as a perfect match (PM) probe. Each PM probe is paired with a mismatch (MM) probe that is created by changing the middle (13th) base pair to control for the possibility of cross-hybridization. RNA samples are prepared and labeled with fluorescent dye. Arrays are scanned, and images are produced to obtain an intensity for each probe. These intensities indicate how much hybridization occurred for each oligonucleotide probe. The level of expression of each gene is calculated as an “average” of the difference between PM and MM, via a procedure provided by Affymetrix through their scanning software. The Affymetrix array is regarded as a one-color platform.

The Affymetrix arrays are constructed on a silicon chip by photolithography and combinatorial chemistry. The target labeling is performed using amplified RNA rather than cDNA. Hybridization to the array is noncompetitive (only a single sample is applied to each chip), and is detected by addition of a fluorescently labeled streptavidin compound that binds to the biotin group in the aRNA molecules.^[4]

cDNA methods are more flexible in design. Researchers have control over the probe content on the array. It can be applied to any organism. The hybridization is based on over a thousand bases rather than 25 bases, thus reducing the chance of cross-hybridization artifacts. In contrast, oligonucleotide arrays can accommodate the genes not represented in a cDNA library. It has lower variability from chip to chip, and can be used by researchers for data comparison across research groups.

Other Microarray Technologies

Recently, a hybrid technology consisting of 50- to 80-mer oligonucleotides representing each gene was introduced. This technology combines the uniformity of GeneChip™ with the specificity of cDNA microarray. That is, the length of sequence is optimized to minimize cross-hybridization compared to shorter-length probes. Also, the oligo-array has higher sensitivity compared to cDNA probes.

Several other microarray-based techniques have also been introduced in recent years. Microarray-based comparative genomic hybridization (array-CGH)^[5,6] is one such technique by which variations in copy numbers between two genomes can be analyzed using DNA microarrays. Array-CGH detects genomic copy number variations more effectively at a higher resolution level than conventional chromosome-based comparative genomic hybridization (CGH). The CGH program developed by Lee et al.^[7] is a user-friendly array-CGH analyzing program that displays analyzed chromosomal data with a corresponding G-banding ideogram. The abnormal DNA copy numbers (gains and losses) can be identified automatically using a user defined window size and sequential student t-tests with sliding windows along with chromosomes.

EXPERIMENTAL DESIGN

A microarray experiment is generally a comparative experiment, in which the experiment of interest is the comparison of the relative expression levels among the samples rather than the determination of absolute intensity measures of each sample. Furthermore, the underlying principle of a two-color experiment is competitive hybridization between two samples. The measured intensities reflect the relative abundance of each sample. Data generated are inherently comparative. In a one-color experiment, the intensity is an absolute measure of gene expression; however, this measure should not be regarded as the precise measure of transcript abundance. Instead, inferences are made about the expression levels for a gene in different samples, not about the level of expression of one gene relative to other genes.

Variability, Replication, and Experimental Unit

The variation of the measured gene expression data can be generally classified into three categories: biological variation, technical variation, and residual variation.^[8,9] Biological variation refers to the variation that arises from different RNA sources, such as different animals or different cell lines or tissues. It reflects differences in host characteristics. Biological variation is due to inherent differences in gene expression, varying from subject to subject due to genetic or environmental factors. Biological variation is estimable only when there are independent

biological replicates. If all biological samples are pooled, the biological variation is minimized, and inference can be made only as to the experimental conditions. The sources of variation associated with the assay of mRNAs in each biological sample and array manufacture are collectively referred to as technical variation. The technical variation accounts for the variation associated with the use of microarray techniques unrelated to the biological samples. The residual variation accounts for sampling or experimental variation or other unaccountable factors. The biological, technical, and residual variations are mutually independent. The variation in a measured intensity is the sum of these three variations.

Experimental design methods are used to identify and minimize experimental variations and provide sound statistical inference. The basic principles of experimental design are randomization, blocking, and replication. Replication is essential in the experimental design. There are two types of replications:^[10] biological replication and technical replication. Biological replication refers to hybridizations that involve mRNA from different extractions. Biological replication is used to allow the generalization of experimental results from sample to population. Technical replication refers to replication in which the mRNA is from the same pool (the same extraction). Technical replication is used to minimize technical artifacts. In microarray quality control experiments, technical replication may be used to assess laboratory or platform reproducibility.

In most microarray experiments, the experimental unit refers to an independent biological RNA source to which the treatments are applied. The experimental unit may be represented by a tissue sample or a sample of cells from a cell culture. The array is not the experimental unit. However, in many studies, an array and an experimental unit have a one to one correspondence, so the array is not distinguishable from the experimental in such studies.

Types of Experimental Design

The choice of experimental design depends on the number of different samples to be compared as well as the aim at the comparisons of interest. Experimental design for one-color microarray experiments such as Affymetrix oligonucleotide array is similar to clinical trials or other biological experiments. The same principle used in the biomedical experiment can be applied to microarray experiments. For example, the biological samples (experimental units) should be randomized to a treatment using a pre-determined scheme. Blocking can be used to increase the precision of estimates. (A block is a subset of experimental units that are more homogeneous than the entire experiment itself.) In the two-color experiment, two samples are labeled differently for competitive hybridization. One important design issue is to determine which samples are to be labeled with a corresponding fluor, and which are to be hybridized together in the same array. In addition, there

can be constraints on the number of slides, the amount of RNA available, or cost considerations.^[9]

The simplest microarray experiment is intended to study the changes in gene expression levels between a control and a treated sample. Each microarray in the experiment is probed with two cDNA samples. The expression levels of the two samples can be compared for each gene in an array. In practice, however, the expression levels from the two measurements are not directly comparable because of dye effects. One dye may be consistently brighter than the other because of different labeling efficiencies, or different scanning sensitivities in the two dyes. Fig. 2,^[11] is a density plot (in log based-2 scale) of a same-same (control vs. control) array. That is, two control samples are labeled separately, one with Cy5 (red dye) and one with Cy3 (green dye). Also, the upper half and lower half of the array are duplicate. In Fig. 2, the solid lines represent the genes from the upper half, and the dotted lines represent the lower half. The dark and light curves represent the red and green fluorescent readings, respectively. The plots show that there were systemic location effect differences in dye intensity across the individual DNA probes. (The left tail from the low intensity spots represents the empty genes.)

To adjust for dye biases, the so-called dye-swap design with two arrays is often used. On array 1, the control

sample is assigned to the green dye, and the treated sample is assigned to the red dye; the dye assignments are reversed on array 2. Dye-swap design is a loop design (described below) for the experiment with a control and a treatment study. This design is useful to reduce dye biases, but it is unlikely to eliminate the bias in every spot of every array. Dye bias and variation can be reduced via the normalization method (see the section “Normalization”).

Many experiments involve more than one treatment sample. Each experiment will have its own study objectives. Some studies may be interested in the comparisons of differences in the expression profiles among different toxins or different cell lines. In other studies, different samples may come from different time points or different dose levels. The main objectives of these two types of studies can be different. Kerr and Churchill^[12] described the two designs for a single factor study: (augmented) reference design and loop design.

In the reference design, all samples of interest (control and treatments) are hybridized on different arrays labeled with the same color dye, while a reference sample labeled with the other color dye is used on every array to hybridize with either a control or a treatment sample. One can denote a dye assignment in which the first sample is labeled with red and the second sample labeled with green as R/G. A reference design for an experiment with four groups—T0, T1, T2, and T3—can be expressed as [T0/Ref, T1/Ref, T2/Ref, T3/Ref]. The reference sample (Ref) can be the same as the control sample (T0) or a universal reference sample. This design has several advantages. The experiment is easy to conduct. The relative expression levels of the treated to control can be directly computed as observed responses. Because each array consists of one measurement of relative change of the sample to a common reference for each gene, statistical methods are readily applicable. In this design, treatment effects are confounded with dye effects. It assumes there are no treatment $\times$ dye interaction effects.

Reference design can be inefficient because fully half of the data are dedicated to an extraneous sample. This type of design is an indirect comparison in which the reference sample serves as an intermediate between two treatments. Suppose the variance of a single log comparison, $\log(T_i/\text{Ref})$, is σ^2 , then the variance of the comparison $\log(T_i/T_j) = \log(T_i/\text{Ref}) - \log(T_j/\text{Ref})$ is $2\sigma^2$.

In the loop design, the control and each treated group are labeled once with the red and once with the green. A loop design for the four groups—T0, T1, T2, and T3—can be [T0/T1, T1/T2, T2/T3, T3/T0] or [T1/T0, T2/T1, T3/T2, T0/T3]. The loop design uses the same number of arrays as a reference design does, but it collects more information on the treatments. The loop design compares two adjacent treatments directly. The variance between two groups hybridized in the same array $\log T_i/T_{(i+1)}$ is σ^2 (but with only one array). Yang and Speed^[7] provided more details on the variance of reference design and loop design, and the time-course experiments and multifactorial designs.

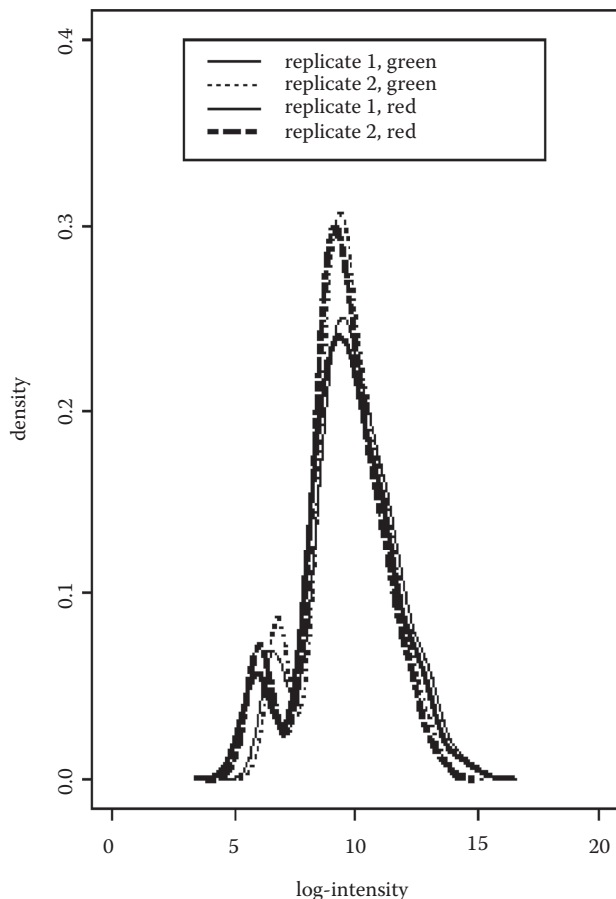


Fig. 2 Density plot of a same-same array

Sample Size Estimation

A fundamental goal in microarray studies is to identify a subset of genes that are differentially expressed under experimental conditions of interest. Before conducting a microarray experiment, one important issue that needs to be decided is the number of arrays (replicates) required in order to have adequate power to identify differentially expressed genes. Lee et al.^[13] suggested three replicates in view of the noise-to-signal ratio. Statistical tests to determine the effects of a treatment on different biological populations should be based on the biological replicate samples. In many studies, an array and a biological sample have a one to one correspondence, the array is not distinguishable from the biological sample. Nevertheless, sample size estimation refers to the number of biological samples needed for the experiment to identify a treatment effect.

Sample size estimation for a single gene is formulated as the number of arrays needed in order to ensure the power $(1 - \beta)$ of detecting the mean difference δ (standardized effect size) at a pre-specified significance cutoff α , where the standardized effect size refers to the desired change of a gene to be detected divided by the estimated standard deviation. For the two sample t -test given α and δ , the number of samples needed in each group in order to achieve the desired power is $n = 2(t_{\alpha/2} + t_{\beta})^2/\delta^2$, where t_{α} and t_{β} are the corresponding percentiles of a t -distribution.

In microarray experiments, identification of differentially expressed genes involves multiple tests of many genes. Additional factors need to be considered: the total number of genes m , the (expected) number of truly differentially expressed genes m_1 (or the proportion of the differentially expressed genes $\pi_1 = m_1/m$), the standardized effect size δ of each gene, the correlation structure among genes, the multiple testing procedure to be used, and pre-specified levels of type I error rate and “power.” The parameters m , m_1 , and the (standardized) effect size $\delta = (\delta_1, \dots, \delta_{m_1})$ for all differentially expressed genes are pre-specified by the investigator. The type I error rate can be specified in terms of the family-wise error rate (FWE),^[14,15] false discovery rate (FDR),^[16–19] or comparison-wise error rate (CWE).^[19–22] The FDR q^* is commonly used in the microarray data analysis. The “power”^[14–23] is commonly referred to as the proportion of truly differentially expressed genes that are correctly declared, the sensitivity. The desired sensitivity is denoted as λ_0 . Given m , $\pi_1 (= m_1/m)$, and effect size $\delta = (\delta_1, \dots, \delta_{m_1})$, the sample size estimation commonly is formulated as the number of arrays needed to detect the specified sensitivity λ_0 , *on average*, at the specified type I error q^* . This formulation of sample size problem is inadequate because the probability of achieving the specified sensitivity can be lower than 50%.

Wang and Chen^[15] and Tsai et al.^[19] formulated the problem as the number of arrays needed to achieve the specified sensitivity λ_0 with a coverage probability ϕ_{λ_0} , where ϕ_{λ_0} is the probability that *at least* λ_0 proportion of

truly differentially expressed genes is detected. In this formulation both λ_0 and ϕ_{λ_0} need to be specified, they are not necessarily equal. The ϕ_{λ_0} can be set at 95%. Tsai et al.^[19] provided a method to estimate the needed sample size for the independent and constant effect size ($\delta_i = \delta_0$ for all i) model as follows.

Given α , δ_0 , and $(1 - \beta)$, the needed sample size n for detecting a differentially gene is

$$n^* = \frac{2(t_{\alpha/2} + t_{\beta})^2}{\delta_0^2} \quad (1)$$

Under the independent model, the total number of truly differentially expressed genes detected U is a binomial random variable with success probability $(1 - \beta)$. The probability ϕ_{λ_0} of identifying at least λ_0 fraction of m_1 differentially expressed genes is the sum of the binomial probabilities^[19]

$$\phi_{\lambda_0} = P(U/m_1 \geq \lambda_0) = \sum_{l=[m_1\lambda_0]}^{m_1} \frac{m_1!}{l!(m_1-l)!} (1-\beta)^l \beta^{m_1-l} \quad (2)$$

For specified q^* and λ_0 , sample size can be estimated by first solving for the constant comparison-wise power $(1 - \beta^*)$ from the equation,

$$\phi_{\lambda_0} = \sum_{l=[m_1\lambda_0]}^{m_1} \frac{m_1!}{l!(m_1-l)!} (1-\beta^*)^l \beta^{*(m_1-l)} \geq 95\%. \quad (3)$$

The calculated power $(1 - \beta^*)$ is then used to estimate the α level corresponding to the specified q^* , $\alpha = [m_1(1 - \beta^*)q^*]/[m_0(1 - q^*)]$. Subsequently, the sample size n is calculated from Eq. 1 by using α and $(1 - \beta^*)$.

Gene expression data are typically positively correlated, which leads to an over-dispersion (relative to the independence) on the distribution of the true positive findings. Therefore, the methods based on the independent model described above will underestimate the true needed sample size. Tibshirani^[22] proposed a permutation method to account for both dependency and unequal effect sizes among genes using a pilot data set. Because the sample size of the pilot data is typically smaller than the needed sample size, the null distributions generated from the pilot data have more variations; simply using the null distributions generated from the pilot data can overestimate the needed sample size. Recently Lin et al.^[23] proposed a permutation procedure modified from the Tibshirani^[22] method. Their method takes the correlation structure and effect size heterogeneity into consideration by using a small pilot data set, four to six samples per groups.

DATA PREPROCESSING AND NORMALIZATION

After data collection, many factors in generating intensity measurements need to be considered prior to data analysis.

There are low signal intensities or inconsistent spots across arrays. In addition, the intensities contain background contribution that can be addressed in diverse ways. Some arrays may have multiple probes that measure the same gene; the intensities from these probes can be combined to generate a single expression level for the gene. Data pre-processing is the first step of data analysis, it attempts to correct technical biases. There are many options and methods for data filtering, local and regional backgrounds, multiple probes, and normalization and transformation. In particular, normalization has been an active research area in microarray data analysis since the microarray technology was introduced.^[24–32] Normalization methods for the Affymetrix arrays have been proposed including dCHIP,^[29] RMA,^[30] Affymetrix Microarray Suite 5.0 (MAS 5), and Affymetrix Probe Logarithmic Intensity Error (PLIER) http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf.

The primary goal of normalization is to properly eliminate or minimize the systematic variation, such as microarray construction in probe printing, sample preparation procedures, hybridization and washing procedures, detecting method, and so forth. Normalization strategies depend on the experimental design and the data collection process. The primary consideration is to determine the (sub)set of genes for use as the baseline for a normalization method. That is, normalization can be based on either an entire set of genes or a certain subset of genes, such as housekeeping genes that are not regarded to change the expression level under the experimental condition or some added set of control genes. Two assumptions are commonly made on normalization: (i) most genes do not change their expression level; and (ii) the total intensity across all spots in the array should be the same. The global normalization method uses all genes on the array for adjustment; it uses the mean or median as the normalization factor based on the two assumptions. For each array, the data are normalized by dividing individual gene intensity by the mean or median intensity of the array. However, the global mean/median normalization method fails to account for spatial- and intensity-dependent biases that have been observed in many experiments.

Two known statistical methods proposed for normalization are the analysis of variance (ANOVA) model^[26,27] and the scatter-plot smoother “lowess” fit.^[31] Although both procedures consider two-color fluorescence array, they can be applied to one-color data with modifications. The ANOVA procedure^[27] recommends modeling the logs of individual red and green intensities. The ANOVA model uses mean estimates to perform global adjustments for array, dye, treatment, gene effects, and their appropriate interactions simultaneously. However, as in the case of global normalization, the ANOVA method is a constant adjustment, with respect to specific effects such as dye biases. Alternatively, the lowess fit^[31] uses local regression to account for intensity- and spatial-dependent dye biases. The lowess fit is a scatter-plot smoother that performs

robust locally linear fits. Yang et al.^[31] recommended performing an array-by-array normalization on the log-ratios instead of the log-ratio of two intensities. Unlike the ANOVA approach, this lowess approach fits different lowess curves for different arrays. These two procedures can be formulated by a generalized additive model.^[32]

In a recent study, in order to minimize the non-concordance of inter-laboratory measurement of Affymetrix GeneChip™ results, Lee et al.^[33] demonstrated that labeling extension values (LEVs), which are correlation coefficients between probe intensities and probe positions, are significantly correlated with the gene expression levels on eukaryotic Affymetrix microarray data. LEVs have proven to be useful in exploring laboratory signatures, which represent the systemic variations uniquely produced in each laboratory during the microarray procedures from probe labeling and hybridization, to signal detection. Lee et al.^[33] Treated LEVs can be used as a filtering parameter for improving inter- and intra-laboratory comparability of gene expression profiles without compromising subsequent functional analyses of biological networks on them.

IDENTIFYING DIFFERENTIALLY EXPRESSED GENES AND GENE SET ANALYSIS

A microarray experiment is conducted to study the changes in gene expressions under different experimental conditions of interest. The simplest microarray experiment is performed to identify a subset of genes that are differentially expressed between control and treated groups, for example, between drug treatment and no drug treatment. The fold-change has been used as a criterion to identify differentially expressed genes in early microarray studies as the number of replicates was small or none. In the fold-change approach, a gene is said to be differentially expressed if the ratio in absolute value of the expression levels between the two conditions exceeds a certain threshold, for example, two-fold or three-fold change. The fold-change has been known to be a poor choice for identifying differentially expressed genes.^[34,35]

Identifying Differentially Expressed Genes (Gene Selection)

Identification of differentially expressed genes can be separated into two steps.^[34] The first step is to calculate a discriminatory score that will rank the genes in order of evidence of differential expressions. The second step is to determine a threshold cutoff from the ranked scores to select the differentially expressed genes.

Gene Ranking

The data are typically transformed in log₂ scale after normalization. Differential expression can be evaluated by

comparing the difference between the mean expressions of the two conditions. The difference, denoted by M , represents the fold-change between two samples. Because of a wide range of magnitudes and variability among different genes in the array, the fold-changes computed for different genes are not directly comparable as measures of evidence of differential expressions. The M statistic can be standardized by dividing a scale parameter s (e.g., standard deviation), denoted by $T = M/s$. The T statistic has the general expression that represents the test statistics for two sample comparisons. The T statistic includes the fold-change M statistic as special case by setting $s = 1$; and T becomes Student's t -statistic when s is the sample standard deviation of the mean difference. Various T statistic variants have been proposed in microarray data analysis.^[36–44] Tusher et al.^[36] proposed the SAM (<http://www-stat.stanford.edu/~tibs/SAM/>) statistic $M/(s_0 + s)$ by adding a penalty s_0 to the sample standard deviation in the denominator to account for a very small standard deviation which results in a large t -value, where s_0 is computed from the data. When the sample size is small, the variance estimate s^2 may be imprecise. To improve the reliability of variance estimates, a number of shrinkage estimators have been proposed.^[41–44] The variance of a shrinkage estimator has the general form: $s = ws_1^2 + (1 - w)s^2$ where s_1^2 is a variance of M based on all genes in the arrays and $0 \leq w \leq 1$ is a weight depending on the sample size. These methods seem to work well and they are particularly useful when the sample size is small.

Assigning a Significance Level

The T statistic is a measure for gene ranking, but it does not provide a criterion to determine a cutoff. Statistical significance testing is designed to provide a cutoff by computing the p -value from a statistic. The p -value of a test statistic is computed either from the distribution of the test statistic or from a re-sampling method. Generally, the distribution of the normalized intensities is not normal. The permutation tests provide an alternative method to determine the significance, particularly for the test statistics that do not have a closed mathematical expression, such as SAM or shrinkage statistics; their probabilities cannot be computed directly. The principle of the permutation test is conceptually straightforward. It first computes the empirical value of the test statistic for the observed data, and then performs all possible permutations under the null hypothesis and computes the test statistic for each permutation. The p -value of the test is calculated by summing the probability of the observed outcome and the probabilities of all outcomes with the test statistic values greater than the observed value. Permutation tests do not require assumption on the distributions; they are easy to perform using high-speed personal computers or workstations. However, the test is not feasible when the number of replicates is fewer than six.^[40]

A p -value cutoff divides the genes into two sets. Ideally, the selected set would contain the differentially expressed genes and the non-selected set would contain the non-differentially expressed genes. However, the selected gene set will have false positives; likewise, the non-selected genes will have false negatives. The false-positive probability or p -value is computed under testing a single gene. Since hundreds or thousand of genes are tested, determining a cutoff should be considered in terms of an overall false-positive error. This issue is referred to as multiple hypotheses testing or testing multiple endpoints. The familywise error (FWE)^[45–47] measure is commonly used in testing multiple endpoints when the number of tests is small. The FWE approach could present a problem in the analysis since the analysis tends to screen out all but a handful of genes that show extreme differential expression if m , the total number of genes, is large. With a large number of genes involved in the comparisons, the false discovery rate (FDR) error measure^[48–50] is a more useful approach for determining a significance cutoff. The FDR error measure considers the *expected proportion of false positives among the selected genes*.^[48] Essentially, FDR considers the probability of false selections among those selected genes. The FDR approach allows the findings to be made, provided that the investigator is willing to accept a small fraction of false-positive findings.

In the FDR approach, the genes are ranked according to the p -values:

$$p_{(1)} \leq \dots \leq p_{(r)} \leq \dots \leq p_{(m)}.$$

The notation $p_{(i)}$ represents the i -th ordered p -value, and m is the number of genes in the analysis. For example, $p_{(1)}$ is the smallest, $p_{(2)}$ is the second smallest, and so on. For the desired FDR level, an FDR-controlled procedure^[51] is to find the largest $p_{(r)}$ such that $(m \cdot p_{(r)})/r \leq \text{FDR}$. Those r genes with p -values less than or equal to $p_{(r)}$ are selected as differentially expressed genes. Conversely, if r genes are selected, then an empirical FDR estimate is $(m \cdot p_{(r)})/r$.^[50] If the number of true null hypotheses, m_0 , were known, an improved FDR estimate, $(m_0 \cdot p_{(r)})/r$, could be used. Since m_0 is not typically known, m is used in practice to ensure proper control of error rates. Several statistical methods for estimation of the number of true null hypotheses have been considered by Hsueh et al.^[52] which could be used to gain additional rejections.

Both the FDR and FWE approaches emphasize controlling the false-positive error. Because of a large number of genes and an effort to maintain a low significance level, an FDR approach often results in a short list of significant genes. Some genes that have good prediction powers might not be captured in the list. In these genomic/genetic applications, both the false-positive error and false-negative error are of concern. Delongchamp et al.^[53] proposed a ROC (receiver operating characteristic) approach to determining an optimal cutoff based on minimizing the total

cost from making false-positive and false-negative errors. This approach requires estimates of m_0 and m_1 . Chen et al.^[34] illustrated an application of the ROC approach to determining a p -value cutoff.

Gene selection in many applications often targets a limited number of candidate differentially expressed genes for confirmation and further study. The differentially expressed genes so selected are referred to as statistically significant. However, if an observed change is small with a very small standard deviation, then the change can be identified as statistically significant even if it may not be significant “biologically.” In a two sample comparison, it is common to use the p -value to decide a cutoff and then focus on the genes that pass the cutoff with a higher fold-change comparison. This analysis can be plotted by the so-called volcano plot.^[34] The approach of using the p -value cutoff as the primary criterion followed by the fold-change, provides the control of false-positive error and, in the meantime, preserves the desired biological significance.

The T and T variant statistics described are for two experimental conditions; many microarray experiments involve multiple experimental conditions. The different experimental conditions can be dose levels, time points, or treatment combinations. Comparisons among several treatments or factorial designs are analyzed in linear model framework. Kerr et al.^[27] proposed a single ANOVA model for an entire microarray experiment. Jin et al.^[54] fitted separate ANOVA models for each gene. Tsai et al.^[32] proposed a generalized ANOVA model and performed the analysis for each gene. The p -value of the test statistics can be computed using the permutation method.

A microarray experiment can generate different gene lists by different filters, normalizations, analysis methods, or study objectives. In some studies, it may be of interest to select a subset of genes from the multiple gene lists. Consider a mouse experiment to study differences in gene expression among the three p53 genotypes: wild-type (+/+), knock-out (-/-), and heterozygous (+/-). A statistical analysis typically consists of a comparison among the three genotypes. An important follow-up analysis is the comparisons between the knock-out and wild-type mouse and between the heterozygous and wild-type mouse. The Dunnett's test is frequently used to generate the differentially expressed gene lists for the two comparisons. Often, it is also of interest to determine the genes that show differences in both comparisons. Chen et al.^[55] recently proposed layer ranking algorithms to provide a single preference gene list from multiple gene lists generated by different ranking criteria.

Gene Set (Enrichment) Analysis

Gene set analysis (GSA) is a statistical approach to determine whether a functionally related set of genes is expressed differently (enrichment and/or deletion) under different experimental conditions. A gene set refers to a group of genes with related functions based on biologically

relevant information, such as a metabolic pathway, protein complex, or GO (gene ontology) category.

The commonly used approach to GSA is first to identify a list of genes that are expressed differently between two experimental conditions using a statistical significance testing described above. After selecting a list of differentially expressed genes, the list is then compared to biologically pre-defined gene sets to determine whether any sets are over-represented in the list of differentially expressed genes compared to the pre-defined gene sets.^[56–58] This process of comparing the number of genes in the differential expression list with the number in a predefined gene set is known as over-representation analysis (ORA). The Fisher's exact test is typically used to assess the significance for an over-representation.^[56] The ORA approach has several shortcomings.^[59,60]

Mootha et al.^[61] and Subramanian et al.^[62] proposed Gene Set Enrichment Analysis (GSEA) to identify a significantly different gene set between the diabetic samples and normal muscles for which the ORA failed to identify any single gene set. The key concept is that the individual genes in a gene set are closely related and often have similar expression patterns. By borrowing strength across the gene set, GSA will increase statistical power. The GSA approach essentially shifts the level of analysis of the microarray experiment from single genes to sets of related genes. The work of Mootha et al.^[61] has inspired the development of various GSA methods for alternatives to the ORA approach.

In GSA, the null hypothesis is all the genes in the gene set are not associated with the experimental conditions. The alternative hypothesis can be either a one-sided or two-sided test. The one-sided test means that gene expression changes in the gene set are in one direction: either all up or all down. The basic idea in this analysis is that the gene sets are closely related and, hence, will have similar expression patterns, either up or down. The two-sided test means that the gene class can exhibit both increased and decreased gene expression.^[60]

There are many one-sided GSA statistics.^[59–63] But, most GSA statistics are for two-sided tests.^[64–67] A typical GSA approach can be summarized as follows: (i) all genes are ranked by computing a test statistic (or p -value) that measures the association between the expressions and the treatment conditions; (ii) for each gene class or functional category, a class score is calculated as a summary measure of the class; (iii) resampling methods are used to generate the null distribution of the class score for each gene class; and (iv) statistical significance is assessed by comparing the observed functional score to the percentile of the null distribution of the gene class. By considering the distribution of the entire set of genes, this approach is more powerful and interpretable. The success of a gene class testing requires development of multivariate statistics and computational algorithms for testing hundreds of variables, and a well-characterized gene class mapped to microarray probes.

CLASSIFICATION AND PREDICTION

Class prediction aims at developing a prediction model that accurately predicts the class membership of a new sample from the available gene expression data set. The prediction models can be used to discriminate between different biologic phenotypes or to predict the diagnostic category, prognostic stage, or treatment response of a patient. Development of a prediction model involves two components: (i) model building and (ii) performance assessment. Typically, the data are divided into a training set and a test set; the prediction model is developed on the training set and then is used to classify samples in the test set to assess its predictive accuracy.

Model Building

The model building step can be summarized into four components: (i) selecting feature predictors; (ii) selecting a classification algorithm; (iii) specifying the training parameters of the prediction model; and (iv) fitting the prediction model to the training samples.

Because of the large number of predictors involved, many predictors are often noisy in nature and irrelevant for prediction. The use of all predictors to build the prediction model can suppress or reduce the performance of a classifier. Selection of a subset of the most informative predictor variables, called feature selection, is commonly addressed in the development of classification algorithms.^[68,69] Feature selection is the most important task in the model building step with high dimensional data. Depending on classification algorithms, there are two general approaches to feature selection: filters and wrappers. The filter approach removes irrelevant predictors according to some pre-determined criterion such as a *p*-value. Classical classification algorithms, such as Fisher's linear discriminant analysis and *k*-nearest neighbor, often use the filter approach to select relevant predictors prior to classification.^[70–73] The filter approach is a stand-alone prior step, regardless of which classification algorithm will be used. The affects of the selected predictors on the performance of the algorithm are not taken into account. The wrapper approach finds a subset of predictors and evaluates its relevance while building the prediction model. The classification algorithms, such as stepwise logistic regression, classification trees,^[74] and support vector machines with recursive feature elimination,^[75] use the wrapper approach.

The second task in the model building is the selection of a classification algorithm. A prediction model is a mathematical function constructed on the training samples from the selected classification algorithm. Numerous classification algorithms have been proposed for applications to microarray data. Several authors have compared various classification algorithms including several variants of linear discriminant analysis, nearest neighbor classification,

logistic regression trees, partial least square analysis, neural network algorithms, naïve Bayes algorithms, shrunken centroids, several variants of classification trees, regression trees, and support vector machines.^[76–78] After selecting the algorithm, the prediction model needs to be specified before model fitting. Model specification means specifying all aspects of the model including feature selection criteria, specific functional form of the prediction model, approach and criterion for performance assessment, and so forth. For example, in many classification algorithms, a cut-point to translate a quantitative predictive index into a class label must be specified. For example, the logistic regression typically uses 0.5 as a cut-point to separate two classes based on an equal misclassification cost. After selecting the prediction algorithm to be used, the prediction model is fitted to the training data set. The objective is to search for a prediction function that optimizes the class separation. In model fitting, the parameters of the prediction model are estimated to minimize the probability of misclassification error when equal misclassification costs are assumed.

Performance Assessment

In the development of a prediction model, the most important question is the ability of the model to predict a future sample. To ensure an unbiased assessment of accuracy, the prediction model is developed in one data set and then applied to another data set to estimate the predictive accuracy. The most straightforward method of estimating the accuracy is the split-sample validation in which one portion of the data is held out (test data set) while the classifier is being developed on the remaining data (training data set) and then is tested on the test data set. In practice, data often are insufficient to split into a training set and a test set for validation. Instead, cross-validation is used to evaluate the performance of a prediction model.

Cross-validation involves repeatedly splitting the data into a training set containing most of the samples and applying the prediction rule to the test set made up of the remaining samples. The prediction accuracy estimate is the average accuracy of the numerous training-test partitions.

In cross-validation, the test data must be completely independent of the training data from which the prediction model is built. The cross-validation needs to repeat all steps of model development, including gene selection and model construction within each stage. In using the filter approach, gene selection must be conducted in the training set to avoid selection bias.^[79,80] The cross-validation estimate is, presumably, the expected accuracy of the developed classifier to predict future samples using the whole data set. Since cross-validation only uses a fraction of samples, its estimates may depend on the number of folds used. As discussed, the accuracy estimate from the leave-one-out cross-validation may be expected to be higher than the estimate from the two-fold cross-validation; also, the

leave-one-out estimate may better reflect the accuracy rate estimated from the whole data set since it uses almost the entire data set in training the classifier.

CLUSTER ANALYSIS AND PATTERN RECOGNITION

One important goal in the analysis of gene expression data is to determine the relationships between genes or gene clusters for identifying biological functions or predicting specific biological outcomes (or diseases) from the analysis of expression patterns. Clustering and visualization methods have been developed and applied to organizing gene expression data by grouping genes with similar patterns of expression. Usually, genes with previously unknown functions are simply annotated with functions of the known genes in the same cluster because they share similar expression profiles across different experimental conditions. Genes expressed with similar patterns may be coregulated by common upstream regulatory motifs or may belong to the same biochemical pathways. A cluster of unknown genes with homogeneous expression profiles may indicate the discovery of a novel function group.

Before any clustering and visualization procedures can be applied to the gene expression matrix, careful extraction of array images and appropriate normalization of system variations are necessary. A screening (data filtering) procedure is often applied in selecting a subset of genes that satisfy certain criteria of expression patterns across all arrays (treatments) for further statistical analyses. A commonly used criterion is to include genes with the expression $|\log_2 \text{Trt/Ref}| \geq 1$ in at least some experimental conditions. Minimum variation (standard deviation or range) criteria are also frequently applied. The screening process is needed for reducing the number of genes from tens of thousands to thousands or hundreds for clearer visualization, faster computation, and better interpretation. In addition to array normalization and centering, individual gene normalization is often used to scale the relative magnitude of a gene's expression profile (vector) to one. Gene normalization does not only prevent clustering procedures from being dominated by genes with extreme expression patterns; they also prevent visualizations from losing color resolution.

In this section, a gene expression time course data set from a serum stimulation of primary human fibroblasts^[81] is adopted to illustrate concepts of statistical visualization and clustering tools in gene expression analyses. The experiment used cDNA arrays with 9,800 spots, representing approximately 8,600 distinct human transcripts. Briefly, fibroblasts were grown in culture and sera-deprived for 48 hours. Serum was added back and samples were taken during various periods: at the start (designated as 0), 15 minutes, 30 minutes, 1 hour, 2 hours, 3 hours, 4 hours, 8 hours, 12 hours, 16 hours, 20 hours, and 24 hours (with one array for each of these 12 samples). Genes were selected

for their analysis if the expression level deviated from time 0 by at least a factor of 3.0 in at least two time points. The selection resulted in a gene expression matrix of size 517 (genes) $\times$ 12 (arrays). A further randomly selected subset of 100 genes is used in the illustration.

Gene Expression Matrix Map with Hierarchical Clustering

The most frequently used visualization tool for gene expression profiles is the gene expression matrix map. There are many synonyms for gene expression matrix map in the literature. Wen et al.^[81] made one of the earliest applications of matrix map to gene expression data, and Eisen et al.^[82] laid down the fundamentals of cluster analysis of gene expression patterns with matrix map. The numerical data matrix of gene expression intensities (single-color experiment) or log ratios (two-color experiment) is projected through properly chosen single directional (single-color) gray spectrum or bi-directional (two-color) gray scales. Fig. 3a is the log ratio (base 2) gene expression matrix with the 100 randomly selected genes on 12 time course arrays. A green-black-red spectrum is used to represent negative-zero-positive log ratio values with brighter (darker) colors representing extreme (mild) differentially expressed gene-array combinations. The 100 row color strips represent 100 gene expression profiles that are randomly permuted, while the order of 12 column strips representing 12 arrays remain intact to depict the time course nature of the experiments. Very little information can be visually extracted from Fig. 3a because rows (genes) are randomly permuted. It is necessary to rearrange the 100 rows to satisfy the concept of relativity^[83] such that genes with similar (different) expression profiles are placed in closer (distant) rows before any interesting pattern can be summarized from this expression matrix map.

The most commonly used permutation mechanism for rearranging row orders (gene) and column order (array) in an expression matrix map is a hierarchical clustering tree (HCT) with a binary dendrogram representing the association structure of pair-wised genes or arrays. A proximity (similarity or distance) matrix, such as a correlation matrix, is first generated. Fig. 3b is the Pearson correlation matrix of the 100 gene expression profiles with a blue-white-red bi-directional spectrum representing the negative-zero-positive correlation coefficients. A similarity proximity matrix is usually converted into a distance matrix before the dendrogram for that matrix is constructed. The second step is to identify the pair of genes with the smallest distance and to group them with a link. Distances between the newly formed group with the rest of the 98 genes are computed resulting in a new proximity matrix of size 99 $\times$ 99. The algorithm proceeds in a recursive manner to build the tree structure step by step, with one less cluster member at each step. The final tree architecture is displayed as the left panel of Fig. 3c.

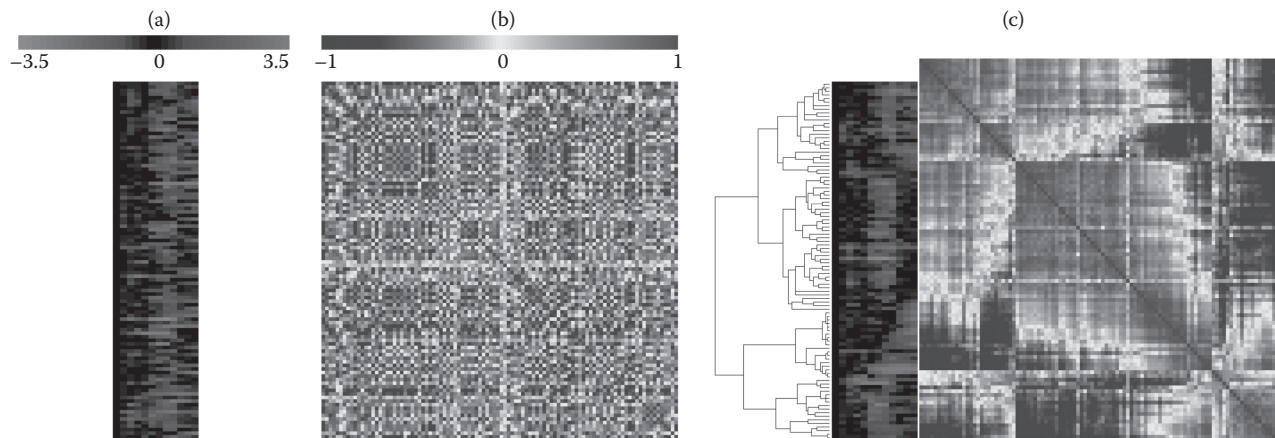


Fig. 3 Gene expression matrix map with hierarchical clustering of 100 randomly selected genes from a time course data set with serum stimulation of primary human fibroblasts.^[82] (a) Gene expression matrix map for the 100 genes on 12 time points (arrays) with a green-black-red color spectrum; (b) correlation matrix map for the 100 genes with a blue-white-red color spectrum; (c) hierarchical clustering tree with sorted expression matrix map and correlation matrix map

The relative positions of terminal leaves in this tree form a linear order of the 100 genes, which is used to sort the randomly arranged input correlation matrix map into the structured matrix in the middle panel. The randomly ordered gene expression map in Fig. 3a is rearranged into the right panel of Fig. 3c. Variants of hierarchical clustering algorithms have been developed, according to the direction (agglomerative vs. divisive) of trees that are constructed and the way new distances are computed (single linkage, complete linkage, average linkage, centroid method, and Ward’s method). Sokal and Sneath^[84] describe methodologies for constructing HCTs and non-hierarchical clustering algorithms in details.

There is one possible flip associated with every intermediate node in the tree architecture. A tree with 100 genes has 99 intermediate nodes with 2^{99} possible flip combinations and 2^{99} different permutations in rearranging the gene orders. The up-regulated half (red) of the simulated

time course data in Fig. 4 is created by shifting the peaking point of a Sine curve in the directions of left-to-right and top-to-down. The down-regulated half (green) simply reverses the process. The true underlying structure of this data is a two-component model in which each component follows a smooth global trend. With different flipping mechanisms in the five panels of Fig. 4, users may observe completely different clustering patterns without seeing the global trends. When noise is added to the original smooth transitional data, as would happen with microarray data, HCT-type tree permutations would not be able to reconstruct the global transitional pattern. We see that the clustering pattern in Fig. 4e actually mimics that of Fig. 3c very well, which suggests that similar global patterns might be embedded in the fibroblast to serum gene expression data.

External and internal references can be applied to these flips, such that certain criteria may be satisfied.^[85,86] Tien et al.^[87] proposed a flipping mechanism for a conventional

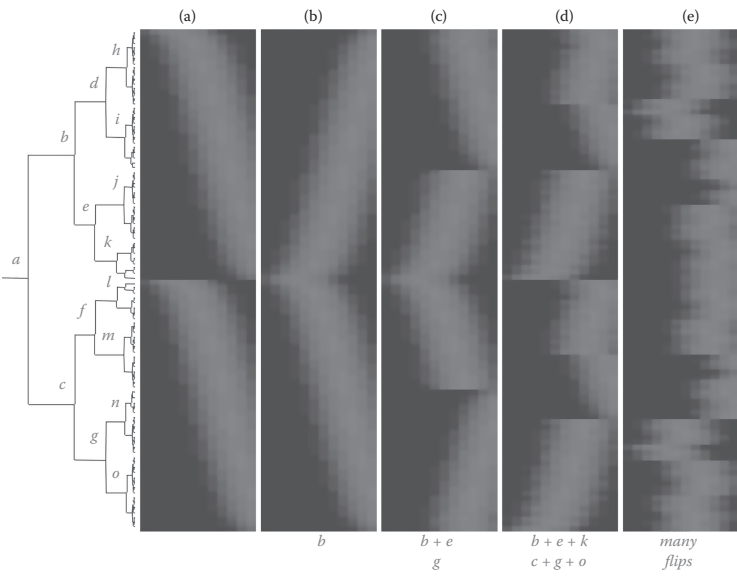


Fig. 4 Flipping effects of intermediate nodes of an HCT. (a) Original simulated data with a two-smooth-component model; (b) With one node flipped; (c) With three nodes flipped; (d) With six nodes flipped; (e) With many nodes flipped

agglomerative HCT using a rank-two ellipse (R2E, an improved singular value decomposition algorithm for sorting purpose) seriation by Chen^[83] as an external reference. Using the same 100 randomly selected genes on 12 time course arrays as in Fig. 3 an average linkage HCT with random flipping of intermediate nodes is constructed in Fig. 5a. Again, the order of 12 arrays remain intact in the gene expression map (100 genes by 12 arrays) and the correlation matrix map (12 arrays by 12 arrays) to depict the time course nature of the experiments. Many gene clusters can be inferred from the gene expression matrix map and the gene-gene correlation matrix maps (100 genes by 100 genes). Fig. 5b has the gene expression and the gene-gene correlation matrix maps sorted by the rank-two elliptical (R2E) order of the 100 genes. A very smooth global trend reflecting the time-related gene expression pattern can now be explored. The R2E order in Fig. 5b is then applied to guide the flipping mechanism of Fig. 5a with the resulting HCT and sorted expression and correlation maps illustrated in Fig. 5c. While HCTs always produce permutations with good local behavior, the rank-two ellipse seriation gives the best global grouping patterns and smooth transitional trends. The resulting algorithm automatically integrates the desirable properties of each method so that users have access to a clustering and visualization environment for gene expression profiles (as well as general high-dimensional data profiles) that preserves coherent local clusters and identifies global grouping trends.

Wu et al.^[88] integrated most commonly used HCT algorithms, R2E, and several flipping methods with other visualization features into the GAP (Generalized Association Plots) package. GAP is a Java-designed exploratory data analysis (EDA) software for matrix visualization (MV) and clustering of high-dimensional data sets. It provides direct visual perception for exploring structure of a given gene

expression profile matrix and its corresponding proximity matrices for genes and arrays. Various matrix permutation algorithms and clustering methods with validation indices are implemented for extracting embedded gene expression information. GAP is more powerful and effective than conventional graphical methods when dimension reduction techniques fail or when data is of ordinal, binary, and nominal type. It can be obtained from <http://gap.stat.sinica.edu.tw/Software/GAP>.

Non-hierarchical Clustering

K-means^[89] and self-organizing maps (SOM)^[90,91] are the two most commonly used non-hierarchical clustering algorithms in gene expression pattern analysis. In the non-hierarchical clustering procedure, genes are divided into k ($k_1 \times k_2$ for SOM) partitions or groups, with each partition representing a cluster of genes. Therefore as opposed to the hierarchical clustering, the number of clusters must be known (decided) a priori.

In K-means clustering, the user first specifies k initial cluster centroids or seeds. The proximities (similarity or distance) from each gene to all k centroids are calculated. Each gene is then assigned to the closest cluster (centroid). The k new centroids are formed with new cluster members, and the genes are reallocated to one of the new k clusters. This iterative process stops if there is no reallocation of genes, or if the reassignment satisfies the criteria set by the stopping rule. Variants of K-means algorithms differ with respect to (i) the method used for obtaining initial cluster centroids or seeds; (ii) the rule used for reassigning genes; and (iii) the proximity for computing the closeness between each gene and every centroid. Because K-means method outputs only the cluster memberships for all genes, this membership information is usually displayed in colors or symbols (Fig. 6a with $k = 5$).

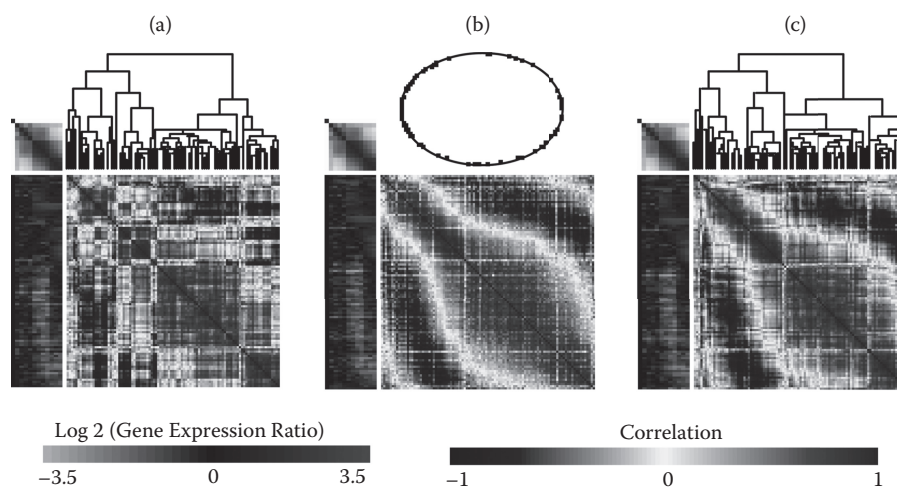


Fig. 5 Matrix visualization for expression profiles map in Fig. 3 with three sorting algorithms. (a) Matrix visualization with hierarchical clustering tree (HCT). (b) Matrix visualization with rank-two ellipse seriation (R2E). (c) Matrix visualization with R2E guided HCT (HCT_R2E)

The SOM method combines the features of dimension reduction in multi-dimensional scaling (MDS) and pre-specified number of clusters in K-means. The goal is to represent genes in the original high-dimensional input expression space by grid points in a low-dimensional output space. Commercial packages usually limit the output space to a one-dimensional grid line or two-dimensional grid net for visualization purpose. Similar to K-means algorithm, these grid points are projected to the original high-dimensional input space as centroids for the gene expression profiles to be iteratively assigned to a cluster membership. The computation of cluster membership is confined to the grid structure in the low-dimensional output space. Final cluster membership is displayed using the output space grid structure. Gene expression profiles for genes belonging to each grid point are displayed as a multivariate statistical plot at the relative grid location. Commonly used statistical plots include parallel coordinate plot, side-by-side box-plots, and gene expression matrix map. An SOM analysis with a two-dimensional 3×3 grid and box-plots for the 100 randomly selected genes is illustrated in Fig. 6b.

Simulation studies have shown that K-mean algorithms and other non-hierarchical clustering algorithms perform poorly when random initial seeds are used. Their performance is improved when the results from hierarchical methods are used to form the initial partition.^[92] In other words, hierarchical and non-hierarchical techniques should

be applied as complementary clustering techniques rather than as competing techniques. The K-means or SOM output membership information can be used in dimension reduction plots, principal component analysis (PCA), or MDS to be discussed next.

Dimension Reduction Analysis and Visualization

There are two major statistical techniques for simultaneously visualizing thousands of gene expression profiles in a single display—dimension reduction visualization and dimension-free visualization. Because of its high-dimensional nature, dimension reduction techniques such as PCA^[93,94] and MDS^[95,96] are often applied to gene expression matrices for projecting the original high-dimensional structure onto a lower dimensional space (usually two or three) for visualizing and computational purposes. Dimension reduction visualization is often adopted for presenting gene grouping structure for methods such as K-means and SOM.

For a given gene expression matrix X of size g (genes) by a (arrays), PCA performs an eigenvalue decomposition of the $a \times a$ covariance matrix $\Sigma = X^*X/g$ as $\Sigma \mathbf{v}_k = \lambda_k \mathbf{v}_k$, where $\mathbf{v}_k$'s are the eigenvector of Σ with eigenvalue λ_k , $k = 1, \dots, a$. The PCA summarizes the dispersion of gene expression profiles as data cloud in a small number of major axes (principal components) of variation among the arrays. Alter et al.^[93] named these major axes as the eigenarrays

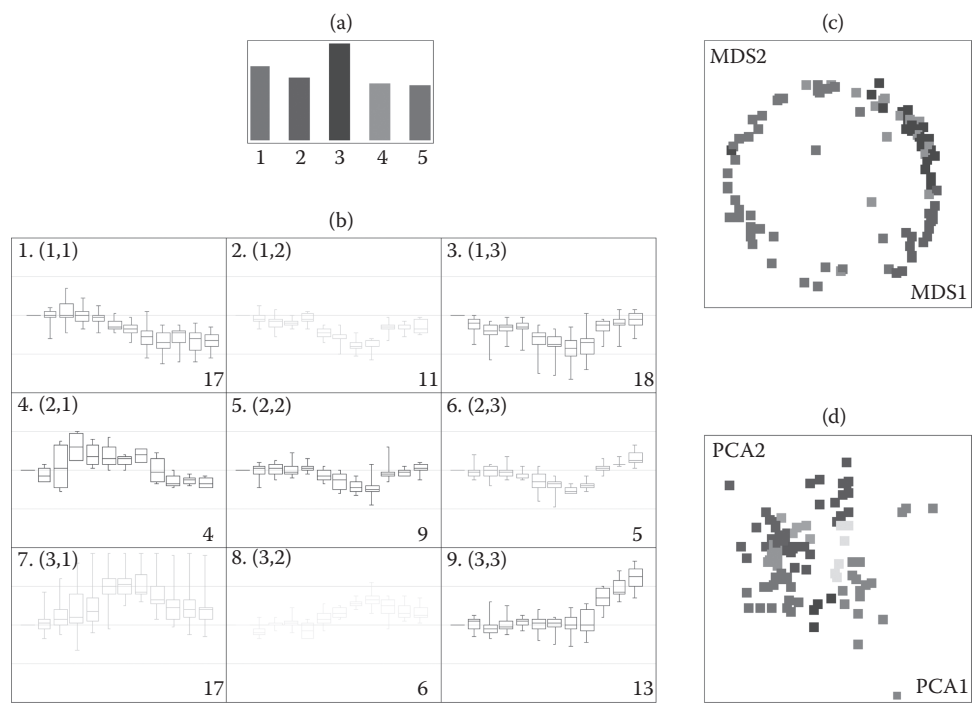


Fig. 6 K-means clustering and self-organizing maps (SOM) with multidimensional scaling (MDS) and principal component analysis (PCA) for the time course data set with 100 genes. (a) K-means ($k = 5$) clustering with cluster membership by symbols; (b) SOM with a 3×3 grid structure, expression profiles for genes grouped at each grid point are displayed as box-plots with cluster membership by symbols; (c) MDS with K-means membership symbols; (d) PCA with SOM membership symbols

of the original expression matrix. They normalized the arrays by filtering out the eigenarrays that were inferred to represent noise or experimental artifacts. Hastie et al.^[97] recursively applied the PCA to gene expression matrices for identifying subsets of genes with coherent expression patterns and large variation across conditions. Fig. 6d is the plot of the first two principal components of the 100 by 12 gene expression profiles with colors correspond to gene clusters identified by the SOM analysis in Fig. 6b.

Given a set of n genes (or n arrays), among which the associations are to be studied, MDS first computes proximity (distance or similarity measurement) for all pairs of objects (genes or arrays) as δ_{rs} , $1 \leq r, s \leq n$. MDS then converts the input proximities into disparities $\hat{d}_{rs}$, $1 \leq r, s \leq n$ with a transformation function f . The function f can be the identity function if proximity δ_{rs} itself is a Euclidean distance. MDS finally represents objects as points in a Euclidean space, so that the perceived distances d_{rs} between points can reflect proximities δ_{rs} between objects. For visualization purposes, the dimension of the projected space is usually kept as low as possible. Thus it is unavoidable that part of the information in the original proximity matrix of similarity (or dissimilarity) will be lost in the MDS configuration plot. Kruskal^[98] proposed a scheme for measuring how much a lower-dimensional geometrical representation falls short of a perfect match. This measure, called STandardized REsidual Sum of Squares (STRESS), is defined as
$$\text{STRESS} = \left\{ \frac{\sum_r \sum_s (d_{rs}^{(q)} - \hat{d}_{rs}^{(q)})^2}{\sum_r \sum_s (d_{rs}^{(q)})^2} \right\}^{1/2}$$
 where q (usually 2 or 3) is the dimension of the solution space. A two-dimensional MDS configuration plot using the Pearson correlation matrix of the 100 gene expression profiles is displayed in Fig. 6c, with colors corresponding to gene clusters identified by the K-means analysis in Fig. 6a. Tzeng et al.^[99] developed a new rapid metric MDS method with a low computational complexity, making metric MDS applicable for large gene expression data sets.

Many useful graphical methods for studying microarray gene expression profiles are illustrated in the *Handbook of Computational Statistics: Data Visualization* edited by Chen, Härdle, and Unwin.^[100]

Gene Network Modeling and Integrated System Biology

One of the fast growing areas of research related to gene expression data analysis is the integration of expression profiles to the existing biological databases for validating and predicting gene networks. The analysis of gene expression profiles becomes much more powerful if knowledge about other biological databases such as interaction (gene-gene, gene-protein, protein-protein) and pathways (metabolic, regulatory) can be incorporated into the modeling frameworks. DeRisi et al.^[101] used global analysis of changes in gene expression to validate known metabolic structure for

diauxic shift in yeast (*Saccharomyces cerevisiae*). Marcotte et al.^[102] proposed integrating information about gene expression, protein homologues, phylogenetic profiles, metabolic function, and experimental data into genome-wide prediction of protein functions. D'haeseleer et al.^[103] discussed applications of clustering algorithms and reverse engineering, such as Boolean networks to measure the output of the gene regulatory network from gene expression data. Jenssen et al.^[104] used pure information retrieval tools to extract biomedical knowledge from publicly available gene and text databases to create a gene-to-gene cocitation network for more than ten thousand named human genes.

CONCLUSION

DNA array experiments are novel biotechnologies that have been increasingly used in biological and medical research. The mass of data generated from a typical DNA array experiment raises numerous statistical and computational challenges in data displays and processing, experimental design, and multiple testing, cluster analysis, and prediction. There are inherent biases in microarray data due to spatial, intensity, and dye effects within an array and array-to-array variation across experimental samples. Normalization is an integral part of microarray data analysis. Design consideration for a one-color experiment is similar to the parallel design in clinical trials. For a two-color experiment, the loop design appears to be more efficient than the reference design when the number of treatments is small. When the number of treatments is large, optimal design will depend on the objectives of the experiment. Several authors have considered other design issues, such as pooling mRNA samples.^[105] The field of microarray expression analysis is booming, and new analytic methods are under development. The literature is filled with novel analysis methods. New procedures for studying gene expression are being continuously developed. Generally accepted normalization procedures and data analysis methods, based on sound statistical principles, may be possible.

REFERENCES

1. The chipping forecast. *Suppl. Nat. Genet.* **1999**, *21*, 1–60.
2. Lockhart, D.J.; Dong, H.; Byrne, M.C.; Follettie, M.T.; Gallo, M.V.; Chee, M.S.; Mittmann, M.; Wang, C.; Kobayashi, M.; Horton, H.; Brown, E.L. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **1996**, *14*, 1675–1680.
3. Lipschutz, R.J.; Fodor, S.; Gingeras, T.; Lockhart, D. High-density synthetic oligonucleotide arrays. *Nat. Genet.* **1999**, *21*, 20–24.
4. Gibson, G.; Muse, S.V. *A Primer of Genome Science*; Sinauer Associates: Sunderland, MA, 2002.

5. Albertson, D.G.; Pinkel, D. Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet.* **2003**, *12*, R145–R152.
6. Shaw-Smith, C.; Redon, R.; Rickman, L.; Rio, M.; Willatt, L.; Fiegler, H.; Firth, H.; Sanlaville, D.; Winter, R.; Colleaux, L.; Bobrow, M.; Carter, N.P. Microarray based comparative genomic hybridisation (aCGH) detects sub-microscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J. Med. Genet.* **2004**, *41*, 241–248.
7. Lee, Y.S.; Chao, A.; Chao, A.S.; Chang, S.D.; Chen, C.H.; Wu, W.M.; Wang, T.H.; Wang, H.S. CGcgh: a tool for molecular karyotyping using DNA microarray-based comparative genomic hybridization (array-CGH). *J. Biomed. Sci.* **2008**, *15*, 687–696.
8. Novak, J.P.; Sladek, R.; Hudson, T.J. Characterization of variability in large-scale gene expression data: implications for study design. *Genomics* **2002**, *79*, 104–113.
9. Chen, J.J.; Delongchamp, R.R.; Tsai, C.A.; Hsueh, H.-M.; Sistare, F.; Thompson, K.; Desai, V.G.; Fuscoe, J.C. Analysis of variance components in gene expression data. *Bioinformatics* **2004**, *20*, 1436–1446.
10. Yang, Y.W.; Speed, T.P. Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* **2002**, *3*, 579–583.
11. Chen, Y.-J.; Kodell, R.L.; Sistare, F.; Thompson, K.; Morris, S.; Chen, J.J. Normalization methods for cDNA microarray data Analysis. *J. Biopharm. Stat.* **2003**, *13*, 54–57.
12. Kerr, M.K.; Churchill, G.A. Experimental design for gene expression microarrays. *Biostatistics* **2001**, *2*, 183–201.
13. Lee, M.L.; Kuo, F.C.; Whitmore, G.A.; Sklar, J. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridization. *Proc. Natl. Acad. Sci.* **2000**, *97*, 9834–9839.
14. Yang, M.C.K.; Yang, J.J.; McIndoe, R.A. Microarray experimental design: power and sample size considerations. *Physiol. Genomics* **2003**, *16*, 24–28.
15. Wang, S.J.; Chen, J.J. Sample size for identifying differentially expressed genes in microarray experiments. *J. Comput. Biol.* **2004**, *11*, 714–726.
16. Jung, S.-H. Sample size for FDR-control in microarray data analysis. *Bioinformatics* **2005**, *21*, 3097–3104.
17. Pounds, S.; Cheng, C. Sample size determination for the false discovery rate. *Bioinformatics* **2005**, *21*, 4263–4267.
18. Li, S.S.; Bigler, J.; Lampe, J.W.; Potter, J.D.; Feng, Z. FDR-controlling testing procedures and sample size determination for microarrays. *Stat. Med.* **2005**, *24*, 2267–2280.
19. Tsai, C.-A.; Wang, S.-J.; Chen, D.-T.; Chen, J.J. Sample size for gene expression microarray experiments. *Bioinformatics* **2005**, *21*, 1502–1508.
20. Lee, M.-L.; Whitmore, G. Power and sample size for DNA microarray studies. *Stat. Med.* **2002**, *21*, 3543–3570.
21. Dobbin, K.; Simon, R. Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics* **2005**, *6*, 27–38.
22. Tibshirani, R. A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics* **2006**, *7*, e106.
23. Lin, W.-J.; Hsueh, H.-M.; Chen, J.J. Power and sample size estimation in microarray studies. *BMC Bioinformatics*, **2010**, *11*, e48.
24. Chen, Y.; Dougherty, E.R.; Bittner, M.L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt.* **1997**, *2*, 364–374.
25. Schuchhardt, S.; Beule, D.; Malik, A.; Wolski, E.; Eickhoff, H.; Lehrach, H.; Herzel, H. Normalization strategies for cDNA microarrays. *Nucleic. Acids. Res.* **2000**, *28* (10), e47.
26. Kerr, M.K.; Martin, M.; Churchill, G.A. Analysis of variance for gene expression microarray data. *J. Comp. Biol.* **2000**, *7*, 819–838.
27. Kerr, M.K.; Afshari, C.A.; Bennett, L.; Bushel, P.; Martinez, J.; Walker, N.J.; Churchill, G.A. Statistical analysis of a gene expression microarray experiment with replication. *Stat. Sin.* **2002**, *12*, 203–217.
28. Wolfinger, R.D.; Gibson, G.; Wolfinger, E.D. Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* **2001**, *8*, 625–637.
29. Li, C.; Wong, W.H. Model based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 31–36.
30. Irizarry, R.A.; Bolstad, B.M.; Collin, F.; Cope, L.M.; Hobbs, B.; Speed, T.P. Summaries of Affymetrix GeneChip® probe level data. *Nucleic. Acids Res.* **2003**, *31*, e15.
31. Yang, Y.W.; Dudoit, S.; Luu, P.; Peng, V.; Ngai, J.; Speed, T.P. Normalization of cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic. Acids Res.* **2002**, *30*, e15.
32. Tsai, C.; Hsueh, H.; Chen, J.J. A Generalized additive model for microarray gene expression data analysis. *J. Biopharm. Stat.* **2004**, *14*, 553–573.
33. Lee, Y.S.; Chen, C.H.; Tsai, C.N.; Chao, A.; Wang, T.H. Microarray labeling extension values: laboratory signatures for Affymetrix GeneChips. *Nucleic. Acids Res.* **2009**, *37*, e61.
34. Chen, J.J.; Wang, S.-J.; Tsai, C.-A.; Lin, C.-J. Selection of differentially expressed genes in microarray data analysis. *Pharmacogenomics J.* **2007**, *7*, 221–220.
35. Chen, J.J.; Hsueh, H.-M.; Delongchamp, R.R.; Lin, C.-J.; Tsai, C.-A. Reproducibility of microarray data: a further analysis of microarray quality control data. *BMC Bioinformatics* **2007**, *8*, e412.
36. Tusher, V.G.; Tibshirani, R.; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 5116–5121.
37. Baldi, P.; Long, A.D. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **2001**, *17*, 509–519.
38. Lönnstedt, I.; Speed, T.P. Replicated microarray data. *Stat. Sin.* **2002**, *12*, 31–45.
39. Dudoit, S.; Yang, Y.H.; Callow, M.J.; Speed, T.P. Statistical methods for identifying differential expressed genes in replicated cDNA microarray experiments. *Stat. Sin.* **2002**, *12*, 111–139.
40. Tsai, C.; Chen, Y.J.; Chen, J.J. Testing for differentially expressed genes with microarray data. *Nucleic. Acids Res.* **2003**, *31*, e52.
41. Wright, G.W.; Simon, R.M. A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **2003**, *19*, 2448–2455.

42. Jain, N.; Thatte, J.; Braciale, T.; Ley, K.; O'Connell, M.; Lee, J.K. Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* **2003**, *19*, 1945–1951.
43. Smyth, G.K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **2004**, *3*, A1–A3.
44. Cui, X.; Hwang, J.T.G.; Qiu, J.; Blades, N.J.; Churchill, G.A. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **2005**, *6*, 59–75.
45. Hochberg, Y.; Tamhane, A.C. *Multiple Comparison Procedures*; John Wiley & Sons: New York, 1987.
46. Westfall, P.H.; Young, S.S. *Resampling-Based Multiple Testing*; John Wiley & Sons: New York, 1993.
47. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
48. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **1995**, *57*, 289–300.
49. Storey, J.D. A direct approach to false discovery rate. *J. R. Stat. Soc. B* **2002**, *64*, 479–498.
50. Tsai, C.A.; Hsueh, H.M.; Chen, J.J. Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics* **2003**, *59*, 1071–1081.
51. Benjamini, Y.; Hochberg, Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stati.* **2000**, *25*, 60–83.
52. Hsueh, H.; Chen, J.J.; Kodell, R.L. Comparison of methods for estimating number of true null hypothesis in multiplicity testing. *J. Biopharm. Stat.* **2003**, *13*, 675–689.
53. Delongchamp, R.R.; Bowyer, J.F.; Chen, J.J.; Kodell, R.L. Multiple testing strategy for analyzing cDNA array data on gene expression. *Biometrics* **2004**, *60*, 774–782.
54. Jin, W.; Riley, R.M.; Wolfinger, R.D.; White, K.P.; Passador-Gurgel, G.; Gibson, G. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat. Genet.* **2001**, *29*, 389–395.
55. Chen, J.J.; Tseng, S.L.; Tsai, C.A.; Chen, C.H. Gene selection with multiple ordering criteria. *BMC Bioinformatics* **2007**, *8*, e74.
56. Draghici, S.; Khatri, P.; Martins, R.P.; Ostermeier, G.C.; Krawetz, S.A. Global functional profiling of gene expression. *Genomics* **2003**, *81*, 98–104.
57. Pavlidis, P.; Qin, J.; Arango, V.; Mann, J.J.; Sibille, E. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem. Res.* **2004**, *29*, 1213–1222.
58. Rivals, I.; Personnaz, L.; Taing, L.; Potier, M.C. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* **2007**, *23*, 401–407.
59. Tian, L.; Greenberg, S.A.; Kong, S.W.; Altschuler, J.; Kohane, I.S.; Park, P.J. Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci. USA* **2005**, *102*, 13544–13549.
60. Chen, J.J.; Lee, T.; Delongchamp, R.; Chen, T.; Tsai, C.A. Significance analysis of groups of genes in expression profiling studies. *Bioinformatics* **2007**, *23*, 2104–2112.
61. Mootha, V.K.; Lindgren, C.M.; Eriksson, K.F.; Subramanian, A.; Sihag, S.; Lehar, J.; Puigserver, P.; Carlsson, E.; Ridderstråle, M.; Laurila, E.; Houstis, N.; Daly, M.J.; Patterson, N.; Mesirov, J.P.; Golub, T.R.; Tamayo, P.; Spiegelman, B.; Lander, E.S.; Hirschhorn, J.N.; Altshuler, D.; Groop, L.C. PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **2003**, *34*, 267–273.
62. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; Mesirov, J.P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **2005**, *102*, 15545–15550.
63. Efron, B.; Tibshirani, R. On testing the significance of set of genes. *Ann. Appl. Stat.* **2007**, *1*, 107–129.
64. Goeman, J.J.; van de Geer, S.A.; de Kort, F.; van Houwelingen, H.C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* **2004**, *20*, 93–99.
65. Dinu, I.; Potter, J.D.; Mueller, T.; Liu, Q.; Adewale, A.J.; Jhangri, G.S.; Einecke, G.; Famulski, K.S.; Halloran, P.; Yasui, Y. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics* **2007**, *8*, e242.
66. Hummel, M.; Meister, R.; Mansmann, U. GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics* **2008**, *24*, 78–85.
67. Tsai, C.-A.; Chen, J.J. Multivariate analysis of variance test for gene set analysis. *Bioinformatics* **2009**, *25*, 897–903.
68. Blum, A.; Langley, P. Selection of relevant features and examples in machine learning. *Artif. Intell.* **1997**, *97*, 245–271.
69. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324.
70. Hastie, T.; Tibshirani, R.T.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, 2001.
71. Vapnik, V.N. *Statistical Learning Theory*; Wiley: New York, 1998.
72. Tsai, C.-A.; Chen, C.-H.; Lee, T.-C.; Ho, I.-C.; Yang, U.-C.; Chen, J.J. Gene Selection for sample classifications in microarray experiments. *DNA Cell Biol.* **2004**, *23*, 607–614.
73. Furey, T.S.; Christianini, N.; Duffy, N.; Bednarski, D.W.; Schummer, M.; Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **2000**, *16*, 906–914.
74. Brieman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *CART: Classification and Regression Trees*; Chapman and Hall/CRC: Boca Raton, FL, 1998.
75. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422.
76. Dudoit, S.; Fridlyand, J.; Speed, T.P. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **2002**, *97*, 77–87.
77. Lee, J.W.; Lee, J.B.; Park, M.; Song, S.H. An extensive evaluation of recent classification tools applied to microarray data. *Comput. Stat. Data. Anal.* **2005**, *48*, 869–885.
78. Moon, H.; Ahn, H.; Kodell, R.L.; Lin, C.-J.; Baek, S.; Chen, J.J. Ensemble methods for classification of patients for personalized medicine with high-dimensional data. *Artif. Intell. Med.* **2007**, *41*, 197–207.

79. Ambrose, C.; McLachlan, G.J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 6562–6566.
80. Dupuy, A.; Simon, R. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J. Natl. Cancer Inst.* **2007**, *99*, 147–157.
81. Wen, X.; Fuhrman, S.; Michaels, G.S.; Carr, D.B.; Smith, S.; Barker, J.L.; Somogyi, R. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 334–399.
82. Eisen, M.B.; Spellman, P.T.; Brown, P.O.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 14863–14868.
83. Chen, C.H. Generalized association plots: information visualization via iteratively generated correlation matrices. *Stat. Sin.* **2002**, *12*, 7–29.
84. Sokal, R.R.; Sneath, P.H. *Principal of Numerical Taxonomy*; Freeman: San Francisco, CA, 1963.
85. Hartigan, J.H. Representation of similarity matrices by trees. *J. Am. Stat. Assoc.* **1967**, *62*, 1140–1158.
86. Alon, U.; Barkai, N.; Notterman, D.A.; Gish, K.; Ybarra, B.; Mack, D.; Levine, A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 6745–6750.
87. Tien, Y.J.; Lee, Y.S.; Wu, H.M.; Chen, C.H. Methods for simultaneously identifying coherent local clusters with smooth global patterns in gene expression profiles. *BMC Bioinformatics* **2008**, *9*, e155.
88. Wu, H.-M.; Tien, Y.-J.; Chen, C.-H. GAP: a graphical environment for matrix visualization and cluster analysis. *Comput. Stat. Data Anal.* **2010**, *54*, 767–778.
89. Tavazoie, S.; Hughes, J.D.; Campbell, M.J.; Cho, R.J.; Church, G.M. Systematic determination of genetic network architecture. *Nat. Genet.* **1999**, *22*, 281–285.
90. Kohonen, T. *Self-Organizing Maps*; Springer: Berlin, 1995.
91. Tamayo, P.; Slonim, D.; Mesirov, J.; Zhu, Q.; Kitareewan, S.; Dmitrovsky, E.; Lander, E.S.; Golub, T.R. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 2907–2912.
92. Panel on Discriminant Analysis and Clustering. Discriminant analysis and clustering. *Stat. Sci.* **1989**, *4*, 34–69.
93. Alter, O.; Brown, P.O.; Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 10101–10106.
94. Holter, N.S.; Mitra, M.; Maritan, A.; Cieplak, M.; Banavar, J.R.; Fedoroff, N.V. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 8409–8414.
95. Borg, I.; Groenen, P. *Modern Multidimensional Scaling: Theory and Applications*; Springer-Verlag: New York, 1997.
96. Bittner, M.; Meltzer, P.; Chen, Y.; Jiang, Y.; Seftor, E.; Hendrix, M.; Radmacher, M.; Simon, R.; Yakhnik, Z.; Ben-Dor, A.; Sampas, N.; Dougherty, E.; Wang, E.; Marincola, F.; Gooden, C.; Lueders, J.; Glatfelter, A.; Pollock, P.; Carpten, J.; Gillanders, E.; Leja, D.; Dietrich, K.; Beaudry, C.; Berens, M.; Alberts, D.; Sondak, V.; Hayward, N.; Trent, J. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **2001**, *406*, 536–540.
97. Hastie, T.; Tibshirani, R.; Eisen, M.B.; Alizadeh, A.; Levy, R.; Staudt, L.; Chan, W.C.; Botstein, D.; Brown, P. Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* **2000**, *2*, 3.1–3.21.
98. Kruskal, J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **1964**, *29*, 1–27.
99. Tzeng, J.; Lu, H.H.-S.; Li, W.-H. Multidimensional scaling for large genomic data sets. *BMC Bioinformatics* **2008**, *9*, e179.
100. In *Handbook of Computational Statistics: Data Visualization*; Chen, C.H.; Härdle, W.; Unwin, A., Eds.; Springer-Verlag: Berlin, 2008.
101. DeRisi, J.L.; Iyer, V.R.; Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **1997**, *278*, 680–686.
102. Marcotte, E.M.; Pellegrini, M.; Thompson, M.J.; Yeates, T.O.; Eisenberg, D.A. Combined algorithm for genome-wide prediction of protein function. *Nature* **1999**, *402*, 1–8386.
103. D'haeseleer, P.; Liang, S.; Somogyi, R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* **2000**, *16*, 707–726.
104. Jenssen, T.K.; Laegreid, A.; Komorowski, J.; Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* **2001**, *28*, 21–28.
105. Zin, W.; Riley, R.M.; Wolfinger, R.D.; White, K.P.; Passador-Gurgel, G.; Gibson, G. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat. Genet.* **2001**, *29*, 389–395.

Microdosing Studies

Fuyu Song

Center for Food and Drug Inspection, China Food and Drug Administration, Beijing, China

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

In recent years, microdosing approach has received much attention in pharmaceutical research and development. The concept of microdosing approach focuses on safety signal detection, which is primary for exploratory purpose (FDA, 2006; ICH, 2009). Burt (2011) indicated that the use of microdosing approach has great potential in shortening the development by making critical go/no-go decision early. In addition, it is believed that microdosing approach will increase the probability of success based on a limited number of subjects available. In practice, however, many scientific factors regarding the design and analysis of microdose studies are commonly encountered, which have limited the potential use of microdosing approach in pharmaceutical development. These scientific factors and/or issues include, but are not limited to, (i) distinction between microdose effect and placebo effect, (ii) microdose versus therapeutic dose, (iii) characterization of dose–response curve, (iv) accuracy and reliability of predictive model, and (v) possible false positive and false negative rates. These issues are discussed. In addition, a predictive index is proposed to assess the accuracy and reliability of predictive model established based on data collected from a microdose study.

Keywords: False positive/negative; Microdose; Placebo effect; Predictive index; Therapeutic dose.

INTRODUCTION

In pharmaceutical development, a microdose is defined as a dose that is less than 1/100th of the dose of a test substance calculated (based on animal data) to yield a pharmacologic effect of the test substance with a maximum dose of $\leq 100 \mu\text{g}$.^[1,2] Due to differences in molecular weights compared to synthetic drugs, the maximum dose for protein products is ≤ 30 nanomoles. Microdosing studies are often designed not only to evaluate pharmacokinetics or imaging of specific targets but also to induce pharmacologic effects. In practice, the harm or risk of microdose to human subjects is considered very limited and not notable, while information adequate to support the initiation of such limited human studies can be derived from limited nonclinical safety studies. In the last decade, microdose is often compared with those observed at a therapeutic dose (see, e.g., the works of Lappin,^[3] Burt,^[4] Burt et al.,^[5] and Chow and Song^[6]). As indicated by Lappin,^[3] about 80% of the microdose pharmacokinetics available in the public domain has been shown to scale to those observed at a therapeutic dose, within a twofold difference. As a result, microdosing approach is now being extended into drug (pharmacokinetics) development in situations where the concentration of a drug in cell or tissue types is relevant to its efficacy. In addition, the concept of microdosing approach has been applied to address drug–drug

interactions by giving human volunteers a microdose of the candidate drug before and after the administration of a drug known to inhibit or induce certain enzymes.^[3]

As indicated in the 2006 FDA guidance on *exploratory Investigational New Drug (IND)*, a typical microdosing study involves very limited human exposure and has no therapeutic or diagnostic intent (see also the FDA guideline^[7]).^[1] A microdosing study is considered an exploratory study that focuses on the detection of safety signals. Thus, it is suggested that preclinical and clinical approaches, as well as chemistry, manufacturing, and controls (CMC) information, should be considered when planning exploratory studies in humans, including studies of closely related drugs or therapeutic biological products. The FDA currently accepts the use of extended single-dose toxicity studies in animals to support single-dose studies in humans. For microdosing studies, a single mammalian species (both sexes) can be used if justified by in vitro metabolism data and by comparative data on in vitro pharmacodynamics effects. The route of exposure in animals should be by the intended clinical route. In these studies, animals should be observed for 14 days post-dosing with an interim necropsy, typically on day 2, and endpoints evaluated should include body weights. Because microdosing studies involve only single exposures to microgram quantities of test materials and such exposures are comparable to routine environmental exposures, routine generic

toxicology testing is not needed. For similar reasons, safety pharmacology studies are also not recommended.

The concept of microdosing approach in pharmaceutical development is encouraging. As indicated by Burt,^[4] the use of human microdosing approach in pharmaceutical development has the following benefits: (1) it takes just 6 months from laboratory bench to completion of clinical studies, (2) efficient lead candidate selection, (3) reduces expensive late-stage attrition (i.e., kill ineffective compound early and cheap), (4) substantially reduced preclinical toxicology package compared to phase I, (5) only gram quantities of non-good manufacturing practice (GMP) drug (typically 10 g) are needed, (6) any route of administration possible, including intravenous (IV), (7) absolute oral bioavailability calculation, (8) drugs can be tested in sensitive populations; renal impaired patients, women of child bearing age, and cancer patients, and (9) reduces the use of animals in research. However, Chow and Song^[6] pointed out that there are statistical issues/concerns regarding the validity of the human microdosing approach in pharmaceutical development. These statistical issues include (1) the selection of microdose (e.g., “how to distinguish the effects due to microdose and the placebo effect?” and “how to select the initial dose and the dose range under study?”), (2) the selection of study endpoint (e.g., “should a clinical endpoint or a surrogate endpoint or a biomarker be used?” and “whether the surrogate or biomarker is predictive of clinical endpoint?”), (3) the determination of sample size (e.g., precision analysis or power analysis), and (4) model selection and validation (e.g., “how to handle inter- and intra-subject variabilities?” and “how to determine the impact of nonlinearity for the dose

range under study?”). In addition, it is a concern that the extrapolation based on the results from microdose to regular dose may not be reliable because the variability associated with the dose is usually proportional to the dose.

In Section “Regulatory Perspectives,” regulatory perspectives^[1,2] regarding the use of microdosing approach in drug development are discussed followed by some scientific factors^[6] that are commonly encountered when applying microdosing approach in drug development. In Section “Prediction of Microdose,” under certain assumptions, a predictive index for comparing the microdose and a therapeutic dose proposed by Chow and Song^[6] is discussed. Some concluding remarks are given in the last section.

REGULATORY PERSPECTIVES

ICH Guideline

In 2009, the International Conference on Harmonization (ICH) published a guideline that provides international standards for the nonclinical safety studies recommended to support human clinical trials for pharmaceuticals.^[2] This guidance further suggests reducing the use of animals and other resources for drug development by considering alternative methods for safety evaluation. These methods such as microdosing approach can be used to replace current standard methods. The ICH M3 (R2) guideline recommends two different microdosing approaches, which are briefly described next (see also Table 1).

The first approach would involve no more than a total dose of 100 μ g that can be administered as a single dose

Table 1 Recommended nonlinear studies to support exploratory clinical trials

Dose to be administered	Clinical		Nonclinical	
	Start and maximum doses	Pharmacology	General toxicity studies	Genotoxicity/other
Approach 1: Total dose $\leq 100 \mu$ g (no interdose interval limitations) AND total dose $\leq 1/100$ th no-observed-adverse-effect-level (NOAEL) and $\leq 1/100$ th pharmacologically active dose (scaled on mg/kg for i.v. and mg/m ² for oral)	Maximal and starting doses can be the same but not exceed a total accumulated dose of 100 μ g.	In vitro target/receptor profiling should be conducted Appropriate characterization of primary pharmacology (mode of action and/or effects) in a pharmacodynamically relevant model should be available to support human dose selection.	Extended single-dose toxicity study in one species, usually rodent, by intended route of administration with toxicokinetic data, or via the i.v. route. A maximum dose of 1000-fold the clinical dose on a mg/kg basis for i.v. and mg/m ² for oral administration can be used.	Genotoxicity studies are not recommended, but any studies or structure–activity relationship (SAR) assessments conducted should be included in the clinical trial application. For highly radioactive agents (e.g., PET imaging agents), appropriate PK and dosimetry estimates should be submitted.

(Continued)

Table 1 Recommended nonlinear studies to support exploratory clinical trials (*Continued*)

Dose to be administered	Clinical	Nonclinical		
	Start and maximum doses	Pharmacology	General toxicity studies	Genotoxicity/other
Approach 2: Total cumulative dose $\leq 500 \mu\text{g}$, maximum of five administrations with a washout between doses (six or more actual or predicted half-lives) and each dose $\leq 100 \mu\text{g}$ AND each dose $\leq 1/100\text{th}$ of the NOAEL and $< 1/100\text{th}$ of the pharmacologically active dose	Maximal daily and starting doses can be the same, but not exceed $100 \mu\text{g}$.	In vitro target/receptor profiling should be conducted. Appropriate characterization of primary pharmacology (mode of action and/or effects) in a pharmacodynamically relevant model should be available to support human dose selection.	7-Day repeated-dose toxicity study in one species, usually rodent, by the intended route of administration with toxicokinetic data, or via the i.v. route. Hematology, clinical chemistry, necropsy, and histopathology data should be included. A maximum dose of 1000-fold the clinical dose on a mg/kg basis for i.v. and mg/m^2 for oral administration can be used.	Genotoxicity studies are not recommended, but any studies or SAR assessments conducted should be included in the clinical trial application. For highly radioactive agents (e.g., PET imaging agents), appropriate PK and dosimetry estimates should be submitted.
Approach 4: Dosing up to 14 days into the therapeutic range but not intended to evaluate clinical MTD	With toxicity in both species, follow appropriate regional guidance for clinical starting dose, if toxicity is not seen in either species (i.e., the NOAELs are the highest dose tested and doses used were not otherwise limited, e.g., not an multifunctional device [MFD]), or is seen only in one species, the clinical starting dose should be the one that gives a predicted clinical area under the curve (AUC) value (based on either interspecies PK modeling or mg/m^2 conversion) that is approximately $1/50\text{th}$ of the AUC at the NOAEL from the species yielding the lower exposure. For other considerations on initial dosing in humans, e.g., predicted PD activity, regional guidance should be consulted. Without toxicity in both species, it is recommended that the maximum clinical dose not exceed $1/10\text{th}$ the lower exposure (AUC) in either species at the highest dose tested in the animals.		2-Week repeated-dose toxicity studies in rodent and non-rodent with standard parameters assessed and where dose selection in animals is based on exposure multiples of anticipated clinical AUC at maximum dose.	Ames assay (or an appropriate alternative assay if Ames is inappropriate, e.g., for an antibacterial product) and an assay (in vitro or in vivo) capable of detecting chromosomal damage in a mammalian system.

or divided doses in any subject. This could be useful to investigate the target receptor binding or tissue distribution in a positron emission tomography (PET) study. A second use could be to assess PK with or without the use of an isotopically labeled agent. The second microdosing approach is the one that involves less than five administrations of a maximum of 100 µg per administration (a total of 500 µg per subject). This can be useful for applications similar to the first microdosing approach described above, but with less active PET ligands.

In some situations, it could be appropriate to carry out a clinical microdosing study using the IV route on a product intended for oral administration and for which an oral nonclinical toxicology package already exists. In this case, the IV microdose can be qualified by the existing oral toxicity studies as described in Table 1, Approach 3, where adequate exposure margins have been achieved. It is not recommended to investigate the IV local tolerance of the drug substance in this situation because the administered dose is very low (100 µg maximum). If a novel IV vehicle is being employed, then local tolerance of the vehicle should be assessed.

FDA Guidance

Similar to the ICH M3 (R2) guideline, in 2006, the FDA published guidance on exploratory IND study, which is intended for clinical trials that are conducted early in phase 1 such as microdosing studies. From the FDA's perspective, a microdosing study usually involves very limited human exposure, and most importantly, it has no therapeutic or diagnostic intent. As indicated by the 2006 FDA guidance, microdosing studies are often conducted prior to the traditional dose escalation, safety, and tolerance studies that ordinarily initiate a clinical drug development program. The duration of dosing in an exploratory IND study is expected to be limited (e.g., 7 days). This guidance applies to early phase 1 clinical studies of investigational new drugs and biological products that assess feasibility for further development of the drug or biological product.

In addition, the FDA published the guidance "PET Drugs—Current Good Manufacturing Practice (cGMP)," which is intended for PET drugs.^[8] The guidance addresses resources, procedures, and documentation for all PET drug production facilities. In some cases, the guidance provides practical examples of methods or procedures that PET drug production facilities can use to comply with the cGMP requirements (see also^[9]).

SCIENTIFIC FACTORS

As mentioned earlier, despite some potential benefits in the early phase of pharmaceutical/clinical development, Chow and Song^[6] pointed out that there are statistical factors regarding the validity of the human microdosing approach

in pharmaceutical development. These statistical issues and concerns are described in the sections that follow.

Microdose versus Placebo

As indicated earlier, a microdose is less than 1/100th of the dose of a test substance calculated (based on animal data) to yield a pharmacologic effect of the test substance with a maximum dose of ≤100 µg. As a result, a microdose is close to a placebo (inactive dose). Thus, an interesting question is how to distinguish whether the effect (or response) observed is due to the microdose or a placebo.

A placebo response is a simulated or otherwise medically ineffectual treatment for a disease or other medical condition intended to deceive the recipient. Sometimes, patients given a placebo treatment will have a perceived or actual improvement in a medical condition; this phenomenon is usually referred to as the placebo effect. In clinical research, placebo effects are often the subject of scientific research aiming to understand the underlying neurobiological mechanisms of action in pain relief, immunosuppression, Parkinson's disease, and depression. For example, Zubieta et al.^[10] indicated that placebos can have real, measurable effects on physiological changes in the brain detected by brain imaging techniques (see also the work of Vul et al.^[11]).

To ensure the success of microdosing studies, it is important to rule out that the observed effect is by chance alone and is due to placebo effect.

Selection of Therapeutic Dose

An effective dose is defined as the dose or amount of drug that produces a therapeutic response or desired effect in some fraction of the subjects taking it. Therapeutic dose is an effective dose that is required in order to produce a desirable (or clinically meaningful) effect. The therapeutic range of a drug is defined by the range between the minimum effective dose (MED) and the maximum tolerable dose (MTD). The MED is the lowest dose level of a drug product that provides a clinically significant response in average efficacy, which is also statistically significantly superior to the response provided by the placebo. Similarly, the MTD is the highest possible but still tolerable dose level with respect to a prespecified clinical limiting toxicity. In general, these limits refer to the average patient population. For instances in which there is a large discrepancy between the MED and the MTD, it is stated that the drug has a large therapeutic window. Conversely, if the range is relatively small, or if the MTD is less than the MED, then the pharmaceutical product will have little to no practical value.

In microdosing studies, a microdose is often compared with those observed at a therapeutic dose. The ultimate goal is to determine whether the treatment effect observed at the microdose is predictive of that at the therapeutic

dose. In practice, however, the following questions raised: (1) How to confirm that the selected therapeutic dose is truly an effective dose? and (2) Is microdose predictive of the selected therapeutic dose?

Animal Model versus Human Model

In order to establish a predictive model for human, the following predictive models are necessarily established:

1. For animal population, it is to determine that microdose is predictive of therapeutic dose.
2. Animal model is predictive of human model at a given dose. For example, at microdose, animal response is predictive of human response, while at therapeutic dose, animal response is predictive of human response.

Let A_M and A_T be the animal responses at microdose and therapeutic dose, respectively, and H_M and H_T be the human responses at microdose and therapeutic dose, respectively. In this case, suppose that we have (1) $H_T = f_1(A_T)$, (2) $A_T = f_2(A_M)$, and (3) $H_M = f_3(A_M)$ or $A_M = f_3^{-1}(H_M)$. As a result, we have $H_T = f_1 f_2 f_3^{-1}(H_M)$ or $H_T = g(H_M)$, where $g = f_1 f_2 f_3^{-1}$. It should be noted that in practice, f_1 , f_2 , and f_3 may not be linear.

Characterization of Dose–Response Curve

To determine whether the microdose is predictive of the therapeutic dose, we first need to characterize the dose–response curve based on several microdoses and perhaps a therapeutic dose or a couple of therapeutic doses. With a limited number of dose levels available, it is difficult, if not impossible, to characterize the dose–response curve. If there are only two dose levels, i.e., microdose and therapeutic dose, extrapolation may not be reliable especially when the therapeutic dose is much larger than the microdose.

In addition, the dose–response curve in a dose range under study could be either linear or nonlinear. In this case, the traditional analysis of variance (ANOVA)-type F-tests with specific contrasts (linear or nonlinear) may not be appropriate. Alternatively, Cheng et al.^[12] proposed the use of slope approach by considering slopes between dose levels to characterize the dose–response curve. The slope approach (based on either slopes between dose levels and initial dose or slopes between adjacent dose levels) was found to be effective and can describe the shape of the dose–response curve and will be able to determine whether there is a clinically meaningful difference in response between the microdose and the therapeutic dose.

Variability Associated with Dose

Heterogeneity in variance between microdose and therapeutic dose is a major concern in establishing a predictive

model based on the microdose for the therapeutic dose. At microdose, the sensitivity associated with the analytical methods then plays an important role for accurate and reliable quantitation of the response of a given microdose, which has an impact on the prediction of therapeutic dose.

In dose–response studies, it is assumed that the dose–response is a monotonic increasing function of dose. In practice, it also recognized that in a dose–response study, the variability associated with the response is often proportional to the dose. Under the monotonic assumption and the fact that variability is proportional to the dose, several statistical methods have been proposed in the literature. These methods include (1) testing for linearity or nonlinearity contrasts such as quadratic contrast in the ANOVA and (2) a slope approach proposed by Cheng et al.^[12] based on normalized responses (i.e., responses adjusted for doses).

False Negative and False Positive

In practice, an established predictive model needs to be validated. Commonly used criteria are the false negative rate and/or false positive rate, which have an impact on the accuracy and reliability of the established predictive model. The false negative rate and/or false positive rate depend(s) upon not only the study design but also assay variability and assay sensitivity. Thus, it is suggested that a pilot study be conducted for determining the false negative rate and/or false positive rate before the conduct of a microdosing study in humans.

PREDICTION OF MICRODOSE

Under certain assumptions, Chow and Song^[6] proposed a predictive index for comparing the microdose and the therapeutic dose in microdosing studies, which is described next.

Predictability Index

Let X and Y be the characteristics at a microdose from animal and human, where X and Y are independent and follow normal distributions with means μ_X and μ_Y and variances σ_X^2 and σ_Y^2 , respectively. The concept used here is that if Y is predictive of X , the probability that the ratio of the standardized responses $V = (X - \mu_X)/\sigma_X$ and $W = (Y - \mu_Y)/\sigma_Y$ is equal to 1 could be close to 1. Similar to the idea of the concept for consistency of raw materials and/or final products from two different sites,^[13] we propose the following probability as an index to assess the predictability of the two characteristics:

$$p = P\left(1 - \delta < \frac{V}{W} < \frac{1}{1 - \delta}\right), \quad (1)$$

where $0 < \delta < 1$ is defined as a limit that allows for prediction. Thus, p tends to 1 as δ tends to 1. We will refer to p

as the predictive index. A small δ implies the requirement of a high degree of consistency at the microdose between animals and humans. Under the normality assumption of X and Y , the ratio of the two centered normal variables, $U = (X - \mu_X)/(Y - \mu_Y)$, is Cauchy variable with probability density function (see the work of Cedilnik et al.^[14]):

$$f_U(u) = \frac{1}{\pi} \cdot \frac{\sigma_X/\sigma_Y}{u^2 + (\sigma_X/\sigma_Y)^2}.$$

Thus, Eq. 1 can be rewritten as

$$p = P\left((1-\delta) \cdot \frac{\sigma_X}{\sigma_Y} < \frac{V}{W} < \frac{1}{1-\delta} \cdot \frac{\sigma_X}{\sigma_Y}\right) \\ = \left[0.5 + \frac{\tan^{-1}(u)}{\pi}\right]_{u=(1-\delta) \cdot \frac{\sigma_X}{\sigma_Y}}^{u=\frac{1}{1-\delta} \cdot \frac{\sigma_X}{\sigma_Y}} \quad (2)$$

Therefore, the predictive index p is a function of the parameters $\theta = (\sigma_X^2, \sigma_Y^2)$. Suppose that observations $X_i, i = 1, \dots, n_X$ and $Y_i, i = 1, \dots, n_Y$ are collected. Thus, using the invariance principle, the maximum likelihood estimate of p can be obtained as

$$\hat{p} = \left[0.5 + \frac{\tan^{-1}(u)}{\pi}\right]_{u=(1-\delta) \cdot \frac{\hat{\sigma}_X}{\hat{\sigma}_Y}}^{u=\frac{1}{1-\delta} \cdot \frac{\hat{\sigma}_X}{\hat{\sigma}_Y}}, \quad (3)$$

where $\hat{\sigma}_X^2 = \sum_{i=1}^{n_X} (X_i - \bar{X})^2 / n_X$, $\hat{\sigma}_Y^2 = \sum_{i=1}^{n_Y} (Y_i - \bar{Y})^2 / n_Y$, and $\bar{X}$ and $\bar{Y}$ are the sample means of X and Y , respectively. In other words, $\hat{p} = h(\hat{\theta}) = h(\hat{\sigma}_X^2, \hat{\sigma}_Y^2)$. Similar to the idea of the concept for consistency of Tse et al.,^[13] it can be verified that

$$\frac{\hat{p} - p - B(\hat{p})}{C(\hat{p})} \longrightarrow N(0, 1),$$

where $B(\hat{p})$ and $C(\hat{p})$ are the expected value and standard deviation of $\hat{p}$. Under this asymptotic result, a $(1 - \alpha) \times 100\%$ confidence interval of p can be obtained. Let $LL(\hat{p})$ be the lower limit of the confidence interval. Following the similar idea of consistency proposed by Tse et al.,^[13] we propose the following quality control (QC) criterion for examination of prediction of microdose. If the probability that the lower limit $LL(\hat{p})$ of the constructed $(1 - \alpha) \times 100\%$ confidence interval of p is greater than or equal to a prespecified QC lower limit—say QC_L exceeds a prespecified number β (say $\beta = 80\%$), then we claim that U and W are consistent or similar. In other words, U and W are consistent or similar if

$$P(QC_L \leq LL(\hat{p})) \geq \beta, \quad (4)$$

where β is a prespecified constant.

Sample Size Requirement

In practice, it is necessary to select a sample size to ensure that there is a high probability, say β , of consistency between U and W when in fact U and W are consistent. It is suggested that the sample size is chosen such that there is more than 80% chance that the lower confidence limit of p is greater than or equal to the QC lower limit, i.e., $\beta = 0.8$. In other words, the sample size is determined such that $P\{QC_L \leq LL(\hat{p})\} \geq \beta$. This leads to

$$P\{QC_L \leq \hat{p} - B(\hat{p}) - z_{\alpha/2} \sqrt{\text{Var}(\hat{p})}\} \geq \beta.$$

Thus,

$$P\{QC_L + z_{\alpha/2} \sqrt{\text{Var}(\hat{p})} - p \leq \hat{p} - p - B(p)\} \geq \beta.$$

This gives

$$P\left\{\frac{QC_L - p}{\sqrt{\text{Var}(\hat{p})}} + z_{\alpha/2} \leq \frac{\hat{p} - p - B(p)}{\sqrt{\text{Var}(\hat{p})}}\right\} \geq \beta.$$

Therefore, the sample size required for achieving a probability higher than β can be obtained by solving the following equation:

$$\frac{QC_L - p}{\sqrt{\text{Var}(\hat{p})}} + z_{\alpha/2} \leq -z_{1-\beta}. \quad (5)$$

Assuming that $n_X = n_Y = n$, then the common sample size is given by

$$n \geq \frac{(z_{1-\beta} + z_{\alpha/2})^2}{(p - QC_L)^2} \left\{ \left(\frac{\partial \hat{p}}{\partial \mu_X} \right)^2 V_X + \left(\frac{\partial \hat{p}}{\partial \mu_Y} \right)^2 V_Y + \left(\frac{\partial \hat{p}}{\partial V_X} \right)^2 (2V_X^2) + \left(\frac{\partial \hat{p}}{\partial V_Y} \right)^2 (2V_Y^2) \right\}. \quad (6)$$

The above result suggests that the required sample size will depend on the choices of α , β , V_X , V_Y , $\mu_X - \mu_Y$, and $p - QC_L$.

CONCLUDING REMARKS

As indicated by Lappin^[3] and Burt,^[4] the application of microdosing approach as a tool for early drug selection and development is growing and can be applied to drug-to-drug interaction by examining drug concentrations in tissues and certain cell types. Both ICH guideline and FDA guidance, however, emphasize that microdosing approach is for exploratory purpose, and it should focus on the safety in the early phase of pharmaceutical/clinical development rather than on the development for efficacy.

In practice, many scientific factors and practical issues, which limit the possible application of microdosing approach in pharmaceutical research and development, remain unsolved. These scientific factors include, but are not limited to, (1) the issue of placebo effect both at the microdose and at the therapeutic dose, (2) the prediction of animal model to human model, (3) the selection of therapeutic dose, (4) the characterization of dose–response curve, (5) accuracy and reliability of extrapolation (prediction) of the microdose to the therapeutic dose, and (6) the assessment of false positive/negative of microdose prediction. These practical issues have limitations in many aspects on drug development for efficacy.

Although the predictive index proposed by Chow and Song^[6] can examine the closeness between the treatment effects at microdose and therapeutic dose under a simple study design, design and analysis of microdosing studies are still not fully understood. In practice, it is suggested that appropriate statistical methods should be developed under a valid design for a rigorous and more accurate and reliable assessment of the performance of microdosing approach in pharmaceutical development.

REFERENCES

1. FDA. *Guidance for Industry, Investigators, and Reviewers—Exploratory IND Studies*. Center for Drug Evaluation and Research (CDER), Food and Drug Administration, Rockville, MD, January 2006.
2. ICH M3(R2). *International Conference on Harmonization Guidance on Nonclinical Safety Studies for the Conduct of Human Clinical Trials and Marketing Authorization for Pharmaceuticals*. June 2009.
3. Lappin, G. Microdosing: Current and the future. *Bioanalysis* **2010**, 2 (3), 509–517.
4. Burt, T. Microdosing and phase 0. Presented at Duke Clinical Research Unit, Duke University Medical Center, Durham, NC, September 15, 2011.
5. Burt, T.; Wu, H.; Layton, A.; Lee, K.; Rouse, D.C.; Hawk, T.C.; Weitzel, D.H.; Chin, B.B.; Cohen-Wolkowicz, M.; Chow, S.C.; Noveck, R.J. Intra-arterial microdosing (IAM): A novel drug development approach, proof-of-concept PET study in rats. *J. Nucl. Med.* **2015**, 56 (11), 1793–1799.
6. Chow, S.C.; Song, F.Y. Scientific factors on microdosing approach in pharmaceutical development. *World J. Pharm. Res.* **2015**, 4 (8), 266–277.
7. FDA. *Guidance for Industry and Researchers—The Radioactive Drug Research Committee: Human Research without an Investigational New Drug Application*. Center for Drug Evaluation and Research (CDER), Food and Drug Administration, Rockville, MD, August 2010.
8. FDA. *Guidance for PET Drugs—Current Good Manufacturing Practice (CGMP)*. Center for Drug Evaluation and Research (CDER), Food and Drug Administration, Rockville, MD, August 2011.
9. Yarkoni, T. Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—Commentary on Vul et al. (2009). *Perspect. Psychol. Sci.* **2009**, 4 (3), 294–298.
10. Zubieta, J.-K.; Yau, W.-Y.; Scott, D.J.; Stohler, C.S. Belief or need? Accounting for individual variations in the neurochemistry of the placebo effect. *Brain Behav. Immun.* **2006**, 20 (1), 15–26.
11. Vul, E.; Harris, C.; Winkielman, P.; Pashler, H. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition 1. *Perspect. Psychol. Sci.* **2009**, 4 (3), 274–290.
12. Cheng, B.; Chow, S.C.; Su, W.L. On the assessment of dose proportionality: A comparison of two slope approaches. *J. Biopharm. Stat.* **2006**, 16, 385–392.
13. Tse, S.K.; Chang, J.Y.; Su, W.L.; Chow, S.C.; Hsiung, C.; Lu, Q. Statistical quality control process for traditional Chinese medicine. *J. Biopharm. Stat.* **2006**, 16, 861–874.
14. Cedilnik, A.; Košmelj, K.; Blejec, A. The distribution of the ratio of jointly normal variables. *Metodološki zvezki* **2004**, 1 (1), 99–108.

Minimization Procedure

Dongsheng Tu

NCIC Clinical Trials Group, Queen's University, Kingston, Ontario, Canada

Abstract

This entry illustrates the minimization procedure through an example and discusses its advantages and disadvantages in comparison with stratification and other procedures.

Keywords: Minimization; Randomization; Stratification.

INTRODUCTION

Randomized clinical trials are considered as “golden standard” in evaluating a new treatment or therapy. From the name, it is easy to know that the randomization is one of the most important components in a clinical trial. Randomization procedure was introduced by R.A. Fisher in the 1920s when designing experiments in agricultural studies, and it was first implemented in medical settings by A.B. Hill in the 1960s.^[1] The importance and usefulness of the randomization in clinical trials are addressed in details in several other entries. Basically, randomization prevents biased allocation of subjects to treatment groups and provides the foundation for statistical tests. A simple randomization procedure can be described as an analog to the “flipping of the coin.” A subject would receive a test treatment if the head of the coin appears, or a standard treatment or placebo if otherwise. In theory, this simple randomization procedure will ensure that the treatment groups will be balanced for all covariates including patient and disease characteristics such as age and extent of disease. However, in practice, a chance imbalance may happen with simple randomization, especially when the sample size of the trial is small. If some unbalanced covariates are strongly correlated with the study outcomes, the interpretation of the treatment comparison results might be difficult, and the credibility of the study is also often under question.

A stratified randomization (or stratification, for short) may be used to avoid potential imbalances for some important covariates. These covariates are also called the stratification factors, which have potentially strong relationships with the outcomes of the study and are identified before the study starts. Within individual strata defined by the levels of these stratification factors, the simple randomization procedure, sometimes together with the blocking technique, is used to allocate subjects into different treatment groups. Kernan et al.^[2] reviewed the advantage and disadvantage of the stratification procedure: It can ensure that

treatment groups are balanced in terms of the important covariates that are stratified and would reduce both type I and type II errors; however, because of incomplete filling of blocks within strata, it reduces to simple randomization when the number of strata is large. Therneau^[3] found that the stratification begins to fail if the total number of strata is greater than approximately $n/2$, where n is the total sample size. Hallstrom and Davis^[4] recommend that with blocked stratification, the number of strata should be less than n/K , where K is the block size. Without blocking, the number of strata should be between $n/50$ and $n/100$.

In practice, however, a large number of strata may have to be included in a clinical trial. For multicenter trials, in section 2.3.2 (titled Randomization) of the ICH guideline on Statistical Principles for Clinical Trials,^[5] it is recommended that randomization procedures should be organized centrally and the center should be a stratification factor. Stratification by center would also sometimes facilitate the process of drug supply and distribution. In this case, for a trial of moderate size with more than five participating centers, the number of strata needed would easily exceed the number that was recommended. The minimization method first proposed by Taves^[6] and Pocock and Simon^[7] would be used to deal with these seemingly conflictual demands. This method was popularized by White and Freedman^[8] to biostatistical and medical audiences. It should be noted that the minimization method is one the dynamic allocation procedures in which the allocation of treatment to a subject is influenced by the current balance of allocated treatments and by the stratum to which the subject belongs and the balance within that stratum.^[5] There have been various extensions and generalizations of minimization procedures proposed in the literature. A recent review of these extensions can be found in the paper by Rosenberger and Sverdlov.^[9] In this article, we will illustrate the procedure as originally proposed by Taves^[6] through an example and discuss its advantages and disadvantages in comparison with stratification and other procedures.

PROCEDURE AND ALGORITHM

As mentioned before, the minimization procedure is a method that controls for imbalance within stratification factors in the patients assigned to the different treatment groups. It balances the marginal treatment totals for each level of a stratification factor. The basic idea of this procedure is to assign the next patient to the treatment arm with the smallest sum of marginal total of his/her corresponding level of each stratification factor. If the totals are equal, the treatment will be randomly assigned using the simple randomization method. We use a clinical trial in patients with advanced breast conducted by the NCIC Clinical Trials Group (NCIC CTG MA.19) to illustrate the algorithm for this procedure.^[10]

MA.19 was designed to randomize 350 patients with metastatic and/or recurrent breast cancer equally to receive two different chemotherapies (Arm A: DDPE + DOX and Arm B: DOX only). The study was stopped earlier than planned after 305 patients were accrued. There were four stratification factors identified before the study was opened: (i) whether the patient received prior cytotoxic treatment (yes, no); (ii) whether the patient was registered to MA.16, another NCIC CTG study (yes, no); (iii) whether the patient had visceral disease (yes, no); and (iv) the center the patient was treated (there were 40 centers at initiation of the trial and more would be added during the process of the trial).

In general, suppose that there are a total of M stratification factors and each has h_i levels, where $i = 1, 2, \dots, M$. For MA.19, assume that $i = 1$ for the factor of prior cytotoxic treatment; $i = 2$ for the factor of registration to MA.16; $i = 3$ for the factor of presence of visceral disease; and $i = 4$ for the factor of the center. Then $h_1 = h_2 = h_3 = 2$ and $h_4 = 40$. Define, at any time of the accrual period, $n(m(i), i, k)$ as the total number of patients who were in the $m(i)$ th level of the i th stratification factor and were allocated to receive the k th treatment, where $m(i) = 1, 2, \dots, h_i$ for $i = 1, 2, \dots, M$ and $k = 1$ or 2 ($k = 1$ means the patient was allocated to arm A; $k = 2$ means the patient was allocated to arm B). For example, in MA.19, assume that for $1 \leq i \leq 3$, $m(i) = 1$ if the level is “yes” and $m(i) = 2$ if the level is “no,” and for $i = 4$, $m(i)$ is the number of the center (labeled from 1–40). Then if $i = 2$, $m(2) = 1$, and $k = 2$, $n(m(2), 2, 2)$ is the total number of patients who had registered to MA.16 and had been allocated to treatment arm B.

The patient allocation process in minimization procedure starts with assigning the first patient to a treatment determined by the simple randomization procedure such as “flipping of a coin.” Then at any future time when a new patient arrives, his/her values of each stratification factor are recorded. Suppose that a new patient arrives with $m(i)$ as the value for the level of the stratification factor i . Then we calculate for each treatment k , $T(k) = \sum_{i=1}^M n(m(i), i, k)$ the sum of the total number of patients with value $m(i)$ for the i th stratification factor, where $i = 1, 2, \dots, M$, and

allocated to treatment k (a given patient may be counted multiple times). The following example illustrates the calculations of $T(k)$. Suppose that a new patient to be allocated had prior cytotoxic treatment ($m(i) = 1$ for $i = 1$), was not registered to MA.16 ($m(i) = 2$ for $i = 2$), was presented with visceral disease ($m(i) = 1$ for $i = 3$), and was from center #12 ($m(i) = 12$ for $i = 4$), the following are the number of patients allocated into each treatment arm just before this patient has arrived:

	Arm A	Arm B
$m(1) = 1$	$n(m(1), 1, 1) = 12$	$n(m(1), 1, 2) = 8$
$m(2) = 2$	$n(m(2), 2, 1) = 11$	$n(m(2), 2, 2) = 12$
$m(3) = 1$	$n(m(3), 3, 1) = 4$	$n(m(3), 3, 2) = 3$
$m(4) = 12$	$n(m(4), 4, 1) = 4$	$n(m(4), 4, 2) = 6$

From the above table, we can immediately get $T(1) = 12 + 11 + 4 + 4 = 31$ and $T(2) = 8 + 12 + 3 + 6 = 29$.

After the calculations of $T(1)$ and $T(2)$, based on the minimization principle, the new patient will be assigned to the treatment group with lower $T(i)$. In the last example, the new patient would be assigned to treatment arm B. If $T(1) = T(2)$, the treatment for the new patient will be assigned using a simple randomization procedure. This process is repeated whenever there is a patient waiting for allocation.

We can also illustrate the details of above process using a small trial with ten patients and one single stratification factor, which is the gender of the patient and has two levels (male vs. female).

Assume the first patient is a male. Because he is the first patient who enters the study, we have to allocate treatment to him using a simple randomization procedure. Assume the outcome from this simple procedure is treatment 1. Now assume the second patient to be allocated is a female. As no female patient has been randomized before, we have $T(1) = T(2) = 0$. Therefore, we also have to allocate this patient using the simple randomization procedure. Assume the outcome is treatment 1 again. If the third patient to be allocated is a female again, she will be allocated to treatment 2 because one female patient has been previously allocated to treatment 1 and no female patient to treatment 2 (i.e., $T(1) = 1$ and $T(2) = 0$). Now assume the fourth new patient is a male. He will also be allocated to treatment 2 because $T(1) = 1$ and $T(2) = 0$. If the fifth new patient is a male again, he will have to be assigned a treatment by the simple randomization procedure because each treatment group has one male patient, which leads to $T(1) = T(2) = 1$. Let the outcome be treatment 1. Now assume the next patient is a female, we will also have to allocate her by the simple randomization procedure since we also have $T(1) = T(2) = 1$. Assume the outcome is treatment 2 and next patient is a male. Because two male patients have now been allocated to treatment 1 and one to treatment 2 (i.e., $T(1) = 2$ and $T(2) = 1$), this patient will be allocated

to treatment 2. If the next patient is a female, she will be allocated to treatment 1 because one female patient has been allocated to treatment 1 and two to treatment 2 (i.e., $T(1) = 1$ and $T(2) = 2$). Now assume last two patients are both male. We will have to assign the treatment to the first one using the simple randomization procedure because two male patients have been allocated to each of two treatments (i.e., $T(1) = T(2) = 2$). If the outcome of the simple randomization procedure is treatment 2, the last patient will be allocated to treatment 1 because now we have $T(1) = 2$ and $T(2) = 3$. At the end, three of six male patients and two of the four female patients are allocated to each of the two treatment groups.

The above process can be easily programmed using any computer language that can generate object-oriented user interfaces, such as Visual Basic, JAVA, or SAS/AF, and implemented with telephone or web-based randomization systems. The procedure can also be easily generalized in the following situations: where there are more than two treatment arms, or it is desired to allocate the patients into treatment groups with unequal probabilities.

COMPARISON WITH OTHER PROCEDURES

Many studies using the minimization procedure have demonstrated the ability of minimization to achieve balance. Table 1 the final balance achieved in the MA.19 study (center is summarized only by Canadian vs. non-Canadian), which shows an excellent balance in all the level of the stratification factors.

Pocock and Simon^[7] showed that minimization might be more effective than stratification when trials are small (<100 patients) and there are many (>3) covariates. Therneau^[3] demonstrated that minimization outperformed stratification under a wide range of plausible conditions with

respect to achieving overall balance on the distribution of stratification factors between treatment groups. He concluded that minimization can accommodate a large number of factors from 10–20 without difficulty; however, as mentioned before, stratification begins to fail if the total number of distinct combinations of factor levels (called also as strata) is greater than approximately $n/2$, where n is the total sample size. Stratification is aimed at achieving balance between treatment groups within each combination of the stratification factor levels. Minimization, however, balances only with the marginal distribution of these factors, i.e., within each level of the stratification factors.

Birkett^[11] examined the comparative effects of stratification and minimization on the type I and II errors. In a series of simulations, he found that minimization performed at least as well as stratification, but pointed out that his conclusions were limited to the condition he examined, i.e., normally distributed and independent variables. Tu et al.^[12] reported results from simulation studies which compared the impact of minimization on statistical efficiency, which are measured by the simulated type I and II errors of statistical tests for the comparisons of treatment groups formed by the following methods: (i) random allocation with blocking, (ii) stratification with blocking, and (iii) minimization. The results showed that whether minimization increases precision (i.e., reduces type I or II errors) depends the presence or the absence of interaction among the stratification factors. The minimization procedure is better than stratification, only if the interaction is not substantial, possibly because of the inability of minimization to achieve balance among all the strata. Signorini et al.^[13] recently proposed a new dynamic allocation method as an alternative to minimization. With this method, balance in each stratum can also be achieved. To use this method, the stratification factors are first divided into different levels. For example, for a trial with two stratification factors—gender and hospital—gender within hospital could be considered as level 1; hospital as level 2; and the overall trial as level 3. Define $D_i = |T_i - C_i|$ as the difference of the numbers of patients allocated to the treatment and the control groups at level i , respectively, where T_i and C_i are the numbers allocated to the treatment and the control groups at level i , respectively. Define a critical value d_i for each level. If level i is the lowest level such that $D_i \geq d_i$, force the allocation of the next patients so as to reduce D_i . If $D_i < d_i$ for all i , then randomly allocate the patient. Simulations showed that this method could achieve stratum balance. However, there are still many practical issues to be resolved, e.g., how d_i could be objectively determined.

DISCUSSION

The ICH guideline on Statistical Principles for Clinical Studies^[5] and the European Committee for Proprietary Medicinal Products (CPMP) Points to Consider

Table 1 Accrual by Stratification Factor in MA.19

	Number of patients (%)		
	DPPE/DOX	DOX	TOTAL
Total randomized	153	152	305
<i>Prior cytotoxic treatment</i>			
Yes	20 (13.1)	19 (12.5)	39 (12.8)
No	133 (86.9)	133 (87.5)	266 (87.2)
<i>Registration to MA.16</i>			
Yes	7 (4.6)	6 (3.9)	13 (4.3)
No	146 (95.4)	146 (96.1)	292 (95.7)
<i>Presence of visceral disease</i>			
Yes	96 (62.7)	95 (62.5)	191 (62.6)
No	57 (37.3)	57 (37.5)	114 (37.4)
<i>Center</i>			
Canadian	24 (62.7)	22 (62.5)	46 (62.6)
Non-Canadian	129 (37.3)	130 (37.5)	159 (37.4)

document on Adjustment for Baseline Covariates^[14] discussed specifically the use of dynamic allocation procedures, of which the minimization procedure is a typical example. The ICH guideline warns that “deterministic dynamic allocation procedures should be avoided and an appropriate element of randomization should be incorporated for each treatment allocation” and CPMP document strongly discourages the use of dynamic allocation methods. The minimization procedure described above could be considered as a deterministic procedure, except in the situation when $T(1) = T(2)$, where a subject will be randomly allocated, because the values of the stratification factors for all the subjects previously allocated will determine the allocation for a new subject. For this reason, some people suggested that a simple randomization procedure be performed after $T(1)$ and $T(2)$ are calculated to decide the allocation for a new subject. In our opinion and from our experience, this may destroy the whole purpose of the minimization procedure and thereby reduces it to a simple randomization procedure. At the end, even the marginal balance may not be achieved. In order to avoid the predictability of a deterministic procedure, such as minimization, we advise that it is used only in multicenter trials where there is a central office responsible for treatment allocations. In this case, a separate system group and a sophisticated computer program will keep not only investigators in the centers but also data management personnel at the central office blind to the treatment code. Another proposal is to allocate the subject to the treatment group of lower marginal total with higher probability rather than certainty.^[7,15] However, it is still not clear what would be the best choice for this probability, which may depend on the size of the trial.^[16]

Even if the deterministic aspect of the dynamic allocation methods is avoided by the methods described above, the CPMP document still discourages the use of dynamic allocation methods because “it remains controversial whether the analysis adequately reflects the randomization scheme.”^[14] There have been claims that the analysis based on permutation tests should be used if minimization is used as the method of treatment allocation instead of asymptotic tests. Buyse^[15] has argued that “simulations show that it does not make any material difference what type of test is used.” Kalish and Begg^[17] have shown by simulations that the treatment allocation method has little, if any, effect on the size and power of the tests used to analyze the results of the trial.

Another warning in the ICH guideline is that “the complexity of the logistics and the potential impact on the analysis should be carefully evaluated when considering dynamic allocation.” Logistically, with a central randomization office, our experience shows that the minimization procedure would make it easier for the management of drug supply and distribution and for the handling of randomization and unblinding requests. Partly due to this, the minimization procedure has become a standard method

for many Cooperative Cancer Research Groups supported by the U.S. National Cancer Institute.

When a clinical trial is designed, however, whether the study will be stratified and which method will be used for the allocation of patients are usually not taken into account in the sample-size calculations. The practical reason in most of the situations is that it is impossible or difficult to estimate correctly the percentages of the patients who will be accrued into each stratum. Another reason behind this practice is that stratified tests that take into consideration of stratification are usually more powerful than those tests that ignored stratification factors. Thus, the sample size calculated without the information of strata is usually more conservative if these stratified tests will be used in the final analysis. It has also been observed that even slight imbalances in important prognostic factors might cause serious bias treatment comparisons and substantial loss of power.^[18] Therefore, it is important to identify important prognostic factors before the study and to use a method, such as minimization, that balances the treatment arms in terms of these important prognostic factors.

CONCLUSION

The minimization procedure would be an alternative to the standard stratification procedure when there are many factors to be stratified, or each factor has a large number of levels, and there is a central randomization office which performs the allocation. As for the stratified randomization, it is important to identify stratification factors that are strongly related to the study outcomes before the study starts. All the levels of stratification factors should be clearly defined, and the assessment of these factors should not contain difficult measures to avoid allocating patients to the wrong strata. The computer program that implements the minimization procedure should be carefully validated according to appropriate regulatory guidance.

REFERENCES

1. Armitage, P. The role of randomization in clinical trials. *Control Clin. Trials* **1982**, *1*, 345–352.
2. Kernan, W.N.; Viscoli, C.M.; Makuch, R.W.; Brass, L.M.; Horwitz, R.I. Stratified randomization for clinical trials. *J. Clin. Epidemiol.* **1999**, *52*, 19–26.
3. Therneau, T.M. How many stratification factors are “too many” to use in a randomization plan? *Control Clin. Trials* **1993**, *14*, 98–108.
4. Hallstrom, A.; Davis, K. Imbalance in treatment assignments in stratified block randomization. *Control Clin. Trials* **1988**, *9*, 375–382.
5. Ich Statistical principles for clinical trials. *Stat. Med.* **1999**, *18*, 1905–1942.
6. Taves, D.R. Minimization: A new method of assigning patients to treatment and control group. *Clin. Pharmacol. Ther.* **1974**, *15*, 443–453.

7. Pocock, S.J.; Simon, R.M. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trials. *Biometrics* **1975**, *31*, 103–115.
8. White, S.J.; Freedman, L.S. Allocation of patients to treatment groups in a controlled clinical study. *Br. J. Cancer* **1978**, *37*, 849–857.
9. Rosenberger, W.F.; Sverdlov, O. Handling covariates in the design of clinical trials. *Stat. Sci.* **2008**, *23*, 404–419.
10. Reyno, L.; Seymour, L.; Tu, D.; Dent, S.; Gelmon, K.; Walley, B.; Pluzanska, A.; Gorbunova, V.; Garin, A.; Jassem, J.; Pienkowski, T.; Dancey, J.; Pearce, L.; MacNeil, M.; Marlin, S.; Lebwohl, D.; Voi, M.; Pritchard, K. Phase III study of N, N-Diethyl-2-[4-(Phenylmethyl) Phenoxy] Ethanamine (BMS217380-01) combined with doxorubicin versus doxorubicin alone in metastatic/recurrent breast cancer: A National Cancer Institute of Canada Clinical Trials Group study MA.19. *J. Clin. Oncol.* **2004**, *22*, 269–276.
11. Birkett, N.J. Adaptive allocation in randomized clinical trials. *Control Clin. Trials* **1985**, *6*, 146–155.
12. Tu, D.; Shalay, K.; Pater, J. Adjusting treatment effects for covariates in clinical trials: Statistical and regulatory issues. *Drug Inf. J.* **2000**, *34*, 511–523.
13. Signorini, D.F.; Leung, O.; Simes, R.J.; Beller, E.; Gebski, V.J.; Gallagher, T. Dynamic balanced randomization for clinical trials. *Stat. Med.* **1993**, *12*, 2343–2350.
14. Committee for proprietary medicinal products (CPMP) points to consider on adjustment for baseline covariates. *Stat. Med.* **2004**, *23*, 701–709.
15. Buyse, M. Centralized treatment allocation in comparative clinical trials. *Appl. Clin. Trials* **2000**, *9* (6), 32–37.
16. Hagino, A.; Hamada, C.; Yoshimura, I.; Ohashic, Y.; Sakamoto, J.; Nakazato, H. Statistical comparison of random allocation methods in cancer clinical trials. *Control Clin. Trials* **2004**, *25*, 572–584.
17. Kalish, L.A.; Begg, C.B. The impact of treatment allocation procedures on nominal significance levels and bias. *Control Clin. Trials* **1987**, *8*, 121–135.
18. Akazawa, K.; Nakamura, T.; Palesch, Y. Power of log-rank test and Cox regression model in clinical trials with heterogeneous samples. *Stat. Med.* **1997**, *16*, 583–597.

Minimum Effective Dose

Keumhee Carriere Chough

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada

Abstract

The minimum effective dose is defined as the lowest dose level of a pharmaceutical product that provides a clinically significant response in average efficacy. This entry briefly summarizes how the determination of the MED is made.

Keywords: Analysis of variance; Dose-response; Model-based approach.

INTRODUCTION

Moore¹ has pointed out that the difference between a drug and a poison is the dose. To provide an effective and safe treatment of a certain disease, it is extremely important to identify the dosing range of a pharmaceutical product. The lower bound of the dosing range is referred to as the minimum effective dose (MED). The MED is then defined as the lowest dose level of a pharmaceutical product that provides a clinically significant response in average efficacy, which is also statistically significantly superior to the response provided by the placebo.^[2–5] According to this definition, the MED must produce a response with a magnitude of clinical superiority over placebo because a small but statistically significant response resulting from either a large sample size or a small variability can be of no clinical application. On the other hand, the MED must also provide a statistically significant clinical response. This is because a large clinical response, while statistically insignificant from the placebo response produced at a dose level, cannot establish the scientific evidence of the effectiveness at that dose level. Similarly, the maximum tolerable dose (MTD) is the highest possible but still tolerable dose level with respect to a pre-specified clinical limiting toxicity.^[6,7] The therapeutic range (window) is then defined as the range of the dose levels from the MED to the MTD. If the difference between the MTD and the MED is large, then the corresponding pharmaceutical product is said to have a wide therapeutic range. On the other hand, if the MTD is very close to or even smaller than the MED, then the product is practically of little therapeutic value. The definition discussed above focuses on the average efficacy of a patient population. Fillon^[3] has defined the MED for a particular patient as the lowest dose level that provides a pre-specified clinical effectiveness above a pre-determined percentage of patients. We refer to the first traditional definition as the population minimum

effective dose (PMED) and the second definition as the individual minimum effective dose (IMED).

DESIGN AND ANALYSIS

The MED is usually estimated based on the data of primary efficacy end points collected from dose-ranging or dose-response clinical trials conducted during the Phase II clinical development of a drug product. These dose-response studies are usually randomized, double-blind, parallel-group designs comparing several doses to a concurrent placebo. Occasionally, a crossover design such as William's design^[8] is employed. Many clinicians, however, find that a variety of titration designs either with or without a concurrent parallel placebo group for dose ranging studies are useful because of the impression that they mimic the clinical practices in the real world. The ICH E4 guidance on "Dose-Response Information to Support Drug Registration"^[9] classifies the dose-response designs as follows: (i) parallel dose-response design, (ii) crossover dose-response design, (iii) forced titration design, and (iv) optional titration design (placebo-controlled titration to end points).

As pointed by the ICH E4 guidance, a randomized multiple crossover design of several doses can be employed if drug effect develops rapidly and patients return to baseline conditions quickly after termination of therapy, if responses are not irreversible (e.g., cure or death), and if patients have reasonably stable disease. The primary advantages of crossover design can be summarized as follows: (i) the estimation of dose-response curve and MED depends upon intra-subject variability and hence requires fewer patients than the parallel-group design; (ii) because each patient receives several different doses, the distribution of individual dose-response curves and IMED, as well as the population MED and average dose-response curve, may be estimated; and (iii) as compared to titration designs,

dose and time are not confounded with each other and carryover effects can be better characterized. However, on the other hand, the potential drawbacks of crossover dose-response designs are as follows: (i) there may be analytic problems, if there are too many dropouts; (ii) the duration of the study may be too long for certain patients to endure; and (iii) there is a possibility of carryover effects, baseline comparability, and treatment-by-period interaction.^[9]

It is crucial to include a concurrent placebo group to estimate the MED because a dose of any drug product cannot establish its effectiveness without the comparison with the placebo. Furthermore, the ICH E4 guidance also suggests the inclusion of one or more doses of an active drug control. The inclusion of both placebo and active concurrent control groups allows an evaluation of assay sensitivity and an assessment for detecting the difference between an ineffective drug and an ineffective study. In addition to choosing an appropriate statistical design, the selection of dose levels, the number of dose levels, and the sample sizes for each dose group are equally important for the estimation of the MED. These issues are not only related to each other but also very difficult to deal with. The dose range should be chosen as wide as possible within the safety limit so that a dose-response relationship can be adequately characterized. Although a linear relationship can be derived from the response of two active doses, the ICH E4 guidance^[9] indicates that this approximation is usually not sufficiently informative. In addition, a positive slope can still provide evidence of a drug effect even without placebo concurrent control. It cannot be used for the estimation of MED that requires a clinically significant and meaningful response which is statistically significant superiority over the placebo. Furthermore, a difference between drug groups and placebo control unequivocally shows the effectiveness of the drug. The number of dose levels, including the active agent and the placebo, therefore should be at least four. Ideally, it is preferable to select a dose level whose response is not expected to be statistically different from the placebo response so that the MED can be more accurately estimated. Sample size determination includes an estimation of the total sample size and its distribution across different dose groups. Ruberg^[4] has suggested that the sample size should be determined based on the statistical test for the hypothesis of interest for primary efficacy clinical end points. However, it should be noted that the sample size required for a dose-response study may be different from that for the estimation of MED. Further research for the sample size for the estimation of MED is needed.

Two approaches are commonly used to estimate the MED. The first approach is the method of hypotheses testing based on the analysis of variance (ANOVA). It includes step-down, step-up, and single-step procedures. The step-down procedures consist of the Dunnett's step-down procedure, Williams's test for ordered alternative,^[10] and the step-down linear contrasts.^[11,12] The Bonferroni procedure proposed by Hochberg,^[13] which was later refined by

Dunnett and Tamhane^[14] in conjunction with the application of Helmert contrasts, is a typical step-up procedure for the estimation of the MED. Ruberg^[15] introduced the use of the step contrasts and basin steps as single-step procedures to estimating the MED. The MED estimated by the method of ANOVA would be one of the dose levels evaluated in the trial. The interval estimation for the MED is not yet developed for the approach by the ANOVA. Non-parametric methods based on the Mann-Whitney statistic for the identification of MED are also available^[16,17] when the normality assumption of the primary end point is in doubt. The other approach is the model-based method. Although a functional form such as the four-parameter logistic function is generally required, both point and interval estimates can be obtained by the model-based approach.^[18] In addition, one can more directly incorporate the information of the effective response into the assumed model for the estimation of the MED. Ruberg^[4,18] has provided a comprehensive review of the design and the estimation of the MED.

CONCLUSION

The clinically meaningful response can be expressed either by the original actual response at a particular dose level or as an additional clinically meaningful improvement over the placebo response. All approaches described above assume the clinically significant response a priori as a known constant. This assumption relies on the external validity of the past history and previous experience of the disease under study. This assumption is reasonable only if the placebo response is well established by adequate, well-controlled studies and the clinical condition for the evaluation of the pharmaceutical product is well understood. On the other hand, with respect to internal validity, the clinically significant response should be determined by the response provided by the current placebo control group. Further research in this area is needed.

REFERENCES

1. Moore, T.J. *Deadly Medicine*; Simon and Schuster: New York, 1995.
2. Rodda, B.E.; Tsianco, M.C.; Bolognese, J.A.; Kersten, M.K. Clinical Development. In *Biopharmaceutical Statistics for Drug Development*; Peace, K.E., Ed.; Marcel Dekker, Inc: New York, 1988.
3. Fillon, T.G. Estimating the minimum therapeutically effective dose of a compound via regression modeling and percentile estimation. *Stat. Med.* **1995**, *14*, 925–932.
4. Ruberg, S.J. Dose response studies: I. Some design considerations. *J. Biopharm. Stat.* **1995**, *5*, 1–14.
5. Chow, S.C.; Liu, J.P. *Design and Analysis of Clinical Trials: Concepts and Methodologies*; John Wiley and Sons: New York, 1998.

6. Storer, B.E. Design and analysis of phase I trials. *Biometrics* **1989**, *45*, 925–937.
7. Korn, E.L.; Midthune, D.; Chen, T.T.; Rubinstein, L.V.; Christian, M.C.; Simon, R. A comparison of two phase I designs. *Stat. Med.* **1994**, *13*, 1799–1806.
8. Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*; Marcel Dekker, Inc: New York, 1999.
9. ICH International Conference on Harmonisation Tripartite Guidance E4 on Dose-Response Information to Support Drug Registration. Fed. Regist. **1994**, *59*, 55972–55976. US Government Printing Office: Washington, DC.
10. Williams, D.A. The comparison of several dose levels with a zero dose control. *Biometrics* **1972**, *28*, 519–531.
11. Tukey, J.W.; Ciminera, J.L.; Heyes, J.F. Test the statistical certainty of a response to increasing dose of a drug. *Biometrics* **1985**, *41*, 295–301.
12. Capizzi, T.; Survill, T.T.; Heyse, J.F. An empirical and simulated comparison of some tests for detecting progressiveness of response with increasing doses of a compound. *Biom. J.* **1992**, *34*, 275–289.
13. Hochberg, Y. A sharper Bonferroni's procedure for multiple tests of significance. *Biometrika* **1988**, *75*, 800–802.
14. Dunnett, C.W.; Tamhane, A.C. A step-up multiple test procedure. *J. Am. Stat. Assoc.* **1992**, *87*, 162–170.
15. Ruberg, S.J. Contrasts for identifying the minimum effective dose. *J. Am. Stat. Assoc.* **1989**, *84*, 816–822.
16. Chen, Y.I. Nonparametric identification of the minimum effective dose. *Biometrics* **1999**, *55*, 1236–1240.
17. Chen, Y.I. Rank-based tests for dose finding in non-monotonic dose-response settings. *Biometrics* **1999**, *55*, 1258–1262.
18. Ruberg, S.J. Dose response studies: II. Analysis and interpretation. *J. Biopharm. Stat.* **1995**, *5*, 15–42.

Mixed Effects Models

Donghui Zhang and Chunpeng Fan

Sanofi Aventis, Bridgewater, New Jersey, U.S.A.

A. Lawrence Gould

Merck Research Laboratories, Blue Bell, Pennsylvania, U.S.A.

Abstract

This entry briefly outlines linear and non-linear mixed effect models by providing background information and estimation models for each.

INTRODUCTION

Mixed effect models have been the subject of active theoretical research for about three decades. The increasing availability of software for performing the intensive calculations required to implement mixed effect model methods has made it practical to implement these methods in a wide variety of applications. One area of application for which mixed effect models have been particularly well suited is that of repeated measurement trials, in which a subject receives a treatment or a sequence of treatments repeatedly over a period of time. Typical examples of repeated measurement trials include longitudinal trials of the effect of long-term administration of a treatment and pharmacokinetic trials in which sequences of blood samples are obtained to study the time course of the concentration of a drug in the blood following one or more applications of the drug. Another area of application for mixed effect models is in accounting for excess variation attributable to intraclass correlation arising from clustering. Typical examples include members of a family, students within a class and classes within a school, patients within a center contained in a multicenter trial, etc. Mixed effect models express an observed response explicitly as a function of fixed effects such as administered treatment, gender, and age, and of random effects such as measurement error, biological variation within a subject, or population variation among subjects. Mixed effect models may be linear, when the response is expressed as the sum of fixed and random effects, or non-linear, when the response is a non-linear function of the fixed and random effects. The application of mixed effect model methods requires assumptions about the distributions of the random effects. The usual assumption is that the effects are normally distributed, although other distributional assumptions may be made depending on the application. This entry briefly outlines linear and non-linear mixed effect models. More comprehensive treatments of the issues may be found in Bryk and Raudenbush,^[1] Khuri et al.,^[2] Rao and Kleffe,^[3] Searle,^[4]

and Searle et al.,^[5] which deal primarily with linear mixed effect models; Gallant,^[6] Carroll et al.,^[7] Davidian and Giltinan,^[8] and Vonesh and Chinchilli,^[9] which address non-linear models; and Andersen^[10] and McCullagh and Nelder,^[11] McCulloch and Searle,^[12] Diggle et al.,^[13] Hardin and Hilbe,^[14] and Dobson and Barnett,^[15] which address generalized linear mixed effect model methods that are appropriate when the data are discrete, such as categories, rather than continuous. A comprehensive coverage on the linear mixed effects model, the generalized linear mixed model, and the non-linear mixed model can be found in Demidenko.^[16] Random effects can be introduced into survival-type models to account for missing covariates and clustering.^[17–34]

LINEAR MODELS

Background

Conventional general fixed effect linear models can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

where $\mathbf{y}$ denotes a vector of observations, $\boldsymbol{\beta}$ denotes a vector of unknown constants (parameters), $\mathbf{X}$ denotes a matrix of (known) covariates or indicator variables, and $\boldsymbol{\varepsilon}$ denotes a vector of unknown residual errors. The elements of $\boldsymbol{\varepsilon}$ are ordinarily assumed to be independently distributed with zero mean and constant variance σ^2 . Circumstances in which a more complex covariance structure for $\boldsymbol{\varepsilon}$ would be appropriate can usually be approached effectively using mixed model methods as described below. Methods for estimating the elements of $\boldsymbol{\beta}$ under a variety of assumptions about the distribution of the elements of $\boldsymbol{\varepsilon}$ are well known.^[4] Non-linear fixed effect models usually replace the $\mathbf{X}\boldsymbol{\beta}$ term in Eq. 1 with a non-linear function $f(\mathbf{X}, \boldsymbol{\beta})$ of parameters and covariates, although they can

also be formulated in other ways such as through link functions.^[11]

In applications such as the longitudinal studies, when the aim is to study the evolution of response over time, the observation vector $\mathbf{y}$ may consist of sequences of values, each sequence corresponding to a different subject or patient. At least two sources of variability arise in such cases. One source of variability corresponds to the variation among a subject's sequence of observations beyond what the fixed effects would account for. Another source of variability is the variation between the observations on different subjects, which often will be greater than the variability among the observations made on the same subject. The idea that the observations could be functions of various kinds of random effects as well as fixed effects underlies mixed effect models.

Random effects can be incorporated into linear models by generalizing Eq. model 1 slightly to express a vector of observations as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon} \quad (2)$$

The essential difference is the addition of the term $\mathbf{Z}\boldsymbol{\delta}$, where $\mathbf{Z}$ is a matrix of known covariates and $\boldsymbol{\delta}$ denotes a vector of random effects, analogous to $\boldsymbol{\varepsilon}$. Expression 2 describes a particularly simple way to incorporate the random effects. Random effects actually can be incorporated into non-linear models in a number of ways, outlined in "Non-linear Models." Ordinarily, the elements of $\boldsymbol{\varepsilon}$ are assumed independently distributed with zero mean and constant variance σ^2 , although more complex covariance structures (e.g., autoregressive) could be assumed. The elements of $\boldsymbol{\delta}$ are also assumed to have zero mean and assumed to be uncorrelated with the elements of $\boldsymbol{\varepsilon}$, but may have a complex covariance structure, with covariance matrix $\mathbf{D}$. More usually, especially in practice, $\mathbf{Z}\boldsymbol{\delta}$ is expressed as

$$\mathbf{Z}\boldsymbol{\delta} = \sum_{i=1}^q \mathbf{Z}_i \boldsymbol{\delta}_i$$

where $\boldsymbol{\delta}$ has m_i elements and covariance matrix $\sigma_i^2 \mathbf{I}_{m_i}$, where $\mathbf{I}_{m_i}$ denotes an $m_i \times m_i$ identity matrix, and the elements $\boldsymbol{\delta}_i$ and $\boldsymbol{\delta}_j$ are uncorrelated if $i \neq j$. Then the covariance matrix $\mathbf{D}$ is block-diagonal with the matrices $\sigma_i^2 \mathbf{I}_{m_i}$, $i = 1, \dots, q$ on its diagonal, and the covariance matrix of the elements of the observation vector $\mathbf{y}$ can be written as

$$\mathbf{V} = \mathbf{V}(\mathbf{y}) = \sum_{i=1}^q \sigma_i^2 \mathbf{Z}_i \mathbf{Z}_i^T + \sigma^2 \mathbf{I}_n \quad (3)$$

Estimation

The objective of the analyses of mixed effect models is inference about the fixed effects $\boldsymbol{\beta}$ and estimation of σ^2 and the distinct elements in the covariance matrix $\mathbf{D}$

(we use vector $\boldsymbol{\theta}$ to represent the variance components). This is commonly accomplished by maximizing the likelihood of the data under a set of distributional assumptions about the random effects such as multivariate normality. The computations are generally not explicit, but instead entail successively more accurate iterative approximations to the solutions of the likelihood equations or estimating equations.^[35–37] The iterative nature of the computations means that a number of numerical analysis issues need to be considered such as convergence (at all, global vs. local), dependence on starting values, potential singularity of $\mathbf{V}$, algorithm used for calculations, effect of boundaries (e.g., variance components must be positive), and computing cost. (See, e.g., Searle et al.,^[5] Chapter 8.)

Maximum Likelihood Estimation

Estimation can be carried out under the assumption of multivariate normality of the observations by the conventional maximum likelihood (ML) or by the restricted maximum likelihood (REML). Conventional maximum likelihood estimates the fixed effect parameters $\boldsymbol{\beta}$ and the variance components σ_i^2 ($i = 1, \dots, q$) as those values that maximize the logarithm of the multivariate normal likelihood

$$L = L(\boldsymbol{\beta}, \mathbf{V} | \mathbf{y}) = -\frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (4)$$

These are usually taken as the values for which the first derivatives of L equal zero; it is necessary to verify that these values really do maximize expression 4 and that the estimates of the variance components are positive. The derivative of L with respect to $\boldsymbol{\beta}$ is

$$L_{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y} - \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \boldsymbol{\beta} \quad (5)$$

and the derivative with respect to σ_i^2 , $i = 0, 1, \dots, q$, is

$$L_{\sigma_i^2} = \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} \mathbf{Z}_i \mathbf{Z}_i^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \text{tr}(\mathbf{V}^{-1} \mathbf{Z}_i \mathbf{Z}_i^T) \quad (6)$$

with the convention that $\sigma_0^2 = \sigma^2$ as defined in Eq. 3 and $\mathbf{Z}_0 = \mathbf{I}_n$. The estimates of the variance components derived by maximizing the log likelihood in Eq. 4 via the first derivatives provided in Eqs. 5 and 6 do not reflect the degrees of freedom involved in estimating $\boldsymbol{\beta}$.

If $\hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\theta}})$ denotes the estimated covariance matrix of the observations using the ML estimator $\boldsymbol{\theta}$ of the variance components, then the maximum likelihood estimator of $\mathbf{X}\boldsymbol{\beta}$ is

$$\text{MLE}(\mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}(\mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{y} \quad (7)$$

and its asymptotic covariance matrix is

$$\mathbf{V}(\text{MLE}(\mathbf{X}\hat{\boldsymbol{\beta}})) = \mathbf{X}(\mathbf{X}^T \hat{\mathbf{V}} - \mathbf{I}_X)^{-1} \mathbf{X}^T \quad (8)$$

REML Estimation

Estimates of the variance components that do reflect the degrees of freedom involved in estimating β are obtained using restricted (or residual) maximum likelihood, usually abbreviated as REML. Restricted maximum likelihood estimates of the variance components are based on the residuals after fitting the fixed effects by ordinary least squares. Thus if $\mathbf{U}^-$ denotes a generalized inverse of a matrix $\mathbf{U}$ ($\mathbf{U}^-$ is any matrix satisfying $\mathbf{U}\mathbf{U}^-\mathbf{U} = \mathbf{U}$), then $\mathbf{r} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y}$ is the vector of residuals corresponding to the observation vector $\mathbf{y}$. The expectation of $\mathbf{r}$ is zero regardless of the value of $\mathbf{V}$. The matrix $\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ has $n - p$ linearly independent rows, where p is the rank of $\mathbf{X}$. Let $\mathbf{M}$ denote an $(n - p) \times n$ matrix whose rows are linearly independent vectors in the row space spanned by the rows of $\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. $\mathbf{M}$ could be any $n - p$ linearly independent rows of $\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. If $\mathbf{w} = \mathbf{M}\mathbf{y}$, then $\mathbf{w}$ represents an observation from a multivariate normal distribution with zero mean and nonsingular covariance matrix $\mathbf{U} = \mathbf{M}\mathbf{V}\mathbf{M}^T$. The logarithm of the likelihood under multivariate normality is

$$L = -\frac{1}{2}(\log|\mathbf{U}| + \mathbf{w}^T\mathbf{U}^{-1}\mathbf{w}) \quad (9)$$

and the derivative with respect to $\sigma_i^2, i = 1, \dots, q$, is

$$L_{\sigma_i^2} = \frac{1}{2}(\mathbf{w}^T\mathbf{U}^{-1}\mathbf{Z}_i\mathbf{Z}_i^T\mathbf{U}^{-1}\mathbf{w} - \text{tr}(\mathbf{U}^{-1}\mathbf{Z}_i\mathbf{Z}_i^T)) \quad (10)$$

The variance components can be estimated from these expressions with attention to the same computational considerations, namely, that the estimates really do maximize expression 9 and the estimates are positive (or at least nonnegative).

Estimates of the fixed effects and the asymptotic covariance matrix of the estimates are given by Eqs. 7 and 8, with $\hat{\mathbf{V}}$ calculated using the REML estimates of the variance components.

Other Estimation Procedures

The parameters can be estimated by means other than the maximum likelihood and the REML. Extended generalized least squares (EGLS) can be used when the observations at least conceptually represent series of observations on individuals.^[38] Write the j th observation on individual i as

$$y_{ij} = f(x_{ij}, \beta_i) + \sigma g(f(x_{ij}, \beta_i), \theta) \varepsilon_{ij} \quad (11)$$

where the parameter vector β_i for individual i is drawn from a distribution (usually assumed p -variate normal) with mean β and covariance matrix Σ . The within-individual errors ε_{ij} are assumed independently and identically

distributed with zero mean and unit variance. The scale parameter σ and the variance function $g()$ allow for heterogeneity because of the intra-individual correlation. An analog to REML estimation can be applied to EGLS estimators as well. The general formulation provides a unified basis for a number of variations on this theme.^[39–45]

The analyses can also be carried out using Bayesian^[46–57] or empirical Bayesian^[58] methods, which start with a likelihood for the observations based on a model such as Eq. 2 or 11, and then append expressions for prior distributions of the various parameters to obtain an expression for the joint distribution of the data and the parameters. Integrating with respect to the model (likelihood) parameters yields a marginal distribution for the observations that depends at most on the (known) parameters of the prior distributions. Dividing the joint distribution of the data and the model parameters by this marginal distribution gives the posterior distribution of the model parameters. Inferences about the parameters and about the functions of the parameters (for example, tolerance limits or predictive limits for future observations) are based on the posterior distribution. The difficult part of the computation is the integration to produce the marginal distribution of the observations. One can partly avoid the problem by approximating the mean of the posterior distribution by its mode.^[55] Current practice employs Markov Chain Monte Carlo techniques^[48,53] that sample from conditional posterior distributions in such a way as to generate samples from the full joint posterior distribution. These can be summarized by appropriate techniques to yield summary statistics and descriptions of the posterior distributions.^[59,60] Markov Chain Monte Carlo methods are powerful tools for analyzing mixed models, both linear and non-linear. However, the results they provide need to be checked carefully to assure that the assumptions on which they are based are met, notably that the realizations from the sampling process comprise independent (or at least uncorrelated) stationary series.

SMALL SAMPLE ADJUSTMENTS

For small sample problems, the ML or REML method for estimating fixed effects β and variance components as well as the inference for fixed effects that are based on the asymptotic distribution of $\hat{\beta}$ may be inadequate when ignoring the variability in the $\hat{\mathbf{V}} = \mathbf{V}(\hat{\theta})$ in Eq. 7. For certain combinations of covariance structure, design, and sample size, this can seriously overestimate the true precision of $\hat{\beta}$. Further, Wald-type test procedures and corresponding confidence intervals that are based on asymptotic chi-squared approximations also ignore the variability in the estimate of variance components.

The investigation of approximating the small sample precision of $\hat{\beta}$ and related distributional issues can be dated back to the work by Harville and coworkers.^[61–66] Kenward and Roger^[67] combined these results and further developed

a method that adjusted the bias of the inference for fixed effects caused by ignoring the variability in the estimate of θ when only limited number of observations was involved in the model.

When EGLS is employed to make the inference for the fixed effects, in addition to $\hat{\mathbf{V}} = \mathbf{V}(\hat{\theta})$, the variability in estimating fixed effects may also contribute to overestimating the true precision through the estimate of its variance-covariance matrix that also depends on $\hat{\beta}$. This problem has been investigated in several studies that have proposed different directions to correct the bias including correcting the empirical residuals utilized to estimate the variance-covariance matrix,^[68–70] correcting the degrees of freedom when making the inference,^[70,71] and using experiencing formulae.^[72]

NON-LINEAR MODELS

As noted in “Introduction,” the essential difference between non-linear and linear models is the expression of the expectation of the observation vector as a non-linear function $f()$ of the covariates and parameters instead of a linear function. The functional form of f may be obtained from previous studies, derived from the theory based on fundamental physical or biological principles, or represent empirical description of observations, e.g., from a graphical display of data. The functional form of f ordinarily is assumed known. Analyses can be carried out without assuming that f is known by using a spline approximation to f or using generalized additive models.^[73]

Random effects can enter the model non-linearly as well as linearly. This occurs, for example, in generalized linear models and in survival analyses, where the random effects scale the hazard function. Inference in these cases can be considerably more complicated than when the effects enter linearly, and standard methods often cannot be applied. Several approaches often used to deal with this problem: extended general linear models as described in “Other Estimation Procedures,” where parameters are estimated for each individual in the sample and are then combined; approximation of the non-linear model by a first-order Taylor series expansion that effectively linearizes it;^[43,74,75] the semi-non-parametric method;^[76] the fully Bayesian analysis,^[77] and Laplace’s approximation.^[78]

SOFTWARE

The computations required for the analysis of mixed models range from intensive to very intensive. Standard software packages usually include some facility for analyzing linear mixed models for continuous, normally distributed data. Some of these packages are commercial, and some are available free of charge. Packages are also available to carry out calculations for non-linear models,

e.g., logistic mixed models, ordinal regression, Poisson regression, etc. The availability of the diagnostic procedures to test the validity of the results of computer-intensive methods is an important consideration in the choice of a software package.

REFERENCES

1. Bryk, A.S.; Raudenbush, S.W. *Hierarchical Linear Models: Applications and Data Analysis Methods*; Sage Publication: Thousand Oaks, CA, 1992.
2. Khuri, A.I.; Mathew, T.; Sinha, B.K. *Statistical Tests in Mixed Linear Models*; John Wiley: New York, 1998.
3. Rao, C.R.; Kleffe, J. *Estimation of Variance Components and Applications*; Elsevier-Science: New York, 1988.
4. Searle, S.R. *Linear Models*; John Wiley: New York, 1997.
5. Searle, S.R.; Casella, G.; McCulloch, C.E. *Variance Components*; John Wiley: New York, 1992.
6. Gallant, A.R. *Nonlinear Models*; Edward Elgar: Northampton, 1997.
7. Carroll, R.J.; Ruppert, D.; Stefanski, L.A. *Measurement Error in Nonlinear Models*; C.R.C. Press: Boca Raton, FL, 1995.
8. Davidian, M.; Giltinan, D.M. *Nonlinear Models for Repeated Measurement Data*; Chapman and Hall: London, 1995.
9. Vonesh, E.F.; Chinchilli, V.M. *Linear and Nonlinear Models for the Analysis of Repeated Measurements*; Marcel Dekker: New York, 1997.
10. Andersen, E.B. *Introduction to the Statistical Analysis of Categorical Data Analysis*; Springer Verlag: New York, 1997.
11. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*; Chapman and Hall: London, 1989.
12. McCulloch, C.E.; Searle, S.R. *Generalized Linear, and Mixed Models*; John Wiley: New York, 2001.
13. Diggle, P.J.; Heagerty, P.; Liang, K.Y.; Zeger, S.L. *Analysis of Longitudinal Data*, 2nd Ed.; Oxford University Press: New York, 2002.
14. Hardin, J.; Hilbe, J. *Generalized Linear Models and Extensions*; Stata Press: Hardin, College Station, 2007.
15. Dobson, A.J.; Barnett, A.G. *Introduction to Generalized Linear Models*, 3rd Ed.; Chapman and Hall: Boca Raton, FL, 2008.
16. Demidenko, E. *Mixed Models Theory and Applications*; John Wiley & Sons, Inc.: New York, 2004.
17. Ducrocq, V.; Casella, G. *Genet. Sel. Evol.* **1996**, *28*, 505–529.
18. Hougaard, P. *Lifetime Data Anal.* **1995**, *1*, 255–273.
19. Keiding, N.; Andersen, P.K.; Klein, J.P. *Stat. Med.* **1997**, *16*, 215–224.
20. King, T.M.; Beaty, T.H.; Liang, K.Y. *Genet. Epidemiol.* **1996**, *13*, 139–158.
21. Lam, K.F.; Kuk, A.C. *J. Am. Stat. Assoc.* **1997**, *92*, 985–990.
22. Liang, K.Y.; Self, S.G.; Bandeen, R.K.; Zeger, S.L. *Lifetime Data Anal.* **1995**, *1*, 403–415.
23. Morris, C.; Christiansen, C. *Lifetime Data Anal.* **1995**, *1*, 347–359.
24. Petersen, J.H. *Biometrics* **1998**, *54*, 646–661.

25. Pickles, A.; Crouchley, R. *Stat. Med.* **1995**, *14*, 1447–1461.
26. Sargent, D.J. *Biometrics* **54**, 1486–1497.
27. Scheike, T.H.; Jensen, T.K. *Biometrics* **1997**, *53*, 318–329.
28. Sinha, D. *Biometrics* **1998**, *54*, 1463–1474.
29. Walker, S.G.; Mallick, B.K. *J. Royal Stat. Soc., Ser. B* **1997**, *59*, 845–860.
30. Xue, X.N.; Brookmeyer, R. *Stat. Med.* **1997**, *16*, 1983–1993.
31. Xue, X.N. *Biometrics* **1998**, *54*, 1631–1637.
32. Yau, K.W.; McGilchrist, C.A. *Stat. Med.* **1998**, *17*, 1201–1213.
33. Yue, H.B.; Chan, K.S. *Biometrics* **1997**, *53*, 785–793.
34. Zehl, P.H. *Stat. Med.* **1997**, *16*, 1573–1585.
35. Burnett, R.T.; Ross, W.H.; Krewski, D. *Environmetrics* **1995**, *6*, 85–99.
36. Kenward, M.G.; Lesaffre, E.; Molenberghs, G. *Biometrics* **1994**, *50*, 945–953.
37. Zeger, S.L.; Liang, K.Y. *Stat. Med.* **1992**, *11*, 1825–1839.
38. Davidian, M.; Giltinan, D.M. *Biometrics* **1993**, *49*, 59–73.
39. Carroll, R.J.; Ruppert, D. *Transformation and Weighting in Regression*; Chapman and Hall: London, 1988.
40. Beal, S.L.; Sheiner, L.B. *Technometrics* **1988**, *30*, 327–338.
41. Peck, C.C.; Beal, S.L.; Sheiner, L.B.; Nichols, A.I. *J. Pharmacokinet. Biopharm.* **1984**, *11*, 303–319.
42. Zeger, S.L.; Liang, K.Y.; Albert, P.S. *Biometrics* **1988**, *44*, 1049–1060.
43. Lindstrom, M.J.; Bates, D.M. *Biometrics* **1990**, *46*, 673–687.
44. Racine-Poon, A. *Biometrics* **1985**, *41*, 1015–1023.
45. Steimer, J.L.; Mallet, A.; Golmard, J.L.; Boisvieux, J.F. *Drug Metab. Rev.* **1984**, *15*, 265–292.
46. Berry, D.A.; Stangl, D.K. *Bayesian Biostatistics*; Marcel Dekker: New York, 1996.
47. Bromeling, L.D. *Bayesian Analysis of Linear Models*; Marcel Dekker: New York, 1984.
48. Carlin, B.P.; Louis, T.A. *Bayes and Empirical Bayes Methods for Data Analysis*; Chapman and Hall: London, 1996.
49. Chaturvedi, A. *J. Stat. Plan. Inference* **1996**, *50*, 175–186.
50. Cowles, M.K.; Carlin, B.P.; Connett, J.E. *J. Am. Stat. Assoc.* **1996**, *91*, 86–98.
51. Gelfand, A.E.; Hills, S.E.; Racine-Poon, A.; Smith, A.M. *J. Am. Stat. Assoc.* **1990**, *85*, 972–985.
52. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*; Chapman and Hall: London, 1995.
53. Gilks, W.R.; Richardson, S.; Spiegelhalter, D.J. *Markov Chain Monte Carlo in Practice*; Chapman and Hall: London, 1996.
54. Harville, D.A.; Zimmermann, A.G. *J. Stat. Comput. Simul.* **1996**, *54*, 211–229.
55. Lindley, D.V.; Smith, A.F.M. *J. Royal Stat. Soc., Ser. B* **1972**, *34*, 1–42.
56. Simon, R.; Freedman, L.S. *Biometrics* **1997**, *53*, 456–464.
57. Waclawiw, M.A.; Liang, K.Y. *J. Am. Stat. Assoc.* **1993**, *88*, 171–178.
58. Maritz, J.S.; Lwin, T. *Empirical Bayes Methods*; Chapman and Hall: London, 1989.
59. Best, N.G.; Cowles, M.K.; Vines, S.K. *CODA: Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output, Version 0.4*; MRC Biostatistics Unit: Cambridge, 1997.
60. Spiegelhalter, D.J.; Thomas, A.; Best, N.G.; Gilks, W.R. *BUGS; Bayesian inference using Gibbs sampling*; MRC Biostatistics Unit: Cambridge, 1997.
61. Kackar, A.N.; Harville, D.A. *J. Am. Stat. Assoc.* **1984**, *79*, 853–862.
62. Harville, D.A. *J. Am. Stat. Assoc.* **1985**, *80*, 132–138.
63. Jeske, D.R.; Harville, D.A. *Commun. Stat. Theory Methods* **1988**, *17*, 1053–1087.
64. Hulting, F.L.; Harville, D.A. *J. Am. Stat. Assoc.* **1991**, *86*, 557–568.
65. Harville, D.A.; Jeske, D.R. *J. Am. Stat. Assoc.* **1992**, *87*, 724–731.
66. Harville, D.A.; Carriquiry, A.L. *Biometrics* **1992**, *48*, 987–1003.
67. Kenward, M.G.; Roger, J.H. *Biometrics* **1997**, *53*, 983–997.
68. Mancl, L.A.; DeRouen, T.A. *Biometrics* **2001**, *57*, 126–134.
69. Kauermann, G.; Carroll, R.J. *J. Am. Stat. Assoc.* **2001**, *96*, 1387–1396.
70. Fay, M.P.; Graubard, B.I. *Biometrics* **2001**, *57*, 1198–1206.
71. Pan, W.; Wall, M.M. *Stat. Med.* **2002**, *21*, 1429–1441.
72. Morel, J.G.; Bokossa, M.C.; Neerchal, N.K. *Biom. J.* **2003**, *45*, 395–409.
73. Hastie, T.J.; Tibshirani, R.J. *Generalized Additive Models*; Chapman and Hall: London, 1990.
74. Boeckmann, A.J.; Sheiner, L.B.; Beal, S.L. *NONMEM Users guide. NONMEM Project Group C255*; University of San Francisco: San Francisco, CA, 1992.
75. Vonesh, E.F.; Carter, R.L. *Biometrics* **1992**, *48*, 1–17.
76. Davidian, M.; Gallant, A.R. *Biometrika* **1993**, *80* (3), 475–488.
77. Wakefield, J.C.; Smith, A.F.M.; Racine-Poon, A.; Gelfand, A.E. *Appl. Statist.* **1994**, *43*, 201–221.
78. Wolfinger, R. *Biometrika* **1993**, *80* (4), 791–795.

Mixed-Outcome Data

A. R. de Leon

Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada

Abstract

In this entry is provided an up-to-date, comprehensive, and unified overview of available strategies for modeling and analyzing mixed outcomes, highlighted by connections between various approaches and research threads in the literature, paying particular attention to their advantages and disadvantages.

Keywords: Copulas; General location model; Logistic regression; Maximum likelihood; Poisson outcomes; Probit regression; Pseudolikelihood; Random-effects models.

INTRODUCTION

Mixed outcomes are ubiquitous in applications in health and medicine, and joint analysis of such outcomes entails specification of models flexible enough to accommodate them. Such joint models are potentially advantageous in several statistical and practical respects. For example, intrinsically multivariate questions concerning relationships between outcomes and the joint influence of covariates on them may be easily answered by fully exploiting the multivariate nature of the data through joint models. From a statistical standpoint, joint analysis avoids multiple testing and naturally leads to global tests, thus resulting in increased power and better control of type I error rates. Significant efficiency gains over separate univariate analyses have also been reported, especially in settings where there are missing data (e.g., Ref. [1]). The following are examples of studies where mixed outcomes are of interest.

Example 1

(Macular degeneration study) Longitudinal data are obtained from a two-armed randomized multi-center clinical trial on visual acuity of patients suffering from macular degeneration. In the trial, a patient's visual acuity was assessed at different time points through his ability to read lines of letters on standardized vision charts displaying line letters of decreasing size that he must read from top (largest letters) to bottom (smallest letters). Each line with at least four letters correctly read is called one line of vision. The data consist of a true count endpoint, corresponding to the total number of letters correctly read, and a surrogate binary endpoint, defined as the loss of at least three lines of vision at one year compared with their baseline performance. The study aim is to investigate the relationship between the mixed endpoints. See also Molenberghs and Verbeke^[2] and Pinto and Normand.^[3]

Example 2

(Irwin's toxicity study) The data come from a three-day repeated dose-toxicity study on neurofunctional effects of a psychotropic drug on rats. The data consist of several binary outcomes relating to behavior, including locomotor activity—characterized by the animal's abnormal biting, restlessness, writhing—and positional passivity—defined as an animal's struggle response to sequential handling—and a number of continuous outcomes, such as body temperature, pupil size, and so forth. The goal of the study is to determine treatment effects as well as the association between certain mixed outcomes. Details are given in Faes et al.^[4]

Example 3

(Veterans' Health Administration health performance monitoring study) The study collected monitoring data on health performance of several geographically defined service networks in the United States, with the objective of characterizing the quality of care of different service networks. The observed data consist of multi-level continuous and discrete outcomes collected repeatedly within clusters over time, and consist of mixed continuous and binary variables on visits, days between visits, readmissions, days between readmissions, and so on, from a sample of veterans. The goal of the study is to characterize trends in quality of individual service networks on the basis of the mixed outcomes and to identify problems in quality for specific networks. Details are found in Daniels and Normand.^[5]

In analyzing such mixed-outcome data, emphasis is usually placed on determining the joint distribution of the mixed outcomes, from which are obtained specific aspects of their relationships, such as marginal and conditional distributions, and associations. Most research dealing with statistical problems associated with joint analysis of mixed outcomes has appeared only recently due to difficulties in

specifying a mixed-outcome joint distribution and the relative lack of standard models. This is further compounded in longitudinal and cluster data settings in medical and health studies, where associations between mixed outcomes at various levels need to be meaningfully delineated in the analysis.

In this contribution, we provide an up-to-date, comprehensive, and unified overview of available strategies for modeling and analyzing mixed outcomes. We highlight connections between various approaches and research threads in the literature, paying particular attention to their advantages and disadvantages. Several examples provide background material on the assorted challenges that arise in the analysis of mixed outcomes.

JOINT MIXED-OUTCOME MODELS

A number of joint modeling strategies for mixed outcomes have been studied in the literature. The general approach first specifies a model for the joint distribution of mixed outcomes, then fits the model to data at hand, and finally uses the model to draw inferences. The challenge is that such multivariate distributions are uncommon.

Consider a mixed-data set-up represented by the vectors $\mathbf{X}_i$ and $\mathbf{Y}_i$ denoting discrete (e.g., nominal or ordinal categorical variables, counts) and continuous outcomes, respectively, observed from subject $i = 1, \dots, n$. Joint analysis of X_i and Y_i requires either direct or indirect specification

of the joint density $f_{\mathbf{X}_i, \mathbf{Y}_i}(\mathbf{x}, \mathbf{y})$. Models for the analysis of continuous outcomes have been well studied; however, only recently have researchers begun tackling the corresponding problems associated with discrete outcomes. Examples of earlier work in this area include the multivariate probit model and the so-called Dale model (see^[2] for more details). Although these have paved the way for advances in mixed-outcome data analysis,^[6,7] there is no standard model similar to the multivariate normal distribution for continuous data or the multinomial distribution for discrete variables, resulting in a dearth of models. A taxonomy of currently available modeling approaches for mixed outcomes is shown in Fig. 1.

In the next section, we survey various models for $f_{\mathbf{X}_i, \mathbf{Y}_i}(\mathbf{x}, \mathbf{y})$, many developed for specific applications and mostly based on some factorization of it into marginal and conditional components. A number of issues concerning their application to joint regression modeling of mixed-outcome data are discussed. These issues involve model specification as well as ensuing inference based on such models.

FACTORIZATION MODELS

One of the earliest proposals of directly specifying the joint distribution factorizes it into a conditional distribution of one set of outcomes and a marginal distribution of the other set.

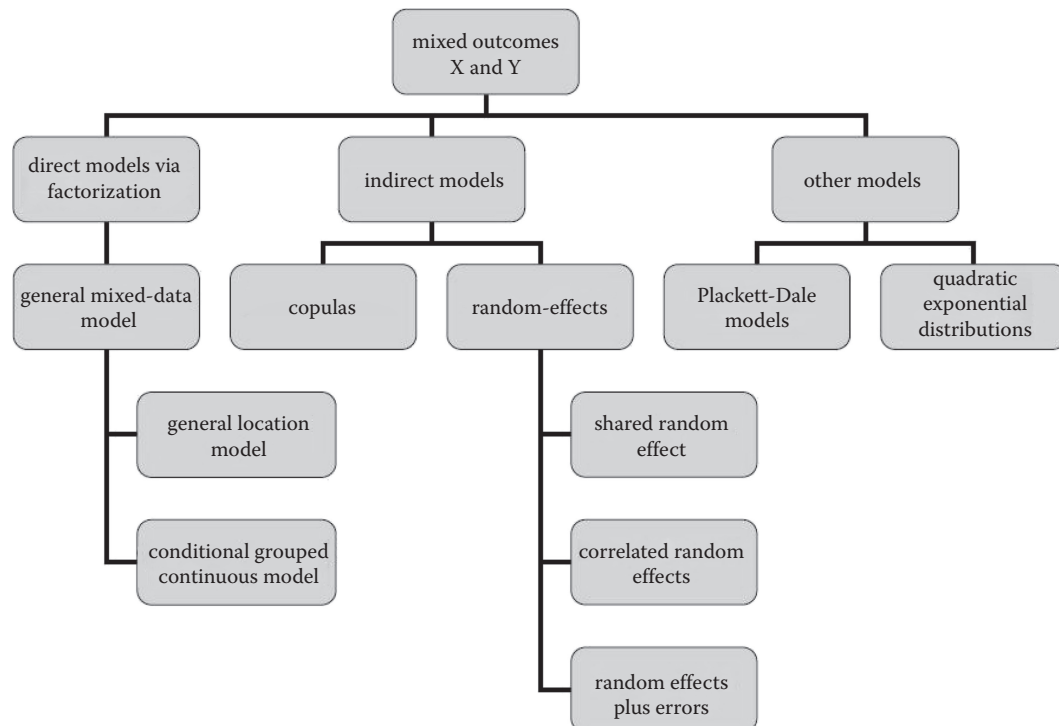


Fig. 1 Taxonomy of models for mixed outcomes X (discrete) and Y (continuous)

This suggested two formulations of mixed-outcome joint distributions: (i) a marginal distribution for discrete outcomes and a conditional distribution for continuous outcomes, given discrete outcomes and (ii) a marginal distribution for continuous outcomes and a conditional distribution for discrete outcomes, given continuous outcomes.

Formulation (i) has received much attention in mixed-data literature, the most popular such approach being Olkin and Tate's^[8] so-called general location model (GLOM). The model is a special case of conditional Gaussian distributions used in graphical association models of mixed data, and assumes different multivariate normal distributions for continuous outcomes given discrete outcomes, and a marginal multinomial distribution for the latter.

Example 4

(Logistic-normal model) Consider a binary outcome X_i and a continuous outcome Y_i . Given outcome-specific covariates $\mathbf{z}_{x_i}$ and $\mathbf{z}_{y_i}$, a marginal regression model for X_i and a conditional regression model for Y_i given X_i are specified as

$$\log\left(\frac{\mu_{x_i}}{1-\mu_{x_i}}\right) = \mathbf{z}_{x_i}^T \boldsymbol{\beta}_x \text{ and } \mu_{y_i|x_i} = \mathbf{z}_{y_i}^T \boldsymbol{\beta}_y + \gamma(X_i - \mu_{x_i}),$$

where $\mu_{x_i} = P(X_i = 1)$ and $\mu_{y_i|x_i} = E(Y_i | X_i)$. With $Y_i | X_i \sim \text{normal}(\mu_{y_i|x_i}, \sigma^2)$ and $X_i \sim \text{bernoulli}(\mu_{x_i})$, the joint density of X_i and Y_i is then

$$f_{X_i, Y_i}(x, y) = f_{X_i}(x) f_{Y_i|x_i}(y | x) = \frac{1}{\sigma} \phi\left(\frac{y - \mu_{y_i|x_i}}{\sigma}\right) \exp\left\{\eta_{x_i} x - \log(1 + e^{\eta_{x_i}})\right\},$$

where $\eta_{x_i} = \log\{\mu_{x_i}/(1-\mu_{x_i})\}$ and $\phi(\cdot)$ is the standard normal density. The association between X_i and Y_i is induced by γ as follows:

$$\text{corr}(X_i, Y_i) = \frac{\gamma}{\sqrt{\gamma^2 + \frac{\sigma^2}{\mu_{x_i}(1-\mu_{x_i})}}}.$$

Note that X_i and Y_i are uncorrelated whenever $\gamma = 0$.

The logistic-normal model extends GLOM to incorporate marginal regression models for binary and continuous outcomes. A number of refinements and extensions of the logistic-normal model have since been studied by several authors. Fitzmaurice and Laird^[9] adapt this model to mixed data with missing responses.

An example of formulation (ii) is the so-called conditional grouped continuous model (CGCM), which assumes a logistic or probit conditional distribution for binary given continuous outcomes and a marginal multivariate normal distribution for the latter. In such models, a continuous latent

vector $\mathbf{Y}_i^*$ underlying $\mathbf{X}_i$ is supposed to exist and a joint distribution $f_{Y_i^*, Y_i}(y^*, y)$ is assumed for $\mathbf{Y}_i^*$ and Y_i . A threshold model is then postulated for $\mathbf{X}_i$ and its corresponding latent vector $\mathbf{Y}_i^*$, so that $f_{X_i, Y_i}(x, y) = f_{Y_i}(y) f_{X_i|Y_i}(x | y)$ is specified through $f_{Y_i^*, Y_i}(y^*, y) = f_{Y_i}(y) f_{Y_i^*|Y_i}(y^* | y)$. Although $f_{Y_i^*, Y_i}(y^*, y)$ is completely arbitrary, modeling it as multivariate normal is convenient because of the latter's convenient marginal and conditional distributional properties.

Example 5

(Probit-normal model) Consider again binary and continuous outcomes X_i and Y_i in Example 4. To model the joint distribution of X_i and Y_i , let $Y_i \sim \text{normal}(\mu_{y_i}, \sigma^2)$ and assume, underlying X_i , a continuous latent variable Y_i^* such that $X_i = I\{Y_i^* > 0\}$, where Y_i and Y_i^* are jointly normal with correlation ρ . With $\mu_{y_i} = \mathbf{z}_{y_i}^T \boldsymbol{\beta}_y$ and $\mu_{y_i^*} = \mathbf{z}_{x_i}^T \boldsymbol{\beta}_{x|y}$, the joint density of X_i and Y_i is

$$f_{X_i, Y_i}(x, y) = \left\{ \Phi\left(\frac{\mathbf{z}_{x_i}^T \boldsymbol{\beta}_{x|y}}{\sqrt{1-\rho^2}} + \gamma(y - \mu_{y_i})\right) \right\}^x \times \left\{ 1 - \Phi\left(\frac{\mathbf{z}_{x_i}^T \boldsymbol{\beta}_{x|y}}{\sqrt{1-\rho^2}} + \gamma(y - \mu_{y_i})\right) \right\}^{1-x} \times \frac{1}{\sigma} \phi\left(\frac{y - \mu_{y_i}}{\sigma}\right),$$

where $\gamma = \rho/(\sigma\sqrt{1-\rho^2})$ and $\Phi(\cdot)$ is the standard normal distribution function. We also get

$$\text{corr}(X_i, Y_i) = \frac{\text{cov}\left(\Phi\left(\frac{\mathbf{z}_{x_i}^T \boldsymbol{\beta}_{x|y}}{\sqrt{1-\rho^2}} + \gamma(Y_i - \mu_{y_i})\right), Y_i\right)}{\sigma \sqrt{\Phi\left(\frac{\mu_{y_i^*} + \gamma\mu_{y_i}}{\sqrt{1+\gamma^2\sigma^2}}\right) \left[1 - \Phi\left(\frac{\mu_{y_i^*} + \gamma\mu_{y_i}}{\sqrt{1+\gamma^2\sigma^2}}\right)\right]}}.$$

For identifiability, it is necessary that Y_i^* has unit variance. In addition, $\boldsymbol{\beta}_{x|y}$ must contain an intercept because the cutpoint was arbitrarily taken as 0. Note that X_i and Y_i are uncorrelated whenever $\gamma = 0$.

Adaptations of the probit-normal model to clustered data settings are discussed in Refs. [6,10].

De Leon and Carrière^[11] describe the general mixed-data model (GMDM) as a hybrid of GLOM and CGCM and provide a unified treatment of these two standard mixed-data models. An attractive feature of GMDM is that it incorporates associations and correlations between different outcomes and accounts for their measurement levels. The model can serve as a platform for extending conventional multivariate methods to the case of mixed data with discrete and continuous outcomes.

Asymmetry in Treatment of Outcomes

It is common to rely on the “arrow of time” to order outcomes and decide the direction of conditioning. However, many mixed-outcome models use a structural approach to classify them into continuous or discrete. Factorization models induce a hierarchy in outcomes, with the conditioning outcome treated as intermediate variable, and the conditioned outcome as the ultimate response.

Example 6

(Probit-normal model) Consider Example 5 again. It is easy to see that

$$\mu_{x_i|y_i} = P(X_i = 1 | Y_i = y) = \Phi \left(\frac{\mathbf{z}_i^T \boldsymbol{\beta}_{x|y}}{\sqrt{1 - \rho^2}} + \gamma(y - \mu_{y_i}) \right),$$

so that the marginal model becomes

$$\mu_{x_i} = P(X_i = 1) = \Phi \left(\frac{\mu_{y_i}^* + \gamma \mu_{y_i}}{\sqrt{1 + \gamma^2 \sigma^2}} \right) = \Phi \left(\mathbf{z}_i^T \boldsymbol{\beta}_x + \mathbf{z}_i^T \bar{\boldsymbol{\beta}}_y \right),$$

where $\boldsymbol{\beta}_x = \boldsymbol{\beta}_{x|y} / \sqrt{1 + \gamma^2 \sigma^2}$. Note that the marginalized probit model for X_i includes the regression coefficient $\bar{\boldsymbol{\beta}}_y = \gamma \boldsymbol{\beta}_y / \sqrt{1 + \gamma^2 \sigma^2} = \rho \boldsymbol{\beta}_y / \sigma$.

Observe that the probit-normal model reverses the factorization in the logistic-normal model. Conditioning on continuous outcomes in this case suggests a predictive model where discrete outcomes are the responses of true interest with continuous outcomes serving as “explanatory” variables. This may not be appropriate in many practical applications, where a more symmetrical treatment of outcomes is needed.

Direction of Conditioning

Factorization models are not invariant to the direction of conditioning taken and the factorization adopted. Although ideally dictated by subject-matter considerations, the choice of the direction of conditioning is mainly for statistical convenience. In toxicological studies, where such models were first developed, the biological mechanism is not well understood, variations of logistic-normal and probit-normal models have thus been studied. The resulting models are not comparable, as parameters have different interpretations depending on the factorization used.

Example 7

(Logistic-normal vs. probit-normal models) From the logistic-normal model, $\mu_{y_i} = E(Y_i) = \mathbf{z}_i^T \boldsymbol{\beta}_y$, so that regression parameters $\boldsymbol{\beta}_x$ and $\boldsymbol{\beta}_y$ have marginal interpretations. This is not true of the probit-normal model. For example, the interpretation

of $\boldsymbol{\beta}_{x|y}$ is conditional on continuous outcome Y_i ; the marginal covariate effect $\boldsymbol{\beta}_x$ is obtained by averaging over Y_i . The logistic-normal model also suggests a normal-mixture marginal distribution for Y_i with

$$\text{var}(Y_i) = \gamma^2 \mu_{x_i} (1 - \mu_{x_i}) + \sigma^2,$$

implying that the variance of Y_i depends on covariate $\mathbf{z}_i$. This contrasts with the homogeneity of Y_i in the probit-normal model.

Another drawback of factorization models is that they do not easily extend to the setting of multiple mixed outcomes. No easy expressions can be obtained for associations between outcomes as well. It is also possible for models to yield very different estimates of the associations.^[10]

RANDOM-EFFECTS MODELS

Indirect approaches to specifying mixed-outcome joint distributions have also been studied. One approach that has found widespread adoption in practice introduces shared or correlated random effects to incorporate correlations between mixed outcomes in the resulting joint model. Daniels and Normand^[5] use this strategy, albeit in a Bayesian framework, to profile health care units. Applications to high-dimensional mixed data analysis are provided by Faes et al.^[4]

The basic idea in this approach is to use random effects, either shared or correlated, to build in correlation between mixed outcomes. The approach does not resort to factorization, and thus yields a symmetrical treatment of mixed outcomes. Its hierarchical structure allows for considerable flexibility in accounting for different measurement levels, delineation of various associations, incorporation of covariate effects, and extension to longitudinal and clustered data settings. Details can be found in Refs. [2,12].

Shared Random Effects

Although random effects provide a general and versatile route for incorporating in the model various correlations between outcomes and differences in measurement levels, they are not without their shortcomings. McCulloch^[12] illustrates a problem for the case of one binary outcome and one continuous outcome, with respective conditional Bernoulli and normal distributions, given a shared normal random effect, which induces correlation between outcomes.

Example 8

(Probit-normal model with shared random effect) Consider a binary outcome $X_i \sim \text{bernoulli}(\mu_{x_i})$ and a continuous outcome $Y_i \sim \text{normally}(\mu_{y_i}, \sigma^2)$. A random effect shared by X_i and Y_i is introduced as follows:

$$\Phi^{-1}(\mu_{x_i|y_i}) = \mathbf{z}_{x_i}^T \boldsymbol{\beta}_{x|u} + U_i \text{ and } \mu_{y_i|u} = \mathbf{z}_{y_i}^T \boldsymbol{\beta}_{y|u} + \lambda U_i,$$

where $U_i \sim \text{normal}(0, \tau^2)$, with X_i and Y_i conditionally independent given U_i . The coefficient λ adjusts for difference in scales between the outcomes. The joint density is

$$f_{X_i, Y_i}(x, y) = \frac{1}{\sigma \tau} \int_{-\infty}^{\infty} \mu_{x_i|u}^x (1 - \mu_{x_i|u})^{1-x} \phi\left(\frac{y - \mu_{y_i|u}}{\sigma}\right) \phi(u) du.$$

We also get

$$\text{corr}(X_i, Y_i) = \sqrt{\frac{\lambda^2 \tau^2}{\sigma^2 + \lambda^2 \tau^2}} \frac{\text{cov}\left(\Phi\left(\mathbf{z}_{x_i}^T \boldsymbol{\beta}_{x|u} + U_i\right), \lambda U_i\right)}{\sqrt{\Phi\left(\mathbf{z}_{x_i}^T \boldsymbol{\beta}_x\right)\left\{1 - \Phi\left(\mathbf{z}_{x_i}^T \boldsymbol{\beta}_x\right)\right\}}},$$

where $\boldsymbol{\beta}_x = \boldsymbol{\beta}_{x|u} / \sqrt{1 + \tau^2}$.

With a conditional probit link for binary outcome X_i given random effect U_i , the marginal regression parameter $\boldsymbol{\beta}_x$ is smaller than the conditional regression parameter $\boldsymbol{\beta}_{x|u}$ by a factor $\sqrt{1 + \tau^2}$. In practical applications where interest lies in marginal effects of covariates, this lack of a marginal interpretation may be considered unattractive. In addition, the model constrains the correlation between X_i and Y_i . McCulloch^[12] shows that $\text{corr}(X_i, Y_i) \rightarrow \sqrt{2/\pi} \approx 0.798$ as $\tau^2 \rightarrow \infty$ with σ^2 fixed.

The situation is somewhat better in the Poisson-normal case in that marginal and conditional regression coefficients, with the exception of intercepts, are identical; however, the correlation becomes even more problematic, being integrally tied to overdispersion in the Poisson outcome. Details are discussed in Ref. [12].

Correlated Random Effects

A possible remedy to the shortcomings of the shared random effect approach is to allow for separate but correlated random effects. This results in models that more flexibly describe correlations between mixed outcomes. The following example involving count and continuous outcomes with correlated random effects is discussed in detail by McCulloch.^[12]

Example 9

(Poisson-normal model with correlated random effects) Consider a count outcome $X_i \sim \text{poisson}(\mu_{x_i})$ and a continuous outcome $Y_{it} \sim \text{normal}(\mu_{y_i}, \sigma^2)$. It is assumed that the continuous outcome is repeatedly observed over time $t = 1, \dots, n_i$, to identify subject-to-subject variation. Correlated random effects for X_i and Y_{it} are introduced as follows:

$$\log(\mu_{x_i|u_t}) = \mathbf{z}_{x_i}^T \boldsymbol{\beta}_{x|u} + U_i \text{ and } \mu_{y_i|u_t} = \mathbf{z}_{y_i}^T \boldsymbol{\beta}_{y|u} + \lambda U_i,$$

where U_{x_i} and U_{y_i} are jointly normally distributed with $E(U_{x_i}) = E(U_{y_i}) = 0$, $\text{var}(U_{x_i}) = \tau_x^2$, $\text{var}(U_{y_i}) = \tau_y^2$, and $\text{corr}(U_{x_i}, U_{y_i}) = \rho$. It is straightforward to show that

$$\text{corr}(X_i, Y_{it}) = \frac{\rho T_x T_y \exp\left(\mathbf{z}_{x_i}^T \boldsymbol{\beta}_{x|u} + \frac{1}{2} |\rho| \tau_x^2\right)}{(\sigma^2 + \tau_y^2) \left\{ \mu_{x_i} + \mu_{x_i}^2 (e^{\tau_x^2} - 1) \right\}},$$

with $\mu_{x_i} = \exp(\mathbf{z}_{x_i}^T \boldsymbol{\beta}_{x|u} + \tau_x^2/2)$.

In the Poisson-normal model in Example 9, correlated normal random effects allow for negative correlations between outcomes. It is also easier to disentangle overdispersion in the Poisson outcome, as this now depends on the covariance matrix of the random effects.^[12] However, allowing for correlated random effects in high-dimensional problems may not be computationally feasible^[4] in practice.

Random Effects Plus Correlated Errors

Still another alternative is to add correlated error terms to models for the outcomes. To see how this can be done, assume $\mathbf{X}_i$ and $\mathbf{Y}_i$ to be conditionally independent, given an unobserved random effect U_i . Generalized linear mixed models for $\mathbf{X}_i$ and $\mathbf{Y}_i$ are then specified, either via an exponential family or a marginal moments specification. Random effect U_i is subject-specific and accounts for correlation patterns of unobserved heterogeneity; to account for residual correlations, marginal models are assumed for the outcomes, in addition to random effect U_i . Specifically, we have

$$\begin{pmatrix} \mathbf{X}_i \\ \mathbf{Y}_i \end{pmatrix} = \begin{pmatrix} \mathbf{h}_{x_i}(\mathbf{U}_{x_i}, \mathbf{z}_{x_i}) \\ \mathbf{h}_{y_i}(\mathbf{U}_{y_i}, \mathbf{z}_{y_i}) \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_{x_i} \\ \boldsymbol{\epsilon}_{y_i} \end{pmatrix} = \mathbf{h}(\mathbf{U}_i, \mathbf{z}_i) + \boldsymbol{\epsilon}_i,$$

where the inverse link functions in $\mathbf{h}(\cdot, \cdot)$ depend on outcome types, and $\mathbf{z}_i$ is a vector of covariates. For discrete outcomes such as binary outcomes and counts, data approximation based on linearization may be adopted. Note that $\mathbf{h}(\cdot, \cdot)$ allows for a non-additive specification of fixed and random effects in the link function; it is also possible to incorporate latent variables for discrete outcomes. Correlations between $\mathbf{X}_i$ and $\mathbf{Y}_i$ are induced by covariance matrices of $\mathbf{U}_i$ and $\boldsymbol{\epsilon}_i$.

Example 10

(Repeated measures logistic-normal model with random effects and errors) Let X_{it} and Y_{it} be binary and continuous outcomes for subject $i = 1, \dots, n$, at time $t = 1, \dots, n_i$. The model is given by

$$\begin{pmatrix} X_{it} \\ Y_{it} \end{pmatrix} = \begin{pmatrix} \frac{\exp(\mathbf{z}_{x_{it}}^T \boldsymbol{\beta}_{x_{it}} + U_{x_{it}})}{1 + \exp(\mathbf{z}_{x_{it}}^T \boldsymbol{\beta}_{x_{it}} + U_{x_{it}})} \\ \mathbf{z}_{y_{it}}^T \boldsymbol{\beta}_{y_{it}} + U_{y_{it}} \end{pmatrix} + \begin{pmatrix} \varepsilon_{x_{it}} \\ \varepsilon_{y_{it}} \end{pmatrix},$$

where $U_{x_{it}}$ and $U_{y_{it}}$ are as defined in Example 9, with independent errors $\varepsilon_{x_{it}}$ and $\varepsilon_{y_{it}}$. Faes et al.^[4] derive a first-order approximation for the correlation between X_{it} and Y_{it} as

$$\text{corr}(X_{it}, Y_{it}) \approx \frac{\rho \tau_x T_y v_{it}}{\sqrt{v_{it} (1 + v_{it} \tau_x^2) (\tau_y^2 + \sigma^2)}},$$

$$\text{var}(\varepsilon_{x_{it}}) \approx v_{it} = \frac{\exp(\mathbf{z}_{x_{it}}^T \boldsymbol{\beta}_{x_{it}}) \{1 - \exp(\mathbf{z}_{x_{it}}^T \boldsymbol{\beta}_{x_{it}})\}}{\{1 + \exp(\mathbf{z}_{x_{it}}^T \boldsymbol{\beta}_{x_{it}})\}^2}$$

and $\text{var}(\varepsilon_{y_{it}}) = \sigma^2$

The inclusion of random effects in the above model accounts for correlations between repeated measurements. In the case of independent random effects, the outcomes are approximately marginally uncorrelated. Incarnations of the above model, with varying levels of generality, are given by Faes et al.,^[4] Gueorguieva and Sanacora,^[1] and Gueorguieva and Agresti.^[13] Bayesian versions of the model have been studied by Daniels and Normand^[5] and Dunson,^[14] among others. This approach is quite attractive; however, for Poisson outcomes, for example, it leads to the same problem of overdispersion being built into the model, as what happens in the shared random effect approach.^[12]

OTHER MODELS

A number of other alternative models have been proposed and studied by various authors. A Plackett-Dale distribution has been used by Molenberghs et al.^[15] as an alternative to multivariate normal distributions for latent variables. The quadratic exponential model, which has a general form that can accommodate mixed outcomes, has been adopted by Sammel et al.^[16] in clustered settings. The model considers conditional log-odds ratios as measures of associations and requires a normalizing constant, the computational demands of which can be prohibitive in certain cases. Regan and Catalano^[17] introduced mixed-probit models for binary and continuous outcomes that, unlike factorization-based models, treat outcomes in a more symmetrical fashion and have marginal interpretations for their regression parameters. However, these models were developed for specific applications, and thus, may lack generality and flexibility to be applied in other mixed-outcome data settings.

Finally, several authors have adopted copula functions to indirectly specify mixed-outcome joint distributions. This is only a recent phenomenon in modeling mixed outcomes, and unresolved issues, both methodological and practical, abound, especially as they apply to discrete data.

Example 11

(Robit-normal model) To model the joint distribution of binary outcome X_i and continuous outcome Y_i , we let $Y_i \sim N(\mu_{y_i}, \sigma^2)$ and assume, underlying X_i , a continuous latent variable $Y_i^* \sim t_1(\mu_{y_i^*}, 1, \nu)$, i.e., the univariate (generalized) t -distribution with mean $\mu_{y_i^*}$, unit scale, and degrees of freedom ν , such that $X_i = I\{Y_i^* > 0\}$.

With $\mu_{y_i} = \mathbf{z}_{y_i}^T \boldsymbol{\beta}_{y_i}$ and $\mu_{y_i^*} = \mathbf{z}_{x_i}^T \boldsymbol{\beta}_{x_i}$, and using the Gaussian copula, de Leon and Wu^[18] derive the joint density of X_i and Y_i as

$$f_{X_i, Y_i}(x, y) = \left\{ \Phi \left(- \frac{\Phi^{-1}(w_i) - \rho \left(\frac{y - \mu_{y_i}}{\sigma} \right)}{\sqrt{1 - \rho^2}} \right) \right\}^x$$

$$\times \left\{ 1 - \Phi \left(- \frac{\Phi^{-1}(w_i) - \rho \left(\frac{y - \mu_{y_i}}{\sigma} \right)}{\sqrt{1 - \rho^2}} \right) \right\}^{1-x} \frac{1}{\sigma} \phi \left(\frac{y - \mu_{y_i}}{\sigma} \right),$$

where $w_i = P(X_i = 1) = P(Y_i^* > 0)$ and ρ is the Pearson correlation between so-called normal scores $\Phi^{-1}\{P(Y_i^* \leq 0)\}$ and $\Phi^{-1}\{P(Y_i^* \leq 0)\}$, a “proxy” for the tetrachoric correlation between X_i and Y_i .

Liu^[19] introduces the term robit regression based on a t -latent distribution as a robust alternative to and extension of logit and probit regression models. Liu^[19] shows that robit models approximate both logit (with $\nu \approx 7$) and probit (with large ν) regressions, and thus provide a general approach to binary regression modeling.

Copula functions, which are common in actuarial and financial applications, have proved useful in practice when the joint distribution of interest is either not available or difficult to specify but marginal distributions can be specified with confidence, like in mixed-outcome data settings. The approach entails specifying marginal distributions $f_{X_i}(x)$ and $f_{Y_i}(y)$, and combining them to form a joint distribution via a suitable copula function. Note that, similar to random effects approaches, copula-based models treat mixed outcomes symmetrically; however, unlike factorization and random effects models, their regression parameters are marginally meaningful. Recent applications of copulas to mixed outcomes are discussed in Song^[20,21] and Hoff.^[22]

MODEL ESTIMATION

A variety of approaches have been considered in the literature for model estimation. Given a full specification of the model, a likelihood-based approach may be considered. However, evaluation and direct maximization of the full likelihood may be computationally prohibitive in practice. In addition, a full likelihood approach raises questions about the robustness of the likelihood specification quite apart from any computational difficulties. To circumvent these, the full likelihood may be replaced by a more computationally tractable function, a so-called pseudo-likelihood function. One such function is obtained by compositing pairwise likelihoods over the data. Faes et al.^[23] adopts this strategy in a high-dimensional mixed-outcome analysis via mixed models implemented using standard software, such as SAS procedures NLMIXED and glimmix. Working with such a modification of the likelihood function generally yields consistent estimates of model parameters, including correlations under a range of possible models for higher-order dependency as captured by the full joint distribution. Computational and statistical performance of these methods has been shown to range from acceptably good to excellent.

Several alternative estimation strategies based on the EM and Monte Carlo methods have also been proposed in the literature.^[16] However, as with conventional full likelihood estimation, these methods have the disadvantage of excessive computational requirements. Non-likelihood based approaches such as generalized estimating equations (GEE)^[3] may also be adapted to this context. These may prove more useful and computationally more feasible than likelihood-based methods in practice.

CONCLUSION

This contribution provides an up-to-date survey of joint analysis of mixed outcomes. Both direct and indirect approaches to joint model specification are discussed along with their advantages and disadvantages. Because correlations between the mixed outcomes are usually of practical interest, particular attention is given to differences in how such correlations are incorporated in the models. Technical and practical issues relating to model estimation are also briefly discussed, with focus on pseudo-likelihood methods.

The use of copulas in modeling mixed outcomes is a recent phenomenon. Because of this, unresolved issues, both methodological and practical, abound. One is the paucity, if not total lack, of copulas that can be used for mixed outcomes; of theoretical, but of lesser practical, concern is the non-uniqueness of copulas for discrete outcomes. Another involves the interpretation of the dependence parameter of copula functions, a serious problem for discrete distributions. De Leon and Wu^[18]

have developed copula-based regression models for mixed outcomes by adopting a latent-variable formulation of the discrete outcomes and using Gaussian copulas to “glue” mixed-outcome marginal regression models. Work on generalizations of their approach to high-dimensional settings with possibility of incorporating random effects would be worthwhile. The challenge here lies in defining models that allow for different levels of association among outcomes, as in longitudinal studies.

Incomplete data are ubiquitous in studies involving mixed outcomes.^[9] This is an area where little research has been done. Adaptations of the models discussed in this contribution to handle missing data would be of great importance in practice.

FURTHER READING

The last decade has seen many remarkable advances in statistical methodology for analyzing mixed data and we give here only a few selected references. Regan and Catalano^[17] provide an excellent early survey of mixed-outcome analysis, with particular focus on toxicological applications. Molenberghs and Verbeke^[2] provide a good introduction to generalized linear mixed models for mixed outcomes and include a number of practical examples. Faes et al.^[23] update Molenberghs and Verbeke^[2] and include recent advances in analysis of longitudinal mixed outcomes in high-dimensions. Bayesian approaches including methods for handling incomplete mixed-outcome data, are described in Daniels and Hogan.^[24] McCulloch et al.^[25] devote a chapter on random-effects models for mixed outcomes. The use of copulas for modeling mixed data is discussed in Song.^[20]

ACKNOWLEDGMENT

This work was supported by grants from the Alberta Heritage Foundation for Medical Research and the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

1. Gueorguieva, R.V.; Sanacora, G. Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Stat. Med.* **2006**, *25*, 1307–1322.
2. Molenberghs, G.; Verbeke, G. *Models for Discrete Longitudinal Data*; Springer: New York, 2005.
3. Teixeira-Pinto, A.; Normand, S.-L.T. Correlated bivariate continuous and binary outcomes: Issues and applications. *Stat. Med.* **2009**, *28*, 1753–1773.
4. Faes, C.; Aerts, M.; Molenberghs, G.; Geys, H.; Teuns, G.; Bijmens, L. A high-dimensional joint model for longitudinal outcomes of different nature. *Stat. Med.* **2008**, *27*, 4408–4427.

5. Daniels, M.J.; Normand, S.-L.T. Longitudinal profiling of health care units based on continuous and discrete patient outcomes. *Biostatistics* **2006**, *7*, 1–15.
6. Catalano, P.J.; Ryan, L.M. Bivariate latent variable models for clustered discrete and continuous outcomes. *J. Am. Stat. Assoc.* **1992**, *87*, 651–658.
7. Fitzmaurice, G.M.; Laird, N.M. Regression models for a bivariate discrete and continuous outcome with clustering. *J. Am. Stat. Assoc.* **1995**, *90*, 845–852.
8. Olkin, I.; Tate, R.F. Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Stat.* **1961**, *32*, 448–465 (correction in 36, 343–344).
9. Fitzmaurice, G.M.; Laird, N.M. Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics* **1997**, *53*, 110–122.
10. Regan, M.M.; Catalano, P.J. Combined continuous and discrete outcomes. In *Topics in Modelling of Clustered Data*; Aerts, M.; Geys, H.; Molenberghs, G.; Ryan, L.; Eds.; Chapman & Hall/CRC Press: Boca Raton, FL, 2002.
11. de Leon, A.R.; Carrière, K.C. General mixed-data model: extension of general location and grouped continuous models. *Canadian J. Stat.*, **2007**, *35*, 533–548.
12. McCulloch, C. Joint modeling of mixed outcome types using latent variables. *Stat. Methods Med. Res.* **2007**, *17*, 1–21.
13. Gueorguieva, R.V.; Agresti, A. A correlated probit model for joint modeling of clustered binary and continuous response. *J. Am. Stat. Assoc.* **2001**, *96*, 1102–1112.
14. Dunson, D.B. Bayesian latent variable models for clustered mixed outcomes. *J. R. Stat. Soc. B* **2000**, *62*, 355–366.
15. Molenberghs, G.; Geys, H.; Buyse, M. Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. *Stat. Med.* **2001**, *20*, 3023–3038.
16. Sammel, M.D.; Ryan, L.M.; Legler, J.M. Latent variable models for mixed discrete and continuous outcomes. *J. R. Stat. Soc. B* **1997**, *59*, 667–678.
17. Regan, M.M.; Catalano, P.J. Likelihood models for clustered binary and continuous outcomes: Application to developmental toxicology. *Biometrics* **1999**, *55*, 760–768.
18. de Leon, A.R.; Wu, B. Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine* **2011**, *30* (2), 175–85.
19. Liu, C. Robit regression: a simple robust alternative to logistic and probit regression. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*; Gelman, A.; Meng, X.-L.; Eds.; Wiley: New York, 2004.
20. Song, P.X.-K. *Correlated Data Analysis: Modeling, Analytics, and Applications*; Springer: New York, 2007.
21. Song, P.X.-K.; Li, M.; Yuan, Y. Joint regression analysis of correlated data using Gaussian copulas. *Biometrics* **2009**, *65*, 60–68.
22. Hoff, P.D. Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.* **2007**, *1*, 265–283.
23. Faes, C.; Geys, H.; Catalano, P.J. Joint models for continuous and discrete longitudinal data. In *Longitudinal Data Analysis*; Fitzmaurice, G.; Davidian, Verbeke G.; Molenberghs, G.; Eds.; Chapman & Hall/CRC: Boca Raton, FL, 2009.
24. Daniels, M.J.; Hogan, J.W. *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*; Chapman & Hall/CRC: Boca Raton, FL, 2008.
25. McCulloch, C.E.; Searle, S.R.; Neuhaus, J.M. *Generalized, Linear, and Mixed Models*, 2nd Ed.; Wiley: New York, 2008.

MMRM with Missing Data

Annpey Pong, Jun Zhao, Allen Lee, Phillip Phiri, and Lev Sverdlov

Merck Research Laboratories, Summit, New Jersey, U.S.A.

Abstract

In clinical research of drug development, missing values in longitudinal data are common due to early drop out. Therefore the method of handling missing values is critical in data analysis of drug development. This entry breaks down the methodology of the Mixed-effects Model Repeated Measurement and provides an example illustrating its role in the problem of missing data.

INTRODUCTION

In clinical research of drug development, missing values in longitudinal data are common due to early drop out. Missing values usually lead to a potential source of bias in data interpretation.^[1] Therefore the method of handling missing values is critical in data analysis of drug development. Particularly it's important to submission of confirmatory trials for New Drug Applications (NDA).^[2,3]

According to the ICH E9, Section 5.3, *Statistical Principles for Clinical Trials*,^[4] the methods of dealing with missing values are to be pre-defined in the protocol, or may be updated in the statistical analysis plan (SAP) during the blind review. However, there is no universally applicable method of handling missing values recommended. The sensitivity of the results to the method of handling missing values should be concerned and investigated. In current practice for NDA submissions, the method of handling missing value to the primary analysis for key efficacy variable(s) is the focus and to be clearly described in the SAP before locking the database.

In many therapeutic areas such as CNS, the general period of observation is divided into the short term acute phase (lasted from days to weeks) and the long term extension phase (lasted from weeks to years). The drop out rate in the long term phase is usually much higher than completion rate in the short term study. Effect of missing data in the long term extension studies is an additional concern and more problematic for the SAP development and statistical analysis.^[5]

Little and Rubin^[1] classified missing values into 3 categories: (i) missing completely at random (MCAR) assumes missing values are randomly distributed across all observations. In other words, missingness does not depend on either observed or unobserved outcomes; (ii) missing at random (MAR) requires missing values are randomly distributed on the observed data only. As a result, the missingness depends on the observed outcomes but not the unobserved; (iii) missing not at random

(MNAR) presumes for non-ignorable missing values. Consequently, the missingness depends on the unobserved outcomes. In reality, we cannot tell from the data at hand whether the missing observations are MCAR, MAR, or MNAR except by obtaining follow-up data from non-respondents. Nevertheless this is a non-practical approach in clinical trials.

In the past decade, the last observation carried forward (LOCF) has been a long-standing and traditional method of handling missing values for longitudinal data especially in psychiatric clinical trials. It's not uncommon that a simple endpoint analysis such as analysis of covariance (ANCOVA) is performed in longitudinal studies. For a commonly used primary efficacy variable, mean change from baseline to endpoint, the LOCF approach assumes the change to endpoint from baseline is the same as the change to the last observed value. In other words, the missing value at endpoint is imputed by the last observed value in the trial for early drop out subjects. This is based on the assumption of MCAR. Researches have shown that LOCF was not robust to handle the potential bias due to missing values and the results were often invalid.^[6–8] In addition, the LOCF does not take response profiles over time into consideration. Except the LOCF, other missing data imputation alternatives are available. The methods include the baseline observation carried forward (BOCF), worst observation carry forward (WOCF), partial LOCF (such as, BOCF for dropout due to AE, and LOCF for other dropouts), and a single imputation by using the mean value. The primary drawbacks of these methods are the underestimation of the standard error and a strong assumption of MCAR. To avoid these problems, the multiple imputation (MI) methods^[9] and maximum likelihood methodology (MLE)^[10] which require the assumption of MAR are usually considered. The primary difference between MI and MLE is MI requires additional data manipulation and modeling choice to impute missing data, but not for the MLE based on the maximum likelihood.

The Mixed-effects Model Repeated Measurement (MMRM) is a likelihood-based method which provides better control of type I and type II errors than ANCOVA based on LOCF for missing data imputation (ANCOVA-LOCF).^[11–13] The MMRM uses all the observed data from repeated measurements assuming to be multivariate normal to adjust the data which was missing on an individual subject. In drug development, the MMRM analysis was initially acceptable for primary analysis to the efficacy measurement, 17-item Hamilton Rating Scale, on confirmatory phase III trials, Duloxetine (as Cymbalta®).^[14,15] This approach took the place of the commonly used ANCOVA-LOCF for primary analysis in antidepressant products.

The mixed models for repeated data were firstly introduced into psychiatric literature by Gibbons et al.^[16] The model assumes a continuous outcome variable which is linearly related to a set of explanatory variables. The fundamental difference of mixed model to the traditional simple endpoint analysis is the repeated observations within subjects are correlated, and this correlation is incorporated into the analysis model. Usually, an unstructured variance-covariance structure is considered, or a goodness-of-fit exploration to be used to disclose the best fit from different structures. In addition, the analysis model contains both fixed and random effects. The commonly used main fixed effects are treatment group, study center, and visit for multi-center clinical trials. The random effect is usually referred to the subject's effect in the study.

The MMRM estimates for missing data are derived using EM algorithm. As a result, the EM algorithm may not converge to the maximum likelihood estimate when data are sparse, or if the rate of missing data is high, or analysis model is over-parameterized.^[17,18] The analysis for MMRM can be easily assessed using, for example, the SAS procedure MIXED or the SPlus or R function, without additional data manipulation. It is to be noted that the estimates from MMRM are valid if MAR assumption is met. The assessment of missing mechanism for MAR may be explored by odds ratio which is illustrated in a numerical example in "Example."

MMRM METHODOLOGY

General Linear Mixed Model

The general linear mixed-effect model was introduced by Laird and Ware,^[17] which contains both fixed and random effects. Let $\mathbf{y} = (y_1, y_2, \dots, y_n)$ be the n -dimensional response vector from n subjects, and the data to be explained by p fixed-effects variables $x_1, \dots, x_p$, and q random-effects variables $z_1, \dots, z_q$. Then the mixed model can be written as the following matrix format:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where

$\mathbf{X}$ is the $(n \times p)$ -dimensional matrix associating to the observations of the fixed-effects

$\mathbf{Z}$ is the $(n \times q)$ -dimensional matrix associating to the random-effects

$\boldsymbol{\beta}$ is the p -dimensional vector of fixed-effects parameters

$\boldsymbol{\gamma}$ is the q -dimensional vector of random-effects parameters

$\boldsymbol{\varepsilon}$ is the $(n \times 1)$ -dimensional vector of errors for observations.

In addition, $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ are independently normal distributed with

$$E \begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

$$Var \begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}$$

where $\mathbf{G}$ is a $q \times q$ -dimensional covariance matrix for the random effects; $\mathbf{R}$ is the $(n \times n)$ covariance matrix for the errors; $\mathbf{0}$ is vector of zeroes; Note that the correlation among data within a subject and heterogeneous variances at different visits is allocated under longitudinal data structure. The variance of $\mathbf{y}$, $\mathbf{V}$, is therefore to be calculated as below:

$$\mathbf{V} = Var(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$$

It is to be noted that the general linear model is a special case of the mixed model with $\mathbf{Z} = \mathbf{0}$ and $\mathbf{R} = \sigma^2 \mathbf{I}$

Estimating $\mathbf{G}$ and $\mathbf{R}$ in the Mixed Model

To estimate $\boldsymbol{\beta}$, generalized least squares (GLS) is used to minimize

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

where $\mathbf{V}$ is the variance of $\mathbf{y}$. It becomes clear that we need to estimate $\mathbf{G}$ and $\mathbf{R}$ first. Under the assumption that $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ are normally distributed, estimation of the covariance parameters ($\mathbf{G}$ and $\mathbf{R}$) can be done through (restricted) maximum likelihood methods. It has been shown^[19,20] that both methods accommodate data that are missing at random (MAR).

In addition to the assumptions to $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$, we may further assume that $\mathbf{R}$ and/or $\mathbf{G}$ follow some variance-covariance structures. The variance-covariance structures like compound symmetry (CS), AR(1), toeplitz (TOEP), unstructured (UN), etc are commonly used. To test one covariance structure against the other for lack of fit, Information Criterion (IC) such as Akaike's Information Criterion^[21] and Schwarz's Bayesian Information Criterion^[22] are frequently used. The rule is the smaller IC the better fit. One can also use the likelihood ratio to test whether one model is significantly better than the other.

Estimating β and γ in the Mixed Model

Let $\hat{G}$ and $\hat{R}$ be the estimates of G and R , respectively. To obtain estimates of β and γ , the standard method is to solve the *mixed model equations*^[23] as below:

$$\begin{bmatrix} X'\hat{R}^{-1}X & X'\hat{R}^{-1}Z \\ Z'\hat{R}^{-1}X & Z'\hat{R}^{-1}Z + \hat{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X'\hat{R}^{-1}Y \\ Z'\hat{R}^{-1}Y \end{bmatrix}$$

The solutions can be written as the empirical Bayes estimators^[17] shown below:

$$\begin{aligned} \hat{\beta} &= (X'\hat{V}^{-1}X)^{-1} X'\hat{V}^{-1}y \\ \hat{\gamma} &= \hat{G}Z'\hat{V}^{-1}(y - X\hat{\beta}) \end{aligned}$$

Note that the mixed model equations are extended normal equations. $\hat{G}$ is nonsingular. If $\hat{G}$ is singular, then the mixed model equations can be modified as shown in Henderson.^[23]

Statistical Properties and Inferences

With estimates of β and γ , hypothesis testing regarding these unknown parameters can be performed to

$$H_0 : L \begin{bmatrix} \beta \\ \gamma \end{bmatrix} = 0$$

If L consists of a single row, a general t -statistic can be computed as

$$t = \frac{L \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix}}{\sqrt{L\hat{C}L'}}$$

where

$$\hat{C} = \begin{bmatrix} X'\hat{R}^{-1}X & X'\hat{R}^{-1}Z \\ Z'\hat{R}^{-1}X & Z'\hat{R}^{-1}Z + \hat{G}^{-1} \end{bmatrix}^{-1}$$

is the approximate variance-covariance matrix of $(\hat{\beta} - \beta, \hat{\gamma} - \gamma)$; $^{-1}$ denotes the generalized inverse of the coefficient matrix of the mixed model. In general, t is

approximately t -distributed and its degrees of freedom can be estimated.^[24,25]

If $\text{rank}(L) > 1$, a general F -statistic can be derived as

$$F = \frac{\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix}^{-1} L'(L\hat{C}L')^{-1} L \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix}}{\text{rank}(L)}$$

The numerator degree of freedom is the rank (L) and the denominator degree of freedom is taken from the corresponding fixed or random effects.

EXAMPLE

In this section, the efficacy data from a schizophrenia study is presented. This example is to describe the rational of using a mixed model for the longitudinal data, instead of a fixed effect ANCOVA model. This is a double-blind, placebo controlled 6-week study.

The efficacy assessment, total PANSS (Positive and Negative Syndrome Scale) score, were done on baseline and post-treatments on Days 4, 7, 14, 21, 28, 35 and 42. The primary interest is the change from baseline in total PANSS score at 42 days (or endpoint) on treatment. A traditional methodology is to analyze the data by using an ANCOVA model on day 42/Endpoint, with missing data imputed by LOCF.

It was observed that 61.3% of subjects finished the 42 days of treatment from four treatment groups. The reasons for discontinuation are summarized in Table 1. It shows that a relatively high percentage of drop out rate of 38.7% and unbalanced drop out due to lack of therapeutic response and withdrew consent between treatment groups. This leads a direction that the assumption of MCAR may be violated. In other words, the validity of using the traditional method of LOCF is a concern. Since this trial was a well-controlled phase III clinical study, the missing mechanism of MNAR was unlikely either.

To assess the missing mechanism of MCAR, the following approach was applied:

If missing depends on observed outcome of interest, then the assumption of MCAR does not hold.

Table 1 Drop Out Rate by Treatment Groups and Reason for Discontinuation

Reason for discontinuation(%)	Treatment 1	Treatment 2	Treatment 3	Treatment 4	Overall
AE/SAE	10.6	9.4	10.4	4.5	8.8
Underlining Disease	7.3	8.5	5.2	1.8	5.7
Lack of Efficacy	17.9	7.5	3.5	10.8	10.1
Withdrew of Consent	10.6	10.4	22.6	18.9	15.6
Lost to Follow-up	1.6	1.9	3.5	2.7	2.4
Other	2.4	3.8	0.9	0.0	1.8
Total	43.1	33.0	40.9	36.9	38.7

To verify the statement above, a generalized linear model of the probability of drop out at Day 42/Endpoint for subjects who had at least 14 days of treatment, was employed. The rationale of using Day 14 is based on the protocol defined minimal length of in-patient period of Day 14. The model is shown as below:

$$P(\text{dropout at Day 42/Endpoint}) = \text{treatment group} \\ + \text{center} + \text{baseline total PANSS score} \\ + \text{change from baseline to Day 14 of total PANSS score,}$$

where P is the probability of including the factors on the right hand side of the equation. Thereafter a 95% confidence limit of Odds Ratio from the logistic regression model was performed (Table 2).

It was observed that the lower bound of the 95% confidence limit of Odds Ratio was larger than 1 and the covariate change from baseline to Day 14 was significant (p -value is 0.002). It concludes that the subjects with a lower change from baseline to Day 14 have a higher probability to drop out on Day 42 (such as, Odds Ratio = 1.033 per PANSS unit). Consequently the assumption of MCAR does not hold since the missingness on Day 42 depends on the observed outcome on Day 14. A MMRM was performed below based on the missing mechanism of MAR.

$$\text{Change from baseline in total PANSS score} = \text{treatment} \\ + \text{center} + \text{baseline total PANSS score} \\ + \text{visit} + \text{visit} \times \text{treatment,}$$

The unstructured (UN) covariance structure was used to model the variance of the repeated PANSS total score measurements over time within subjects. The fixed

Table 2 The Odds Ratio Estimate for Change from Baseline to Day 14

Point estimate	95% Wald confidence limits	
1.033	1.012	1.055

effects in the model are treatment group, study center, visit, baseline total PANSS score and visit by treatment interaction. The random effect is referred to the subject's effect in the study. The results of comparing of 3 treatments to the reference of treatment 1 were presented in the table below on both LOCF and MMRM. The factors in ANCOVA model with LOCF imputation are treatment group, study center, and a covariate of baseline total PANSS score.

CONCLUSIONS

Missing data are an almost ever-present problem in clinical trials.^[26] A common choice in many therapeutic areas is to assess the mean change from baseline to endpoint via analysis of variance (ANOVA), with missing data imputed by carrying the last observation forward (LOCF).^[11] The LOCF has been used as the method of choice because it is simple and easy to use. LOCF assumes that the condition of the patient whose data is missing remains unchanged from the last observed value to the endpoint of the trial. In many cases, patient's condition either improves or worsens and it seldom remains constant.^[26] Carrying observation forward may therefore bias estimates of treatment effects and associated errors.^[8,16,26,27]

As noted above, Little and Rubin^[1] classified missing values into 3 categories, namely; MCAR, MAR and MNAR based on process leading to missingness and under what process different analysis methods would be appropriate to use. This implies that the method of analysis should take account the reasons leading to the missingness. However, it is difficult to tell from the data which category the missingness falls unless there is a follow-up data from non-respondents. Nevertheless, this is not an easy practice in a well-control clinical trial.

In many clinical research settings, the MAR assumption is more reasonable than the MCAR and the likelihood-based mixed-effects models offer a general

Table 3 Mean Change from Baseline to Study Endpoint using LOCF and MMRM Analysis for Intent-To-Treat Population

Methods	Statistics	Treatment 1	Treatment 2	Treatment 3	Treatment 4
LOCF	Mean Change	-10.7	-14.9	-15.4	-16.2
	SE	1.57	1.69	1.63	1.66
	Difference to Treatment 1		-4.11	-4.70	-5.48
	SE of difference		2.25	2.21	2.23
	P-value		0.068	0.034	0.014
MMRM	Mean Change	-14.6	-19.4	-20.0	-21.3
	SE	1.61	1.68	1.70	1.70
	Difference to Treatment 1		-4.86	-5.47	-6.77
	SE of Difference		2.32	2.33	2.33
	P-value		0.038	0.020	0.004

SE: Standard Error.

framework for the longitudinal analyses.^[26] These methods are more robust to potential bias from missing data than LOCF^[16,26] and other MCAR methods. The advantages of using MMRM include; using all the available data to compensate for the data missing on a particular patient,^[12] no additional data manipulation is required to accommodate the missing data; also the analyses can be implemented using standard software such as the SAS Procedure Mixed (SAS Institute, Cary, NC).^[28]

Since missing data in clinical research will always be present and we cannot tell from data which category the missingness falls, the MMRM is suitable for both MAR and MCAR, it would be safe to use MMRM in clinical research where data is missing and use MNAR methods as sensitivity analysis. It is important to note that MMRM is one example from the family of likelihood-based analyses developed under MAR.

REFERENCES

1. Little, R.; Rubin, D.B. *Statistical Analysis with Missing Data*, John Wiley & Sons: New York, 1987.
2. The European Agency for the Evaluation of Medical Products (EMA). Points to Consider on Missing Data, 2001.
3. O'Neil, R. Considerations Regarding Choice of Primary Analysis in Longitudinal Trials with Dropouts: An FDA Presented at FDA/Industry Workshop, 2003.
4. International Conference on Harmonization (ICH). Guidelines on Statistical Principles for Clinical Trials (E9), 1998.
5. Mallinckrodt, G.; Chuang-Stein, C.; McSorley, P.; Schwatz, J.; Archibald, D.; Perahia, D.; Detka, M.; Alphs, L. A case study comparing a randomized withdrawal trial and a double-blind long-term trial for assessing the long-term efficacy of an antidepressant. *Pharm. Stat.* **2007**, *6*, 9–22.
6. Liu, G.; Gould, L. Comparison of alternative strategies for analysis of longitudinal trials with dropouts. *J. Biopharm. Stat.* **2002**, *12* (2), 207–226.
7. Shao, J.; Zhong, B. Last observation carry-forward and last observation analysis. *Stat. Med.* **2003**, *22*, 2429–2441.
8. Siddiqui, O.; Ali, M.W. A Comparison of the random-effects pattern mixture model with last-observation-carried-forward (LOCF) analysis in longitudinal clinical trials with dropouts. *J. Biopharm. Stat.* **1998**, *8* (4), 545–563.
9. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; John Wiley & Sons: New York, 1987.
10. Cnaan, A.; Laird, N.M.; Slasor, P. Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data. *Stat. Med.* **1997**, *16*, 2349–2380.
11. Mallinckrodt, C.H.; Clark, W.S.; David, S.R. Accounting for dropout bias using mixed-effects models. *J. Biopharm. Stat.* **2001**, *11* (1&2), 9–21.
12. Mallinckrodt, C.H.; Sanger, T.M.; Dube, S.; DeBrotta, D.J.; Molenberghs, G.; Carroll, R.J.; Potter, W.Z.; Tollefson, G.D. Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biol. Psychiatry.* **2003**, *53*, 754–760.
13. Lane, P. Handling drop-out in longitudinal clinical trials: a comparison of the LOCF and MMRM approaches. *Pharm. Stat.* **2008**, *7*, 93–106.
14. Detke, M.J.; Lu, Y.; Goldstein, D.J.; Hayes, J.R.; Demitrack, M.A. Duloxetine, 60 mg once daily, for major depressive disorder: a randomized double-blind placebo-controlled trial. *J. Clin. Psychiatry.* **2002**, *63* (4), 308–315.
15. Mallinckrodt, C.H.; Raskin, J.; Wohlreich, M.M.; Watkin, J.G.; Detke, M.J. The efficacy of duloxetine: a comprehensive summary of results from MMRM and LOCF-ANCOVA in eight clinical trial. *BMC Psychiatry.* **2004**, *4*, 26.
16. Gibbons, R.D.; Hedeker, D.; Elkin, I.; Waternaux, C.; Kraemer, H.C.; Greenhouse, J.B.; Shea, M.T.; Imber, S.D.; Sotsky, S.M.; Watkins, J.T. Some conceptual and statistical issues in analysis of longitudinal psychiatric data. *Arch. Gen. Psychiatry.* **1993**, *50*, 739–750.
17. Laird, N.M.; Ware, J.H. Random-effects models for longitudinal data. *Biometrics.* **1982**, *38*, 963–974.
18. Schafer, J.L. *Analysis of Incomplete Multivariate Data*; London: Chapman & Hall/CRC, 1997.
19. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592.
20. Little, R.J.A. Modeling the drop-out mechanism in repeated-measures studies. *J. Am. Stat. Assoc.* **1995**, *90*, 1112–1121.
21. Akaike, H.A. new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **1974**, *19* (6), 716–723.
22. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464.
23. Henderson, C.R. *Applications of Linear Models in Animal Breeding*; University of Guelph: Ontario, Canada, 1984.
24. McLean, C.R.; Sanders, W.L. Approximating Degrees of Freedom for Standard Errors in Mixed Linear Models. *Proceedings of the statistical Computing Section*; American Statistical Association: New Orleans, LA, 1988, 50–59.
25. Stroup, W.W. Predictable Functions and Prediction Space in the Mixed Model Procedure. *Applications of Mixed Models in Agriculture and Related Disciplines, Southern Cooperative Series Bulletin No. 343*; Louisiana Agricultural Experiment Station: Baton Rouge, 1989, 39–48.
26. Verbeke, G.; Molenberghs, G. *Linear Mixed Models for Longitudinal Data*; Springer: New York, 2000.
27. Heyting, A.; Tolboom, J.; Essers, J. Statistical handling of dropouts in longitudinal clinical trials. *Stat. Med.* **1992**, *11*, 2043–2061.
28. Little, R.; Yau, L. Intent-to-treat analysis for longitudinal studies with dropouts. *Biometrics.* **1996**, *52*, 1324–1333.

Modified Large Sample Method

Yonghee Lee

Forest Research Institute, Jersey City, New Jersey, U.S.A.

Abstract

Many researchers have suggested various methods for setting approximate confidence intervals on linear combinations of variance components. This entry focuses on the modified large sample method and works to detail its components and applications.

INTRODUCTION

Linear combinations of variance components are frequently considered when the parameter of interest is individual or gross variability in statistical experimental designs. Because of the lack of an exact confidence interval for a general linear combination of variance components, many researchers have suggested various methods for setting approximate confidence intervals on linear combinations of variance components. Among those methods, the Modified Large Sample (MLS) method has its merit in terms of simplicity, flexibility, and good empirical performance. Generally, the MLS method can give an approximate confidence bound on any linear combination of variance components of the following form:

$$c_1\sigma_1^2 + c_2\sigma_2^2 + \cdots + c_p\sigma_p^2 \quad (1)$$

where $\sigma_1^2, \dots, \sigma_p^2$ are variance components and $c_1, \dots, c_p$ are fixed nonzero real numbers. The idea of the MLS method was independently proposed by Howe^[1] and Graybill and Wang.^[2] The main idea of the MLS method was well described by Howe^[1] as follows: When each of the marginally exact quantiles of estimators of variance components are known, we can incorporate this knowledge into setting a good approximate confidence interval for a linear combination of variance components by interpolating the given exact quantiles.

In most clinical trials comparing a test drug and a control (e.g., a placebo control or an active control), treatment effect is usually established by comparing changes in mean responses from baseline of some primary study endpoints assuming that their corresponding variabilities are comparable. However, in practice, variabilities associated with the test drug and the control could be very different. When the variability of the test drug is much larger than that of the reference drug, safety of the test drug could be

a concern. Thus in addition to comparing mean responses between treatments, it is also of interest to compare the variabilities associated with the responses between treatments. Especially, in bioequivalence trials, the intrasubject (or within-subject) variability, the intersubject (or between-subject) variability, and the total variability become the primary endpoints if a drug company is to establish individual and/or population bioequivalence as well as average bioequivalence.

In the next section, the main idea of the original MLS method will be introduced when all coefficients are positive and estimators of variance components are independent. The extensions of the MLS method for the general cases such that (1) signs of coefficients are unrestricted; (2) setting confidence bound on a ratio of linear combinations of variance components; (3) its extension when estimators are not independent, will be discussed in subsequent sections, respectively. The last section will discuss applications of the MLS method in bioequivalence studies based on the guidances published by the Food and Drug Administration (FDA).

OVERVIEW

Denote the distribution of a random vector $\mathbf{x}$ by $\mathcal{L}(\mathbf{x})$. Hence $\mathcal{L}(\mathbf{x}) = \mathcal{L}(\mathbf{y})$ means that $\mathbf{x}$ and $\mathbf{y}$ have the same distribution. We use $N(\boldsymbol{\mu}, \mathbf{V})$ to denote the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$ and use χ_r^2 to denote a random variable having the central chi-square distribution with r degrees of freedom. The α th quantile of $N(0,1)$, $\mathcal{L}(\chi_r^2)$, the central t -distribution with degrees of freedom r , and F -distribution with degrees of freedom r_1, r_2 are denoted by z_α , $\chi_{\alpha,r}^2$, $t_{\alpha,r}$, and F_{α,r_1,r_2} , respectively. Note that α th quantile is the lower α th quantile such as $P(Z \leq z_\alpha) = \alpha$. The transpose of the vector $\mathbf{x}$ is denoted by $\mathbf{x}'$, and the Kronecker product of two matrices, $\mathbf{A}$ and $\mathbf{B}$, is denoted by $\mathbf{A} \otimes \mathbf{B}$.

MAIN IDEA OF MODIFIED LARGE SAMPLE METHOD

Consider the problem of setting a $1 - \alpha$ upper confidence bound for $\eta = c_1\sigma_1^2 + \dots + c_p\sigma_p^2$, where $\sigma_1^2, \dots, \sigma_p^2$ are unknown variance components and $c_1, \dots, c_p$ are known positive constants. Let $\hat{\sigma}_i^2$ be an unbiased estimator of σ_i^2 such that $\mathcal{L}(\hat{\sigma}_i^2) = \mathcal{L}(\sigma_i^2 n_i^{-1} \chi_{n_i}^2)$, $i = 1, \dots, p$. In practice, $\hat{\sigma}_i^2$ is typically a quadratic form of an observed normal random vector (e.g., a sample variance or a mean square error). For any fixed i , an exact $1 - \alpha$ upper confidence bound for σ_i^2 is

$$\frac{n_i}{\chi_{\alpha, n_i}^2} \hat{\sigma}_i^2 = \hat{\sigma}_i^2 + \sqrt{\hat{\sigma}_i^4 \left(\frac{n_i}{\chi_{\alpha, n_i}^2} - 1 \right)^2} \quad (2)$$

However, when $p \geq 2$, an exact upper confidence bound for η usually does not exist.

Suppose that $\hat{\sigma}_1^2, \dots, \hat{\sigma}_p^2$ are independent. Then

$$\begin{aligned} \text{Var}(c_1\hat{\sigma}_1^2 + \dots + c_p\hat{\sigma}_p^2) &= c_1^2 \text{Var}(\hat{\sigma}_1^2) + \dots + c_p^2 \text{Var}(\hat{\sigma}_p^2) \\ &= c_1^2 \sigma_1^4 n_1^{-2} \text{Var}(\chi_{n_1}^2) + \dots + c_p^2 \sigma_p^4 n_p^{-2} \text{Var}(\chi_{n_p}^2) \\ &= c_1^2 \sigma_1^4 2n_1^{-1} + \dots + c_p^2 \sigma_p^4 2n_p^{-1} \end{aligned} \quad (3)$$

An estimator of the variance in Eq. 3 is then obtained by replacing σ_i^4 in Eq. 3 by its estimator $\hat{\sigma}_i^4$. For large n_i 's, these results and the Central Limit Theorem lead to the following approximate $1 - \alpha$ upper confidence bound for η :

$$\begin{aligned} c_1\hat{\sigma}_1^2 + \dots + c_p\hat{\sigma}_p^2 + z_{1-\alpha} \sqrt{c_1^2 \hat{\sigma}_1^4 2n_1^{-1} + \dots + c_p^2 \hat{\sigma}_p^4 2n_p^{-1}} \\ = c_1\hat{\sigma}_1^2 + \dots + c_p\hat{\sigma}_p^2 + \sqrt{c_1^2 \hat{\sigma}_1^4 2z_{1-\alpha}^2 n_1^{-1} + \dots + c_p^2 \hat{\sigma}_p^4 2z_{1-\alpha}^2 n_p^{-1}} \end{aligned} \quad (4)$$

In view of Eq. 2, Graybill and Wang^[2] proposed to replace $2z_{1-\alpha}^2 n_i^{-1}$ in Eq. 4 by $\left(\frac{n_i}{\chi_{\alpha, n_i}^2} - 1 \right)^2$, $i = 1, \dots, p$, and referred this method as the Modified Large Sample (MLS) method. The resulting upper confidence bound for η is then

$$\begin{aligned} c_1\hat{\sigma}_1^2 + \dots + c_p\hat{\sigma}_p^2 \\ + \sqrt{c_1^2 \hat{\sigma}_1^4 \left(\frac{n_1}{\chi_{\alpha, n_1}^2} - 1 \right)^2 + \dots + c_p^2 \hat{\sigma}_p^4 \left(\frac{n_p}{\chi_{\alpha, n_p}^2} - 1 \right)^2} \end{aligned} \quad (5)$$

Using a similar argument, we can obtain the following MLS lower confidence bound for η :

$$\begin{aligned} c_1\hat{\sigma}_1^2 + \dots + c_p\hat{\sigma}_p^2 \\ - \sqrt{c_1^2 \hat{\sigma}_1^4 \left(\frac{n_1}{\chi_{1-\alpha, n_1}^2} - 1 \right)^2 + \dots + c_p^2 \hat{\sigma}_p^4 \left(\frac{n_p}{\chi_{1-\alpha, n_p}^2} - 1 \right)^2} \end{aligned} \quad (6)$$

An equal-tail, two-sided MLS confidence interval for η can be obtained by using the upper bounds (Eq. 5) and the lower bound (Eq. 6) as the interval limits.

The MLS confidence bound has the following properties:

Property 1. The confidence bound in Eq. 5 is asymptotically correct; that is, the coverage probability of the confidence bound converges to $1 - \alpha$ when all n_i 's increase to infinity. This is because

$$\lim_{n \rightarrow \infty} \frac{n}{2} \left(\frac{n}{\chi_{\alpha, n}^2} - 1 \right)^2 = z_{1-\alpha}^2$$

which can be proven using the Cornish–Fisher expansion.

Property 2. When $\sigma_i^2 > 0$ and $\sigma_j^2 = 0$ for all $j \neq i$, the confidence bound in Eq. 5 reduces to the confidence bound in Eq. 2. Hence it is an exact confidence bound. This property is not enjoyed by the confidence bound in Eq. 4.

Property 1 implies that the coverage probability of the MLS confidence bound converge to the intended confidence coefficient $1 - \alpha$ as sample sizes become large. The finite sample evaluation of coverage probability of the MLS confidence bounds were investigated by using the numerical integration in Ref. [2]. The numerical evaluation in Ref. [2] indicates that the MLS method is comparable to the method in Refs. [3,4] and is also conservative in the sense that the finite sample coverage probabilities are larger than $1 - \alpha$ for most cases. Property 2 is the main and unique advantage of the MLS method over other competing methods. When one of the variance components is relatively larger than the others, the MLS method gives an almost exact confidence bound. This property was also discussed by Howe.^[1] Other than these two properties of the MLS method, another good property is its simplicity.

AN EXTENSION OF THE MODIFIED LARGE SAMPLE METHOD TO DIFFERENCE OF VARIANCE COMPONENTS

In this section, we will consider an extension of the MLS method when the signs of coefficient in Eq. 1 are unrestricted. Originally, Howe^[1] considered an approximate confidence bound on intersubject (between-group) variance in one-way classification model by using a similar idea to Graybill and Wang.^[2] First, Howe's idea will be introduced and, second, its generalizations will be discussed.

Consider two mean squares $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, which are distributed independently and

$$\mathcal{L}(\hat{\sigma}_1^2) = \mathcal{L}\left(\frac{(\sigma_a^2 + \sigma_e^2)\chi_{n_1}^2}{n_1}\right), \quad \mathcal{L}(\hat{\sigma}_2^2) = \mathcal{L}\left(\frac{\sigma_e^2 \chi_{n_2}^2}{n_2}\right)$$

Therefore an unbiased estimator of σ_a^2 is given by $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$. For this example, if

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \leq F_{\alpha, n_1, n_2}$$

then the upper confidence bound of σ_a^2 is zero. Howe^[1] considered this as an extra piece of readily available information and incorporated this information with the idea of the MLS method to construct more accurate confidence bound on σ_a^2 . The $1 - \alpha$ MLS upper confidence bound of σ_a^2 can be defined as

$$\hat{\sigma}_1^2 - \hat{\sigma}_2^2 + \sqrt{V}$$

where

$$V = \hat{\sigma}_1^4 \left(\frac{n_1}{\chi_{\alpha, n_1}^2} - 1 \right)^2 + A \hat{\sigma}_2^4 \quad (7)$$

Now the constant A in Eq. 7 can be determined by forcing the upper confidence bound $\hat{\sigma}_1^2 - \hat{\sigma}_2^2 + \sqrt{V}$ to be 0 when $\hat{\sigma}_1^2 / \hat{\sigma}_2^2 = F_{\alpha, n_1, n_2}$. This imposed condition leads to the equation for A . By using the solution for A from the given equation, $1 - \alpha$ MLS upper confidence bound of σ_a^2 is given by

$$\begin{cases} 0 & \text{if } F \leq F_{\alpha, n_1, n_2} \\ \hat{\sigma}_1^2 - \hat{\sigma}_2^2 + \sqrt{V} & \text{if } F > F_{\alpha, n_1, n_2} \end{cases} \quad (8)$$

where $F = \hat{\sigma}_1^2 / \hat{\sigma}_2^2$ and

$$V = \hat{\sigma}_1^4 \left(\frac{n_1}{\chi_{\alpha, n_1}^2} - 1 \right)^2 + \hat{\sigma}_2^4 \left[(F_{\alpha, n_1, n_2} - 1)^2 - F_{\alpha, n_1, n_2}^2 \left(\frac{n_1}{\chi_{\alpha, n_1}^2} - 1 \right)^2 \right]$$

Note that the confidence bound in Eq. 8 is using an extra information such as $\sigma_a^2 > 0$ in the model considered.

Consider a more general parameter $c_1 \sigma_1^2 - c_2 \sigma_2^2$, where c_1 and c_2 are known positive real numbers. However, for this general case, if the parameter $c_1 \sigma_1^2 - c_2 \sigma_2^2$ can be negative, the MLS confidence upper bound in Eq. 8 is no longer valid. For this case, Lu et al.^[5] extended Howe's confidence bound in Eq. 8 as

$$\begin{cases} c_1 \hat{\sigma}_1^2 - c_2 \hat{\sigma}_2^2 + \sqrt{V^*} & \text{if } F \leq F_{\alpha, n_1, n_2} \\ c_1 \hat{\sigma}_1^2 - c_2 \hat{\sigma}_2^2 + \sqrt{V} & \text{if } F > F_{\alpha, n_1, n_2} \end{cases} \quad (9)$$

where F and V are the same as in Eq. 8 except using $c_i \hat{\sigma}_i^2$ instead of $\hat{\sigma}_i^2$ and

$$V^* = c_1^2 \hat{\sigma}_1^4 \left[\left(\frac{1}{F_{\alpha, n_1, n_2}} - 1 \right)^2 - \left(\frac{1}{F_{\alpha, n_1, n_2}} \right)^2 \left(\frac{n_2}{\chi_{1-\alpha, n_2}^2} - 1 \right)^2 \right] + \hat{\sigma}_2^4 \left(\frac{n_2}{\chi_{1-\alpha, n_2}^2} - 1 \right)^2$$

Note that the bound in Eq. 9 is exact when 1) $\sigma_1 = 0$; 2) $\sigma_2 = 0$; 3) $c_1 \sigma_1^2 = c_2 \sigma_2^2$. Furthermore, Ting et al.^[6] considered MLS confidence bounds for more general cases such that the parameter of interest can be described as

$$\sum_{i=1}^q c_i \sigma_i^2 - \sum_{j=q+1}^p c_j \sigma_j^2$$

where $q < p$ and $c_1, \dots, c_p$ are all positive real numbers. Also, the alternative form of the MLS upper confidence bound was proposed by Ting et al.^[6] without considering the F test in Eq. 9, and it can be described as

$$\sum_{i=1}^q c_i \hat{\sigma}_i^2 - \sum_{j=q+1}^p c_j \hat{\sigma}_j^2 + \sqrt{W} \quad (10)$$

where

$$W = \sum_{i=1}^q c_i^2 \hat{\sigma}_i^4 L_i^2 + \sum_{j=q+1}^p c_j^2 \hat{\sigma}_j^4 L_j^2 + \sum_{i=1}^q \sum_{j=q+1}^p c_i c_j \hat{\sigma}_i^2 \hat{\sigma}_j^2 L_{ij}$$

and L_i, L_j, L_{ij} are appropriately chosen to make the MLS upper confidence bound exact for some special cases. Note that we can simplify the confidence bound in Eq. 10 by ignoring the cross-product terms $\sum_{i=1}^q \sum_{j=q+1}^p c_i c_j \hat{\sigma}_i^2 \hat{\sigma}_j^2 L_{ij}$. Moreover, if we consider Property 1 in the previous section as the imposed condition for special cases (i.e., an exact confidence bound when only one variance component is positive and the others are zero), the following simplified $1 - \alpha$ MLS confidence upper bound can be obtained:

$$\sum_{i=1}^q c_i \hat{\sigma}_i^2 - \sum_{j=q+1}^p c_j \hat{\sigma}_j^2 + \sqrt{W^*} \quad (11)$$

where

$$W^* = \sum_{i=1}^q c_i^2 \hat{\sigma}_i^4 L_i^2 + \sum_{j=q+1}^p c_j^2 \hat{\sigma}_j^4 L_j^2$$

and

$$L_i^2 = \left(\frac{n_i}{\chi_{\alpha, n_i}^2} - 1 \right)^2, \quad i = 1, \dots, q;$$

$$L_j^2 = \left(\frac{n_j}{\chi_{1-\alpha, n_j}^2} - 1 \right)^2, \quad j = q+1, \dots, p$$

Note that the form in Eq. 11 is the most simplified form of the MLS confidence bound and it is used for tests of Individual/Population Bioequivalence, which will be discussed in the last section.

AN EXTENSION OF THE MODIFIED LARGE SAMPLE METHOD TO RATIO OF VARIANCE COMPONENTS

In many experimental designs, a certain ratio of variance components becomes the parameter of interest. For example, under one-way random effect model, a ratio of intersubject variance and total variance is often considered as an important parameter, which is called as the heritability measure in genetics. The MLS confidence bound on the ratio of variance components have been extensively developed by Burdick et al.^[7-10] In this section, the MLS confidence bound by Gui et al.^[8] will be introduced.

Consider the problem of setting a confidence upper bound on the ratio

$$\rho = \frac{\sigma_2^2 + \sigma_3^2}{\sigma_1^2 + \sigma_2^2}$$

The $1 - \alpha$ upper bound on the ratio ρ can be defined by U such that $P(\rho < U)$ is close to $1 - \alpha$. Note that $P[\rho < U] = P[(\sigma_2^2 + \sigma_3^2)/(\sigma_1^2 + \sigma_2^2) < U] = P(\gamma_U < 0)$ such that

$$\gamma_U = -U\sigma_1^2 + (1-U)\sigma_2^2 + \sigma_3^2 \quad (12)$$

Because γ_U in Eq. 12 is a linear combination of variance components, the MLS method discussed in the previous section can be applied to find $1 - \alpha$ MLS confidence upper bound on γ_U such as $P(\gamma_U < U^*) \approx 1 - \alpha$. On the other hand, we also desire $P(\gamma_U < 0) \approx 1 - \alpha$. Therefore we have the quadratic equation $U^* = 0$ in terms of U and the solution for U can be considered as $1 - \alpha$ MLS confidence upper bound on the ratio ρ .

However, there is one difficulty in constructing confidence bound on γ_U in Eq. 12 because the sign of U in Eq. 12 cannot be determined prior to setting the bound. For example, if $\sigma_3^2 > \sigma_1^2$, then $\rho > 1$, which, in turn, implies $U > 1$. Therefore we can determine the sign of $1 - U$, which is the coefficient of σ_2^2 in Eq. 12. On the other hand, if $\sigma_3^2 < \sigma_1^2$, then $\rho < 1$. For this case, Gui et al.^[8] recommend to restrict the range of U by $0 < U < 1$. For either case, as discussed in the previous section, we can apply the MLS method by Lu et al.^[5] to construct $1 - \alpha$ MLS confidence upper bound on γ_U and its form is given in Eq. 9. Gui et al.^[8] proposed the procedure to find MLS confidence bound on ρ such that 1) assume $\sigma_3^2 > \sigma_1^2$; 2) compute an upper bound U ; 3) if $U > 1$, then stop; 4) if $U < 1$, then assume $\sigma_3^2 \leq \sigma_1^2$ and find U again.

Gui et al.^[8] considered constructing the MLS confidence bound on a more general parameter such that

$$\rho = \frac{\sum_{i=1}^q c_i \sigma_i^2 - \sum_{j=q+1}^p c_j \sigma_j^2}{\sum_{k=1}^r c_k \sigma_k^2}$$

The explicit form of confidence interval and more details can be found in Ref. [8].

AN EXTENSION OF THE MODIFIED LARGE SAMPLE METHOD WHEN ESTIMATORS ARE DEPENDENT

One of the important assumptions for the MLS method is that estimators of variance components should be independent. However, in practice, this independent assumption might not be satisfied. For example, if we consider a linear combination of variance components in the multivariate normal distribution, the sample variances are not independent. Furthermore, in replicated crossover design, the estimators of intersubject variability (or total variability) between two treatments are not independent.

Note that when $\hat{\sigma}_1^2, \dots, \hat{\sigma}_p^2$ are dependent, result Eq. 3 does not hold and, consequently, Eqs. 4–6 are asymptotically incorrect; that is, their coverage probabilities do not converge to the nominal level $1 - \alpha$, although confidence bounds Eqs. 5 and 6 still enjoy Property 2 in the “Main Idea of Modified Large Sample Method” section. Lee^[11] considered the extension of the MLS method when estimators are dependent. We will consider a main idea of the extension of the MLS method and present an example with multivariate normal distribution to illustrate an application of the extension.

Assume that $\hat{\sigma}_i^2$'s are quadratic forms of an observed normal random vector and c_i 's are known nonzero real numbers. Our extension builds on the fact that the quadratic form $c_1 \hat{\sigma}_1^2 + \dots + c_p \hat{\sigma}_p^2$ may have a *chi-square representation* in the sense that

$$\mathcal{L}(c_1 \hat{\sigma}_1^2 + \dots + c_p \hat{\sigma}_p^2) = \mathcal{L}(\lambda_1 n_1^{-1} \chi_{n_1}^2 + \dots + \lambda_q n_q^{-1} \chi_{n_q}^2) \quad (13)$$

where $\chi_{n_1}^2, \dots, \chi_{n_q}^2$ are independent chi-square random variables and λ_i 's are unknown parameters that are function of variance components and coefficients c_i 's. Also, assume that the signs of λ_i 's are known. However, in each application, one has to derive the explicit form of the chi-square representation. When $\hat{\sigma}_1^2, \dots, \hat{\sigma}_p^2$ are independent, Eq. (13) holds with $\lambda_i = c_i \sigma_i^2$.

Although the chi-square random variables $\chi_{n_1}^2, \dots, \chi_{n_q}^2$ in Eq. (13) are not observable, the coefficients $\lambda_1, \dots, \lambda_q$ usually have consistent estimators that can be obtained from observed sample. Using Eq. 13 and applying the argument by Lu et al.^[5] to the independent linear combination

$\lambda_1 n_1^{-1} \chi_{n_1}^2 + \dots + n_q^{-1} \lambda_q \chi_{n_q}^2$, we obtain the following extended MLS upper and lower confidence bounds:

$$c_1 \hat{\sigma}_1^2 + \dots + c_p \hat{\sigma}_p^2 + \sqrt{\hat{\lambda}_1^2 \left(\frac{n_1}{u_1} - 1 \right)^2 + \dots + \hat{\lambda}_q^2 \left(\frac{n_q}{u_q} - 1 \right)^2} \quad (14)$$

and

$$c_1 \hat{\sigma}_1^2 + \dots + c_p \hat{\sigma}_p^2 - \sqrt{\hat{\lambda}_1^2 \left(\frac{n_1}{l_1} - 1 \right)^2 + \dots + \hat{\lambda}_q^2 \left(\frac{n_q}{l_q} - 1 \right)^2} \quad (15)$$

where $\hat{\lambda}_i$ is a consistent estimator of λ_i , $i = 1, \dots, q$,

$$u_i = \begin{cases} \chi_{\alpha, n_i}^2 & \lambda_i > 0 \\ \chi_{1-\alpha, n_i}^2 & \lambda_i < 0 \end{cases} \quad \text{and} \quad (16)$$

$$l_i = \begin{cases} \chi_{\alpha, n_i}^2 & \lambda_i < 0 \\ \chi_{1-\alpha, n_i}^2 & \lambda_i > 0 \end{cases}$$

Note that, when $\hat{\sigma}_1^2, \dots, \hat{\sigma}_p^2$ are independent, $\hat{\lambda}_i = c_i \hat{\sigma}_i^2$. The extended MLS confidence bounds Eq. 14 and Eq. 15 enjoy Properties 1 and 2 of the original MLS method.

Because most of the estimators of variance components under the normal assumption can be described as a quadratic form of multivariate normal distribution, we will consider multivariate normal distribution to illustrate an idea of extension of the MLS method. Let $\mathbf{y}_1, \dots, \mathbf{y}_n$ be independent and identically distributed p -dimensional random vectors with $\mathcal{L}(\mathbf{y}_i) = N(\mu, \mathbf{V})$. Suppose that the parameter η is a linear combination of the p diagonal elements of $\mathbf{V}$ such that $\eta = c_1 \sigma_1^2 + \dots + c_p \sigma_p^2$. Let be the sample covariance matrix and $\hat{\sigma}_i^2$ be its i th diagonal element. Note that $\hat{\sigma}_1^2, \dots, \hat{\sigma}_p^2$ are dependent unless $\mathbf{V}$ is diagonal. Define $\mathbf{y} = (\mathbf{y}_1', \dots, \mathbf{y}_n')'$. Then, a linear combination of estimators of variance components can be described as a quadratic form such that

$$c_1 \hat{\sigma}_1^2 + \dots + c_p \hat{\sigma}_p^2 = \mathbf{y}' \mathbf{A} \mathbf{y}$$

is a quadratic form, where

$$\mathbf{A} = \mathbf{C} \otimes \mathbf{J}_n, \quad \mathbf{J}_n = (n-1)^{-1} (\mathbf{I}_n - n^{-1} \mathbf{1}_n \mathbf{1}_n')$$

$\mathbf{I}_n$ is the identity matrix of order n , $\mathbf{1}_n$ is the n -dimensional vector of 1's, and $\mathbf{C}$ is a diagonal matrix whose i th diagonal element is c_i . Note that $\mathcal{L}(\mathbf{y}) = N(\tilde{\mu}, \tilde{\mathbf{V}})$, where $\tilde{\mu} = \mu \otimes \mathbf{1}_n$ and $\tilde{\mathbf{V}} = \mathbf{V} \otimes \mathbf{I}_n$. The distribution of quadratic form $\mathbf{y}' \mathbf{A} \mathbf{y}$ is not straightforward, but it can be shown that the distribution of this quadratic form is equivalent to that of a linear combination of independent chi-square random variables by using the moment generating function of $\mathbf{y}' \mathbf{A} \mathbf{y}$. The moment generating function of $\mathbf{y}' \mathbf{A} \mathbf{y}$ is given by

$$E[\exp(t \mathbf{y}' \mathbf{A} \mathbf{y})] = \left(1 - \frac{2t\lambda_1}{n-1}\right)^{-\frac{n-1}{2}} \dots \left(1 - \frac{2t\lambda_p}{n-1}\right)^{-\frac{n-1}{2}}$$

where $\lambda_1, \dots, \lambda_p$ are the roots of the equation

$$|\lambda \mathbf{I}_p - \mathbf{C} \mathbf{V}| = |\lambda \mathbf{V}^{-1} - \mathbf{C}| = 0 \quad (17)$$

i.e., $\lambda_1, \dots, \lambda_p$ are the eigenvalues of $\mathbf{C} \mathbf{V}$.

Hence the moment generating function of the quadratic form $\mathbf{y}' \mathbf{A} \mathbf{y}$ is same as that of a linear combination of independent chi-square random variable so that the quadratic form has chi-square representation such as

$$\mathcal{L}(\mathbf{y}' \mathbf{A} \mathbf{y}) = \mathcal{L}(\lambda_1 (n-1)^{-1} \chi_{n-1,1}^2 + \dots + \lambda_p (n-1)^{-1} \chi_{n-1,p}^2)$$

where $\chi_{n-1,i}^2$, $i = 1, \dots, p$ are independent chi-square random variables with $n-1$ degrees of freedom. This means that $c_1 \hat{\sigma}_1^2 + \dots + c_p \hat{\sigma}_p^2 = \mathbf{y}' \mathbf{A} \mathbf{y}$ has the chi-square representation (Eq. 13) with $n_i = n-1$ and λ_i 's as the solutions of Eq. 17. Let $\hat{\lambda}_1, \dots, \hat{\lambda}_p$ be the solution of Eq. 17 with $\mathbf{V}$ replaced by, which is any consistent estimator of $\mathbf{V}$. Then $\hat{\lambda}_i$ is a consistent estimator of λ_i as $n \rightarrow \infty$ and the extended MLS confidence bounds Eq. 14 and Eq. 15 can be obtained.

We will consider the bivariate normal distribution with simulation studies for further illustration. The covariance matrix $\mathbf{V}$, its estimator $\hat{\mathbf{V}}$, and the coefficient matrix $\mathbf{C}$ are given by

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}, \quad \hat{\mathbf{V}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 \end{pmatrix},$$

$$\mathbf{C} = \begin{pmatrix} c_1 & 0 \\ 0 & c_2 \end{pmatrix}$$

Note that $\eta = c_1 \sigma_1^2 + c_2 \sigma_2^2$, which is estimated by $c_1 \hat{\sigma}_1^2 + c_2 \hat{\sigma}_2^2$. Without loss of generality, let $\lambda_1 = \lambda_-$ and $\lambda_2 = \lambda_+$ be the eigenvalues of $\mathbf{C} \mathbf{V}$, i.e.

$$\lambda_{\pm} = \frac{c_1 \sigma_1^2 + c_2 \sigma_2^2 \pm \sqrt{(c_1 \sigma_1^2 - c_2 \sigma_2^2)^2 + 4 c_1 c_2 \rho^2 \sigma_1^2 \sigma_2^2}}{2} \quad (18)$$

Estimators $\hat{\lambda}_1$ and $\hat{\lambda}_2$ can be obtained by replacing σ_i^2 and $\rho \sigma_1 \sigma_2$ in the previous equation by the corresponding estimators $\hat{\sigma}_i^2$ and $\hat{\sigma}_{12}$, respectively.

APPLICATION OF THE MODIFIED LARGE SAMPLE METHOD TO BIOEQUIVALENCE STUDIES

Bioequivalence between two different formulations of a drug product is established when the two formulations have equivalent bioavailability. Bioavailability is described

as the rate and extent that a drug in some dosage form is absorbed and becomes available at the site of the pharmacological effect. Pharmacokinetic (PK) response such as the area under the concentration time curve (AUC) or the maximum concentration ($C_{\max}$) is usually considered as measures of bioavailability.^[12–16]

The recent FDA guidance^[16] describes average bioequivalence (ABE), individual bioequivalence (IBE), and population bioequivalence (PBE) for establishment of bioequivalence between a test formulation and a reference formulation. Average bioequivalence can be claimed if the 90% confidence interval of the ratio of population means between the test formulation and the reference formulation falls within 80% and 125%. However, ABE has limitations because it only focuses on the average of bioavailability, and ignores the variability of bioavailability when assessing bioequivalence. As a result, it does not guarantee drug interchangeability. Drug *interchangeability* include the concepts of drug *switchability* and *prescribability*. Drug switchability can be established if a patient who receives the reference formulation can expect similar bioavailability when switches in the test formulation occur. The FDA recommends IBE be used to assess drug switchability. Individual bioequivalence takes into consideration of within-subject variability and variability due to subject by formulation interaction. Drug prescribability can be claimed if the distribution of PK responses to the reference formulation is sufficiently similar to the distribution of PK responses to the test formulation. The FDA suggests PBE be used to assess drug prescribability. Population bioequivalence includes comparison of total variabilities between formulations. Hence if PBE is claimed, the physician can prescribe one of two formulations to the patient with the same expectation for the therapeutic effect.

The measures for IBE and PBE are described in the guidance.^[16] Let y_T be the log-transformed PK response of a subject that is treated by the test formulation, and y_R, y_R^* be two identically distributed log-transformed PK responses to the reference formulation. Motivated by Schall and Luus,^[17] the FDA^[16] defines a measure of PBE by the quantity

$$\theta = \begin{cases} \frac{E(y_R - y_T)^2 - E(y_R - y_R^*)^2}{E(y_R - y_R^*)^2/2} & \text{if } E(y_R - y_R^*)^2/2 \geq \sigma_0^2 \\ \frac{E(y_R - y_T)^2 - E(y_R - y_R^*)^2}{\sigma_0^2} & \text{if } E(y_R - y_R^*)^2/2 < \sigma_0^2 \end{cases} \quad (19)$$

where σ_0^2 is a constant specified by the FDA.

Assume that y_T, y_R , and y_R^* in Eq. 19 are from the same subject. Then, IBE can be claimed if the following null hypothesis H_0 is rejected at the 5% level of significance:

$$H_0: \theta \geq \theta_l \quad \text{vs.} \quad H_1: \theta < \theta_l \quad (20)$$

where θ_l is an upper limit specified by the FDA. Note that if y_T, y_R , and y_R^* in Eq. 19 are independent observations from different subjects, θ becomes the measure to establish PBE. In Ref. [16] 2×4 crossover design is recommended to assess the measure of PBE and IBE.

Hyslop et al.^[18] propose a statistical test for IBE by using the MLS method after the original hypothesis (Eq. 20) is transformed to an equivalent hypothesis in which new parameter can be written as a linear combination of variance components. The FDA^[16] adopted the statistical test for IBE by Hyslop et al.^[18] and applied the same argument to construct a statistical test for PBE. Alternatively, Quiroz et al.^[19,20] proposed tests to assess IBE and PBE by applying the MLS confidence upper bound on the original ratio criteria in Eq. 19. The FDA's tests for IBE/PBE are simpler than those by Quiroz, but the FDA's test for PBE is conservative so that it gives very low power in some situations. The test for PBE by Quiroz et al.^[20] is better than the FDA's test in terms of its size, which is more closer than nominal level 0.05 than the FDA's test.

The key assumption for the original MLS method is that estimators of variance components should be independent. However, it is shown by Chow et al.^[21] that the estimators of the variance components in the test for PBE by the FDA^[16] are not independent so that the asymptotic size of the test is always strictly less than the nominal level except for some special cases. This implies that the FDA test is too conservative and gives unnecessarily low powers. By using normal approximation, Chow et al.^[21] proposed an asymptotic test that gives the correct asymptotic size. Lee^[11] proposed a test for PBE by using the extended MLS, which can be applied to the case when estimators are dependent. The test based on the extended MLS method has more power than the FDA's test for PBE and its size is very close the nominal level. Furthermore, Lee et al.^[22] showed that the extended MLS method can be used to compare intersubject and total variability under replicate crossover designs.

For locally acting drug product such as nasal aerosol and nasal spray, which are not intended to be absorbed into the bloodstream, bioavailability and the corresponding bioequivalence can be assessed in in vitro studies alone.^[23,24] Statistical methods for assessment of in vitro bioequivalence testing for locally acting drug can be classified into two categories: the nonprofile analysis and the profile analysis. The FDA^[23] used the same criteria as that for PBE and applied the MLS method to construct a test for the nonprofile analysis. Note that the experimental design for the nonprofile analysis is a parallel design so that estimators of variance components are independent, which implies that the original MLS method can be applied without consideration about dependency among estimators.

CONCLUSION

The MLS method is useful in providing statistical inference for any linear combination of variance components of the form:

$$c_1\sigma_1^2 + c_2\sigma_2^2 + \cdots + c_p\sigma_p^2$$

Its applications in biopharmaceutical research include the assessment of in vivo bioequivalence (i.e., IBE and PBE) and in vitro bioequivalence. In recent years, the assessment of reproducibility in clinical research has attracted much attention. The concept of the MLS method can be applied for obtaining statistical inference on reproducibility of study endpoints of interest in clinical research as well.

REFERENCES

- Howe, W.G. Approximate confidence limits on the mean of $x + y$ where x and y are two tabled independent random variables. *J. Am. Stat. Assoc.* **1974**, *69*, 789–794.
- Graybill, F.A.; Wang, C.-M. Confidence intervals on non-negative linear combinations of variances. *J. Am. Stat. Assoc.* **1980**, *75*, 869–873.
- Satterthwaite, F.E. Synthesis of variance. *Psychometrika*. **1941**, *6*, 309–316.
- Satterthwaite, F.E. An approximate distribution of estimates of variance components. *Biom. Bull.* **1946**, *2*, 110–114.
- Lu, T.-F.C.; Graybill, F.A.; Burdick, R.K. Confidence intervals on a difference of expected mean squares. *J. Stat. Plan. Inference*. **1988**, *18*, 35–43.
- Ting, N.; Burdick, R.K.; Graybill, F.A.; Jeyaratnam, S.; Lu, T.-F.C. Confidence intervals on linear combinations of variance components that are unrestricted in sign. *J. Stat. Comput. Simul.* **1990**, *35*, 135–143.
- Lu, T.-F.C.; Graybill, F.A.; Burdick, R.K. Confidence intervals on the ratio of expected mean squares $(\theta_1 - d\theta_2)/\theta_3$. *J. Stat. Plan. Inference*. **1989**, *21* (2), 179–190.
- Ting, N.; Burdick, R.K.; Graybill, F.A. Confidence intervals on ratio of positive linear combinations of variance components. *Stat. Probab. Lett.* **1991**, *11* (6), 523–528.
- Burdick, R.K.; Graybill, F.A. *Confidence Intervals on Variance Components*; Marcel Dekker, Inc.: New York, 1992, Vol. 127.
- Gui, R.; Graybill, F.A.; Burdick, R.K.; Ting, N. Confidence intervals on ratios of linear combinations for non-disjoint sets of expected mean squares. *J. Stat. Plan. Inference*. **1995**, *48*, 215–227.
- Lee, Y. Confidence Intervals and Tests for Linear Combinations of Variance Components When Estimators Are Dependent. In Ph.D. Dissertation; University of Wisconsin at Madison: 2002.
- FDA. *Guidance of Statistical Procedure for Bioequivalence Studies Using a Standard Two-Treatment Crossover Design*; Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, 1992.
- FDA. *In Vivo Bioequivalence Studies Based on Population and Individual Bioequivalence Approaches*; Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, 1999.
- FDA. *Guidance for Industry: Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Considerations*; Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, 2000.
- FDA. *Statistical Approaches to Establishing Bioequivalence*; Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration; 2001.
- Chow, S.-C.; Liu, J.-P. *Design and Analysis of Bioavailability and Bioequivalence Studies*, 2nd Ed.; Marcel Dekker, Inc.: New York, 2000.
- Schall, R.; Luus, H.G. On population and individual bioequivalence. *Stat. Med.* **1993**, *12*, 1109–1124.
- Hyslop, T.; Hsuan, F.; Holder, D.J. A small sample confidence interval approach to assess individual bioequivalence. *Stat. Med.* **2000**, *19* (20), 2885–2897.
- Quiroz, J.; Ting, N.; Wei, G.C.G.; Burdick, R.K. A modified large sample approach in the assessment of population bioequivalence. *J. Biopharm. Stat.* **2000**, *10* (4), 527–544.
- Quiroz, J.; Ting, N.; Wei, G.C.G.; Burdick, R.K. Alternative confidence intervals for the assessment of bioequivalence in four-period cross-over designs. *Stat. Med.* **2002**, *21* (13), 1825–1847.
- Chow, S.-C.; Shao, J.; Wang, H. Statistical test for population bioequivalence. *Statistica Sinica*. **2003**, 539–554.
- Lee, Y.; Shao, J.; Chow, S.-C.; Wang, H. Test for inter-subject and total variabilities under crossover design. *J. Biopharm. Stat.* **2002**, *12* (4), 503–534.
- FDA. *Statistical Information for In Vitro Bioequivalence Data*; Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, 1999.
- FDA. *Bioavailability and Bioequivalence Studies for Nasal Aerosols and Nasal Sprays for Local Action*; Office of Generic Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, 2003.

Multicenter Trials

William J. Huster

Eli Lilly and Company, Indianapolis, Indiana, U.S.A.

Abstract

A multicenter clinical trial is a clinical trial involving more than one clinical center that is conducted to evaluate the effectiveness of a therapeutic agent according to a common protocol. This entry reviews the statistical design and analysis involved with multicenter trials and discusses the regulatory requirements for these trials.

INTRODUCTION

A multicenter clinical trial is a clinical trial (see *Clinical Trials* entry) involving more than one clinical center (i.e., study site, investigator, field site, clinic) that is conducted to evaluate the effectiveness of a therapeutic agent according to a common protocol. Multicenter trials have been utilized in clinical medicine since the 1940s, as documented by Bradford Hill^[1] in his seminal work on the statistical aspects of trials of antihistamines, cortisone, and streptomycin. Multicenter trials are employed primarily for two reasons: (1) the use of multiple centers allows adequate and/or rapid accrual of patients, and (2) the involvement of multiple centers enhances the generalizability of the results because of the anticipated heterogeneity in patient populations and center practices. This latter rationale is attractive because it more closely resembles how the therapeutic agent will be utilized when commercially available. Other reasons for utilizing multicenter trials are to introduce state-of-the-art therapies to the medical community as well as to solicit a wider range of clinical opinions concerning the utility of therapeutic agent.

OVERVIEW

Many of the aspects of multicenter trials run by the pharmaceutical industry are similar to those of multicenter trials run by collaborative efforts sponsored by government agencies (e.g., National Institutes of Health). However, there are several important differences. Pharmaceutical industry trials evaluate new therapeutic agents for regulatory approval in a specific patient population. Often, these multicenter trials are used in phase 3 of drug development (see *Drug Development* entry), sometimes as the pivotal studies in the new drug application to the Food and Drug Administration (see *Food and Drug Administration* entry). Collaborative trials address outstanding medical questions for approved therapeutic agents in a broader patient population; the results of these trials profoundly affect current medical

practice.^[2] In addition, because pharmaceutical industry trials evaluate new therapeutic agents, there is a need for timely safety monitoring and periodic safety reporting to regulatory agencies during the trial. Collaborative trials are not subject to such regulatory oversight, because they evaluate approved therapeutic agents about which the safety profile is known. See “Multicenter Trials” in the *Encyclopedia of Biostatistics*^[3] for a more complete description of collaborative multicenter trials. The following discusses aspects of multicenter trials in the pharmaceutical industry.

By its nature, a multicenter clinical trial is organizationally complex, requiring careful protocol development (see *Protocol Development* entry) and implementation as well as clear organizational structure and decision making.

Protocol Development and Implementation

The protocol should be clearly written so that it can be easily understood and uniformly implemented across clinical centers. In particular, entry and therapy/patient discontinuation criteria should be rigorously standardized using verifiable decision rules. The pharmaceutical company (i.e., drug sponsor) should hold investigator meetings and training sessions for study coordinators to assure common understanding of evaluation procedures, especially for the efficacy endpoint (see entries on *Clinical Endpoint* and *Surrogate Endpoint*). Manuals of standard operating procedures should be created to enhance uniformity and protocol adherence further. Ongoing monitoring by the sponsor of each clinical center’s compliance to the protocol should occur throughout the trial. The foregoing concerns are critically important to the success of the study.

Many of the data-collection and data-management issues associated with multicenter trials are similar to those associated with single-center clinical trials. For example, the data collection tool should be designed to collect essential data in a consistent manner. Also, there should be a quality control process in place to ensure that the collected data match the patient’s medical records.

Several new data-management issues are introduced by multicenter trials compared to a single-center trials. For example, a central resource center can be used to perform the evaluation of important objective safety and efficacy endpoints as well as entry criteria. Use of a central resource center ensures standardized, blinded review of these critical data. The logistics of sample collection and flow from the clinical centers to the central resource centers (e.g., collection of blood sent for evaluation to a central laboratory) and the subsequent transfer of the data to the data-management center should be clearly spelled out, facilitated, and monitored by the sponsor.

It is the sponsor's responsibility to have in place a process for monitoring the study for emerging safety issues and for communicating any issues in a timely manner to the regulatory agencies and, if necessary, to the clinical investigators.

Organizational Structure and Decision Making

The organizational structure of and decision-making process for the multicenter trial should be determined and agreed to prior to study initiation. In particular, the roles and responsibilities of each committee should be defined. Usually, there is a study chair (e.g., the principle investigator) and a committee of investigators. The sponsor facilitates communication between investigators by conducting investigator meetings as well as frequent mailings. It is often the role of the investigator committee, in consultation with the sponsor, to plan publications and to resolve authorship.

The study coordinating center is responsible for managing the clinical centers (e.g., monitoring patient accrual, ensuring protocol compliance) and the data-coordinating center is responsible for managing the data and analysis. These coordinating centers may be a contracted research organization (see entry on *Contract Research Organization (CRO)*), an academic center, or medical/data-management/statistical personnel of the sponsor. Also, as already mentioned, there may be several additional resource centers, such as central laboratories and evaluation centers, which are usually external to the sponsor so as to maintain perceived integrity of the results.

Often, Data Safety Monitoring Boards (see also entry on *Data Monitoring Committees (DMC)*) are used to monitor accumulating safety data. In addition, the DSMB may perform an interim analysis (see entry on *Interim Analysis*) of efficacy endpoints, if specified in the protocol. The membership and operation of the DSMBs are kept separate from the study coordinating center so as to minimize risks to study integrity.^[4]

REGULATORY REQUIREMENTS

There are no specific regulatory requirements beyond those noted for clinical trials.^[5]

STATISTICAL DESIGN AND ANALYSIS

Statistical Design

One drawback to the use of multicenter trials is the presence of additional sources of variability which are not present in single-center trials. In particular, a major source of variability is the clinical center. Clinical centers often vary on factors that may be related to either the therapy or clinical endpoint: characteristics of the patients enrolled at the center, specific interests/skills of the center and physician (e.g., family practice vs. referral center), and implementation of the protocol (e.g., how the physician evaluates the patient).^[6] To minimize this variability, the design of multicenter trials usually considers the clinical center as a block. Prospective stratification by clinical center with randomization is generally preferred to achieve balance among therapy groups. However, there are occasions where stratification is not possible or is not necessary. For example, in mortality or cancer trials (see entry on *Cancer Trials*), there may be many centers, each with very few patients, precluding stratification by center. Also, other factors may be felt to be more predictive of the clinical outcome than clinical center. In these cases, randomization may be performed using a stratification scheme administered by a central system (e.g., interactive voice randomization system—see entry on *Interactive Voice Randomization System (IVRS)*) that stratifies on other important factors besides center.^[7]

One should consider the number of centers and the number of patients allocated per center in the design. The guidance from the International Conference on Harmonization (see entry on *International Conference on Harmonization (ICH)*) on Statistical Principles for Clinical Trials^[8] recommends that one should try to avoid excessive variation in the number of patients per center and to avoid centers with very few patients. Such variation, if present, complicates the interpretation of the therapy-by-center interaction (to be discussed in the section “Statistical Analysis”).

The sample size determination (see entry on *Sample Size Determination*) should be specified in the protocol with enough details so that an independent reviewer (e.g., regulatory reviewer) can verify the calculations. The sample size estimated should take into account the design of the trial (e.g., comparative vs. equivalence, parallel vs. crossover), anticipated noncompliance, possibly varying hazards, and the multicenter nature of the study (i.e., the variability may be larger than that seen in single-center trials).^[9] Centralized readings of objective safety and efficacy endpoints may allow a reduction in variability. Sample size and power calculations for the multicenter trial are based on the assumption that the therapy differences from each center are estimating the same quantity. Often, clinical trials are not designed to consider the presence of therapy-by-center interaction,^[10] leaving it to the statistical analysis to explore the robustness of the results to this assumption.

Statistical Analysis

The statistical model for the analysis should be specified in the protocol. If the trial has been stratified on center, this effect should be included in the model to account for variability associated with center as well as to ensure that the appropriate error term is used.^[11] It is generally accepted in the pharmaceutical industry that center is a fixed effect (V.A. Uthoff, personal communication, 1986). This assumption limits the generalizability of the trial results to the centers employed in the trial. Alternatively, the center effect could be considered a random effect so that inference can be more generalizable. However, random-effects models often require greater sample sizes because of the reduced degrees of freedom in the error term. In addition, there is concern that centers do not represent a random sampling of all investigative centers but, rather, are selected on the basis of past performance and the current set of entry criteria.

The protocol should state a priori how therapy-by-center interaction is to be evaluated. In particular, the protocol should state the level of significance to use, whether interaction will be evaluated for all safety and efficacy measures, and time points, how to pool centers with few patients, and what will be carried out if interaction occurs.

If therapy-by-center interaction is not significant, the protocol should specify how interference is to be carried out; e.g., will the full or the reduced model be used? The reduced model is a parsimonious model having greater power for the detection of therapy effects, whereas the full model provides unbiased estimates (i.e., one can never be sure that the interactions really zero). In the latter case, there is controversy whether sequential or partial sums of squares should be used for inference.^[12] The choice ultimately depends on the hypothesis to be tested.

If the therapy-by-center interaction is significant, the interpretation of the main effect is controversial. As already mentioned, a critical assumption for multicenter trials is that the therapy effect is consistent from center to center. When this is not valid, the nature of the interaction should be explored in an attempt to characterize it. Generally, there are two types of interaction: quantitative and qualitative.^[13] *Quantitative* interaction refers to therapy differences that are in the same direction but vary in magnitude across the centers; *qualitative* interaction refers to therapy differences that vary in direction across the centers. Graphical presentation of the therapy effect for each center, as well as identification of the centers having the greatest statistical contribution to the interaction,^[14] is useful to gain insight into the cause of the interaction.

SCIENTIFIC ISSUES

One rationale for the use of multicenter trials is increased heterogeneity in patients and centers, enhancing the generalizability of the trial conclusions. However, there is a

cost associated with multicenter trials. Greater heterogeneity can imply greater variability, leading to compromised power and larger sample sizes for the detection of therapy effects. This increased variability may be reduced through the use of a central reading of objective endpoints.

It is often difficult to define *center*. Is it the region, which may have several investigative centers, is it the hospital/practice, which may have several investigators, or is it the investigator? The driving principle is to define *center* in such a way that one achieves homogeneity in the important factors that affect the outcome variable and therapy.

RECENT DEVELOPMENTS

As mentioned in the statistical design section, there are vague notions concerning the optimal number of centers and number of patients per center as well as concerning the effect of differential sample sizes among centers on the analysis. Shao and Chow^[15] design a multicenter trial using a two-stage sampling plan normally used for United States Pharmacopeia (USP) tests (see entry on *USP Tests*). They show that this plan is optimal in minimizing the variance of the associated statistical testing procedures. Using this sampling plan, one can select the number of centers and number of patients per center to maintain the optimality criterion. Salsburg (1991, personal communication) addresses the issue of minimum sample size per center from the standpoint of the power of the tests for interaction in a two-way fixed-effects analysis of variance (ANOVA). He concludes that, when there is no therapy-by-center interaction or when therapy allocations are balanced within center, differential sample sizes among centers have no effect on alpha level power for hypothesis tests.

A recent work^[16] considers an alternative to sequential or partial sums of squares for inference. Ciminera et al. propose weighting of the therapy effect from each center, depending on the variability observed within each center. Centers are grouped by the observed variability, and more variable centers are given less weight, and vice versa. The authors show that this procedure provides a reasonable estimate of the therapy effect in the presence of heteroskedasticity and “modest” therapy-by-center interaction.

CONCLUSION

Nonparametric statistical tests (specifically, extended Mantel–Haenszel tests) for analyzing data from multicenter trials have proliferated in recent years. Simulation studies have shown that these tests have excellent performance, even under optimal conditions for normal-theory tests.^[17]

In an attempt to bring speed and uniformity to the drug decision-making process, Enas et al.^[18] present a collection of standard statistical and graphical techniques for analyzing multicenter clinical trials. These techniques, easily

automated, facilitate better understanding of the efficacy findings. In addition, strategies evaluating the robustness of inference to varying assumptions are also presented. These strategies are consistent with the use of sensitivity analysis to evaluate robustness as promulgated by Jones.^[19]

The FDA Guidelines for the Clinical and Statistical Sections of New Drug Applications (NDA)^[20] state that the approval of a new therapeutic agent should be “supported by more than one well-controlled trial. ... This interpretation is consistent with the general scientific demand for replicability.” However, when a multicenter trial is viewed as a collection of single-center trials using a common protocol, replication of the therapy effect within the multicenter trial is feasible. Huster and Louv^[21] present a min-max statistic that assesses whether the therapy effect has been replicated in a multicenter trial. They describe the operating characteristics of this statistic under the null and several alternative hypotheses. They also observe that, as described by Salsburg earlier, the power for the detection of a therapy effect is not affected by different sample sizes.

REFERENCES

1. Hill, B. *Statistical Methods in Clinical and Preventive Medicine*; Oxford University Press: New York, 1962.
2. Herson, J. *Stat. Med.* **1993**, *12*, 555–564.
3. Bryant, J.; Cronin, W.; Wieand, S. *Encyclopedia of Biostatistics: Multicenter Trials*; Armitage, P., Colton, T., Ed.; Wiley: New York, 1997, 2710–2716.
4. Huster, W.J.; Shah, A.S.; Dere, W.; Kaiser, G.V.; DiMarchi, R.D. Harmonization of Requirements: Case Study of EVISTA. In *Drug Information Association Meetings, Boston, Massachusetts*; 1998.
5. *International Conference on Harmonization Harmonized Tripartite Guideline*; Guidelines for Good Clinical Practice (E6), 1996.
6. Fleiss, J.L. *The Design and Analysis of Clinical Experiments: Multicenter Trials*; Wiley: New York, 1986, 176–180.
7. Lachin, J.M. *Control. Clin. Trials.* **1981**, *2*, 93–113.
8. *International Conference on Harmonization Harmonized Tripartite Guideline*; Statistical Principles for Clinical Trials (E9), 1998.
9. Halperin, M.; DeMets, D.; Ware, J. *Stat. Med.* **1990**, *9*, 881–892.
10. Chinchilli, V.M.; Bortey, E.B. *J. Biopharm. Stat.* **1991**, *1*, 67–80.
11. Fleiss, J.L. *Control. Clin. Trials.* **1986**, *7*, 267–275.
12. Speed, F.M.; Hocking, R.R.; Hackney, O.D. *JASA.* **1978**, *73*, 105–112.
13. Byar, D.P. *Stat. Med.* **1985**, *4*, 255–263.
14. Hwang, D.S.; Barry, E.; Hamot, H. Treatment-by-Center Interaction: Resolving the Question by Presenting Evidence. In *Proceedings of Biopharmaceutical Section for the ASA Meetings in Toronto, Canada*; 1983, 40–44.
15. Shao, J.; Chow, S.-C. *Stat. Med.* **1993**, *12*, 1999–2008.
16. Ciminera, J.L.; Heyse, J.F.; Nguyen, H.H.; Tukey, J.W. *Stat. Med.* **1993**, *12*, 1033–1045.
17. Davis, C.S.; Chung, Y. *Biometrics.* **1995**, *51*, 1163–1174.
18. Enas, G.G.; Sanger, T.M.; Huster, W.J. *Biopharm. Rep.* **1993**, *2*, 1–12.
19. Jones, D.R. *Drug Inf. J.* **1993**, *27*, 833–836.
20. *Food and Drug Administration Guideline for the Format and Content of the Clinical and Statistical Sections of New Drug Applications*; 1988.
21. Huster, W.J.; Louv, W.C. *J. Biopharm. Stat.* **1992**, *2*, 219–238.

Multicollinearity

K. Van Steen

*Montefiore Institute – Systems and Modeling Unit; Bioinformatics and Modeling,
Université de Liège, Liège; Department of Human Genetics, Catholic University of Leuven,
Leuven, Belgium*

G. Molengerghs

*Leuven Biostatistics and Statistical Bioinformatics Centre, Department of Public Health,
Leuven, Belgium*

Abstract

This entry summarizes some of the important issues in detecting multicollinearity problems, assessing its severity and locating causal determinants.

Keywords: Diagnostics; High-throughput data; Many predictor variables; Multicollinearity; Remedial measures.

INTRODUCTION

The high throughput of data arising from microarrays, or the complete sequence of the human genome has left scientists with a rich and extensive information source. Moreover, the wide availability of software and the increase in computer power have improved the possibility to access and process, not only genomic data, but also data from many other fields. Therefore it is not surprising that an increasing number of studies involve the simultaneous evaluation of a large pool of predictors (prognostic or risk factors). It is well known that with multiple factors being considered, analyses may be complicated by a multitude of problems. These include investigating many interactions from the same data set and an elevated possibility of finding associations because of chance alone, or “multicollinearity,” among others. In this paper, we summarize some of the important issues in detecting multicollinearity problems, assessing its severity and locating causal determinants.

TWO DISTINCT SETTINGS WHERE MULTICOLLINEARITY MATTERS

Prognostic Factor Analysis Using Quality of Life (QOL) Measures

When evaluating the efficacy of medical treatment on cancer, prolongation of life expectancy and tumor shrinkage have traditionally been taken as outcome measures. Despite the substantial side effects and functional impairment often associated with cancer treatment, it is only recently that attention has been given to the assessment of QOL. Since

the advent of methods for measuring the QOL of patients with cancer, several studies have been published in which QOL variables derived from visual analogue scales and patient-completed questionnaires have been identified as important prognostic factors, in addition to clinical factors.^[1] This finding has considerable importance, particularly in advanced disease where treatment is generally palliative and the aim is to optimize QOL. However, our knowledge of the structure of QOL questionnaires alerted us to difficulties leading to incorrect model selection or ambiguities in determining the direction or magnitude of effects of the predictor variable on the response variable, even with a correct model specification (e.g., Refs. [1] and [2].) It is self-evident that these issues need to be addressed before drawing conclusions from a prognostic factor analysis.

Prognostic factor analysis is widely used in oncology to identify variables that are “independent” predictors of survival and response to treatment. Prognostic factors are used to stratify patients in the design and analysis of clinical trials, to assist in the interpretation of the data generated by such trials, and to aid in the clinical management of individual patients. In the last decade, many methods for measuring QOL have been developed, but in the clinical trial setting, multidimensional patient-completed questionnaires are most widely used. Of these, the EORTC Quality of Life Questionnaire QLQ-C30 is one of the best-known and validated questionnaires, and is available in many languages.^[2] It is a 30-item questionnaire that consists of subscales measuring aspects of function (physical, role, cognitive, emotional, and social), symptoms (pain, dyspnea, nausea/vomiting, loss of appetite, constipation, diarrhea, insomnia, and fatigue), the financial impact of illness and treatment, and overall health and QOL. The last two items

(overall health and QOL) are combined to form the global QOL subscale. This global measure is frequently reported in prognostic factor analyses because of its apparent simplicity and utility as an overall indicator or summary of the patients' QOL. However, it has been known since the first validation study of the QLQ-C30 that the subscales are highly correlated.^[2] This is confirmed by the ordinal measures for correlatedness listed in Table 1.^[3] These findings are to be expected because the questionnaire was specifically designed to measure various components of "QOL." Unfortunately, the impact of correlation between subscales on the outcome of prognostic factor analyses incorporating QOL variables is not always acknowledged in analyses.

Genetic Association Studies and Epistasis

The Human Genome Project, having generated a complete sequence of the human genome in April 2003^[4] and its spinoffs such as the International SNP Map Working Group (Ref. [5]; single nucleotide polymorphism, or SNP), the Allele Frequency/Genotype Project, determining the frequency of 60,000 SNPs in three major world populations, the Genomes to Life Program,^[6] or the Copy Number Variation (CNV) Project,^[7-9] are making it increasingly possible to disentangle the genetic basis of a given disease using genome-wide association studies. The contribution of Copy Number Variants (CNVs) to common complex diseases is largely unknown, but it is suspected that their role in disease etiology may be considerable. The CNV-expression association is complex, in that gene expression levels do not necessarily increase with copy number. CNVs

may be in linkage disequilibrium (LD) with different regulatory variants in different populations or may interact with other CNVs or SNPs.^[10] SNPs are the most common type of genetic variation in humans. They have a dense distribution across the genome (on average about every 1000 base pairs) and a low mutation rate, which makes them ideal markers in genetic association studies. Application of these markers in association analysis is based on the principle that tight linkage can generate allelic association at the population level, provided that most of the disease-bearing chromosomes in the population are descended from one or a few ancestral chromosomes. A popular design, whether SNPs or CNVs are involved as genetic markers, uses cases and controls.^[11]

In a classic logistic modeling framework (case-control status is taken as outcome variable), the search for functional variants can be carried out by constructing a model for the probability of disease (e.g., including only main effects at all variants under investigation). Quantifying the effects of a single locus is achieved by interpreting the corresponding regression coefficients, conditional on the fixed status at the remaining loci. However, if the single locus is involved in a complex multicollinearity pattern with other loci included in the model, it is questionable how much value can be placed on this interpretation. In addition, an increasing number of examples become available in which it is shown that accounting for interaction can lead to improved power to locate disease genes^[12]. In addition, the main effect of a locus may be virtually undetectable because the partner locus has an opposite or reciprocal effect.^[13] These settings point toward the necessity

Table 1 Multicollinearity Tests Based on Chi-Square (Overall Test), *F* (Single Variable), or *t* (Pairs of Variables) Statistics, as Explained in Theil

Test Overall test	Value 606.359 Variable names											
	DY	PF	RF	EF	CF	SF	QL	FA	NV	PA	SL	AP
Single variable	5.801	13.713	13.735	4.393	4.276	10.969	16.830	15.051	7.087	6.094	2.429	7.731
Pair of variables	0	3.020	1.847	2.237	2.218	2.133	2.181	2.960	2.492	0.359	-0.725	2.579
		0	4.248	1.031	1.471	2.383	3.536	3.540	2.262	2.793	0.337	2.512
			0	1.366	1.914	3.391	3.634	3.382	2.362	3.473	1.239	2.265
				0	3.242	2.546	1.620	2.231	0.333	1.485	2.412	1.680
					0	2.790	1.876	2.360	0.812	2.281	-0.170	2.771
						0	3.986	3.391	2.513	3.249	1.210	2.770
							0	3.908	2.101	3.186	1.685	2.373
								0	3.048	2.905	1.170	3.230
									0	1.933	1.115	4.326
										0	0.559	1.907
											0	0.180
												0

No *p* values are reported because variables are dichotomous. Therefore the multivariate normal assumption is violated and it makes little sense to rely on distributional properties of the test statistics.

DY = dyspnea, PF = physical functioning, RF = role functioning, EF = emotional functioning, CF = cognitive functioning, SF = social functioning, QL = global quality of life, FA = fatigue, NV = nausea/vomiting, PA = pain, SL = insomnia, and AP = appetite loss.

(From Ref. 16).

to accommodate interactions, hereby drastically increasing the number of parameters in the model. Although Clayton and Jones^[14] argue that higher-order associations fall away at progressively more rapid rates than first-order associations, even the inclusion of only main effects and all two-way interactions might introduce a severe problem of multicollinearity.

Multicollinearity: What is in a Name?

There does not seem to exist a precise operating definition of “multicollinearity.”^[15] Far too often, the term is used for “harmful” multicollinearity and is described in terms of symptoms it might bring about. The symptomatic descriptive view on multicollinearity of Kotz and Johnson^[16] is only one out of many such “definitions” that have passed in review during history. It refers to multicollinearity as “the situation which arises when some of all of the explanatory variables are so highly correlated one with another that it becomes very difficult, if not impossible, to disentangle their influences and obtain a reasonably precise estimate of their separate effects.” Clearly, symptomatic definitions like these have some flaws. In general, they do not offer a full coverage of potentially harmful effects, nor do they give a general setting where problems of multicollinearity can occur. For instance, in the description above, high correlations are sufficient but not necessary for multicollinearity problems to arise.

The underlying mathematical definition for “multicollinearity” is quite simple though: k predictors are collinear if the vectors that represent them lie in a subspace of dimension less than k . In this definition, the sample or observation space is used as a data representation platform. Here, every point in space refers to a predictor or variate. The n axes refer to the items or individuals in the study on which information was collected. This alternative way of representing data is quite convenient to give geometric interpretations to sampled data in terms of lengths, projections, and angles.^[17] It differs from the more classical covariate space, in which every point in space indicates an item or individual and the p axes refer to the measured variates.

The mathematical definition of multicollinearity (“perfect” multicollinearity) rarely applies to practical settings. The linear dependency requirement is not necessary for a collinearity problem to exist; therefore a broader notion is warranted, allowing for *imperfect* multicollinearity. Belsley et al.^[15] further specify: “collinearity exists if there is a high multiple correlation when one of the variates is regressed on the others.” Heuristically, we can say that one variable is very nearly a linear combination of other variables. Obviously, this interferes with a classical assumption of regression analysis that the columns of the design matrix $\mathbf{X}$ are linearly independent.

Despite the fact that nonlinear models and limited dependent variable models are likely to be affected by

multicollinearity as well, multicollinearity as an issue is generally discussed in terms of least squares linear regressions and matrix algebra.^[18] With $\mathbf{y}$, an n -dimensional vector of observations on a dependent variate of interest; $\mathbf{X}$, an $n \times p$ matrix of measurements on p predictors x_i , $i = 1, \dots, p$ ($\mathbf{X}$ is the design matrix and contains all dependent variables included in the analysis); $\boldsymbol{\beta}$, a $p \times 1$ vector of parameters; and e , an $n \times 1$ vector of error terms, the linear regression equation can be written as: $y = \mathbf{X}\boldsymbol{\beta} + e$, $e \sim N(0, \sigma^2 I)$. Using standardized (dependent and independent) variables throughout this paper for ease of exposition, we can write the estimated parameter estimates $\hat{\boldsymbol{\beta}}$ in terms of correlations among covariates $\mathbf{R}_{xx}$, and correlations between covariates and dependent variable $\mathbf{R}_{xy}$, in the following way (e.g., Ref. [19]):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y}) = \mathbf{R}_{xx}^{-1}\mathbf{R}_{xy} \quad (1)$$

$\mathbf{R}_{xx} = (r_{ij})_{1 \leq i, j \leq p}$, r_{ij} is the correlation between the variates x_i and x_j ; $\mathbf{R}_{xy} = (r_{ij})_{1 \leq i \leq p}$, r_{yi} is the correlation between y and x_i . Note that in the observation space, the orthogonal projection of $\mathbf{y}$ onto the (hyper)plane determined by the vectors that represent the covariates $x_1, \dots, x_p$ corresponds with $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, that the coefficient of determination R^2 (proportion of the total variation in the dependent variable y that is explained or accounted for by the variation in the independent variable x_i) is the cosine between $\mathbf{y}$ and $\hat{\mathbf{y}}$, and that correlation r_{ij} is determined by the cosine between the vectors $\mathbf{x}_i$ and $\mathbf{x}_j$.

The variance of $\hat{\boldsymbol{\beta}}$ can be estimated via:^[19]

$$\hat{\sigma}^2 \left(\frac{1}{n} \mathbf{R}_{xx}^{-1} \right) = \left(\frac{n}{n-p} \right) \frac{|\mathbf{R}_{yx}|}{|\mathbf{R}_{xx}|} \left(\frac{1}{n} \mathbf{R}_{xx}^{-1} \right) \quad (2)$$

involving the $(p+1) \times (p+1)$ matrix $\mathbf{R}_{yx}$ of both correlations among covariates and correlations between covariates and dependent variable. Expression 1 clearly shows that when linear dependencies occur among covariates, $\mathbf{X}'\mathbf{X}$ will have a zero determinant, and the inverse will not exist; therefore parameter estimates will be undefined. Moreover, the relationship between covariates and response may modify the effects of multicollinearity in terms of precision of the regression parameter estimates. This is indicated by the occurrence of $\mathbf{R}_{yx}$ in Expression 2, and further discussed in Ref. [18]. These authors show that under certain conditions, multicollinearity may deflate parameter variance estimates, rather than inflate them (as usually mentioned by many authors, e.g., Ref. [20] or Ref. [21]).

TWO FACES OF THE SAME COIN

Dohoo et al.^[22] mention two sources of multicollinearity: sample-based and structural-based. The first source simply refers to all settings in which collected measurements on predictor variables appear to be correlated with a potential

for harmful multicollinearity (e.g., because they truly are, or because the same variables expressed in different units are used, etc.). The second source is because of ad hoc (after the data are collected) introduction of collinearity in the analysis (e.g., by including power terms in polynomial regression). Glantz and Slinker^[23] suggest to mean-center covariates before taking powers because this will usually be very effective in destroying strong relationships between the variates and their powers.

Suppose two variables x_1 and x_2 are measures of the same concept, but expressed in different units. In this extreme situation, the variables are maximally correlated ($r_{ij} = 1$) and essentially convey the same information. Many planes can be drawn through the line determined by the corresponding vectors x_1 (x_2), each of which yields the same residual vector $e (= y - \hat{y})$ and hence the same fit. However, many parameters exist to describe $\hat{y}$ as a linear combination of x_1 and x_2 . Suppose a third variable is added, say x_3 , geometrically orthogonal to x_1 but in the plane determined by y and x_1 . Then the overall model, including all three variables x_1 , x_2 , x_3 , fits the data perfectly. Neither x_1 nor x_2 will contribute significantly to the model after the other one is included. The overall fit is not affected by the multicollinearity. Removing both variables from the model would result in a much worse fit (coefficient of determination < 1).

Another extreme situation of multicollinearity arises in the presence of a suppressor variable. Such a variable is highly correlated with another independent variable but uncorrelated with the dependent variable.^[16] It may substantially increase the multiple correlation R^2 when combined with a variable that is only moderately correlated with the dependent variable. Suppressor variables may even lead to artificial interactions, as was shown by Sithisarankul et al.^[24]

HARMFUL MULTICOLLINEARITY

Despite the fact that multicollinearity is more a data problem than a statistical problem, it can cause severe problems in statistical modelling or model diagnostics. Multicollinearity “does not hurt as long as it does not bite.”^[15]

Slinker and Glantz^[25] specify that it “bites” when the analyses yield nonsensical results (in terms of clinical interpretation), parameter estimates of incorrect magnitude or incorrect sign (according to prior expectations), and huge standard errors leading to small observed test statistics, and accept too many null hypotheses. These symptoms are often combined as indicators for “harmful multicollinearity.”

When model selection is emphasized, multicollinearity is not much of a problem because predictions will still be accurate, and the overall R^2 can be seen as a quantification of how well the model predicts the y values. When the goal is to understand how the x variables influence y ,

then multicollinearity should be taken seriously. Because the confidence intervals may become extremely wide, excluding a subject (or adding a new one) may lead to dramatic changes in the estimated regression parameters or may even change their signs. In prognostic factor analysis, both issues are important. Here, the first step is to find a limited number of prognostic factor with a high impact on the response. The second step is to utilize these factors in categorizing patients to certain disease groups, as accurate as possible.

When the model is correctly specified, parameters are unbiased, even in the presence of multicollinearity. However, when model fit is based on a sample of the population at large, collinearity may result in inflated^[26] or deflated^[20] variance estimates. In reality, chances are high to select the wrong (i.e., incomplete) model. One believes that the model has the “correct” structure but parameter estimates may lead to erroneous interpretations about direction or magnitude of predictor variable effect on the response of interest. On a population basis, selecting the wrong model will lead to biased parameter estimates.^[21] On sample data, not only variable omission bias, but also problems with respect to proper variance estimation are to be expected, as Mela and Kopalle^[18] point out. Therefore it is important to properly detect and deal with multicollinearity.

The developed diagnostic tools in “Diagnostic Tool Kit” can be seen as ways to determine unacceptable multicollinearity and “harmful” multicollinearity, or a means to distinguish “acceptable” departures from orthogonality. In this setting, the aim is to detect evidence against the null hypothesis of orthogonality between independent variables. Note that orthogonal variables, being variables with zero covariance, are represented in the observation space as geometrically orthogonal vectors.

DIAGNOSTIC TOOL KIT

Unfortunately, there is no exclusive test that can tell you whether multicollinearity is or is not a problem. In the following, we briefly comment on some commonly used and less commonly used warning signals.

DETERMINANT

The absence of orthogonality immediately translates into $X'X > 0$. Transforming $\#X'X\#$ into an approximate chi-square statistic provides a statistical quantification for the interdependency between predictors.^[27] When x_i is orthogonal to the remaining set of predictors, the diagonal element xx_i of $(X'X)^{-1}$ satisfies $xx_i = [((X'X)^{-1})_{ii}]/(X'X) = 1$. The larger is the departure from orthogonality, the closer xx_i will become to infinity. The latter can be used as a means to locate a “troublesome” covariate. Alternatively, $[(n-p)/(p-1)](xx_i - 1)$, F -distributed with $n-p$ and $p-1$

degrees of freedom (under normality assumption of the dependent variables) can be taken to localize sources of multicollinearity.^[28]

CORRELATIONS

Klein^[29] indicates that a pairwise correlation r_{ij} bigger than the coefficient of determination R^2 is indicative of harmful multicollinearity. Extended to more than two predictors, Farrar and Glauber^[3] specify that a variable is harmfully multicollinear if its multiple correlation with other members of the independent set is greater than the dependent variables' multiple correlation with the entire predictor set.

However, no matter the strength of the relationship between dependent and independent variables, harmful multicollinearity is, in essence, an artifact of the independent data, via characteristics of the design matrix $\mathbf{X}$ (or $\mathbf{X}'\mathbf{X}$). Because large correlations between standardized variables will lead to large covariances and hence large departures from orthogonality, the correlation matrix is the first obvious object to look at in detail. In biological and physicochemical studies, intercorrelations of 0.90 or higher in absolute values are often considered as unacceptable.^[15,24,25,30] Some authors indicate pairwise absolute correlations of 0.70 or larger as signals for harmful multicollinearity.^[15]

Three important remarks are in place here. First, as mentioned before, "high" correlations are sufficient but not necessary for multicollinearity problems to arise. Second, the correlation matrix only allows for looking at pairwise relations between variates. Multicollinearities among three or more predictor variables can exist even though pairwise correlations are small. Third, relationships between dependent and independent variables can modify the "effects" of multicollinearity.

VARIANCE INFLATION FACTORS(VIFS)

The severity of multicollinearity may be suggested by the magnitude of the VIFs for the regression of each predictor on all remaining predictor variables. The VIF can be interpreted as the ratio of a parameter estimate's variance to what it would be in the absence of multicollinearity. It measures the inflation in the variance of the parameter estimate because of collinearity between the explanatory variable and other variables. In a three-variable regression setting, the VIF for β_1 is given by $VIF(\hat{\beta}_1) = 1/(1 - R_1^2)$, with R_1^2 , the unadjusted R^2 , regressing from x_1 onto x_2 and x_3 . A VIF larger than 10 is generally considered indicative of harmful multicollinearity (e.g., Ref. [31]), but this value is arbitrary, and relatively small VIFs may still point to instability. Some authors (e.g., Ref. [25]) accept a VIF > 4 (corresponding to an auxiliary regression R_1^2 of 0.75).

Note that the covariance matrix of the estimated regression parameters is $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$. When the variables are centered and scaled, $\sigma^2 = 1$ and the tolerance for the j th predictor is simply the j th diagonal element $xx_j = (1 - R_j^2)$ of $(\mathbf{X}'\mathbf{X})^{-1}$. Hence, when R_j^2 is close to one, the variable x_j does not add any further information to the remaining variables, the tolerance is 0, and the VIF for the j th parameter is infinitely large.

CONDITION NUMBERS

Alternatively, the structure of $\mathbf{X}'\mathbf{X}$, in particular the eigenvalues, can be examined. The literature suggests that eigenvalues smaller than 0.01 indicate a serious problem.^[25] The square root of the ratio of the largest to each individual eigenvalue is known as the condition index, and allows one to assess the relative magnitudes of the eigenvalues. The largest condition index (square root of the ratio of the largest to the smallest eigenvalue) is the condition number. A large such number is seen as evidence toward harmful multicollinearity. However, criteria for a condition number to signify serious multicollinearity are arbitrary, with values of 30–100 often quoted.^[31] But sometimes numbers as "small" as 2,015 or 1,025 are suggested as a cutoff.

BOOTSTRAP RESAMPLING

Bootstrap resampling is a less commonly used, but promising, diagnostic tool in the present context.^[32] Its usefulness is motivated on several levels. First, previously described diagnostic tools are derived from a linear regression model with normal error terms. However, regression is not the only technique that may be affected by harmful multicollinearity. Second, often, predictor variables are not continuous and are entered into the regression model in dichotomized form. This is particularly true in prognostic factor analyses. Third, the potential in variability (decrease in stability) for multicollinear regression coefficients impacts automated variable selection procedure. Backward selection may discard from the model what might be important variables. Forward selection may "fail" when neither of two correlated factors is identified as statistically significant, although both have an influence on the outcome.^[33] Therefore in the light of potential harmful multicollinearity, it is important to examine the stability of the chosen regression model. Stability may be defined by the replication stability for the choice of variables included in the model and the predictive ability of the model itself, as was suggested by Sauerbrei and Schumacher^[34] and further discussed in Schumacher et al.^[35]

This technique generates a number of samples (each the same size as the original data set) by randomly selecting

patients and replacing them before selecting the next patient (i.e., bootstrap resampling). The frequency of inclusion of the component variables in the regression models, fitted to each of these data sets (e.g., using automatic backward selection), can be considered to be indicative of the importance of the factors. Alternatively, these inclusion frequencies can be seen as estimates of the posterior probabilities that the regression coefficients are different from zero.^[34]

ACCOUNTING FOR MULTICOLLINEARITY OR AVOIDING THE PROBLEM?

A drastic way of avoiding the harmful effects of multicollinearity is to delete offending predictor variables from the model, based on some variables being redundant. However, despite overlap, variables may be explicitly designed to measure different aspects of a feature. So, it is not at all straightforward to argue for redundancy in such a situation.

Instead of deleting predictor variables, a weighted sum of the variables causing multicollinearity can be used. Substituting mathematically transformed variables for harmful ones or rescaling them may mitigate harmful multicollinearity.^[25] Nevertheless, such procedures may be hard to justify in the presence of nonnumerical data.

Alternatively, the number of explanatory variables is reduced by relying on multivariate data techniques and artificial orthogonalization of the covariate data (e.g., factor analysis or principal component analysis).^[31,36,37] By using principal components as new predictors, the initial intercorrelations between predictors are removed, yet a maximal amount of “information” in the data is retained. Although mathematically convenient, the question is how many components to retain or how to interpret the new constructs. In the majority of the cases, interpretation is not straightforward, not to say impossible. The same reasoning applies for factor analysis, in which the original variables are assumed to be a linear combination of factors plus an error term. Farrar and Glauber^[3] illustrate how backtransforming q factors (the “new” predictors) to p original variables (the “old” predictors), $q < p$, might actually compound multicollinearity.

Another route is to explicitly account for the presence of multicollinearity in the statistical technique chosen, such as Bayesian analysis or constrained regression.^[20] In a linear regression setting, multicollinearity can be accounted for by performing ridge regression, first proposed by Hoerl and Kennard.^[38] It is based on adding a (biasing) constant to the (nearly) singular matrix $\mathbf{X}'\mathbf{X}$. Although it leads to biased parameter estimates, a sensible choice of biasing constant may lead to a substantial reduction in the variance of the parameter estimates. Apart from being a remedial technique, it can also serve as a diagnostic tool for detecting predictors that elevate a multicollinearity problem. Sudden changes in the ridge estimates as the biasing

constant approaches zero can be seen as warning signals. Note that ridge estimators can also be used in logistic regression to improve the parameter estimates and to diminish error made by further predictions.^[39]

In “Alternative Ways to Go About Multicollinearity Issues,” we reconsider the examples introduced before, and offer some alternative ways to address multicollinearity.

ALTERNATIVE WAYS TO GO ABOUT MULTICOLLINEARITY ISSUES

Prognostic Factor Analysis Using QOL Measures

In a prognostic factor analysis^[32] incorporating clinical and QOL variables from the QLQ-C30 in a clinical trial of patients with advanced breast cancer, we observed instability in the model predicting response to chemotherapy because forward selection and backward elimination produced substantially different final multivariate models. This led us to suspect that harmful multicollinearity between QOL subscales was influencing model selection. Therefore we undertook a thorough examination of the QOL data used in the prognostic factor analysis to diagnose the presence and degree of multicollinearity, and to determine the stability of the predictive models obtained. In particular, replication stability of the final prognostic factor model was illustrated via a bootstrap resampling procedure based on works of Sauerbrei and Schumacher^[34] and Schumacher et al.^[35] and inclusion frequencies of predictor variables (based on 1000 simulated data sets) were reported (refer to Ref. [32]).

However, when we are interested in the “best set” of prognostic factors (which we usually are) rather than in the “best independent” prognostic factor, we need to account for the correlation structure of the potential prognostic factors under consideration. Therefore we calculated the model selection probabilities based on how many times a permissible model was selected in the bootstrap samples. The results are listed in Table 2. The relatively low frequency of each model (only the top 10 are shown) indicates that there is a serious inherent instability, without any model or group of models convincingly standing out. The top model seems to be indicative of dyspnea, with emotional functioning and fatigue being potentially important prognostic factors. However, caution is in place. The interpretation of the observed prognostic factors is not clearcut, even though the same three factors show up as good candidates when variable inclusion frequencies are used prior to bootstrap resampling (step 2; Table 2). Rather than believe in a single model, it is much more appropriate to use the model selection probabilities as weights, so as to obtain weighted averaged parameters,^[40] hereby incorporating effects of multicollinearity on model uncertainty.

Table 2 Results of the Bootstrap Model Selection Approach

Prognostic factors		10 models with the highest $h(M_j)$ % out of 233									
Step 1											
DY	79.20	*	*	*	*	*	*	*	*	—	—
PF	10.90	—	—	—	—	—	—	—	—	—	—
RF	7.20	—	—	—	—	—	—	—	—	—	—
EF	74.00	*	—	*	*	*	*	*	—	—	*
CF	10.00	—	—	—	—	—	—	—	—	—	—
SF	12.90	—	—	—	—	—	—	—	*	—	—
QL	14.80	—	—	—	*	—	—	—	—	—	—
FA	48.90	*	—	—	—	*	*	—	*	*	*
NV	31.40	—	*	*	—	*	—	—	—	—	—
PA	5.90	—	—	—	—	—	—	—	—	—	—
SL	8.30	—	—	—	—	—	—	—	—	—	—
AP	17.60	—	—	—	—	—	*	*	—	—	—
$h(M_j)$ %	$h(x_i) > 0.0\%$	13.4	5.4	5.0	3.1	2.9	2.8	2.5	2.4	2.2	2.2
Step 2											
$h(M_j)$ %	$h(x_i) > 0.1\%$	18.7	5.2	6.7	3.7	4.6	3.5	4.8	0.1	2.1	1.5
$h(M_j)$ %	$h(x_i) > 0.2\%$	39.1	8.3	11.7	0.0	8.6	0.0	0.0	0.0	3.7	3.2
$h(M_j)$ %	$h(x_i) > 0.3\%$	39.1	8.3	11.7	0.0	8.6	0.0	0.0	0.0	3.7	3.2
$h(M_j)$ %	$h(x_i) > 0.4\%$	59.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.4	5.9
$h(M_j)$ %	$h(x_i) > 0.5\%$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

The inclusion frequencies ($h(x_i)$) are based on the logistic regression models for response selected in 1000 bootstrap samples (automated forward selection), with treatment arm (ARM) included in the model and retaining the clinical variables disease-free interval (DFI) and dominant site of disease (DS). The corresponding model selection probabilities $h(M_j)$ resulting from different selection levels are also given. DY = dyspnea, PF = physical functioning, RF = role functioning, EF = emotional functioning, CF = cognitive functioning, SF = social functioning, QL = global quality of life, FA = fatigue, NV = nausea/vomiting, PA = pain, SL = insomnia, and AP = appetite loss.

GENETIC ASSOCIATION STUDIES AND EPISTASIS

In “Two Distinct Settings Where Multicollinearity Matters,” we indicated some of the potential drawbacks in terms of multicollinearity involved in using univariate logistic regression models for case-control genetic association studies. Alternatively, a multivariate approach can be adopted, in which genotype information is modeled conditional on disease status. It allows investigating the dependence of interactions on disease status and environmental factors. Missingness and error measurements in the covariates now become outcome problems; whereas the first can be integrated in the framework of missing data of Rubin,^[41] the second involves dealing with the residual errors. Although the problem of having many parameters to estimate remains, the problem of multicollinearity is alleviated. Van Steen et al.^[42] discuss merits of a particular multivariate model in this context, namely the multivariate Dale model. It extends the bivariate global odds ratio model described by Dale^[43] and McCullagh and Nelder.^[44] Joint probabilities are decomposed into main effects (described

by marginal probabilities) and interactions (described by odds ratios of second and higher orders).

Obviously, the aforementioned alternative ways to circumvent multicollinearity problems are of little use in the context of genomewide association studies (GWAs). The multi-dimensional character of GWAs, in particular when performing large-scale epistasis screening, and the gene-gene or protein-protein interaction networks that may impose complex dependencies between genetic markers, have paved the way for the revival of non-parametric pattern searching methods. Popular examples of such methods, which at the same time aim to reduce dimensionality with minimal information loss, include the non-parametric Multifactor Dimensionality Reduction method of Ritchie et al.^[45] and the semi-parametric Model-Based Multifactor Dimensionality Reduction approach of Calle et al.^[46]

DISCUSSION

There exists a vast literature on multicollinearity, which shows that the subject has occupied and still occupies

the minds of many scientists. This does not only relate to deriving an operable definition—it also relates to reconciling differing views on the severity of multicollinearity. In other words, is it really a problem? Is it harmful? Obviously, one can only talk about “harmful” multicollinearity when methods have been developed and used to detect it. This is not the end of the story though. Once detected, one of the key issues is to decide on ways to deal with the problem in further analyses—a matter of “avoiding” or “handling?”

Some people argue that the game is not worth the candle and that harmful multicollinearity, in terms of generating large standard errors of parameter estimates, is of the same order as having too small a sample size for modelling purposes.^[47] In a sense, this is true. However, it all depends on the purpose of your analysis. Whether it is in economics, clinical trials, or genetic epidemiology, the question is whether we are most interested in model selection or model testing. If the emphasis is on providing good predictions of the response variable, it does not matter whether or not selected factors are actually statistically significant—although, ideally, one hopes that the selected predictors (e.g., SNPs) in the model are true causal factors, or at least highly correlated to true causal variants.^[48] If the emphasis is on controlling the number of factors falsely included in the model, then multicollinearity issues do matter because, for example, large standard errors may lead to wide confidence intervals or small observed test statistics, and therefore reduce power to assess the significance of a factor.

In these cases, it is crucial to (i) identify, locate, and quantify harmful multicollinearity; and (ii) to overcome the problem or to use statistical tools that explicitly deal with multicollinearity issues.

The diagnostic tool kit at our disposal contains several signaling flags based on the matrix $X'X$ and its properties, or on measures of associations such as pairwise association r_{ij} or the coefficient of determination R^2 . In general, larger correlations, higher VIFs and condition indices, and lower determinants may all indicate an increased severity in harmful multicollinearity. The problem is how large should these quantities be for multicollinearity to be “unacceptable?” Cutoff values may differ from one field to the other and need to be “relaxed” because they are often derived under the assumption of multivariate normality. In addition, the majority of these diagnostics do not distinguish the direction of the correlations, nor do they include information about the strength of the relationship between explanatory variables and response. Mela and Kopalle^[18] show that the effects of multicollinearity are moderated by the relationship between the outcome variable and the predictor variables. Clearly, it can be very misleading to rely on a single diagnostic tool.

Because estimates of standard errors and parameters tend to be sensitive to changes in the data, an alternative route (complementing standard approaches) is to assess

model uncertainty or instability by adopting a bootstrap resampling approach. This technique generates a number of samples (each of the same size as the original data set) by randomly selecting observations and by replacing them before selecting the next observation (i.e., bootstrap resampling). Model averaging methodology naturally serves the purpose of explicitly incorporating model uncertainty in the analysis results.

CONCLUSION

One can always attempt to mitigate problems of multicollinearity. However, it is not always possible to overcome them. One such setting is a prognostic factor analysis, where most of the initial variables need to be entered into the model. Alternatively, strategies that explicitly account for harmful multicollinearity in the analysis are used. Here, the use of bootstrap models (to obtain greater insight into the stability of the models) and averaging procedures remain useful approaches, despite their computational intensive nature.

ACKNOWLEDGMENTS

This work was carried out within the framework of the Belgian IUAP/PAI network “Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data” and the Belgian IUAP/PAI network “BIOMAGNET,” and was partially supported by grant MH59532 of the National Institutes of Health. The authors would like to thank Fabio Efficace and Andrew Bottomley (Quality of Life Unit at the European Organization for Research and Treatment of Cancer, EORTC) for their invaluable help during the preparation of this document.

REFERENCES

1. Coates, A.; Porzsolt, F.; Osoba, D. Quality of life in oncology practice: Prognostic value of EORTC QLQ-C30 scores in patients with advanced malignancy. *Eur. J. Cancer* **1997**, *33* (7), 1025–1030.
2. Aaronson, N.K.; Ahmedzai, S.; Bergman, B.; Bullinger, M.; Cull, A.; Duez, N.J.; Filiberti, A.; Flechtner, H.; Fleishman, S.B.; de Haes, J.C.J.M.; Kaasa, S.; Klee, M.; Osoba, D.; Razavi, D.; Rofe, P.; Schraub, S.; Scneeuw, K.; Sullivan, M.; Takeda, F. The European Organization for Research and Treatment of Cancer QLQ-C30: A quality of life instrument for use in international clinical trials in oncology. *J. Natl. Cancer Inst.* **1993**, *85*, 365–376.
3. Farrar, D.E.; Glauber, R.R. Multicollinearity in regression analysis: The problem revisited. *Rev. Econ. Stat.* **1967**, *49*, 92–107.
4. Collins, F.S.; Morgan, M.; Patrinos, A. The Human Genome Project: Lessons from large-scale biology. *Science* **2003**, *300*, 286.

5. International SNP Map Working Group A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **2001**, *409*, 928–933.
6. Frazier, M.E.; Johnson, G.M.; Thomassen, D.G.; Oliver, C.E.; Patrinos, A. Realizing the potential of the genome revolution: The Genomes to Life Program. *Science* **2003**, *300*, 290.
7. Iafrate, A.J.; Feuk, L.; Rivera, M.N.; Listewnik, M.L.; Donahoe, P.K.; Qi, Y.; Scherer, S.W.; Lee, C. Detection of large-scale variation in the human genome. *Nat. Genet.* **2004**, *36*, 949–955.
8. Redon, R.; Ishikawa, S.; Fitch, K.R.; Feuk, L.; Perry, G.H.; Andrews, T.D.; Fiegler, H.; Shapero, M.H.; Carson, A.R.; Chen, W.; Cho, E.K.; Dallaire, S.; Freeman, J.L.; González, J.R.; Gratacòs, M.; Huang, J.; Kalaitzopoulos, D.; Komura, D.; Macdonald, J.R.; Marshall, C.R.; Mei, R.; Montgomery, L.; Nishimura, K.; Okamura, K.; Shen, F.; Somerville, M.J.; Tchinda, J.; Valsesia, A.; Woodwark, C.; Yang, F.; Zhang, J.; Zerjal, T.; Zhang, J.; Armengol, L.; Conrad, D.F.; Estivill, X.; Tyler-Smith, C.; Carter, N.P.; Aburatani, H.; Lee, C.; Jones, K.W.; Scherer, S.W.; Hurler, M.E. Global variation in copy number in the human genome. *Nature* **2006**, *444*, 444–454.
9. Marioni, J.C.; Thorne, N.P.; Valsesia, A.; Fitzgerald, T.; Redon, R.; Fiegler, H.; Andrews, T.D.; Stranger, B.E.; Lynch, A.G.; Dermitzakis, E.T.; Carter, N.P.; Tavaré, S.; Hurler, M.E. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol.* **2007**, *8*, R228.
10. Stranger, B.E.; Forrest, M.S.; Dunning, M.; Ingle, C.E.; Beazley, C.; Thorne, N.; Redon, R.; Bird, C.P.; de Grassi, A.; Lee, C.; Tyler-Smith, C.; Carter, N.; Scherer, S.W.; Tavaré, S.; Deloukas, P.; Hurler, M.E.; Dermitzakis, E.T. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **2007**, *315*, 848–853.
11. Terwilliger, J.D.; Weiss, K.M. Linkage disequilibrium mapping of complex disease: Fantasy or reality?. *Curr. Opin. Biotechnol.* **1998**, *9*, 578–594.
12. Culverhouse, R.; Suarez, B.K.; Lin, J.; Reich, T. A perspective on epistasis: limits of models displaying no main effect. *Am. J. Hum. Genet.* **2002**, *70*, 461–471.
13. Frankel, W.N.; Schork, N.J. Who's afraid of epistasis?. *Nat. Genet.* **1996**, *14*, 371–373.
14. Clayton, D.; Jones, H. Transmission/disequilibrium tests for extended marker haplotypes. *Am. J. Hum. Genet.* **1999**, *65*, 1161–1169.
15. Belsley, D.A.; Kuh, E.; Welsch, R.E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; Wiley: New York, 1980.
16. *Encyclopedia of Statistical Sciences*; John Wiley and Sons: New York, 1985.
17. Johnson, R.A.; Wichern, D.W. *Applied Multivariate Statistical Analysis*, 3rd Ed.; Prentice-Hall: New Jersey, 1992.
18. Mela, C.F.; Kopalle, P.K. The impact of collinearity on regression analysis: The asymmetric effect of negative and positive correlations. *Appl. Econ.* **2002**, *34*, 667–677.
19. Neter, J.; Kutner, M.H.; Nachtsheim, C.J.; Wasserman, W. *Applied Linear Regression Methods*, 3rd Ed.; Irwin: Chicago, IL, 1996.
20. Theil, H. *Principles of Econometrics*; John Wiley and Sons, Inc.: New York, 1971.
21. Johnston, J. *Econometric Methods*, 3rd Ed.; McGraw-Hill Publishing Company: New York, 1984.
22. Dohoo, I.R.; Ducrot, C.; Fourichon, C.; Donald, A.; Hurnik, D. An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies. *Prev. Vet. Med.* **1996**, *29*, 221–239.
23. Glantz, S.A.; Slinker, B.K. *Primer of Applied Regression and Analysis of Variance*; McGraw-Hill: New York, 1990.
24. Sithisarankul, P.; Weaver, V.M.; Diener-West, M.; Strickland, P.T. Multicollinearity may lead to artificial interaction: An example from a cross sectional study of biomarkers. *Southeast Asian J. Trop. Med. Public Health* **1997**, *28*, 404–409.
25. Slinker, B.K.; Glantz, S.A. Multiple regression for physiological data analysis: The problem of multicollinearity. *Am. J. Physiol.* **1985**, *249*, R1–R12.
26. Lehmann, D.R.; Gupta, S.; Steckel, J. *Marketing Research*; Addison-Wesley Educational Publishers, Inc.: Reading, MA, 1988.
27. Bartlett, M.S. Tests of significance in factor analysis. *Br. J. Psychol.* **1950**, *3*, 77–85.
28. Graybill, F. *An Introduction to Linear Statistical Models*; McGraw-Hill: New York, 1961.
29. Klein, L.R. *An Introduction to Econometrics*; Prentice-Hall: Upper Saddle River, NJ, 1962.
30. Mager, P.P.; Rothe, H. Obscure phenomena in statistical analysis of quantitative structure-activity relationships: Part 1. Multicollinearity of physicochemical descriptors. *Pharmazie* **1990**, *45*, 758–764.
31. Krzanowski, W.J. *Principles of Multivariate Analysis, A User's Perspective*; Oxford Statistical Science Series 3: New York, 1988.
32. Van Steen, K.; Curran, D.; Kramer, J.; Molenberghs, G.; Van Vreckem, A.; Bottomley, A.; Sylvester, R. Multicollinearity in prognostic factor analyses using the EORTC QLQ-C30: Identification and impact on model selection. *Stat. Med.* **2002**, *21*, 3865–3884.
33. Gunst, R.F.; Mason, R.L. Advantages of examining multicollinearities in regression analysis. *Biometrics* **1997**, *33*, 249–260.
34. Sauerbrei, W.; Schumacher, M. A bootstrap resampling procedure for model building: Application to Cox regression model. *Stat. Med.* **1992**, *11*, 2093–2109.
35. Schumacher, M.; Hollander, N.; Sauerbrei, W. Resampling and crossvalidation techniques: A tool to reduce bias caused by model building. *Stat. Med.* **1997**, *16*, 2813–2827.
36. Kendall, M.G. *A Course in Multivariate Analysis*; Charles Griffin: London, 1957.
37. Kendall, M.G. *Multivariate Analysis*; Charles Griffin: London, 1975.
38. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67.
39. Cessie, S.Le.; Van Houwelingen, J.C. Ridge estimators in logistic regression. *Appl. Stat.* **1992**, *14*, 191–201.
40. Augustin, N.H.; Sauerbrei, W.; Schumacher, M. *Incorporating Model Selection into Prognostic Factor Model Predictions. Freiburg Center for Data Analysis and Modeling*;—Preprint no. 76 at; <http://www.fdm.uni-freiburg.de/pub/preprints.php> (accessed November 2003).

41. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592.
42. Van Steen, K.; Tahri, N.; Molenberghs, G. Introducing the multivariate Dale model in population-based genetic association studies. *Biom. J.* **2004**, *46* (2), 187–202.
43. Dale, J.R. Global odds ratio models for bivariate, discrete, ordered responses. *Biometrics* **1986**, *42*, 909–917.
44. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*; Chapman and Hall: London, 1989.
45. Ritchie, M.D.; Hahn, L.W.; Roodi, N.; Bailey, L.R.; Dupont, W.D.; Parl, F.F.; Moore, J.H. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* **2001**, *69*, 138–147.
46. Calle, M.L.; Urrea, V.; Malats, N.; Van Steen, K. MB-MDR: Model-Based Multifactor Dimensionality Reduction for detecting interactions in high-dimensional genomic data. Technical Report, Department of Systems Biology, Universitat de Vic, 2008.
47. Achen, C.H. *Interpreting and Using Regression*; SAGE: Beverly Hills, CA, 1982.
48. Devlin, B.; Roeder, K.; Wasserman, L. Analysis of multi-locus models of association. *Genet. Epidemiol.* **2003**, *25*, 36–47.

Multidimensional Data Analysis: Penalized Regression Methods

Keumhee Carriere Chough

University of Alberta, Edmonton, Alberta, Canada

Xiaoming Wang

Shanghai University of Finance and Economics, Shanghai, China

Abstract

This entry provides an overview of and comments on regularization regression procedures by first reviewing traditional penalized methods, and then Lasso and its extensions, and also discussing some penalty function designs that lead to bias reduction methods.

Keywords: Bias-reduction technique; Multi-dimensional data; Penalized regression; Two-step model building; Variable selection.

INTRODUCTION

Developments in bio-imaging technology have made it possible to widely use genome-wide experiments to do whole-genome analyses of Single Nucleotide Polymorphism (SNP) patterns, microarray-based gene expression analyses, and modern-mass-spectrometry-based protein expression analyses to study many health related issues. These new technologies allow observing thousands of SNP, mRNA expression, or protein expression measurements in a single assay, thus providing massive amounts of biological data on study samples. There has been a substantial impact on disease subtyping as a result of the discovery of biomarkers that can be used in a wide range of applications, for example, quantifying individual risk and/or its modification by non-host (environmental) factors based on host genetic (SNP) patterns, detecting disease in an earlier stage than currently possible, enabling molecular classifications of diseases for accurate targeting by therapies, and predicting patient prognosis and/or determining the best (tailored) therapy based on the expression profiles of the patient and his or her disease. While these new technologies offer tremendous promise, a number of critical issues need to be resolved in order to achieve and maximize their utilities for use in therapeutic and preventive medicine.

Statistics methods play a vital part of these new biological analyses. Statistics methods are required, for example, in the normalization of expression from multiple arrays, in selecting significant differentially expressed features, in tumor classifications, and in expression pathways and networks analysis. The statistical challenges involved in defining the characteristics of expression-based studies include: (i) there is a much larger number of features

contributing expression measures, which far exceeds the number of observations available, and (ii) by virtue of pathway (or network) relationships, the expression measures tend to be highly correlated. These challenges have spawned several new methodological approaches to adopt penalization techniques for model/feature selection, for example, Lasso^[1] and adaptive Lasso^[2], and SCAD^[3] and non-negative garrote.^[4] Commonly used algorithms include LARS,^[5] LQA,^[3] and MM.^[6] In this chapter, we provide a overview of and comments on these regularization regression procedures.

PENALIZED APPROACHES

The general form of a penalization technique can be described as *loss + penalty*. Under a selected *penalty* function, change of *loss* leads to various types of regression models. The least-square loss leads to a regularized linear model; a negative log-likelihood loss leads to a regularized likelihood model; and a hinge-loss function leads to a support-vector machine model. In this section we address a key aspect of using a penalization technique: choosing a penalty function. We first review traditional penalized methods, and then *Lasso* and its extensions. Finally, we discuss some penalty function designs that lead to bias reduction methods.

Traditional Penalized Methods

L_0 Penalty (or Entropy)

The entropy penalty function can be described as $p_\lambda(|\theta|) = \frac{1}{2} \lambda^2 I(|\theta| \neq 0)$. Here and after, λ (or λ_1 and λ_2)

is used to present tuning parameter(s). If we denote $|M|$ as the number of variables selected in a model and focus on finding a model with $|M| \leq t$, then we can use $\frac{1}{2} \lambda^2 \sum_{j=1}^p I(|\beta_j| \neq 0) = \frac{1}{2} \lambda^2 |M|$ in the following penalized least-square problem:

$$\min \frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \frac{1}{2} \lambda^2 |M|, \quad (1)$$

where $\{(y_i, x_{i1}, \dots, x_{ip}), i = 1, \dots, n\}$ is a sample set of n observations, and $\beta_j, j = 1, \dots, p$ are coefficients to be estimated.

This optimization problem leads to the traditional ordinary least-square model selection methods including *stepwise* and *all-subsets regression* procedures. These procedures pick predictors and then use least squares or maximum likelihood to estimate the coefficients. In other words, they focus on variable selection and do nothing specific to estimate parameters. Although these methods produce a sparse model, they have several drawbacks: (i) theoretical properties can be hard to understand, (ii) estimates are extremely variable because of their inherent discreteness, as addressed by Breiman,^[7] and (iii) computational costs are very high as p gets large.

L_2 Penalty (or Ridge)

The L_2 penalty function $P(|\theta|) = \theta^2$ leads to the popular ridge regression.^[8] It achieves improved prediction accuracy via shrinkage. In a linear regression situation, ridge regression achieves coefficients shrinkage by a constraint $\sum_{j=1}^p \beta_j^2 \leq t$. An equivalent formulation is afforded by L_2 penalized regression:

$$\min \frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \frac{1}{2} \lambda \sum_{j=1}^p \beta_j^2, \quad (2)$$

In contrast to the entropy, ridge regression is not concerned with variable selection (it uses all candidate predictors). Rather, it deals with how coefficients are estimated. Although ridge penalization does not produce unbiased estimators and a parsimonious model, it achieves better prediction performance than a pure least-square estimation through a bias-variance tradeoff, especially when the predictors in the model are highly correlated. Numerous studies have shown that ridge penalization is an effective tool for dealing with co-linearity among predictors. Ridge regression can be obtained efficiently via standard quadratic programming when $p \leq n$ or dual algorithms when $p > n$.

Lasso and Its Extensions

Lasso. In a linear regression situation, the L_1 penalty function $p_\lambda(|\theta|) = \lambda|\theta|$ leads to Lasso^[1] (the least absolute

shrinkage and selection operator). Lasso is a shrinkage method similar to ridge, with subtle but important differences. The Lasso estimator is a solution to the minimization of the residual sum of squares, subject to a constraint on the sum of absolute values of the regression coefficients $\sum_{j=1}^p |\beta_j| \leq t$. It is equivalent to the following regularized least-square problem with an L_1 penalty

$$\min \frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j|. \quad (3)$$

The Lasso method combines the desirable features of ridge regression and subset selection procedures. Firstly, similar to subset selection, the L_1 penalty results in variable selection. Because of the nature of the constraint, making λ large enough or equivalently t sufficiently small will cause some of the coefficients to be (estimated as) exactly zero, as variables with coefficients of zero are effectively eliminated from the model. Another important feature of the L_1 is that its shrinkage property is similar to ridge. The big difference between ridge and lasso is that ridge shrinks the coefficients toward zero proportionally to the value of the coefficients and lasso exerts the same force on all nonzero coefficients. Hence for significant variables, the shrinkage towards zero of an L_1 penalty is less than that of an L_2 penalty. This is important for providing accurate predictions of future values in high-dimensional situations.

Efron et al.^[5] developed Lars (least angle regressions) as an efficient algorithm for computing the whole Lasso solution path as λ is varied. The Lars method can be viewed as a stagewise method, which accelerate the computations using mathematical formulas. Their algorithm exploits the fact that the coefficient profiles are piecewise linear, which leads to computational costs as low as the full least-squares fit on the data. Lars has some remarkable properties. The first one is speed: the entire sequence of the Lars steps with $p < n$ variables requires $O(p^3 + np^2)$ computations—the cost of a least-square fit on p variables. Second, the basic Lars algorithm, which is based on the geometry of angle bisection, can be used efficiently to fit Lasso and stagewise models, with certain modifications in higher dimensions.

Elastic net. Zou and Hastie have proposed the elastic-net approach:^[9]

$$\min \frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \frac{\lambda_2}{2} \sum_{j=1}^p \beta_j^2 \quad (4)$$

The elastic-net penalty can be viewed as a compromise between ridge and Lasso. The shrinkage behavior of the elastic-net penalty on small coefficients is much like that of Lasso, with almost the same shrinkage amount toward zero; the behavior on large coefficients is as heavy as ridge, proportional to the absolute value of the coefficients.

An intuitive aspect of this method is that the L_1 penalty eliminates redundant variables and the L_2 penalty deals with possible high correlation among predictors.

The elastic-net approach has the following properties: (i) similar to Lasso, it produces a sparse model by variable selection; (ii) similar to ridge, it shrinks toward zero for non-zero coefficients and shrinks together the coefficients of correlated predictors; and (iii) for a fixed λ_2 , the elastic net solution path as λ_1 varied can be computed efficiently using the Lars algorithm. With a suitable L_2 penalization, elastic net can provide a more stable predictive performance than Lasso when predictors are highly correlated. This is the main difference between Lasso and elastic net.

Extensions or Alternative to Lasso

The Lasso approach laid down the groundwork for high-dimensional analysis and thus became a popular regression method with attractive features. This idea has been applied broadly to generalized linear models and Cox's proportional hazard models.^[10] Recently, a variety of research activities have been devoted to related penalization methods. These include (i) the fused Lasso^[11] for a sequence of predictors where there is a natural ordering of the predictors and the coefficients for nearby predictors are normally similar; (ii) the group Lasso,^[12] which handle groups of predictors with a natural grouping, for example, sets of bases for smoothing splines and sets of dummy variables representing levels of factors; (iii) the L_1 realization path for the SVM^[13] (support-vector machine); (iv) the Dantzig selector,^[14] a modified version of the Lasso; and (v) the graphical Lasso^[15] for sparse covariance estimation and undirected graphs.

Bias Reduction Methods

SCAD and Its Oracle Property

Fan and Li^[3] observe that a good penalty function results in an estimator with three properties: *unbiasedness*, so that the resulting estimator is nearly unbiased when the unknown coefficient of the regression is large; *sparsity*, so that the resulting estimator has an automatic thresholding rule that sets some estimated coefficients to zero to reduce the complexity of the model if the estimated values are small; *continuity*, so that the resulting estimator is a continuous function of data to avoid instability in model prediction.

Fan and Li^[3] discuss various penalty functions in terms of the above three properties and establish the conditions for a penalty function that satisfies them. Their work has led to the development of a new penalty function, SCAD, a non-concave penalty function that has been widely used in penalization techniques in different statistical contexts. Its related penalization technique is considered to produce estimators that are less biased than estimators produced by the Lasso approach.

The proposed SCAD is a continuous differentiable function defined by

$$P'_\lambda(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)}{(a-1)\lambda} I(\theta > \lambda) \right\},$$

for $a > 2$ and $\theta > 0$ (where A_+ is defined as the positive part of A). Different from various L_q penalties, SCAD is a non-concave penalty function. It leads to the following penalized least square

$$\min \frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \sum_{j=1}^p P_\lambda(|\beta_j|). \quad (5)$$

Fan and Li^[3] note that SCAD results in an estimator with three properties: *unbiasedness*, *sparsity*, and *continuity*. The most attractive aspect of SCAD is that it does not penalize coefficients with large values. This property makes the SCAD estimator asymptotically equivalent to an oracle estimator, which is obtained by deleting all irrelevant predictor variables. This result implies that SCAD can achieve consistent model selection and optimal prediction simultaneously as the sample size increases. Their work has laid down the groundwork for variable selection problems in the finite parameter setting. Further studies related to SCAD penalization, including theoretical analysis and extensions to other loss situations, can be found in the literature,^[6,16,17] among others.

The solution to (5) can be calculated using LQA^[3] (local quadratic approximation), which turns (5) into an iterative high-dimensional quadratic program problem.

Combined Penalty (CP)

Wang et al.^[18] proposed a new combined penalization technique. In a linear regression situation, the regularized least-square can be expressed as follows:

$$\min \frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \sum_{j=1}^p J_{\lambda_1, \lambda_2}(|\beta_j|). \quad (6)$$

The combined penalty function $J_{\lambda_1, \lambda_2}(|\beta_j|) = \sum_{j=1}^p \left(P_{\lambda_1}(|\beta_j|) + \frac{\lambda_2}{2} \beta_j^2 \right)$ is a compromise between ridge and SCAD.

The penalty in Eq. (6) has two separate functions. As described,^[18] the SCAD is adopted mainly for eliminating redundant features, and the ridge can give suitable shrinkage to non-zero coefficients and reduce the influence of highly correlated predictors. This flexible design allows us to adjust the penalty and make it suitable for various data settings, for example, when $p \gg n$ and the strength of the data is too weak to provide distinct and true features from a highly correlated feature cluster. In this case, we should use relatively heavier

L_2 penalization or heavier shrinkage to non-zero coefficients. At the other extreme, if $n \gg p$ and the data is strong enough to detect all the significant features, reducing the unnecessary estimation bias of the coefficients using a very small value of λ_2 may help to reduce predictive error.^[18]

Here, we emphasize that incorporating the ridge penalty to deal with colinearity among predictors is very important in high-dimensional data analysis. As pointed out,^[17] even when the predictors are independent, the maximum sample correlation can become high as the dimension becomes large. Hence, one crucial aspect of choosing a penalty function is deciding how to deal with high-colinearity among predictors. In other words, shrinkage of large coefficients is necessary in high dimensional settings.

Wang et al.^[18] proved that, similar to SCAD, the combined penalization technique results in an estimator with three properties: *sparsity*, *continuity*, and asymptotically enjoying the so-called *oracle* property under some regular conditions. As $\lambda_2 > 0$, the estimator of (6) is no longer unbiased, but it is *asymptotically unbiased* as $\lambda_2 \rightarrow 0$.

The solution to (6) can be calculated using LQA to turn (6) into an iterative quadratic program problem and using a dual algorithm to overcome the computational burden related to the high-dimensional quadratic program.

Two-Step Strategy

Lasso is an attractive regularization method for high-dimensional data analysis that combines variable selection with an efficient computational procedure. However, the rate of convergence of the Lasso is slow for some sparse high-dimensional data, as the number of predictor variables grows with the number of observations. The reason can be traced back to the biased nature of the Lasso. Fan and Li^[3] first pointed out that asymptotically the Lasso has a non-ignorable bias for estimating non-zero coefficients. To overcome this difficulty, two-step model building procedures are proposed to reduce the bias in estimation.

Relaxed Lasso

One example is the relaxed Lasso proposed by Meinshausen.^[19] For $\lambda \in [0, \infty)$ and $\phi \in (0, 1]$, in the first step, select a subset of variables $M_\lambda = \{1 \leq j \leq p \mid \hat{\beta}_j^\lambda\}$ using Lasso

$$\hat{\beta}^\lambda = \arg \min_{\beta} \frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j|. \quad (7)$$

Then in the second step, the relaxed Lasso estimator is defined as the solution of the following penalized least-square problem:

$$\min \frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j I_{M_\lambda}(j) \right)^2 + \phi \lambda \sum_{j=1}^p |\beta_j|, \quad (8)$$

where $I_{M_\lambda}(\cdot)$ is an indicator function on the subset of variables M_λ in $\{1, \dots, p\}$. The first step can be considered as a feature-selection step, while in the second coefficient-estimation step the parameter $\phi \leq 1$ relaxes the penalty and reduces the bias of the coefficient estimation.

When we have enough information from the data, we can expect that with a suitably chosen λ , the subset M_λ can cover all or most of the significant features. Then relaxing the penalty can give a better parameter estimation. Meinshausen^[19] proved that, as n tends to ∞ , the convergence rate of $\text{IE} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \hat{\beta}_j \right)^2 - \sigma^2$ is almost unaffected by the number of predictors if the tuning parameters λ and ϕ are chosen appropriately.

Adaptive Lasso and Elastic Net

Other examples of the two-step strategy include adaptive Lasso and adaptive elastic net. Zou^[2] proposed the adaptive Lasso estimator

$$\min \frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p \hat{w}_j |\beta_j| \quad (9)$$

where $\hat{w}_j, j = 1, \dots, p$ are the adaptive data-driven weights and can be computed by $\hat{w}_j = (|\beta_j^{\text{ini}}|)^{-\gamma}$, with γ a positive constant and β_j^{ini} an initial root-n consistent estimate of the coefficient β_j . Setting the adaptive data-driven weights lets the penalty enforce different levels of penalization on coefficients based on their performance in the initial estimate. The essential idea behind this approach is similar to that of SCAD: larger coefficients receive fewer penalties. Zou^[2] showed that under some regularity conditions the adapted Lasso performs asymptotically as well as the oracle.

The adaptive elastic-net^[20] combines elastic-net with the technique of adaptive data-driven weight used in adaptive Lasso. Similar to adaptive Lasso, the adaptive elastic net can be viewed as a two-step model building procedure. In the first step we use elastic net

$$\min \frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \sum_{j=1}^p (\lambda_1 |\beta_j| + \lambda_2 \beta_j^2). \quad (10)$$

to get initial estimates of the coefficients $\beta_j^{\text{ini}}, j = 1, \dots, n$. In the second step an adaptive elastic net can be estimated from the following adaptive weighted optimal problem:

$$\min \frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \sum_{j=1}^p (\lambda_1^* \hat{w}_j |\beta_j| + \lambda_2 \beta_j^2) \quad (11)$$

where $\hat{w}_j, j = 1, \dots, p$ are the adaptive data-driven weights and can be computed by $\hat{w}_j = (|\beta_j^{\text{ini}}|)^{-\gamma}$, and λ_1^* is a new L_1 tuning parameter value different from that of λ_1 in (10).

Similar to adaptive Lasso, the adaptive elastic-net also performs asymptotically as well as the oracle.

TUNING AND FEATURE SELECTION ISSUES

Tuning issues are crucially important in feature/model selection in the context of high-dimensional regression. The distinguishing characteristics of genome-wide studies ($p \gg n$ and highly correlated expression measures) make feature selection and tuning very difficult. Here, we discuss issues related to the two principal means of tuning: one is criteria-based and the other is prediction error-based. Similar to the previous section, our discussion focuses on the setting with least-squares loss.

A variety of criteria developed for classical model selection studies (in the $n \geq p$ setting) have been used for model selection problems in high-dimensional settings, including AIC^[21] (Akaike Information Criterion), C_p ^[22], BIC^[23] (Bayesian Information Criterion), and ϕ -criterion.^[24] Common to all these approaches is penalization of training error estimates as well as the need to estimate the residual variance σ^2 of the full model. However, estimating σ^2 is problematic, because a full model will yield $\hat{\sigma}^2 = 0$ as $p \gg n$.

The merit of K-fold CV (Cross-Validation), which bases model selection on prediction error determination, has been convincingly advanced (see Breiman^[25]). But CV can also be problematic, as $p \gg n$, especially when n is small. The difficulties mainly come from the large variability of CV estimates of the prediction error.

The distinguishing characteristics of genome-wide studies ($p \gg n$ and highly correlated expression measures) contribute to the instability of feature and model selection, that is, the existence of the so-called “Rashomon effects” (see Berman^[25])—many competing distinct models with comparably good performance.

In a regression problem with a limited dimension and a sufficiently large number of observations, a small disturbance of data will not alter the final selected model very much. In a complex high-dimensional problem, however, a small perturbation of data could lead to a selection of a totally different feature subset, because the sample size is often limited relative to the dimension and noise levels of data. This is one of the potential reasons why different studies with an identical biomarker-discovery aim could lead to inconsistent results in gene/protein expression analysis.

All these difficulties in both tuning and model selection arise from the distinguishing characteristics of genome-wide studies ($p \gg n$ and highly correlated expression measures). The data used for regression usually does not have enough information strength. This implies that the difficulties are not from lack of statistical method developments, and hence cannot be overcome easily by developing even more powerful statistical techniques.

CONCLUSIONS

The purpose of this entry is twofold: the first is to review and comment on the current state of penalization regression techniques and second is to explain the “sensitivity”—a common phenomenon related to expression-based biological studies and other high-dimensional data analyses.

All these penalization methods are designed for high-dimensional analysis. The Lasso and elastic-net have considerable promise, offering speed, relatively stable predictions and interpretability. Alternate penalization techniques such as SCAD, CP, adaptive Lasso, and adaptive elastic-net have advantages for certain kinds of data. Adopting “bias-reduction” strategy allows them asymptotically enjoying the so-called *oracle* property under some regular conditions. While whether these “bias-reduction” techniques can get higher predictability than Lasso/Enet depends on “information strength” included in data. Further work is still needed to investigate performance of these methods.

The difficulties in tuning and “sensitivity” in model selection arise from the distinguishing characteristics of genome-wide studies ($p \gg n$ and highly correlated expression measures). Since these issues arise due to the fact that the data used for regression does not have enough information strength, it is impossible to completely overcome these difficulties by developing more powerful statistical techniques. We believe that an essential remedy is to add new information to the data, either providing a larger data set by combining information from several genome-wide experiments or providing new structure/cluster information on the expression data.

ACKNOWLEDGMENTS

This work was supported by grants from Alberta Heritage Foundation for Medical Research, Natural Sciences and Engineering Council of Canada and Shanghai University of Finance and Economics through Project 211 Phase III and Shanghai Leading Academic Discipline Project (Project Number: B803).

REFERENCES

1. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. Royal Stat. Soc. Ser B* **1996**, 58, 267–288.
2. Zou, H. The adaptive Lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, 101, 1418–1429.
3. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, 96, 1348–1360.
4. Breiman, L. Better subset regression using the nonnegative garrote. *Technometrics* **1995**, 37, 373–384.
5. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, 32, 407–499.

6. Hunter, D.; Li, R. Variable selection using MM algorithms. *Ann. Stat.* **2005**, *33*, 1617–1642.
7. Breiman, L. Heuristics of Instability and Stabilization in Model Selection. *Ann. Stat.* **1996**, *24*, 2350–2383.
8. Hoerl, A.; Kennard, R. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67.
9. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. Royal Stat. Soc. Ser. B* **2005**, *67* (2), 301–320.
10. Tibshirani, R.J. The lasso method for variable selection in the Cox Model. *Stat. Med.* **1997**, *16*, 385–395.
11. Tibshirani, R.; Saunders, M.; Rosset, S.; Zhu, J.; Knight, K. Sparsity and smoothness via the fused lasso. *J. Royal Stat. Soc. Ser. B* **2005**, *67*, 91–108.
12. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. Royal Stat. Soc. Ser. B* **2006**, *68*, 49–68.
13. Hastie, T.; Rosset, S.; Tibshirani, R.; Zhu, J. The entire regularization path for the support vector machine. *J. Mach. Learn. Res.* **2004**, *5*, 1391–1415.
14. Dantzig, G.B. *Linear Programming and Extensions*; Princeton University Press: Princeton, NJ, 1963.
15. Friedman, J.H.; Hastie, T.; Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **2008**, *9* (3), 432–441.
16. Antoniadis, A.; Fan, J. Regularization of wavelets approximations. *J. Am. Stat. Assoc.* **2001**, *96*, 939–967.
17. Fan, J.; Lv, J. *Sure Independence Screening for Ultra-High Dimensional Feature Space*. Technical report, Department of Operations Research and Financial Engineering, Princeton University: Princeton, NJ, 2008.
18. Wang, X.; Park, T.; Carriere, K.C. Variable selection via mixed penalization for high dimensional data analysis with application to micro-arrays **2009**, Submitted.
19. Meinshausen, N. Lasso with relaxation. *Comput. Stat. Data Anal.* **2007**, *52*, 374–393.
20. Zou, H.; Zhang, H.H. On the adaptive elastic-net with a diverging number of parameters. *Ann. Stat.* **2009**, *37*(4), 1733–1751.
21. Akaike, H. Information theory and an extension of the maximum likelihood principle. *Second Int. Symp. Info. Theory* **1973**, 267–281.
22. Mallows, C.L. Some comments on Cp. *Technometrics* **1973**, *15*, 661–675.
23. Schwartz, G. Estimating the dimension of a model. *Ann. Stat.* **1979**, *6*, 461–464.
24. Hannan, E.J.; Quinn, B.G. The determination of the order of autoregression. *J. Royal Stat. Soc. Ser. B* **1979**, *41*, 190–195.
25. Breiman, L. Statistical modeling: the two cultures. *Stat. Sci.* **2001**, *16*, 199–215.

Multinational Clinical Trials

Thomas R. Weihrauch

Bayer AG, Wuppertal, Germany

Abstract

Multinational clinical trials are, in most cases, set up in order to complete the trial more rapidly when one country cannot provide an adequate number of patients to meet the sample size requirements. This entry discusses the procedural details, advantages, and obstacles of a multinational clinical trial.

INTRODUCTION

Multinational clinical trials of investigational drugs or of approved drugs for new indications have become increasingly popular over these past several years. In most cases, they are set up in order to complete the trial more rapidly, particularly when the condition under study has a low incidence and any one country cannot provide an adequate number of patients to meet the sample size requirements and the study time table. Experience at Bayer indicates that the number of such trials has increased as has the number of participating centers and the overall costs per trial. Moreover, the preparation, evaluation, and documentation for such trials is more extensive than it is for multicenter trials performed in one country. The obstacles to multinational trials are manifold and they arise from differences in standard and style of medical practice, differences in drug regulations and differences of medical culture, not to mention the difficulties inherent in organizing the trial itself. There are, however, many advantages to this type of investigation because it permits a more rapid enrollment of patients by a wide range of treating physicians in a variety of international medical institutions. Such trials are best performed, however, in countries and centers that have similar medical standards and practice. From the drafting of the protocol to the publication of the final results, the successful performance of a multinational trial requires an optimal level of teamwork, a strong study coordination, and most importantly, a committed and highly motivated expert panel of international investigators.

OVERVIEW

Clinical trials provide the basis for assessing the efficacy and safety of investigational drugs or of approved drugs for new indications. In order to gain approval by International Regulatory Authorities and acceptance by the medical community, large randomized, controlled, double blind trials

are often performed in Phase III of clinical development. In many instances, investigational centers from several countries participate in such trials and become part of a multinational effort. There are many obvious advantages of such an approach, but there are also some drawbacks or obstacles. Despite these, however, multinational clinical trials have become increasingly common in clinical research. Reviews on the more general aspects of this approach have been published by Mondan,^[1] and Spilker,^[2] and our group.^[3,4] This paper will address some of the more pragmatic issues applicable to this subject based on the experience that the Pharmaceutical Division of Bayer has gained over these past several years. The subject is of no small interest because it should be recognized that major advances in the therapy of such devastating diseases as acquired immunodeficiency syndrome (AIDS) and cancer will likely only come about through the performance of high-quality, multicenter trials, which in many instances, will be multinational.

ADVANTAGES

The availability of suitable patients for a specific study may be compromised because of the low incidence of disease under investigation or because of the stringent inclusion or exclusion criteria of the study protocol. For example, multinational clinical trials have been used to investigate new drugs for the treatment of angina pectoris because of the detailed and strict guidelines that must be followed to prove efficacy in this condition.

The enrollment of patients into a study may also be difficult because the protocol may require special testing or equipment, which may not be standard in all hospitals. For a medication that was intended for use in emergency circumstances, one of our studies could only be performed in centers that had ambulances with special equipment for drug delivery and specially trained personnel. Because of these prerequisites, the study had to include centers from several countries in Europe.

The enrollment of a large number of patients from a multitude of centers from different countries permits the demonstration of a drug's efficacy and safety under varying conditions such as lifestyle, diet, and concomitant therapy. Moreover, larger trials provide a greater opportunity for physician investigators to participate in innovative clinical research and to gain firsthand experience with new agents. Their patients profit likewise, first from the opportunity of receiving a new potentially important therapeutic modality, and second, by the special attention and care that must be given to all patients enrolled in study protocol.

OBSTACLES

Anyone who has experienced the challenge of organizing a multinational clinical trial is aware of the many obstacles that must be overcome to successfully complete the trial. Obstacles common to multicenter trials within one country are usually compounded or amplified when the attempt is made to carry out a multinational effort. Some of these factors are addressed below.

Regulatory

Although much progress has been made regarding the harmonization of the regulatory framework for clinical development of new drugs, there are still discrepancies from one country to another. For example, if a transatlantic trial is planned with a treatment duration of 3 months, the clinical investigators in Europe must await completion and clearance of a 6-month toxicology study whereas their U.S. colleagues may commence on the basis of only 3 months' toxicology data.

A further obstacle may arise from the difficult legal requirements of informed consent. For example, in Germany, patients with altered mental function or those in coma with stroke or head injury may not be able to give consent and therefore are not eligible for study.

For new therapeutic principles in life-threatening conditions, it is not uncommon that ethical review committees have varying opinions. This may be a result of different standard of therapy or may be related to different views on the ethics of treatment procedures. Although, in some countries such as Germany, ethics committees do not formally approve protocols but rather give advice, it is highly recommended that agreement on the protocol be obtained from all committees that are involved in protocol evaluation.

Medical Culture and Practice

Differences in cultural background among the participating countries may provide stumbling blocks to the smooth progression of the trial. For example, in Germany the condition impaired brain function of old age is not recognized

in Anglo-Saxon countries and requires different inclusion criteria and treatment than, for example, Alzheimer's disease. The diversity of language alone may impact on the classification of the condition to be treated. A number of misunderstandings may arise when Anglo-Saxon abbreviations are used which are abstruse to investigators who are not thoroughly familiar with the English language. These have the potential to cause confusion and often miscommunication.

In a major multinational clinical trial of stroke sponsored by Bayer, centers from Great Britain had to be excluded because of the different therapeutic strategies practiced there as compared with continental Europe. In Great Britain conservative management was more prevalent, whereas in Europe a more aggressive approach to therapy was practiced.

It is well accepted that the exposure to risk factors for coronary artery disease because of diet differs between people of nordic and mediterranean countries in Europe. In studies of metabolic diseases, such as diabetes or hypolipoproteinemia, diet may have to be standardized with the disadvantage that the results then may not be applicable to the daily routine of patients in individual countries.

The diversity in medical culture and practice should not be underestimated in the establishment of a multinational trial because this could have serious consequences on the outcome. This must be considered at an early stage and may lead to stratification of patients, the exclusion of certain patients, or even the exclusion of certain centers or countries from participation.

Clinical Parameters

Within a multinational clinical trial, key elements of the study protocol must be standardized. This sometimes poses problems in the area of clinical parameters because they may vary greatly from one region to another in established indications, as in the case of angina pectoris (time to onset of angina vs. onset of change in the electrocardiogram). Psychometric tests are difficult to standardize and, in order to address this difficulty, we have had to perform studies using two or three parallel scores for important efficacy variables.

Reference Drugs

Reference drugs are included in Phase IIB and Phase III clinical trials as active controls and are required in a number of countries to provide price comparisons for health authorities. For a multinational clinical trial, the inclusion of reference drugs is often a problem because of factors such as labeling, formulations, appearances, names, and even countries of origin. The same holds true for comedication or standard therapy, if the investigational drug is add-on.

Adverse Events

Reporting of adverse events (AEs) may pose a problem because of the different approaches used in different countries, particularly with respect to the formulation, format, and the time period allowed for reporting. All AEs are of interest but the prompt reporting of serious or unexpected events must be assured; they must also be evaluated immediately as they may be the first indication that a study must be stopped because of an unfavorable risk/benefit ratio. The standard reporting guidelines and the standard operating procedures carefully developed by most sponsors of major multinational trials are now streamlining this process. Moreover, the international introduction and implementation of Good Clinical Practices (GCPs) is doing much to improve the timely reporting of more precise information on AEs.

Health Care Systems

Health care systems differ widely among countries and may influence the standard therapy used for specific diseases (reimbursement policy). In addition to possibly affecting the overall outcome of a specific illness, these differences may render multinational clinical trials on the cost effectiveness of new therapy very difficult or even impossible.

POINTS TO CONSIDER

Listed below are some of the important points to consider when preparing for and carrying out a multinational clinical trial.

Coordination

Critical to the success of the trial is the establishment of an effective coordination function which plays a pivotal role in all aspects of the project. At Bayer, this function resides with the coordinator of the multinational study (CMS). The CMS is appointed by medical research management and reports functionally to the International Clinical Project Manager. The CMS is the intermediary between the clinical project leaders (CPLs) of the different medical departments involved in the trial and the central services departments of Bayer.

Depending upon the size of the trial and other factors, it may be desirable to establish committees or individuals responsible for selected functions. The following are a list of some of the functions that may require specific supervision:

- Drafting and finalizing the trial protocol.
- Regulatory activities.
- Tracking study supplies.
- Clinical Quality Assurance (CQA).

- Data management/statistical analysis.
- Safety committee (an individual or individuals not associated with the trial, but usually within Bayer, who will assess AEs and their reportability to regulatory authorities as well as their impact on the course of the study).
- An independent Extramural Review Committee (experts in the study area responsible for reviewing each Case Report Form (CRF) to determine validity and other aspects of patient management and to provide expert advice on efficacy and safety issues that may arise during the study).
- Data Analysis and Review Team/Manuscript Review Committee (usually composed of key investigators who will provide input on data analysis and on the generation of the study manuscript).

Whether or not committees are formed or whether these functions are taken over by individuals, the responsibilities must be clearly defined very early in trial development so that confusion, inefficiency, and delays may be avoided.

The CMS for the trial must work in concert with all of these functions and with the CPLs in the participating countries who maintain full local responsibility. An effective and continued communication between the CMS and CPL is essential not only in the planning phase, when the national characteristics of each country may be discussed and considered in protocol development, but also during patient recruitment so that any problems which may arise may be rapidly resolved.

The overall manpower needs for the trial will depend principally on the complexity of the protocol and CRF, and on the number of centers and patients to be included as well as the expected rate of patient enrollment. To provide optimal human resources, the needs of both the coordinating country and all participating countries must be carefully determined.

Outline of the Trial

To establish the feasibility of a multinational clinical trial, a concrete study outline should be prepared early and provided to CPLs in countries of interest. This document permits each CPL to evaluate the potential for participation and to determine resource requirements, which can then be discussed with local Medical Direction. The trial outline must include the condition and population of patients to be studied (inclusion and exclusion criteria, the nature of the comparative drugs, their dosage, and conditions of use).

Time lost in clinical trials is often caused by overoptimistic estimation of the number of patients available for study. This is usually a function of the inclusion and exclusion criteria defined in the protocol and can make the fraction of suitable patients sometimes surprisingly small. Very slow patient enrollment can become demoralizing not only

to the sponsor but also to the investigator team. Realistic estimates of patient availability must be made and, in some cases, it may even be desirable to do a pilot study to test a protocol before embarking on any major project.

Protocol Development

Considering the potential problems that may arise as a result of differences in medical culture and practice, it is critical that all participating investigators reach precise agreement on the study protocol. The objectives of the study must be clearly understood and investigators must reconcile these with the given conditions within their own institutions. The progression from the rough outline of a study to the final version of the protocol for a multinational clinical trial may take months of careful deliberation among experts of various disciplines. This process should be effectively led by a company staff member usually in concert with a principle investigator.

The Case Report Form

In developing the CRF for a multinational clinical trial, there are several issues that need special attention. For example, when one common language is used, it is imperative that all terms be clearly understood along the entire chain of personnel involved in the trial. It may appear self-evident but the understanding must first start with the team that prepares or designs the form and proceed to the local CPLs, investigators, and very importantly, to the investigational team member who will complete the CRF on site. Of course, those individuals monitoring the trial including any Contract Research Organization (CRO) personnel must also be thoroughly knowledgeable about the form. Finally, the data analysis or biometrics team must be resident experts. One of the more critical aspects of the CRF is the coding of free text entries such as AEs or investigator comments. It is important that these entries be clear and consistent with a commonly used glossary and that any ambiguities of language be clarified before the data are computer key punched.

Prestudy investigator meetings, during which the CRF is carefully reviewed, will do much to foster common understanding. However, a comprehensive written CRF Guide will be an extremely useful document to have available at study start up. The guide should be updated periodically by the CMS during the study as new and unanticipated issues arise which need clarification. Such a guide will do much to not only ensure consistency and quality of data, but will also serve to facilitate the eventual data processing.

If CRFs are translated into different languages, it is imperative that the translation be carefully checked by bilingual medical experts to assure accuracy. Ample time for designing, reviewing, printing, and shipping of forms must be provided.

Regulatory

The starting point for meeting the regulatory requirements of all countries is consultation with the standard operating procedures of the sponsor and with GCP guidelines recognized worldwide. These documents provide the basis for generating a common set of investigator regulatory responsibilities. However, one should keep in mind that there may be specific requirements for certain countries. Listed below are some examples:

- Prior agreement from authorities via submission of comprehensive preclinical documentation, e.g. Clinical Trial Exemption (CTX), Investigational New Drug (IND).
- Permission from authorities to import a drug.
- Prescribed content and format of documents to be submitted to Ethical Review Committees/Institutional Review Boards.
- Requirements for patient information (written or oral).
- Obligatory drug labeling and procedures for drug labeling.
- Special procedures for AE reporting to national authorities.
- Liability insurance.

Some of these local requirements may be specifically addressed in the study protocol.

Informed Consent

The informed consent text should be, as far as possible, identical among the participating countries. The CMS is responsible for drafting the original form, then the subject must work closely with all CPLs to reach a consensus on a medically and ethically solid document that complies with the needs of all local ethical review boards and regulatory authorities. A clear and concise patient information sheet that may accompany the informed consent form should also be developed.

Clinical Quality Assurance

In the course of preparing a multinational clinical trial the coordination of quality assurance must be worked out early. This function may be assigned to the CQA manager of the coordinating country who will then work closely with the CQAs of participating countries and the CQA function in Germany. Details of this activity are provided in standard operating procedures.

Biometrics

The biometrician plays a pivotal role in the multinational clinical trial from the conceptualizing of the study protocol through the entire period of patient enrollment and finally

to the critical stage of data processing and analysis. Listed below are some of the more specific biometric activities:

- Development of the statistical plan including sample size calculation, statistical methods, and participation in the identification of the primary and secondary efficacy variables.
- Describe plans for interim analyses if any are to be performed.
- Coordinate development of the CRF.
- Participate in the development of the CRF Guide. (The guide is an indispensable tool that helps harmonize the multiplicity of methods and procedures that may be used in different countries. It also serves to remove ambiguities that may be linked to differences in language and medical culture inherent in each country.)
- Control for consistency of data recorded on the CRF. (Based on early review of a sample of CRFs from different countries, biometrics may identify inconsistencies in data recording which may stem from misunderstanding of CRF or the protocol itself. These must be rectified promptly and appropriate guidelines must be provided in the CRF Guide in order to avert further misunderstandings).
- Data management—very soon after patient enrollment starts, particularly for a complex study including numerous countries and centers, data management must rapidly identify inconsistencies in CRF data through well planned plausibility tests. These will inevitably lead to queries, even in a carefully monitored study, which must be clarified. The clarification process is sometimes a circuitous one from coordinator to CPL to study-site coordinator to investigator and may even have to be repeated more than once. This is very time-consuming and therefore, the earlier it is started the sooner the database may be finalized.
- Statistical analysis in the strict sense.
- Statistical component of the study reports.

Clinical Trial Supplies

The CMS is responsible for ordering clinical trial supplies. In a timely fashion, the CMS must obtain from participating countries all pertinent information on the legal requirements for importation of the study drug. This will ensure an early completion of order forms and the shipping of drug supplies on schedule. The following are important activities pertaining to this subject.

- Procedure for planning and ordering trial supplies. The plan for trial supplies is developed by the CMS in consultation with the International Clinical Project Manager at a very early stage. The plan must address bulk as well as all specially packaged materials including drugs used in positive control studies, test tubes, and even diluents or intravenous (IV) fluids when required.

When coordinating a multinational trial, the procedures to be followed are similar to those used for regional studies. However, considering the larger quantities of medication and materials often required, the coordinator must provide the Department of Clinical Supplies at Bayer with sufficient lead time.

- Label text—To facilitate preparation of labels at the local level, it is useful to have a detailed example of a label text in English. This may be translated and supplemented with any locally required labeling.
- Import license—At an early stage of trial preparation, the CMS must identify those countries that require importation licenses. Drug supplies may not be shipped until these are obtained. In Spain, for example, the study protocol must be approved by the Health Ministry and the Ethics Committee before an application for the import of test drugs may be filed. This stage may take 6 months or more.
- Randomization list—The coordinator must assure that the Department of Clinical Supplies will receive, on time, any applicable randomization lists.
- Bottlenecks—The coordinator should provide the Department of Clinical Supplies with a shipping list prioritized by country and center so that, in case of bottlenecks in the processing of orders, priorities are already established. This should be updated and revised as required.
- Tracking of supply requirements—The CMS is responsible for updating or revising orders based on the level of activity in participating countries (e.g., very slow or rapid enrollment, revised number of patients, new centers), or in case of an amendment to the study protocol (e.g., extension of the study).

Budget

It is no surprise that multinational clinical trials are expensive. The bulk of expense arises from the grants, which must be made to investigators, to assiduously carry out the trial in full compliance with the study protocol. However, other substantial expenses are incurred and include the following: overhead charges by hospital administration, which may exceed 50% of the investigator grant; payment to third-party providers of health care for special biochemical or other tests; investigator meetings; meetings and training sessions for investigator staff; international team meetings. More recently, charges for CROs have become prominent if the intramural capacity of the sponsoring company must be supplemented by extramural services to effectively monitor and rapidly process data from the trial. Other intramural costs include those associated with staffing for GCP, regulatory affairs, and those to support manufacturing, packaging, and shipping of study drug, control medications, and other materials.

In regard to the investigator grant for study, it is most desirable to provide investigators in each country with the same sum on a per patient basis. Occasionally, slight

differences may be necessary because of the individual needs of certain institutions. Early in the study planning, the CMS should develop a budget concept based on the requirements of the protocol and estimation of their current costs; this should constitute the framework for grant discussions with individual investigators. The contracts between sponsor and investigator within each country should be identical as far as possible.

There are certain costs associated with multinational trials that are sometimes subtle and not readily apparent beforehand. Included among these are the following:

- Local budget requirements which, in addition to grant support, must cover travel and other expenses of investigators when attending investigator and other meetings, and costs associated with logistical support of preparing and shipping any special biological specimens which may be required per protocol.
- Expenses incurred in the organizing of international activities, such as planning meetings, expert and investigator meetings, and costs of supporting very active extramural efficacy and safety review committees.

The costs for these functions must be planned for early and built into any overall budget for the trial.

DISCUSSION

The many advantages of a multinational clinical trial would seem to outweigh the potential disadvantages. The pros and cons for such an undertaking are listed in Table 1. Many clinical researchers feel that a multinational clinical trial should only be performed if it is not possible to successfully perform the trial in one country within an acceptable time frame. There are exceptions to this rule—for example, when it is an explicit objective of the trial to have investigators from a variety of countries gain experience with a new therapy, or if it is the intention of the sponsor to obtain data on drug efficacy and safety in a more internationally diverse cohort of patients.

Countries considered for participation in any multinational clinical trial should be carefully selected with the aim of minimizing differences in medical culture and practice. In addition, the resources and capacity of the sponsor's local medical departments must be carefully assessed.

Special attention must be paid to the appointment of the study coordinator who will be one the key individuals impacting on the quality and success of the trial. The coordinator should have clearly defined functions and responsibilities and be given sufficient authority and support by management to effectively carry out his/her assignment. The coordinator should be made responsible for the coordination of all tasks within the company and with participating country representatives who should functionally report to the coordinator regarding trial activities. The coordinator

Table 1 PROS/Requirements and CONS/Drawbacks for Multinational Clinical Trials

PROS/requirements

- Large sample size
- Increased speed of recruitment
- Broad acceptability of study results
- Broad impact on physician experience
- Increased access to patients with rare diseases
- Ample resources
- Simple trial design
- Simple, brief data collection sheet (CRF)

CONS/drawbacks

- Differences in
- Culture
- Medical training and practice
- Classification of diseases
- Prevalence of diseases
- Genetic Background of patients
- Therapy
- Language
- Diet, customs, religion
- Regulations (e.g., use of placebo)
- Writing numbers, dates and abbreviations
- Using central laboratories
- Interpretation of adverse events

must interact closely with the International Clinical Project Manager and usually with one or two senior and respected investigators who will play a leadership role in the trial.

The tasks and responsibilities of the individuals or committees described earlier—such as protocol development, drug supplies, regulatory, CQA, data management/statistics, safety, extramural review, data analysis, and manuscript preparation—must be clearly described and promulgated at an early stage of study preparation.

The study protocol and CRF should be made as clear as possible in order to prevent opportunities for misinterpretation which can easily occur given the different perspectives that investigators from various countries bring to the trial. This is a demanding task and starts with the individual who drafts the initial documents and proceeds to the next group of individuals to review them. As a general rule, before the documents progress to another level of review, there should be complete understanding and agreement among those who have already completed their review and provided input. This process may be tedious and time consuming but following this course will greatly assist in generating a final draft that is clear and unambiguous. Thus by the time investigators receive the protocol and CRF, the study concept and design have been thoroughly examined and sharply defined.

It is of paramount importance that input for both protocol and CRF be provided by the medical and biometrics representatives from all participating countries. These individuals bring not only their medical and biometric expertise to the subject, but also their international perspective, which will make both documents—particularly the CRF—more universal and user-friendly.

Thorough training of the intramural study team with respect to the protocol and CRF must be carried out. Any complicated or new procedures called for in the protocol should be explained by in-house experts so that staff are reasonably familiar with these before interfacing with the extramural investigator team. Training of the investigator teams should only be given after the intramural staff are thoroughly knowledgeable about both the protocol and CRF.

Occasionally, special medical/surgical procedures, which may not be standard to all centers, must be carried out as key components of a trial. These may also be the subject of special training seminars given by consulting experts. Training in these procedures will assist in assuring that all investigators bring to the study a high level of competence in these key procedures. This competence across study centers and countries will do much to minimize the center or country effects which must be statistically analyzed at the end of the trial.

Procedures for quality assurance should also be agreed upon in advance. For example, compliance with protocol, procurement of informed consent, accurate completion of CRFs, reporting of all serious or unexpected AEs, and availability of source data for review by study monitors must be assured. The sponsor may identify a single individual who will be responsible for coordinating all CQA activities with the local CQA representatives.

Early during patient enrollment, the trial coordinator with other clinical colleagues will team up with the biometrics group to define the approach to data analysis and table set generation. This is a very challenging and time-consuming task, but a satisfying one because it begins to address the outcome of the trial and therefore the fruits of one's labor.

Considering the high cost of a multinational trial compared to the cost of a multicenter trial performed in one country—in terms of translations, struggles to harmonize, logistics, communications, organizing of meetings, and travel—once a multinational trial is initiated the sponsor should make every effort to achieve successful completion. It should be borne in mind, however, that even if all of the preparatory and organizational aspects of a trial are performed perfectly, for the trial to succeed, it must have the full commitment and motivation of all of the investigators. Although the scientific task itself may be motivating the investigators, commitment must be sustained throughout the study. This may be encouraged by meaningful communications among the members of the study group particularly on such aspects as study status, events, actions by individual committees, results (even before the end of the trial, without code breaking, for example on the demographics

or other features of the patient population under study), and scientific updates in the relevant therapeutic areas. Such communications can be achieved through monitoring contacts, newsletters, and importantly, through investigator meetings. These communications should not be restricted to principle investigators alone but should include assistants and other pertinent supporting research staff.

CONCLUSION

The increased use of multinational clinical trials in Phase IIB and Phase III of clinical drug development or following the introduction of a new drug demonstrates that the advantages of access to a larger number of patients and usually a more rapid recruitment, and earlier completion of the trial, and the experience gained from use of the drug in differing patient populations, outweigh the disadvantages or drawbacks of varying medical practice and culture and the increased costs associated with the trial. The sometimes perceived glamor of such an undertaking should not distract from the careful monitoring and meticulous attention that must be devoted to detail in the overall management of the trial.

It should be noted that there are still many scientific/practical issues associated with multinational clinical trials remain unsolved. These issues include, but are not limited to, translation/back translation of the informed consent form and case report forms, sample size requirements in specific countries, the use of central laboratory, and possible treatment-by-country interaction due to difference in ethnic factor.^[4,5]

ACKNOWLEDGEMENTS

Minor updates have been made by the Editor-in-Chief in order to reflect developments since publication of the *Encyclopedia of Biopharmaceutical Statistics, Third Edition*.

REFERENCES

1. Modan, B. Logistics of a multinational population study. *Biomed. Pharmacother.* **1987**, *41*, 373–376.
2. Spilker, B. National versus multinational drug development and clinical trials. *Drug News Perspect.* **1990**, *3* (8), 469–474.
3. Philipp, E.; Weihrach, T.R. Multinational drug development and clinical research: A bird's eye view of principle and practice. *Drug Inf. J.* **1993**, *27*, 1121–1132.
4. Suwelack, D.; Weihrach, T.R. Practical issues in design and management of multinational trials. *Drug Inf. J.* **1992**, *26*, 371–378.
5. Chow, S.C.; Hsiao, C.F. Bridging diversity: Extrapolating foreign data to a new region. *Pharm. Med.* **2010**, *24*, 349–362.
6. Liu, J.P.; Chow, S.C.; Hsiao, C.F. (Eds) *Design and Analysis of Bridging Studies*. Taylor & Francis: New York, 2012.

Multiple Comparisons

Mani Y. Lakshminarayanan

Biostatistics and Research Decision Sciences, Merck & Co., Inc., Rahway, New Jersey, U.S.A.

Abstract

Efficacy and safety claims of an experimental drug are never based on one endpoint, but on several endpoints. This entry reviews the importance of multiple comparisons among treatments in clinical research.

INTRODUCTION

Generally, a phase II/III trial involves collecting data for evidence to show the efficacy and safety of an experimental drug. The integrity of this evidence relies heavily on such factors as the design, the patients recruited, the overall conduct of the study, and the appropriateness of the endpoints and the analysis. Descriptions of these factors must be clearly specified a priori in the protocol, as required by the regulatory guidelines.

Efficacy and safety claims of an experimental drug are never based on one endpoint, but on several endpoints. For example, in a trial involving patients with schizophrenia, a set of efficacy endpoints may include more than one derived measure from the Positive and Negative Symptom Scales (PANSS) and Clinical Global Assessment Scales (CGI-severity and CGI-improvement). Also, for most of the trials, a safety claim is based on adverse events, vital signs, laboratory measurements, and electrocardiogram. In addition to several endpoints, most of the phase II/III trials are typically comparative trials—comparison against placebo or an active therapy, or dose-response trials that involve more than one dose of the experimental drug. There is also a time factor that enters into the overall equation; viz., patient data for various endpoints and doses are collected across several time points in a longitudinal fashion. It is now not difficult to imagine the dimensionality of the accumulated data collected with one overall objective of showing that the drug is either efficacious or safe or both.

OVERVIEW

Statisticians play an important role in assuring their clients—the clinicians—that the design and the methods chosen are appropriate and most sensitive to the primary questions that are asked in the protocol. If the trial is an exploratory one, then it is typical to present and interpret the data using data summaries such as means and standard

deviations, without straying into any inferential procedures. For a confirmatory trial, final conclusion about the success or failure of the study is generally based on statistical inference and a presentation of the overall p -value. With the unavoidable multiplicity in a trial such as just described, if the overall conclusion is based on several comparisons, it can then result in the presentation of several p -values from various statistical tests. There is always some chance that not all of these tests would result in a statistical significance, typically considered as $p \leq 0.05$.

Suppose, in a confirmatory pivotal trial with three doses of an experimental drug and a placebo with one primary endpoint, that out of three pairwise comparisons with placebo, only two yield significant results. For this example, suppose we would like to make an overall claim that the trial has proven the efficacy of the drug. Regulators (such as the U.S. Food and Drug Administration, FDA) may not accept as evidence of efficacy, especially if it is not specified a priori, which comparison is more important than the other. From their perspective, regulators are not comfortable with an inefficacious drug reaching the market, compared with an efficacious drug not reaching the market. That is, a false-positive result is more of an issue to the regulators; this idea will be covered further in the section “Error Rates and Need for Multiple Comparisons.”

ERROR RATES AND NEED FOR MULTIPLE COMPARISONS

When there is more than one treatment, an overall F -test can be performed to determine whether there is any difference in the treatments. If the overall F -test is significant, then individual comparisons between the treatments, generally pairwise comparisons, are carried out to investigate which treatment is different from what. At this point, it must be made clear that considering these pairwise comparisons only if the overall F -test is significant might be erroneous.^[1] Suppose, in a trial involving three doses of an

experimental drug (low, medium, and high) and a placebo, that there are six possible pairwise comparisons in this trial (or experiment). In general, if there are k treatments, then there are $k(k-1)/2$ possible pairwise comparisons.

A *comparisonwise error rate* (CWE) is a type I error rate for each comparisons; that is, it is the probability of erroneously rejecting the null hypothesis between treatments involved in the comparison. On the other hand, an *experimentwise error rate* (EWE, or *familywise error rate*, *FWER*) is the error rate associated with one or more type I errors for all comparisons included in the experiment. When multiple statistical tests are performed using the same data set, the *EWE* can be much larger than the significance level associated with each test itself. For illustration, assume that $P\{\text{type I}\} = \alpha$ for each test or comparison. Then, for k comparisons,

- $\text{CWE} = \alpha$
- $\text{EWE} = 1 - (1 - \alpha)^k$

To illustrate the inflation of the *EWE*, these rates can be expressed in numbers as (with $\alpha = 0.05$):

k	CWE	EWE
1	0.05	0.05
5	0.05	0.25
10	0.05	0.40
20	0.05	0.64
50	0.05	0.92

As can be seen in the table, even with $k = 5$ comparisons, the *EWE* increases to 25%, which may make the overall conclusion based on these $k = 5$ comparisons questionable. A property that is generally required is the strong control of the *FWER*, that is, the probability of making one or more errors on the whole of the considered hypotheses Marcus et al.^[2] On the other hand, a weak control of the *FWER* means simply controlling the α for the global test. Although the latter is a more lenient control, it does not allow the selection of active variables because it simply produces a global p -value that does not allow interesting hypotheses to be selected, so the former is usually preferred because it makes inference on each (univariate) hypothesis.

Several procedures have been suggested for controlling *FWER*. Some authors have grouped the most commonly used *FWER* multiple comparison methods into three categories: (i) single step methods (e.g., Bonferroni correction), (ii) step-down methods (e.g., Holm's method), and (iii) step-up methods (e.g., Hochberg's method).

An alternative approach to multiplicity control is given by the False Discovery Rate (FDR). This is the maximum proportion of type I errors in the set of elementary hypotheses. The *FWE* guarantees a more severe control than the *FDR*, which in fact only controls the *FWE* in the case of global null hypotheses, that is, when all involved hypotheses are under the null.^[3]

Preference of choosing *FDR* over *FWER* for controlling multiplicity depends on the experimental hypotheses. In confirmatory studies, it is usually better to strongly control the *FWER*, which will guarantee an adequate inference when it is preferable not to commit even a single error. In the case of exploratory studies or experiments involving large number of hypotheses (e.g., micro-array data), the aim is to highlight a pattern of potentially involved variables. In such cases, requiring that not a single error be made may be too stringent, and as a result, *FDR* would appear to be a more reasonable approach. On the other hand, Huang and Hsu^[4] provide arguments explaining why controlling *FDR* is inappropriate in clinical trials.

More recently, Storey et al.^[5] have developed a new procedure called the Optimal Discovery Procedure (ODP) to deal with simultaneous significance testing. The Neyman-Pearson lemma has been a standard in providing a simple rule for optimally testing a single hypothesis when the null and alternative hypotheses are known. The majority of the research involving strategies extending a single testing procedure to multiple tests tends to be based on p -values that are calculated from individual tests as a mean to control *FWER* or *FDR*. The ODP method proposed by Storey provides a platform where one can borrow strength from multiple tests to improve the performance of simultaneous testing. This is very similar to shrinkage estimation in which one borrows strength across point estimates to improve the overall performance.

Descriptions of some of the multiple comparisons procedures as well as the control of *FDR* are provided below:

DESCRIPTION

P-Value Based Procedures

The p -value based procedures can be classified into single-step procedures and stepwise procedures. Single-step procedures (or global test procedures)—for example, procedures based on Bonferroni or Sidak inequality and their modification—test each null hypothesis of interest independently without referring to other hypotheses, and thus the order in which the null hypotheses are examined is not important. In contrast, stepwise procedures, such as the Holm procedure, the Hochberg procedure, and the Hommel procedure, test each hypothesis separately, and the decision to reject or accept is made by implication in a sequential manner. It is generally true that stepwise procedures are more powerful than single-step procedures in the sense that they reject more null hypotheses without inflating the *FWER*.

Global Test Procedures

Suppose the trial objective is to determine the overall treatment effect on a primary endpoint. In this situation, it is

appropriate to apply a global test procedure (as described below), even though other procedures (such as step-down approach) can also be applied on an endpoint basis. Some of the global test procedures include likelihood ratio (LR) test, Ordinary Least Squares (OLS) and Generalized Least Squares (GLS) tests, and Bonferroni and Simes approaches. When there are multiple groups (e.g., treatments), at the conclusion of a global test it is imperative to perform all-pairwise comparisons (described below).

All Pairwise Comparisons

Assume that there are k treatment means (or LSMEANS in SASTM terminology), $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_k$, and let s^2 be an estimate of the variance, typically obtained from an analysis of variance (ANOVA) table. In the case of an unbalanced one-way model, with n_i observations for treatment i ,

$$\bar{y}_i = \frac{\sum y_{ij}}{n_i}$$

$$s^2 = \frac{\sum \sum (y_{ij} - \bar{y}_i)^2}{\sum (n_i - 1)} \quad (1)$$

where $v = \sum (n_i - 1)$ is the degrees of freedom associated with the estimate s^2 .

Fisher's Least Significant Difference

Fisher's least significant difference (LSD) method is a two-step process. As indicated in section "Error Rates and Need for Multiple Comparisons," one starts with an overall F -test for testing the equality of treatment means, $H_0: \mu_1 = \mu_2 = \dots = \mu_k$. If H_0 is not rejected, then no further steps are taken. Otherwise, this method concludes that the treatment means μ_i and μ_j are different, for every $i \neq j$, if

$$|\bar{y}_i - \bar{y}_j| > t_{\alpha/2}(v) \left[s^2 \left(n_i^{-1} + n_j^{-1} \right) \right]^{1/2} \quad (2)$$

where $t_{\alpha/2}(v)$ denotes a critical value for the t -distribution having v degrees of freedom and an upper-tail probability of $\alpha/2$.

This two-step Fisher's LSD procedure, also known as the *protected LSD method*, controls the CWE at the level α if the overall null hypothesis is true. If the overall null hypothesis is false, then it does not control the CWE unless $k \leq 3$. The unprotected LSD procedure is used with a significance level that is chosen to control the EWE; that is, the rate used to assert the simultaneous correctness of the comparisons is of no consequence. If a moderate or large number of comparisons ($k > 3$) is to be carried out, then neither of the LSD procedures is recommended, and the Bonferroni method can be considered as an alternative.

Bonferroni Method

The *Bonferroni method* uses the Eq. inequality (2) with a CWE of α/m , where α is the desired EWE and m is the number of pairwise comparisons that are to be carried out within the experiment. This procedure is not recommended if there is a large number of pairwise comparisons, which can render the CWE too small to be of any value as α/m becomes small when m is increased. In such situations, multiple range tests provide some alternatives, some of which are discussed next.

Tukey's Multiple Range Test

The method suggested by Tukey controls the EWE incurred in testing for treatment differences, especially when the averages involved in those differences are based on the same number of observations, that is, when they are taken in pairs. Using this test, we can declare the two means $\bar{y}_i$ and $\bar{y}_j$ to be significantly different if

$$|\bar{y}_i - \bar{y}_j| > q(\alpha, k, v) \left\{ s^2 \frac{(n_i^{-1} + n_j^{-1})}{2} \right\}^{1/2} \quad (3)$$

where $q(\alpha, k, v)$ is the "studentized range statistic," k is the number of averages being compared, s^2 is the mean squared error (MSE) from the ANOVA model with v degrees of freedom, and α is the EWE. Tables of critical values for the studentized range statistic are widely available.

This multiple range test suggested by Tukey can also be used to construct simultaneous confidence intervals on all pairs of mean differences $\mu_i - \mu_j$, and it has been shown that the confidence intervals have a confidence coefficient $100(1 - \alpha)\%$ collectively; that is, 4

$$P \left\{ \mu_i - \mu_j \in \bar{y}_i - \bar{y}_j \pm |q| \left[s^2 \frac{(n_i^{-1} + n_j^{-1})}{2} \right]^{1/2} \text{ for all } i \neq j \right\} = 1 - \alpha \quad (4)$$

Duncan's Multiple Range Test

For comparison of pairs of treatment means, Duncan has developed a test, called a *multiple range test*; its CWE is a function of the number of comparisons. It is a good compromise between Fisher's LSD and Tukey's multiple range test, but it does not control the CWE. To use this test, first the treatment means, obtained from a balanced design with a sample size n , are arranged in ascending order. The criterion is to conclude that the largest and smallest of the treatment means, $\bar{y}_i$ and $\bar{y}_j$, are significantly different if

$$|\bar{y}_i - \bar{y}_j| > q(\alpha_p, p, v) \left\{ \frac{\text{MSE}}{n} \right\}^{1/2} \quad (5)$$

where p = number of averages, $q(\alpha_p, p, v)$ is the critical value from the “studentized range statistic” with an EWE of α_p , and MSE is the mean squared error with v degrees of freedom. When α is the CWE, then α_p is related to α as $\alpha_p = 1 - (1 - \alpha)^{p-1}$.

Holm's method

Holm's^[6] sequentially rejective procedure represents an example of a step-down method. Holm's method improves the sensitivity of Bonferroni's method to detect real differences. The method sequentially contrasts ordered unadjusted p -values with a set of critical values and rejects a null hypothesis if the p -value and each of the smaller p -values are less than their corresponding critical values. The increase in power with Holm's approach is substantial when some null hypotheses are rejected with a high probability while the remaining p -values are compared with less stringent critical values and provides a strong control of the FWER.

Simes Method

With the Simes Method, one would reject the global null hypothesis if $p_{(i)} \leq \alpha/m$, for at least one $i = 1, 2, \dots, m$. The adjusted p -value for the global hypothesis is $p^* = m \cdot \min\{p_{(1)}/1, \dots, p_{(m)}/m\}$. Simes method provides an improvement over Bonferroni approach in controlling the global type I error rate under independence. Recent research in this area indicates that the Simes method does provide satisfactory control of the FWER but can sometimes exceeds α .^[7] Finally, Simes method cannot be used to make inferences on individual hypothesis, since it test only the global hypothesis.

Other Tests

There are other tests available for pairwise comparisons. These are not discussed here, but detailed discussions on these procedures are available in the literature.^[1]

Multiple Comparisons with a Control

In many clinical studies, when there are several treatment groups (e.g., multiple doses of an experimental drug) and a control group, the primary objective specified in the protocol may dictate comparison of the treatment group against the control group, but within-treatment group comparisons may not be required. For example, if the interest is only to show that the new therapy is “better” than no therapy (placebo), then there is no need to make comparisons between the dose groups of the new therapy. Comparisons of such types are known as *multiple comparisons with a control*.

Dunnett's test is the most popularly used multiple comparison procedure for comparison against a control.

Dunnett's Test

Suppose the $k - 1$ treatment groups are given as μ_i , $i = 1, 2, \dots, k - 1$, and the control group is denoted as μ_k . Furthermore, the treatment groups are modeled in terms of a balanced one-way model as

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, 2, \dots, k; \quad j = 1, 2, \dots, n \quad (6)$$

Assume that the error terms ε_{ij} are normally distributed with mean 0 and unknown variance σ^2 . Eq. 1 can now be used to estimate μ_i and σ^2 for this balanced model.

Dunnett's method provides both one-sided and two-sided simultaneous confidence intervals for $\mu_i - \mu_k$.^[1] For the one-sided alternatives, the lower bound of the confidence interval is given as

$$\mu_i - \mu_k > \hat{\mu}_i - \hat{\mu}_k - T \hat{\sigma} \sqrt{\frac{2}{n}}, \quad \text{for } i = 1, 2, \dots, k - 1 \quad (7)$$

where $T = T_{k-1, v}(\{\rho_{ij}\})(\alpha)$ is solved using the following equation:

$$\int_0^\infty \int_{-\infty}^\infty [\Phi(z + \sqrt{2}Tu)]^{k-1} d\Phi(z) \gamma(u) du = 1 - \alpha \quad (8)$$

In Eq. 8, γ is the density of $\hat{\sigma}/\sigma$, Φ is the distribution function of the standard normal, and T depends only on α , k , $v = k(n - 1)$, and $\rho_{ij} = 1/2$ for all $1 \leq i \neq j \leq k - 1$. Also, ρ_{ij} is the correlation coefficient of the estimators $\hat{\delta}_i = \hat{\mu}_i - \hat{\mu}_k$ ($1 \leq i \leq k - 1$). It is widely known^[8] that $T = T_{k-1, v}(\{\rho_{ij}\})(\alpha)$ are the critical points from the distribution of the random variable $\max T_i$, where $T_1, T_2, \dots, T_k$ have a multivariate t -distribution with v degrees of freedom and correlation matrix $\{\rho_{ij}\}$; its tabulated values are widely available for the case $\rho_{ij} = \rho$.

For two-sided simultaneous confidence intervals, $\hat{\mu}_i - \hat{\mu}_k \pm |h| \hat{\sigma} \sqrt{2n}$, the solution to $|h|$ is obtained through the equation

$$\int_0^\infty \int_{-\infty}^\infty [\Phi(z + \sqrt{2}|h|t) - \Phi(z - \sqrt{2}|h|t)]^{k-1} d\Phi(z) \Upsilon(t) dt$$

In other words, $|h|$ are critical points from the distribution of $\max |T_i|$. For the unbalanced case, tabulation of the critical values may be impractical in many cases. But for some special correlation structures, the integration in Eq. 8 can be simplified to produce iterative solutions. For example, in the case of an unbalanced one-way design, $\rho_{ij} = \lambda_i \lambda_j$ ($1 \leq i \neq j \leq k - 1$), where $\lambda_i = (1 + n_k/n_i)^{-1/2}$ for $i = 1, 2, \dots, k - 1$; tables are available for this case.^[8]

Other Tests

Other procedures, such as step-up and step-down procedures, available for multiple comparisons against a control^[1] are not discussed here.

Multiple Comparisons in Dose-Finding Studies

Dose-response studies are performed to evaluate the relationship between the effect (both efficacy and safety) and the dose and to evaluate the recommended dose regimen and frequency. In such studies, several doses of the experimental drug are compared against the zero-dose (or placebo) control. If there is no monotonicity assumed in the effect (that is, group means are not ordered), then Dunnett's test is recommended to compare the group means against placebo. But if the group means are expected to follow an increasing or decreasing order, then there are several procedures available for comparing the doses against placebo. One of the main benefits of some of these procedures is the control of familywise error rate.

Dose-response studies may have multiple endpoints, e.g., primary and secondary endpoints. Under this scenario, inference on the secondary endpoints is performed only after confirming efficacy in the primary endpoint at that dose. This ordering guides the selection of test statistics for each intersection hypothesis in closed testing. Liu et al.^[9] provide a different perspective showing that ordering guides the partitioning of the parameter space in using the partitioning principle to construct multiple tests that control the appropriate FWER.

Closed Test Procedures

Suppose there is a family of hypotheses denoted by $\{H_i, 1 \leq i \leq k\}$ and its closure is represented by taking all nonempty intersections $H_p = \bigcap_{j \in P} H_j$ for $P \subset \{1, 2, \dots, k\}$. A *closed testing method*^[2] asserts that any hypothesis H_p is rejected if and only if every H_Q is rejected by its associated α -level test for all $Q \supset P$, assuming that an α -level test for each hypothesis H_p is available. For tests of this type, Marcus et al.^[2] showed that it controls the type I familywise error rate. In order to use this method, if we have α -level tests individually for each H_p , then they can be applied in a step-down manner. A simple example is provided by a modified Dunnett's test.

Suppose that, in a dose-response study with several doses of an experimental drug and a placebo, it is assumed that the primary objective is to make one-sided comparisons with placebo. For this example, consider a family of hypotheses: $\{H_i : \mu_i - \mu_k \leq 0 (1 \leq i \leq k-1)\}$ against upper one-sided alternatives, where placebo is the k th group. Further, let us assume that the treatment groups have the same size n and that the number of patients in placebo is n_k , and let $\rho = n/(n + n_k)$. The step-down procedure is carried out as follows:

- Calculate T_p , the t -statistics for $1 \leq i \leq k-1$, and let the ordered t -statistics be $T_{(1)} \leq T_{(2)} \leq \dots \leq T_{(k-1)}$ with their corresponding hypotheses denoted as $H_{(1)}, H_{(2)}, \dots, H_{(k-1)}$.
- Reject $H_{(j)}$ if $T_{(i)} > T_{j,v,p}(\alpha)$ for $i = k-1, k-2, \dots, j$. If $H_{(j)}$ is not rejected, then conclude that $H_{(j-1)}, \dots, H_{(1)}$ are also to be retained.

This step-down version of Dunnett's test is more powerful than the test given in Eq. 7, which is based on a single critical value.^[2]

Partitioning Principle

Similar to the closure method described above, the partitioning principle partitions the parameter space into disjoint sets with or without some logical ordering of the hypotheses. Inferential tests are then performed sequentially at different partitioning stages, and the inferential testing stops upon failure to reject a given null hypothesis at the pre-determined partitioning stages. Because the sets of null hypothesis are disjointed at most one set can be true. As a result, even though no multiplicity adjustment is needed, the probability of rejecting at least one true null hypothesis is at most α . Huang and Hsu^[4] have shown that both Holm's method and Hochberg's approach to multiple testing, viewed as step-down and step-up versions of the Bonferroni test, are special cases of testing that uses partitioning principle. They have also shown geometrically that Hochberg's step-up method "cuts corners" off the acceptance regions of Holm's step-down method by making assumptions on the joint distribution of the test statistics.

Inference via the partitioning principle uses the data to decide in which subspaces the true parameter may lie. It has been successfully applied to derive practically useful methods for various areas, including clinical trials. However, the formulation of disjointed sets of hypotheses may be cumbersome when the number of endpoints is moderately large. Moreover, similar to the closure method, confidence intervals on the parameters are generally unavailable.

Other Tests

Other multiple test procedures for dose-finding studies are described in Refs. [10,11]

DESIGN ISSUES

Active Control Studies

In many areas of clinical research (e.g., anti-infectives), standard therapies (positive or active controls) are commonly available in the market. Inclusion of one or more doses of the positive control would serve as reference standard for efficacy comparisons with the experimental

drug, while comparison with placebo would validate the clinical trial. In other words, it would serve to differentiate between an “ineffective” drug vs. an “ineffective” study.

When a trial involves one or more doses of an experimental drug, placebo, and one or more doses of an active control, always the first dilemma is to establish a family of hypotheses a priori in the protocol. Based on the study design and various underlying hypotheses, strategies should be explored for testing various hypotheses. One such set of hypotheses, {drug vs. placebo} and {positive control vs. placebo}, would help to conclude whether both the drug and the positive control are superior to placebo. For each one of these hypotheses, multiple comparison procedures would then have to be proposed in the protocol, depending on the objective of either controlling *comparisonwise* or *familywise* error rate.^[12]

Sample Size

As study designs in most of the clinical trials involve an ANOVA or an analysis of covariance (ANCOVA) model, it is not uncommon to perform sample size calculations based on an overall *F*-test, which may not be appropriate if the primary objective involves multiple comparisons. When multiple comparisons are involved, another approach used is to adjust the Type I error level with a Bonferroni correction. Again, this may lead to more patients than what is actually required. Hsu^[13] suggests a confidence interval approach as follows: Given a confidence interval approach with level $1 - \alpha$, do the sample size computations so that with a prespecified power $1 - \beta (< 1 - \alpha)$, the confidence intervals will cover the true parameter value and be sufficiently narrow. Technical details for exact sample size computations are given in [Appendix C](#) of Hsu's excellent textbook.^[1] Certain computer software, including SASTM procedures, is also available for sample size computations that require adjustment for multiple comparisons and multiple tests.^[14]

CONCLUSION

Multiple comparisons are commonly encountered in clinical research. Multiple comparisons may involve pairwise comparisons among treatments or multiple comparisons with a control when there are more than two treatments in the study. In this case, statistical methods for controlling error rates such as CWE or EWE or FDR are necessary for multiple comparisons. In practice, closed test procedures are useful for addressing multiple comparisons in dose-finding studies. If there are large number of tests are involved (e.g., safety data), then it would be worthwhile to investigate controlling the false discovery rate (FDR).

It should be noted that FDA circulated a draft guidance on multiple endpoints in clinical trials for comments in

January 2017.¹⁵ This draft guidance, when finalized, will represent FDA's current thinking about the problems posed by multiple endpoints in the analysis and interpretation of study results and how these problems can be managed in clinical trials for human drugs, including drugs subject to licensing as biological products.

ACKNOWLEDGEMENTS

Minor updates have been made by the Editor-in-Chief in order to reflect developments since publication of the *Encyclopedia of Biopharmaceutical Statistics, Third Edition*.

REFERENCES

1. Hsu, J.C. *Multiple Comparisons—Theory and Methods*; Chapman & Hall: London, 1996.
2. Marcus, R.; Peritz, E.; Gabriel, K.R. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **1976**, *63*, 655–660.
3. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A new and powerful approach to multiple testing. *J Royal Stat Soc, Ser B.* **1995**, *57*, 1289–1300.
4. Huang, Y.; Hsu, Jason C. Hochberg's step-up method: cutting corners off Holm's step-down method. *Biometrika*. **2007**, *94*, 965–975.
5. Storey, J.D.; Dai, J.Y.; Leek, J.T. The optimal discovery procedure for large scale significance testing, with applications to comparative microarray experiments. *Biometrics* **2007**, *8*, 414–432.
6. Holm, S. A simple sequentially rejective multiple test procedure. *Scand J Stat.* **1979**, *6*, 65–70.
7. Sarkar, S.; Chang, C.K. Simes method for multiple hypothesis testing with positively dependent test statistics. *J. Am. Stat. Assoc.* **1997**, *91*, 1601–1608.
8. Hochberg, Y.; Tamhane, A.C. *Multiple Comparison Procedures*; Wiley & Sons: New York, 1987.
9. Liu, Y.; Hsu, J.C.; Ruberg, S. Partition testing in dose-response studies with multiple endpoints. *Pharm. Stat.* **2007**, *61*, 181–192.
10. Tamhane, A.C.; Hochberg, Y.; Dunnett, C.W. Multiple test procedures for dose finding. *Biometrics* **1996**, *52*, 21–37.
11. Chow, S.C.; Liu, J.P. *Design and Analysis of Animal Studies in Pharmaceutical Development*; Marcel Dekker, Inc.: New York, 1998.
12. D'Agostino, R.B.; Massaro, J.; Kwan, H.; Cabral, H. Strategies for dealing with multiple treatment comparisons in confirmatory clinical trials. *Drug Inf. J.* **1993**, *27*, 625–641.
13. Hsu, J.C. Sample size computation for designing multiple comparison experiments. *Comput. Stat. Data Anal.* **1998**, *7*, 79–91.
14. Westfall, P.H.; Tobias, R.D.; Rom, D.; Wolfinger, R.D.; Hochberg, Y. *Multiple Comparisons and Multiple Tests Using the SAS System*; SAS Institute: Cary, NC, 1999.
15. FDA *Guidance for Industry, Multiple Endpoints in Clinical Trials*. Food and Drug Administration, Silver Spring, MD, January, 2017.

Multiple Endpoints

Mario Comelli

Università di Pavia, Pavia, Italy

Abstract

The aim of this entry is to give a synthetic account of several actually known methods of analysis of multiple endpoints.

INTRODUCTION

A common problem arising in medical research is that of comparing some groups of patients (for instance, the arms of a clinical trial) based on multiple outcome measures called *endpoints*, all of which must be considered of primary importance and have no hierarchical structure. The aim of this entry is to give a synthetic account of several actually known methods of analysis of multiple endpoints. Each method is described with its merits and the situations where it is most appropriate. The procedures of application will be detailed wherever possible. However, sometimes the reader will be referred to the original article for the steps of computation. Most of the published material on the topic concerns the comparison of two groups. So for most of this entry, only the comparison of two arms of a study will be considered. They will be identified as experimental and control. Consistent with the informative nature of this entry, mathematical formalism will be kept to a minimum.

In the analysis of a multiple endpoint study, two types of questions are possible:

1. Are there any endpoints for which the experimental treatment is more effective than the control (or vice versa)? Which are these endpoints?
2. Does the combined evidence of all endpoints support the global superiority of the experimental treatment, even if no clear conclusions can be drawn for any determined endpoint?

By answering the first question, one can often answer the second question as well. Indeed, if one proves that the experimental treatment is more effective than the control for determined endpoints, and no less effective for the others, one has also proved that the treatment is globally superior to the control. The reverse, however, is not true.

NOTATION

C	is the control treatment
T	is the experimental treatment.
k	is the number of endpoints.
n_T and n_C	are the numbers of subjects in the T and C samples, respectively.
S_i or s_i	is a test statistic that compares the effect of T and C on the i th endpoint. (Uppercase indicates a random variable, lowercase an observed value.)
P_i (or p_i)	is the unadjusted P -value of the univariate test on the i th endpoint. (Uppercase indicates a random variable, lowercase an observed value.) P_i does not take the multiplicity of endpoints into account. Therefore P_i (or p_i) is called <i>raw</i> or <i>nominal</i> .
P_{ia} (or p_{ia})	is the P -value of the test on the i th endpoint adjusted by taking into account the multiplicity of endpoints. Therefore P_{ia} (or p_{ia}) is called <i>adjusted</i> .
S or s	is a test statistic that compares the effect of T and C simultaneously on all endpoints.
P or p	is the P -value of the global comparison, testing the global difference of the effect of T and C on all endpoints, without identifying the endpoints where the difference is “significant.”
H_0	is the global null hypothesis. It states that the difference of effect to T and C is zero on all endpoints.
H_{0i}	is the null hypothesis that the effect of T and C on the i th endpoint is the same.
$\sim H_{0i}$	is an alternative hypothesis to H_{0i} and to H_0 . It states that the effect of T and C on the i th endpoint is not the same; it is the complement of H_{0i} .
m	is the number of true H_{0i} in the trial.
$\mathcal{P}(E)$	is the probability of an event E .

$\bar{E}$	is the complementary event of E .
FWER	is the familywise error rate in the strong sense. It is the probability of rejecting one (or more) true null hypotheses when a mixed set (a family) of true and false null hypotheses is tested. In other words, it is the probability of making at least one Type I error in the whole set of comparisons.
α_c	is the Type I error probability required for each comparison to obtain $\text{FWER} = \alpha$.

If not otherwise specified, all P -values are taken two-sided.

IDENTIFICATION OF THE ENDPOINTS WHERE T IS MORE EFFECTIVE THAN C

Identifying the endpoints affected by T is a multiple comparison problem. It requires the test of one hypothesis for each endpoint. Because of the many hypotheses tested, the main difficulty is to avoid the multiple testing bias without losing too much power. There are different types of methods to attain this aim. Some of them compute adjusted P -values using the p_i s obtained via the univariate test (even if the tests are different for different endpoints). These methods can be also used by the reader of a scientific paper reporting p_i s. Their drawback is that they are often difficult to apply to the computation of confidence intervals or of the required sample size of a clinical trial.

Some other procedures use the distribution or at least the correlation matrix of the endpoints under H_0 . They either postulate a known distribution or empirically estimate it. This type of technique has some specific developments for binary endpoints.

P -Value–Based Techniques

The simplest technique of this type is the *Bonferroni correction*, which consists of multiplying all p_i s by k . The p_{ia} s so obtained are then compared to the chosen Type I error probability: α (e.g., 0.05). This technique keeps $\text{FWER} \leq \alpha$. Therefore from the p_{ia} s, one can draw inferences concerning each single endpoint exactly as if it were the only one tested. The Bonferroni method is based on the Bonferroni probability inequality,^[1] which is true for every set of k events E_j :

$$1 - \mathcal{P}\left(\bigcap_{j=1}^k E_j\right) \leq \sum_{j=1}^k \mathcal{P}(\bar{E}_j) \quad (1)$$

Let us consider a family of k endpoints. Let E_j be the event that the test statistic S_j of the j th endpoint assumes a more probable (or with a higher probability density) value than a certain s_j having P -value p . Then, the left-hand side is the probability of obtaining, at least once, a P -value equal to or lower than p in the whole set of k comparisons.

Let $\bar{E}_j$ be the complement of E_j . The P -value p of the statistic s_j , by definition, is the probability of obtaining a value of S_j equally probable to or less probable (or with an equal or lower density) than s_j if H_{0j} is true. Therefore, if H_{0j} is true, $\mathcal{P}(\bar{E}_j) = p$. Moreover, if there are $m \leq k$ true H_{0j} s, then the probability of all the corresponding events $\bar{E}_j$ is equal to p . Therefore applying the Bonferroni inequality only to the endpoints unaffected by the treatment, one obtains

$$1 - \mathcal{P}\left(\bigcap_{j=1}^m E_j\right) \leq mp \leq kp \quad (2)$$

Now one can assume p as equal to the observed raw P -value p_i of a certain H_{0i} . Subsequently, the inequality states that the probability of obtaining once or more, for the set of m true H_{0j} s, a test statistic S_j equally probable to or less probable (or with an equal or lower density) than a value s_j having P -value p_i is lower than kp_i . By a slight generalization of the definition of P -value, we call that quantity p_{ia} , i.e., the P -value of the i th endpoint, computed by taking into account the error probability of all the comparisons. In other words, p_{ia} is the P -value adjusted for multiplicity.

Another very simple formula for the computation of adjusted p_{ia} s is obtained under the condition that the endpoints are stochastically independent. It can be derived as follows: The probability of obtaining, for all P_j s of the m true H_{0j} s, values larger than p_i is $(1 - p_i)^m$, under independence. Therefore the probability of obtaining for at least one P_j corresponding to a true H_{0j} a value equal to or smaller than p_i is

$$1 - (1 - p_i)^m \leq 1 - (1 - p_i)^k \quad (3)$$

So one can obtain $p_{ia} = 1 - (1 - p_i)^k$.

Many improvements to the Bonferroni correction to test single endpoints have been proposed; they obtain a greater power by taking into account all the p_i 's, instead of just the one to correct.^[2–4] Some of these are very simple to use. The general logic of Holm's step-down technique^[2] is as follows: Put the p_i s in increasing order and adjust the smallest via the Bonferroni correction. If the p_{ia} so obtained is larger than the chosen Type I error probability α , retain all H_{0i} s. If, on the contrary, its value is smaller than α , then reject the corresponding H_{0i} . This H_{0i} is then to be excluded from the set of possible true null hypotheses. Therefore adjust the second smallest p_i by multiplying it by the correction factor $k - 1$. Indeed, recall from Eq. (2) that to adjust p_i , it is enough to multiply it by the number of true H_{0j} s; now, after the rejection of one hypothesis, we know that the number of true H_{0j} s is not larger than $k - 1$. If the second p_{ia} is larger than α , then retain the corresponding hypothesis and all those having p_i s. If it is equal or smaller than α , then reject the corresponding H_{0i} , and go on with the same logic to compute the third p_{ia} by multiplying the third smallest p_i by $(k - 2)$. Compare

the third p_{ia} with α , and so on. The procedure stops when a p_{ia} turns out to be nonsignificant: then retain the H_{oi} under scrutiny and the remaining H_{oi} s.

Hochberg's step-up procedure^[4] theoretically requires the stochastic independence of the test statistics used to detect the effect of T on different endpoints. However, simulation results^[5,6] confirm that it is conservative, i.e., it keeps the FWER smaller than the declared α , also when the test statistics are positively correlated. Hochberg and Rom^[7] study its behavior when the p_i s are obtained from Z-tests (the simplest test for Gaussian endpoints with known variance). They theoretically prove that Hochberg's procedure is conservative when the test statistics are either positively correlated or examined by two-sided tests. When Z values are negatively correlated and the tests are one-sided with $\alpha = 0.05$, the theoretical upper bound of FWER is 0.0525.

The derivation of Hochberg's method is rather complex and is explained in three papers.^[3,4,8] However, its application is simple: arrange the p_i s in decreasing order. If the largest p_i is equal to or smaller than the required α , all the H_{oi} s are rejected. Otherwise, multiply the second largest p_i by 2: If the product is equal to or smaller than α , reject the corresponding H_{oi} and all the H_{oi} s with smaller p_i s. Otherwise, retain the corresponding H_{oi} , multiply the third largest p_i by 3, and compare it with α . If it is equal to or smaller than α , reject the corresponding H_{oi} and all the H_{oi} s with smaller p_i s. If the value is larger, then retain the corresponding H_{oi} , and so on. The procedure continues until an H_{oi} is rejected; then all H_{oi} s with a smaller p_i than the first rejected H_{oi} are rejected as well. The p_{ia} of each endpoint is obtained from the corresponding raw p_i multiplied by its rank (in decreasing order). If this product is larger than the p_{ia} computed from the proceeding p_i , it is substituted by the preceding p_{ia} . This is carried out to account for the nonincreasing order of the p_{ia} s.

Hochberg's technique is always more powerful than Holm's. For an example of its application to a real clinical trial data as compared to other adjustment techniques, see Ref. [9].

Hochberg and Benjamini^[10] improved Hochberg's method using an estimate of the number of the true H_{oi} s in the trial. The value of is obtained by applying the graphical technique of Schweder and Spjøtvoll^[11] (described in the section "Inferring the Global Superiority of the Treatment from the Combined Evidence of All Endpoints") to two-sided p_i s. At the beginning of the procedure, arrange the p_i s in decreasing order and retain all H_{oi} s having a p_i larger than α . Let a be the number of accepted H_{oi} s. Then multiply the largest p_i of the H_{oi} s not retained by the smaller of the values $a + 1$ and $\hat{m}$. If the product is smaller than α , reject the corresponding H_{oi} and all the ones with smaller p_i s. Otherwise, retain the corresponding H_{oi} and multiply the next p_i by the smaller of the values $a + 2$ and $\hat{m}$. If the product is smaller than α , reject the corresponding H_{oi} and all the H_{oi} s with smaller p_i s. Otherwise, retain the

concerned H_{oi} and multiply the next p_i by the smaller of the values $a + 3$ and $\hat{m}$, and so on. The process stops when a hypothesis is rejected; subsequently, all the succeeding ones are also rejected.

The reason this procedure is considered as an improvement of Hochberg's technique is that one never needs to multiply a p_i by a correction factor larger than the number of true H_{oi} s. The p_{ia} s can be defined in the following way: Each p_{ia} is the product that would test the corresponding H_{oi} , if hypothetically there would be no rejections until the smallest p_i . Whenever such product is determined to be larger than the one preceding it in the testing procedure, the corresponding p_{ia} is assumed to be equal to the previous one.

Brown and Russell^[5] studied both the FWER control and the power of 17 different P -value-based adjustment techniques (which includes nine step-down or step-up procedures and eight methods based on the parametric or nonparametric estimation of the distribution function of the P_i s). They simulated two different numbers of endpoints, three patterns of correlation among endpoints, three distributions of truth and falseness among the H_{oi} s, and two different amounts of deviation from the null hypotheses. Each statistical technique was applied 10,000 times in each of the 36 simulated situations, with $\alpha = 0.05$. The authors found that the improved Hochberg's procedure (see earlier discussion)^[10] falsely rejects any true H_{oi} at most 5.16% of the time. As for power, that method is consistently superior to seven other techniques in all experimental situations. No other examined procedure approaches this level of power achievement. Thus concluding their review, the authors recommended the improved Hochberg's step-up procedure over other P -value-based techniques.

Adjusted confidence intervals of outcome corresponding to step-down and step-up procedures are generally not easily available. Indeed, the estimated number of true null hypotheses is not immediately useful for confidence interval computation. Some particular results are worked out in Hayter and Hsu.^[12]

It is difficult to imagine the algorithm to compute the optimal sample size of a future trial to be analyzed by the step-up or step-down procedures. However, if one plans to use the improved Hochberg method, a reasonable guess could be made in the following way. From prior knowledge, one can obtain an estimate $\hat{m}$ of the number of true H_{oi} s in the future trial (it should be less than $k/2$; otherwise, the choice of the primary endpoints of the trial needs to be reconsidered). First, one must identify the endpoint on which the expected standardized effect of T , if present, is the smallest. Then one should determine the minimum clinically relevant effect of T on this endpoint (it should be less than the expected effect; otherwise, the endpoint should not have been chosen as a primary one). The next step would be to compute the required sample size to achieve the desired power (e.g., 0.8), as if the analysis were made on just this endpoint with a Type I error probability of $\alpha/\hat{m}$ (e.g., 0.05/ $\hat{m}$).

Techniques Based on the Distribution or the Correlation Structure of the Endpoint

A method based on the estimation of the correlation between endpoints is the one proposed by Tamhane.^[13] The adjusted p_i of each endpoint is computed by the following equation:

$$p_{ia} = 1 - (1 - p_i)^{k(1-\bar{p})} \quad (4)$$

where k is the number of endpoints and $\bar{p}$ is their average sample correlation (this can be computed from the sample correlation matrix of the endpoints, which is produced by any standard statistical program). The rationale behind the equation is largely commonsense: when the endpoints are completely independent, $\bar{p} = 0$, and Tamhane's formula turns out to be the simple correction Eq. (3) for independent tests. However, when all the correlations are 1—that is, all the endpoints are really just one endpoint (but for additive and multiplicative constants)—then the equation fittingly does not correct the raw p_i s at all. Eq. (4) has the advantage of simplicity, but it is not supported by any theoretical evidence.

Sankoh et al.^[6] empirically studied the performance of two similar correction equations, based on the estimation of the correlation between endpoints. Sankoh's simulations show that they are rather liberal. For both equations, the percentage of false significant results can be as high as 12% when the declared significance level of the procedure is 0.05. Sankoh proposes a method to correct their error rate, based on simulations (see later discussion).

Tamhane's formula can be used to construct confidence intervals for the endpoint differences adjusted for multiplicity. The per-interval confidence level, $1 - \alpha_c$, can be determined by the following equation:

$$(1 - \alpha_c)^{k(1-\bar{p})} = 1 - \alpha \quad (5)$$

If each confidence interval is computed at level $1 - \alpha_c$, then the probability of obtaining all intervals to include the true parameters should be $1 - \alpha$. However, this value of FWER could be optimistic, depending on the validity of Tamhane's formula. Empirical validation of the technique is needed.

From Tamhane's formula, one can compute the approximate sample size required for a multiple-endpoint clinical trial: one needs to know, or estimate from prior knowledge, the average correlation between endpoints. An example of sample size computation is as follows: Imagine a trial with eight endpoints, with a foreseen mean correlation of 0.7. Use Eq. (5) to compute the per-comparison significance level α_c required to achieve the desired global α . Compute α_c by solving the following equation:

$$1 - (1 - \alpha_c)^{8(1-0.7)} = 0.05$$

The solution is $\alpha_c = 0.027$. Then compute the sample size required for a level 0.027 test (with the desired power) applied to the endpoint with the smallest expected standardized treatment effect. This should also be the required sample size of the trial.

Tamhane^[13] also describes the *bootstrap* (or resampling) technique, which adjusts the p_i s by empirically estimating their distribution under H_0 . The rationale behind the bootstrap technique goes as follows: The probability distribution of any quantity Q meant to compare the groups of the trial can be estimated from its observed frequency distribution. If the null hypothesis is true for all endpoints, then the T and C samples of the trial differ in no systematic way; they come from the same population. So under H_0 , the two treatment groups can be thought of as random subsamples extracted from the total trial sample. One can artificially extract, with replacement, thousands more couples of random subsamples from the total trial sample; under H_0 , they all belong to the same population as the real data. In those artificial couples of subsamples, one can observe the empirical frequency distribution of the minimum P_i of the endpoints. So the frequency with which the minimum P_i is equal to or lower than a certain p_j obtained in the original data can be observed. It estimates the probability of obtaining such a p_j or a smaller P -value for at least one endpoint if H_0 is true. In other words, it is an estimate of p_{ia} . Pooling the samples does not change the correlation of the endpoints. Therefore the bootstrap adjustment estimates p_{ia} s based on exact correlations. However, unlike the P -value-based techniques, it fails to take into account the complete set of p_i s when adjusting each of them. In other words, the adjustment is made assuming that all H_{0i} s are true, which is probably conservative. Step-down and set-up versions of the procedure are available.^[14–16] The step-down version follows the logic of Holm's method: arrange the p_i s in increasing order. Resample data containing all endpoints to adjust the minimum p_i . Then exclude the corresponding endpoint from the data resampled to adjust the second smallest p_i . Third, also exclude the endpoint corresponding to the second p_i from the resampling, and adjust the third smallest endpoint. Go on until all the endpoints are adjusted. This method has been implemented in SAS, to correct the P -values of the t -test and the Fisher exact test for proportions, among others. It is invoked with the PROC MULTTEST command with the options BOOTSTRAP and STEPDOWN.^[17]

Techniques to Analyze Categorical and Ordinal Endpoints, Based on Discrete Probability Distributions

Some interesting methods based on the distribution of the observations are available to analyze dichotomous multiple endpoints; they are based on the "exact" Fisher test for the equality of two proportions. To use them, this test must be applied to compare the frequencies of *good* outcomes

of the two arms of the trial for each endpoint. It produces discrete P -values, with the consequence that p_i s lower than certain bounds are impossible. If, for a given endpoint i , one acquires a p_i smaller than the minimum possible P -value for every other endpoint, then no correction is needed for that p_i . Indeed, it can arise by chance only in one comparison: the one where it was obtained. Hence no multiple-testing bias exists. If only a subset of q comparisons can generate P -values equal to or smaller than p_i , then the Bonferroni correction should take into account only these q comparisons. Indeed, p_i or a smaller P -value can arise by chance only in these q comparisons but not in the whole set of k tests. Therefore one can accept $p_{ia} = qp_i$. To establish the minimum possible (one-sided) value of P_i for each comparison, the following equation is available:

$$\frac{\binom{g}{n_T}}{\binom{n}{n_T}} \text{ if } g \geq n_T \text{ and } \frac{\binom{n-g}{n_T-g}}{\binom{n}{n_T}} \text{ if } g < n_T \quad (6)$$

where n is the total number of subjects, n_T is the number treated, and g is the number of observations with the *good* outcome.

Following this logic, Mantel^[18] proposed a more precise multiple comparison procedure for the Fisher test. To adjust a p_i , one adds together the largest theoretically possible P -values of all other comparisons, which are equal to or smaller than p_i . Of course, in order to follow this procedure, one has to know all the possible values P_j of each comparison. They are obtained from the hypergeometric probability distribution of g_T : the number of subjects in the treated group experiencing the *good* outcome:

$$\mathcal{P}(g_T) = \frac{\binom{g}{g_T} \binom{n-g}{n_T-g_T}}{\binom{n}{n_T}} \quad (7)$$

Consider the following example. Imagine a trial with five dichotomous endpoints (say) A, B, C, D, E. The observed p_i of endpoint A is 0.019. To correct it, obtain the total number g of subjects with *good* outcome, observed for endpoint B. Compute all its possible P -values by substituting each possible value of g_T into Eq. (7). Do the same for endpoints C, D, and E. Now among all possible P -values computed for endpoints B, C, D, and E, look for those equal to or smaller than 0.019. Imagine that the largest of such values for endpoint B is 0.016 and that the largest for endpoint D is 0.012. Endpoints C and E have no possible P -values smaller than 0.019. Then the adjusted P -value of endpoint A is $0.019 + 0.016 + 0.012 = 0.047$.

This procedure can be used only when the theoretically possible values of P_i of the k comparisons are known. This takes place when small frequencies are analyzed via the Fisher exact test, but it also happens when small data

sets are analyzed by nonparametric techniques, e.g., the Wilcoxon test. Therefore Mantels' adjustment method can also be applied to ordinal data.

The discrete distribution of the endpoints can be exploited together with Hochberg's improved set-up procedure. One could proceed in the following way: Start Hochberg's improved procedure as usual, by putting the p_i s in decreasing order and by retaining all H_{0i} s corresponding to p_i 's $> \alpha$. From the second step onward, multiply the p_i under scrutiny by the minimum of these three numbers: $\hat{m}$ (the graphically estimated number of true H_{0i} s), $a_p + 1$ (the number of retained hypotheses in all preceding steps + 1), and q (the number of comparisons that can generate P -values equal to or smaller than p_i). If the obtained product is larger than α , then retain the corresponding hypothesis and examine the next p_i . Otherwise, reject the corresponding hypothesis and all the succeeding ones. This combined procedure should allow interesting power gains, because q tends to be small when a is large, and vice versa.

False Discovery Rate

An interesting alternative to the usual approach is the one proposed by Benjamini and Hochberg.^[19] They argue that it is sometimes more interesting to control the proportion of false rejections than to keep low the probability of at least one false-positive result. This seems to be a reasonable point of view when some tens of endpoints are to be tested and the researcher has no strong interest in any single one but rather looks for clusters of affected endpoints. The authors define the false-discovery rate as the expected proportion of rejected true H_{0i} s among all rejected H_{0i} s. They describe a test procedure and prove that (under stochastic independence of the endpoints unaffected by T) it warrants an expected false-discovery rate lower than a chosen value α_R (e.g., $\alpha_R = 0.05$). The procedure goes as follows: Arrange the raw p_i s in decreasing order. Compare the largest p_i with α_R : If it is smaller or equal to α_R , then reject all hypotheses. Otherwise, retain the corresponding hypothesis and compare the second largest p_i with $(k-1)\alpha_R/k$. If its value is smaller or equal, reject the corresponding H_{0i} and all the succeeding H_{0i} s. Otherwise, retain the corresponding hypothesis and compare the third largest p_i with $(k-2)\alpha_R/k$, etc. Continue in this manner until either a rejection ensues or the last p_i has been compared to α_R/k . Not surprisingly, the simulations show that, mostly when $k > 10$, this method applied within $\alpha_R = 0.05$ has a much higher power than the techniques that keep the FWER below 0.05.

INFERRING THE GLOBAL SUPERIORITY OF THE TREATMENT FROM THE COMBINED EVIDENCE OF ALL ENDPOINTS

Establishing the global superiority of T can be thought of as a single-hypothesis-testing problem. The multivariate

null hypothesis to be rejected H_0 declares that the true differences between the two arms are zero for *all* endpoints. Many alternative hypotheses are conceivable; it is important to choose the one that is relevant to prove in the trial at hand. The different multivariate alternative hypotheses that we consider are the following.

1. $\sim H_0$ is an alternative to H_0 , which contains and mixes up every possible difference of the effect of T and C . Inequalities of opposite sign for different endpoints are included in $\sim H_0$. It is the complement of H_0 .
2. $H_{T \geq C}$ is an alternative to H_0 , which states that T has a better effect than C on at least one endpoint and has no worse effect on any endpoint.
3. H_A is an alternative hypothesis to H_0 , which states that the average (defined somehow) of the effects of T on all endpoints is better than the average effect of C .

All these hypotheses state that the effects of T and C are different, but they do not identify the affected endpoints. Thus they are less informative than the alternative hypotheses concerning single endpoints. However, they are proved by just one test, whereas the $\sim H_0$ s are many and require many tests. For that reason, the procedures designed to prove $\sim H$, $H_{T \geq C}$, or H_A are often much more powerful than those used to test the H_0 s. They are to be considered especially in the final stage of the research on a treatment, when one is often more interested in assessing its multidimensional efficacy than in describing accurately the pattern of its clinical effects. For each of the three alternative hypotheses just mentioned, different statistical procedures are recommended.

To use the procedures described next correctly, be sure to measure all endpoints in such a way that positive differences always correspond to a more favorable outcome.

Techniques Powerful in Detecting $\sim H_0$

The alternative hypothesis $\sim H_0$ is rarely an interesting one; the discovery that some endpoints are favorably (and others unfavorably) affected by T , without identifying the concerned endpoints, can rarely have practical consequences. In any case, two procedures can be recommended to prove $\sim H_0$. Hotelling's T^2 is the uniformly most powerful test over $\sim H_0$ (subject to some invariance conditions) when the endpoints are multivariate Gaussian.^[20,21] This means that it detects violations of the null hypothesis, when the truth lies anywhere in the specified alternative hypothesis, with higher probability than every other test. This test is routinely performed by any program of multivariate analysis of variance (such as multiple analysis of variance, MANOVA, of SPSS or GLM of SAS). To apply it, one just has to compare the two arms by such programs, inputting all the endpoints as response variables.

If the endpoints are far from Gaussian, a “discriminant” logistic regression might work better, as suggested

by Cupples et al.^[22] The procedure, which can also be used with categorical endpoints, goes as follows: First, fit a logistic model with the allocated treatment as the *response* variable and no explanatory variables. Then fit another logistic model with the same response variable, with all the endpoints input as *explanatory* variables. The P -value of the multivariate comparison between treatments is equal to the P -value that compares the fit of the two logistic models. An example of a work that applies this technique can be found in Ref. [9]. However, note that, with this method, one cannot include prognostic covariates and stratification factors in the analysis.

Techniques Powerful in Detecting $H_{T \geq C}$

$H_{T \geq C}$ is certainly an interesting hypothesis to prove; however, the significance of the tests (powerful enough to detect it) is not necessarily a proof of it. Indeed, they can reject H_0 with a probability much larger than α , also when $H_{T \geq C}$ is far from true (this happens even if the Type I error rate is correctly kept to α). A typical situation where the tests described in this section reject H_0 when $H_{T \geq C}$ is false is when T has a strong favorable effect on some endpoints and a weaker unfavorable impact on others. Therefore to use the tests described in this paragraph as a proof of $H_{T \geq C}$, one must be sure that the difference of effect between T and C is zero or is favorable to T for all endpoints.

Schweder and Spjøtvoll^[11] proposed a *graphical technique*. This method is not a proper statistical test but allows one to estimate the number of endpoints that are favorably influenced by T . It is based on the raw one-sided p_i s computed on the single endpoints. (In case one-sided p_i s are not directly available, as happens with a χ^2 test, they can be computed in the following way: Observe in the data if the effect of T on the concerned endpoint appears to be favorable—e.g., the frequency of patients free of some discomfort appears to be increased by T .) If it appears as such, then: $p_{i \text{ one-sided}} = p_{i \text{ two-sided}}/2$; otherwise, $p_{i \text{ one-sided}} = 1 - (p_{i \text{ two-sided}}/2)$. The authors propose to plot the $(1 - p_i)$ s vs. their rank for all endpoints. If H_0 is true, then the plot should be a line through the origin with slope equal to $1/(k + 1)$. The reasons for this fact are as follows: From the definition of P -value, it derives that P_i is equal to or smaller than a determined value p_i , with probability equal to p_i . This means that P_i and $1 - P_i$ have a uniform probability distribution between 0 and 1. Now when k values are obtained from a random variable uniformly distributed in the interval 0–1, they are expected to divide the unit interval into $k + 1$ equal subintervals. Thus if one sequentially plots k points having as ordinates the $(1 - p_i)$ s in increasing order and as abscissas the k increasing numbers, which divide the 0–1 interval into $k + 1$ equal subintervals, then one expects to find each ordinate equal to the corresponding abscissa. Hence the interpolating curve is expected to be a line with slope 1. If one changes the scale of the x -axis and extends each subinterval to measure 1, the abscissas

become equal to the natural numbers from 1 to k . Then they are the ranks of the corresponding ordinates, i.e., of the $(1 - p_i)s$. Moreover, the former 0–1 interval extends to measure $k + 1$, so the slope of the line becomes $1/(k + 1)$ (because an increase of $k + 1$ on the x -axis corresponds to an increase of 1 on the y -axis).

If the number of true $H_{0i}s$ is $m < k$, the plot should deviate somewhere from a straight line. Indeed, some $(1 - p_i)s$ should be larger than expected, corresponding to the small p_i s of the false $H_{0i}s$. However, all of the leftmost points of the plot probably correspond to the large p_i s of the true null hypotheses. Hence they should plot a line. The expected slope of this line is $1/(m + 1)$. To understand such a development, remember that if two points of the plot correspond to true $H_{0i}s$, the expected difference of their ordinate is $1/(m + 1)$. Indeed, the $(1 - p_i)s$ that correspond to true $H_{0i}s$ are expected to generate $m + 1$ equal intervals between 0 and 1. Thus if one takes two points that correspond to true $H_{0i}s$ and whose abscissas differ by 1 (two consecutive points of the leftmost part of the plot), then they lie on a line with the expected slope $1/(m + 1)$. This fact gives a clue on how to estimate the number of true $H_{0i}s$: One can fit a straight line to the leftmost points of the plot that do not consistently deviate from a linear trend and compute its slope b . An estimate $\hat{m}$ of the number of the true $H_{0i}s$ is then yielded by $1/b - 1$. This estimate is slightly biased toward larger-than-true values.^[10]

Perlman^[23] found the likelihood ratio test for $H_{T \geq C}$ and derived its distribution for multivariate Gaussian endpoints with unknown covariance matrix. This test should have optimal properties; Lehmann^[21] writes that the likelihood ratio tests are often asymptotically most powerful in a uniform manner over the relevant alternative hypothesis, even in nonclassical situations. The power of Perlman's test has been empirically investigated by Follmann.^[24] He found that the test is actually preferable to both O'Brien's and Follmann's tests (see later discussion) only when some endpoints are truly favorably affected by T and a few number are unfavorably affected, a situation outside the range of $H_{T \geq C}$. Perlman's test is difficult to compute and has not yet found its way into most known statistical programs. The following are simpler tests for multivariate Gaussian endpoints.

O'Brien's test^[25] is particularly powerful if the true effects are equal on all endpoints.^[24] To use it, one has to compute a numerator and a denominator of a t -test statistic in the following way: From a computer program, obtain the k -dimensional vector $\bar{y}_T$ of the outcomes of the endpoints averaged on the T sample and the analog vector $\bar{y}_C$ computed as on the C sample. Get the estimated covariance matrix $\hat{V}$ of the endpoints as well. Then, to obtain the numerator, multiply the inverse of the estimated covariance matrix ($\hat{V}^{-1}$) by the vector of the mean differences of outcome: $\bar{y}_T - \bar{y}_C$, and sum up the elements of the resulting vector [formally, the numerator is $1' \hat{V}^{-1}(\bar{y}_T - \bar{y}_C)$,

where 1 is the vector of 1s]. The denominator is given by the usual correcting factor of t -tests: $\sqrt{1/n_T + 1/n_C}$ multiplied by the square root of the sum of the elements of $\hat{V}^{-1}$. [Formally, the denominator is $\sqrt{1/n_T + 1/n_C(1' \hat{V}^{-1} 1)}$]. For two endpoints, these computations can easily be performed by hand. If the endpoints are more than three, some matrix algebra program must be used. The obtained t is compared with the usual tables of critical values of t with $n_T + n_C - 2k$ degrees of freedom.

O'Brien's test can be one-sided or two-sided with the same modalities of a usual t -test. If the one-sided version turns out to be significant, the conclusion does not necessarily state that $H_{T \geq C}$ is true (although this is a very likely outcome). One has just proved that the expected value of the numerator of t , i.e., $1' \hat{V}^{-1} E(y_T - y_C)$ is positive. To see what this means, consider the case when the endpoints are independent. In that case, the numerator of O'Brien's test becomes the sum of the endpoint mean differences, each divided by the respective endpoint variance. So when the endpoints are independent, the hypothesis proved by the test states that the sum of the expected effects of T on all the endpoints, divided by their variances, is positive.

A very interesting feature of O'Brien's test is that, in its two-sided version, it can be applied to the comparison of more than two groups in stratified trials with prognostic covariates. (This can be carried out under the condition that the covariance matrix of the endpoints is the same within all strata, and conditional on all values of the prognostic covariates). To apply O'Brien's test in this general form, compute for each patient the quantity $u = 1' \hat{V}^{-1} y$ (where y is the vector of outcomes of the patient for all endpoints). Then, use u as the response variable in a usual program of analysis of variance.

Simulations show that this procedure keeps the FWER fairly under control, even when the data have exponential or Cauchy distributions.^[25] Follmann^[24] compared its power with that of Perlman's and his own test. He found that O'Brien's test is by far the most powerful version to detect $H_{T \geq C}$ when the endpoints are independent. Hence its use is recommended for nearly independent endpoints when one is certain that, if T is truly effective, it is so for all endpoints. O'Brien's test is also preferable with moderately positively correlated endpoints, if the true effects on all endpoints are very similar.

Tang et al.^[26] give a simplified approximate version of the likelihood ratio test of Perlman. The computation of Tang's test involves the Cholesky decomposition of the covariance matrix of the endpoints and other moderately complex matrix manipulations. The details are given in the original paper, together with the table of the critical values of the test statistic for up to 10 endpoints. If that table is not sufficient to analyze someone's data, the authors provide a simple formula (4.1 of their paper) for the computation of the P -value. It should not be excessively

difficult to implement that formula if the number of endpoints is not too high. The simulations performed by Tang and colleagues compare the power of their test to that of O'Brien's. They show that O'Brien's test is superior when the true effects of T on all endpoints are similar and that their test is superior in the opposite case.

Tang's test is intrinsically one-sided, but its conception is such that it can be significant in situations in which T is not superior to C . For instance, imagine that T and C are compared on uncorrelated endpoints. In this case, the test statistic is a weighted mean of the squared positive endpoint differences. If T has a strong favorable effect on one endpoint, but also an equally strong unfavorable one on the other endpoint, the test may be significant if we use it to demonstrate the superiority either of T or that of C . Therefore to apply Tang's one-sided test, it is not sufficient to adopt the pragmatic "approve or disapprove" point of view on a treatment evaluation, i.e., to merge countereffects and no effects in one null hypothesis (the logic that justifies one-sided tests in a single-endpoint trial). Indeed, overwhelming countereffects do not avoid Tang's tests stating significant favorable effects.

Follmann^[27] has proved that when Tang's method is applied to two positively correlated endpoints, it can become significant, even when the sample means of the two endpoints are both negatively affected by the treatment. This is not a crippling flaw of the test, however, because one could apply Tang's test at the condition that at least one of the endpoints is improved in the T sample. If this condition is not fulfilled, the null hypothesis could be accepted without further computations.

Follmann^[24] conceived a *one-sided modified T^2 test* for multivariate Gaussian endpoints, which joins ease of use, theoretically exact control of the α level, and good power to detect a large spectrum of alternative hypotheses within $H_{T \geq C}$. Its procedure with significance level α is as follows: Compute the mean difference between T and C of each endpoint, $\bar{y}_T - \bar{y}_C$, and sum them up over all endpoints. If the sum is negative, then retain H_0 . Otherwise, compare the two samples by a usual Hotelling's T^2 test: Reject H_0 if the test statistic is larger than the 2α critical value. In other words, the P -value of Follmann's test is *half* the P -value written on the output of the T^2 computation. Follmann proves that the Type I error rate of his test is always below α . With extensive simulations, he shows that his test is more powerful in detecting $H_{T \geq C}$ than the theoretically optimal Perlman's likelihood ratio test when the endpoints are moderately positively correlated ($\rho = 0.75$). In this situation, if the true effect of T , although favorable everywhere, is rather heterogeneous among the endpoints, Follmann's test is also superior to O'Brien's and should be recommended.

Rom^[28] proposes a global one-sided test for multiple dichotomous endpoints based on their joint distribution. The author gives an example of an application of his procedure

to the case of two endpoints with hypergeometric distribution. However, for three or more endpoints, Rom's method probably needs the application of bootstrap techniques to estimate the joint probability distribution of the endpoints.

Techniques Powerful in Detecting H_A

When one expects T to be superior to C on average but also to have some countereffects, the tests described in this section become relevant. Of course, according to the meaning given to the term *average superiority*, the chosen test will be different.

Follmann^[24] quotes the likelihood ratio test for a H_A , defined as follows: $E(y_T - y_C) > 0$ (i.e., the arithmetic mean of the effect of T on all endpoints is positive). However, he writes that this test is not yet available if the covariance matrix of the endpoints is unknown.

Follmann^[27] proposes a very simple and appealing test for multivariate normal endpoints. A slight generalization of his method is as follows: For each subject, multiply each endpoint measure by a predetermined fixed weight, and sum all the products to obtain a unique score. Then perform a usual analysis of variance on this score. This procedure is correct because, if the endpoints are Gaussian, every linear combination of these endpoints is likewise Gaussian. The test has a one-sided version, easily obtained from the two-sided one, by considering the observed score mean difference between T and C . If this difference is favorable to T , then $P_{\text{one-sided}} = P_{\text{two-sided}}/2$; otherwise, $P_{\text{one-sided}} = 1 - (P_{\text{two-sided}}/2)$. The alternative hypothesis proved by this test, if it is significant, is that the predetermined linear combination of the endpoint outcomes has a more favorable expected value for T than for C .

Follmann affirms that his method may be used when the multiple endpoints are surrogate of one "hard" endpoint difficult to measure directly. The linear combination of endpoint outcomes should be a linear model previously fitted to independent data, which turned out to be reasonably predictive of the "hard" endpoint. However, his method should also be considered when the "hard" endpoint aimed at is *impossible* to measure directly, e.g., cognitive impairment or quality of life. In this type of situation, the weights given to each endpoint should be attributed by experts of the field. The subjective choice of the weights should not be a problem if they are clearly stated in the protocol of the analysis.

In Follmann's simulations of a two-endpoint trial, the power of his test is always superior to the mean power of the univariate tests applied to the single endpoints when both endpoints are truly favorably affected by the treatment. Moreover, Follmann finds that the form of the rejection region of his test makes sense from a substantive point of view. Indeed, whatever is the number and the correlation structure of the endpoints, all the observed outcomes leading to the rejection of the null hypothesis correspond

to a better outcome score than all the outcomes leading to retain H_0 .

If applied to an independently measurable “hard” endpoint, such as mortality, Follmann’s procedure allows one to compute confidence intervals for it. Moreover, his score can be used as the response variable of a general linear model, allowing all the trial design features—more than two arms, stratification, random factors, prognostic covariates, repeated measures, etc.—to be taken into account in the statistical analysis.

From Follmann’s method, a formula can be deduced for the computation of the required sample size of a trial. Proceed as follows: Ask the field experts the minimum clinically relevant effect difference of each endpoint, assuming other endpoints stay unaffected. Then compute the weighted mean of the minimum clinically relevant differences, using the same weights chosen for the analysis. This mean should be a meaningful minimum clinically relevant score difference. Compute the variance of the score from the variances of the endpoints and their correlation (for the computation formula, see Ref. [29]). After knowing the minimum clinically relevant difference and the variance of the score, obtain the required sample size from the usual tables published for univariate trials.

O’Brien^[25] proposed a rank-based test: the measures of each endpoint are ranked among all subjects. (A better outcome has higher rank than a worse one.) Then, for each subject, the sum of the ranks of all endpoint outcomes is computed. This sum is asymptotically Gaussian and can be used as response variable of a general linear model. The test can be one- or two-sided. The alternative hypothesis proved by the one-sided version is as follows: “The probability that a subject of the T population has a better outcome on any endpoint than a subject of the C population is on average larger than 0.5.” O’Brien’s simulations show that the rank-sum test has a reasonable control of the Type I error rate; it rejects the true null hypothesis a maximum of 6.3% of the times for a declared $\alpha = 0.05$ in the set of situations considered by the author. The power of the rank-sum test is inferior to the power of the parametric test proposed by same author (see earlier discussion) in situations where the last should perform at its best, i.e., when the endpoints are multivariate Gaussian and all are truly improved by similar amounts by T . When the endpoints are not Gaussian, the rank-sum test turns out to be superior. Hence the rank-sum test is recommended when the endpoints are certainly far from the Gaussian distribution.

In the particular case where only one endpoint is actually affected by T , Follmann^[27] empirically proved that the Bonferroni correction is optimal. However, this situation should not arise if the endpoints of the trial are well chosen. If the Bonferroni correction, or other techniques devised to test the single H_{0i} s, are used to test the global H_0 , the multivariate P is equal to the smallest of the p ias computed for the H_{0i} s.

REGULATORY AGENCIES AND THE MULTIPLE-ENDPOINT ISSUE

Regulatory agencies do not give very detailed guidelines for the analysis of multiple endpoints in clinical trials. They require a fair control of the Type I error rate without giving any indication of a preferred technique. For instance, in the guidance for rheumatoid arthritis treatment evaluation,^[30] the FDA considers the following alternatives. 1) Three out of the four primary endpoints relevant to the pathology are significantly improved by T . 2) A multivariate statistical method is applied. 3) The efficacy variables are combined within patients (composite endpoint), and the weights used are clearly defined in the study protocol.

An example of the concern of the statisticians in the regulatory agencies for the analysis of multiple-endpoint clinical trials is expressed, for instance, by Sankoh et al.^[6] They give a fair and clear account of various adjustment procedures. The authors also propose a simulation-based method to correct the FWER of some P -value-based and ad hoc techniques. Their method is as follows: Estimate the sample correlation matrix among endpoints without breaking the blind code of the trial. Simulate thousands of trials with the same correlation structure and no effect of T . Apply the adjustment defined in the protocol to each virtual trial sample, and count the frequency of rejections. If the proportion of rejections is larger than the chosen rate (e.g., 0.05), apply the same adjustment again but using a smaller nominal α level (e.g., 0.04) in the formulae. Continue trying different nominal values of α until the proportion of rejections is pared down to the desired value (in our example, 0.05). Use the so-identified nominal α , instead of 0.05, in the adjustment procedures for the analysis of the real trial.

The authors also discuss the use of global test procedures and underscore the necessity that a global alternative hypothesis should be meaningful. As Sankoh points out, global tests are certainly not appropriate when each endpoint has a completely different pathologic meaning.

REFERENCES

1. Bickel, P.J.; Doksum, K.A. *Mathematical Statistics*; Hoen-Day: San Francisco, CA, 1976, 439.
2. Holm, S. *Scand. J. Statist.* **1979**, *6*, 65–70.
3. Hommel, G. *Biometrika* **1988**, *75*, 383–386.
4. Hochberg, Y. *Biometrika* **1988**, *75*, 800–802.
5. Brown, B.W.; Russell, K. *Stat. Med* **1997**, *16*, 2511–2528.
6. Sankoh, A.J.; Huque, M.F.; Dubey, S.D. *Stat. Med* **1997**, *16*, 2529–2542.
7. Hochberg, Y.; Rom, D. *J. Stat. Plan. Inference* **1995**, *48*, 141–152.
8. Simes, R.J. *Biometrika* **1986**, *73*, 751–754.
9. Comelli, M.; Klersy, C. *JBS* **1996**, *6*, 115–125.
10. Hochberg, Y.; Benjamini, Y. *Stat. Med* **1990**, *9*, 811–818.

11. Schweder, T.; Spjøtvoll, E. *Biometrika* **1982**, *69*, 493–502.
12. Hayter, A.J.; Hsu, J.C. *JASA* **1994**, *89*, 128–136.
13. Tamhane, A.C. *Handbook of Statistics: Design and Analysis of Experiments*; Ghosh, S. Rao, C.R., Eds.; North-Holland: Amsterdam, 1996, Vol. 13, 587–630.
14. Westfall, P.H.; Young, S.S. *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*; Wiley: New York, 1993, Vol. 1.
15. Troendle, J.F. *JASA* **1995**, *90*, 370–378.
16. Troendle, J.F. *Biometrics* **1996**, *52*, 846–859.
17. SAS Institute Inc. *SAS/STAT Software: Changes and Enhancements Through Release 6.11*; SAS Institute: Cary, NC, 1996, 733–739.
18. Mantel, N. *Biometrics* **1980**, *36*, 381–399.
19. Benjamini, Y.; Hochberg, Y. *J. R. Stat. Soc* **1995**, *57*, 289–300.
20. Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*, 1st Ed.; Wiley: New York, 1958, 115.
21. Lehmann, E.L. *Testing Statistical Hypotheses*, 2nd Ed.; Wiley: New York, 1986, 462–477.
22. Cupples, L.A.; Heeren, T.; Schatzkin, A.; Colton, T. *Ann. Intern. Med* **1984**, *100*, 122–129.
23. Perlman, M.D. *Ann. Math. Stat* **1969**, *40*, 549–567.
24. Follmann, D. *JASA* **1996**, *91*, 854–861.
25. O'Brien, P.C. *Biometrics* **1984**, *40*, 1079–1087.
26. Tang, D.I.; Gnecco, C.; Geller, N.L. *Biometrika* **1989**, *76*, 577–583.
27. Follmann, D. *Stat. Med* **1995**, *14*, 1163–1175.
28. Rom, D.M. *Stat. Med* **1992**, *11*, 511–514.
29. Hogg, R.V.; Craig, A.T. *Introduction to Mathematical Statistics*; Macmillan: New York, 1978, 177.
30. U.S. Department of Health and Human Services. *Food and Drug Administration Guidance for Industry (Draft Guidance)*, 1998, 28–29.

Multiple-Dose Bioequivalence Studies

Vernon M. Chinchili

Penn State College of Medicine, Hershey, Pennsylvania, U.S.A.

Abstract

The purpose of a bioequivalence study is to determine if two products are sufficiently similar in their rate and extent of drug absorption so that they can be interchangeably used in a patient to produce the same therapeutic outcome. The focus of this entry is the statistical assessment of average bioequivalence in multiple-dose studies.

INTRODUCTION

The purpose of a bioequivalence study is to determine if two products, containing the same amount of therapeutic agent, are sufficiently similar in their rate and extent of drug absorption so that they can be interchangeably used in a patient to produce the same therapeutic outcome.

Methods currently required by the Food and Drug Administration (FDA) generally involve a single measurement of the metrics of rate and extent of absorption [area under the curve (AUC), $C_{\max}$, $T_{\max}$] in 16–32 normal, healthy volunteers. This procedure is referred to as “average bioequivalence” because only the means of the two products are statistically compared. There are several assumptions inherent in this method:

1. The sample of individuals is homogeneous.
2. The volunteer study population is representative of the target patient populations.
3. For each metric, the variances of the test and reference products are the same.

OVERVIEW

The modification of average bioequivalence is population bioequivalence, in which estimated means and variances are compared. Generally, it is recognized that population bioequivalence is adequate to show that the test product is safe and effective in the population and is therefore prescribable. However, this process does not examine differences between the test and reference products in individual subjects (or patients). Therefore it may not be a good indicator of whether the test and reference products can be interchangeably used in individual subjects. This has been referred to as the “switchability” of products.^[1] Anderson and Hauck^[2] proposed an early approach for assessing the switchability, or individual bioequivalence, of test and reference products.

To better assess the interchangeability or switchability of products in individual subjects, there has been increasing scientific interest in developing new statistical procedures that allow the assessment of the following questions, which are not adequately examined by current methods of population bioequivalence.^[1,3,4] Although regulatory interest in this issue was intense during the 1990s,^[5] recent regulatory guidelines have shelved the pursuit of requiring individual bioequivalence.^[6]

However, scientific interest remains for addressing some of the following questions:

1. Do the intrasubject variances of the test and reference products differ?
2. Are there subsets of the study population (and thus, possibly, the patient population) that show greater differences between test and reference products?
3. Can these differences be statistically detected as a subject $\times$ treatment interaction?

These questions can be answered only if replicate measures in the same individual are available. There are basically two experimental approaches to obtaining replicate measures: replicate designs in single-dose studies and repeated-measurements designs in multiple-dose studies.

A replicate design in a single-dose study refers to the situation in which a subject receives each of the test and reference products in a single-dose administration, and in which each event is replicated on more than one occasion. Thus a replicate design in a single-dose study provides replicate observations of the important metrics.

For subjects in a multiple-dose study, each formulation is administered on repeated regular intervals in order to attain steady state conditions. Blood sampling is performed after the steady state conditions have supposedly occurred. However, multiple-dose studies provide the opportunity to conduct blood sampling on repeated occasions for each subject within each period of the crossover design. For example, if subjects in a multiple-dose study

are to be administered a formulation on each of d consecutive days, and it takes $d-q+1$ days to attain steady state conditions, then this multiple-dose study can provide multiple observations of the important metrics on q days of that period (days $d-q+1$ through d).

The focus of the following statistical section is the assessment of average bioequivalence, although multiple-dose studies provide an appropriate framework for the assessment of individual bioequivalence. Wang et al.^[7] describe an approach for estimating AUC in a multiple-dose study in which there is noncompliance. However, the following section assumes full compliance.

STATISTICAL METHODS

We consider a general $s \times p$ crossover design in which subjects are randomized to s sequences containing test (T) and reference (R) formulations administered over p periods. Within sequence i , $i = 1, \dots, s$, the test formulation appears in p_{Ti} periods and the reference formulation appears in p_{Ri} periods, with $p_{Ti} + p_{Ri} = p$. Suppose that measurements are taken on q occasions within each period after steady state conditions have been attained. Let

$$\begin{aligned} \mathbf{Y}_{Tijk} &= [\mathbf{Y}_{Tijk1} \dots \mathbf{Y}_{Tijkq}]' \\ \mathbf{Y}_{Rijl} &= [\mathbf{Y}_{Rijl1} \dots \mathbf{Y}_{Rijlq}]' \end{aligned} \quad (1)$$

denote the q vector of observations for the k th occurrence of the test formulation and the l th occurrence of the reference formulation, respectively, for the j th subject within the i th sequence, $i = 1, \dots, s$, $j = 1, \dots, n_i$, $k = 1, \dots, p_{Ti}$, and $l = 1, \dots, p_{Ri}$. For example, the $\mathbf{Y}_{Tijk}$ and $\mathbf{Y}_{Rijl}$ could denote measurements of AUC or $C_{\max}$ on the natural log scale. We use the natural log transformation because there is reason to believe that AUC and $C_{\max}$ are usually log-normally distributed.^[8]

The model we assume is

$$\mathbf{Y}_{Tijk} = \mu_T + \mu_{Tm}^* + U_{Tij} + \gamma_{Tikm} + E_{Tijk} \quad (2a)$$

$$\mathbf{Y}_{Rijl} = \mu_R + \mu_{Rm}^* + U_{Rij} + \gamma_{Rilm} + E_{Rijl} \quad (2b)$$

for $i = 1, \dots, s$, $j = 1, \dots, n_i$, $k = 1, \dots, p_{Ti}$, $l = 1, \dots, p_{Ri}$, and $m = 1, \dots, q$, where

μ_T = population mean for formulation T

μ_R = population mean for formulation R

μ_{Tm}^* = deviation from population mean at occurrence m for formulation T

μ_{Rm}^* = deviation from population mean at occurrence m for formulation R

γ_{Tikm} = nuisance parameter (occurrence $\times$ sequence nested within formulation T)

γ_{Rilm} = nuisance parameter (occurrence $\times$ sequence nested within formulation R)

$\mathbf{U}_{ij} = [U_{Tij} \ U_{Rij}]'$ = bivariate random subject effect

$\mathbf{E}_{Tijk} = [E_{Tijk1} \dots E_{Tijkq}]'$ = q vector of random errors under formulation T

$\mathbf{E}_{Rijl} = [E_{Rijl1} \dots E_{Rijlq}]'$ = q vector of random errors under formulation R

Constraints on the nuisance parameters are necessary to prevent overparameterization of the model with respect to the fixed location parameters:

$$\sum_{m=1}^q \mu_{Tm}^* = 0 \quad \text{and} \quad \sum_{m=1}^q \mu_{Rm}^* = 0 \quad (3a)$$

$$\sum_{i=1}^s \sum_{k=1}^{p_{Ti}} \gamma_{Tikm} = 0 \quad \text{and} \quad (3b)$$

$$\sum_{i=1}^s \sum_{l=1}^{p_{Ri}} \gamma_{Rilm} = 0$$

for each $m = 1, \dots, q$.

With the constraints in place, the model in Eqs. 2 and 3 is said to be saturated because there are a total of spq unrestricted location parameters that corresponds exactly to the number of cells in the $s \times p$ crossover design with q occurrences within each of the p periods. It is not necessary to include all the nuisance parameters in the model and force it to be saturated. Although a saturated model is conservative, it protects against confounding of the formulation means with any excluded nuisance parameters.

The nuisance parameters $\mu_T^* = [\mu_{T1}^* \ \mu_{T2}^* \dots \mu_{Tq}^*]'$ and $\mu_R^* = [\mu_{R1}^* \ \mu_{R2}^* \dots \mu_{Rq}^*]'$ are of special interest because they represent deviations from test and reference formulation means, respectively, over the q occurrences within each period. If steady-state conditions have not been achieved, then it is anticipated that the estimates of μ_T^* and μ_R^* will significantly differ from zero and the variability of the estimated μ_T and μ_R will be relatively large. In such a situation, there will be decreased statistical power for bioequivalence testing.

In addition, we assume that carryover effects do not exist in the proposed model. This is a reasonable assumption in bioequivalence trials^[8] if the washout period is of adequate length; however, it may not be so for other types of crossover trials. If the existence of carryover effects is suspected, then the model can be modified to include them.

The distributional assumptions for the j th subject within the i th sequence, $i = 1, \dots, s$, and $j = 1, \dots, n_i$, are as follows:

$$\mathbf{U}_{ij} = \begin{bmatrix} U_{Tij} \\ U_{Rij} \end{bmatrix} \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Omega = \begin{bmatrix} \omega_{TT} & \omega_{TR} \\ \omega_{TR} & \omega_{RR} \end{bmatrix} \right) \quad (4)$$

where Ω is an unknown positive definite matrix;

$$\mathbf{E}_{Tijk} \sim N_m(\mathbf{0}, \sigma_{Tm}^2 \mathbf{I}_m) \quad \text{and} \quad \mathbf{E}_{Rijl} \sim N_m(\mathbf{0}, \sigma_{Rm}^2 \mathbf{I}_m) \quad (7)$$

where σ_T^2 and σ_R^2 are unknown positive scalars and $\mathbf{I}_m$ is the $m \times m$ identity matrix. Further, we assume that the random subject effects and the random errors are mutually independent.

The model and assumptions described in Eqs. 1–7 are a generalization of those presented by Chinchilli et al.^[9] for multiple-dose studies with a 2×2 crossover design, and those given by Chinchilli and Esinhart^[10] for single-dose studies with an $s \times p$ crossover design. Other researchers, such as Ekbohm and Melander^[11] and Sheiner,^[1] have considered similar types of models with the distributional assumptions as Eqs. 6 and 7 for the case $q = 1$. The distributional assumptions in Eqs. 6 and 7 differ from those typically assumed for a crossover experiment, such as univariate random effects, i.e., $U_{Tij} = U_{Rij}$ for $i = 1, 2$ and $j = 1, \dots, n_j$, and homogeneous intrasubject variances, i.e., $\sigma_T^2 = \sigma_R^2$, such as the books by Jones and Kenward^[12] and Ratkowsky et al.^[13]

The model and assumptions in Eqs. 1–7 fall into the general realm of a mixed-effects linear model. Therefore standard statistical software packages can be invoked to perform a likelihood analysis of a particular data set. A special case of interest is a design that is uniform within sequences, defined by Laska et al.^[14] as a situation in which the test formulation appears the same number of times within each sequence ($p_{T1} = \dots = p_{Ts}$) and the reference formulation appears the same number of times within each sequence ($p_{R1} = \dots = p_{Rs}$). If the crossover design is uniform within sequences and the model is saturated in terms of the location parameters, then, as Chinchilli and Esinhart^[10] showed, the maximum likelihood (ML) and restricted maximum likelihood (REML) estimates of the location parameters and the variance–covariance components have a closed-form solution. Under these circumstances, if $\delta = \mu_T - \mu_R$, then

$$\hat{\delta} = \bar{Y}_{T\dots} - \bar{Y}_{R\dots} \quad (6)$$

is the ML and REML estimator of the formulation difference. The estimator in Eq. 8 is intuitive because it simply involves averaging the response over all appropriate observations. The variance expression for the estimator in Eq. 6 is

$$\text{var}(\hat{\delta}) = \left(\frac{1}{n_1} + \dots + \frac{1}{n_s} \right) \times \left(\omega_{TT} + \omega_{RR} - 2\omega_{TR} + \frac{1}{p_T q} \sigma_T^2 + \frac{1}{p_R q} \sigma_R^2 \right) / s^2 \quad (7)$$

The variance formula in Eq. 7 is useful for sample-size calculations, as described in Ref. [9] for the 2×2 crossover design.

The typical approach for assessing bioequivalence with respect to average bioavailability based on pharmacokinetic endpoints such as AUC and $C_{\max}$ is the two 1-sided testing approach introduced by Schuirmann^[15] and the

corresponding $100(1 - \alpha)\%$ confidence interval for $\mu_T - \mu_R$ described by Berger and Hsu,^[16] denoted as $[\hat{\delta}_\alpha, \hat{\delta}_{1-\alpha}]$. For the confidence-interval approach, the endpoints of the confidence interval are exponentiated. It is then determined whether this exponentiated confidence interval lies within (0.80, 1.25). The null hypothesis of bioinequivalence at the α significance level is rejected in favor of the exponentiated $100(1 - \alpha)\%$ confidence interval that lies entirely within (0.80, 1.25).

RESULTS AND DISCUSSION

A multiple-dose study was conducted by the Department of Pharmacy and Pharmaceutics at Virginia Commonwealth University, supported by Mova Pharmaceutical Company, in order to assess the average bioequivalence of test and reference formulations of levethyroxine sodium. Their study consisted of 24 healthy male volunteers who were randomized to receive either sequence TR or sequence RT. Blood samples were drawn from each subject on days 41 and 43 of steady state within each period of formulation administration; so $q = 2$ in the notation of our model in Eqs. 2a, and 2b. For purposes of illustration, we focus on the analysis of AUC only.

We applied the natural log transformation to the data and proceeded with the statistical model and analysis as described earlier. The SAS code for the analysis using PROC MIXED is as follows:

```
PROC MIXED;
CLASS subject formulat;
MODEL log_auc = mu_t mu_r mustar_t mustar_r
gamma_t1 gamma_t2 gamma_r1 gamma_r2/NOINT
SOLUTION;
RANDOM mu_t mu_r/SUBJECT = subject TYPE =
UNG;
REPEATED/SUBJECT = subject TYPE = SIM R
GROUP = formulat;
ESTIMATE 'formulation diff' mu_t 1 mu_r -1 /DF = 22;
```

The uppercase components of the SAS code represent statements and keywords, whereas the lowercase components represent user-provided variable names and options. Variables corresponding to the two formulation means and the six nuisance parameters (mustar_t, mustar_r, gamma_t1, gamma_t2, gamma_r2, gamma_r2) are created using IF–THEN statements within an SAS DATA step. The RANDOM and REPEATED statements are constructed in a manner to provide the intersubject and intrasubject variance components, as described in the model and assumptions of the previous section. The ESTIMATE statement yields the estimated formulation difference and its standard error, which are needed to construct the 95% confidence interval for assessing average bioequivalence.

The estimated test and reference formulation means for the logarithm of the AUC are 6.892 and 6.886, respectively. None of the six estimated nuisance efforts are significantly different from zero ($p > 0.15$). The estimated variance components are as follows:

Parameter	Estimate
ω_{TT}	0.0518
ω_{RR}	0.0422
ω_{TR}	0.0400
σ_T^2	0.0035
σ_R^2	0.0039

With respect to the assessment of average bioequivalence, the ratio of the geometric means on the original scale is 1.001, and the 95% confidence interval using the Berger and Hsu^[16] approach is (0.96, 1.05). Therefore average bioequivalence is concluded for AUC. As an aside, the intrasubject variances of the two formulations, 0.0035 and 0.0039, appear to be very similar.

CONCLUSION

A bioequivalence study is used to determine if the rate and extent of drug absorption for a test product is similar to that of a reference product. A multiple-dose bioequivalence study provides the opportunity for more data on the rate and extent metrics, hence, it should yield more statistical power than a single-dose bioequivalence study. This paper describes a mixed-effects linear model for analyzing the data from a multiple-dose bioequivalence study and illustrates the analysis with a real-life example.

ACKNOWLEDGMENTS

I thank Dr. William H. Barr for his contribution as coauthor to the original version of this article published in 2000.

REFERENCES

1. Sheiner, L.B. Bioequivalence revisited. *Stat. Med.* **1992**, *11*, 1777–1788.

2. Anderson, S.; Hauck, W.W. Consideration of individual bioequivalence. *J. Pharmacokinet. Biopharm.* **1990**, *18*, 259–273.
3. Wellek, S. A comment on so-called individual criteria of bioequivalence. *J. Biopharm. Stat.* **1997**, *7*, 17–21.
4. Marzo, A.; Balant, L.P. Bioequivalence. An updated reappraisal addressed to applications of interchangeable multi-source pharmaceutical products. *Arzneimittelforschung* **1995**, *45*, 109–115.
5. Chen, M.L. Individual bioequivalence—A regulatory update. *J. Biopharm. Stat.* **1997**, *7*, 5–11.
6. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research. *Guidance for Industry: Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Considerations*; 2000, <http://www.fda.gov/cder/guidance>
7. Wang, W.; Hsuan, F.; Chow, S.-C. An adjusted two one-sided t-test for the assessment of bioequivalence with multiple doses. *J. Biopharm. Stat.* **1997**, *7*, 157–170.
8. Westlake, W.J. Bioavailability and Bioequivalence of Pharmaceutical Formulations. In *Biopharmaceutical Statistics for Drug Development*; Peace, K.E., Ed.; Marcel Dekker: New York, 1988; 329–352.
9. Chinchilli, V.M.; Esinhart, J.D.; Barr, W.H. Analysis of multiple-dose bioequivalence studies. *J. Biopharm. Stat.* **1994**, *4*, 423–435.
10. Chinchilli, V.M.; Esinhart, J.D. Design and analysis of intra-subject variability in crossover experiments. *Stat. Med.* **1996**, *15*, 1619–1634.
11. Ekbohm, G.; Melander, H. The subject-by-formulation interaction as a criterion of interchangeability of drugs. *Biometrics* **1989**, *45*, 1249–1254.
12. Jones, B.; Kenward, M.G. *Design and Analysis of Cross-Over Trials*; Chapman and Hall: New York, 1989.
13. Ratkowsky, D.A.; Evans, M.A.; Alldredge, J.R. *Cross-Over Experiments: Design, Analysis, and Application*; Marcel Dekker: New York, 1993.
14. Laska, E.M.; Meisner, M.; Kushner, H.B. Optimal crossover designs in the presence of carryover effects. *Biometrics* **1983**, *39*, 1087–1091.
15. Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* **1987**, *15*, 657–680.
16. Berger, R.L.; Hsu, J.C. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Stat. Sci.* **1996**, *11*, 283–319.

Multiple-Stage Designs: Phase II Cancer Trials

Masha Kocherginsky and Shang P. Lin

Department of Health Studies, The University of Chicago, Chicago, Illinois, U.S.A.

Abstract

This entry provides a review of several widely used multi-stage designs, along with some recently proposed modifications, improvements, and extensions. Binomial and trinomial response models are considered in detail, and some additional multi-stage designs are discussed.

Keywords: Cancer trials; Clinical trials; Group sequential methods; Multi-stage design.

INTRODUCTION

The primary objective of a Phase II cancer trial is to determine if a treatment is “active” or efficacious enough to warrant a definitive Phase III clinical trial. Traditionally, the development of Phase II trial designs has focused on cytotoxic chemotherapeutic agents, which are presumed to have clinical benefit by shrinking tumors. Typically, treatment in Phase II trials is administered to patients with metastatic disease who will be considered responsive or not based on tumor shrinkage. Past experience suggests that efficacy of a drug is disease-site specific; therefore, a Phase II trial is limited to a specific type of cancer. The kind of cancer against which the treatment is to be tested is determined through pre-clinical animal studies and Phase I clinical trials.

Tumor shrinkage remains the most common primary endpoint in Phase II cancer trials of cytotoxic agents. Tumor response can be determined in the first few months of treatment and usually has good correlation with survival. In order to obtain an objective evaluation of the response, the patients should have a disease whose severity in terms of tumor size can be measured through diagnostic tools, most commonly radiographic imaging such as CT or MRI.

Usually, a Phase II cancer trial is based on a one-sample statistic with binary response (treatment success or failure). The classical design is a single-stage procedure in which a fixed number of patients are accrued to receive the treatment. Based on the number of responses (i.e., successes) observed, one would decide whether the treatment holds sufficient promise to warrant further testing, typically in a Phase III trial. In this process, false-positive and false-negative errors (α and β) should be kept within pre-specified limits. The key design issue in this situation is determining the minimum sample size required and the number of responses needed to declare that the treatment is active.

Ethically the design should allow for early termination so that the number of patients exposed to treatment with poor activity is minimized (this is commonly referred to as “stopping for futility”). In this regard, a full sequential design in which analysis is performed after the outcome of each new patient becomes available is most efficient, but is difficult to implement. Multi-stage design is a compromise between the simplicity of single-stage design and the efficiency of full sequential design. Such design reduces, on the average, the number of patients required while keeping the trial logistics manageable.

Since Gehan^[1] proposed the use of a two-stage design in Phase II studies, a number of multi-stage designs, mostly for two stages, have been proposed in the literature. In general, they differ mainly in the number of stages, stopping rule, and optimality criterion. Some are not based on any optimization such as the Fleming’s^[2] designs, which focus on the spending of error rates at each stage. In terms of statistical properties and practical considerations, Mariani and Marubini^[3] have reviewed a number of multi-stage designs for Phase II cancer trials. Kramar et al.^[4] have described and compared designs of Gehan,^[1] Fleming,^[2] Simon,^[5] and Ensign et al.^[6] and Mariani and Marubini,^[7] and Perrone et al.^[8] have studied usage of the available statistical designs for Phase II cancer trials. Estimation of the binomial proportion at the end of a multi-stage trial is addressed by Jung et al.^[9] and Tsai et al.^[10] Jung et al.^[11] and Koyama and Chen^[12] discuss p-value calculation at the end of the trial. Advantages and disadvantages of randomized Phase II trials including a prospective control have been discussed (for example, Ref. [13]). Some new randomized multi-stage trial designs have been proposed by Jung^[14] for binomial outcome, and by Sun et al.^[15] for trinomial outcome, and the use of randomized discontinuation design has been proposed by Rosner et al.^[16] Seamless Phase II/III designs, which can also be considered multi-stage designs, have been reviewed by Thall.^[17]

Recent discoveries in molecular biology of cancer have led to the development of novel molecular targeted anti-cancer agents which target specific proteins involved in tumor growth, angiogenesis, metastases, and other cell processes. The effect of many such therapies is thought to be cytostatic, meaning that the drug slows down or stops tumor growth rather than shrinks existing tumors. The implicit assumption for such drugs is that tumor growth inhibition will translate into clinical benefit of increased progression-free or overall survival. Clearly, for such therapies the traditional dichotomous response-based endpoint is no longer appropriate because response rate is expected to be low. The need for alternative approaches to clinical trial design and endpoint selection for such agents has been well established.^[18] Some of the proposed endpoints include multinomial response categories, continuous endpoints (e.g., changes in tumor size), and time to progression or progression-free survival at a fixed time point.^[19,20] The use of randomized trials with a prospective control group to compensate for the lack of reliable historical control data for many cytostatic agents has also been discussed.^[13,21]

This paper provides a review of several widely used multi-stage designs, along with some recently proposed modifications, improvements, and extensions. Binomial and trinomial response models are considered in detail, and some additional multi-stage designs are discussed. We present the frequentist approach, while the reader is referred to Tan and Machin^[22] and references therein for discussion of the Bayesian approach. The rest of the chapter is as follows: the section “Hypothesis Testing” describes the hypothesis to be tested; “Single-Stage Design” discusses designs based on fixed-sample procedure, which serves to lay the foundation for the discussions of multi-stage designs given in section “Multi-stage Design,” and finally, some concluding remarks are given in “Conclusion.”

HYPOTHESIS TESTING

Because we would like to make a decision whether the treatment is active enough to warrant a Phase III trial, we consider the design within the hypothesis testing framework in which a null hypothesis (H_0) is compared to an alternative hypothesis (H_1). $H_0(H_1)$ states that the true response rate is not better (worse) than a pre-specified undesirable (desirable) level. A decision rule for acceptance or rejection of H_0 is then defined based on the number of responses. The design must take into account two possible errors that occur in the determination of activity or the response rate, namely, the probability of claiming that the treatment is active (or response rate is high) when it is inactive, and the probability of claiming the treatment is inactive (or tumor response rate is low) when it is active. Statisticians refer to these two errors as the type I error (false-positive error) and type II error (false-negative error)

respectively, and seek to control them at prescribed low levels α and β .

Binomial Response

Standard medical practice classifies the tumor response into four categories: (i) complete response (CR); (ii) partial response (PR); (iii) stable disease (SD); and (iv) progression of disease (PD) according to the Response Evaluation Criteria in Solid Tumors (RECIST).^[23] Often this initial classification is further grouped into two categories yielding a binomial response. For example, for solid tumors, response includes both complete and partial responses (50% or more reduction in tumor size), and non-response includes both stable disease and progression of disease.

When the efficacy of a treatment is defined as a binary endpoint (response/no response), the typical hypothesis we wish to test is

$$H_0 : p \leq p_0 \quad \text{versus} \quad H_1 : p \geq p_1 \quad (1)$$

where p is the true probability of response, p_0 is a selected undesirable response rate, and $p_1(>p_0)$ is a selected desirable response rate. The value of p_0 is selected as the response rate above which the treatment is considered as sufficiently active. The alternative value p_1 is selected as the value of the active response rate that needs to be detected with high probability (power). Note that rejection of H_0 only leads to the conclusion that $p > p_0$ and not that $p \geq p_1$. Hypothesis (1) remains the same when the role of “response” is exchanged with that of “no toxicity;” hence designs developed for the binomial response rate are also applicable for toxicity monitoring.

Trinomial Response

A CR, defined as complete disappearance of the tumor, is more desirable than a PR since a CR is more likely to signal an important anti-tumor effect and result in a survival improvement. To differentiate between CR and PR, one could consider a trinomial response model where a patient’s response has three possible outcomes: CR, PR, or NR (no response) with probabilities p_c , p_p and $(1-p_c-p_p)$ respectively. Panageas et al.^[24] have considered the following hypothesis testing:

$$\begin{aligned} H_0 : p_c \leq p_{0c} \text{ and } p_p \leq p_{0p} \quad & \text{versus} \\ H_1 : p_c \geq p_{1c} \text{ or } p_p \geq p_{1p} \end{aligned} \quad (2)$$

where p_{0c} and p_{0p} are selected undesirable CR rate and PR rate respectively, and $p_{1c}(>p_{0c})$ and $p_{1p}(>p_{0p})$ are selected desirable CR rate and PR rate respectively. If a more restrictive alternative hypothesis $H_1 : p_c \geq p_{1c}$ and $p_p \geq p_{1p}$ is considered, then hypothesis (2) reduces to hypothesis (1) when the CR and PR rates are combined. The combination

of complete and partial response rates (CR + PR) is often referred to as the objective response rate (OR).

Zee et al.^[25] and Freidlin et al.^[26] considered a different trinomial response model where a patient's response takes three possible outcomes: R(CR or PR), P(early progression), and S(stable disease) with probabilities p_r , p_p , and $(1 - p_r - p_p)$ respectively. This model differentiates between patients with stable disease and early progressive disease, rather than combining them as a non-response group. The hypothesis to be tested is:

$$\begin{aligned} H_0 : p_r \leq p_{0r} \text{ and } p_p \geq p_{0p} & \text{ versus} \\ H_1 : p_r \geq p_{1r} \text{ or } p_p \leq p_{1p} \end{aligned} \quad (3)$$

where p_{0r} and p_{1r} are the undesirable and desirable response rates, and p_{0p} and p_{1p} are the undesirable and desirable progression rates. Note that for the response rate the higher rate is more desirable, while for the progression rate the lower rate is more desirable. Thus in the above hypothesis we assume $p_{1r} > p_{0r}$ and $p_{1p} < p_{0p}$.

The hypothesis formulation in Zee et al. is particularly relevant to clinical trials of cytostatic agents because the alternative hypothesis of either improved response or decreased early progression represents the clinical perspective. However, Freidlin et al.^[26] note that from the design point of view, the more restricted alternative hypothesis $H_1 : p_r \geq p_{1r}$ and $p_p \leq p_{1p}$ is more appropriate because for some points in the alternative space (e.g., $p_r = 0$ and $p_p = p_{1p}$) defined with an "or" the hypothesis test will not have the stated false-negative error rate, i.e., will have low power.

Pragmatically, testing hypothesis (3) is equivalent to testing the hypothesis

$$\begin{aligned} H_0 : p_r \leq p_{0r} \text{ and } p_s \leq p_{0s} & \text{ versus} \\ H_1 : p_r \geq p_{1r} \text{ or } p_s \geq p_{1s} \end{aligned} \quad (3a)$$

where $p_s \equiv (1 - p_r - p_p)$ denotes the probability of S (stable disease), and p_{0s} and p_{1s} ($> p_{0s}$) are the undesirable and desirable stable disease rates. However, because disease progression is more critical than stable disease, hypothesis (3) specified in terms of progressive disease rate may be more directly relevant for making the decision of whether to accept or reject a new treatment. Furthermore, note that hypothesis (3a) is essentially the same as hypothesis (2) with the roles of R, S, and P are exchanged with CR, PR, and NR respectively.

SINGLE-STAGE DESIGN

The single-stage procedure is the classical approach to design a Phase II cancer trial. The analysis is performed after the planned number of patients has been accrued. Although its simplicity is desirable, for ethical reasons it is usually considered unacceptable for Phase II cancer

trials since it does not allow early termination of the trial when the treatment does not appear to be active (stopping for futility). Discussion of the single-stage design here serves to provide a foundation for multi-stage designs, and to highlight basic differences between single-stage and multi-stage designs.

Binomial Response

Let X be a random variable representing the number of responses in n patients. We denote the mass and cumulative binomial probability functions at $X = r$ as $b(r; p, n)$ and $B(r; p, n)$; respectively. Thus, $b(r; p, n) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$ and $B(r; p, n) = \sum_{j=0}^r b(j; p, n)$.

When one designs a single-stage trial to test hypothesis (1), once the parameters (p_0 , p_1 , α , β) are specified, the required sample size, n , and the critical point r at or above which to reject H_0 can be determined by a search over all possible (r, n) 's. For any given (r, n) , let $\beta(p) \equiv B(r-1; p, n)$, $\alpha' = 1 - \beta(p_0)$ and $\beta' = \beta(p_1)$. In practice, a small-sized n is used as the starting value, then the maximum value of X , say x , for which $\beta' \leq \beta$ is found; if $\alpha' \leq \alpha$ at x , n is retained and the critical point r is set equal to x , otherwise n is increased by 1 and the cycle is repeated. Note that due to the discrete nature of the binomial response variable α' and β' thus obtained may differ substantially in some cases from their nominal levels α and β .

For moderate to large samples, $z(p) = (\hat{p} - p) / [p(1-p)/n]^{1/2}$ follows approximately the standard normal distribution, where $\hat{p} = X/n$. One would then reject H_0 whenever $z(p_0) \geq z_{1-\alpha}$; that is, whenever $X \geq np_0 + z_{1-\alpha} \{np_0(1-p_0)\}^{1/2}$, where z_u denotes the u -quantile of the standard normal distribution. Due to the discreteness of the binomial distribution, the more appropriate critical value is (see Table 1 in Fleming^[2]):

$$r = \left[np_0 + z_{1-\alpha} \{np_0(1-p_0)\}^{1/2} \right]^* \quad (4)$$

where $[f]^*$ denotes the nearest integer to f . The sample size required to satisfy the error constraints on α and β is approximately (see Fleming^[2]):

$$n = \left\{ [z_{1-\beta}(p_1(1-p_1))^{1/2} + z_{1-\alpha}(p_0(1-p_0))^{1/2}] / (p_1 - p_0) \right\}^2 \quad (5)$$

When the sample size is small, the design based on the normal approximation can yield anomalous results (see A'Hern^[27]). The n obtained from the exact search is in general larger than that using the normal approximation.

Trinomial Response

Because hypotheses (3) and (3a) are essentially equivalent to hypothesis (2), we shall focus our discussion on testing hypothesis (2). Here the acceptance of H_0 is defined

by a two-dimensional region R of lattice points in the first quadrant, that is $R = \{(i, j) | 0 \leq i \leq I, 0 \leq j \leq Y_i, \text{ and } Y_1 > Y_2 > \dots > Y_I \geq 0\}$ where the first coordinate i refers to the number of CRs and the second coordinate j that of PRs (for a graphical depiction see Panageas et al.^[25]). The boundary of such a region can be represented as $\{(i, Y_i), 0 \leq i \leq I\}$, where I is the maximum number of CRs for which at least one point of the acceptance region R occurs, and Y_i is the maximum number of PRs in R at each value of j . Let x_c and x_p denote the numbers of observed CRs and of PRs, respectively, in n patients. The decision rule then is to accept H_0 if (x_c, x_p) falls on or below the boundary of R , otherwise the treatment is deemed effective and warranting further testing.

The mass trinomial probability function of observing x_c complete responses and x_p partial responses in n patients is given by:

$$t(x_c, x_p; p_c, p_p, n) = \frac{n!}{x_c! x_p! (n - x_c - x_p)!} p_c^{x_c} p_p^{x_p} (1 - p_c - p_p)^{n - x_c - x_p}$$

Let $\beta(p_c, p_p)$ denote the cumulative trinomial probability in the region R , that is,

$$\beta(p_c, p_p) \equiv P\{(x_c, x_p) \in R | (p_c, p_p)\} = \sum_{i=0}^I \sum_{j=0}^{Y_i} t(i, j; p_c, p_p, n) \quad (6)$$

Then, we can calculate the overall type I and type II errors as $\alpha' = 1 - \beta(p_{0c}, p_{0p})$ and $\beta' = \beta(p_{1c}, p_{1p})$ respectively. Similarly, one can define the acceptance region for testing hypothesis (3) or (3a), and calculate their associated error rates.

Given the parameters $(p_{0c}, p_{0p}, p_{1c}, p_{1p}, \alpha, \beta)$ or $(p_{0r}, p_{0p}, p_{1r}, p_{1p}, \alpha, \beta)$ for testing hypothesis (2) or (3) respectively, the key issue of the single-stage design for trinomial response is the determination of the total sample size required, n , and of the acceptance region R , subject to the error constraints α and β . Typically, the solution is obtained by an intensive search algorithm that enumerates over all feasible two-dimensional regions using exact trinomial probabilities. The search algorithms developed in^[24,25] for two-stage designs can be adapted to single-stage designs discussed here.

Lin and Chen^[28] considered the weighted-score method for testing hypothesis (2). Given x_c and x_p , the numbers of observed CRs and of PRs respectively in n patients, it can be shown that the likelihood ratio test criterion for testing a simple trinomial distribution against a simple alternative is a monotone function of the weighted score $s = \omega x_c + x_p$, where ω is a function of $(p_{0c}, p_{0p}, p_{1c}, p_{1p})$. This suggests the following decision rule: accept H_0 if s is at or below the critical value r ; otherwise the treatment is deemed effective and warranting further testing. The larger weight $\omega > 1$ for CR corresponds to $p_{1c}/(p_{1c} + p_{1p}) > p_{0c}/(p_{0c} + p_{0p})$, that

is, the proportion of CR among objective responses (OR = CR + PR) is greater under H_1 than under H_0 . This is consistent with the notion that if higher proportion of CR among OR indicates a more desirable outcome, then CR should be given more weight in deciding whether to accept or reject H_0 . Let $t_\omega(s; p_c, p_p, n)$ be the probability that the weighted score is s in n patients, that is

$$t_\omega(s; p_c, p_p, n) = \sum_{\omega x_c + x_p = s} t(x_c, x_p; p_c, p_p, n)$$

When ω is an integer different (x_c, x_p) 's may give rise to the same score s . In general, ω is a non-integer and each weighted score s is determined by a unique (x_c, x_p) pair, that is, there exists only one (x_c, x_p) pair such that $s = \omega x_c + x_p$. Let $\beta(p_c, p_p)$ denote the cumulative trinomial probability of observing a weighted score that is less than or equal to r in n patients, that is:

$$\beta(p_c, p_p) = \sum_{s \leq r} t_\omega(s; p_c, p_p, n)$$

Then, we can calculate the overall type I and type II errors as $\alpha' = 1 - \beta(p_{0c}, p_{0p})$ and $\beta' = \beta(p_{1c}, p_{1p})$ respectively. The weighted-score method translates the two dimensional search for the acceptance region to a one-dimensional search for the critical point, following the same search strategy for the binomial response model. After obtaining the critical point for the weighted score, it can be converted to the acceptance region in the two-dimensional space of (CR, PR)'s (see example below). It is easy to adapt algorithm developed in^[28] for two-stage designs to single-stage designs discussed here.

Example

For the binomial response with $(p_0, p_1, \alpha, \beta) = (0.3, 0.5, 0.05, 0.10)$, the single-stage design requires a sample size of 53, and if there are 22 or more responses then the treatment would be considered as effective to move on to Phase III trial. Using the weighted-score approach^[28] the single-stage design for the trinomial response with $(p_{0c}, p_{0p}, p_{1c}, p_{1p}, \alpha, \beta) = (0.1, 0.2, 0.2, 0.3, 0.05, 0.10)$ requires a sample size of 49 and a critical value of 22.90 of the weighted score with $\omega = 1.39$. This critical value defines an acceptance region in the two-dimensional space of (CR, PR)'s whose boundary consists the set of lattice points: $\{(0,22), (1,21), (2,20), (3,18), (4,17), (5,15), (6,14), (7,13), (8,11), (9,10), (10,9), (11,7), (12,6), (13,4), (14,3), (15,2), (16,0)\}$. For example, since (0,22) is on the boundary and (13,5) is above the boundary, one would conclude that the treatment is ineffective with 22 PRs, and effective with 13 CRs and 5 PRs. The weighted-score method is more efficient since it requires only 49 patients, less than 53 required by the design based on the binomial model.

MULTI-STAGE DESIGN

In a multi-stage design, patients are entered into the trial in stages; at the end of each stage, a decision based on the number of responses observed thus far is made either to stop the trial or to continue to the next stage. Thus the multi-stage design differs from the single-stage design in a simple but important way: in the single-stage design a fixed number of patients is accrued for the trial, whereas in the multi-stage design, although the maximum number of patients required is fixed, the actual number accrued can vary greatly, depending on the stage at which the trial is terminated. Hence, on average the sample size required in a multi-stage design will generally be smaller than that of the corresponding single-stage design. Most multi-stage designs, e.g., Simon's^[5] two-stage designs, aim to minimize the average sample size according to some optimization criterion. Others, such as Fleming's^[2] multi-stage designs, focus on controlling the spending of error rates among stages and are not based on any optimization criterion. Most multi-stage designs have been limited to two stages, mostly because the additional stages would cause too many unacceptable disruptions during the trial and the logistics of the trial may become too complicated and infeasible.

Binomial Response

We describe several widely used multi-stage designs for testing hypothesis (1). We also describe various modifications, improvements, and extensions recently published in the literature.

Gehan's Two-Stage Design

Gehan's^[1] design is only partially based on the hypothesis testing framework. It controls the error β of rejecting an effective treatment only at the first stage; the final sample size is determined by the precision requirement of estimation. It does not specify an undesirable response rate p_0 , which would be used to control the probability α of accepting an ineffective treatment.

In the first stage, n_1 patients are treated. The trial is terminated and the treatment is abandoned if x_1 , the number of responses, among the n_1 patients is zero. If there is at least one response ($x_1 \geq 1$), then the trial proceeds to the second stage by accruing n_2 additional patients to be treated. Let x_2 denote the number of responses among the n_2 patients. The goal is to obtain an estimator of the true response rate p whose standard error is within a pre-specified limit, σ .

In the first stage, n_1 is chosen to control the probability β (type II error) of abandoning the treatment when the true response rate is at the minimum desirable level p_1 . That is, $\beta \equiv \text{prob}(x_1 = 0 | p = p_1) = (1 - p_1)^{n_1}$. Hence, $n_1 = \log(\beta) / \log(1 - p_1)$. Given that n_1 thus obtained will in general not be an integer, in practice the sample size n_1

required for the first stage is set to be the next integer bigger than this solution. At the end of the trial, the estimate of the response rate is $\hat{p} = x/n$ and $\text{var}(\hat{p}) = p(1-p)/n$, where $x = x_1 + x_2$ and $n = n_1 + n_2$. The precision requirement, $\text{var}(\hat{p}) = \sigma^2$, implies that $n = n_1 + n_2 = p(1-p)/\sigma^2$. Because the true rate p is unknown, we estimate it by x_1/n_1 from the first stage sample, or as suggested in Gehan,^[1] by the more conservative upper 75% confidence limit for p based on the statistic x_1/n_1 .

The sample sizes n_1 and n_2 required for the Gehan's design can be easily obtained by the procedure described above. They are also available in Gehan^[1] for commonly used (β, p_1, σ) 's. As an example, if $(\beta, p_1, \sigma) = (0.05, 0.20, 0.1)$, then the first stage sample size is 14 and the maximum second stage sample size is 11 (note that $n = n_1 + n_2 = p(1-p)/(0.1)^2$ is maximized at $p = 0.50$ with $n_{\max} = 25$; hence $n_2 = 25 - 14 = 11$). If the number of responses at the first stage is $x_1 = 3$, then the upper 75% confidence limit derived from $x_1/n_1 = 3/14$ is 0.34, which gives rise to a total sample size $n = (0.34)(0.66)/(0.01) \approx 23$; hence $n_2 = 9$. On the other hand, if $x_1 = 4$, then a similar calculation leads to $n_2 = 11$.

Fleming Multi-Stage Design

Following O'Brien and Fleming^[29] for the two-sample situation, Fleming^[2] proposed a one sample multi-stage testing procedure for Phase II clinical trials. Let K be the number of stages, $(n_1, n_2, \dots, n_K)$ the number of patients accrued at each stage, $(r_1, r_2, \dots, r_K)$ the set of rejection points, and $(\alpha_1, \alpha_2, \dots, \alpha_K)$ a corresponding set of acceptance points. The acceptance and rejection points form the boundaries of a test region in which $\alpha_i < r_i$. At the terminal stage $\alpha_K = r_K - 1$, so that a decision in favor of or against H_0 is necessarily taken by the end the study. Let $s_j = \sum_{i=1}^j x_i$, where x_i is the number of responses among n_i patients. At the end of each stage, say, stage j , the decision rule is to stop the trial for effectiveness (reject H_0) if $s_j \geq r_j$, or ineffectiveness (accept H_0) if $s_j \leq \alpha_j$, or otherwise to continue the study to the next stage.

The total sample size $n \equiv n_1 + n_2 + \dots + n_K$ is determined by Eq. 5 and is then apportioned arbitrarily, in general equally, among the K stages. Hence, Fleming's^[2] design does not optimize with regard to the sample size within stages. The rejection and acceptance points are determined as follows:

$$r_j = \left\lceil \sum_{i=1}^j n_i p_0 + z_{1-\alpha} \{n p_0 (1 - p_0)\}^{1/2} \right\rceil + 1, j = 1, \dots, K \quad (7)$$

$$\alpha_j = \left\lfloor \sum_{i=1}^j n_i \tilde{p}_1 - z_{1-\alpha} \{n \tilde{p}_1 (1 - \tilde{p}_1)\}^{1/2} \right\rfloor, \quad j = 1, \dots, K-1 \quad (8)$$

where $\tilde{p}_1 = [(n p_0)^{1/2} + (1 - p_0)^{1/2} z_{1-\alpha}]^2 / (n + z_{1-\alpha}^2)$. Note that all r_j and α_j points here depend on the total sample size n .

They are obtained by adjusting the critical values for the standard single-stage test procedure based on the normal approximation. For example, r_j in Eq. 7 is obtained by replacing the standard critical value $z_{1-\alpha}$ for rejecting H_0 at the end of j th stage by $(n / \sum_{i=1}^j n_i)^{1/2} z_{1-\alpha}$. This adjustment makes H_0 difficult to reject at the early stages but easier later on. Since Eq. 7 coincides with Eq. 4 at the K th stage, the test at the end of the trial is performed as the single-stage design. The nominal type I and type II error levels of the single-stage procedure are essentially preserved whenever $K \leq 5$ (see Table 2 in Fleming^[2]).

Remark. Chang et al.^[30] have considered the same multiple testing problem as Fleming^[2] and produced designs that minimized the average of the average sample sizes under the null and alternative hypotheses. They proposed a likelihood ratio method to determine optimal decision boundaries when the sample size at each stage is given. This approach allows them to search for the optimal design from a restricted set without having to do an exhaustive search. In all cases they investigated the optimal design thus obtained was the true global optimum. They were also able to obtain the optimal allocation of sample sizes among K stages and corresponding optimal decision boundaries. They used multiples of 5 as sample size at each stage and presented optimized results for three-stage designs (see Tables 1–3 in Chang et al.^[30]).

Simon's Two-Stage Designs

Simon^[5] restricted the attention to two-stage designs due to practical considerations in the management of multi-center clinical trials. Unlike Fleming^[2] and Chang et al.,^[29] early termination in Simon's^[5] designs occurs for inefficacy only. Let (n_1, n_2) denote the numbers of patients accrued at the first and second stages, and (x_1, x_2) the corresponding numbers of responses observed. The decision rule for Simon's two-stage designs is: At stage 1, stop the trial and reject H_1 if $x_1 \leq r_1$, otherwise continue to stage 2; at stage 2, reject H_1 if $x_1 + x_2 \leq r_2$, otherwise reject H_0 . Here we refer to (r_1, r_2) as the rejection points of H_1 at the end of stage 1 and stage 2. In Fleming's^[2] design these critical values were referred to as acceptance points of H_0 and denoted as a_1 and a_2 .

Let $\beta_1(p)$ and $\beta_2(p)$ denote the probabilities of rejecting H_1 at the end of stages 1 and 2 respectively, that is,

$$\beta_1(p) = P[x_1 \leq r_1 | p] = B(r_1; p, n_1)$$

and

$$\begin{aligned} \beta_2(p) &= P[x_1 > r_1 \text{ and } x_1 + x_2 \leq r_2 | p] \\ &= \sum_{x=r_1+1}^{\min[n_1, r_2]} b(x; p, n_1) B(r_2 - x; p, n_2) \end{aligned}$$

The overall probability of rejecting H_1 is $\beta(p) = \beta_1(p) + \beta_2(p)$. Hence, the overall type I error is $\alpha' = 1 - \beta(p_0)$ and type II error is $\beta = \beta(p_1)$. The expected sample size required is $EN = n_1 + [1 - \beta_1(p)]n_2$.

Given $(p_0, p_1, \alpha, \beta)$, there are many (n_1, n_2, r_1, r_2) 's that satisfy the error constraints on α and β . Within these, the optimal design is the one with the minimum $EN(p_0)$, and the minimax design is the one that has the minimum total sample size n and within this fixed n it has the minimum $EN(p_0)$. By enumeration using exact binomial probabilities, Simon^[5] has described an extensive search procedure to determine both design solutions. Sample sizes and critical values for both optimal and minimax designs, expressed as $(r_1/n_1, r_2/n)$ with $r = r_2$ and $n = n_1 + n_2$, are available in Simon^[5] and also in Chow et al. (Tables 5.3.1 and 5.3.2 in Ref. [31]).

Remark 1. Although the optimal design has the minimum expected sample size, it requires a larger total sample size than the minimax design. In some cases, especially when the patient accrual rate is low, an alternative design between the minimax and optimal designs may be more attractive if its expected sample size (EN) is close to that of the optimal design but with much smaller total sample size than the optimal design. Due to the discreteness of the binomial distribution such alternative designs often exist and can be identified by comparing minimum EN for each n between that of the minimax design and that of the optimal design (or any accruable sample size). Jung et al.^[32] suggested a graphical method to help search for such an alternative design.

Remark 2. Hanfelt et al.^[33] have argued that the relevance of Simon's optimal design depends to some extent on the size of a given institution's program of Phase II trials. If an institution conducts a large number of independent Phase II trials, then it makes sense trying to minimize the average sample size required. They proposed a modification to Simon's optimal design with the objective of minimizing the median sample size. The modified design is appropriate at institutions conducting one or a small number of Phase II trials. The modified optimal design tends to be identical to the Simon's optimal design when the hypothesized response rates are small. In most other cases, the modified design substantially reduces the median sample size. It also tends to have a smaller initial cohort of patients (n_1) than under Simon's design. Generally, the modified design has an expected sample size that exceeds that of Simon's design by one or two patients.

Remark 3. Like Fleming^[2] and Chang et al.,^[30] Shuster^[34] has considered the decision problem where one is allowed to stop the trial at the end of the first stage, either for significance or for non-significance of response rate. This feature is important in some clinical trial settings, such

as Phase II pediatric cancer trials, in which it is in the patients' and investigators' best interests to allow early termination of trials for efficacy. Because one will never know the true value of the response probability p , Shuster proposed a two-stage minimax design that minimizes the globally maximized expected sample size. When early stopping due to significance is not allowed and $p = 1$ (guaranteed response), the trial continues to the second stage and the globally maximized expected sample size reduces to the total sample size. Hence, Shuster's minimax design includes Simon's minimax design as a special case. This new minimax design has lower average sample size than Simon's optimal design under the alternative and at the global maximum. (Note: Simon's^[5] optimal design has minimum average sample size under the null p_0).

Remark 4. At the end of the trial a decision can be made about whether the therapy warrants further evaluation, but an estimate of the true probability of response may also be desired. However, the naïvely used MLE estimator $\hat{p} = (x_1 + x_2)/n$ is generally biased because it does not account for the stopping rule (see Ref. [35]). Jung and Kim^[9] evaluated the uniformly minimum variance unbiased estimator (UMVUE) of p when used with Simon's optimal and minimax designs. They found that the MLE is generally more biased with optimal than minimax designs, and that the bias of the MLE is greatest when $p_0 < p < p_1$; although they did not encounter bias < -0.03 , this difference could be important in subsequent trial design with target effect size of 0.15 to 0.20. They also found some efficiency loss for the UMVUE relative to the MLE, but pointed out that bias is the bigger problem. In a related article, Jung et al.^[11] discussed the issue of p -value calculation at the end of the trial. Koyama and Chen^[12] discussed estimation and p -value calculation when the actual number of patients enrolled deviates from the planned.

Remark 5. Recently, Jung^[14] has proposed a randomized analog of the Simon's two-stage optimal and minimax designs with a concurrent control arm. Sample size in each treatment arm is comparable to that of a single arm trial, thus the overall sample size of the trial is approximately twice that of a single arm study. In order to keep the sample size reasonably small and not to preclude a definitive randomized Phase III trial, they suggest using α of 0.15 or 0.20 and β of 0.20.

Three-Stage Designs

Chen^[36] extended both the optimal and minimax two-stage designs in Simon^[5] to three-stage designs. For the case of an ineffective treatment, this extension can reduce the expected sample size by an average of 10% from that of a two-stage design. The three-stage design is useful when

accrual is slow, so that the temporary suspension of the study and examination of cumulative data does not create an onerous administrative burden.

The decision rule for stopping or continuing the study is as follows: At stage 1, stop and reject H_1 if $x_1 \leq r_1$, otherwise continue to stage 2; at stage 2, stop and reject H_1 if $x_1 + x_2 \leq r_2$, otherwise continue to stage 3; at stage 3, reject H_1 if $x_1 + x_2 + x_3 \leq r_3$, otherwise reject H_0 . Let $\beta_1(p)$, $\beta_2(p)$ and $\beta_3(p)$ denote the probabilities of rejecting H_1 at the end of stages 1, 2, and 3 respectively, that is,

$$\begin{aligned}\beta_1(p) &= P[x_1 \leq r_1 | p] = B(r_1; p, n_1) \\ \beta_2(p) &= P[x_1 > r_1 \text{ and } x_1 + x_2 \leq r_2 | p] \\ &= \sum_{x_1=r_1+1}^{\min[n_1, r_2]} b(x_1; p, n_1) B(r_2 - x_1; p, n_2)\end{aligned}$$

and

$$\begin{aligned}\beta_3(p) &= P[x_1 > r_1, x_1 + x_2 > r_2 \text{ and } x_1 + x_2 + x_3 \leq r_3 | p] \\ &= \sum_{x_1=r_1+1}^{\min[n_1, r_2]} \sum_{x_2=r_2+1-x_1}^{\min[n_2, r_3-x_1]} b(x_1; p, n_1) b(x_2; p, n_2) \\ &\quad \times B(r_3 - x_1 - x_2; p, n_3).\end{aligned}$$

The overall probability of rejecting H_1 is $\beta(p) = \beta_1(p) + \beta_2(p) + \beta_3(p)$. Hence, the overall type I and type II errors are $\alpha' = 1 - \beta(p_0)$ and $\beta = \beta(p_1)$ respectively. The expected sample size required is $EN(p) = n_1 + (1 - \beta_1(p))n_2 + (1 - \beta_1(p) - \beta_2(p))n_3$.

Given $(p_0, p_1, \alpha, \beta)$, there are many $(r_1, n_1, r_2, n_2, r_3, n_3)$'s that satisfy the error constraints on α and β . Among them, the optimal solution is the one with the minimum $EN(p_0)$, and the minimax solution is the one that has the minimum total sample size n and within this fixed n has the minimum $EN(p_0)$. Chen^[36] wrote a FORTRAN program to search exhaustively for these two solutions based on enumeration of exact binomial probabilities. Sample sizes and critical values under a variety of design parameters are available in Chen (Tables I–IV in Ref. [36]) and also in Chow et al. (Tables 5.3.9–5.3.12 in Ref. [31]).

Ensign et al.^[6] have observed that Simon's^[5] optimal two-stage design does not allow early termination when there are many failures at the start. Therefore, motivated by ethical reasons to reduce the number of patients receiving treatment with very low response rate, they proposed a three-stage design as a combination of the first stage in Gehan's^[1] design and Simon's^[5] optimal two-stage design. This design can be viewed as a special case of Chen's^[36] optimal three-stage design with the first stage sample size n_1 set to be at least 5 and the rejection point r_1 to be 0. Sample sizes and critical values of Ensign et al.'s^[6] optimal three-stage designs, under a variety of design parameters,

are available in Ensign et al. (Tables 1 and 2 in Ref. [6]) and also in Chow et al. (Tables 5.3.7 and 5.3.8 in Ref. [31]). Chen^[36] corrected errors in many entries of these tables. In general, the average sample size of Ensign et al.'s^[6] optimal three-stage design is smaller than that of Simon's optimal two-stage design, although the maximum sample size can be smaller or larger when $p_0 \leq 0.30$. It has the greatest utility when the true response rate of the treatment is low. However, the restrictions on (r_1, n_1) increase $EN(p_0)$ and make the first stage redundant for large p_0 .

Flexible Designs

In the cooperative group setting consisting of many centers, practical considerations make it difficult to arrive at the planned sample size exactly. Sometimes, the patient accrual may fall short of the required sample size; other times, a center would have to accept more patients even though their original quota of patients has been met. In such cases, actual sample size at each stage can deviate slightly from the exact design. To overcome such difficulties, Green and Dahlberg^[37] have proposed various adaptive two-stage designs to address the decision-making issues based on the attained or actual sample sizes in each stage, not the planned sample size. Herndon^[38] has proposed a two-stage design that allows patient accrual to continue while the results of the first stage are reviewed. Unlike Simon's optimal and minimax designs, these designs do not aim to minimize the number of patients tested on an ineffective treatment.

Chen and Ng^[39] have proposed a flexible design as a collection of two-stage designs where the first stage size is in a set of consecutive values $(n_1, \dots, n_k)$ and the total sample size is also in another set of consecutive values $(N_1, \dots, N_k)$, and each of k^2 possible designs has the same probability of occurrence. Denote the critical value for the first stage as r_i when the sample size is n_i , and for the second stage as R_j when the sample size is N_j . Each member, say (r_i, n_i, R_j, N_j) , of the design collection follows the same decision rule as Simon's two-stage designs, that is as follows: At stage 1, stop and reject H_1 if there are r_i or fewer responses in the n_i patients, otherwise continue to stage 2; at stage 2, reject H_1 if there are R_j or fewer responses out of N_j patients, otherwise reject H_0 and conclude that the treatment is a candidate for further testing. For each member design, type I and type II errors and expected sample size (EN) are defined as in the Simon's two-stage design. Type I and type II errors and EN of the flexible design are then defined as the average of these respective measures over all members within the collection. There are many collections of (r_i, n_i, R_j, N_j) 's that satisfy the α and β constraints. The optimal flexible design is the one that has the minimum average EN when $p = p_0$. The minimax flexible design is the one that has the minimum N_k , and the minimum average EN within this fixed N_k when $p = p_0$. In application, $k = 8$ is suggested

with target sizes at the middle of ranges. Tables providing sample sizes and critical values for flexible two-stage optimal designs and minimax designs are available in Chen and Ng (Tables I–IV in Ref. [39]) and also in Chow et al. (Tables 5.3.3–5.3.6 in Ref. [31]). Masaki et al. [40] proposed an extension of Chen and Ng's method in which the actual sample size is allowed to deviate from the planned sample size symmetrically by up to two, and the probability of a given sample size decreases as it deviates from the target sample size.

Remark. Instead of the average, it would seem reasonable to define the error rates and EN of a flexible design as the median or maximum of these measures within the collection. However, the statistical property of using these different criteria would need to be further studied.

Designs that Jointly Evaluate Efficacy and Toxicity

For some treatments, evaluation of toxicity is equally important to assessment of efficacy, and it is desirable to incorporate toxicity monitoring into the early stopping rule of the trial. Recognizing this need, Bryant and Day^[41] proposed two-stage designs, similar in structure to Simon's^[5] optimal two-stage designs, which jointly evaluate clinical response and toxicity. From current treatment results, both undesirable and desirable levels for the true response rate p_r , (p_{0r}, p_{1r}) , and for the true non-toxicity rate p_t , (p_{0t}, p_{1t}) , are specified. The null hypothesis $H_0: p_r \leq p_{0r}$ or $p_t \leq p_{0t}$ is to be tested against the alternative hypothesis $H_1: p_r \geq p_{1r}$ and $p_t \geq p_{1t}$.

Let (n_1, n_2) denote the numbers of patients accrued at the first and second stages, (x_1, x_2) be the corresponding numbers of clinical responses, and (y_1, y_2) be that of patients who do not experience toxicity. The decision rule for the trial is as follows: At stage 1, stop the trial and reject H_1 if either $x_1 \leq r_1$ or $y_1 \leq t_1$, otherwise continue to stage 2; at stage 2, reject H_1 if either $x_1 + x_2 \leq r_2$ or $y_1 + y_2 \leq t_2$, otherwise reject H_0 and recommend the treatment for further testing. The probabilities of rejecting at the end of stages 1 and 2 respectively and the expected sample size (EN) required are evaluated (Eqs. 4.6–4.9 in Bryant and Day^[41]), based on the joint distribution of response and toxicity under various assumptions on the bivariate association. Type I errors at (p_{0r}, p_{1r}) and (p_{1r}, p_{0r}) are subjected to constraints α_r and α_t . Given $(p_{0r}, p_{1r}, p_{0t}, p_{1t}, \alpha_r, \alpha_t, \beta)$, there are many $(n_1, n_2, r_1, r_2, t_1, t_2)$'s that satisfy the error constraints on α_r, α_t and β . Within these, the optimal solution is the one that minimizes $\max_{\phi > 0} \max \{E_{01}(\phi), E_{10}(\phi)\}$, where ϕ is an unspecified association parameter, and $E_{01}(\phi)$ and $E_{10}(\phi)$ denote EN evaluated at (p_{0r}, p_{1r}) and (p_{1r}, p_{0r}) , respectively, under ϕ . Bryant and Day^[41] provide a table for sample sizes and critical points under a limited set of design parameters. The $\max \{E_{01}, E_{10}\}$ is always bigger than the corresponding EN from the Simon's optimal two-stage design, reflecting the additional cost of addressing the uncertainty in toxicity.

Remark. Conaway and Petroni^[42] and Thall and Cheng^[43] have also considered multi-stage designs to monitor anti-tumor activity and toxicity. Unlike,^[41] designs proposed in Ref. [42] controlled type I error at (p_{0r}, p_{0r}) and at the maximum over the entire null hypothesis region. Exhaustive search algorithm based on enumeration of exact bivariate binomial probabilities was used in Refs. [41,42] to search for the optimal designs. On the other hand, optimal two-stage designs developed in Ref. [43] were based on normal theory, using the arcsin square-root transformations on the bivariate binomial probabilities. Designs for the Phase II trials allowing for a trade-off between response and toxicity was addressed in Refs. [43,44].

Randomized Discontinuation Design

Randomized discontinuation design (RDD) is a two-stage design in which all enrolled patients receive the experimental treatment in the first stage, and a subset of patients is selected at the end of the first-stage based on response to treatment. In the second stage, only this subset of patients are randomized to continue their current treatment or to switch to placebo. This design is an example of an enrichment design, in which the objective is to select a more homogeneous group of patients who are likely to show benefit from treatment and to exclude others who may dilute the effect if it exists. For example, an RDD can be used to study the effects of a long-term cytotoxic therapy,^[45] where patients responding after the first stage are randomized, and those with stable or progressive disease are taken off study. More recently, however, RDD has been proposed as an alternative Phase II trial design for the study of molecular targeted (cytostatic) agents in an attempt to identify a subpopulation that may have a clinical benefit without experiencing objective response. An example of this is a trial in renal cell cancer^[16] in which patients with stable disease (SD) at the end of the first stage were randomized to discontinue therapy, whereas responders and patients with progressive disease were taken off study (responders were allowed to continue treatment for ethical reasons). Different types of endpoints, such as response rate with longer therapy, proportion of patients with progressive disease, or time to disease progression, can be used to compare the randomized treatment arms at the end of the second stage.

Power and sample size for RDDs are usually determined using simulations. For response or progression endpoints, tumor size data can be generated using a tumor growth model,^[46] and changes in tumor size at a fixed time point can be classified as response, stable disease or progression. For time to event endpoints, either a tumor growth model or Weibull distribution can be used to generate event times. Compared to clinical trials with upfront randomization, Freidlin and Simon^[48] and Fu et al.^[47] found some scenarios where RDD is more efficient, especially if the proportion of “sensitive” patients (e.g., patients who express the presumed molecular target of the agent) is relatively small

in the population and no reliable assay for identifying such patients otherwise is available. They note, however, that their conclusions are sensitive to the data generating model.

Although the RDD has been proposed as an alternative trial design for targeted cytostatic agents, some of its disadvantages include complicated interpretation of the results and estimation, and potentially biased results due to the unexpected carry-over effect. Finally, if the proportion of sensitive patients and the effect size are both small, both the RDD and the upfront-randomized Phase II trials may require too large a sample size to be feasible.

Trinomial Response

A multi-stage design for trinomial response can be considered as an extension of a single-stage design for trinomial response or of a multi-stage design for binomial response. Lin and Chen^[28] and Panageas et al.^[24] extended Simon's^[5] two-stage designs for binomial response to trinomial response where complete and partial responses are differentiated. On the other hand, Zee et al.,^[25] Dent et al.,^[48] and Freidlin et al.^[26] extended Fleming's^[2] multi-stage design to trinomial response where stable disease and disease progression are differentiated.

The various hypotheses for trinomial response have been described in the section “Trinomial Response.” We shall focus on testing hypothesis (2). Let (n_1, n_2) denote the numbers of patients accrued at the first and second stages and let $n = n_1 + n_2$. Also let x_{1c} and x_{1p} denote the numbers of CRs and PRs respectively among n_1 patients, and x_c and x_p the numbers of CRs and PRs respectively among n patients. The decision rule for the trial is based on acceptance regions R_1 and R . Each of these regions is defined as in section “Single-Stage Design.” At the end of the first stage, if the (x_{1c}, x_{1p}) pair falls in the region R_1 , the trial will be terminated early. If the pair falls above the region, the trial will proceed to the second stage. At the end of the second stage, if the (x_c, x_p) pair falls in the region $R_2 = R - R_1$, then the treatment will be deemed ineffective. If it falls anywhere above the region R , the treatment will be recommended for further testing.

Let $\beta_1(p_c, p_p)$ and $\beta_2(p_c, p_p)$ denote the cumulative trinomial probabilities in the regions R_1 and R_2 respectively. These probabilities can be obtained by replacing R in Eq. 6 with R_1 and R_2 respectively. Let $\beta(p_c, p_p) = \beta_1(p_c, p_p) + \beta_2(p_c, p_p)$. Then the overall type I and type II errors are $\alpha' = 1 - \beta(p_{0c}, p_{0p})$ and $\beta' = \beta(p_{1c}, p_{1p})$. The expected sample size required is $EN(p_c, p_p) = n_1 + [1 - \beta_1(p_c, p_p)]n_2$. Given the parameters $(p_{0c}, p_{0p}, p_{1c}, p_{1p}, \alpha, \beta)$, the design problem involves finding both regions R_1 and R , subject to error constraints α and β . Among all feasible (R_1, R) 's the optimal design is the one with the minimum $EN(p_{0c}, p_{0p})$, and the minimax design is the one with the minimum total sample size n , and within this fixed n with the minimum $EN(p_{0c}, p_{0p})$. Optimal and minimax designs can be obtained using exhaustive search

algorithm based on enumeration of exact trinomial probabilities. Such algorithm is computationally highly intensive. Panageas et al.^[24] considered only optimal designs and tabulated sample sizes and boundaries of R_1 and R under a variety of design parameters.

Lin and Chen^[28] used the weighted-score approach to differentiate between the complete and partial responses. From the number of complete responses x_c and the number of partial responses x_p , a weighed score $s^* = \omega x_c + x_p$, $\omega > 1$, is computed. The decision rule is as follows: At stage 1, if the weighted score of responses among n_1 patients is s_1 or less, the study is terminated and the treatment is declared to be ineffective. Otherwise, the trial is continued to stage 2. At stage 2, if the weighted score out of n patients is s or less, the study is terminated and the treatment is declared ineffective. Otherwise, the treatment warrants to be tested further.

Let $\beta_1(p_c, p_p)$ and $\beta_2(p_c, p_p)$ denote the probabilities of rejecting H_1 at the end of stage 1 and stage 2 respectively, that is

$$\beta_1(p_c, p_p) \equiv T_\omega(s_1; p_c, p_p, n_1) = \sum_{s^* \leq s_1} t_\omega(s^*; p_c, p_p, n_1).$$

and

$$\beta_2(p_c, p_p) = \sum_{s_1 < s^* \leq \min[\omega n_1, s]} t_\omega(s^*; p_c, p_p, n_1) T_\omega(s - s^*; p_c, p_p, n_2).$$

The overall probability of rejecting a treatment is $\beta(p_c, p_p) = \beta_1(p_c, p_p) + \beta_2(p_c, p_p)$. Hence, the overall type I and type II errors are $\alpha' = 1 - \beta(p_{0c}, p_{0p})$ and $\beta' = \beta(p_{1c}, p_{1p})$ respectively. The expected sample size $EN = n_1 + [1 - \beta_1(p_c, p_p)]n_2$.

There are many $(s_1/n_1, s/n)$'s which satisfy α and β requirements for a specific $(p_{0c}, p_{0p}, p_{1c}, p_{1p})$. Within these the optimal solution is defined as the one which has the minimum EN when $p_c = p_{0c}$ and $p_p = p_{0p}$. A minimax solution is the one which has the minimum n , and within this fixed n the minimum EN when $p_c = p_{0c}$ and $p_p = p_{0p}$. An extensive search procedure based on exact calculation of trinomial probabilities is used to find the optimal and minimax designs. As indicated in the single-stage design, the weighted-score method only needs to search for the critical values s_1 and s which can then be used to determine boundaries of acceptance regions R_1 and R respectively. Hence, the search based on weighted score is computationally much less intensive compared to the two-dimensional search of Panageas et al.^[24] Design solutions $(s_1/n_1, s/n)$'s under a variety of design parameters are available in Lin and Chen (Tables I and II in Ref. [28]).

Example: Given $(p_{0c}, p_{0p}, p_{1c}, p_{1p}) = (0.033, 0.067, 0.12, 0.18)$ and $(\alpha, \beta) = (0.10, 0.10)$, the optimal two-stage design of Panageas et al.^[24] gives rise to $n_1 = 13$, $n = 29$, $EN(p_{0c}, p_{0p}) = 19.06$, and acceptance regions R_1 and R

with boundaries $\{(0,1), (1,0)\}$ and $\{(0,5), (1,4), (2,2), (3,1)\}$ respectively. On the other hand, the optimal two-stage design based on the weighted score approach of Lin and Chen^[28] gives rise to $(s_1/n_1, s/n) = (1.23/13, 5.23/29)$ and $EN(p_{0c}, p_{0p}) = 19.06$, where scores s_1 and s are computed with weight $\omega = 1.23$. The scores $s_1 = 1.23$ and $s = 5.23$ determine respectively the boundaries of acceptance regions R_1 and R to be $\{(0,1), (1,0)\}$ and $\{(0,5), (1,4), (2,2), (3,1), (4,0)\}$. Both approaches have the same n_1 , n , $EN(p_{0c}, p_{0p})$, and R_1 . However, the weighted-score approach includes an additional boundary point (4,0) in the acceptance region R . Hence, with four CRs one would conclude that the treatment is effective under the design of Panageas et al.^[24] but ineffective under the design based on the weighted score approach. Both designs are more efficient than Simon's two-stage optimal design when CR and PR rates are combined (summed) which requires $n_1 = 12$, $n = 35$ with $EN(p_0) = 19.84$.

Multi-stage design for testing hypothesis (3) was originally proposed by Zee et al.^[25] and was applied in Phase II trials by Dent et al.^[48] Such design extends Fleming's^[2] design that focuses on controlling the spending of the error rate among stages and is not based on any optimization criterion with respect to the average sample size. Zee et al.^[25] also use exhaustive search algorithm to obtain the design solution. As indicated earlier, hypothesis (3) is pragmatically equivalent to hypothesis (3a), and hypothesis (3a) is mathematically equivalent to hypothesis (2). Hence, the designs discussed above for hypothesis (2) are applicable for either hypothesis (3) or (3a).

Kocherginsky et al.^[49] used a similar approach to design a two-stage trial which satisfies α and β constraints for multiple alternative hypotheses chosen to represent different scenarios of what would be considered drug activity. More specifically, they chose four alternative hypotheses to represent (i) a clinically meaningful increase in response rate and no change in early progression relative to H_0 , (ii) no change in response rate and a clinically meaningful decrease in progression rate only, (iii) moderate improvement in both response and progression rates, and (iv) strong improvement in both response and progression. For practical considerations, the overall sample size was chosen to match Simon's optimal design for testing H_0 vs alternative hypothesis (i), and an exhaustive search was used. The design minimizing the expected sample size was then chosen among all designs satisfying α and β constraints for multiple alternative hypotheses.

Sun et al.^[15] proposed a randomized multi-stage clinical trial for trinomial outcome, and also consider multiple alternative hypotheses representing activity in either endpoint (response or early progression)

$$H_{A1} : p_T > p_C \quad \text{and} \quad q_T = q_C$$

$$H_{A2} : p_T = p_C \quad \text{and} \quad q_T < q_C$$

$$H_{A3} : p_T > p_C \quad \text{and} \quad q_T < q_C$$

where p_T and p_C are response rates in treatment (T) and control (C) arms, and q_T and q_C are the respective progression rates. In this design, $2n_1$ patients are randomized during the first stage, and the decision to continue to stage 2 is based on satisfying either of the following mutually exclusive conditions:

Condition1: $X_{1T} - X_{1C} \geq D_1$ and $Y_{1T} - Y_{1C} < D_1 / 2$

Condition2: $-D_1 / 2 < X_{1T} - X_{1C} < D_1$ and $Y_{1T} - Y_{1C} \leq -D_1$

where X_{1T} and X_{1C} represent the number of responses in T and C arms at the end of stage 1, Y_{1T} and Y_{1C} represent the number of early progressions, cutoff D_1 is a boundary parameter of the design, and cutoff $D_1/2$ serves as the non-inferiority margin. If the trial continues to stage 2, an additional n_2 patients are enrolled in each treatment arm, and a similar decision rule is applied at the end of the trial with different cutoff D_2 . For this design, the total sample size needed to detect a 0.20 difference in response and progression rates between the treatment arms ranges between 46–58 for $\alpha = 0.20$ and $\beta = 0.20$ under H_{A3} , and larger sample sizes will be required if all three alternative hypotheses are to maintain $\beta = 0.20$.

CONCLUSION

For ethical reasons, multi-stage designs have become indispensable in conducting Phase II cancer trials. The most important feature of such designs is that they allow for early termination when the treatment is ineffective (or effective in some trial settings). As a consequence, the average sample size is reduced even without formal optimization with respect to the sample size within stages. However, most multi-stage designs were developed with the specific goal to optimize efficiency of the trial.

The gain in efficiency of a multi-stage design does not come without cost. For example, problems and difficulties are encountered with the suspension of patient accrual at the end of each stage of a trial. During the accrual down time, the interest and enthusiasm of participating investigators may wane, and accrual momentum during the subsequent stage of the study suffers. Thus, multi-stage designs with more than three stages are usually not recommended.

Even with only two stages, study suspensions in multi-stage designs can result in the actual (attained) sample size being different from the planned sample size, especially in multi-center studies. Thus, development of designs with greater flexibility and innovation, with improved statistical properties and practical considerations is an important area that requires further research. Other important areas for further research include possible extensions, modifications, or improvements on designs based on trinomial response and designs that jointly evaluate efficacy and toxicity.

Although efficient designs that address ethical concerns are important for Phase II cancer trials, many factors, such as deviations from planned sample sizes or proper analysis and inference in the end of the trial, may influence outcome and the decision on whether to proceed to a Phase III trial. These and other issues specific to the scientific context of the disease and experimental agent need to be discussed among clinical investigators and biostatisticians conducting the trials.

Selection of the primary endpoint has become an important factor in clinical trial design due to the development of targeted molecular therapies. Tumor response may no longer be an appropriate endpoint for trials involving such agents because their mechanism of action, which is inhibition of tumor growth or disruption of other cellular processes, may lead to clinical benefit without tumor response. Several innovative clinical trial designs and general guidance on choosing the primary endpoint have been recently discussed in the literature,^[19,20,50] but this remains an active research area.

Finally, although investigators are becoming more aware of the necessity for rigorous statistical methods, the number of Phase II studies that report a statistical plan is still disappointingly low (19.7% and 35.2% as reported in Refs. [7] and [8], respectively), which leads to concerns about the quality of such studies. We hope that this review of multiple-stage designs for phase II cancer trials will help investigators to better understand and appreciate the issues involved in phase II clinical trial design and conduct, and will guide them in making a proper choice among the many available designs.

REFERENCES

1. Gehan, E.A. The determination of the number of a patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Therapeutics* **1961**, *13* (1), 346–353.
2. Fleming, T.R. One sample multiple testing procedure for Phase II clinical trials. *Biometrics* **1982**, *38* (1), 143–151.
3. Mariani, L.; Marubini, E. Design and analysis of Phase II cancer trials: A review of statistical methods and guidelines for medical researchers. *Intl. Stat. Rev.* **1996**, *64* (1), 61–88.
4. Kramar, A.; Potvin, D.; Hill, C. Multistage designs for Phase II clinical trials: statistical issues in cancer research. *Br. J. Cancer* **1996**, *74* (8), 1317–1320.
5. Simon, R. Optimal two-stage designs for Phase II clinical trials. *Control Clin. Trials* **1989**, *10* (1), 1–10.
6. Ensign, L.G.; Gehan, E.A.; Kamen, D.S.; Thall, P.F. An optimal three-stage design for Phase II clinical trials. *Stat. Med.* **1994**, *13* (17), 1727–1736.
7. Mariani, L.; Marubini, E. Content and quality of currently published Phase II cancer trials. *J. Clin. Oncol.* **2000**, *18* (2), 429–436.
8. Perrone, F.; Di Maio, M.; De Maio, E.; Maione, P.; Ottaviano, A.; Pensabene, M.; Di Lorenzo, G.; Lombardi,

- A.V.; Signoriello, G.; Gallo, C. Statistical design in Phase II clinical trials and its application in breast cancer. *Lancet Oncol.* **2003**, *4* (5), 305–311.
9. Jung, S.; Kim, K.M. On the estimation of the binomial probability in multistage clinical trials. *Stat. Med.* **2004**, *23* (6), 881–896. Available at: <http://dx.doi.org/10.1002/sim.1653>.
 10. Tsai, W.; Chi, Y.; Chen, C. Interval estimation of binomial proportion in clinical trials with a two-stage design. *Stat. Med.* **2008**, *27* (1), 15–35. Available at: <http://dx.doi.org/10.1002/sim.2930>.
 11. Jung, S.H.; Owzar, K.; George, S.L.; Lee, T. P-value calculation for multistage phase II cancer clinical trials. *J. Biopharm. Stat.* **2006**, *16* (6), 765–775. Available at: http://pdfserve.informaworld.com/225313_751311230_762501449.pdf.
 12. Koyama, T.; Chen, H. Proper inference from Simon's two-stage designs. *Stat. Med.* **2008**, *27* (16), 3145–3154. Available at: <http://dx.doi.org/10.1002/sim.3123>.
 13. Rubinstein, L.; Crowley, J.; Ivy, P.; Leblanc, M.; Sargent, D. Randomized phase II designs. *Clin. Cancer Res.* **2009**, *15* (6), 1883–1890. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19276275>.
 14. Jung, S. Randomized phase II trials with a prospective control. *Stat. Med.* **2008**, *27* (4), 568–583. Available at: <http://dx.doi.org/10.1002/sim.2961>.
 15. Sun, L.Z.; Cong, C.; Patel, K. Optimal two-stage randomized multinomial designs for phase II oncology trials. *J. Biopharm. Stat.* **2009**, *19* (3), 485–493. Available at: DOI: 10.1080/10543400902802417.
 16. Rosner, G.L.; Stadler, W.M.; Ratain, M.J. Randomized discontinuation design: application to cytostatic anti-neoplastic agents. *J. Clin. Oncol.* **2002**, *20* (22), 4478–4484. Available at: <http://jco.ascopubs.org/cgi/content/abstract/20/22/4478>.
 17. Thall, P.F. A review of phase 2–3 clinical trial designs. *Lifetime Data. Anal.* **2008**, *14* (1), 37–53. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/17763973>.
 18. Korn, E.L.; Arbuck, S.G.; Pluda, J.M.; Simon, R.; Kaplan, R.S.; Christian, M.C. Clinical trial designs for cytostatic agents: are new approaches needed? *J. Clin. Oncol.* **2001**, *19* (1), 265–272. Available at: <http://jco.ascopubs.org/cgi/content/abstract/19/1/265>.
 19. Booth, C.M.; Calvert, A.H.; Giaccone, G.; Lobbezoo, M.W.; Eisenhauer, E.A.; Seymour, L.K. Design and conduct of phase II studies of targeted anticancer therapy: recommendations from the task force on methodology for the development of innovative cancer therapies (MDICT). *Eur. J. Cancer.* (Oxford, 1990). **2008**, *44* (1), 25–29. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/17845846>.
 20. Dhani, N.; Tu, D.; Sargent, D.J.; Seymour, L.; Moore, M.J. Alternate endpoints for screening phase II studies. *Clin. Cancer. Res.* **2009**, *15* (6), 1873–1882. Available at: <http://clincancerres.aacrjournals.org/cgi/content/abstract/15/6/1873>.
 21. Rubinstein, L.V.; Korn, E.L.; Freidlin, B.; Hunsberger, S.; Ivy, S.P.; Smith, M.A. Design issues of randomized phase II trials and a proposal for phase II screening trials. *J. Clin. Oncol.* **2005**, *23* (28), 7199–7206. Available at: <http://jco.ascopubs.org/cgi/content/abstract/23/28/7199>.
 22. Tan, S.B.; Machin, D. Bayesian two-stage designs for Phase II clinical trials. *Stat. Med.* **2002**, *21* (14), 1991–2012.
 23. Therasse, P.; Arbuck, S.; Eisenhauer, E.; Wanders, J.; Kaplan, R.; Rubinstein, L., et al. New guidelines to evaluate the response to treatment in solid tumors. *J. Natl. Cancer Inst.* **2000**, *92* (3): 205–216. Available at: <http://jnci.oxford-journals.org.proxy.uchicago.edu/cgi/reprint/92/3/205>.
 24. Panageas, K.S.; Smith, A.; Gonen, M.; Chapman, P.B. An optimal two-stage Phase II design utilizing complete and partial response information separately. *Control Clin. Trials* **2002**, *23* (4), 367–379.
 25. Zee, B.; Melnychuk, D.; Dancy, J.; Eisenhauer, E. Multinomial Phase II cancer trials incorporating response and early progression. *J. Biopharm. Stat.* **1999**, *9* (2), 351–363.
 26. Freidlin, B.; Dancy, J.; Korn, E.L.; Zee, B.; Eisenhauer, E. Multinomial Phase II trial design (correspondence). *J. Clin. Oncol.* **2002**, *20* (2), 599.
 27. A'Hern, R.P. Sample size tables for exact single-stage Phase II designs. *Stat. Med.* **2001**, *20* (5), 859–866.
 28. Lin, S.P.; Chen, T.T. Optimal two-stage designs for Phase II clinical trials with differentiation of complete and partial responses. *Comm. Stat. – Theory Method* **2000**, *29* (5/6), 923–940.
 29. O'Brien, P.C.; Fleming, T.R. A multiple testing procedure for clinical trials. *Biometrics* **1979**, *35* (3), 549–556.
 30. Chang, M.N.; Therneau, T.M.; Wieand, H.S.; Cha, S.S. Designs for group sequential Phase II clinical trials. *Biometric* **1987**, *43* (4), 865–874.
 31. Chow, S.C.; Shao, J.; Wang, H. *Sample Size Calculations In Clinical Research*; Marcel Dekker: New York, 2003.
 32. Jung, S.H.; Carey, M.; Kim, K.M. Graphical search for two-stage designs for Phase II clinical trials. *Control Clin. Trials* **2001**, *22* (4), 367–372.
 33. Hanfelt, J.J.; Slack, R.S.; Gehan, E.A. A modification of Simon's optimal design for Phase II trials when the criterion is median sample size. *Control Clin. Trials* **1999**, *20* (6), 555–566.
 34. Shuster, J. Optimal two-stage designs for single arm Phase II cancer trials. *J. Biopharm. Stat.* **2002**, *12* (1), 39–51.
 35. Chang, M.; Wieand, H.; Chang, V. The bias of the sample proportion following a group sequential phase II clinical trial. *Stat. Med.* **1989**, *8* (5), 563–570. Available at: <http://www3.interscience.wiley.com/journal/113396405/abstract>.
 36. Chen, T.T. Optimal three-stage designs for Phase II cancer clinical trials. *Stat. Med.* **1997**, *16* (23), 2701–2711.
 37. Green, S.J.; Dahlberg, S. Planned versus attained design in Phase II clinical trials. *Stat. Med.* **1992**, *11* (7), 853–862.
 38. Herndon, J.E. II. A design alternative for two-stage, Phase II, multicenter cancer clinical trials. *Control Clin. Trials* **1998**, *19* (5), 440–450.
 39. Chen, T.T.; Ng, T.H. Optimal flexible designs in Phase II clinical trials. *Stat. Med.* **1998**, *17* (20), 2301–2312.
 40. Masaki, N.; Koyama, T.; Yoshimura, I.; Hamada, C. Optimal two-stage designs allowing flexibility in number of subjects for phase II clinical trials. *J. Biopharm. Stat.* **2009**, *19* (4), 721–731. Available at: <http://proxy.uchicago.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=41998459&loginpage=Login.asp&site=ehost-live&scope=site>.

41. Bryant, J.; Day, R. Incorporating toxicity considerations into the design of two-stage Phase II clinical Trials. *Biometrics* **1995**, *51* (4), 1372–1383.
42. Conaway, M.R.; Petroni, G.R. Bivariate sequential designs for Phase II trials. *Biometrics* **1995**, *51* (2), 656–664.
43. Thall, P.F.; Cheng, S.C. Optimal two-stage designs for clinical trials based on safety and efficacy. *Stat. Med.* **2001**, *20* (7), 1023–1032.
44. Conaway, M.R.; Petroni, G.R. Designs for Phase II trials allowing for a trade-off between response and toxicity. *Biometrics* **1996**, *52* (4), 1375–1386.
45. Kopec, J.; Abrahamowicz, M.; Esdaile, J. Randomized discontinuation trials: utility and efficiency. *J. Clin. Epidemiol.* **1993**, *46* (9), 959–971. Available at: <http://cat.inist.fr/?aModele=afficheN&cpsidt=4862483>.
46. Freidlin, B.; Simon, R. Evaluation of randomized discontinuation design. *J. Clin. Oncol.* **2005**, *23* (22), 5094–5098. Available at: <http://jco.ascopubs.org/cgi/content/abstract/23/22/5094>.
47. Fu, P.; Dowlati, A.; Schluchter, M. Comparison of power between randomized discontinuation design and upfront randomization design on progression-free survival. *J. Clin. Oncol.* **2009**, JCO.2008.19.6709. Available at: <http://jco.ascopubs.org/cgi/content/abstract/JCO.2008.19.6709v2>.
48. Dent, S.; Zee, B.; Dancey, J.; Hanauske, A.; Wanders, J.; Eisenhauer, E. Application of a new multinomial Phase II stopping rule using response and early progression. *J. Clin. Oncol.* **2001**, *19* (3), 785–791.
49. Kocherginsky, M.; Cohen, E.E.; Karrison, T. Design of phase II cancer trials for evaluation of cytostatic/cytotoxic agents. *J. Biopharm. Stat.* **2009**, *19* (3), 524–529. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19384693>.
50. Cannistra, S.A. Phase II trials in journal of clinical oncology. *J. Clin. Oncol.* **2009**, *27* (19), 3073–3076. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19451415>.

Multiplicative Intensity Models

Keumhee Carriere Chough

University of Alberta, Edmonton, Alberta, Canada

Abdulkadir Hussein

University of Windsor, Windsor, Ontario, Canada

Abstract

Multiplicative intensity (MI) models are used to analyze recurrent event data, where subjects repeatedly experience an event, or events of various types, throughout the study. This entry discusses the Andersen-Gill model of multiplicative intensity, the most commonly used model.

INTRODUCTION

Multiplicative intensity (MI) models are used to analyze recurrent event data, where subjects repeatedly experience an event, or events of various types, throughout the study, for example, a certain type of tumor (or different types of tumors) occurring in laboratory animals under observation over a period of time. In such situations, the objective of the analysis is to estimate the rate (intensity) at which the events occur and the effect of certain covariates on the intensity.

Multiplicative intensity models are extensions of Cox's proportional hazards model,^[1] in that they allow multiple events per subject and time-dependent covariates. Multiplicative intensity models were first introduced in their simplest form by Aalen.^[2] Then Anderson and Gill^[3] extended the models to apply to situations where the intensity function can have time-dependent and/or time-independent covariates. Since these seminal papers were published, MI models have been used in many applications and extended in many directions. For example, Tuli and coworkers^[4] used an MI model to analyze the repeated failures of shunts from three different causes among children with hydrocephalus. Hussein and Carrière^[5] used an MI model to identify covariates that cause discontinuity of care among a cohort of patients. Lin et al.^[6] used an MI model to identify factors contributing to motorcycle crashes among junior college students.

In this entry, we discuss the most commonly used MI model, i.e., the Andersen-Gill model. For detailed theoretical developments and practical applications, see Refs. [7–9].

THE MULTIPLICATIVE INTENSITY MODEL AND NOTATIONS

Consider recurrent event data collected from n subjects. Let $N_{hi}(t)$ be the number of type h events that have occurred to

the i th subject over a time period $(0, t]$. This is a counting process in the sense that it keeps its value until the type h event happens to the i th individual, at which time it adds 1 (a jump with size equal to 1) to its previous value. Therefore, at any given time t , the difference between the value of the processes at that time ($N_{hi}(t)$) and its value right before that time ($N_{hi}(t^-)$) is either 0 or 1. We denote this counting process at time t by

$$dN_{hi}(t) = N_{hi}(t) - N_{hi}(t^-) \quad (1)$$

assuming that the events are counted for each subject from a well determined origin of time. The origin can be the beginning of the study if all subjects were randomized at the same time. If subjects enter the study at staggered times and are randomized at their entry time, the origin for each subject could be the subject's actual time of entry. Also, we assume that no subject is observed (followed) after time τ , so that the study duration is $[0, \tau]$. In that period of time, the times of events and the subject's relevant covariate information are recorded. Assume that the counting processes are 0 at time 0 and that no two processes jump at the same time, i.e., $N_{hi}(0) = 0$ for all h and i , and there are no tied event times. If ties occur, some conditions are needed to ensure that the processes $N_{hi}(t)$ form a multivariate counting process, and to apply all theoretical inferences developed for the MI model. One could add a small amount such as 0.01 to displace the tied event times or apply the usual tie-handling methods such as the Breslow, Efron, or exact methods used for Cox's proportional hazards model, available in both SAS and Splus.

Let $Y_{hi}(t)$ be a risk indicator process having a value of 1 if the i th subject is under observation and at risk of suffering from the type h event just before t , and a value of 0 otherwise. Unlike the death-censoring indicator in the usual one-event (death) survival analysis, the process $Y_{hi}(t)$ allows subjects to be in and out of risk. For example, a patient hospitalized for a workplace-related injury is not at

risk of suffering from workplace-related injury during hospitalization. Thus, $Y_{hi}(t)$ for such a patient is 0 just before release from hospital and 1 afterwards.

Define $z_{hi}(t)$ as the type h event specific covariate process for the i th subject. Also, let $H(t)$ denote the history containing all available information about the three processes N , z , and Y up to time t . The covariate process $z(t)$ and the risk indicator process $Y(t)$ are assumed to be *predictable*. This means that their values at time t are fixed once the history is given up to but not including t , i.e., $H(t^-)$. The predictability condition for the covariates entails allowing no measurements at the time when an event takes place.^[10] This condition excludes the so-called internal time-dependent covariates.

Now that we have defined the processes N , z , Y , and the history H , we can define the functions of the MI model. The process $\lambda_{hi}(t, \theta)$ is called the intensity function of $N_{hi}(t)$, if the probability that $N_{hi}(t)$ has a jump in the small (infinitesimal) interval $(t^-, t]$ of length dt , given the history up to t^- , is approximately $\lambda_{hi}(t, \theta)dt$, so that

$$P[N_{hi}(t) - N_{hi}(t^-) = 1 | H(t^-)] \approx \lambda_{hi}(t, \theta)dt \quad (2)$$

or, equivalently, the expected jump in that interval given the history up to t^- is approximately $\lambda_{hi}(t, \theta)dt$, so that

$$E[N_{hi}(t) - N_{hi}(t^-) | H(t^-)] \approx \lambda_{hi}(t, \theta)dt \quad (3)$$

It follows from this definition that the process

$$M_{hi}(t) = N_{hi}(t) - \int_0^t \lambda_{hi}(s, \theta)ds \quad (4)$$

is a martingale. Loosely speaking, a martingale is a stochastic process whose expected value in the future, given “history” up to the present, is equal to its present value.

Equivalently, one could start from the martingale property to define intensity and say that the intensity of $N_{hi}(t)$ is a predictable process $\lambda_{hi}(t, \theta)$, such that (4) is a martingale. In addition to the above, one would also require the process

$$\Lambda_{hi}(t) = \int_0^t \lambda_{hi}(s; \theta)ds \quad (5)$$

to be “locally bounded.”

Under this setup, we say that the counting process $N_{hi}(t)$ follows a multiplicative intensity model if its intensity function is given by

$$\lambda_{hi}(t, \theta) = \alpha_{h0}(t, \gamma)Y_{hi}(t)e^{\beta'z_{hi}(t)} \quad (6)$$

where $\theta = (\gamma', \beta')'$. The term $e^{\beta'z_{hi}(t)}$ can be replaced by any positive function $f(t, z_{hi}(t))$, although the exponential

function is most commonly used (see Ref. [11] for other forms of f). The risk indicators $Y_{hi}(t)$ ensure that the subject is currently under observation and at risk of suffering from a type h event just before time t , in which case $Y_{hi}(t)$ has a value of 1 and hence does not contribute anything to the product in (6). The function $\alpha_{h0}(t, \gamma)$ is called a baseline intensity and is a non-negative function of some unknown parameters γ . It is interpreted as the probability of a type h event occurring to a subject at time t even in the absence of a covariate effect.

Note that the dependence of the intensity on the past of the event process, i.e., $N_{hi}(t^-)$, which is included in the past history $H(t^-)$, does not appear explicitly in Eq. (6). It is, however, implicitly assumed that such dependence is incorporated via time-dependent covariates $z(t)$. In general, the dependence structure among recurrent events is not easy to incorporate into the model, especially when attempting to accommodate them through $z(t)$. One alternative is to include the number of past events as time varying covariates, although there are no simple interpretations of these parameters. In practice, many analysts have neglected modeling this dependence, effectively assuming that the multiple events within a subject are independent. However, using this assumption leads to a mis-specified MI model. In such cases, the use of robust variance estimates could recover some of the lost efficiency.

The quantities

$$\Lambda_{hi}(t, \theta) = \int_0^t \lambda_{hi}(s, \theta)ds \quad (7)$$

and

$$A_{h0}(t, \gamma) = \int_0^t \alpha_{h0}(s, \gamma)ds \quad (8)$$

are called cumulative intensity and cumulative baseline intensity functions, respectively. By taking expectation for both sides of (4) and noting that the martingale will have an expected value of 0, they can be interpreted as the number of type h events expected to occur to the i th subject in $[0, t]$ in the presence or absence, respectively, of covariates.

Parametric Multiplicative Intensity Models

The MI model is parametric if, apart from the γ in (6), the functions $\alpha_{h0}(t, \gamma)$ and $z_{hi}(t)$ are known functions of time. Examples of parametric MI models include the Weibull, exponential, and Gompertz–Makeham survival analysis regression models. If covariates are time-independent, a Weibull-like parametric MI model would have $\alpha_{h0}(t, \gamma) = \alpha t^\rho$, where $\gamma = (\alpha, \rho)'$. Similarly, an exponential-like parametric MI model would have $\alpha_{h0}(t, \gamma) = \gamma$. The latter is also a Poisson-like MI model when each subject experiences only one event.

In addition to specifying the parametric form of the baseline intensity in a parametric MI model, one can assume that $N_{hi}(t)$ is a nonhomogeneous Poisson process. This means that, for each t , $dN_{hi}(t)$ has a Poisson distribution with the rate $\lambda_{hi}(t, \theta)$ and is independent of its history. Thus, $H(t^-)$ in (2) and (3) need not contain information on the process $N_{hi}(s)$ itself. The above Weibull and exponential specifications of baseline intensity can still be used along with the Poisson assumption to obtain parametric Poisson MI models for recurrent events. For the Weibull-like Poisson MI model for dealing with recurrent events, see Ref. [12].

Semiparametric Multiplicative Intensity Models

If, on the other hand, $\alpha_{h0}(t)$ is completely unknown, we can use the Andersen–Gill MI model, also known as a semiparametric MI model, without having to depend on unknown parameters.

The most well-known example of a semiparametric MI model is Cox's proportional hazards model for right-censored survival data. This model is obtained when only one type of a right-censored event occurs only once to each subject. Dropping the subscript h , we redefine the observed data as (X_i, δ_i) where X_i is either the event time or the censoring time for the i th subject, depending on which comes first, and δ_i is the status indicator, taking a value of 1 if X_i is the event time and 0 otherwise. We define $N_i(t) = I(X_i \leq t, \delta_i = 1)$, where $I(A)$ is 1 if A is true, and 0 otherwise. Similarly, the process $Y_i(t)$ can be redefined as

$$Y_i(t) = I(X_i \geq t) = I(\text{subject } i \text{ was alive and uncensored until just before } t)$$

Note that the nonhomogeneous Poisson assumption mentioned above for parametric MI models can still be used with a completely unspecified baseline intensity, thus obtaining a semiparametric Poisson MI model.^[12]

In general, the Poisson assumption (also known as memoryless property of the process) does not hold, since the re-occurrence of an event usually depends on previous occurrences. However, the Poisson assumption, both in the parametric and semiparametric set-up, has the advantage that (2) and (3) are interpretable as marginal probabilities and marginal means.

ESTIMATION AND HYPOTHESES TESTING

Given an MI model with an intensity function as in (6), the parameters $(\gamma', \beta')'$ have to be estimated in the parametric case. In the semiparametric case, the baseline cumulative intensity, $A_{h0}(t)$, and the vector of covariate effects, β , need to be estimated.

Before writing a likelihood for the model, we have to bear in mind that the independent censoring assumption is crucial in the case of right censoring, which terminates observation of the subject. Technically, a censoring mechanism is called “independent” if the intensity in (2) and (3) is not altered by knowing that the subject was not censored up until t^- and including this knowledge in the history $H(t^-)$. To write a “likelihood” for the model, one could use the heuristic argument that, given the history $H(t^-)$, $dN_{hi}(t)$ are independent Bernoulli random variables taking a value of 1 with probability $\lambda_{hi}(\theta, t)dt$. Hence, the joint likelihood of the data (excluding contributions from censoring times) is the product over time, event type and subjects of $[\lambda_{hi}(\theta, t)dt]^{dN_{hi}(t)}[1 - \lambda_{hi}(\theta, t)dt]^{1-dN_{hi}(t)}$.

$$L(\tau, \theta) = \prod_{0 \leq t \leq \tau} \prod_h \prod_i [\lambda_{hi}(\theta, t)dt]^{dN_{hi}(t)} \times [1 - \lambda_{hi}(\theta, t)dt]^{1-dN_{hi}(t)} \quad (9)$$

Manipulating (9) using product limit arguments, or treating $dN_{hi}(t)$ as a random variable following Poisson distribution with mean $\lambda_{hi}(\theta, t)dt$, given the history $H(t^-)$, one can rewrite the likelihood as

$$L(\tau, \theta) = \prod_h \prod_i \left[\prod_{0 \leq t \leq \tau} [\lambda_{hi}(\theta, t)dt]^{dN_{hi}(t)} \right] \times \exp \left(- \int_0^\tau \lambda_{hi}(\theta, t)dt \right) \quad (10)$$

These quantities are partial likelihoods, since the piece contributed by the censoring times is being ignored and it may carry some information about the parameters. If the independent censoring mechanism is also noninformative, (9) and (10) will be full likelihoods carrying all the information contained in the observed data about θ . Nevertheless, these partial likelihoods can be treated as full likelihoods with little loss of efficiency.^[3]

Parametric Multiplicative Intensity Models

To obtain score functions for the parametric MI model, we take the log of (10) with $\lambda_{hi}(\theta, t) = \alpha_{h0}(t, \gamma)Y_{hi}^*(t)e^{\beta'z_{hi}(t)}$. Note that the log of a continuous product over t becomes a continuous sum, and by differentiating this log partial likelihood with respect to (γ, β) , we get the system of score equations (see Ref. [7], p. 585)

$$\sum_h \sum_i \int_0^\tau \frac{(\partial/\partial \gamma_j) \alpha_{h0}(t, \gamma_j)}{\alpha_{h0}(t, \gamma_j)} dN_{hi}(t) - \sum_h \int_0^\tau \frac{\partial}{\partial \gamma_j} \alpha_{h0}(\gamma, t) S_h^{(0)}(t, \beta) dt = 0 \quad (11)$$

and

$$\sum_h \sum_i \int_0^\tau z_{hij}(t) dN_{hi}(t) - \sum_h \int_0^\tau \alpha_{h0}(\gamma, t) S_{hj}^{(1)}(t, \beta) dt = 0 \quad (12)$$

where $S_h^{(0)}(\beta, t) = \sum_{i=1}^n Y_{hi}(t) e^{\beta' z_{hi}(t)}$, and $S_{hj}^{(1)}(\beta, t) = \sum_i z_{hij} Y_{hi}(t) e^{\beta' z_{hi}(t)}$ is the first-order partial derivative of $S_h^{(0)}(\beta, t)$ with respect to β_j .

Solving Eqs. (11) and (12) with respect to (γ, β) , for example by the Newton–Raphson method, would render the maximum partial likelihood estimators of the parameters in the model, $\hat{\gamma}$ and $\hat{\beta}$. Since $dN_{hi}(t)$ is a step function with jumps of size 1, the integrals in the first terms of (11) and (12) are essentially discrete sums over ordered event times. For example, if $\{t_{hij}; j = 1, \dots, n_{hi}\}$ denotes ordered type h event times of the i th subject, then the first term of (11) becomes

$$\int_0^\tau \frac{\partial}{\partial \gamma} [\log \alpha_{h0}(t, \gamma)] dN_{hi}(t) = \sum_{j=1}^{n_{hi}} \frac{\partial}{\partial \gamma} [\log \alpha_{h0}(t_{hij}, \gamma)] \quad (13)$$

Similar arguments apply to the first term of (12).

The Fisher observed information can be partitioned according to the components of the parameter vector $(\gamma', \beta')'$ as follows:

$$I(\tau; \gamma, \beta) = \begin{pmatrix} I_{\gamma\gamma} & I_{\gamma\beta} \\ I_{\beta\gamma} & I_{\beta\beta} \end{pmatrix}$$

The quantities $I_{\gamma\gamma}$ and $I_{\gamma\beta}$ are the negatives of the derivative of the left-hand side of (11) for γ and β , respectively, and similarly, $I_{\beta\gamma}$ and $I_{\beta\beta}$ are the negatives of the derivatives of the left-hand side of (12) for γ and β , respectively. The inverse of this matrix evaluated at $(\hat{\gamma}', \hat{\beta}')'$ is a consistent estimator for the asymptotic covariance matrix of $(\hat{\gamma}', \hat{\beta}')'$.^[7,13]

The testing of hypotheses about $(\gamma', \beta')'$ can be based on the usual test statistics such as the likelihood ratio, Wald, and Rao's score. These are asymptotically χ^2 test statistics.^[13]

Semiparametric Multiplicative Intensity Models

For the semiparametric MI model, we again use (10) but without γ , i.e., $\alpha_{h0}(t, \gamma) = \alpha_{h0}(t)$. One can take the log-partial likelihood and differentiate it with respect to $\alpha_{h0}(t)$ and solve the score, explicitly, with respect to $\alpha_{h0}(t)dt$ to obtain

$$\hat{\alpha}_{h0}(t)dt = \sum_i dN_{hi}(t) / \sum_i Y_{hi}(t) e^{\beta' z_{hi}(t)} = d\hat{A}_{h0}(t, \beta) \quad (14)$$

By inserting this expression back into the partial likelihood (10), one essentially gets a partial likelihood involving β only,

$$L_\tau(\beta) = \prod_{0 \leq t \leq \tau} \prod_h \prod_i \left(\frac{e^{\beta' z_{hi}(t)}}{\sum_{i=1}^n Y_{hi}(t) e^{\beta' z_{hi}(t)}} \right)^{dN_{hi}(t)} \quad (15)$$

This is exactly the same as Cox's partial likelihood,^[1] and it has been shown^[3] that it can be treated as a full likelihood and used for inferences about β under some regularity conditions. The partial likelihood in (15) can also be thought of as the product over time, event-type, and subjects of the conditional probability of a type h event happening to the i th subject at time t given that a type h event has happened at that time to one of those at risk among the n subjects under study.^[1]

Taking the log of (15) and using the same arguments as above, we get

$$\log L_\tau(\beta) = \sum_h \sum_i \int_0^\tau [\beta' z_{hi}(t) - \log S_h^{(0)}(\beta, t)] dN_{hi}(t) \quad (16)$$

where $S_h^{(0)}(\beta, t)$ is defined as above. Therefore, components of the score vector [first-order partial derivatives of (16)] are

$$\begin{aligned} U_j(\beta, \tau) &= \frac{\partial}{\partial \beta_j} \log L_\tau(\beta) \\ &= \sum_h \sum_i \int_0^\tau \left[z_{hij}(t) - \frac{S_{hj}^{(1)}(\beta, t)}{S_h^{(0)}(\beta, t)} \right] dN_{hi}(t) \\ &= \sum_h \sum_i U_{hij}(\beta, \tau) \end{aligned} \quad (17)$$

where $S_{hj}^{(1)}(\beta, t)$ is the first-order partial derivative of $S_h^{(0)}(\beta, t)$ with respect to β_j . Thus, the j th component of the score vector can be written as the sum of contributions due to the h th event-type from the i th individual. Solving the system of equations, $U_j(\beta, \tau) = 0$ for $j = 1, \dots, p$ gives $\hat{\beta}$, the maximum partial likelihood estimator of β . The negative of the matrix of second-order partial derivatives of (16) is the observed Fisher information matrix $I(\hat{\beta}, \tau)$. The inverse of $I(\hat{\beta}, \tau)$ is a consistent estimator for the covariance matrix of $\hat{\beta}$ for large samples. To recognize the dependence of future events on past events, we can use the robust covariance of $(\hat{\beta})$, obtainable from *coxph* in *Splus* by using the *cluster(id)* argument, where *id* refers to the subject index.

Using (16), tests of hypotheses can be based on the usual likelihood ratio, Wald or score statistics, all of which have asymptotic χ^2 distributions. Both estimation and hypothesis testing for the semiparametric MI model are

implemented in the PROC PHREG and *coxph* functions of SAS and Splus, respectively.^[14]

Since statistical inferences for the parametric and semi-parametric MI models are based on partial likelihoods, tied event times from different subjects can be accommodated through the partial likelihoods using the techniques of Cox's proportional hazards model as discussed earlier.

Estimating the Cumulative Intensity Functions

Integrating $\hat{\alpha}_{h0}(t)dt$ and replacing β with its maximum partial likelihood estimator, we find that the Breslow estimator (also known as Tsiatis, Link, and Nelson–Aalen) for the cumulative baseline intensity function is given by

$$\hat{A}_{h0}(t, \hat{\beta}) = \int_0^t \frac{\sum_i dN_{hi}(s)}{\sum_{i=1}^n Y_{hi}(s)e^{\hat{\beta}'z_{hi}(s)}} \quad (18)$$

Using (14) or (18) and the definition of the cumulative intensity function, we can obtain an estimator for the cumulative intensity $\Lambda_{hi}(t, \beta)$ of a subject with covariate values $z_{hi}(t)$ as

$$\hat{\Lambda}_{hi}(t, \hat{\beta}) = \int_0^t Y_{hi}(s)e^{\hat{\beta}'z_{hi}(s)} d\hat{A}_{h0}(s, \hat{\beta}) \quad (19)$$

This expression reduces in practice to a discrete sum over the ordered event times of the subject, i.e.,

$$\hat{\Lambda}_{hi}(t, \hat{\beta}) = \sum_{t_{hij} \leq t} e^{\hat{\beta}'z_{hi}(t_{hij})} [\hat{A}_{h0}(t_{hij}, \hat{\beta}) - \hat{A}_{h0}(t_{hi(j-1)}, \hat{\beta})] \quad (20)$$

where $t_{hij}; j = 1, 2, \dots$ are the ordered type h event times for the i th subject.

The baseline cumulative intensity for a parametric MI model is estimated by replacing γ with its maximum partial likelihood estimator to get

$$\hat{\Lambda}(t, \hat{\gamma}) = \int_0^t \alpha_{h0}(s, \hat{\gamma}) ds \quad (21)$$

MODEL ASSESSMENT AND DIAGNOSTICS

To facilitate the discussion, let us assume there is only one type of recurrent event and drop the subscript h from the MI model. There are several assumptions underlying the MI model specification. Both parametric and semi-parametric MI models share the proportional intensity, the log-linearity, and the martingale assumptions, whereas in parametric MI models one is also concerned with the specific shape of the parametric cumulative baseline intensity,

$\Lambda_0(t, \theta)$. Most of the goodness-of-fit and diagnostic methods in the literature have been developed for semiparametric MI models, and especially for the Cox regression model. Therefore, our main focus in this section is on describing some of the techniques for semiparametric MI models, while briefly mentioning how they could be adapted to the parametric set-up.

The following residuals are useful for diagnosing MI regression models and assessing the model assumptions.

1. Martingale residuals for the i th subject,

$$\hat{M}_i(\tau) = N_i(\tau) - \hat{\Lambda}_i(\tau, \hat{\beta}) \quad (22)$$

where $\hat{\Lambda}_i(\tau, \hat{\beta}) = \int_0^\tau Y_i(t)e^{\hat{\beta}'z_i(t)} d\hat{A}_0(t, \hat{\beta})$ and $\hat{A}_0(t)$ is the Breslow estimator for the cumulative baseline intensity. Obviously, $\hat{\Lambda}_i(\tau, \hat{\beta})$ is essentially a discrete sum over event times since $\hat{A}_0(t, \hat{\beta})$ itself is available only at event times. From (4), this residual is usually interpreted as the observed number of events at time t minus the predicted number of events (from the MI model) at that time. Similarly, for parametric MI models,

$$\hat{M}_i(\tau) = N_i(\tau) - \hat{\Lambda}_i(\tau, \hat{\gamma}, \hat{\beta}) \quad (23)$$

where $\hat{\Lambda}_i(\tau, \hat{\gamma}, \hat{\beta}) = \int_0^\tau \alpha_0(t, \hat{\gamma}) Y_i(t) e^{\hat{\beta}'z_i(t)} dt$.

2. Schoenfeld residuals for the i th subject and j th covariate at time t ,

$$\hat{r}_{ij}(t) = z_{ij}(t) - \left[S_j^{(1)}(\hat{\beta}, t) / S^{(0)}(\hat{\beta}, t) \right] \quad (24)$$

which is exactly the integrand in (17) with estimated β .

3. Score residuals for the i th subject and j th covariate,

$$U_{ij}(\hat{\beta}, \tau) = \int_0^\tau \left[z_{ij}(t) - \frac{S_j^{(1)}(\hat{\beta}, t)}{S^{(0)}(\hat{\beta}, t)} \right] dN_i(t) \quad (25)$$

which is the contribution of subject i to the score function (17), evaluated at $\hat{\beta}$. These residuals are the sums of the Schoenfeld residuals of (24) over event times in $[0, \tau]$. Score residuals for β in the parametric MI model are slightly different and are given by

$$\int_0^\tau z_{ij} d \left[N_{hi}(t) - \hat{\Lambda}_i(t, \hat{\gamma}, \hat{\beta}) \right] \quad (26)$$

This expression is obtained by rearranging and estimating the left-hand side of (12).

Assessment of Influential Observations

We can measure the influence of the i th subject on the regression parameter estimates by computing

$\Delta^{(i)}\beta = \hat{\beta} - \hat{\beta}_{i-}$, where $\hat{\beta}$ is obtained by fitting the model using data from all subjects, and $\hat{\beta}_{i-}$ is the parameter estimator from a model not containing the i th subject. Plotting the elements of the vector $\Delta^{(i)}\beta$ against subject index would then reveal influential subjects (subjects with high leverage). Recomputing $\Delta^{(i)}\beta$ after removing each subject from the model can be computationally costly. Therefore, an often-used approximation to $\Delta^{(i)}\beta$ is the vector $I^{-1}(\hat{\beta}, \tau)(U_{i1}, \dots, U_{ip})'$, where $I^{-1}(\hat{\beta}, \tau)$ is the estimated covariance of $\hat{\beta}$ and U_{ij} are the score residuals (see Ref. [15] for a modified version of this approximation). These residuals are called *dfbetas* in SAS and Splus and can be saved for plotting purposes.

These methods are also applicable to parametric MI models with score residuals given in (26). See Ref. [15] for a discussion of the validity of these methods in the parametric case.

Assessment of the Log-Linearity Assumption

For time-independent covariates, the log-linearity assumption is that the log of the ratio of the intensity function to the baseline intensity function is linear in the covariates. That is, $\log \lambda_i(t, \theta)/\alpha_0(t, \gamma) = \beta'z_i$. This assumption would be violated if, for instance, the covariates enter the model in quadratic or other functional form. Also, certain covariates, say the first covariate z_1 , might need to be transformed before entering the model through the exponent in (6). These assumptions can be assessed by plotting the martingale residuals, $\hat{M}_i$, obtained from a model excluding the covariate z_1 , against the values of the covariate itself, z_{i1} . A smoothed-out version of such a scatter plot would give an indication of type of transformation required, as long as z_1 is not strongly correlated with the remaining covariates. This method and some modifications of it are suggested in Refs. [15,16] and implemented in the Splus function *cox.zph* (see Splus manual). In SAS, the martingale residual for a subject can be calculated by summing over all the rows corresponding to the subject, and can be saved using the RESMART option of the OUTPUT statement of PROC PHREG.

Proportional Intensity Assumption

The proportional intensity (PI) assumption is relevant when the covariates are not time dependent. In that case, the PI assumption is that the ratio of intensities for two subjects with distinct covariate values, say z_i and z_j , is time independent (constant), i.e., $\lambda_i(\beta, t)/\lambda_j(\beta, t) = e^{\beta'(z_i - z_j)}$. Note that the proportional intensity assumption becomes the proportional hazards (PH) assumption in the Cox model. This assumption is violated if β depends on time. Writing $\beta_j(t) = \beta + \theta g_j(t)$ for $j = 1, \dots, p$, where $g_j(t)$ are predictable processes, Grambsch and Therneau^[17] have shown that many tests appearing in the literature for checking the PI (or PH)

assumption are just tests for $H_0: \theta = 0$ with different choices of $g_j(t)$. These authors propose the use of a smoothed plot of rescaled Schoenfeld residuals against event times to reveal the functional form of the coefficients [i.e., the form of $g_j(t)$]. To do so, one would first extract the p -dimensional vector, $r(t_k)$, containing Schoenfeld residuals for all covariates of the only subject having an event at time t_k (if there are no ties). Then, we would form the scaled p -dimensional Schoenfeld residuals, $r^*(t_k) = [I^{-1}(\hat{\beta}, \tau)/d]r(t_k)$, where d is the total number of events in the sample and $\hat{\beta}$ is estimated under H_0 . If there are tied event times, then the Schoenfeld residuals in (24) use the average of covariate values, $z_{ij}(t_k)$, over all subjects having events at t_k . A smoothed scatter plot of the components of $r^*(t_k) + \hat{\beta}$ against t_k would then reveal the functional form of the corresponding coefficients, $\beta(t)$. The authors also provide χ^2 tests for checking the global hypothesis $\theta = 0$ and the hypotheses $\theta_j = 0$, regarding each parameter β_j for $j = 1, \dots, p$. Obviously, these tests are against alternatives with specific $g_j(t)$. Several $g_j(t)$, including identity, are available in the *cox.zph* function of Splus.

To use a simple graphic method of checking whether a given covariate with several levels (e.g., gender or age classes) complies with the PI assumption, one can stratify the MI model over the levels of the covariate and obtain estimated cumulative baseline intensities for each strata using (18). If the covariate in question complies with the PI assumption, then plots of the log of these estimated cumulative baseline intensities should give approximately parallel curves. Another useful method is to plot the estimated cumulative baseline intensities one against another (or each one against one stratum declared as a reference). Such plots are expected to be straight lines through the origin if the PI assumption is correct for the covariate. These plotting techniques can be used with parametric or semiparametric MI models.

Omnibus Goodness-of-Fit Methods

The MI model relies, globally, on an assumption that the cumulative intensity process is the predictable process that makes (4) a martingale. If this process is a martingale, then the model is correct, regardless of the form of the intensity function, $\lambda_i(\beta, t)$. Tests and diagnostic plots based on this assumption can be called “omnibus” in the sense that we examine all the assumptions including PI and functional forms of the covariates by checking only the martingale assumption.

For instance, if the martingale assumption is correct, then taking the expected value of both sides of (4), as mentioned above, will imply that the expected number of events up to time t for a subject is $\Lambda_i(t, \theta)$. Thus, the estimated version of this cumulative intensity summed over all subjects should approximate the observed number of events in the sample up to time t . If the martingale assumption is correct, then a plot of $\sum_i \Lambda_i(t, \theta)$ against the cumulative observed number of events in the sample up to t will

give a straight line through the origin with a unit slope (see Refs. [7,18]).

Recently, Jones and Harrington^[19] proposed tests for this martingale assumption by exploiting the fact that, if the $M_i(t)$ processes defined in (4) were martingales, then their estimated versions at successive event times would be approximately martingales and hence would have uncorrelated increments with mean 0. They developed two statistics, Q'' and Q_{cor} , for testing the mean 0 and uncorrelated increments properties, respectively. These tests are asymptotically χ^2 and are useful for checking the global fit of the model.

Checking Parametric Baseline Intensity Form

To check the validity of a presumed baseline intensity function for parametric MI models, one can plot the Breslow estimate of the semiparametric cumulative baseline intensity in (18) against (21) obtained from fitting the presumed parametric model. If the assumed parametric baseline is reasonable, then this plot should give a straight line through the origin.

Alternatively, the log of the Breslow estimator can be plotted against time or a transformation of time for specific parametric baseline shapes, thus bypassing the need to fit a parametric model. For example, for a presumed Weibull-like baseline intensity with $\Lambda_0(t, \gamma) = \alpha t^\rho$, the plot of $\log \Lambda_0(t, \hat{\beta})$ against $\log t$ would give approximately a straight line with intercept $\log \alpha$ and slope ρ .

Nested parametric MI models can also be identified through the usual -2 log-likelihood difference in the two models. For example, to check whether an exponential-like parametric MI model is better than a Weibull-like model, one can use the usual G^2 , which would follow a χ^2 distribution with 1 degree of freedom.

Multiplicative Intensity Models with Frailty

In practice, data exhibit more heterogeneity than is usually explainable by the covariates in the model. This heterogeneity can be due to unobserved characteristics of the individuals in the study. One way to tackle this problem is to assume that each subject belongs to a family (a cluster) and each family is characterized by a random unobserved frailty, X_l ; $l = 1, 2, \dots$. The family can be made up of the subject itself, in which case the subject ID is used to index the frailty. Usually, the x_l are assumed to be from a gamma distribution with mean 1 and variance $1/\nu$ and they enter the model through (6) in a multiplicative way, i.e., $\lambda_{hi}(t, \theta) = \hat{x}_l \alpha_{h0}(t, \gamma) \gamma_{hi}(t) e^{\beta' z_{hi}(t)}$. As $\nu \rightarrow \infty$, the variance goes to 0 and the model becomes a frailty-less MI model.

The estimation procedure uses the EM algorithm (see Ref. [7]), whose E-step replaces the unobserved realizations of the X_l with their expected values given the observed data, $\hat{x}_l = E[X_l | N, Y, Z]$, assuming that censoring

is noninformative about $\alpha_{h0}(t, \gamma)$ and β , given the realizations of the frailties. Then the log-partial likelihood in (16) is maximized with respect to the unknown parameters, including ν , in the M-step of the EM algorithm using intensities $\lambda_{hi}(t, \theta) = \hat{x}_l \alpha_{h0}(t, \gamma) \gamma_{hi}(t) e^{\beta' z_{hi}(t)}$. The partial likelihood so obtained is actually equivalent to the one obtained by integrating frailties out of the partial likelihood for the complete data, i.e., both observed and unobserved data (N, Y, Z, X). Therefore, the resulting likelihood is usually called a “marginal partial likelihood,” and is used as a partial likelihood for testing hypotheses about θ and ν . Multiplicative intensity models with gamma frailties can be fitted using the *coxph* function in Spls by adding the term *frailty(family, dist=“Gamma”)* to the model specification, where family is a variable indexing the cluster (family), which could be equal to the subject ID. Recently, Therneau, Grambsch, and Pankratz^[20] provided an algorithm for fitting MI frailty models using a penalized likelihood approach.

CONCLUSIONS

This entry is an overview of multiplicative intensity models. We presented parametric and semiparametric MI models, along with established strategies for estimating and testing the model parameters. We then discussed a number of related technical issues for applications. These include the use of frailties for modeling dependence among recurrent events of the same subject. Other well established methodologies for analyzing recurrent event data are those of Wei, Lin, and Weissfeld (WLW) and Printice, William, and Peterson (PWP). For a discussion of these methods, see Ref. [21]. Also, Yau and McGilchrist^[22] proposed a strategy for completely bypassing the MI models. Another issue of great importance in using MI models is the presence of dependent censoring. For example, under dependent censoring, the interpretation of the intensity function as an instantaneous net hazard rate is not possible using Cox’s proportional hazards model.^[10] Alternative MI models for dealing with dependent censoring in the recurrent event analysis are proposed in Ref. [23]. However, we note that most of these techniques are based on an asymptotic theory, requiring large samples. A comprehensive simulation study that examines finite sample performance of the inference procedures would be of great importance in practice.

REFERENCES

1. Cox, D.R. The analysis of multivariate binary data. Appl. Stat. **1972**, 21, 113–120.
2. Aalen, O. Nonparametric inference for a family of counting processes. Ann. Stat. **1978**, 6, 701–726.
3. Andersen, P.K.; Gill, R.D. Cox’s regression model for counting processes: a large sample study. Ann. Stat. **1982**, 10, 1100–1120. (comment: 1121–1124).

4. Lawless, J.F.; Wigg, M.B.; Tuli, S.; Drake, J.; Lamberti-Pasculli, M. Analysis of repeated failures or durations, with application to shunt failures for patients with pediatric hydrocephalus. *Appl. Stat.* **2001**, *50* (4), 449–465.
5. Hussein, A.; Carrière, K.C. A measure of continuity of care based on the multiplicative intensity model. *Stat. Med.* **2002**, *21* (3), 457–465.
6. Lin, M.R.; Chang, S.H.; Pai, L.; Keyl, P.M. A longitudinal study of risk factors for motorcycle crashes among junior college students in Taiwan. *Accid. Anal. Prev.* **2003**, *35* (2), 243–252.
7. Andersen, P.K.; Borgan, O.; Gill, R.D.; Keiding, N. *Statistical Models Based on Counting Processes*; Springer-Verlag Inc.: New York, 1993, 767.
8. Fleming, T.R.; Harrington, D.P. *Counting Processes and Survival Analysis*; John Wiley & Sons: New York, 1991, 429.
9. Gill, R.D. Understanding Cox's regression model: a martingale approach. *J. Am. Stat. Assoc.* **1984**, *79*, 441–447.
10. Self, S.G.; Prentice, R.L. Comments on Andersen and Gill's "Cox's regression model for counting processes: a large sample study." *Ann. Stat.* **1982**, *10*, 1121–1124.
11. Prentice, R.L.; Self, S.G. Asymptotic distribution theory for Cox-type regression models with general relative risk form. *Ann. Stat.* **1983**, *11*, 804–812.
12. Lawless, J.F. Regression methods for Poisson process data. *J. Am. Stat. Assoc.* **1987**, *82*, 808–815.
13. Borgan, O. Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scand. J. Stat.* **1984**, *11*, 1–16. (corrigendum: 11, 275).
14. Therneau, T.M.; Hamilton, S.A. RhDNase as an example of recurrent event analysis. *Stat. Med.* **1997**, *16*, 2029–2047.
15. Therneau, T.M.; Grambsch, P.M.; Fleming, T.R. Martingale-based residuals for survival models. *Biometrika* **1990**, *77*, 147–160.
16. Grambsch, P.M.; Therneau, T.M.; Fleming, T.R. Diagnostic plots to reveal functional form for covariates in multiplicative intensity models. *Biometrics* **1995**, *51*, 1469–1482.
17. Grambsch, P.M.; Therneau, T.M. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **1994**, *81*, 515–526. (corrigendum: **1995**, *82*, 668).
18. Arjas, E. A graphical method for assessing goodness of fit in Cox's proportional hazards model. *J. Am. Stat. Assoc.* **1988**, *83*, 204–212.
19. Jones, C.L.; Harrington, D.P. Omnibus tests of the martingale assumption in the analysis of recurrent failure time data. *Lifetime Data Anal.* **2001**, *7* (2), 157–171.
20. Therneau, T.M.; Grambsch, P.M.; Pankratz, V.S. Penalized survival models and frailty. *J. Comput. Graphical Stat.* **2003**, *12* (1), 156–175.
21. Wei, L.J.; Glidden, D.V. An overview of the statistical methods for multiple failure time data in clinical trials. *Stat. Med.* **1997**, *16*, 833–839.
22. Yau, K.K.W.; McGilchrist, C.A. ML and REML estimation in survival analysis with time dependent correlated frailty. *Stat. Med.* **1998**, *17*, 1201–1213.
23. Wang, M.C.; Qin, J.; Chiang, C.T. Analyzing recurrent event data with informative censoring. *J. Am. Stat. Assoc.* **2001**, *96* (455), 1057–1065.

Multiplicity in Clinical Trials

Peter Westfall

Area of Information Systems and Quantitative Sciences, Texas Tech University, Lubbock, Texas, U.S.A.

Frank Bretz

Statistical Methodology and Consulting, Novartis, Basel, Switzerland; Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Wien, Austria

Abstract

Multiplicity, loosely defined, refers to multiple inferences that are made in a simultaneous context. Much attention has been given to multiplicity issues in clinical trials over the past few decades, culminating in recent guidance documents from the European Medicines Agency's Committee for Human Medical Products and the US Food and Drug Administration. As discussed in these regulatory documents, multiplicity arises in almost any clinical trial: Potential sources include comparison of several treatment or dose groups, multiple endpoints, multiple time points, interim analyses, multiple tests of the same hypothesis (e.g., parametric and nonparametric), variable and model selection, and subgroup analysis. In this review article, some questions and problems related to multiplicity are highlighted, and some of the more recent solutions involving multiple comparisons procedures are discussed as they appear in the recent guidance documents.

Keywords: Bayesian methods; Dose–response function; False discovery rate; Family-wise error rate; Primary endpoints; Replicability; Safety endpoints; Secondary endpoints; Subgroup analysis.

INTRODUCTION

Multiplicity, loosely defined, refers to multiple inferences that are made in a simultaneous context. Much attention has been given to multiplicity issues in clinical trials over the past few decades,^[1–7] culminating in recent guidance documents from the European Medicines Agency's Committee for Human Medical Products (CHMP)^[8] and the US Food and Drug Administration (FDA).^[9] As discussed in these regulatory documents, multiplicity arises in almost any clinical trial: Potential sources include comparison of several treatment or dose groups, multiple endpoints, multiple time points, interim analyses, multiple tests of the same hypothesis (e.g., parametric and nonparametric), variable and model selection, and subgroup analysis. In this review article, some questions and problems related to multiplicity are highlighted, and some of the more recent solutions involving multiple comparisons procedures (MCPs) are discussed as they appear in the recent guidance documents.

Statistical practices of the pharmaceutical industry are largely driven by expectations of the regulatory agencies. The CHMP guideline states^[8]:

In clinical studies it is often necessary to answer more than one question about the efficacy (or safety) of the experimental treatment in a specific disease, because the success of a drug development programme may depend on a

positive answer to more than a single question. It is well known that the likelihood of a positive chance finding increases with the number of questions posed, if no actions are taken to protect against the inflation of false positive findings from multiple statistical tests.

The FDA guideline states similarly^[9]

Most clinical trials performed in drug development contain multiple endpoints to assess the effects of the drug and to document the ability of the drug to favorably affect one or more disease characteristics. As the number of endpoints analyzed in a single trial increases, the likelihood of making false conclusions about a drug's effects with respect to one or more of those endpoints becomes a concern if there is not appropriate adjustment for multiplicity.

While these documents acknowledge that multiplicity adjustment is sometimes unnecessary, they also suggest strongly that such non-adjustment be carefully justified. As a result, simple adjustment procedures that control the family-wise error rate (FWER) are routinely incorporated in the study protocols of pharmaceutical companies. There are many FWER-controlling procedures, and the simple Bonferroni method can be pointed out as a useful (although somewhat conservative) standard. This method simply requires that if there are k inferences in a family,

then all inferences should be performed at the α/k significance level rather than at the α level. The result is that the FWER, or probability of declaring one or more false positives, is no more than α .

THE PROBLEM: REPLICATION

The scientific concern surrounding multiplicity is whether the reported effects from the clinical trial will replicate in the general patient population. Lack of replication can occur in three ways: (i) the drug is ineffective in the general patient population, despite being “efficacious” in the clinical trial study; (ii) the drug has an effect in the opposite direction in the patient population as it had in the clinical study; (iii) the drug’s effect size in the patient population is substantially lower than it is in the clinical study.

Specific explanations of these problems follow:

- Errors of declaring effects when none exist
The classical characterization is in terms of the “1 in 20” significance criterion: In 20 tests of hypotheses, all of which are (unknown to the trialist) truly null, one false positive result is expected. Thus, repeated testing can increase the likelihood of false positives.
- Errors of declaring effects in the “wrong direction”
One might not believe point-zero null hypotheses can truly exist, in which case the “1 in 20” explanation has little impact. Another point of view, similar to the testing point of view, is that it is likely that directions of the claims can be erroneous when multiple tests are used without multiplicity adjustment. For example, if a test of a two-sided hypothesis is performed at the .05 level, then there is (at most) a .025 probability that the test will be declared significant, but in the wrong direction. When multiple tests are performed, this probability accumulates, so that if 40 tests are performed, one directional error (in a worst case) may be expected. Even when multiplicity adjustments are performed, some do not control the probability of erroneous sign declaration.^[10]
- Errors of declaring inflated effect sizes
A third characterization of the multiplicity problem is in terms of selection effects. In this scenario, directional errors need not be postulated. In fact, it can be believed with a priori certainty that all effects are in their expected directions. Nevertheless, when a single, “most significant” comparison is isolated from this collection, it can only be presumed that the effect size is biased upward because of selection effects.^[11]

In all these three characterizations, there is a concern that the presentation of the scientific findings from the study may be exaggerated. By adopting a more conservative stance, MCPs aim to avoid such replication errors.

Is multiplicity really a cause for concern? Does the multiple comparisons problem truly cause exaggerated claims that do not replicate in the follow-up studies? The answer is “yes”; several such documented instances are reported.^[12–14] It is not at all uncommon in pharmaceutical research to find that effect sizes either do not replicate or replicate at a much lower level than was previously expected. Such failure to replicate has similarly been a recent concern in virtually all of science.^[15]

While problems of lack of replication are often attributed to poor designs and data collection procedures,^[12] it is as plausible that they result from selection effects related to the multiplicity problem. In many cases, such effects can be subtle. For example, a clinical efficacy measure might be taken one month after the administration of the drug. The efficacy can be determined by (i) using the raw measure, (ii) using the percentage change from baseline, (iii) using the actual change from baseline, or (iv) using the baseline covariate-adjusted raw measure. If an aggressive strategy where the “best” (most significant) measure is chosen, then the reported effect size measure will be clearly inflated because the maximal statistic capitalizes (unfairly) on random variations in the data. In such a case, it is not surprising that the follow-up studies produce less stellar results; this phenomenon is simply an example of regression to the mean.

SKEPTICISM: A GUIDING PRINCIPLE FOR MULTIPLICITY MANAGEMENT IN CLINICAL TRIALS

Pharmaceutical companies and research institutions that conduct clinical trials (hereafter, simply “sponsors”) are naturally loath to adopt conservative strategies for the analysis of efficacy data; after all, if the data fail to clear the more conservative hurdle, then drug approval might be delayed, at great cost to the company. On the other hand, reviewers from regulatory agencies or journals, and members of monitoring committees (hereafter, simply “reviewers”) have a right and an obligation to be skeptical, as motives of a sponsor can sometimes override scientific objectivity. Thus, it is sensible that reviewers should request a multiplicity correction, one that is within reasonable limits of skepticism that apply to the case at hand. Consider the following simple example. Suppose a sponsor decides to claim efficacy by establishing a significant treatment effect in any one of the various patient subgroups. The reviewer may adopt the skeptical position that the drug might not work at all, in any subgroup, in which case a multiplicity correction that considers the given set of subgroups (which would determine the “family” and the Bonferroni k) is appropriate. Examples showing how careless, non-skeptical analysis of subgroups can lead to type I errors in clinical trials are reported in the literature.^[13,14]

It is not always obvious what the appropriate “family” should be, but skepticism can help to resolve some of the difficulties. As another example, suppose several clinical compounds are to be evaluated against a standard. The experiment can be performed in two ways: (i) all compounds are compared with a standard at a given research center, or (ii) compound i is compared against the standard at research center i . In either case, the goal is to discover which compounds beat the standard.

Duncan^[16] suggests that multiplicity adjustment is not necessary in case (ii) since there are separate experiments. However, the guiding principle of skepticism dispels such thinking. A skeptical position is that few to none of the compounds are better, regardless of whether the experiments are concurrent or separate. Either way, there is an inflated probability of one or more type I errors over the set of tests when the skeptical position is, in reality, correct. Assuming an interest in protecting FWER for the collection of compounds tested, the a priori need for multiplicity correction is identical for cases (i) and (ii).

Cook and Farewell^[17] argue that no adjustment for multiplicity is necessary if a few, preselected endpoints in a clinical trial are to be “marginally” interpreted. Such a strategy may not satisfy the skeptical reviewer. Consider the case of a drug that is tested against an active competitor, and that significance with respect to each of a collection of endpoints is tested. A skeptical point of view is that the two formulations are a priori equivalent. In such a case, isolation of the “most significant” endpoint, and claiming improved efficacy on that measure, is sure to overstate the evidence. A skeptical reviewer has every right to demand a stronger standard of evidence than the unadjusted comparisons that the sponsor might offer.

SKEPTICISM AND BAYESIAN STATISTICS

The skeptical point of view naturally falls within the paradigm of Bayesian statistics, as “skepticism” is simply a prior opinion. Being skeptical can be instantiated by choosing a skeptical Bayesian prior distribution. Acceptance of multiple comparisons methods is largely determined by a person’s degree of skepticism rather than their philosophical orientation (Bayesian or frequentist). Because it is skepticism that motivates a concern for multiple comparisons and selection effects, it can be said that a lack of concern for multiple comparisons reflects a lack of skepticism.

Another FDA guideline that concerns the use of Bayesian statistics in medical device clinical trials^[18] echoes these sentiments, at least indirectly, with regard to (skeptical) Bayesian analyses:

While the Bayesian approach is often favorable to the investigator with good prior information, the approach can be more conservative than a frequentist approach.

Bayesian methods can be more conservative when skeptical priors are used. Spiegelhalter et al.^[19] define a “skeptical prior” to be one where the prior distribution of the effect size is centered at 0, while an “aggressive prior” places the prior mean of the effect size at a positive value, with small probability to the left of zero, suggesting an a priori belief in a treatment effect. This notion of a skeptical prior is also used to justify the ethics of performing clinical trials^[20]; after all, if there is no doubt about efficacy, then there is no ethical reason for conducting a placebo-controlled trial.

Spiegelhalter’s “skeptical” prior is unrealistic because it suggests that the effect is as likely to be positive as

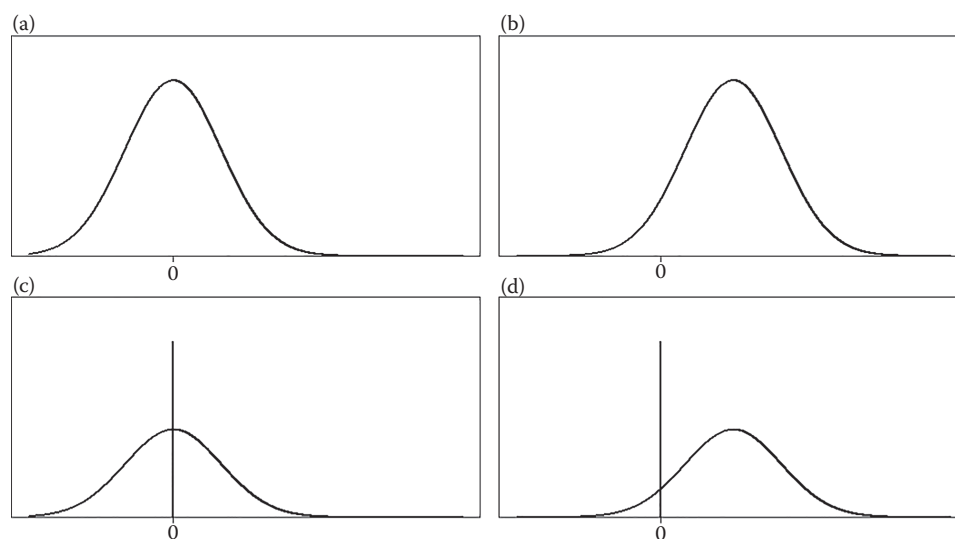


Fig. 1 Skeptical and aggressive priors for treatment difference: (a) semiskeptical; (b) aggressive; (c) skeptical; (d) combines elements of skeptical and aggressive priors

negative. However, the “aggressive” prior is also unrealistic as it contradicts the ethics of running a clinical trial. More flexible priors are needed to model skepticism.

Gönen et al.^[21] estimate that, in oncology trials, a prior that places approximately 50% probability on the null hypothesis has empirical support; Berger and Sellke^[22] discuss the use of such mixture priors that place a positive probability on an effect size of zero. Such a prior is flexible enough to allow that there is (i) residual doubt concerning the existence of an effect and (ii) prior evidence of a positive effect (otherwise, why would an expensive trial be entertained?). Figure 1 displays various prior distributions for effect sizes, including the skeptical and aggressive priors of Spiegelhalter et al., and the corresponding priors that allow a greater degree of skepticism through the use of point masses.

Skepticism also motivates the use of Bayesian hierarchical models that employ shrinkage to downweight extreme, selected effects.^[23] In conjunction with skeptical priors, shrinkage estimates provide excellent solutions to the problem-inflated estimates due to selection effects. Extensive literature exists that discusses these priors, mathematical details surrounding the methodology for clinical trials, and the software used for its implementation.^[21,24,25]

THE CONTROVERSIES

There are several controversial aspects about the use of MCPs in clinical trials. We detail a few of the common ones here.

Controversy 1: In Multiple-Dose Studies, Why Is There a Penalty for “Doing a Better Job”?

Compare, for instance, the following: (i) a placebo-controlled trial with a single active treatment arm and (ii) a placebo-controlled trial with two active treatment arms at different dose levels. Surely the design (ii) is scientifically more informative, so why must the investigator pay a multiplicity penalty?

The reason is that having two doses instead of one increases the likelihood of declaring a significant effect when there is none, if not adjusted properly. The need for MCPs in multiple-dose studies can thus be justified, again, from the skeptical viewpoint. If the goal of the study is to take the “most significant” contrast forward to the approval process, then the evidence will be clearly overstated. A skeptical reviewer may take the point of view that the estimates should be appropriately “shrunk,” which typically requires Bayesian modeling with an appropriately skeptical prior, or that an MCP be adopted. In either case, the hurdle to establish efficacy is higher.

Consider now the case where many doses, say 20, are used. If the goal is to make individual statements about all of the doses, the multiplicity problem is even more evident

than in the previous case with two doses. It is likely that that sample dose–response function will suggest various kinds of non-monotonicities, purely as a result of chance variation in the data. This extreme case suggests that, with increasing doses, the goal should not be multiple comparisons among the doses, but should rather be dose–response curve estimation using regression methods. However, the multiplicity problem still persists because of the uncertainty about the true dose–response model and the inherent need for model selection.^[26]

Controversy 2: Why Not Adjust for Multiplicity of All Tests Performed in the Entire Trial? or in Your Lifetime for That Matter?

This is a common criticism of MCPs. If one is concerned about ever making a type I error, then we may as well give up all hope of making any claim whatsoever, because, as the question suggests, we would have to adjust every test in our lifetime, and the Bonferroni k (or size of the family) would be virtually infinite.

The question reflects a genuine concern and thoughtfulness about the multiplicity issue, but also reflects an avoidance strategy. Most people would agree that multiplicity needs to be managed at some level—it is simply bad science to indiscriminately “data snoop,” with the hope that comparison-wise significance levels provide an appropriate type I error screen.

The “no multiplicity adjustments are necessary” message that has been presented by some statisticians^[27–29] is a potentially dangerous and irresponsible message. If statisticians give the “no adjustment needed” message to our scientist colleagues, they will acquire the idea that it is not an issue, and that “anything goes.” The negative consequences of the “no adjustments necessary” message are that spurious associations are published, results fail to replicate, and inappropriate therapies are recommended.

Multiplicity cannot be controlled, but it can be managed. All competent statisticians will agree that there are places where multiplicity is a concern. But statisticians differ as to where exactly to draw the line, between places where multiplicity is a concern and places where it is not a concern. Statisticians also differ over whether and how to select prior distributions. It turns out that there is a surprising relationship between this issue and the multiplicity issue.^[24,25]

Controversy 3: Suppose That We Wish to Use Multiplicity Adjustment. How Do We Choose “Families” of Tests?

While every clinical trial is unique, Fig. 2 provides some suggested guidance concerning family selection in a “generic” trial. In most trials, there are various tests that are performed; these can be loosely divided into families of “efficacy,” “safety,” and “exploratory” tests. The

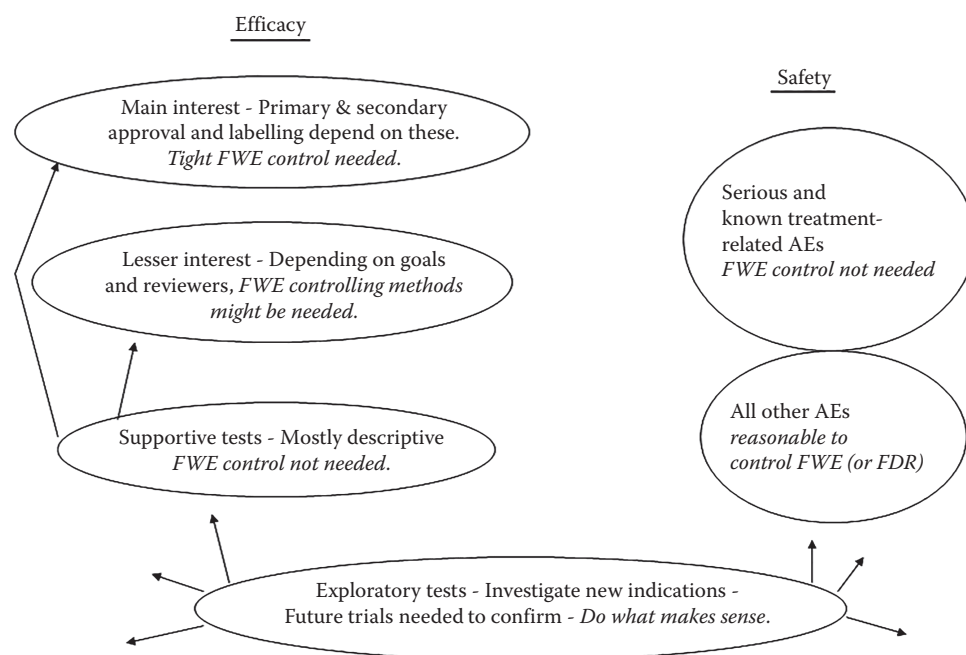


Fig. 2 A categorization of families in clinical trials, with recommended degrees of error rate control within each family

“efficacy” family is further subdivided into “primary efficacy,” “secondary efficacy,” and “supportive” tests. In the primary efficacy family, there are primary endpoint(s) and main secondary endpoints, for which regulatory approval is desired. In this family, the evidence for any claimed significances should be convincing enough to satisfy a skeptical reviewer, and therefore, it is desirable to have strong family-wise control for this set. Moving down the “efficacy side” to the “lesser interest tests,” for which an example may be multiple time point analyses for onset of therapeutic benefit, it is not recommended that this group be included in the primary (top level) family. Nevertheless, some claims may be desired for this set as well; if such tests are placed within a family and multiplicity adjustment is performed over such a family, then it goes without saying that any resulting significances are on stronger ground than if no multiplicity adjustment was performed at all. Finally, the supportive tests (e.g., alternative statistical analyses, alternative surrogate measures) are simply meant to support the primary findings. The intent of these tests is never to claim significance; that is, there is no “rescuing” of a poor primary result here. Rather, these tests can only cast doubt on significances obtained from the primary analyses, and therefore, no type I error adjustment is needed. The categorization of efficacy analyses in Fig. 2 roughly mirrors the rationale clinical endpoint classification by FDA^[9] into primary, secondary, and exploratory endpoints. This classification comes together with the recommendation to control the FWER across all primary and secondary endpoint analyses as they support drug approval and product label, respectively.

On the safety side, the adverse events (AEs) can be divided into two “families,” one including AEs that are

serious and with known treatment-related effects, and the other including the remainder. Type II errors are of much greater concern in the first group. Therefore, no multiplicity adjustment is needed. In fact, even simple trends may be reasonable “flags” here, even if the result is statistically insignificant without multiplicity adjustment.

Alternatively, one may formulate the set of hypotheses regarding serious AEs as multiple equivalence tests in which case it also makes sense to control FWER for that family.^[30]

However, for the remaining set of AEs, there is clearly a need to recognize the multiplicity problem, as one may expect, a priori, that there will be spurious significances in this set when no multiplicity adjustment is adopted. Recognizing the seriousness of type II errors for this case, a reasonable strategy is to present multiplicity-adjusted results side by side with unadjusted results. Reviewers may then make decisions that properly acknowledge the multiplicity problem.

False discovery rate (FDR) controlling methods^[31] have gained enormously in popularity since they offer more power, at the expense of increased type I error frequency. Use of FDR may be helpful here for the collection of secondary AEs.^[32]

At the bottom of the list is the family of exploratory tests. This family is placed in between the safety and efficacy sets, as both types of tests are included. The family may include, as an example, genetic subgroup analyses determined by combinations of single-nucleotide polymorphism markers, and thus may be huge, with the family size k in the thousands. Approval of product is not intended here; rather, the tests are indicative of needs for future study. Standard multiplicity adjustment here seems unreasonable, as power

will be very low; on the other hand, spurious p values less than .0001 are common with so many tests. The cost of type I errors is high here, whether on the safety side or on the efficacy side. On the safety side, there are costs of frivolous lawsuits and lost marketing opportunity costs; on the efficacy side, there are costs of spending resources on ineffective products. FDR controlling methods are useful here as well, and there is currently an enormous stream of literature on the use of such methods for large, exploratory MCPs.

While Fig. 2 provides a high-level guidance on how to select suitable families of hypotheses and error rates, more detailed suggestions still have to be established in many areas of practical concerns, as the boundary between confirmatory inference and exploratory analysis is not always clear. One example is the postmarketing investigation of a drug, where multiple trials may be conducted to understand the effect of alternative outcome measures, in additional subgroups, etc. These trials aim either for label extensions or for publications to better market the new product. One may argue that instead of running separate trials, it is more efficient and ethical to have a single trial (if feasible). The question is whether the type I error rate should be controlled across all endpoints, subgroups, etc., which otherwise would have been addressed in separate studies, each performed at level α . In such situations, however, the boundary versus an exploratory study can be blurred, as multiple study objectives may be prespecified in the study protocol, and one might be tempted to publish or submit only the significant results at an unadjusted level.^[33]

CLOSED TESTING METHODS AND APPLICATIONS TO CLINICAL TRIALS

Despite the fact that it has been more than 40 years since the initial development of the closed testing methodology,^[34] clinical trial designs that utilize closed testing have

only recently become more popular. The reason for this popularity is the incredible flexibility of the closed testing paradigm and its specific applicability to clinical trials^[9]: within its boundaries are found fixed sequence (or a priori ordered) methods, gatekeeping methods, dose–response methods, weighted methods, and methods that consider multiple endpoints and multiple doses simultaneously. An additional reason for the popularity is a guarantee of FWER control. No matter how complex a statistical analysis plan, FWER can be assured, provided that the plan is a special case of a closed testing procedure. Finally, closed testing procedures often are more powerful than classic MCPs, such as the Bonferroni, Tukey, and Dunnett procedures.

Closed testing is easy to describe: First, form all intersections of the elementary hypothesis H_1 , H_2 , etc., and then test all intersections using unadjusted tests. An elementary hypothesis H_i is then declared significant if all intersections that include the elementary hypothesis as a component of the intersection are significant. There is considerable flexibility in the choice of tests for the intersection hypotheses, leading to the wide variety of procedures that fall within the closed testing umbrella.

FIXED SEQUENCE OF OR A PRIORI ORDERED CLINICAL OBJECTIVES

A popular method in clinical trials, the fixed sequence procedure^[9] allows the analyst to impose an a priori order on the hypotheses such as H_1 , H_2 , and H_3 , presuming a natural ordering such as primary and secondary objectives in a clinical trial.^[35] This procedure and the closure method are illustrated in Fig. 3.

Notice that all intersection hypotheses are shown in Fig. 3, above the bottom row where the elementary hypotheses are shown. The flexibility of closed testing allows

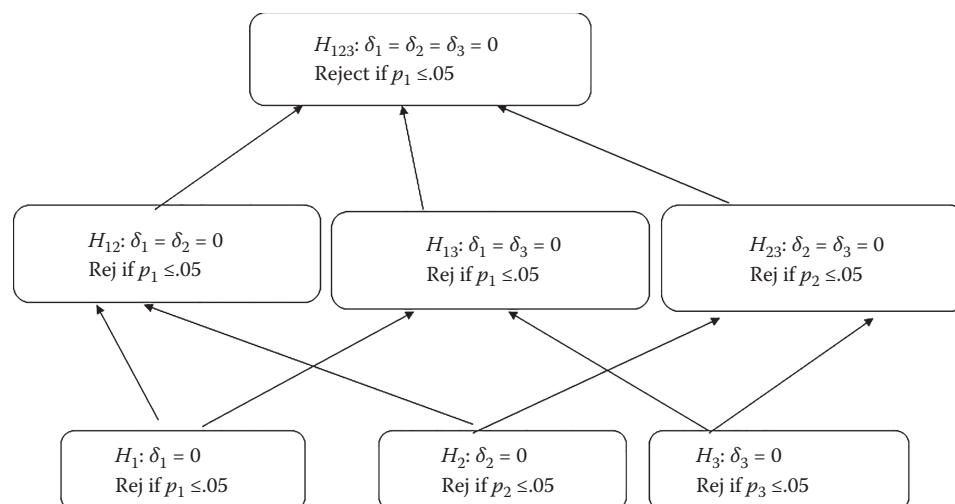


Fig. 3 The fixed sequence procedure as a closed testing procedure

many possible tests for the intersections; in Fig. 3, the tests that give rise to the fixed sequence procedure are shown. Following closed testing, H_1 is rejected if $H_1, H_{12}, H_{13},$ and H_{123} are all rejected using the chosen tests for the intersection hypotheses. If all are tested using the p -value for H_1 only as shown, then one rejects H_1 simply if $p_1 \leq .05$. The remaining hypotheses are rejected using the same procedure; the choice of tests for the intersections thus require that H_2 is rejected if $p_1 \leq .05$ and $p_2 \leq .05$; and that H_3 is rejected if H_2 is rejected if $p_1 \leq .05, p_2 \leq .05,$ and $p_3 \leq .05$. Thus, the fixed sequence procedure is one in which testing continues in the a priori order, stopping as soon as a “fail to reject” decision is reached.

GATEKEEPING PROCEDURES FOR PRIMARY AND SECONDARY ENDPOINTS

Like the fixed sequence method, gatekeeping procedures are also commonly used for clinical trials with ordered objectives.^[9] The prototypical application considers a single primary objective and several, equally weighted secondary objectives. The method allows claims for secondary endpoints once the primary objective of the trial has been met. While other means for declaring secondary variables exist, the gatekeeping strategy can be considered a “gold standard” for placing secondary indications in the label, as full FWER control is assured across the entire collection.

Specifically, let H_1 denote the hypothesis that there is no effect in the primary measure, and let $H_2, H_3, \dots, H_k$ denote a set of secondary endpoints. To make claims about the hypotheses $H_1, H_2, \dots, H_k$ (the primary and secondary endpoints) with full FWER control, the gatekeeping procedure is as follows^[41]:

STEP 1

Test H_1 at (unadjusted) significance level α . If significant, proceed. Otherwise stop.

STEP 2

Test $H_2, H_3, \dots, H_k$ using a closed testing procedure specific to these $k-1$ secondary endpoints. Specifically, if the Bonferroni procedure is used, then the critical point is $\alpha/(k-1)$. Declare all secondary endpoints that meet the more stringent criterion to be statistically significant.

At Step 2, there are a variety of possible tests that could be used instead of the Bonferroni method. Reasonable possibilities include closed testing with the Simes,^[36] O’Brien,^[37] or min- P test.^[38] One can show that gatekeeping procedures are a special case of closed testing,^[39] which often lead to sequentially rejective testing procedures.^[40]

That is, shortcuts can be derived such that the closed hypotheses set is reduced to testing k hypotheses in a data-driven sequence. Because gatekeeping procedures mostly rely on the closed testing principle, they allow mapping of the difference in importance as well as the relationship between the various research questions onto an adequate multiple testing procedure. As a consequence, there has been an explosion of extensions of basic gatekeeping procedure described above that allow multiple doses, multiple endpoints, and weighted hypotheses.^[9]

Consider as an example a Bonferroni-based gatekeeping procedure for testing two primary hypotheses (H_1 and H_2) and, if at least one of them is significant, two secondary hypotheses (H_3 and H_4).^[41] In the simplest case, the two primary hypotheses are initially assigned the local level $\alpha/2$ each, whereas the secondary hypotheses are assigned the local level 0. If one of the primary hypotheses is rejected (i.e., either H_1 or H_2), the local level $\alpha/2$ is split into half and passed on to the secondary hypotheses. Testing continues with the remaining, nonrejected hypotheses at the updated local significance levels until no further hypothesis can be rejected. Further extension to gatekeeping procedures addressing complex questions is available.^[42–46]

In the meantime, the development of gatekeeping procedures has become rather technical, and the underlying test strategies have often become difficult to communicate. Graphical approaches to sequentially rejective multiple testing procedures have been proposed instead and recommended in the recent FDA guideline,^[9] which allow better communication with clinical teams.^[47,48] Using a graphical approach, null hypotheses are represented by a set of vertices with associated weights representing local significance levels. The weight associated with a directed edge between any two vertices indicates the fraction of the (local) significance level at the initial vertex (tail) that is added to the significance level at the terminal vertex (head), if the hypothesis at the tail is rejected. See Fig. 4 for an example, which visualizes the gatekeeping procedure discussed previously.^[41]

While the technical development of gatekeeping procedures is still evolving, attention needs to be given to potential “illogical problems”^[49] and the need to ensure “consistency”^[50] when selecting a suitable testing strategy.

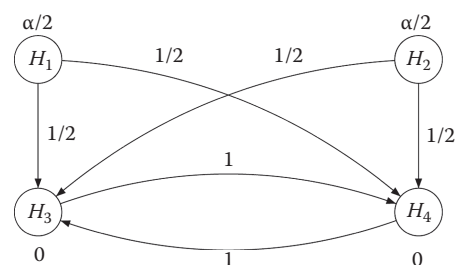


Fig. 4 Graphical illustration of a gatekeeping procedure^[41] with four hypotheses

CONCLUSION

Spurred by the latest regulatory guidelines,^[8,9] this article provided a current look at the issues, controversies, and methodologies that concern multiple comparisons in clinical trials. An insight into the need for multiple comparisons from the skeptical point of view, a point of view that also provides guidance for analysis strategies, was given. Some MCPs are currently popular; gatekeeping procedures were described in detail. Finally, the use of Bayesian methods and FDR controlling methods, both of which are gaining popularity for addressing certain aspects of multiplicity in clinical trials, was suggested. A glaring omission in this article is group sequential and adaptive designs; we instead refer to the literature for the available statistical methods and regulatory guidance.^[51–53]

REFERENCES

1. Tamhane, A.C. Multiple comparisons. *Handbook of Statistics*; North-Holland: Amsterdam, 1996, Vol. 13, 587–630.
2. Hochberg, Y.; Westfall, P.H. On some multiplicity problems and multiple comparisons procedures in biostatistics. *Handbook of Statistics*; Elsevier Sciences B.V.: Amsterdam, 2000, Vol. 18, 75–113.
3. Lakshminarayanan, M.Y. Multiple comparisons. *Encyclopedia of Biopharmaceutical Statistics*, 1st Ed.; Marcel Dekker, Inc.: New York, 2000, 325–329.
4. Bauer, P.; Roßhmel, J.; Maurer, W.; Hothorn, L. Testing strategies in multi-dose experiments including active control. *Stat. Med.* **1998**, *17* (18), 2133–2146.
5. Koch, G.G.; Gansky, S.A. Statistical considerations for multiplicity in confirmatory protocols. *Drug Inf. J.* **1996**, *30* (2), 523–534.
6. Zhang, J.; Hui, Q.; Ng, J.; Stepanavage, M.E. Some statistical methods for multiple endpoints in clinical trials. *Control. Clin. Trials* **1997**, *18* (3), 204–221.
7. Dmitrienko, A.; Tamhane, A.; Bretz, F. *Multiple Testing Problems in Pharmaceutical Statistics*; Chapman & Hall: Boca Raton, FL, 2009.
8. Committee for Human Medicinal Products. *Guideline on Multiplicity Issues in Clinical Trials*. European Medicines Agency, 2017.
9. U.S. Food and Drug Administration. *Multiple Endpoints in Clinical Trials: Guidance for Industry*. U.S. Department of Health and Human Services Food and Drug Administration, 2017.
10. Westfall, P.H.; Bretz, F.; Tobias, R.D. Directional error rates of closed testing procedures. *Stat. Biopharm. Res.* **2013**, *5* (4), 345–355. doi: 10.1080/19466315.2013.818575.
11. Bretz, F.; Westfall, P.H. Multiplicity and replicability: Two sides of the same coin. *Pharm. Stat.* **2014**, *13* (6), 343–344.
12. Ioannidis, J.P.A. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* **2005**, *294* (2), 218–228.
13. Assmann, S.F.; Pocock, S.J.; Enos, L.E.; Kasten, L.E. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* **2000**, *355* (9209), 1064–1069.
14. Fleming, T.R. Evaluating therapeutic interventions: Some issues and experiences (with rejoinders). *Stat. Sci.* **1992**, *7* (4), 428–456.
15. Harris, R. *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions*; Basic Books: New York.
16. Duncan, D.B. Multiple range and multiple F tests. *Biometrics* **1955**, *11* (1), 1–42.
17. Cook, R.J.; Farewell, V.T. Multiplicity considerations in the design and analysis of clinical trials. *J. R. Stat. Soc. Ser. A Stat. Soc.* **1996**, *159* (1), 93–110.
18. U.S. Food and Drug Administration. *Guidance for Industry and FDA Staff: Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials*. U.S. Department of Health and Human Services Food and Drug Administration, 2010.
19. Spiegelhalter, D.J.; Freedman, L.S.; Parmar, M.K.B. Bayesian approaches to randomized trials. *J. R. Stat. Soc. Ser. A Stat. Soc.* **1994**, *157* (3), 357–416.
20. Lee, S.J.; Zelen, M. Clinical trials and sample size considerations: Another perspective. *Stat. Sci.* **2000**, *15* (2), 95–110.
21. Gönen, M.; Johnson, W.O.; Lu, Y.; Westfall, P. The Bayesian two-sample t test. *Am. Stat.* **2005**, *59*, 252–257.
22. Berger, J.O.; Sellke, T. Testing a point null hypothesis: The irreconcilability of *p* values and evidence. *J. Am. Stat. Assoc.* **1987**, *82*, 112–122.
23. Gelman, A.; Hill, J.; Yajime, M. Why we (usually) don't have to worry about multiple comparisons. *J. Res. Educ. Eff.* **2012**, *5*, 189–211.
24. Westfall, P.H.; Tobias, R.; Rom, D.; Wolfinger, R.; Hochberg, Y. *Multiple Comparisons and Multiple Tests Using the SAS® System*; SAS® Institute Inc.: Cary, NC, 1999.
25. Gönen, M.; Westfall, P.H.; Johnson, W.O. Bayesian multiple testing for two-sample multivariate endpoints. *Biometrics* **2003**, *59* (1), 76–82.
26. Bretz, F.; Pinheiro, J.C.; Branson, M. Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* **2005**, *61* (3), 738–748.
27. Saville, D.J. Multiple comparison procedures: The practical solution. *Am. Stat.* **1990**, *44* (2), 174–180.
28. Rothman, K.J. No adjustments are needed for multiple comparisons. *Epidemiology* **1990**, *1* (1), 43–46.
29. Longford, N.T.; Nelder, J.A. Statistics versus statistical science in the regulatory process. *Stat. Med.* **1999**, *18* (17–18), 2311–2320.
30. Hsu, J.C.; Berger, R.L. Stepwise confidence intervals without multiplicity adjustment for dose-response and toxicity studies. *J. Am. Stat. Assoc.* **1999**, *94* (446), 468–482.
31. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate—A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1995**, *57* (1), 289–300.
32. Mehrotra, D.V.; Heyse, J.F. Use of the false discovery rate for evaluating clinical safety data. *Stat. Methods Med. Res.* **2004**, *13* (3), 227–238.
33. Bretz, F.; Maurer, W.; Gallo, P. Discussion of “some controversial multiple testing problems in regulatory applications” by H.M.J. Hung and S.J. Wang. *J. Biopharm. Stat.* **2009**, *19* (1), 25–34.
34. Marcus, R.; Peritz, E.; Gabriel, K.B. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **1976**, *63* (3), 655–660.

35. Maurer, W.; Hothorn, L.; Lehmacher, W. Multiple comparisons in drug clinical trials and preclinical assays: A-priori ordered hypotheses. In *Biometrie in der Chemisch-Pharmazeutischen Industrie*; Fischer Verlag: Stuttgart, Germany, 1995, Vol. 6, 3–18.
36. Hommel, G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **1988**, *75* (2), 383–386.
37. Lehmacher, W.; Wassmer, G.; Reitmeir, P. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* **1991**, *47* (2), 511–532.
38. Westfall, P.H.; Wolfinger, R.D. *Closed Multiple Testing Procedures and PROC MULTTEST*; 2nd Ed.; SAS Observ: Cary, NC, June 2000; http://support.sas.com/kb/22/addl/fusion22950_1_multtest.pdf.
39. Westfall, P.H.; Krishen, A. Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *J. Stat. Plan. Inference* **2001**, *99* (1), 25–40.
40. Hommel, G.; Bretz, F.; Maurer, W. Powerful short-cuts for multiple testing procedures with special reference to gate-keeping strategies. *Stat. Med.* **2007**, *26* (22), 4063–4073.
41. Dmitrienko, A.; Offen, W.W.; Westfall, P.H. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Stat. Med.* **2003**, *22* (15), 2387–2400.
42. Wiens, B.; Dmitrienko, A. The fallback procedure for evaluating a single family of hypotheses. *J. Biopharm. Stat.* **2005**, *15* (6), 929–942.
43. Dmitrienko, A.; Wiens, B.L.; Tamhane, A.C.; Wang, X. Tree structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Stat. Med.* **2007**, *26* (12), 2465–2478.
44. Huque, M.F.; Alosh, M. A flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials. *J. Stat. Plan. Inference* **2008**, *138* (2), 321–335.
45. Guilbaud, O. Simultaneous confidence regions corresponding to Holm's stepdown procedure and other closed-testing procedures. *Biom. J.* **2008**, *50* (5), 678–692.
46. Strassburger, K.; Bretz, F. Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni based closed tests. *Stat. Med.* **2008**, *27* (24), 4914–4927.
47. Bretz, F.; Maurer, W.; Brannath, W.; Posch, M. A graphical approach to sequentially rejective multiple test procedures. *Stat. Med.* **2009**, *28* (4), 586–604.
48. Burman, C.-F.; Sonesson, C.; Guilbaud, O. A recycling framework for the construction of Bonferroni-based multiple tests. *Stat. Med.* **2009**, *28* (5), 739–761.
49. Hung, H.M.J.; Wang, S.J. Some controversial multiple testing problems in regulatory applications. *J. Biopharm. Stat.* **2009**, *19* (1), 1–11.
50. Alosh, M.; Bretz, F.; Huque, M. Advanced multiplicity adjustment methods in clinical trials. *Stat. Med.* **2014**, *33* (4), 693–713.
51. Bauer, P.; Bretz, F.; Dragalin, V.; Koenig, F.; Wassmer, G. Twenty-five years of confirmatory adaptive designs: Opportunities and pitfalls (with discussion). *Stat. Med.* **2016**, *35* (3), 325–367.
52. Committee for Human Medicinal Products. Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. EMEA Doc. Ref. CHMP/EWP/2459/02, October 2007. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003616.pdf.
53. Wassmer, G.; Brannath, W. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer: Heidelberg, Germany, 2016.

Multi-Regional Clinical Trials: Consistency Test

Jiayin Zheng

Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, U.S.A.

Lisa Ying

Bristol-Myers Squibb Company, Princeton, New Jersey, U.S.A.

Na Zeng

National Clinical Research Center for Digestive Disease, Beijing Friendship Hospital, Capital Medical University, Beijing, China

Shein-Chung Chow

Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

In recent years, multiregional clinical trials (MRCTs) that conduct clinical trials simultaneously in Asia-Pacific region, Europe and the United States have become very popular for global pharmaceutical development. The purpose of MRCTs is to shorten the time for pharmaceutical development and regulatory development, submission, and approval around the world. In practice, however, clinical results observed from some regions (subpopulation) may not be consistent with the results from other regions and/or all regions combined (entire population). It is a concern for regulatory review and approval of the clinical results of the region that is inconsistent with all other regions. The inconsistency observed may be due to ethnic differences in different regions and/or similar but different regulatory requirements, study protocol/designs, study endpoints, or time points of assessment, or by random chance due to small sample size for the region. In this article, critical issues that are commonly encountered in an MRCT are discussed. Several statistical tests for consistency (similarity) between clinical results observed from a specific subpopulation and the entire population are presented. As most published articles discussed consistency evaluation for superiority situations, we discussed consistency evaluation for non-inferiority situation in this article through a simulated example concerning consistency in some countries.

Keywords: Assurance probability; Consistency index; Non-inferiority; Sensitivity index; Similarity; Superiority; Test for consistency.

INTRODUCTION

In recent years, multi-regional (or multi-national) multi-center clinical trials have become very popular in global pharmaceutical and/or clinical development. The main purpose of a multiregional clinical trial (MRCT) is not only to assess the efficacy of the test treatment over all regions in the trial but also to bridge the overall effect of the test treatment to each of the region in the trial. Most importantly, an MRCT is to shorten the time for pharmaceutical development and regulatory development, submission, and approval around the world. Although MRCTs provide the opportunity to fully utilize clinical data from all regions to support regional (local) registration, some critical issues such as regional differences (e.g., culture and medical practice/perception) and possible treatment-by-region

interaction that may have an impact on the validity of the multiregional trials inevitably occur.

In multiregional trials for global pharmaceutical/clinical development, one of the commonly encountered critical issues is that clinical results observed from some regions (e.g., Asia-Pacific region) are inconsistent with clinical results from other regions (e.g., the European Community) or global results (i.e., all regions combined). The inconsistency in clinical results between different regions (e.g., Asia-Pacific region and the European Community and/or the United States) could be due to the difference in ethnic factor. In this case, the dose or dose regimen may require adjustment or a bridging study may be required before the data can be pooled for an overall combined assessment of the treatment effect. As a result, evaluation of consistency between specific regions (subpopulation) and all regions

combined (global population) is necessarily conducted before regional registration (e.g., Japan and China). It should be noted that different regions may have different requirements regarding sample sizes at specific regions in order to fulfill with regulatory requirement for registration.^[1]

In practice, consistency between clinical results observed from a subpopulation (specific region such as Japan or China) and the entire population (all regions combined) is often interpreted as similarity and/or equivalence between the two populations in terms of treatment effect (i.e., safety and/or efficacy). Along this line, several statistical methods including test for consistency,^[2] assessment of consistency index,^[3] evaluation of sensitivity index,^[4] achievement of reproducibility and/or generalizability,^[5] Bayesian approach,^[6,7] and the Japanese approach for evaluation of assurance probability^[1] have been proposed in the literature.^[8] The purpose of this article is to provide an overview of these methods.

In the next section, some critical issues that are commonly encountered in multiregional, multicenter trials are discussed. Several statistical methods for testing consistency or similarity/equivalence in clinical results observed from a subpopulation compared to that of the entire population are described in Section “Statistical Methods.” In Section “An Example (for a Non-Inferiority Situation),” an example concerning a multiregional study involving a regional regulatory submission is discussed to illustrate the proposed statistical methods. Some concluding remarks are given at the end of this article.

ISSUES IN MRCTS

A multiregional clinical study is a trial conducted at more than one distinct region where the data collected from these centers are intended to be analyzed as a whole. Within a given region, the trial may be conducted as either a single-site study or a multi-center trial. In a multiregional trial, some practical issues are commonly encountered due to potential differences among centers within region and differences between regions.

Multicenter Trials

A multi-center trial is a trial conducted at more than one distinct center where the data collected from these centers are intended to be analyzed as a whole. At each center, an *identical* study protocol is used. A multicenter trial is a trial with a center or site as a natural blocking or stratified variable that provides replications of clinical results. It should permit an overall estimation of the treatment difference for the targeted patient population across various centers. For a multi-center trial, the Food and Drug Administration (FDA) guideline suggests that individual center results should be presented. In addition, the FDA suggests that statistical tests for homogeneity across

centers (i.e., for detecting possible treatment-by-center interaction) be performed. Any extreme or opposite results among centers should be noted and discussed. If there is no treatment-by-center interaction, the data can be pooled for analysis across centers.

Another issue is related to the use of a central laboratory for testing of samples collected from different centers, which may have a significant impact on the assessment of efficacy and safety of the test treatment under investigation. A central laboratory provides a consistent assessment for laboratory tests. If a central laboratory is not used, the assessment of laboratory tests may differ from center to center depending on the equipment, analyst, and laboratory normal ranges used at that center. In such a case, possible confounding and interaction effects makes it difficult to combine the laboratory values obtained from the different centers for an unbiased assessment of the safety and efficacy of the test treatment.

For statistical analysis of data collected from a multi-center trial, if there is no evidence of treatment-by-center interaction, the data can be pooled for analysis for an overall assessment of the treatment effect across centers. Along this line, Nevius^[9] proposes a set of four conditions under which evidence from a multi-center trial would provide sufficient statistical evidence of efficacy. These conditions are summarized as follows:

1. The combined analysis shows significant results.
2. There is consistency over centers in terms of direction of results.
3. There is consistency over centers in terms of producing nominally significant results in centers with sufficient power.
4. Multiple centers show evidence of efficacy after adjustment for multiple comparisons.

Multiregional, Multi-center Trials

Unlike a single multi-center trial within a given region, a multiregional trial is much more complicated, which involves several multi-center trials from different regions. In addition to those practical issues that are commonly seen in a multi-center trial, there are more critical issues that may have an impact on the assessment of treatment effect in global pharmaceutical/clinical development. These issues include, but are not limited to, potential differences in (1) study requirements (due to different regulatory requirements, see [Table 1](#) as an example), (2) ethnic factors, (3) culture, and (4) medical practice or perception in different regions. Thus, in practice, it is suggested that critical issues such as representativeness, bias/variation control, heterogeneity (similarities and dissimilarities), consistency, and poolability across regions, which may have an impact on the validity of the multiregional trial, must be carefully evaluated in order to provide an overall assessment of the test treatment under investigation.

Table 1 Examples of different regional requirements

Regional requirement	European Medicines Agency	US FDA
NI margin (HIV-infected patients)	10%–15%	10%–12%
Time points (HIV-RNA level)	16 weeks	24 weeks
Endpoint (atrial fibrillation)	Prevention of any recurrence	Delay in symptomatic recurrence
Endpoint (hepatitis B virus)	Combined composite of virological, histological, and biochemical responses	Complete virological response at 52 weeks

In practice, it is often a concern that only positive clinical results are observed in global population (i.e., all regions combined), and yet some regions (i.e., subpopulations) fail to show positive results or show positive results with insufficient power. This inconsistency, if it is not purely due to chance, is a concern for regional (local) registration. As inconsistency may be due to the fact that studies conducted at different regions may be conducted with similar but different (1) study requirements, (2) drug products and doses, (3) patient populations with different ethnic factors, (4) sample sizes, and (5) evaluability criteria, regional regulatory agencies often request the inconsistency be carefully evaluated during the regulatory review and approval process.

For evaluation of consistency, the sponsors can use a null hypothesis of consistency, and then the alternative hypothesis of inconsistency is concluded if the null hypothesis is rejected (often with insufficient power for establishing non-inferiority (NI) or similarity between the subpopulation and the entire population). It should be noted that regional regulatory agencies are not likely to accept this assessment with limited number of subjects available in a region. On the other hand, the regional regulatory agencies may accept the conclusion that there is little evidence to demonstrate that the subpopulation is inconsistent with the entire population. As a result, an additional clinical study may be required to be conducted on the subpopulation in the region in order to confirm the consistency or similarity between the subpopulation and the entire population. As an alternative, we can use a null hypothesis of inconsistency or use one of the other estimation-based methods described in this article.

STATISTICAL METHODS

As mentioned in the “Introduction” section, several methods that have been proposed in the literature in different context could be useful for evaluation of consistency between a subpopulation and the entire population in

multiregional studies. These methods are briefly outlined in the subsequent sections.

Test for Consistency

For an MTCT involving K regions, denote by $W = \{W_1, \dots, W_K\}$ the results of the multiregional trial, where W_i is the standardized test statistics observed from the i th region. To determine whether the local results from a given region is consistent with the global results, Shih^[2] proposed to construct a predictive probability function, $P(v|W)$, where v is the study conducted at a given region. The predictive probability $P(v|W)$ provides a measure of the plausibility of v given the results of W . Thus, Shih's test for consistency is summarized next.

For a given region i , we first construct the predictive probability function, $P(W_i|W)$, of the results observed from the region. We then compare $P(W_i|W)$ with the plausibility of each of the results from other regions, i.e., $P(W_j|W)$, where $j \neq i$ and $j = 1, \dots, K$. We conclude that the results of W_i is consistent with $W = \{W_1, \dots, W_K\}$ if and only if

$$P(W_i|W) \geq \min \{P(W_j|W), j \neq i, j = 1, 2, \dots, K\}. \quad (1)$$

To accommodate the superiority test with the case of the smaller the better, the method can be modified as

$$W_i - W \leq \max \{W_j - W, j \neq i, j = 1, 2, \dots, K\}.$$

If the above inequality holds, we conclude that the results of W_i are consistent with $W = \{W_1, \dots, W_K\}$.

To accommodate the superiority test with the case of the larger the better, the method can be modified as

$$W_i - W \geq \min \{W_j - W, j \neq i, j = 1, 2, \dots, K\}.$$

If the above inequality holds, we conclude that the results of W_i are consistent with $W = \{W_1, \dots, W_K\}$.

Assessment of Consistency Index

Let U and W be the measures of the primary study endpoint (treatment effects) of a given MRCT from a subpopulation (e.g., a specific region) and the global population (i.e., all regions combined), respectively, where $X = \log U$ and $Y = \log W$ follow normal distributions with means μ_x , μ_y , and variances V_x , V_y , respectively. Similar to the idea of using $P(X < Y)$ to assess reliability in statistical quality control,^[10,11] Tse et al.^[3] proposed the following probability as an index to assess the consistency of clinical results observed from the sub-population as compared to the global population

$$p = P\left(1 - \delta < \frac{U}{W} < \frac{1}{1 - \delta}\right), \quad (2)$$

where $0 < \delta < 1$ and is defined as a limit that allows for consistency. Tse et al.^[3] referred to p as the consistency index.

Thus, p tends toward 1 as δ tends toward 1. For a given δ , if p is close to 1, materials U and W are considered to be similar or approximately identical. It should be noted that a small δ implies the requirement of a high degree of consistency between measure U and measure W . In practice, it may be difficult to meet this narrow specification for consistency. Under the normality assumption of $X = \log U$ and $Y = \log W$, Eq. 2 can be rewritten as

$$\begin{aligned} p &= P(\log(1-\delta) < \log U - \log W < -\log(1-\delta)) \\ &= \Phi\left(\frac{-\log(1-\delta) - (\mu_x - \mu_y)}{\sqrt{V_x + V_y}}\right) \\ &\quad - \Phi\left(\frac{\log(1-\delta) - (\mu_x - \mu_y)}{\sqrt{V_x + V_y}}\right). \end{aligned}$$

where $\Phi(z_0) = P(Z < z_0)$ with Z being a standard normal random variable. Therefore, the consistency index p is a function of the parameters $\theta = (\mu_x, \mu_y, V_x, V_y)$, i.e., $p = h(\theta)$. Suppose that observations $X_i = \log U_i, i = 1, \dots, n_x$ and $Y_i = \log W_i, i = 1, \dots, n_y$ are collected in an assay study. Then, using the invariance principle, the maximum likelihood estimator of p can be obtained as

$$\hat{p} = \Phi\left(\frac{-\log(1-\delta) - (\bar{X} - \bar{Y})}{\sqrt{\hat{V}_x + \hat{V}_y}}\right) - \Phi\left(\frac{\log(1-\delta) - (\bar{X} - \bar{Y})}{\sqrt{\hat{V}_x + \hat{V}_y}}\right), \quad (3)$$

where $\bar{X} = \frac{1}{n_x} \sum_{i=1}^{n_x} X_i$, $\bar{Y} = \frac{1}{n_y} \sum_{i=1}^{n_y} Y_i$, $\hat{V}_x = \frac{1}{n_x} \sum_{i=1}^{n_x} (X_i - \bar{X})^2$, and $\hat{V}_y = \frac{1}{n_y} \sum_{i=1}^{n_y} (Y_i - \bar{Y})^2$. In other words, $\hat{p} = h(\hat{\theta}) = h(\bar{X}, \bar{Y}, \hat{V}_x, \hat{V}_y)$. Furthermore, it can be easily verified that the following asymptotic result holds. Tse et al.^[3] showed that $\hat{p}$ is asymptotically normal with mean $E(\hat{p})$ and variance $Var(\hat{p})$, where $E(\hat{p}) = p + B(p) + o\left(\frac{1}{n}\right)$ and $Var(\hat{p}) = C(p) + o\left(\frac{1}{n}\right)$. The detailed expressions of $B(p)$ and $C(p)$ are given in the work of Tse et al.^[3] In other words,

$$\frac{\hat{p} - E(\hat{p})}{\sqrt{Var(\hat{p})}} \rightarrow N(0, 1).$$

Hypotheses testing of the consistency index p can also be conducted based on the asymptotic normality of $\hat{p}$. Consider the following hypotheses:

$$H_0 : p \leq p_0 \text{ vs } H_1 : p > p_0.$$

We can choose p_0 as a pre-specified constant, say 0.8. Also, we recommend the value of some proportion (say 0.8)

multiplying the $\hat{p}$ of the corresponding reference–reference test. Intuitively, the latter has more reference value, i.e., it is obtained from the evidence of the corresponding reference–reference test that can be seen as some reference standard, and we name it as the R–R choice.

We would reject the null hypothesis in favor of the alternative hypothesis of consistency. Under H_0 , we have

$$\frac{\hat{p} - p_0 - B(\hat{p})}{\sqrt{Var(\hat{p})}} \sim N(0, 1). \quad (4)$$

Thus, we reject the null hypothesis H_0 at the α level of significance if

$$\frac{\hat{p} - p_0 - B(\hat{p})}{\sqrt{Var(\hat{p})}} > Z_\alpha.$$

This is equivalent to rejecting the null hypothesis H_0 when

$$\hat{p} > p_0 + B(\hat{p}) + Z_\alpha \sqrt{Var(\hat{p})}.$$

For a superiority test, we are probably more concerned with $p = P\left(\frac{U}{W} < \frac{1}{1-\delta}\right)$. A similar procedure to testing the consistency can be derived. For simplicity, we do not give the details here.

Note that this method only applies to the endpoints following lognormal distribution.

Evaluation of Sensitivity Index

In an MRCT, denote the mean and standard deviation of the target endpoint of the entire population (all regions combined) by μ_2 and σ_2 , respectively. Similarly, denote these of the subpopulation (a or some specific region(s)) by μ_1 and σ_1 . Since the two populations are similar but slightly different, it is reasonable to assume that $\mu_1 = \mu_2 + \varepsilon$ and $\sigma_1 = C\sigma_2$ ($C > 0$), where ε is referred to as the shift in location parameter (population mean) and C is the inflation factor of the scale parameter (population standard deviation). Thus, the (treatment) effect size adjusted for standard deviation of the subpopulation (μ_1, σ_1) can be expressed as follows:

$$\delta_1 = \left| \frac{\mu_1}{\sigma_1} \right| = \left| \frac{\mu_2 + \varepsilon}{C\sigma_2} \right| = |\Delta| \left| \frac{\mu_2}{\sigma_2} \right| = |\Delta| \delta_2,$$

where $\Delta = \frac{1 + \varepsilon/\mu_2}{C}$ and δ_2 is the effect size adjusted for the standard deviation of the entire population (μ_2, σ_2). Δ is referred to as a sensitivity index measuring the change in effect size between the subpopulation and the entire population.^[4]

As it can be seen, if $\varepsilon = 0$ and $C = 1$ then $\delta_1 = \delta_2$. That is, the effect sizes of the two populations are identical.

Applying the concept of bioequivalence assessment, we can claim that the effect sizes of the two patient populations are consistent if the confidence interval of $|\Delta|$ is within (80%, 120%).

In practice, the shift in location parameter (ε) and/or the change in scale parameter (C) could be seen as random. If both ε and C are fixed, the sensitivity index can be assessed based on the sample means and the sample variances obtained from the entire population and the subpopulation. As indicated by,^[4] ε and C can be estimated by $\hat{\varepsilon} = \hat{\mu}_1 - \hat{\mu}_2$ and $\hat{C} = \hat{\sigma}_1 / \hat{\sigma}_2$, respectively, where $(\hat{\mu}_2, \hat{\sigma}_2)$ and $(\hat{\mu}_1, \hat{\sigma}_1)$ are some estimates of (μ_2, σ_2) and (μ_1, σ_1) based on the entire population and the subpopulation, respectively. Consequently, the sensitivity index Δ can be estimated by $\hat{\Delta} = \frac{1 + \hat{\varepsilon} / \hat{\mu}_2}{\hat{C}}$, and the corresponding confidence interval can be obtained based on normal approximation.^[4,12]

In real-world problems, however, ε and C could be either fixed or random variables. In other words, there are three possible scenarios: (1) the case where ε is random and C is fixed, (2) the case where ε is fixed and C is random, and (3) the case where both ε and C are random. Statistical inference of Δ under each of these possible scenarios has been studied by Lu et al.^[12]

Besides, we also give a simplified version: if $\hat{\Delta}$ is within (80%, 120%), the consistency is claimed. For superiority tests, the criterion can be $\hat{\Delta} > 80\%$.

Achieving Reproducibility and/or Generalizability

Alternatively, Shao and Chow^[5] proposed the concept of reproducibility and generalizability probabilities for assessing consistency between specific subpopulations and the global population. If the influence of the ethnic factors is negligible, then we may consider the reproducibility probability to determine whether the clinical results observed in a patient population are reproducible in a different patient population. If there is a notable ethnic difference, the concept of generalizability probability can be used to determine whether the clinical results observed in a patient population can be generalized to a similar but slightly different patient population due to the differences in ethnic factor.

Modified from the original definitions of reproducibility and generalizability probabilities, we proposed three methods for normal-based test procedure and power calculation. The versions for t -distribution-based test procedure and power calculation can similarly be derived.

Specificity reproducibility probability for inequality test: If we are concerned with the null hypothesis $H_0: \mu_1 = \mu_2$ and the alternative hypothesis $H_1: \mu_1 \neq \mu_2$, where μ_1 and μ_2 are the means of treatment effects of a region and the global area (excluding the region of interest), respectively. Denote σ_1^2 and σ_2^2 as the variances of

treatment effects of the region and the global area (excluding the region of interest), respectively. Consider the case without the assumption of equal variance. When both sample sizes n_1 and n_2 are large, an approximate α level test rejects H_0 when $|T| > z_{1-\alpha/2}$, where

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. Since T is approximately distributed as $N(\theta, 1)$ with

$$\theta = \frac{\mu_1 - \mu_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}$$

with a pre-specified level of β (say 0.2), we get the two-sided $1-\beta$ confidence interval for θ as

$(\hat{\theta}_- - \hat{\theta}_+) = \left(T - z_{1-\beta/2}, T + z_{1-\beta/2}\right)$. Then we get two types of

one-sided $1-\beta$ confidence upper bound for $|\theta|$, denoted as $|\hat{\theta}_+^1|$ and $|\hat{\theta}_+^2|$: the first is $\max\{|\hat{\theta}_-|, |\hat{\theta}_+|\}$; the second is the non-negative value v satisfying $P\{|\theta| \leq v | T\} = 1 - \beta$, which is equivalent to the equation of $\Phi(v - T) - \Phi(-v - T) = 1 - \beta$, with $\Phi(\Delta)$ being the distribution function of the standard normal distribution. We define specificity reproducibility probability (with $|\hat{\theta}_+^j|$ taken as the true value of $|\theta|$, and standing for a unfavorable case for $|\theta|$ to some extent) as

$$P\left\{|T| \leq z_{1-\frac{\alpha}{2}} | |\hat{\theta}_+^j|\right\}, j = 1, 2,$$

which is equal to $\Phi\left(z_{1-\frac{\alpha}{2}} - |\hat{\theta}_+^j|\right) - \Phi\left(-z_{1-\frac{\alpha}{2}} - |\hat{\theta}_+^j|\right)$. Compare the specificity reproducibility probability with a pre-specified constant (say 0.8), or the value of the R-R choice mentioned previously. If the specificity reproducibility probability is larger, we conclude that the consistency holds.

Superiority reproducibility probability: With respect to the superiority test with the case of the smaller the better, consider the null hypothesis $H_0: \mu_1 \geq q * \mu_2$ and the alternative hypothesis $H_1: \mu_1 < q * \mu_2$, where q is a pre-specified constant (say 0.5). The notations here have the same meanings as those in the previous paragraph. We have

$$T = \frac{\bar{x}_1 - q\bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{q^2 s_2^2}{n_2}\right)}}$$

and

$$\theta = \frac{\mu_1 - q\mu_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{q^2\sigma_2^2}{n_2}\right)}}.$$

The rejection region is $\{T < z_\alpha\}$. The one-sided $1 - \beta$ confidence upper bound for θ , denoted as $\hat{\theta}_+$, is $T + z_{1-\beta}$. The superiority reproducibility probability is defined as

$$P\{T < z_\alpha | \hat{\theta}_+\} = \Phi(z_\alpha - \hat{\theta}_+) - \Phi(z_\alpha - T - z_{1-\beta}).$$

Compare the superiority reproducibility probability with a pre-specified constant (say 0.8), or the value of the R-R choice mentioned previously. If the superiority reproducibility probability is larger, we conclude that the consistency holds.

Reproducibility probability ratio for inequality test: For the null hypothesis $H_0: \mu_1 = \mu_2$ and the alternative hypothesis $H_1: \mu_1 \neq \mu_2$, where μ_1 and μ_2 are the means of the treatment and the reference for a specific region (or the global area). Denote σ_1^2 and σ_2^2 as the variances of the treatment and the reference for a specific region (or the global area), respectively. Consider the case without the assumption of equal variance. When both sample sizes n_1 and n_2 are large, an approximate α level test rejects H_0 when $|T| > z_{1-\alpha/2}$, where

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

and $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution. Since T is approximately distributed as $N(\theta, 1)$ with

$$\theta = \frac{\mu_1 - \mu_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}},$$

the reproducibility probability by using the estimated power approach is given by

$$\hat{P} = \Phi(T - z_{1-\alpha/2}) - \Phi(-T - z_{1-\alpha/2}).$$

Denote $\hat{P}_S$ and $\hat{P}_G$ as the reproducibility probabilities for a region and the global area, respectively. If $\hat{P}_S/\hat{P}_G \geq q$, where q is a pre-specified constant, then the consistency of the region's results and the global results is concluded.

Besides, the reproducibility probability by using the confidence bound approach can be used as in the literature

of Shao and Chow,^[5] and the same criterion for testing consistency as above applies.

Reproducibility probability ratio for superiority test: Similar to the reproducibility probability ratio for inequality test, we can also consider the ratio for superiority test (with the case of the larger the better): $H_0: \mu_1 \geq \mu_2$ vs. $H_1: \mu_1 < \mu_2$. After obtaining the ratios for the region and the global area, the same criterion for testing consistency applies.

Bayesian Approach

For testing consistency or similarity between clinical results observed from a subpopulation and those obtained from the global population, Chow and Hsiao^[7] proposed the following Bayesian approach through the following hypothesis:

$$H_0: \Delta \geq 0 \text{ vs. } H_a: \Delta < 0,$$

where $\Delta = \mu_1 - \mu_2$, in which μ_1 and μ_2 are the population means of the treatment effects of the subpopulation and the global population, respectively. Let X_i and Y_j , $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, N$, be some efficacy responses observed in the subpopulation and the global population, respectively. For simplicity, assume that both X_i s and Y_j s are normally distributed with known variance σ^2 . When σ^2 is unknown, it can generally be estimated by the sample variance. Thus, Δ can be estimated by

$$\hat{\Delta} = \bar{x} - \bar{y},$$

where $\bar{x} = \sum_{i=1}^n x_i/n$ and $\bar{y} = \sum_{j=1}^N y_j/N$. Under the following mixed prior information for Δ ,

$$\pi = \gamma\pi_1 + (1-\gamma)\pi_2,$$

which is a weighted average of two priors, where $\pi_1 \equiv c$ is a non-informative prior and π_2 is a normal prior with mean θ_0 and variance σ_0^2 , and γ is defined as the weight with $0 \leq \gamma \leq 1$. For the consistency test between a specific region and the global area, the choice of (θ_0, σ_0^2) can be derived from the values of $\hat{\Delta}$ from all regions except the specific region. Thus, the marginal density of $\hat{\Delta}$ is given by

$$m(\hat{\Delta}) = \gamma + (1-\gamma) \frac{1}{\sqrt{2\pi(\sigma_0^2 + \tilde{\sigma}^2)}} \exp\left\{-\frac{(\hat{\Delta} - \theta_0)^2}{2(\sigma_0^2 + \tilde{\sigma}^2)}\right\},$$

where $\tilde{\sigma}^2 = \sigma^2\left(\frac{1}{n} + \frac{1}{N}\right)$. Given the clinical data and prior distribution, the posterior distribution of Δ is

$$m(\Delta|\hat{\Delta}) = \frac{1}{m(\hat{\Delta})} \left\{ \gamma \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \exp\left[-\frac{(\Delta - \hat{\Delta})^2}{2\tilde{\sigma}^2}\right] + (1-\gamma) \frac{1}{2\pi\sigma_0\tilde{\sigma}} \exp\left[-\frac{(\Delta - \theta_0)^2}{2\sigma_0^2} - \frac{(\Delta - \hat{\Delta})^2}{2\tilde{\sigma}^2}\right] \right\}.$$

Thus, given the data and prior information, similarity (consistency) on efficacy in terms of a positive treatment effect for the subpopulation can be concluded if the posterior probability of consistency (similarity) is

$$p_c = P(\mu_1 - \mu_2 > 0 | \text{clinical data and prior}) \\ = \int_{-\infty}^0 \pi(\Delta | \hat{\Delta}) d\Delta > 1 - \alpha.$$

For some pre-specified $0 < \alpha < 0.5$. In practice, α is determined such that the subpopulation is generally smaller than 0.2 to ensure that the posterior probability of consistency is at least 80%. Besides, the R-R choice can be compared with p_c .

Based on the discussion given in the work of Chow and Hsiao,^[7] the marginal density of $\hat{\Delta}$ can be re-expressed as

$$m(\hat{\Delta}) = \gamma + (1 - \gamma) \frac{1}{\sqrt{2\pi(\sigma_0^2 + 2\sigma^2/N)}} \\ \exp \left\{ -\frac{(\hat{\Delta} - \theta_0)^2}{2\left(\sigma_0^2 + \frac{2\sigma^2}{N}\right)} \right\}.$$

As a result, the posterior distribution of Δ is therefore given by

$$m(\Delta | \hat{\Delta}) = \frac{1}{m(\hat{\Delta})} \left\{ \gamma \frac{1}{\sqrt{4\pi\sigma^2/N}} \exp \left[-\frac{(\Delta - \hat{\Delta})^2}{4\sigma^2/N} \right] + (1 - \gamma) \frac{1}{\sqrt{8\pi\sigma_0^2\sigma^2/N}} \exp \left[-\frac{(\Delta - \theta_0)^2}{2\sigma_0^2} - \frac{(\Delta - \hat{\Delta})^2}{4\sigma^2/N} \right] \right\}.$$

Consequently, we have

$$p_c = \int_{-\infty}^0 \pi(\Delta | \hat{\Delta}) d\Delta > 1 - \alpha.$$

Japanese Approach

In a multiregional trial, to establish the consistency of treatment effects between the Japanese population and the entire population under study, the Ministry of Health, Labour and Welfare (MHLW)^[1] of Japan suggested evaluating the relative treatment effect (i.e., treatment effect observed in Japanese population compared to that of the entire population) to determine whether the Japanese population has achieved a desired assurance probability for consistency.

Let μ be the treatment effect of the test treatment under investigation. Denote by μ_1 and μ_2 the treatment effect of the Japanese population and the treatment effect of the entire population (all regions), respectively. Let $\hat{\mu}_1$ and $\hat{\mu}_2$

be the corresponding estimators. For an MRCT, given μ , the sample size ratio (i.e., the size of Japanese population compared to the entire population), and the overall result being is significant at the α level of significance, the assurance probability, p_a , of consistency is defined as

$$p_a = P_\delta \left(\hat{\mu}_1 \leq \frac{1}{\rho} \hat{\mu}_2 | Z \right) z_{1-\alpha} > 1 - \gamma,$$

where Z represents the overall test statistic, ρ is the minimum requirement for claiming consistency, and $0 < \gamma \leq 0.2$ (denote $p_0 = 1 - \gamma$) is a pre-specified desired level of assurance. The Japanese MHLW suggests that ρ be ≥ 0.5 . The determination of ρ , however, should be different from product to product.

In practice, the above required sample size may either not be easy to be achieved or not pre-planned. Instead, consistency may be evaluated using the above idea of the effect of a subpopulation achieving a specified proportion (usually $\geq 50\%$) of the observed overall effect. This can be seen as a simplified version of the Japanese approach.

The Applicability of Those Approaches

Considering the data type and which term the consistency is tested based on, the approaches have different applicability. With respect to the data type, the normal data and

binary data are two widely used types. As for the term of consistency, the following two can be considered: (1) comparing the two effects directly, e.g., comparing $\mu_{T1} - \mu_{R1}$ and $\mu_{T2} - \mu_{R2}$ directly, where (μ_{Ti}, μ_{Ri}) are the treatment mean and the reference mean for the i region, $i = 1, 2$; (2) testing the consistency in terms of the original aim, e.g., testing the consistency between $\{\mu_{T1} - \mu_{R1} < 0\}$ and $\{\mu_{T2} - \mu_{R2} < 0\}$ when the original hypothesis is $H_0: \mu_T - \mu_R \geq 0$ vs. $H_1: \mu_T - \mu_R < 0$. Table 2 shows the applicability of the approaches.

An example of direct comparison and original aim is as follows: assume a global clinical trial was carried out to compare the effect of a new drug and the effect of a reference drug globally and regionally. Denote μ_{GT} , μ_{GR} , μ_{ST} , and μ_{SR} as the means of the new drug globally, the reference drug globally, the new drug for a specific region, and the reference drug for the specific region, respectively. We were concerned with the hypothesis testing $H_0: \mu_{GT} - \mu_{GR} \geq 0$ vs $H_1: \mu_{GT} - \mu_{GR} < 0$, as well as $H_0: \mu_{ST} - \mu_{SR} \geq 0$ vs $H_1: \mu_{ST} - \mu_{SR} < 0$. To test the consistency between the global effect $\mu_{GT} - \mu_{GR}$ and the specific

Table 2 The applicability of the approaches

Approach	Data type		Aim	
	Binary	normal	Compare directly	Original aim
3.1 Shih	✓	✓	✓	×
3.2 Tse	×	✓	✓	×
3.3 Sensitivity Index	×	✓	✓	×
3.4 Specificity reproducibility probability	×	✓	✓	×
3.4 Superiority reproducibility probability	×	✓	✓	×
3.4 Reproducibility probability ratio	×	✓	×	✓
3.5 Bayesian approach	×	✓	✓	✓
3.6 Japanese approach	✓	✓	✓	×

region effect $\mu_{ST} - \mu_{SR}$, we may consider a variety of hypothesis testing. One is just to test $H_0: \mu_{GT} - \mu_{GR} = \mu_{ST} - \mu_{SR}$ vs $H_1: \mu_{GT} - \mu_{GR} \neq \mu_{ST} - \mu_{SR}$, which can be seen as direct comparison and is not directly related to whether $\mu_{ST} - \mu_{SR}$ is smaller than 0. Another one is directly related to $H_0: \mu_{ST} - \mu_{SR} \geq 0$ vs $H_1: \mu_{ST} - \mu_{SR} < 0$, such as the Bayesian method introduced previously which is used to estimate the posterior probability of $\{\mu_{ST} - \mu_{SR} < 0\}$ using both the specific regional data and the global data.

AN EXAMPLE (FOR A NI SITUATION)

Compared with a superiority situation, it is more difficult to evaluate the consistency for NI situations since some methods may not work directly for a NI situation and the concept for consistency is less intuitive. Let us use an event-driven trial with two treatment groups as an example to evaluate the consistency for a NI situation.

We will designate the observed hazard ratio between test drug and control drug as HR_G for the global population and HR_S for a subpopulation. The two-sided type I error rate is 5%. The trial is successful based on the global population: the upper bound of 95% confidence interval (CI) for the $HR_G < 1$ (if it is a superiority trial) or the upper bound of 95% CI for the $HR_G < NI$ margin denoted as Δ (if it is a NI trial). Conditional on the success of the overall trial, the results for the subpopulation analyses could be divided into four categories in Table 3 and interpreted as follows:

- For results in Category 1, it is clear that the subpopulation performed better than the global population. Consistency evaluation is usually not necessary.
- For results in Category 2, it is reasonable to consider the subpopulation results consistent with the global results, since a global result of the same magnitude as the subpopulation result would have given a positive study.
- For results in Category 3, consistency is less clear, since a global result of the same magnitude as the subpopulation result would not have given a positive study.
- For results in Category 4, inconsistency is observed.

Table 3 Four Categories of the Test Results

Category	Superiority trial	NI trial
1	$HR_S < HR_G$ (clearly met superiority)	$HR_S < HR_G$ (clearly met NI)
2	$HR_G < HR_S < 1$ and the upper bound of 95% CI for HR_S based on the overall number of events < 1	$HR_G < HR_S < \Delta$ and the upper bound of 95% CI for HR_S based on the overall number of events $< \Delta$
3	$HR_G < HR_S < 1$ but the upper bound of 95% CI for HR_S based on the overall number of events ≥ 1	$HR_G < HR_S < \Delta$ but the upper bound of 95% CI for HR_S based on the overall number of events $> \Delta$
4	$HR_G < 1 \leq HR_S$	$HR_G < HR_S$ and $HR_S \geq \Delta$

Categories 2 and 3 can be interpreted as answering the question: If the global study point estimate was the same as the observed subpopulation point estimate, would the global study have been positive? Please note that if there are only a few events in a subpopulation, then no meaningful consistency evaluation can be done.

Example (this does not represent the results of a particular clinical trial): A phase 3, active (denoted as drug B) controlled, randomized (1:1 ratio), double-blind, event-driven parallel arm study to evaluate the efficacy and safety of a test drug (denoted as drug A). The primary objective is to demonstrate if the test drug A is non-inferior (NI with a NI margin of 1.40 for the hazard ratio of drug A vs drug B) to the control drug B for a time-to event endpoint. The secondary objectives include superiority tests for the primary endpoint and all cause death. Thirty-eight countries (denoted as C1–C38) from four regions participated in the study. The results for the primary endpoint by country based on the intent-to-treat (ITT) are plotted in Fig. 1 and Table 4 (for selected countries). “NON ESTIMABLE” indicates that the hazard ratio could not be estimated because events did not occur in both treatment arms.

It is clear that no one will question the consistency for Country 18, e.g., since its HR is smaller than that for the

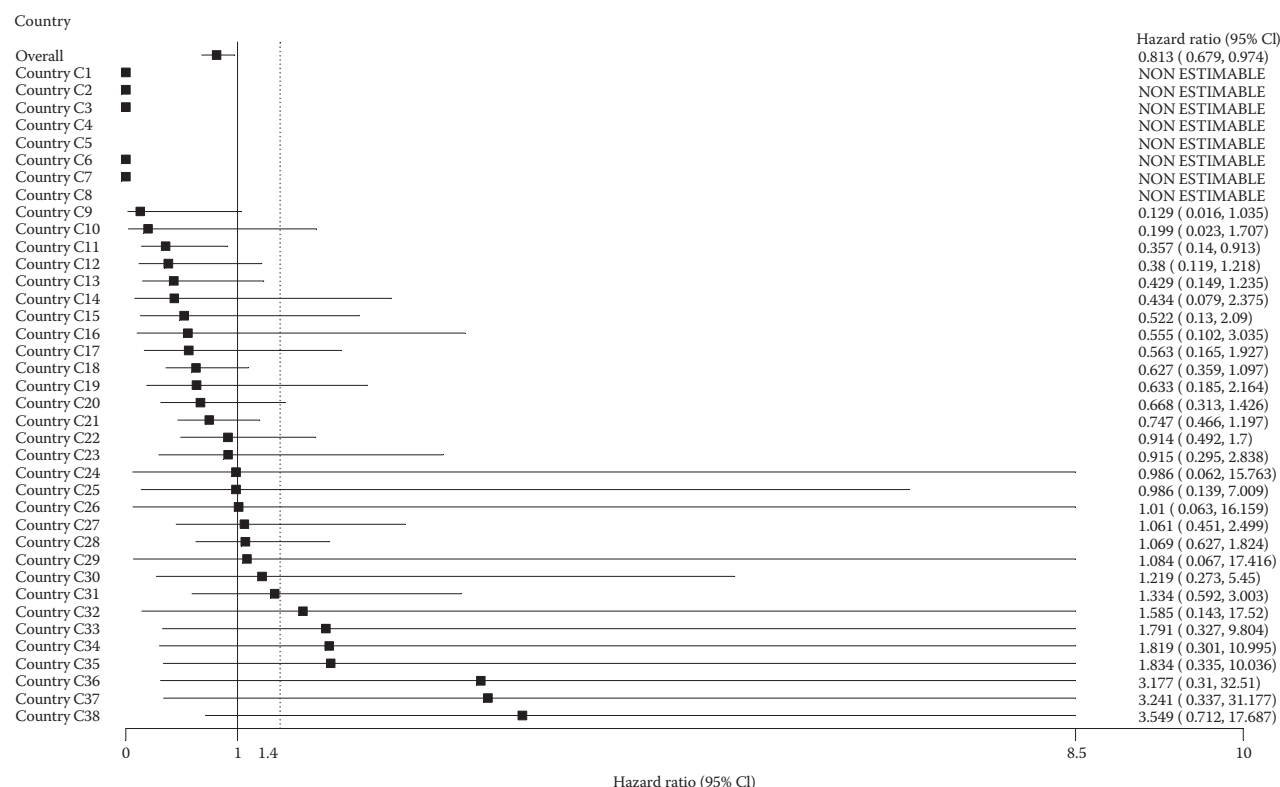


Fig. 1 Adjudicated stroke or systemic embolism during the intended treatment period, overall and by country—randomized subjects

Table 4 Results for the primary efficacy endpoint for selected subpopulations

Population	Total number of subjects, <i>N</i>	Drug A: number of subjects with events (event rate), <i>n</i> (%/year)	Drug B: number of subjects with events (event rate), <i>n</i> (%/year)	HR of drug A/drug B (95% CI)
Global	15,421	183 (1.13)	230 (1.39)	0.81 (0.68, 0.97)
Country C18	1,609	18 (0.96)	29 (1.52)	0.63 (0.36, 1.10)
Country C22	1,255	16 (1.13)	16 (1.24)	0.91 (0.49, 1.70)
Country C28	752	25 (3.47)	23 (3.24)	1.07 (0.63, 1.82)
Country C31	830	12 (1.35)	9 (1.01)	1.33 (0.59, 3.00)
Country C34	289	3 (0.80)	2 (0.44)	1.82 (0.30, 10.99)

global population. Although the HR for Country 22 is less than 1 and the HR for Country 28 is greater than 1, both countries belong to Category 2 since this is a NI trial (hazard ratio should be compared with the NI margin instead of 1). Country 31 belongs to Category 3. Country 34 belongs to Category 4. However, since the number of events isare very small, the results for Country 34 are inconclusive.

Let us use Country 28 and Country 31 as examples to illustrate how to use the statistical methods described in the previous sections to evaluate the consistency. First of all, the HR for Country 28 is far less than the NI margin. If this result had been observed for the global population, then the NI claim would still have been established. The consistency evaluation for Countries 28 and 31 based on

the Bayesian approach described in Section “Bayesian Approach” are listed in Table 5. The results are as expected, i.e., the posterior probabilities of consistency for Country C28 are all >80% (mean=0.90), which indicates consistency. The posterior probabilities of consistency for Country C31 are not all >80% (mean=0.69), which indicates inconsistency may exist.

Some statistical methods such as the method of Tse and Japanese approach that are described in Section “Statistical Methods” for evaluating consistency may not work directly for the NI situations when the subpopulation results and the global result are not on the same side of 1 (for hazard ratio) (or 0 for mean difference), e.g., the hazard ratios for Countries C28 and C31 are >1, but the hazard ratio for the global population is <1. However, if

Table 5 Values of posterior probability (P_c) for various values of weight (γ) for countries C28 and C31

γ	P_c for C28	P_c for C31
0.0	1.00	1.00
0.1	0.98	0.90
0.2	0.96	0.83
0.3	0.94	0.77
0.4	0.92	0.72
0.5	0.91	0.68
0.6	0.89	0.65
0.7	0.88	0.62
0.8	0.86	0.59
0.9	0.85	0.57
1.0	0.84	0.55

we know the HR and its corresponding CI for the control drug vs placebo, then indirect comparison results between the test drug and placebo could be used to evaluate consistency using these methods. For example, the hazard ratio (HR_{BvsP}) and the corresponding 95% CI between the control drug B and placebo from historical trials are 0.38 (0.26, 0.56). Assuming HR_{BvsP} for Countries 28 and 31 is the same as that for the global population, the HRs between drug A and placebo can be derived as 0.31, 0.41, and 0.51 for global, Country C28, and Country 31, respectively. Then Country C28 reserves 85.7% of the global effect (as measured by the relative risk reduction) and Country C31 reserves 71.5% of the global effect. Based on the principle of Japanese guidelines, Both Country 28 and Country 31 results are consistent with the global result. In addition, there is absence of treatment-by-region (the subpopulation X vs the rest population) interactions (the p -values for the interaction test are >0.15) for both Countries C28 and C31.

Therefore, there is no indication that Country C28 is inconsistent with the global population results from a statistical point of view. However, such a conclusion could not be made for Country 31.

OTHER CONSIDERATIONS/DISCUSSIONS

In the previous sections, we have mainly discussed the statistical methods for consistency evaluation. There are other areas for consideration of consistency evaluation. For example, if a numerical inconsistency was observed for the subpopulation data, then one could consider the following from a clinical point of view:

- Consider the severity and subtype of the events, if any, to understand the nature of the difference. For example, consider the types of hemorrhagic strokes and ischemic strokes for the stroke endpoint; cardiovascular (CV) vs non-CV deaths for all-cause death endpoint.

- Conduct sensitivity analysis such as on-treatment analysis vs ITT analysis which may help to clarify the differences. For example, if the on-treatment result for a subpopulation is consistent with the global result but the ITT results are not, then examine the post-treatment medication differences and assess if there is a dose adjustment and stabilization period needed for the post-treatment medication which may explain the difference.

Since the subpopulations are usually not powered in the MRCT, it is likely to observe numerical differences in the results for some subpopulations. It is difficult to identify the causes for the difference between the subpopulation and the global population. Therefore, we suggest that if a numerical difference is observed between a subpopulation of interest and global population, and the sample size is not very small, then perform a complete consistency evaluation (possible difference in population, pharmacokinetic/pharmacodynamics (PK/PD), statistical tools, and clinical aspects) as discussed in this article. If no appropriate causes could be found to explain the numerical difference or to show inconsistency, then the conclusion could be drawn if no inconsistency is found. It is likely due to a random chance, especially when the sample size (or total events) for the subpopulation is small. In such cases, the global results should be used for the regional approval and/or more data could be obtained either prior to or after regional approval.

CONCLUSION

In this article, we have discussed several statistical methods to evaluate consistency between a specific region (i.e., subpopulation) and all regions combined (i.e., global population or entire population). We have also illustrated how to use those methods for NI situations through an example as it is more difficult to evaluate consistency for a NI situation.

Statistical evaluation is only part of the consistency evaluation between a subpopulation and the global population. Possible differences in population, PK/PD, and clinical aspects should also be evaluated.

Since the subpopulations are usually not powered in the MRCT, it is likely to observe numerical differences in the results for some subpopulations. It is difficult to identify the causes for the difference between the subpopulation and the global population, especially when the sample size is small for the subpopulation. Therefore, for evaluation of consistency in clinical results between a specific region and all regions combined, it is suggested that power calculation for sample size should take into consideration the minimum sample size required at the specific region for achieving a desired level of consistency or assurance probability as suggested by MHLW^[1] to ensure consistency in terms of treatment effect between a subpopulation and the entire population.

REFERENCES

1. MHLW. Basic Principles on Global Clinical Trials. Notification No. 0928010, September 28, 2007, Ministry of Health, Labour and Welfare of Japan. Available at <http://www.pmda.go.jp/english/service/pdf/notification/0928010-e.pdf>.
2. Shih, W.J. Clinical trials for drug registration in Asian-Pacific countries: Proposal for a new paradigm from a statistical perspective. *Control. Clin. Trials* **2001**, *22* (4), 357–366.
3. Tse, S.K.; Chang, J.Y.; Su, W.L.; Chow, S.C.; Hsiung, C.; Lu, Q.S. Statistical quality control process for traditional Chinese medicine. *J. Biopharm. Stat.* **2006**, *16* (6), 861–874.
4. Chow, S.C.; Shao, J.; Hu, O.Y.P. Assessing sensitivity and similarity in bridging studies. *J. Biopharm. Stat.* **2002**, *12* (3), 385–400.
5. Shao, J.; Chow, S.C. Reproducibility probability in clinical trials. *Stat. Med.* **2002**, *21* (12), 1727–1742.
6. Hsiao, C.F.; Hsu, Y.Y.; Tsou, H.H.; Liu, J.P. Use of prior information for Bayesian evaluation of bridging studies. *J. Biopharm. Stat.* **2007**, *17* (1), 109–121.
7. Chow, S.C.; Hsiao, C.F. Bridging diversity: Extrapolating foreign data to a new region. *Pharm. Med.* **2010**, *24* (6), 349–362.
8. Liu, J.P.; Chow, S.C.; Hsiao, C.F. *Design and Analysis of Bridging Studies*. Chapman & Hall/CRC Press, Taylor & Francis Group: New York, 2013.
9. Nevius, S.E. Assessment of evidence from a single multicenter trial. *Proceedings of Biopharmaceutical Section of the American Statistical Association*: Alexandria, VA, 1988, 43–45.
10. Enis, P.; Geisser, S. Estimation of the probability that $Y < X$. *J. Am. Stat. Assoc.* **1971**, *66* (333), 162–168.
11. Church, J.D.; Harris, B. The estimation of reliability from stress-strength relationships. *Technometrics* **1970**, *12*, 49–54.
12. Lu, Y.; Kong, Y.Y.; Chow, S.C. Analysis of sensitivity index for assessing generalizability in clinical research. **2016**, Submitted.

Multiregional Clinical Trials: Qualitative Consistency of Treatment Effects

Yoko Tanaka

Eli Lilly and Company, Indianapolis, Indiana, U.S.A.

Hiroshi Nishiyama

Eli Lilly Japan K.K, Kobe, Japan

Abstract

There is a rich literature available on the methods used to estimate sample size for region/country in a multi-regional clinical trial (MRCT) including the guidance from Japan's Ministry of Health, Labour and Welfare (MHLW). While a variety of methods have been studied, two approaches in the MHLW guideline, Methods 1 and 2, are the most commonly used in practice to estimate sample size. When the trial data become available and the consistency of treatment effects across regions is evaluated, the criteria used for regional sample size estimation may not be the only criteria needed to determine consistency or inconsistency. In this entry, the criteria used to estimate regional sample size and to examine consistency are both briefly reviewed. They are then contrasted to identify in what ways they are similar or dissimilar.

Keywords: Heterogeneity assumption; Regulatory and international guidelines; Trend; Planning; analysis.

INTRODUCTION

There is a rich literature available on the methods used to estimate sample size for region/country in a multi-regional clinical trial (MRCT) including the guidance from Japan's Ministry of Health, Labour and Welfare (MHLW). While a variety of methods have been studied, two approaches in the MHLW guideline, Methods 1 and 2, are the most commonly used in practice to estimate sample size. When the trial data become available and the consistency of treatment effects across regions is evaluated, the criteria used for regional sample size estimation may not be the only criteria needed to determine consistency or inconsistency. In this entry, the criteria used to estimate regional sample size and to examine consistency are both briefly reviewed. They are then contrasted to identify in what ways they are similar or dissimilar.

BACKGROUND

Traditionally, clinical trials had been conducted mainly in Western regions such as North America and Western Europe. However, recently clinical trials have become more widely globalized and often include the so-called emerging regions such as Eastern Europe, Latin America, and Asia, which is considered relatively new paradigm in simultaneous drug development. The objective of these

trials, known as multi-regional clinical trials (MRCT), is to show the effectiveness of a new drug and obtain nearly simultaneously regulatory approval in multiple regions by applying the overall trial results in each region. Regulatory guidelines⁽¹⁻³⁾ have been published to provide principles for planning and designing MRCTs.

One of the above regulatory guidelines prepared by MHLW, "Basic Principles on Global Clinical trials"⁽¹⁾ was published when the drug development paradigm in Japan shifted from the stand-alone clinical trials to MRCT with Japan as one of the participating countries to synchronize the submission timing with that of the first country to file for regulatory approval, typically the US. To guide the design of MRCTs which include Japan, the MHLW guideline⁽¹⁾ describes fundamental principles for global clinical trials including two methods to estimate Japanese sample size (Method 1 and 2).

Method 1 estimates the number of subjects in the target region to achieve a specified probability (e.g., 80%) of a consistent treatment effect by taking the ratio of that region's treatment effect to the overall treatment effect compared with a predetermined threshold (e.g., 0.5). Method 2 is a way to determine the number of subjects for each region to achieve a specified probability so that all treatment effects across regions show a similar tendency (e.g., at least favor in study drug compared to placebo).

These two methods triggered research in estimating sample sizes for Japan (and other countries), and currently

there is an abundant literature available for sample size estimation for regions in a multi-regional clinical trial. While a variety of methods have been studied, Methods 1 and 2 are the most commonly used in practice to estimate the sample size for the region of interest. The draft ICH-E17 guideline⁽³⁾ on MRCTs, which became available for public comment in July 2016, includes examples for sample size allocation to regions using Methods 1 and 2 (see the section titled, ICH-E17 Draft Guidance 2016).

When the trial is completed and the data reviewed, the consistency of treatment effects across regions is evaluated. In some cases the criteria used to determine regional sample size may not be the criteria needed to conclude whether treatment effects are consistent or inconsistent across regions. In evaluating this question of consistency, there is a certain common theme that can be delineated from various different sources such as regulatory guidelines, ICH-E17 draft guidance, scientific presentations, and past examples of evaluating consistency.

In this entry, current methods researched to estimate regional sample size will be briefly reviewed to highlight the context of the criteria. These will be compared with the criteria used to evaluate consistency of treatment effects across regions to show that these consistency criteria are not necessarily the same as the criteria used to estimate regional sample size.

CRITERIA FOR SAMPLE SIZE FOR REGION/COUNTRY

Design considerations for MRCTs, including sample size for region/country is well described in the details in the book, *Multiregional Clinical Trials for Simultaneous Global New Drug Development*, edited by Chen and Quan⁽⁴⁾ (MRCT book). In this section, some concepts of consistency which are used to estimate regional sample size in the clinical trial design stage are summarized briefly. The relationship between regional sample size and the probability for demonstrating consistency can be followed, and the details of derivations will be presented for the first criteria, and can be found in the references cited for other criteria.

Assumptions and Notations

- X_{ij} and Y_{ij} : the control and experimental treatment group endpoint values for the j th patient within region i with normal distributions:

$$X_{ij} \sim N(\mu_{ix}, \sigma_i^2) \text{ and } Y_{ij} \sim N(\mu_{iy}, \sigma_i^2), \text{ for } i = 1, \dots, s.$$

- s : number of region
- Variances are the same across groups and regions, $\sigma_1^2 = \sigma_2^2 = \dots = \sigma^2$

- N_i : the number of patients per treatment group for the i th region (1:1 treatment ratio).
- N : the total number of patients in each treatment group, $N = \sum_{i=1}^s N_i$
- f_i : the fraction of patients from region i , $f_i = \frac{N_i}{N}$
- δ_i : true treatment effect within region i , $\delta_i = \mu_{iy} - \mu_{ix}$, positive δ_i indicates better outcome for experimental group
- $\hat{\delta}_i$: estimated true treatment effect δ_i within region i , $\hat{\delta}_i = \hat{Y}_i - \hat{X}_i \sim N\left(\delta_i, \frac{2\sigma^2}{N_i}\right)$
- δ : overall treatment effect based on patients from all regions, $\delta = \sum_{i=1}^s f_i \delta_i$
- $\hat{\delta}$: a common estimate of an overall treatment effect, $\hat{\delta} = \frac{\sum_{i=1}^s N_i \hat{\delta}_i}{N} \sim N\left(\delta, \frac{2\sigma^2}{N}\right)$
- u_i : ratio of the treatment effect for region i to the overall treatment effect, $u_i = \delta_i / \delta$, and $\sum_{i=1}^s f_i u_i = 1$
- z_α : the $(1 - \alpha)$ quantile of the standard normal distribution.

Region as Fixed-Effects

Quan, Li et al.,⁽⁵⁾ classified into 5 ways to define consistency as shown below.

A Specified Proportion of the Observed Overall Effect

This criterion is similar to that used in Method 1, which estimates the number of subjects in the target region to achieve the specified probability to obtain an consistent treatment effect by taking the ratio of that region's treatment effect $\hat{\delta}_i$ to the overall treatment effect $\hat{\delta}$ compared with a predetermined threshold π (e.g. $\pi \geq 0.5$).

$$\hat{\delta}_i / \hat{\delta} > \pi.$$

Uesaka⁽⁶⁾ showed how to determine the regional sample size to achieve the required probability that would simultaneously ensure both the consistency criterion when there is a specific region of interest R_s and a significant overall treatment effect.

$$\begin{aligned} \gamma &= \Pr\left(\hat{\delta} \geq z_\alpha \sigma \sqrt{2/N}, \hat{\delta}_{R_s} > \pi \hat{\delta}\right) \\ &= \int_{z_\alpha \sigma \sqrt{2/N}}^{\infty} \frac{1}{2\sigma \sqrt{\pi/N}} \exp\left(-\frac{(x - \delta)^2}{4\sigma^2/N}\right) \Phi(-y_x) dx \end{aligned}$$

where the joint distribution of $(\hat{\delta}, \hat{\delta}_{R_s})'$ is

$$N \left(\begin{pmatrix} \delta \\ \delta_{R_S} \end{pmatrix}, 2\sigma^2 / N \begin{bmatrix} 1 & 1 \\ 1 & 1/f_{R_S} \end{bmatrix} \right),$$

Φ is the cumulative standard normal distribution function, and

$$y_x = \frac{\delta - \delta_{R_S} + (\pi - 1)x}{\sigma \sqrt{2/N f_{R_S} (1 - f_{R_S})}} \lambda.$$

The conditional probability given there is a significant overall treatment effect ^(5,7) for determining the sample size in a region of interest is

$$\Pr(\hat{\delta}_i > \pi\hat{\delta}, 1 \leq i \leq s | \hat{\delta} \geq z_\alpha \sigma \sqrt{2/N}; \delta_i, 1 \leq i \leq s).$$

Li, Quan et al.,⁽⁸⁾ developed R-code to compute the conditional probability when it is replaced with

$$\frac{\Pr(Z_i > 0, 1 \leq i \leq s; Z_{s+1} > 0 | \delta_i, 1 \leq i \leq s)}{1 - \Phi\left(z_\alpha - \delta / \sigma \sqrt{2/N}\right)}$$

where

$$Z_i = (1 - \pi f_i) \hat{\delta}_i - \pi \sum_{j \neq i} f_j \hat{\delta}_j, 1 \leq i \leq s \text{ and } Z_{s+1} = \hat{\delta} - \sigma \sqrt{2/N}.$$

The joint distribution of $(Z_1, \dots, Z_{s+1})'$ is a multivariate normal random vector with mean

$$\delta \begin{pmatrix} u_1 - \pi \\ u_2 - \pi \\ \vdots \\ u_s - \pi \\ 1 - z_\alpha \sigma \sqrt{2/N} / \delta \end{pmatrix}$$

and covariance matrix

$$\frac{2\sigma^2}{N} \begin{pmatrix} (\pi^2 - 2\pi + 1/f_1) & (\pi^2 - 2\pi) & \cdots & (\pi^2 - 2\pi) & (1 - \pi) \\ (\pi^2 - 2\pi) & (\pi^2 - 2\pi + 1/f_2) & \cdots & (\pi^2 - 2\pi) & (1 - \pi) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ (\pi^2 - 2\pi) & (\pi^2 - 2\pi) & \cdots & (\pi^2 - 2\pi + 1/f_s) & (1 - \pi) \\ (1 - \pi) & (1 - \pi) & \cdots & (1 - \pi) & 1 \end{pmatrix}$$

Without repeating similar details, the rest of the criteria for regional sample size estimation are listed to highlight the concept.

Region Effect Exceeds a Prespecified Size

This criterion is similar to that used in Method 2, which estimates the number of subjects for each region to achieve a specified probability that all treatment effects across regions show a similar tendency (e.g., at least, results favor the study drug compared with placebo).

$$\hat{\delta}_1 > b, \hat{\delta}_2 > b, \dots \text{and } \hat{\delta}_s > b, \text{ where } b \geq 0.$$

Kawai, Chuang-Stein et al.,⁽⁹⁾ showed that the probability to ensure the criteria above depends on the treatment allocation balance within each region, treatment effect size, total sample size, number of regions, and allocation of patients among regions. Unconditional probabilities to ensure the criteria and the conditional probabilities to ensure the criteria given a significant overall treatment effect were evaluated in case of common true treatment effects across regions, i.e., $\delta_1 = \delta_2 = \dots = \delta_s = \delta$.

Region Effects Exceed a Proportion of the Overall Effect Using Hypothesis Testing

Another way to assess consistency of regional effect is to use a hypothesis testing framework similar to Liu and Chow⁽¹⁰⁾. Consistency is considered if H_0 is rejected and H_1 is accepted.

$$H_0 : \delta_1 \leq \pi\delta \text{ or } \dots \text{or } \delta_s \leq \pi\delta \text{ versus, } H_1 : \delta_1 > \pi\delta \text{ and } \dots \text{and } \delta_s > \pi\delta$$

Absence of Significant Treatment-by-Region Interaction

This popular approach to assess treatment-by-subgroup interaction is also applicable to test inconsistency of regional effect when H_0 is rejected and H_1 is accepted.

$$H_0 : \delta_1 = \delta_2 = \dots = \delta_s = \delta \text{ versus, } H_1 : \delta_i, \\ i = 1 \dots s, \text{ are not all the same.}$$

This can be tested using Cochran's Q type statistic test statistic^(4,5) derived from the same manner as in the meta-analysis.⁽¹¹⁾

Lack of Significant Difference for Any Regions from the Overall

This is a one-sided version of previous criteria. If H_{0i} is rejected and H_{1i} is accepted the region i effect is not considered consistent with the overall effect, and therefore consistency across all regions has not been shown.

$$H_{0i} : \delta_i \geq \delta \text{ versus } H_{1i} : \delta_i < \delta, i = 1, \dots, s.$$

ICH-E17 Draft Guidance 2016

The ICH guideline on MRCTs has been developed, and the draft guideline⁽³⁾ became available for public comment in July 2016. The ICH-E17 draft guideline⁽³⁾ lists five examples for sample size allocation to regions.

- i. show similar trends in treatment effects across regions
- ii. determine the regional sample size to show that the regional treatment effect meets a pre-specified proportion of the overall treatment effect
- iii. determine the regional sample size in proportion to region size and disease prevalence without requiring a prespecified allocation for regions
- iv. determine the regional sample sizes to achieve significant results within a region
- v. require a specific number of subjects.

These approaches are not entirely new and have been introduced or discussed in guidelines or at professional/scientific forums. Approaches i. and ii. share concepts similar to those of Japanese Method 2 and Method 1, respectively. Approach iii. was discussed at the ICH scientific and regulatory workshop in 2015⁽¹²⁾ held by the MRCT center of Brigham and Women's hospital, and at Harvard and Peking University. Approach iv. is the most conservative approach, which may be required for stand-alone local trials. Finally, approach v. is sometimes accepted as a feasible number of subjects when recruitment is considered difficult due to prevalence and site availability for the particular disease in a certain region.

Region as Random-Effects

In MRCTs, a random-effects model is introduced to estimate robust total sample size^(13,14). The random effects model treats true regional treatment effects as random variables. δ_i 's are assumed to have variability, exchangeable, and follow the same distribution:

$$\hat{\delta}_i | \delta_i \sim N\left(\delta_i, \frac{2\sigma^2}{N_i}\right) \text{ and } \delta_i \sim N(\delta, \tau^2)$$

where

δ : overall population mean

τ^2 : between region variability.

Or unconditionally

$$\hat{\delta}_i \sim N\left(\delta, \frac{\tau^2 + 2\sigma^2}{N_i}\right).$$

The observed overall sample treatment effect from a random-effect model follows a distribution

$$\hat{\delta} = \sum_{i=1}^s f_i \hat{\delta}_i \sim N\left[\delta, \sigma^2 \left(\frac{2}{N} + \left(\frac{\tau}{\sigma} \right)^2 \sum_{i=1}^s f_i^2 \right)\right].$$

The regional sample size can then be obtained from the following formula of the total sample size per treatment group with the fraction of patients for region i , $f_i = N_i/N$.

$$N_R = \left[\frac{1}{2} \left(\left(\frac{\delta}{\sigma(z_\alpha + z_\beta)} \right)^2 - \left(\frac{\tau}{\sigma} \right)^2 \sum_{i=1}^s f_i^2 \right) \right]^{-1}.$$

If $\tau^2 = 0$ is assumed, then the resulting total sample size per treatment group

$$N_F = \frac{2\sigma^2(z_\alpha + z_\beta)^2}{\delta^2}.$$

Quan, Li et al.,⁽¹⁴⁾ proposed an empirical shrinkage estimator based on the random-effects model for individual regions. They showed the probabilities for simultaneously showing significance of the overall treatment effect and consistency of treatment effect determined based on the criteria of Japanese Method 1 or Method 2 by using a simple random-effects model or an empirical shrinkage estimator, respectively. Information from the other regions is used, with the shrinkage estimator being a weighted average of the estimators of the regional treatment effect and the overall treatment effect. Quan, Mao et al.,⁽¹⁵⁾ proposed the James–Stein type shrinkage estimator associate with smaller variability compared with the regular sample estimator. The relationship between the value of τ^2 and the sample size for a random-effects model can be found in Quan, Mao et al.,⁽¹⁵⁾. They showed the number of regions should not be too small to obtain a desirable type I error.

To allow between-region variability in the model, the overall sample size using the random effects model is typically larger than that of the fixed effects model.

CONSISTENCY EXAMINATION OF TREATMENT EFFECTS ACROSS REGIONS

In examining consistency of treatment effects across regions, a typical approach is to test treatment-by-region interaction as in the general subgroup examination. Chen, Quan et. al.,⁽¹⁶⁾ provided a systematic review of the existing methods to assess consistency of treatment effects across regions and classified the methods into global methods (single statistic), quantitative methods, and qualitative methods. Quantitative methods assess variation in the magnitude but not in the direction of the treatment effects among regions, and qualitative methods assess the opposite direction of treatment effects among regions. We reviewed recommendations on how to examine consistency of treatment effects across regions from multiple sources (ICH-E5, ICH-E17 draft, EMA draft guideline on subgroup, publication, presentation). A point to notice is that the criteria used for calculating the regional sample size in the planning stage may not be the determining criteria to conclude consistency in the data interpretation stage. The EMA draft guideline on the investigation of subgroups⁽¹⁷⁾ (2014) states that, “There is no widely accepted definition for consistency”. This might be due to an essential difficulty in pre-determining one dimensional measurement to judge consistency. While many considerations for the bridging criteria were proposed after ICH-E5 guideline⁽²⁾ was introduced (1998), no established method has been realized. The tenth question and answer in the [appendix D](#) (in Japanese) for the ICH-E5 states that *it’s impossible to suggest a concrete, always applicable standard to demonstrate that the data is “similar” or “not significantly different”*. This discussion could similarly apply to the consistency criteria for MRCTs. Uyama⁽¹⁸⁾ (2012) presented a review of PMDA’s evaluations of NDAs submitted in Japan based on MRCTs, including 28 submissions between 2006 and 2012. These discussions indicated that successful bridging and consistency in MRCT are based on a comprehensive, careful review of the data both on efficacy and safety. This could be rephrased as “consensus”.

Here are the main points we identified in examining consistency of treatment effects across regions from different sources.

At a series of scientific forums held to discuss different perspectives on MRCTs, O’Neil⁽¹⁹⁾ recommended the following points to be considered when addressing issues of treatment consistency:

- Any examination of inconsistency should consider
 - Multiplicity issues
 - Baseline characteristics, medical practice, and other intrinsic/extrinsic factors that may confound the results
- Totality of data/evidence
 - Overwhelming vs marginal overall effect
 - Consistency in other important endpoints and subgroups
 - External data (same class, same patient population, etc.).

From the Japanese regulatory review on MRCT data for new drug applications, Uyama⁽¹⁸⁾ recommended

- Necessity of a sufficient number of Japanese subjects for appropriate evaluation. When the consistency is not observed, important to distinguish whether it was due to either small Japanese sample size or true ethnic difference.
- To consider the variability (e.g., SD) as well as a point estimate when taking account of small regional sample size
- To confirm the consistency of secondary endpoint(s) and safety
- When inconsistency is observed, to carefully evaluate whether it is possible to utilize the MRCT results for evidence of efficacy and safety in Japanese population by investigating the cause using subpopulation analyses.

Recommendations from regulatory guidelines include the following:

Section 2.2.7 in the ICH-E17 draft guideline⁽³⁾ (four examples for consistency examination) recommends

“In the statistical analysis plan, to include a strategy for evaluating consistency of treatment effects across regions, and for evaluating how any observed differences across regions may be explained by intrinsic and/or extrinsic factors using approaches such as

1. Descriptive summaries
2. Graphical displays (e.g., forest plots, funnel plots)
3. Model-based estimation including covariate-adjusted analysis
4. Test of treatment by region interaction.”

Section 6.1 in the EMA draft guideline⁽¹⁷⁾ has the following recommendations on using subgroups to assess consistency or inconsistency

- Absence of statistical significance in an interaction test for treatment-by-covariate or directional consistency (i.e., all points estimates to be going in the same direction) are by themselves not sufficient to imply consistency
- Full consideration of other important factors in assessing inconsistency
- Visual inspection of a Forest plot to flag subgroup for further examination
- Lack of a formal rule to detect heterogeneity of potential interest.

There remains the challenge to resolve conflicting perspectives so that regional sample size allocation can be determined 1) to provide sufficient information for an evaluation of the extent to which the overall treatment effect applies to subjects from different regions in each regional

context which may increase total and regional numbers, and 2) to plan feasible total and regional numbers of subjects without compromising the reason for conducting MRCT. As ICH-E17 is expected to be finalized in the near future, it would be helpful guideline to remember that examining consistency is a collective, comprehensive review of all of the evidence. Current draft guideline lists approach of the examination of consistency without providing a sense of the totality of evidence, something that matters more than just the primary endpoint. This narrow approach could mislead the reader into thinking that examination of consistency should be performed in a quantitative manner, when in fact it has an essential qualitative nature.

CONCLUSION

Various criteria used to estimate regional sample size in a MRCT and consistency examination of treatment effects across regions were reviewed. In looking for the criteria for consistency examination of treatment effects across regions, various different sources such as regulatory guidelines, ICH-E17 draft guidance, scientific presentations, and past examples of evaluating consistency were evaluated. Common set of criteria identified from these various sources do not appear to be limited to any single dimension criteria used for regional sample size allocations, but rather to the totality of evidence from the trial data.

REFERENCES

1. *Basic Principles on Global Clinical Trials*, Notification No.0928010; Ministry of Health, Labour and Welfare: Tokyo, 2007.
2. *ICH-E5 Guideline on Ethnic Factors in the Acceptability of Foreign Clinical Data*, The International Conference on Harmonisation (ICH): Geneva, 1998.
3. *ICH-E17 Draft Guideline on General Principles for Planning and Design of Multi-Regional Clinical Trials*, The International Council for Harmonisation (ICH): Geneva, 2016.
4. Chen, J.; Quan, H. *Multiregional Clinical Trials for Simultaneous Global New Drug Development*, Chapman and Hall/CRC: Boca Raton, FL, 2016; Section II, Chapters 4, 7, 10, 11.
5. Quan H.; Li M.; Chen J.; Gallo P.; Binkowitz B.; Ibia E.; Tanaka Y.; Ouyang SP.; Luo X.; Li G.; Menjoge S.; Talerico S.; Ikeda K. Assessment of Consistency of Treatment Effects in Multiregional Clinical Trials. *Drug information Journal* **2010**, *44* (5), 617–632.
6. Uesaka, H. Sample Size Allocation to Regions in a Multiregional Trial. *Journal of Biopharmaceutical Statistics* **2009**, *19* (4), 580–594.
7. Sekiguchi, R.; Ogawa, S.; Uesaka, H. Sample Size Determination of Japanese Patients for Multiregional Clinical Trial in Oncology. In *Proceedings of Joint Statistical Meeting 2007*, 1044–1046, Salt Lake City.
8. Li, M.; Quan, H.; Chen, J.; Tanaka, Y.; Ouyang, S.; Luo, X.; Li, G. R Functions for Sample Size and Probability Calculations for Assessing Consistency of Treatment Effects in Multi-Regional Clinical Trials. *Journal of Statistical Software* **2012**, *47*.
9. Kawai, N.; Chuang-Stein, C.; Komiyama, O.; Li, Y. An Approach to Rationalize Partitioning Sample Size into Individual Regions in a Multiregional Trial. *Drug information Journal* **2008**, *42* (2), 139–147.
10. Liu, JP.; Chow, SC. Bridging Studies in Clinical Development. *J Biopharm Stat* **2002**, *12*, 359–367.
11. Cochran, WG. The combination of estimates from different experiments. *Biometrics* **1954**, *10*, 101–129.
12. ICH scientific and regulatory workshop 2015: <http://mrctcenter.org/news/mrct-center-holds-regulatory-and-scientific-workshop-at-peking-university/#more-614>.
13. Hung, HMJ.; Wang, SJ.; O'Neill, RT. Consideration of Regional Difference in Design and Analysis of Multi-Regional Trials. *Pharmaceutical Statistics* **2010**, *9*, 173–178.
14. Quan H.; Li M.; Shih WJ.; Ouyang SP.; Chen J.; Zhang J.; Zhao PL. Empirical Shrinkage Estimator for Consistency Assessment of Treatment Effects in Multi-Regional Clinical Trials. *Statistics in Medicine* **2013**, *32*, 1691–1706.
15. Quan H.; Mao X.; Chen J.; Shih WJ.; Ouyang SP.; Zhang J.; Zhao PL.; Binkowitz B. Multi-Regional Clinical Trial Design and Consistency Assessment of Treatment Effects. *Statistics in Medicine* **2014**, *33*, 2191–2205.
16. Chen J.; Quan H.; Binkowitz B.; Ouyang SP.; Tanaka Y.; Li, G.; Menjoge, S.; Ibia, E. Assessing Consistent Treatment Effect in a Multi-Regional Clinical Trial: A Systematic Review. *Pharm Stat* **2010**, *9*, 242–253.
17. Guideline on the Investigation of Subgroup in Confirmatory Clinical Trials (Draft), European Medicines Agency: London, 2014.
18. Uyama, Y. Current Status of Multi-Regional Clinical Trial and Evaluation of Oversea's Data in PMDA (original title in Japanese), Society for Regulatory Science of Medical Products, Tokyo, Japan, Sept. 2–3, 2012. (Available at: <https://www.pmda.go.jp/files/000163966.pdf>) [Only in Japanese, Accessed date: Nov. 12, 2016].
19. O'Neill, R. A Regulatory Perspective on MRCT's and Potential Strategies to Synergize Initiatives, MRCT Center Annual Meeting, Cambridge, MA, Nov. 28, 2012. (Available at: <http://mrctcenter.org/wp-content/uploads/2016/02/2012-11-28-MRCT-Center-Annual-Meeting-Slides.pdf>) [Accessed date: Nov 12, 2016].

Multivariate Meta-Analysis

Elizabeth Stojanovski

School of Mathematics and Physical Sciences, University of Newcastle, Callaghan, New South Wales, Australia

Kerrie L. Mengersen

School of Mathematical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia

Abstract

This entry discusses different sources of multivariate data, the steps involved in meta-analysis, and models in the multivariate meta-analysis framework.

INTRODUCTION

Meta-analysis is the joint statistical analysis of results from a number of related studies to obtain an overall perspective on an effect or outcome of interest. There are often multiple dimensions to the same outcome and so it is common for outcomes to be determined by at least two measures. Studies typically report multiple estimates of the same effect. Moreover, the trend toward exhaustive decision analysis to assess health technology demands examination of multiple outcomes. There are numerous forms of multivariate effect size data, which can arise in several ways, including as a comparison of one or more treatments or interventions with a control group on at least one outcome, as multiple outcomes per study, or as the measurement of an outcome at several time points. Multivariate methods in meta-analysis account for the dependence between related outcomes and are generally preferred to univariate methods that assess each outcome separately as they typically provide more information about outcomes of interest, by borrowing strength across effects both within and across studies.

The chapter begins with a discussion of different sources of multivariate data. The steps involved in any meta-analysis are discussed, followed by a description of common univariate approaches for analysis of multivariate data in meta-analysis. Models in the multivariate meta-analysis framework are then reviewed.

SOURCES OF MULTIVARIATE DATA

Meta-analysis has become a standard statistical technique to evaluate single or multiple outcomes or to evaluate sources of variation in study results.

A meta-analysis of any form entails a range of decisions to be made at each of several steps, including

identification of relevant outcomes from studies, choice of suitable response variables to measure this outcome and how to estimate these from studies, characterization of the model relating study-specific and global covariates to these responses, identification of fixed effects and random effects, and the description of within- and between-study heterogeneity. These choices then assist with forming the specific meta-analysis model, method of estimation, and the choice of univariate or multivariate methods.

Univariate meta-analysis is usually condoned on the grounds of there being a single ideal parameter that describes the response of interest. The simplest type of meta-analysis in the univariate setting comprises calculating a weighted mean of estimates as an aggregate effect size estimate. This can be generalized to incorporate random effects to permit heterogeneity across studies, meta-regressions, and generalized linear mixed models to examine associations between effect sizes and variables that define the individual studies. These have been established in both frequentist^[1,2] and Bayesian^[3–6] frameworks.

In practice, a single parameter may not be justified to assess an outcome of interest. Often reported in studies are multiple estimates of the same effect or different dimensions of the same outcome. Moreover, the current move toward comprehensive decision analysis for assessing health technology^[7] demands examination of multiple endpoints, which provides further motivation for multivariate meta-analysis.

There are numerous forms of multivariate effect size data. A common form of dependency emerges when the studies under examination compare a particular treatment with a particular control on several outcomes. Dependency is incurred because of all outcomes being measured on each subject, which creates correlated effect sizes because of their derivation from the correlated outcome measures. This can arise, for example, when comparing cholesterol and blood pressure measures between a group treated with cholesterol

lowering medication and a group treated with a placebo. Multivariate data may also emerge from the dependence between groups for the situation comparing a particular treatment with a particular control or from a comparison of several treatment groups on a single outcome.

Multivariate data may arise when a common control is compared with at least two treatments for a particular outcome in the compared studies, consequently generating several treatment effects. The common control group in all comparisons infers dependency via the correlated effect size estimates. Similarly, multivariate data may emerge when numerous treatments are each compared with a common control in terms of their effectiveness but not directly to each other. Such mixed comparisons involve decisions concerning indirect relationships.

Multivariate data similarly emerges if an outcome is measured at several points in time. This would produce correlated scores on subjects as well as dependent control and treatment comparisons based on these scores. Dependency would also arise in a situation that an individual study assesses several independent samples with related measures, or if separate studies with similar methods or samples are compared.

Multivariate methods in meta-analysis account for the dependence between related outcomes and are generally preferred to univariate methods that assess each outcome separately.^[8,9] These methods are expected to provide more information about outcomes, by borrowing strength across effects within a study as well as across studies.

STEPS IN A META-ANALYSIS

Any meta-analysis, whether univariate or multivariate, frequentist or Bayesian involves several steps, each of which requires one or more decisions to be made:

1. Design the meta-analysis.
 - a. Identify the aim of the meta-analysis. Is it simply an aggregate point or interval estimate? Are subgroup estimates of interest? Is it important to identify sources of variation, differences in study quality, extreme events, or study rankings? Is the analysis a foundation for further studies or an end in itself? The answer to these questions will impact on all of the following decisions.
 - b. Define the outcomes of interest and corresponding covariates, confounders, and other information relevant to the meta-analysis.
 - c. Construct the meta-analysis model. The model will depend on theoretical assumptions, past results, and exploratory data analyses. It will characterize the relationships between multiple outcomes, describe within- and between-study heterogeneity, and allow for sources of uncertainty, bias, and so on.
2. Collect the data.
 - a. Define the criteria by which studies will be included in the meta-analysis. These can include accessibility, peer reviewed status, the type of information reported, the standard of bias control, and so on.
 - b. Identify the sources used to collect relevant studies. These may include databases, citations, Internet searches, personal requests, and so on.
 - c. Decide on the strategy to deal with missing data. Often the reported information is not consistently available in the required form and some imputation, calculation, or other accommodation needs to be made.
 - d. Collect the relevant studies and extract the relevant information from each study. This is often easier said than done! As the data extracted forms the foundation for the subsequent meta-analysis, it is vitally important to clearly document all such activities.
3. Undertake the analysis.
 - a. Estimate correlations between outcomes and other required measures if these are not reported.
 - b. Undertake the meta-analysis and interpret the results.
 - c. Check the inferences via a sensitivity analysis of the model inputs, and of the model itself.
 - d. As part of the sensitivity analysis, assess the impact of publication bias and other important biases and confounders.

Note that the analysis includes the crucial critical interrogation of the model inputs and outputs, and of the model structure itself. The construction of the model and the corresponding methods of analysis form the main focus of the discussion in the following sections. Before proceeding, however, we discuss two important aspects of all meta-analyses.

Fixed or Random Effects

An important task for a meta-analysis is identification of fixed effects and random effects, particularly with respect to between-study variation. This will then guide the decision between fixed-, random- or mixed-effects models. The decision can be based on empirical considerations, including heterogeneity tests of data to be analyzed or previous results from similar analyses. Alternatively, it can be justified through theoretical or structural descriptions of the variation sources in the data, or it can be purely philosophical. While some argue a fixed-effects model on the basis that overall estimates from heterogeneous studies are uninterpretable,^[10] others argue that between-study variation is a reality in almost all meta-analyses and should be taken into account in the model.^[1,11]

A common set of population effects is assumed to underlie every study for a fixed-effects model, so that each effect estimate from a study is considered a random draw from this common distribution, and all observed variation is a result of study-specific sampling variation alone. A random-effects model, in comparison, assumes that a different set of population effects underlie each study, so that a study estimate is a random draw from a study-specific population, and these population parameters are themselves drawn from an overall population with a global mean (effect) of interest. Thus the observed variation is because of sampling error within studies as well as differences among studies in design, sampling and implementation methods, demographics, treatment-specific information, and so on.

If particular between-study differences can be identified, both fixed-effects and random-effects models can be extended to include regression coefficients that relate these covariates to the study outcomes. These meta-regression models are classified as fixed when all between-study differences are explained by this set of covariates, and as random or mixed when between-study variation remains following the inclusion of predictors. Many researchers have highlighted the scientific and clinical importance of meta-regressions,^[11] especially for multivariate meta-analysis.^[12,13]

Multivariate random-effects models were found to be more realistic than the corresponding fixed-effects models in a simulation study conducted by Berkey et al.^[14] because of a failure of the fixed-effects model to explain all between-study variability. Standard errors of regression coefficients for the fixed-effects parameters were consistently underestimated, hence underestimating widths of corresponding confidence intervals and producing incorrect inferences as the null hypothesis was too often rejected.

Hence the models described are presented according to whether they comprise fixed effects or random effects. Some multivariate models apply to studies with multiple outcomes or treatments but do not generalize to both cases while other methods, which although designed for one situation, easily generalize to the other.

Correlation Estimation

Most multivariate techniques necessitate estimated correlations between outcomes, which are not typically reported in studies.^[15] A study by Chiu (Chiu, C.W. Synthesizing multiple outcome measures with missing data: a sensitivity analysis on the effect of metacognitive reading intervention. Department of Counseling, Educational Psychology, and Special Education, Michigan State University, East Lansing, MI, 1997, unpublished manuscript) reported that less than 10% of multivariate studies that assessed the effect of metacognitive therapy instruction on reading ability reported intercorrelations between the outcomes of interest.

Unreported correlations between effects can be estimated from observed data. Gleser and Olkin^[16] provide relevant formulas for multiple treatment and multiple endpoint studies separately. Variance and covariance estimates can be derived using statistical packages that perform matrix calculations such as Proc IML in SAS, SPSS, R, and so on. Numerous imputation methods are also available for missing correlations (see Ref. [17] and references therein), such as substituting the average of correlations reported in other studies. Finally, correlations among outcomes in a particular study in the meta-analysis can be utilized; Berkey, Anderson, and Hoaglin^[18] claim that correlations from one study are often closer to those unreported for the remaining studies than to zero, which is the inferred correlation when outcomes are examined separately. Alternatively, unreported correlations can be estimated from sources external to the study such as authors, published manuscripts, or funding sources.^[19] It is imperative that positive definiteness of the estimated or imputed correlation matrix be confirmed before proceeding.

UNIVARIATE APPROACHES TO MULTIVARIATE META-ANALYSIS

Meta-Analysis Based on Separate Outcomes

Multiple outcomes are often assessed through separate univariate meta-analyses.^[20] This approach may generate sensible results if the outcomes are independent or weakly correlated, or if such comparisons are not of interest.^[10,21] However, if outcomes are highly related, this approach may lead to biased estimates of the Type I error rate and standard errors.^[8,10,22,23] The type and extent of this bias depends on the type and strength of dependency between outcomes. For example, in a Monte Carlo study reported by Sohn,^[15] the effect of ignoring potential relationships between estimated effects of a study was not significant, whereas the magnitude of the variation term resulting from random errors of the multivariate effects was substantial.

Meta-Analysis Based on Summary Outcomes

It is common in practice to ignore dependence within studies by creating a summary effect size estimate per study by combining related effect sizes per study and combining this with values from other studies using methods of meta analysis for independent data.^[10,20,24] Methods of summarizing effect sizes for each study include pooling outcomes to create composite measures, or pooling raw data to create summary statistics such as averages across outcomes for each study, or choosing an individual ideal or most common outcome.^[19]

A test for homogeneous effect sizes in a study is proposed by Hedges and Olkin.^[10] The relevant test statistic is assumed to follow a chi-squared distribution under

homogeneity, supporting the adequacy of a common effect size. Optimal weights to generate pooled estimates are also presented. Gleser and Olkin^[16] present univariate approaches and composite effect sizes to combine effect sizes for both multiple treatment and multiple endpoint scenarios.

The benefit of a combined estimate is that an overall probability statement regarding treatment differences is provided. This may be reasonable when the sample size is large and the outcome measures are representative of the same underlying construct, so that the effect direction of a particular treatment is the same for all outcomes. This may not be the situation, for example, if one treatment is expected to be better for one outcome than for another.^[18] Moreover, the ability to assess effect differences for different constructs is lost when outcomes are combined or when some constructs are omitted from the data set. If estimates are highly correlated, however, pooling estimates are only slightly superior to selecting a single estimate.^[10] As a separate treatment may have differing impacts on various outcomes, it might be misleading to average effects for each study or to choose a single outcome and ignore the remaining outcomes.

Meta-Analysis Based on Data Subsets

Multiple treatments are often considered. Consider a placebo that is compared to each of two treatments. If the objective of the study is to establish the superior treatment, the trials provide only indirect information, and are referred to as mixed comparisons. Discarding data from some treatment groups is often performed in practice^[25] rather than using multivariate models for such data.

Another univariate method forms subsets of data that are independent, and individual analyses are conducted on each independent outcome. The same problems as discussed above arise because comparisons between data subsets are impacted by correlations in the original data.

The impact of dependency on meta-analysis results can be assessed by a sensitivity analysis,^[26] which involves analyses both with the inclusion and with the exclusion of multiple outcomes in each study. A rational approach comprises analysis of the initial data set with one outcome per study followed by repetitions of the analysis with additions of extra outcomes. If results are similar, the analysis involving one outcome can be reported, along with details of the sensitivity analysis. Otherwise, a multivariate approach that accounts for dependence is needed.

MULTIVARIATE APPROACHES

Notation

Effect sizes or correlations are often the measures combined in a meta-analysis. The effect size is commonly a

measure of the difference between two treatments (or a treatment and a control) and is often measured in terms of the standardized or unstandardized mean difference for continuous data, or the odds ratio, relative risk, or rate difference for binary outcomes. The vector of effects for the i th study is denoted by $\theta_i = (\theta_{i1}, \dots, \theta_{ip})$, and the corresponding vector of estimated effects by $Y_i = (Y_{i1}, \dots, Y_{ip})$, based on $n_i = (n_{i1}, \dots, n_{ip})$ observations and $i = (1, \dots, k)$.

The notation is equivalent for studies with multiple time points, with Y_{ij} here denoting measures on occasion $j = 1, \dots, p$ instead of a measure of outcome j in study i . Outcomes from different studies may have different interpretations if timing of occasions differs between studies. Consider a case in which three effects are presented for each outcome, in particular pretreatment, one month post-treatment, and six months posttreatment measures. Effect interpretations at posttreatment may depend on treatment length and timing of posttreatment measures.

The study-specific effect estimates are often assumed multivariate normal (MVN) based on the assumed distribution of the corresponding population effect (or an appropriate transformation) or on the central limit theorem, so that $Y_i \sim \text{MVN}(\theta_i, \Sigma_i)$, where Σ is the variance-covariance matrix. Gleser and Olkin^[16] derived formulas for effect sizes and variance-covariance matrices under assumptions of both homogeneity and heterogeneity for multiple-endpoint studies.

Frequentist Fixed-Effects Approaches

An early fixed-effects multivariate meta-analysis method was proposed by Hedges and Olkin^[10] in which each study was assumed to measure all outcomes. This model was extended to a meta-regression by Raudenbush, Becker, and Kalaian^[12] for data from multiple-endpoint studies. Further extensions have been proposed for survival data with measurements taken at multiple time points,^[13] data from multiple treatment studies,^[16] and data from studies that simultaneously consider multiple endpoints and multiple treatments.^[18] These approaches are briefly reviewed in this section.

The well-known ordinary least-squares (OLS) regression estimation of parameters assumes constant error variance across observations for independent data, such as for univariate data from different studies in a meta-analysis. When several correlated effect sizes are taken on each subject, heteroscedasticity arises, and the OLS estimator is still unbiased but inefficient as it is no longer the minimum variance estimator. An alternative efficient estimation method is the generalized least-squares (GLS) approach, which is of the form $Y = XB + e$, where $\text{var}(e_i) = \sigma^2(x_i)$ is not constant, and is transformed by dividing all terms by $\sqrt{x_i}$. The error term now has constant variance satisfying OLS assumptions. GLS approaches to meta-analysis have been discussed by Dear,^[13] Kalaian and Raudenbush,^[27] and Timm,^[28] among others. Multivariate

meta-analysis methods that utilize GLS with covariates have also been proposed, for example by Raudenbush, Becker, and Kalaian,^[12] Becker,^[29] and Hedges and Olkin.^[10]

Building on earlier models by Hedges and Olkin^[10] and Rosenthal,^[24] Raudenbush, Becker, and Kalaian^[12] present a fixed-effects multiple regression model for analyzing multiple continuous outcomes. The model is $Y = X\beta$, where X is a design matrix, and the regression coefficients, β , represent common effect sizes across studies, and any error in the fitted model is assumed due to sampling variation alone and is assumed normally distributed around zero. Benefits of the method include incorporating associations between outcomes in each study via study-level variance-covariance matrices that can differ between studies and adjustment for covariates at study and treatment level that may explain variation in effect sizes. For each study, the model additionally accommodates different combinations of outcomes and covariates, hence enabling effects of various treatments on different outcomes to be assessed. One stacked response variable is used for each meta-analysis, thus obviating missing values and the necessity to discard or impute these values as is typically performed with classical multivariate methods in the case that studies do not report all outcomes. Disadvantages include the need for correlations among outcomes, which typically need to be estimated. Additionally, frequentist hypothesis tests are based on the assumption of normally distributed effect sizes; so moderately large sample sizes are required to ensure this.

The model of Raudenbush, Becker, and Kalaian^[12] is extended by Dear^[13] to data at multiple time points, in particular survival proportions in a follow-up study, after adjusting for covariates at study and treatment level. This procedure involves iterative modeling of the fitted values to generate correlations among times within studies. The method also enables differing treatments per subject and trials with only one treatment or control group, which act as non-randomized historical controls.^[13] As an example, a study based on three years of data for a particular treatment and a year of data for an alternative treatment provides four outcome components in the form of survival proportions for the proposed meta-analysis.^[13] The method is still based on fixed-effects, so cannot adequately account for between-study heterogeneity. Dear^[13] recommends a logistic model using generalized estimating equations as a potential resolution.

The meta-analysis method proposed by Raudenbush, Becker, and Kalaian^[12] is used by Gleser and Olkin^[16] to model multiple continuous endpoints. The model, like the earlier method, enables differing numbers of effect sizes (from multiple treatments or endpoints) to be combined between studies. Gleser and Olkin^[30] extended this method to allow at least one treatment to be compared to a common control when information for each comparison is reported in a contingency table. Here, effect sizes are computed as

proportion differences or log odds ratios. As with Gleser and Olkin,^[16] different subsets of treatments may be utilized in different studies, creating models with correlated effect sizes and missing data. The authors suggest that effect size estimates can be obtained by weighted least-squares regression. The associated models can hence be fit using most standard statistical software packages. Alternative random-effects and Bayesian hierarchical approaches have also been proposed; see the following.

Berkey, Anderson, and Hoaglin^[18] present a fixed-effects meta-regression model to collate studies that each report multiple continuous outcomes and/or multiple treatments, allowing for study- and treatment-level covariates for one or more outcomes. As with the above methods, correlations between outcomes are required, each study can report some or all of the outcomes, and may consider some or all treatments. Comparisons that do not involve a common control group are allowed and, similar to Ref. [13], studies that comprise one treatment or control group can be included, thus allowing use of additional data. It differs from previously described methods as outcomes and differences between treatments are measured in original units instead of in terms of effect sizes, thus simplifying interpretations and allowing a wider range of applications. As this method does not apply to trials reporting only effect size data, a study reporting, for example, the proportion of patients experiencing a particular illness^[13] could not be considered.

Frequentist Random-Effects Models

A random-effects model allows potential associations among population effects and covariates to be examined. The fixed-effects GLS models are extended to include random-effects representing variation among studies among population effects that are not accounted for by the model's predictor(s). A regression model that includes covariates as fixed-effects and a random-effects error term is also referred to as a mixed-effects meta-regression, so that the general model becomes $Y = XB + u + e$, where u is the random effect.

For example, the model that combines multiple continuous endpoints across studies proposed by Gleser and Olkin^[16] has been extended to the random-effects framework by Berkey, Anderson, and Hoaglin,^[18] Kalaian and Raudenbush,^[27] Higgins and Whitehead,^[31] and Berkey et al.^[14] A similar model for surrogate markers is used by Daniels and Hughes.^[32]

Van Houwelingen, Arends, and Stijnen^[33] argue that modeling two treatments separately in a multivariate framework is more informative than the alternative of combining the treatment outcomes in a summary outcome such as a log odds ratio and combining these in the meta-analysis. Favorable comparisons are also made with DerSimonian and Laird's random-effects model^[1] in terms of distributional assumptions and trial weightings. Van Houwelingen,

Arends, and Stijnen^[33] demonstrate that this model permits baseline risk examination or control group subjects events risk as a source of heterogeneity in treatment effects across trials. This may assist, for example, with the prognosis of a subject exposed to a control treatment.

As an illustration, a multivariate random-effects model comprising two outcomes might be depicted as follows

$$\begin{pmatrix} Y^A \\ Y^B \end{pmatrix} \sim N \left(\begin{pmatrix} \theta^A \\ \theta^B \end{pmatrix}, (\Sigma + C_i) \right)$$

where Y denotes the two study-specific outcomes, Σ , the between-study covariance matrix, and C_i , the within-study diagonal covariance matrix as treatments are applied to different subjects and assumed uncorrelated. For the case that treatment difference measures are made on multiple outcomes, this is similar to the model adopted by Berkey et al.,^[14] with Y comprising a difference between the measure of two treatments on outcome A with the corresponding outcome B measure. The two outcomes are measured on the same subject and hence the covariance matrix, C_i , is now non-diagonal. Covariates can be added to this model as follows

$$\begin{pmatrix} Y^A \\ Y^B \end{pmatrix} \sim N(BX_i, \Sigma + C_i)$$

This representation can be extended easily to involve more outcomes or treatment groups.

A similar meta-analysis model was developed by McIntosh^[34] based on a two-level bivariate hierarchical regression model for treatment effects, which uses the observed rates in the control group as covariates. These models can be estimated with both maximum likelihood via the expectation-maximization algorithm and Bayesian estimation via Markov Chain Monte Carlo algorithms.^[34] Arends, Vokó, and Stijnen^[35] propose a tri-variate random-effects model, providing the first example of combining more than two outcomes. The model accommodates covariates, independent, or correlated outcomes within trials and missing outcomes.

A more general mixed model was developed by DuMouchel^[36] for the meta-analysis of dose-response estimation. This paper and later ones by the same author allow studies comprising multiple outcomes or multiple treatments to be combined, accommodating both binary and continuous outcomes, where all outcomes in a model are required to be on a common scale. This method differs to those described above as correlations between outcomes for each study are not required, because these are estimated by the model. This model extends the random-effects method of DerSimonian and Laird,^[1] enabling a combination of heterogeneous study designs. Unlike the method proposed by Gleser and Olkin^[16] for multiple treatments, no treatment or control is required to be common to

all studies. Different treatment groups can comprise results from separate groups of subjects, or from groups that are subject to several treatments. Van Houwelingen, Arends, and Stijnen^[33] also adopted a mixed model extension to their 1993 approach^[37] to investigate if the covariates at study level are related to the baseline risk via multivariate regression.

Bayesian Models

The Bayesian framework provides a flexible means to model construction via hierarchies representing different relationships between the data and parameters, and different degrees and levels of uncertainty. It permits data to borrow strength from sources of evidence across studies and outcomes, and from other sources of information such as expert opinion or other studies.^[38] Such models allow more natural inference through the posterior distributions, particularly if expectations other than the mean are of interest.^[39] For example, ranking of studies is straightforward in the Bayesian framework but much more difficult under a frequentist approach.^[38]

Dominici et al.^[40] propose a hierarchical Bayesian group random-effects model and report an applied complex meta-analysis of multiple treatments and multiple outcomes. This analysis evaluates 18 different treatments for migraine headache relief, which are grouped within three classes. Multiple heterogeneous outcomes (including continuous effect sizes, differences between pairs of continuous, and dichotomous outcomes) are reported, and multiple treatments are incorporated via relationships between different classes of treatments.

Nam, Mengersen, and Garthwaite^[41] detail a Bayesian approach to the same problem considered by Raudenbush, Becker, and Kalaian^[12] and Berkey et al.,^[14,18] focusing on multiple outcomes after adjusting for study-level covariates. Modifications and reasonable assumptions are suggested to simplify computations and results, and to conform to the type of prior distributions that are able to be handled using the WinBUGS software package (<http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml>). In particular, WinBUGS currently only handles conjugate multivariate distributions, so the commonly used non-conjugate Wishart requires approximations for inclusion in analysis. Various constraints may also be imposed on correlations to ensure positive definiteness of the corresponding variance covariance matrix.

Three Bayesian multivariate models are proposed and compared by Nam, Mengersen, and Garthwaite.^[41] The preferred model is a full hierarchical representation of the form $Y_i \sim \text{MVN}(\theta_i, \Sigma_i)$; $\theta_i \sim \text{MVN}(XB, \Gamma)$ for $i = 1, \dots, k$, where Y_i and θ_i are p -dimensional vectors of outcomes. Assuming results of different studies to be independent, this model simplifies to a univariate Bayesian model for dependent studies. An alternative model considered is $Y_j \sim \text{MVN}(\theta_j, \Sigma_j)$, $j = 1, \dots, p$, where Y_j and θ_j are

k -dimensional vectors of outcomes across the k studies. The variance–covariance matrix Σ_j has dimension equal to the number of studies, which can be large, and information may not be available to assist with estimation of the matrix elements. Moreover, the model is less able to adjust for estimate differences from studies reporting all outcomes and those reporting a subset of outcomes. The third proposed model is a Bayesian adaptation of the mixed model for multicenter trials that allow more than one summary statistic per outcome in each study.

A similar Bayesian meta-analysis model to analyze the same outcome measured at two times is extended by Abrams et al.^[42] to investigate potential associations between risk at baseline and relative risk. For an application presented to assess management of patients who have acute coronary syndrome in terms of an outcome (mortality) as measured at two follow-ups after a treatment was applied, a proportion of trials reported the outcome six months after treatment, and most reported it 30 days after treatment. Opinions from coronary care consultants were obtained regarding their beliefs on mortality at 30 days and six months after treatment, in an attempt to extend on the limited data available.

General Bayesian mixed model approaches for meta-analysis have been promoted and illustrated by DuMouchel, with particular focus on combining studies of interest to the pharmaceutical sciences.^[4]

Other Frameworks

The above hierarchical (frequentist random-effects and Bayesian) models are often described in a multilevel modeling (MLM) framework.^[27] Analysis is typically performed using restricted maximum likelihood estimation (MLE)^[43] or GLS regression as described above. MLMs in meta-analysis are described as comprising two stages: the within-study or measurement model that predicts effect sizes from each study (relating estimated effect sizes from each study to the true effect sizes) and the study-level model that predicts effects across studies. The model extends to multivariate meta-analysis using other outcomes such as correlations or summary effects such as odds ratios, and to studies with multiple treatments.^[27]

Becker^[29,44] describes methods of combining correlations. Inter-relationships among variables can be analyzed by fixed-effects or random-effects models. A set of linear models is derived from the correlation matrix that describes both direct and indirect relationships between outcomes and explanatory variables. Inferences are made on the coefficients of the resultant standardized regression equations.

The definition of “multivariate” can also be extended to include multiple endpoints from dose–response studies. In this case, each study provides a series of dose-specific relative risks compared to a common reference group. Again, fixed-effects and random-effects models can be constructed.^[45]

CONCLUSIONS

There are a range of methods to select from when dealing with a meta-analysis for multivariate data. It is important that relationships between outcomes be acknowledged either in the modeling procedure or in a sensitivity analysis with the intention to quantify the effect for the situation that it is excluded from the model. A number of factors will determine the method to apply including the number, importance of in addressing the objective of the meta-analysis, and degree of similarity in outcomes. Univariate meta-analysis methods can be used when few studies report multiple outcomes or when the multivariate component is not of importance. If the multivariate nature of the data is a major component of the meta-analysis, associated multivariate methods in meta-analysis that account for the dependence between outcomes should be adapted. Correlations between outcomes, which are typically not reported in studies, must be estimated or imputed from alternative sources. The corresponding multivariate method has advantages in terms of providing additional information about relevant outcomes by borrowing strength across effects both within and between studies. Decisions must additionally be made in the multivariate framework regarding identification of fixed and random effects, describing within- and between-study heterogeneity, and choosing between the frequentist or the more flexible Bayesian framework within which to model these features.

REFERENCES

1. DerSimonian, R.; Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* **1986**, *7*, 177–188.
2. Platt, R.W.; Leroux, B.G.; Breslow, N. Generalized linear mixed models for meta-analysis. *Stat. Med.* **1999**, *18*, 643–654.
3. DuMouchel, W.H.; Harris, J.E. Bayes methods for combining the results of cancer studies in humans and other species. *J. Am. Stat. Assoc.* **1983**, *78*, 293–308.
4. DuMouchel, W. Bayesian meta-analysis. In *Statistical Methodology in the Pharmaceutical Industry*; Berry, D.A., Ed.; Dekker: New York, 1990, 509–529.
5. Carlin, J.B. Meta-analysis for 2×2 tables: a Bayesian approach. *Stat. Med.* **1992**, *11*, 141–158.
6. Smith, T.C.; Spiegelhalter, D.J.; Thomas, S.L. Bayesian approaches to random-effects meta-analysis: a comparative study. *Stat. Med.* **1995**, *14*, 2685–2699.
7. Cooper, N.; Abrams, K.; Sutton, A.; Turner, D.; Lambert, P. A Bayesian approach to Markov modelling in cost-effectiveness analyses: application to taxane use in advanced breast cancer. *J. Roy. Stat. Soc. (A)* **2003**, *166*, 389–405.
8. Ades, A.E. A chain of evidence with mixed comparisons: models for multi-parameter synthesis and consistency of evidence. *Stat. Med.* **2003**, *22*, 2995–3016.
9. Becker, J.B. Multivariate meta-analysis. In *Handbook of Applied Multivariate Statistics and Mathematical Modelling*; Tinsley, H.E.A.; Brown, S., Eds.; Academic Press: San Diego, CA, 2000.

10. Hedges, L.V.; Olkin, I. *Statistical Methods for Meta-analysis*; Academic Press: Orlando, FL, 1985.
11. Berkey, C.S.; Hoaglin, D.C.; Mosteller, F.; Colditz, G.A. A random-effects regression model for meta-analysis. *Stat. Med.* **1995**, *14*, 395–411.
12. Raudenbush, S.W.; Becker, B.J.; Kalaian, H. Modeling multivariate effect sizes. *Psychol. Bull.* **1988**, *103*, 111–120.
13. Dear, K.B. Iterative generalized least squares for meta-analysis of survival data at multiple times. *Biometrics* **1994**, *50*, 989–1002.
14. Berkey, C.S.; Hoaglin, D.C.; Antczak-Bouckoms, A.; Mosteller, F.; Colditz, G.A. Meta-analysis of multiple outcomes by regression with random effects. *Stat. Med.* **1998**, *17*, 2537–2550.
15. Sohn, S.Y. Multivariate meta-analysis with potentially correlated marketing study results. *Naval Res. Logistics* **2000**, *47*, 500–510.
16. Gleser, L.J.; Olkin, I. Stochastically dependent effect sizes. In *The Handbook of Research Synthesis*; Cooper, H.; Hedges, L.V., Eds.; Russell Sage Foundations: New York, 1994, 339–355.
17. Little, R.J.; Rubin, D.B. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu. Rev. Public Health* **2000**, *21*, 121–145.
18. Berkey, C.S.; Anderson, J.J.; Hoaglin, D.C. Multiple-outcome meta-analysis of clinical trials. *Stat. Med.* **1996**, *15*, 537–557.
19. Becker, B.J. Multivariate meta-analysis. In *Handbook of Applied and Multivariate Statistics and Mathematical Modeling*; Tinsley, H.E.A.; Brown, S., Eds.; Academic Press: San Diego, CA, 2000.
20. Rosenthal, R.; Rubin, D.B. Interpersonal expectancy effects: the first 345 studies. *Behav. Brain Sci.* **1978**, *3*, 377–386.
21. Rosenthal, R.; Rubin, D.B. Comparing effect sizes of independent studies. *Psychol. Bull.* **1982**, *92*, 500–504.
22. Walsh, J.E. Concerning the effect of the intraclass correlation on certain significance tests. *Ann. Math. Stat.* **1947**, *18*, 88–96.
23. Holmes, C.T.; Mathews, K.M. The effects of non-promotion on elementary and junior high school pupils: a meta-analysis. *Rev. Educ. Res.* **1984**, *54*, 225–236.
24. Rosenthal, R. Meta-analytic procedures and the nature of replication: the ganzfeld debate. *J. Parapsychol.* **1986**, *50*, 315–336.
25. Song, F.; Altman, D.G.; Glenny, A.; Deeks, J.J. Validity of indirect comparisons for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *Br. Med. J.* **2003**, *326*, 472–484.
26. Greenhouse, J.B.; Iyengar, S. Sensitivity analysis and diagnostics. In *Handbook of Research Synthesis*; Cooper, H.; Hedges, L., Eds.; Sage: New York, 1993; 383–398.
27. Kalaian, H.A.; Raudenbush, S.W. A multivariate mixed linear model for meta-analysis. *Psychol. Methods* **1996**, *1*, 227–235.
28. Timm, N.H. Testing multivariate effect sizes in multiple-endpoint studies. *Multivar. Behav. Res.* **1999**, *34*, 132–145.
29. Becker, B.J. Using results from replicated studies to estimate linear models. *J. Educ. Behav. Stat.* **1992**, *17*, 341–362.
30. Gleser, L.J.; Olkin, I. Meta-analysis for 2×2 tables with multiple treatment groups. In *Meta-analysis in Medicine and Health Policy*; Stangl, D.K.; Berry, D.A., Eds.; Marcel Dekker: New York, 2000.
31. Higgins, J.P.T.; Whitehead, J. Borrowing strength from external trials in a meta-analysis. *Stat. Med.* **1996**, *15*, 2733–2749.
32. Daniels, M.J.; Hughes, M.D. Meta-analysis for the evaluation of potential surrogate markers. *Stat. Med.* **1997**, *16*, 1965–1982.
33. Van Houwelingen, H.C.; Arends, L.R.; Stijnen, T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat. Med.* **2002**, *21*, 589–624.
34. McIntosh, M.W. The population risk as an explanatory variable in research synthesis of clinical trials. *Stat. Med.* **1996**, *15*, 1713–1728.
35. Arends, L.R.; Voko, Z.; Stijnen, T. Combining multiple outcome measures in a meta-analysis: an application. *Stat. Med.* **2003**, *22*, 1335–1353.
36. DuMouchel, W. Meta-analysis for dose-response models. *Stat. Med.* **1995**, *14*, 679–685.
37. Van Houwelingen, H.C.; Koos, H.Z. A bivariate approach to meta-analysis. *Stat. Med.* **1993**, *12*, 2273–2284.
38. Spiegelhalter, D.J.; Best, N.G. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Stat. Med.* **2003**, *22*, 3687–3709.
39. Gelman, A.; Carlin, J.G.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*; Chapman and Hall: London, 1995.
40. Dominici, F.; Parmigiani, G.; Wolpert, R.L.; Hasselblad, V. Meta-analysis of migraine headache treatments: combining information from heterogeneous designs. *J. Am. Stat. Assoc.* **1999**, *94*, 16–28.
41. Nam, I.; Mengersen, K.; Garthwaite, P. Multivariate meta-analysis. *Stat. Med.* **2003**, *22*, 2309–2333.
42. Abrams, K.R.; Sutton, A.J.; Cooper, N.J.; Sculpher, M.; Palmer, S.; Ginley, L.; Robinson, M. Populating economic decision models using meta-analyses of heterogeneously reported studies augmented with expert beliefs Developing economic evaluation methods (DEEMS). 4 Workshop, Oxford, 2003.
43. Frost, C.; Clarke, R.; Beacon, H. Use of hierarchical models for meta-analysis: experience in the metabolic ward studies of diet and blood cholesterol. *Stat. Med.* **1999**, *18*, 1657–1676.
44. Becker, B.J. Corrections to “Using results from replicated studies to estimate linear model”. *J. Educ. Behav. Stat.* **1995**, *20*, 100–102.
45. Berlin, J.A.; Santanna, J.; Schmid, C.H.; Szczech, L.A.; Feldman, H.I. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect. *Stat. Med.* **2002**, *21*, 371–387.

Noninferiority Trials: Type I Error Rates

Seung-Ho Kang

Department of Applied Statistics, Yonsei University, Seoul, South Korea

Abstract

The purpose of this entry is to help people avoid the ambiguity of type I error rates in non-inferiority trials that compare a new treatment to an active control without a placebo arm.

Keywords: Within-trial type I error; Across-trial type I error; Within-trial power; Across-trial power.

INTRODUCTION

Non-inferiority trials that compare a new treatment to an active control without a placebo arm are popular. These types of trials are mainly classified into two categories, depending on their objectives.^[1] In the first category, the objective is to compare the effectiveness of two treatments whose efficacy has already been proven in past trials. In the second category, the objective is to gain approval for a new treatment from a regulatory body by comparing the new treatment to an active control without a placebo arm for ethical reasons. This entry concentrates on the second objective.

REVIEW OF STATISTICAL METHODS FOR AN ANALYSIS OF NON-INFERIORITY CLINICAL TRIALS

Let $X_{T,i}$ and $X_{C,i}$ denote the continuous primary end points from the test treatment and the active control treatment, respectively, in the current non-inferiority clinical trial, with $i = 1, 2, \dots, n$. Let $X_{CH,i}$ and $X_{PH,i}$ represent the continuous primary end points from the active control treatment and placebo product, respectively, in the historical trial, with $i = 1, 2, \dots, n_H$. It is assumed that the sample sizes (n, n_H) are large enough to employ the Central Limit Theorem. It is also assumed that:

$$\begin{aligned} E(X_{T,i}) &= \mu_T, E(X_{C,i}) = \mu_C, \text{Var}(X_{T,i}) \\ &= \text{Var}(X_{C,i}) = \sigma^2 < \infty \end{aligned} \quad (1)$$

and

$$\begin{aligned} E(X_{CH,i}) &= \mu_{CH}, E(X_{PH,i}) = \mu_{PH}, \text{Var}(X_{CH,i}) \\ &= \text{Var}(X_{PH,i}) = \sigma_H^2 < \infty \end{aligned} \quad (2)$$

The Fixed-Margin Method (the Two Confidence Interval Method with a Fixed Margin)

Given that the two methods described in this section and the following section are both occasionally referred to as “the fixed-margin method;” we need different names here to distinguish the two methods presented in the two sections. In order to avoid confusion. We will refer to the fixed-margin method introduced in section and the following section as the two confidence interval (TCI) method with a fixed margin, as the non-inferiority margin is treated as a constant.^[2] This method is clearly described by Hung, Wang, and O'Neill.^[3]

Suppose that larger values of the primary end points lead to favorable outcomes. Then, hypotheses of the fixed-margin method (TCI with a fixed margin) are then expressed as follows:

$$H_{01} : \mu_T - \mu_C \leq -\delta \text{ vs } H_{A1} : \mu_T - \mu_C > -\delta \quad (3)$$

where $\delta(>0)$ is a non-inferiority margin. It is recommended that a non-inferiority margin is determined based on the statistical margin and clinical margin.^[4] The statistical margin is often determined as the 95% lower confidence limit for the active control effect from historical trials. Once the margin is determined, it is treated as a fixed constant, although a lower confidence limit may be involved with the determination of a non-inferiority margin.^[3] If the lower confidence limit is considered as a random variable, the hypotheses in Eq. 3 include a random variable, which violates the basic frequentist statistical principle that hypotheses must not have any random variables.^[5] Therefore, the variability of the historical data is not taken into account in this method.

If:

$$(\bar{X}_T - \bar{X}_C) - 1.96\hat{\sigma}_{TC} > -\delta \quad (4)$$

subsequently, we reject the null hypothesis H_{01} in Eq. 3 where $\hat{\sigma}_{TC}$ is an estimate of the standard error of $(\bar{X}_T - \bar{X}_C)$ and:

$$\bar{X}_T = \frac{1}{n} \sum_{i=1}^n X_{T,i}, \bar{X}_C = \frac{1}{n} \sum_{i=1}^n X_{C,i} \quad (5)$$

The Fixed-Margin Method (The Two Confidence Interval Method with a Random Margin)

The method introduced in this section is also sometimes called “the fixed-margin method.”^[6,7] In order to avoid confusion, we refer to the fixed-margin method in this section as the TCI method with a random margin, because the margin is treated as a random variable.^[2] Therefore, this method incorporates the uncertainty associated with the estimation of the non-inferiority margin.

One approach for defining a non-inferiority margin in TCI with a random margin is to make the margin preserve $100\lambda\%$ ($0 \leq \lambda \leq 1$) of the active control effect. Therefore, the hypotheses of interest are expressed as given by:

$$\begin{aligned} H_{02} : (\mu_T - \mu_{P,put}) / (\mu_C - \mu_{P,put}) &\leq \lambda \text{ versus} \\ H_{A2} : (\mu_T - \mu_{P,put}) / (\mu_C - \mu_{P,put}) &> \lambda \end{aligned} \quad (6)$$

where $\mu_{P,put}$ denotes the population mean of the putative placebo in the current non-inferiority trial. It is important to recognize that $\mu_{P,put}$ cannot be estimated because there is no placebo arm in the current non-inferiority trial. If we present the hypotheses in Eq. 6 in the form of the hypotheses in Eq. 3, the hypotheses of the TCI with a random margin are given by:

$$H_{03} : \mu_T - \mu_C \leq -\delta \text{ versus } H_{A3} : \mu_T - \mu_C > -\delta \quad (7)$$

where $\delta = (1 - \lambda)(\mu_C - \mu_{P,put})$ is the non-inferiority margin. The minimal requirement for approval of a new treatment from a regulatory agencies ($\mu_T > \mu_{P,put}$) is obtained from the alternative hypothesis H_{A3} when $\lambda = 0$. The superiority of a new treatment over the active control ($\mu_T > \mu_C$) is obtained when $\lambda = 1$. It is assumed that the effect size of the active control with respect to the placebo does not change from the historical trial to the current non-inferiority trial. This is known as the constancy assumption:

$$\mu_C - \mu_{P,put} = \mu_{C|H} - \mu_{P|H} \quad (8)$$

The non-inferiority margin δ is estimated with the estimate of $(\mu_{C|H} - \mu_{P|H})$ from the historical trials based on the constancy assumption. The lower limit of the 95% confidence interval of $(\mu_{C|H} - \mu_{P|H})$ is often chosen as a conservative estimate of $(\mu_C - \mu_{P,put})$. A meta-analysis is performed when there are several historical trials. The TCI with a random margin method then rejects the null hypothesis H_{03} , if:

$$(\bar{X}_T - \bar{X}_C) - 1.96\hat{\sigma}_{TC} > - (1 - \lambda)(\bar{X}_{C|H} - \bar{X}_{P|H} - 1.96\hat{\sigma}_{PC|H}) \quad (9)$$

where,

$$\bar{X}_{P|H} = \frac{1}{n_H} \sum_{i=1}^{n_H} X_{P|H,i}, \bar{X}_{C|H} = \frac{1}{n_H} \sum_{i=1}^{n_H} X_{C|H,i} \quad (10)$$

and $\hat{\sigma}_{PC|H}$ is the estimate of the standard error of $(\bar{X}_{C|H} - \bar{X}_{P|H})$.

The Synthesis Method

The synthesis method uses hypotheses which are identical to those of the TCI with a random margin in Eq. 6. Based on the constancy assumption in Eq. 8, the hypotheses in Eq. 6 can be rewritten as follows:

$$\begin{aligned} H_{04} : (\mu_T - \mu_C) &\leq (\lambda - 1)(\mu_{C|H} - \mu_{P|H}) \text{ versus} \\ H_{A4} : (\mu_T - \mu_C) &> (\lambda - 1)(\mu_{C|H} - \mu_{P|H}) \end{aligned} \quad (11)$$

Although the current non-inferiority trial and the historical trial are two different clinical trials conducted at different times and different protocols, the synthesis method treats both trials as if they came from the same randomized clinical trial with the same protocol. Therefore, the test statistic is given as follows:

$$T = \frac{(\bar{X}_T - \bar{X}_C) - (\lambda - 1)(\bar{X}_{C|H} - \bar{X}_{P|H})}{\sqrt{\hat{\sigma}_{TC}^2 + (\lambda - 1)^2 \hat{\sigma}_{PC|H}^2}} \quad (12)$$

If $T > 1.96$, the null hypothesis H_{04} in Eq. 11 is rejected. Although there may exist several placebo-controlled historical trials, for the sake of simplicity, we assume here that the historical trials are collectively a single placebo-controlled trial whose hypotheses and test statistics can be expressed as shown below:

$$H_{0,H} : \mu_{C|H} = \mu_{P|H} \quad T_H = \frac{(\bar{X}_{C|H} - \bar{X}_{P|H})}{\sqrt{\hat{\sigma}_{PC|H}^2}} \quad (13)$$

If $|T_H| > 1.96$, the null hypothesis $H_{0,H}$ in Eq. 13 was rejected.

Differences between the Three Methods

Only the data in the current non-inferiority trial is treated as random variables in the fixed-margin method (TCI with a fixed margin) in section “The Fixed Margin Method (the Two Confidence Interval Method with a Fixed Margin).” Although historical data can be used for the determination of the non-inferiority margin, these data are not regarded as random variables.

On the other hand, both data in the current non-inferiority trial and historical data are treated as random variables in

the fixed-margin method (TCI with a random margin) in section “The Fixed-Margin Method (the Two Confidence Interval Method with a Random Margin)” and in the synthesis method in section “The Synthesis Method.” Therefore, all data from the two trials are involved in statistical inference. It is very important to understand which data are treated as random variables and which data are involved in statistical inference, as this is directly related to which type I error rate should be evaluated and controlled.

There is asymptotic equivalence between TCI with random margin and the synthesis method, apart from the critical value. From Eq. 9, we have :

$$\begin{aligned} \bar{X}_T - \bar{X}_C - 1.96\hat{\sigma}_{TC} &> \\ &- (1-\lambda)(\bar{X}_{C|H} - \bar{X}_{P|H} - 1.96\hat{\sigma}_{PC|H}): \\ \Leftrightarrow (\bar{X}_T - \bar{X}_C) + (1-\lambda)(\bar{X}_{C|H} - \bar{X}_{P|H}) &> \\ 1.96[\hat{\sigma}_{TC} + (1-\lambda)\hat{\sigma}_{PC|H}] & \end{aligned} \quad (14)$$

$$\Leftrightarrow T > 1.96\hat{f} \quad (15)$$

where T is the test statistic in the synthesis method in Eq. 12 and $\hat{f} = \frac{\hat{\sigma}_{TC} + (1-\lambda)\hat{\sigma}_{PC|H}}{\sqrt{\hat{\sigma}_{TC}^2 + (1-\lambda)^2\hat{\sigma}_{PC|H}^2}}$.

Furthermore, we have:

$$T \xrightarrow{d} N(0,1) \text{ under } H_{04} \quad \hat{f} \xrightarrow{P} f = \frac{\sigma_{TC} + (1-\lambda)\sigma_{PC|H}}{\sqrt{\sigma_{TC}^2 + (1-\lambda)^2\sigma_{PC|H}^2}} \quad (16)$$

Hence, the TCI with a random margin is asymptotically equivalent to the synthesis method except that the cutoff value 1.96 is replaced with $1.96\hat{f}$. Note that the value of f is greater than 1, which shows that TCI with a random margin in section “The Fixed-Margin Method (the Two Confidence Interval Method With a Random Margin)” is more asymptotically conservative than the synthesis method in the section “The Synthesis Method.”

TYPE I ERROR RATES

The Within-Trial Type I Error Rate

The within-trial type I error rate is calculated with only random variables in the current non-inferiority trials. The non-inferiority margin obtained from the historical trial is considered as a fixed constant. Therefore, the within-trial type I error rate should be used for the fixed-margin method (TCI with a fixed margin) in section “The Fixed Margin Method (the Two Confidence Interval Method with a Fixed Margin).” The specific formula of the within-trial type I error rate for the hypotheses in Eq. 3 is given below:

$$P(\text{reject } H_{01} | H_{01}) = P\left(\frac{\bar{X}_T - \bar{X}_C + \delta}{\hat{\sigma}_{TC}} > 1.96 | H_{01}\right) \quad (17)$$

Here, $\hat{\sigma}_{TC}$ is an estimate of the standard error of $(\bar{X}_T - \bar{X}_C)$. From Eq. 17, it is clear that only the random variables $\bar{X}_T$ and $\bar{X}_C$ in the current non-inferiority trial are utilized in the calculation of the within-trial type I error rates. Within-trial type I error rates were studied by Brittain and Hu,^[8] Hilton,^[9] Sanchez and Chen,^[10] and Sheng and Kim,^[11] Hung et al.,^[1,3,12] Kang and Tsong,^[13] and Wang and Kang^[2] discussed the characteristics of the within-trial type I error rate.

It is necessary to understand the following characteristics of the within-trial type I error rate.

1. If we want to conduct simulation studies that evaluate the within-trial type I error rate, random numbers should be generated only from the current non-inferiority trials. Although the non-inferiority margin can be calculated from the historical trials, such a procedure is not reflected in the simulation. Therefore, we do not need to generate random numbers from the historical trials.
2. As a minimum requirement for the approval of a test treatment by a regulatory body, the non-inferiority trial should demonstrate that the test treatment would be more effective than a placebo, if a placebo was present in the trial.^[4] Accordingly, the hypotheses of such a minimum requirement are expressed as follows:

$$H_{05} : \mu_T \leq \mu_{P,put} \text{ versus } H_{A5} : \mu_T > \mu_{P,put} \quad (18)$$

Note that the hypotheses in Eq. 18 are actually untestable, as there is no placebo group in the current non-inferiority trial. The within-trial type I error rate does not evaluate the type I error rate associated with the hypotheses in Eq. 18, instead assessing only the type I error rate associated with the hypotheses in Eq. 3.

3. If $H_{A1} : \mu_T - \mu_C > -\delta$ in Eq. 3 is accepted and the non-inferiority margin δ is chosen such that:

$$\mu_{C|H} - \mu_{P|H} \geq \delta \quad (19)$$

based on indirect inference, the constancy assumption in Eq. 8 leads to the conclusion that the test treatment is more effective than a putative placebo.

$$\mu_T - \mu_C > -\delta \geq -(\mu_{C|H} - \mu_{P|H}) \text{ from (3) and (19)}$$

$$\Rightarrow (\mu_T - \mu_C) + (\mu_{C|H} - \mu_{P|H}) > 0$$

$$\Leftrightarrow (\mu_T - \mu_C) + (\mu_C - \mu_{P,put}) > 0$$

from constancy assumption in (8)

$$\Leftrightarrow \mu_T - \mu_{P,put} > 0 \quad (20)$$

However, because this conclusion does not have statistical inference, we cannot assess the type I error rate associated with the conclusion from Eq. 20 with

the within-trial type I error rate. Hung, Wang, and O'Neill^[3] write, “This within-trial type I error rate of the non-inferiority trial cannot help quantify the error rate associated with such an indirect inference.”

4. The within-trial type I error rate represents the type I error rate of the falsely concluded non-inferiority in Eq. 3 with the fixed margin when the non-inferiority trial is repeated infinitely many times with an identical fixed margin.^[14] The within-trial type I error rate appears to be the most familiar type I error rate to the statisticians and clinicians involved in clinical trials.
5. When the primary end point is continuous, the well-known sample size calculation formula for the two-arm non-inferiority trial is expressed as shown below:^[15]

$$n_1 = kn_2, n_2 = \frac{(z_{\alpha/2} + z_{\beta})^2 \sigma^2 (1 + 1/k)}{(|\mu_T - \mu_C| - \delta)^2} \quad (21)$$

This formula is obtained from the within-trial power because it considers only the data in the current non-inferiority trial as random variables. Therefore, this formula produces appropriate sample sizes when the fixed-margin method (TCI with a fixed margin) in the section “The Fixed Margin Method (the Two Confidence Interval Method with a Fixed Margin)” is used. However, if the fixed-margin method (TCI with a random margin) in the section “The Fixed-Margin Method (the Two Confidence Interval Method with a Random Margin)” is employed, the formula given in Eq. 21 should not be used. Instead, another formula based on the across-trial power should be employed. The point (9) of section “The unconditional Across-Trial Type I Error Rate” will address this issue.

6. Although the constancy assumption does not hold, this does not affect the within-trial type I error rate, as it assesses only the type I error rate related to the hypotheses in Eq. 3 for a given fixed margin $\delta > 0$, regardless of how the margin is chosen. A violation of the constancy assumption can affect the conclusion in Eq. 20, but the within-trial type I error rate cannot assess the degree to which it does statistically.
7. Hung, Wang, and O'Neill^[14] refer to the within-trial type I error rate as “the conditional type I error rate.” However, this wording is slightly misleading. It is important to note that the within-trial type I error rate is an “unconditional” probability, not a “conditional” probability, because the non-inferiority margin is treated as a constant and not as a random variable.
8. When one studies some of the statistical properties of the fixed-margin method (TCI with a fixed margin) in the section “The Fixed Margin Method (the Two Confidence Interval Method with a Fixed Margin),” it is also possible to study only the within-trial type I error rate and the within-trial power but not the across-trial type I error rate and across-trial power.

The Unconditional Across-Trial Type I Error Rate

In this section, we describe the unconditional across-trial type I error rate of the synthesis method. Due to the existence of asymptotic equivalence between the synthesis method and the fixed-margin method (TCI with a random margin), as shown in the section “Differences between the Three Methods,” a similar description can be obtained for the fixed-margin method (TCI with a random margin).

The unconditional across-trial type I error rate of the synthesis method is given by:

$$P(T \geq 1.96 | H_{04}) = \int P(T \geq 1.96 | T_H = t_H, H_{04}) w_{T_H}(t_H) dt_H \quad (22)$$

where T is determined in Eq. (12) and $w_{T_H}(t_H)$ is the probability density function of T_H . Eq. 22 can be derived from the following well-known properties of a conditional expectation.

$$E[A] = E_{T_H}[E[A | T_H]], A = I_{[T \geq 1.96]} \quad (23)$$

This equation shows that the unconditional across-trial type I error rate is a weighted average of the conditional across-trial type I error rates. The unconditional across-trial type I error rate was evaluated by Kang and Ryu,^[16] Odem-Davis and Fleming,^[17] Snapinn and Jiang,^[6] Soon et al.,^[7] Wang, Hung, and Tsong,^[18] Wang and Hung,^[19] and Wang and Kang.^[2]

The unconditional across-trial type I error rate has the following characteristics:

1. The unconditional across-trial type I error rate is appropriate for the synthesis method and the fixed-margin method (TCI with a random margin) because both data in the current non-inferiority trial and data in the historical trial are treated as random variables. If we want to conduct simulation studies that evaluate the unconditional across-trial type I error rate, random numbers should be generated from both current non-inferiority trial and historical trial. If the non-inferiority margin is chosen from the historical data, this procedure should be reflected in simulation studies.
2. The paradigm of the within-trial type I error rate differs from that of the unconditional across-trial type I error rate, and they are as a result not comparable.^[3] Therefore, the values of these two type I errors should not be compared.
3. The fixed-margin method (TCI with a random margin) is more conservative than the synthesis method in terms of the unconditional across-trial type I error rate.^[4,18] When one asserts that a particular method is conservative, one should state clearly the type of type I error in relation to which this is so.

4. When $\lambda = 0$ in Eq. 6, we can assess the unconditional across-trial type I error rate for the hypotheses in Eq. 18, while we cannot evaluate the within-trial type I error rate. This is advantageous, as the hypotheses in Eq. 18 represent the minimal requirement for the approval of the test treatment.
5. The current non-inferiority trial and the historical trial are in fact different studies conducted with different protocols at different times. Furthermore, randomization and blinding are not used. However, in the evaluation of the unconditional across-trial type I error rate, all four arms (the test treatment and active control in this study and the active control and placebo in the historical study) are treated as if they come from the same clinical trial, which may cause substantial bias.^[20]
6. If there are several historical trials, there is some concern over the occurrence of selection and publication bias.^[20] These types of bias can affect the unconditional across-trial type I error rate. Hung and Wang^[21] pointed out that subjective judgment when selecting the relevant historical trials among several trials and the construction of a meta-analytic estimate of the control effect in multiple ways may produce multiplicity issues and inflate the family-wise unconditional across-trial type I error rate because the historical data are treated as random variables.
7. Even before we begin the non-inferiority trial, the historical trial is already finished and the historical data have already been observed. However, the unconditional across-trial type I error rate is calculated by incorporating all possible (even unobserved) values of historical trial data, which may not make sense.^[13] The unconditional across-trial type I error rate represents the type I error rate of the falsely concluded non-inferiority in Eq. 7 or 11 when both the non-inferiority trial and the historical trial are repeated infinitely.^[14] However, given that the historical trial has already finished, it may be unreasonable to assume that it is repeated infinitely.
8. Because the historical data are treated as random variables and given that T and T_H share the same historical data, the following correlation exists between them.^[2,13,16]

$$\rho = \frac{-(\lambda - 1)\sigma_H^2}{\sqrt{\sigma_H^2[\sigma^2 + (\lambda - 1)^2\sigma_H^2]}} = \frac{-(\lambda - 1)}{\sqrt{r^2 + (\lambda - 1)^2}}, \quad (24)$$

$$r^2 = \sigma^2 / \sigma_H^2$$

In other words, the correlation affects the extent to which T is significant; therefore, the appropriate type I error rate should depend on the correlation. However, the unconditional across-trial type I error rate does not consider this correlation and has an identical value regardless of the value of the correlation.^[2,13,16]

This arises because all possible (even unobserved) values of the historical trial data are incorporated into the calculation of the unconditional across-trial type I error rate. On the other hand, the conditional across-trial type I error rate considers only the observed value of the historical data; hence, it has a desirable property. This will be discussed in detail in point (2) of the section “The Conditional Across-Trial Type I Error Rate.”

9. If the unconditional across-trial type I error rate is employed, the formula given in Eq. 21 should not be used, as it is based on the within-trial power. Rothman, Wiens, and Chan^[22] provide a sample size calculation method based on the unconditional across-trial power.
10. Suppose that in the fixed-margin method (TCI with a fixed margin) in the section “The Fixed Margin Method (the Two Confidence Interval Method with a Fixed Margin),” the non-inferiority margin was selected as a fraction of the lower confidence limit of the effect size of the active control in the historical trial.

$$\delta = (1 - \lambda)(\bar{X}_{CH} - \bar{X}_{PH} - 1.96\hat{\sigma}_{PC|H}) \quad (25)$$

This margin is treated as a fixed constant once it is determined. Whenever the null hypothesis H_{01} is rejected by the fixed-margin method (TCI with a fixed margin) in the section “The Fixed Margin Method (the Two Confidence Interval Method with a Fixed Margin),” the null hypothesis H_{03} is also rejected by the fixed-margin method (TCI with a random margin) in the section “The Fixed-Margin Method (the Two Confidence Interval Method with a Random Margin)” and vice versa. This most likely is why the names of these two methods are identical. However, these two methods should assess different type I error rates because they employ different paradigms. The fixed-margin method (TCI with a fixed margin) in the section “The Fixed Margin Method (the Two Confidence Interval Method with a Fixed Margin)” should assess the within-trial type I error rate, while the fixed-margin method (TCI with a random margin) in the section “The Fixed-Margin Method (the Two Confidence Interval Method with a Random Margin)” should evaluate the (either unconditional or conditional) across-trial type I error rate.

The Conditional Across-Trial Type I Error Rate

The conditional across-trial type I error rate is defined as follows:

$$P(T \geq 1.96 | T_H = t_H, H_{03} \text{ or } H_{04}) \quad (26)$$

When the across-trial type I error rate is assessed, the historical data are treated as random variables. However, because the historical data have already been observed, a conditional probability rather than an unconditional probability may be more appropriate as a type I error rate. This is the main motivation of the conditional across-trial type I error rate.^[13]

The conditional across-trial type I error rate has the following characteristics.

1. The conditional across-trial type I error rate incorporates only the observed values of the historical data, whereas the unconditional across-trial type I error rate takes into account all possible (even unobserved) values of the historical trial data. Therefore, the conditional across-trial type I error rate appears to be more suitable than the unconditional across-trial type I error rate.^[13]

The conditional across-trial type I error rate is truly a conditional probability given that the observed value of T_H is equal to t_H . The conditional across-trial type I error rate is the type I error rate of the falsely concluded non-inferiority in Eq. 7 or 11 with an identical fixed value of T_H when the only non-inferiority trial (not historical trial) is repeated infinitely with the same fixed value of the random variable margin. This idea is appealing because the historical trial was finished and cannot be repeated.^[13]

2. With the asymptotic property of T given T_H , the conditional across-trial type I error rate is given by:^[13]

$$P(T \geq 1.96 \mid T_H = t_H, H_{04}) \rightarrow 1 - \Phi\left(\frac{1.96 - \rho t_H}{\sqrt{1 - \rho^2}}\right) \text{ as } n \rightarrow \infty \quad (27)$$

where Φ is the cumulative distribution function of the standard normal distribution. From the formula given in Eq. 27, we see that the conditional across-trial type I error rate depends on the correlation coefficient ρ between T and T_H . As a result, the conditional across-trial type I error rate appears to be more appropriate than the unconditional across-trial type I error rate due to the reason described in point (8) of the section “The Unconditional Across-Trial Type I Error Rate.” A major problem of the conditional across-trial type I error rate is that it depends on the unknown nuisance parameter $(\mu_{CH} - \mu_{PH})$.^[13]

CONCLUSION

Based on the following reasons, the within-trial type I error rate should be controlled such that it remains under the nominal level in order to receive approval for a new

treatment from regulatory agencies in two-arm non-inferiority trials.

1. In Statistics 101: In the “Introduction,” we learned the frequentist statistical principle, which states that hypotheses should be established prior to conducting an experiment. Hypotheses consist of unknown parameters and known constants. Information such as historical data can only be reflected in the constants in the hypotheses and/or in the form of the hypotheses. The hypotheses are not allowed to have any random variables from historical data. Therefore, as long as we observe the frequentist statistical principle, the non-inferiority margin should be treated as a fixed constant. A problem with this method is that it ignores the variability of the historical data when calculating the non-inferiority margin. However, this is the price that we should pay if we follow the frequentist statistical principle.
2. As described in point (5) of the section “The Unconditional Across-Trial Type I Error Rate,” to calculate the across-trial type I error rate, both current non-inferiority trial and historical trial are treated as an identical clinical trial that was conducted with identical protocols at the same time. However, this is not true. Ignoring this truth may produce considerable bias, which cannot be quantified.
3. Treating historical data as random variables during the calculation of the across-trial type I error rate may cause a serious problem. Questions of which historical trials are selected and how to use these historical trials may arise when there are several historical trials. Selection and publication biases may occur, as described in point (6) of section “The Unconditional Across-Trial Type I Error Rate,”^[20] which may affect the across-trial type I error rate. The family-wise across-trial type I error rate may also be inflated due to the subjective selection of historical trials and the multiple methods of calculating the meta-analytic estimate of the active control effect.^[21]
4. The sample size formula given in Eq. 21 has been widely employed in non-inferiority trials. Because the formula is based on the within-trial power, it is the within-trial type I error rate, which should be held under the nominal level. If one controls the across-trial type I error rate, the sample size should also be computed based on the across-trial power.
5. A disadvantage of the unconditional across-trial type I error rate is that it does not take into account the correlation between T and T_H , as described in point (8) of section “The Unconditional Across-Trial Type I Error Rate.” Hence, the unconditional across-trial type I error rate exceeds the nominal level as the correlation increases.^[13] A problem with the conditional across-trial type I error rate is that it cannot be calculated owing to the unknown nuisance parameter.

ACKNOWLEDGMENT

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1D1A1A09916819).

REFERENCES

- Hung, H.M.; Wang, S.J.; O'Neill, R. A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. *Biometrical J.* **2005**, *47*, (1), 28–36.
- Wang, S.; Kang, S.H. Strength of evidence of non-inferiority trials with the two confidence interval method with random margin. *J. Biopharm. Stat.* **2013**, *23*, 307–321.
- Hung, H.M.; Wang, S.J.; O'Neill, R. Challenges and regulatory experiences with non-inferiority trial design without placebo arm. *Biometrical J.* **2009**, *51*, (2), 324–334.
- US FDA. *Guidance for Industry: Non-Inferiority Clinical Trials to Establish Effectiveness*; Center for Drug Evaluation and Research, U.S. Food and Drug Administration: Silver Spring, MD, 2016.
- Chow, S.C.; Shao, J. On non-inferiority margin and statistical tests in active control trials. *Stat. Med.* **2006**, *25*, 1101–1113.
- Snapinn, S.; Jiang, Q. Controlling the type I error rate in non-inferiority trials. *Stat. Med.* **2008**, *27*, 371–381.
- Soon, G.; Zhang, Z.; Tsong, Y.; Nie, L. Assessing overall evidence from noninferiority trials with shared historical data. *Stat. Med.* **2013**, *32*, (14), 2349–2363.
- Brittain, E.; Hu, Z. Noninferiority trial design and analysis with an ordered three-level categorical endpoint. *J. Biopharm. Stat.* **2009**, *19*, 685–699.
- Hilton, J.F. Non-inferiority trial designs for odds ratios and risk differences. *Stat. Med.* **2010**, *29*, 982–993.
- Sanchez, M.M.; Chen, X. Choosing the analysis population in non-inferiority studies: Per protocol or intent-to-treat. *Stat. Med.* **2006**, *25*, 1169–1181.
- Sheng, D.; Kim, M.Y. The effects of non-compliance on intent-to-treat analysis of equivalence trials. *Stat. Med.* **2006**, *25*, 1183–1199.
- Hung, H.M.; Wang, S.J.; Tsong, Y.; Lawrence, J.; O'Neill, R. Some fundamental issues for noninferiority testing in active controlled trials. *Stat. Med.* **2003**, *22*, 213–225.
- Kang, S.H.; Tsong, Y. Strength of evidence of non-inferiority trials – the adjustment of the type I error rate in non-inferiority trials with the synthesis method. *Stat. Med.* **2010**, *29*, 1477–1487.
- Hung, H.M.; Wang, S.J.; O'Neill, R. Issues with statistical risks for testing methods in non-inferiority trial without a placebo arm. *J. Biopharm. Stat.* **2007**, *17*, 201–213.
- Chow, S.C.; Shao, J.; Wang, H. *Sample Size Calculations in Clinical Research*; CRC Press: Boca Raton, FL, 2003.
- Kang, S.H.; Ryu, Y. The adjustment of the type I error rate in non-inferiority trials with lambda-margin approach: Each of two different new drugs is approved with two independent trials with the same active control. *J. Biopharm. Stat.* **2011**, *21*, 498–510.
- Odem-Davis, K.; Fleming, T.R. Adjusting for unknown bias in noninferiority clinical trials. *Stat. Biopharm. Res.* **2013**, *5* (3), 248–258.
- Wang, S.J.; Hung, H.M.; Tsong, Y. Utility and pitfalls of some statistical methods in active controlled clinical trials. *Control. Clin. Trials.* **2002**, *23*, 15–28.
- Wang, S.-J.; Hung, H.-M. Assessing treatment efficacy in non-inferiority trials. *Control. Clin. Trials.* **2003**, *24*, 147–155.
- Carroll, K.J. Statistical issues and controversies in active-controlled, “non-inferiority” trials. *Stat. Biopharm. Res.* **2013**, *5* (3), 229–238.
- Hung, H.M.; Wang, S.J. Statistical considerations for non-inferiority trial designs without placebo. *Stat. Biopharm. Res.* **2013**, *5* (3), 239–247.
- Rothman, M.D.; Wiens, B.L.; Chan, I.S.F. *Design and Analysis of Non-inferiority Trials*; Chapman & Hall/CRC: Boca Raton, FL, 2012.

Nonparametric Regression

Kongming Wang

Pfizer Inc., New York, New York, U.S.A.

Theo Gasser

University of Zurich, Zurich, Switzerland

Abstract

Clinical Trial Simulations for Later Development Phases Clinical trial simulation plays an important role in drug development and, in recent years, it has become increasingly popular in the pharmaceutical industry. The focus of this entry is the simulation for adaptive trial designs, which will be defined in detail through several statistical examples.

INTRODUCTION

Non-parametric regression is a fundamental tool in data analysis and provides an effective and simple way of finding data structure. The goal is to recover an unknown function based on sampled data that are contaminated with noise. Hess et al.^[1] have compared several non-parametric methods for estimating hazard functions. Parametric regression models (e.g., linear regression) have finite number of parameters to be estimated, whereas non-parametric models assume only very general assumptions about the unknown function such as that it belongs to a certain class of smooth functions. When estimating r , “smooth” means typically that r is twice continuously differentiable, so there are infinitely many parameters in a non-parametric model.

Given data points (x_i, y_i) where y_i is the observed response at x_i , $i = 1, \dots, n$, we want to study the relationship between x and y . The relationship can be written in general as

$$y_i = r(x_i) + \sigma \varepsilon_i$$

for some unknown function r . The random errors ε_i are assumed to be independently distributed with mean $E(\varepsilon_i) = 0$ and variance $V(\varepsilon_i) = 1$. For random designs where x_i are random variables, the function r is the expectation $r(x) = E(y|x)$. Usually it is assumed that ε_i and x_i are independent. The function r has to be estimated at each x , without the help of a parametric model which specifies qualitatively the relationship between x and y .

Non-parametric regression has been a very active field of statistics, and many methods have been proposed for estimating the function r . The commonly used methods (kernel estimators, smoothing splines, local polynomials, linear wavelet shrinkage) are linear methods in the sense that they are weighted averages of the observed responses

$$\hat{r}(x) = \sum_{i=1}^n w_i(x|x_1, \dots, x_n, b) y_i \quad (1)$$

where w_i are weights depending on design data points x_i , $i = 1, \dots, n$, a smoothing parameter b (“bandwidth”) indicating the extent of smoothing, and the point x at which the function is being estimated. A focal point in these estimators is the selection of the smoothing parameter b , and many methods (global, local, data driven) have been proposed for this purpose. The selection of b depends on the noise level σ and the smoothness of r (see below).

We will describe linear estimators in the section “Linear Estimators.” Spatially adapted wavelet shrinkage/threshold methods, which are similar to the kernel estimators with locally determined smoothing parameters, will also be described there. Analyzing a sample of similar curves (e.g., growth curves, dose-response curves) will be discussed in the section “Analyzing Sample of Curves.” An iterative algorithm based on kernel smoothing and dynamic time warping will be applied to analyze the Zurich growth data in the section “Application to Growth Data” and brain potentials in the section “Application to Brain Potentials.” Multivariate regression and density estimation will be briefly discussed in the section “Kernel Density Estimation and Multivariate Regression.” Some software resources are mentioned in the section “Software,” which is followed by a short conclusion.

LINEAR ESTIMATORS

Without loss of generality we assume that the design points satisfy $0 \leq x_1 < \dots < x_n \leq 1$.

Kernel Estimators

The weights w_i in (1) are derived from a kernel function K , which is usually chosen to satisfy $K(x) = K(-x)$ and to have compact support. Nadaraya^[2] and Watson^[3] proposed the weights

$$w_i(x|x_1, \dots, x_n, b) = K\left(\frac{x-x_i}{b}\right) / \sum_{j=1}^n K\left(\frac{x-x_j}{b}\right).$$

Convolution type weights were proposed by Gasser and Muller^[4,5] as

$$w_i(x|x_1, \dots, x_n, b) = \frac{1}{b} \int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{b}\right) du$$

where $s_i = (x_i + x_{i+1})/2$, $s_0 = 0$, and $s_{n+1} = 1$.

The kernel and the smoothing parameters have to be selected in order to apply the estimator in Eq. (1). The estimator $\hat{r}$ is as smooth as the kernel K . The kernel has to satisfy certain conditions such as $\int K(x)dx=1$ for estimating the function r and $\int K(x)xdx=0$ together with $\int K(x)xdx=-1$ for estimating the first order derivative of r .^[6] For estimating the function r , the Epanechnikov kernel $K(x) = 3/[4(1-x^2)_+]$ is optimal in the sense that it minimizes the mean square error (MSE), which is the sum of variance and square of the bias.⁶ The derivatives of the unknown function r can be estimated by a similar weighted average with specially constructed kernels. Optimal kernels for estimating derivatives and special boundary kernels were also presented in Gasser et al.^[6] The boundary region consists of the intervals $(0, b)$ and $(1-b, 1)$ and a naive application of the kernels for the interior leads there to bias problems, as they are no longer symmetric in these regions.

The smoothing parameter b determines the extent of smoothing. Because K has compact support, a small b means very few data points are used in the weighted average in Eq. (1) and therefore little smoothing is done. Globally selected smoothing parameter does not depend on the location where the function is estimated. For a smooth function without peaks and sharp changes, a globally selected smoothing parameter is appropriate. Locally selected smoothing parameter adapts to the fine structure of the unknown function r . Local smoothing parameter is larger where r is flat and smaller where r has peaks. The optimal parameter, which minimizes the asymptotic MSE, is given by^[7]

$$\text{Local: } b(x) = \left(\frac{\sigma^2}{cn r^{(2)}(x)^2} \right)^{1/5}$$

$$\text{Global: } b = \left(\frac{\sigma^2}{cn \int r^{(2)}(x)^2 dx} \right)^{1/5}$$

The constant c depends only on the kernel and $r^{(2)}$ is the second order derivative of r . These formulae depend on two parameters σ and $r^{(2)}$ which are unknown. Gasser et al.^[8] proposed a simple estimator of σ and Brockmann et al.^[7] proposed a “plug-in” method for estimating the optimal smoothing parameter. The plug-in method estimates $r^{(2)}$

first from data with oversmoothing. Wang and Gasser^[9] discussed the best relative rate for estimating the locally optimal bandwidth. A popular alternative for bandwidth selection is cross-validation (and related methods): a disadvantage is that cross-validation is more variable in theory and practice, an advantage is that it is generally applicable and relatively simple to implement.

Smoothing Splines

A smoothing spline is the solution to the penalized regression problem $\min_r S_b(r)$ where

$$S_b(r) = \sum_{i=1}^n (y_i - r(x_i))^2 + b \int_0^1 [r^{(2)}(x)]^2 dx.$$

Here b is a roughness penalty analogous to the bandwidth of the kernel estimator and $r^{(2)}$ is the second order derivative of the regression function. Without penalty ($b=0$), solving the minimization problem results in interpolating the data. With heavy penalty ($b \rightarrow \infty$), the solution approaches the least square linear fit. For a given smoothing parameter $0 < b < \infty$, the resulting estimator $\hat{r}$ is a natural cubic spline (piecewise cubic polynomial with continuous second order derivative) with knots at $x_1, \dots, x_n$.^[10] Silverman^[11] expresses smoothing spline as a kernel estimator with locally adaptive smoothing parameter asymptotically and took advantage of the explicit formula of kernel estimators in studying asymptotic properties of smoothing splines. Kelly and Rice^[12] propose restricted smoothing splines for monotone regression functions such as dose-response functions and growth curves. One could use a high order penalty term $\int_0^1 [r^{(m)}(x)]^2 dx$ to obtain a spline estimator that has $m-1$ absolutely continuous derivatives $r^{(p)}$, $p=1, \dots, m-1$, and bounded $r^{(m)}$.

The smoothing splines are defined implicitly via a minimization problem and many methods have been proposed for computing smoothing splines. One often-used method is to use a set of basis functions known as *B-splines* which are defined by the recurrence formula

$$B_{j,k}(x) = \frac{x-x_j}{x_{j+k-1}-x_j} B_{j,k-1}(x) + \frac{x_{j+k}-x}{x_{j+k}-x_j} B_{j+1,k-1}(x), k \geq 2$$

with $B_{j,1}(x) = 1$ for $x_j < x < x_{j+1}$.^[13] The smoothing cubic spline can be written as a linear combination of the cubic *B-splines* ($k=4$) as

$$r(x) = \sum_j \gamma_j B_j(x)$$

which leads to $(r(x_1), \dots, r(x_n))' = B\gamma$ with $B_j = B_j(x_i)$. Solve the minimization problem to get the coefficients

$$\hat{\gamma} = (B'B + bM)^{-1} B'\gamma, \text{ with } M_{ij} = \int_0^1 B_i^{(2)}(x) B_j^{(2)}(x) dx.$$

The *B-splines* are almost orthogonal so $B'B$ is almost diagonal, which makes the calculation fast.

Cross-validation can be applied to select the penalty parameter b from data to minimize the MSE.

Local Polynomials

The local polynomial estimator $\hat{r}(x)$ is obtained by locally weighted polynomial regression to minimize the weighted sum of squared residuals

$$\min_{a(x), b_1(x), \dots, b_k(x)} \sum_{i=1}^n K\left(\frac{x-x_i}{b}\right) [y_i - a(x) - b_1(x) \times (x-x_i) - \dots - b_k(x)(x-x_i)^k]^2.$$

Here K is a kernel and b a smoothing parameter as in kernel estimators. This leads to the estimator $\hat{r}(x) = a(x)$. Local linear fit ($k = 1$) has been used most frequently and the weights are

$$w_i(x|x_1, \dots, x_n, b) = g_i(x|x_1, \dots, x_n, b) / \sum g_j(x|x_1, \dots, x_n, b)$$

where

$$g_i(x|x_1, \dots, x_n, b) = K\left(\frac{x-x_i}{b}\right) (s_2(x|x_1, \dots, x_n, b) - (x-x_i) \times s_1(x|x_1, \dots, x_n, b))$$

$$s_p(x|x_1, \dots, x_n, b) = \sum_{j=1}^n K\left(\frac{x-x_j}{b}\right) (x-x_j)^p, \dots, p = 1, 2.$$

Local polynomials automatically adjust to the boundary, and in general the bias of local polynomial estimator declines and the variance increases with the order of the polynomial. Fan^[14] proves minimax properties of local polynomial estimators.

Similar to kernel estimators, a kernel and a smoothing parameter have to be selected before applying local polynomials for non-parametric regression. The Epanechnikov kernel is optimal over all non-negative symmetric functions.^[15] Cross-validation and plug-in methods can be applied to select the smoothing parameter. Seifert and Gasser^[16] have studied the variance properties of local polynomial fitting and proposed modifications to control variability when the design is sparse.

Fig. 1 illustrates the effect of smoothing parameter. This figure is produced using SAS procedure LOESS with local linear fitting. The globally optimal smoothing parameter, $b = 0.135$, is computed by the LOESS procedure. With undersmoothing ($b = 0.05$), noise is not removed. With oversmoothing ($b = 0.3$), detailed features are lost.

Wavelet Estimators

Smoothing or denoising can also be carried out in the frequency domain such as Fourier analysis. Fourier transform gives the spectral content of a signal, but it gives no information regarding where in time those spectral components appear. Therefore it is not suitable for non-stationary signal analysis. Wavelet transform provides time and frequency information simultaneously, hence giving a time frequency representation of the signal.

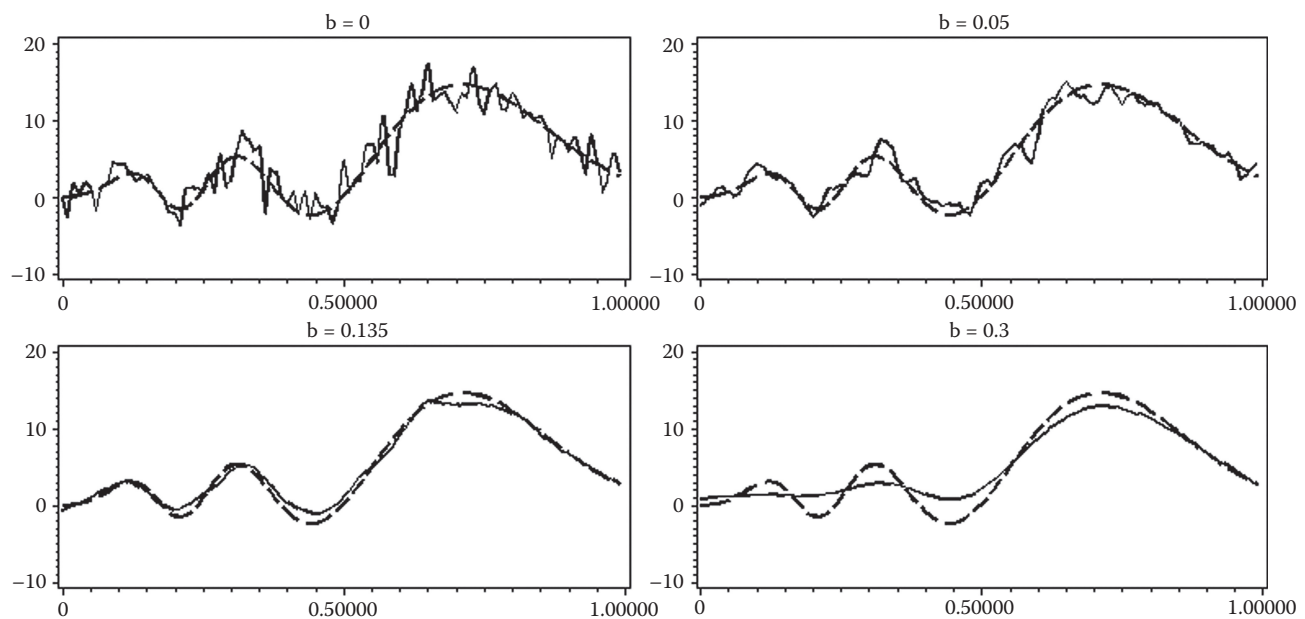


Fig. 1 Illustration of the effect of smoothing parameter. Dashed line is the true function and solid line is smoothed function (estimate). $b = 0$: data plot (no smoothing); $b = 0.05$: undersmoothing; $b = 0.135$: optimal smoothing parameter chosen by SAS procedure LOESS; $b = 0.3$: oversmoothing

Wavelets refer to a set of orthonormal basis functions generated by dilation and translation of a compactly supported scaling function (or father wavelet), ϕ , and a mother wavelet, ψ . A variety of different wavelet families now exist that combine compact support with various degrees of smoothness and numbers of vanishing moments, see Daubechies.^[17] Many types of functions can be sparsely and uniquely represented in terms of a wavelet series. Let

$$\phi_{jk}(x) = 2^{j/2} \phi(2^j x - k), \psi_{jk}(x) = 2^{j/2} \psi(2^j x - k), x \in [0, 1].$$

Then for any $j_0 \geq 0$,

$$\{\phi_{j_0 k}, k = 0, 1, \dots, 2^{j_0} - 1; \psi_{jk}, j \geq j_0, k = 0, 1, \dots, 2^j - 1\}$$

is an orthonormal basis of $L^2([0, 1])$. The basis with $\phi(x) = 1, x \in [0, 1]$, and $\psi(x) = 1, x \in [0, 1/2]$ and $\psi(x) = -1, x \in [1/2, 1]$, is the Haar wavelet family.

For any function $r \in L^2([0, 1])$, it can be represented as

$$r(x) = \sum_{k=0}^{2^{j_0}-1} \langle r, \phi_{j_0 k} \rangle \phi_{j_0 k}(x) + \sum_{j=j_0}^{\infty} \sum_{k=0}^{2^j-1} \langle r, \psi_{jk} \rangle \psi_{jk}(x).$$

The first term is the trend at scaling level j_0 and the second term is the sum of the fluctuation terms.^[18]

In non-parametric regression applications, the function is discretely sampled $R = (r(x_1), \dots, r(x_n))'$. If $n = 2^J$ for some positive integer J and $\{x_i\}$ are equally spaced, fast algorithm using filter banks exists^[19] for implementing the discrete wavelet transform (DWT) to calculate wavelet coefficients $\alpha_{j_0 k} = \langle r, \phi_{j_0 k} \rangle, k = 0, 1, \dots, 2^{j_0}-1$, and $\beta_{jk} = \langle r, \psi_{jk} \rangle, j = j_0, \dots, J-1; k = 0, 1, \dots, 2^j-1$. Note that DWT calculates only n wavelet coefficients. If these coefficients are known, then inverse discrete wavelet transform (IDWT) can be used to calculate the functional values at $\{x_i\}$ using n wavelets.

The wavelet coefficients calculated from noisy data $\{y_i\}$ can be represented as

$$\hat{\alpha}_{j_0 k} = \alpha_{j_0 k} + \sigma \tau_{j_0 k}, \hat{\beta}_{jk} = \beta_{jk} + \sigma \tau_{jk}.$$

Because the wavelets form an orthonormal basis of $L^2([0, 1])$, the transformed noise τ_{jk} are also independently distributed with mean $E(\tau_{jk}) = 0$ and variance $V(\tau_{jk}) = 1$. Wavelet methods are designed to estimate the true wavelet coefficients $\{\alpha_{j_0 k}, \beta_{jk}\}$ from the noisy ones $\{\hat{\alpha}_{j_0 k}, \hat{\beta}_{jk}\}$ and use IDWT to recover the true function.

For regression and denoising, two classes of wavelet estimates are often used. One is linear wavelet shrinkage which estimates the trend term coefficient $\alpha_{j_0 k}$ by $\hat{\alpha}_{j_0 k}$ and the fluctuation term coefficients β_{jk} by shrinking $\hat{\beta}_{jk}$ towards 0: $\tilde{\beta}_{jk} = s(j, b) \hat{\beta}_{jk}$ with $0 < s(j, b) < 1$. The shrinking factor $s(j, b)$ depends on a smoothing parameter b and possibly on the scaling level j . The smoothing

parameter depends on the estimated noise parameter σ . The other class is nonlinear wavelet thresholding which also estimates the trend term coefficient $\alpha_{j_0 k}$ by $\hat{\alpha}_{j_0 k}$, but it estimates the fluctuation term coefficients β_{jk} by thresholding $\tilde{\beta}_{jk} : \tilde{\beta}_{jk} = I(|\hat{\beta}_{jk}| > t(j, b)) \hat{\beta}_{jk}$ where the threshold $t(j, b)$ depends on a smoothing parameter b and possibly on the scaling level j . Antoniadis et al.^[20] have summarized various methods for selecting $s(j, b)$ and $t(j, b)$, and they selected the trend level j_0 as $j_0(n) = \log_2(\log(n)) + 1$ based on asymptotic considerations. Donoho and Johnstone^[21] have proposed an estimate of the noise level σ based on the wavelet coefficients given by

$$\hat{\sigma} = \frac{\text{median}(|\hat{\beta}_{j-1, k}| : k = 0, 1, \dots, 2^{j-1} - 1)}{0.6745}$$

ANALYZING SAMPLE OF CURVES

Many practical applications involve a sample of curves, or functional data. Examples are dose-response relationships from a group of subjects in early phase clinical trials, growth curves of children from birth to adulthood, speech signals, and brain potentials recorded from many channels and from many subjects. Such data can be modeled in the form

$$y_{ij} = r_i(x_{ij}) + \sigma \epsilon_{ij}, \quad j = 1, 2, \dots, n_i, \quad i = 1, 2, \dots, I.$$

There are I curves and curve i is sampled at n_i points.

Each curve can be estimated with the methods described in the section "Linear Estimators." Gasser et al.^[22] has suggested using the same smoothing parameter for all curves. They proposed to use the average of the optimal smoothing parameters derived for each individual curves.

The curves are similar in shape because these curves represent similar relationship between $\{x_{ij}\}$ and $\{y_{ij}\}$, yet the curves are different in characteristics (e.g., location and magnitude of peaks) due to inter-individual variations. Many methods such as principal component analysis deal with variations in amplitude while ignoring dynamic variations such as the timing of the features. There are sensitive and insensitive subjects in dose-response studies, children with earlier and later pubertal spurts, people speaking slowly or quickly, and subjects with different responses to stimulus when recording brain potentials. The aim is to model both similarity and individual variations, in particular when estimating a mean curve.

Many ideas and techniques for such data, including linear regression, principal components analysis, linear modeling, and canonical correlation analysis, as well as specifically functional techniques such as curve registration and principal differential analysis, are discussed in Ramsay and Silverman.²³ The following sections present several semiparametric and non-parametric methods for structural analysis of curves.

Shape-Invariant Modeling

The one-component shape-invariant model (SIM) is a semi-parametric model with a common “shape function” for all curves.^[22]

$$r_i(x_{ij}) = a_i s\left(\frac{x_{ij} - b_i}{c_i}\right) + d_i$$

where a_i , b_i , c_i , and d_i are individual parameters. Shape function and individual parameters have to be estimated from data. The Gompertz model of $s(x) = \exp[-\exp(-x)]$ is a parametric version of the SIM.

A sophisticated algorithm for estimating the shape function and individual parameters is sketched in Gasser et al.^[22] The algorithm starts from an approximate shape function (e.g., logistic function). Then individual parameters are estimated from data using the approximate shape function and nonlinear least squares. With the shape function and individual parameters at hand, the estimated curves can be calculated from the SIM model. An update for the shape function is then derived by fitting the residuals with smoothing or regression splines. This procedure is repeated until convergence.

Structural Analysis

Landmark-based methods are used in signal and image registrations.^[24,25] Landmarks are features of images. For curves, landmarks could be peaks of curves and their derivatives. Kneip and Gasser^[26] have proposed a landmark method to construct a mean curve of the sample curves. They used the term “structural average curve” because it was applied to growth curve analysis. The structural average curve can be an estimate of the shape function of the SIM model.

The structural averaging proceeds as follows:

- i. Estimates of individual curves and their first- and second-order derivatives are obtained with Kernel smoothing: $\hat{r}_i, \hat{r}'_i, \hat{r}''_i$.
- ii. Individual structural points are identified from $\hat{r}_i, \hat{r}'$, and/or $\hat{r}''$. Roughly speaking, structural points (called “landmarks” in shape analysis) are features that are common to all or most curves.
- iii. Shift functions $\hat{u}_i$ are constructed from the locations of the structural points such that individual structural points are shifted to the respective average location. In between, smooth monotonic interpolation is used.
- iv. A structural average $\hat{r}_0$ (in contrast to a cross-sectional average) is obtained by averaging aligned (smoothed) curves:

$$\hat{r}_0(x) = \sum_{i=1}^I \hat{r}_i(\hat{u}_i(x)) / I$$

Step (ii) sounds simple but it is rather delicate. It is not easy to define structural points that are common to most curves and to determine them unequivocally from noisy data, such that they have an equivalent meaning. Structural points could be defined based on the applications. For evoked brain potentials, the structural points might be P100, N100, P200, N200, and P300 (the positive and negative peaks occurring at about 100, 200, and 300 milliseconds after applying a stimulus). For general curve analysis, the structural points can be simply the peaks of $\hat{r}_i, \hat{r}'$, and $\hat{r}''$, which are common to most of the curves.

Step (iii) eliminates the dynamic variations among individuals by aligning the features across individuals. Once the features are identified in step (ii), this is a simple procedure.

Step (iv) is the average of the aligned curves.

Because of the averaging in step (iv), the average curve is much smoother than the smoothed individual curves and has little variation. Therefore, it is a good estimate of the shape function. Kneip and Gasser^[26] have shown that the structural average is closer to the underlying growth model than the mean of individually determined peak velocity of the pubertal spurt.

Dynamic Time Warping

Dynamic time warping was developed to align two signals with different dynamics.^[27–29] This method has been mainly applied to speech analysis and speech recognition. Suppose that two sequences $\{f(i), i = 1, \dots, M\}$ and $\{g(j), j = 1, \dots, N\}$ characterize two signals respectively. We want to find the best match between the two signals by some alignment w , based on minimizing a cost function. The classical cost function is given by

$$\inf_w \sum_{(i,j) \in w} (f(i) - g(j))^2.$$

Here $w = \{(i, j)\}$ is a warping path connecting $(1, 1)$ and (M, N) in a two dimensional square lattice and satisfying monotonicity and connectedness. This means that both coordinates of the parametrized path $w = \{(i(k), j(k)): k = 1, \dots, K; i(1) = j(1) = 1, i(K) = M, j(K) = N\}$ have to be nondecreasing, and that they can only increase by 0 or 1 when going from k to $k + 1$. Obviously, this is a minimization problem that can be solved efficiently by using dynamic programming. Fig. 2 is a demonstration of aligning one curve to another using dynamic time warping.

What controls the warping result is the cost function. In principle, a cost function could be any functional of the two input signals and a warping path, depending on the purpose of the application. For the purpose of aligning maxima, minima, and inflection points of curves, we incorporate derivatives of the signals into the cost function.^[30] Apart from heuristics, theoretical analysis and

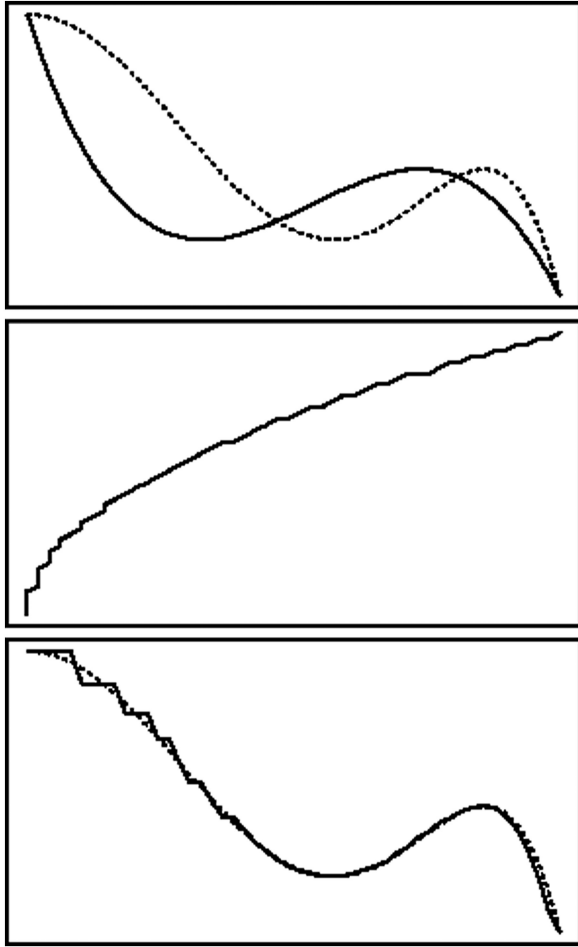


Fig. 2 Dynamic time warping demonstration. Panels from top to bottom: two signals; warping function to align one curve to another; aligned signals

simulation analysis lend support to this idea. Incorporating higher order derivatives (the second derivative in particular) is possible in principle, but estimating higher order derivatives from noisy data is difficult. The cost function is given by

$$C(w) = \int F(f, f', g, g', w, \alpha)(x) \phi(w'(x)) dx$$

where f' is the derivative of f , α is a parameter to balance the alignment of curves and their derivatives, and the functional F is given by

$$F(f, f', g, g', w, \alpha)(x) = \alpha^2 \left(\frac{f(x)}{\|f\|} - \frac{g(w(x))}{\|g\|} \right)^2 + (1 - \alpha)^2 \left(\frac{f'(x)}{\|f'\|} - \frac{g'(w(x))}{\|g'\|} \right)^2.$$

Here $\|f\|$ is the infinite norm of f . The normalization is for reducing alignment error due to difference in amplitude. The penalty function ϕ is for reducing excessive alignment

and can be defined by a convex function with minimum $\phi(1) = 0$. This also makes the optimal alignment unique. To reduce the alignment error due to smoothing errors from estimating derivatives, the parameter α can be chosen from (0.5, 1).

Compared to the structural analysis as described in the section “Structural Analysis,” the advantage of dynamic time warping is that it automatically aligns structural points of two functions. It needs thus less a priori knowledge and less manpower, but structural analysis could still be preferable in difficult situations. Simulations had shown that dynamic time warping is an appropriate method for aligning two functions.^[30] The proposed cost function outperforms the classical cost function because of the use of derivatives, in particular when there are nonlinear shifts between functions. Furthermore, dynamic time warping with the proposed cost function produces smoother shift functions. As expected, the performance of warping depends mainly on the amount of true time shifts between two functions. The noise level does not affect the warping results very much since data are smoothed before warping.

Dynamic time warping produces a relative shift function between two curves. To align a set of sample curves to a common time scale, we need a reference curve. Then a warping function can be estimated by dynamic time warping and all curves can be aligned to this reference curve. Wang and Gasser^[31] have proposed the following iterative algorithm for aligning a set of curves.

Step (i): set the initial warping functions $\hat{w}_i^0(x) = x$ (noalignment).

Step (ii): compute the initial aligned average as

$$\hat{r}^0(x) = \frac{\sum_{i=1}^I \hat{r}_i(\hat{w}_i^0(x))}{I}.$$

Step (iii): update the warping functions by solving the minimization problems

$$C(\hat{w}_i^1) = \min_w C(w) = \int F(\hat{r}_i, \hat{r}_i', r^0, r^{0'}, w, \alpha)(x) \phi(w'(x)) dx.$$

Repeat steps (ii) and (iii) till convergence.

Under some assumptions, it can be shown that the algorithm converges and the aligned average is the shape function if the sample curves are generated from a SIM model.^[31] Because the algorithm estimates the differences in timing between the sample curves and the average curve, confidence regions can be calculated at feature points of the average curve.^[32]

Simulation results and some computational considerations discussed in Wang and Gasser^[31] are briefly described here. The parameter α can be chosen such that it produces the minimum cost $\min_w C(w)$. Choosing α from a few values {0.3, 0.5, 0.7} worked well in our simulations. On one hand, the penalty function ϕ should be strictly convex such that the solution of the minimization

problem of step (iii) is unique. On the other hand, ϕ should be flat so that it does not produce artificial warping functions by pushing toward the identity function. To avoid estimating derivatives of the warping functions, we have replaced the single penalty term $\phi(w'(x))$ with three conditions which are easy to handle in programming. Our simulations show that a few iterations (less than or equal to five iterations) produce nice results. In our simulations and applications, we use three iterations.

APPLICATION TO GROWTH DATA

The Zurich growth data is analyzed with the methods described in the sections “Linear Estimators” and “Analyzing Sample of Curves.” The data are the measurements of the growth (height, shoulder width, etc) of children from birth to adulthood. There are 120 boys and 112 girls. For each person, the measurements are taken quarterly from birth to one year old, semi-annually from one to two years old, annually from two to ten years old, and then semiannually from ten till pubertal growth has stopped. This results in at most 36 measurements per person. The actual number of measurements per person

ranges from 27 to 35, with an average of 32 measurements per person.

Here we discuss growth velocity of the shoulder widths of the children. Growth velocity is more informative because growth curves are simply increasing functions of age. From birth to two years old, the growth speed is high but it is simply decreasing. Therefore, we will analyze the growth speed of the children’s shoulder widths from 2 to 20 years old. The measurement unit is in centimeter.

Individual growth velocities (the derivatives of the growth curves) and second-order derivatives of the growth curves were estimated by kernel smoothing. The iterative algorithm described in the section “Dynamic Time Warping” was applied to estimate an average growth velocity of shoulder widths. Figs. 3 and 4 plot the results for boys and girls respectively.^[31]

From subgraph (b) of the Figs. 3 and 4, we see that the growth speed of a child’s shoulder widths changes quite a bit between individuals. The simple average of the growth speed of more than 100 children does catch the growth spurt at about age 14, but the amplitude of this peak is reduced. Furthermore, it misses the early spurt at about age 7 (see subgraph (d) of the Figs. 3 and 4). The mid growth spurt occurs at about the same age for boys and girls, but the major spurt

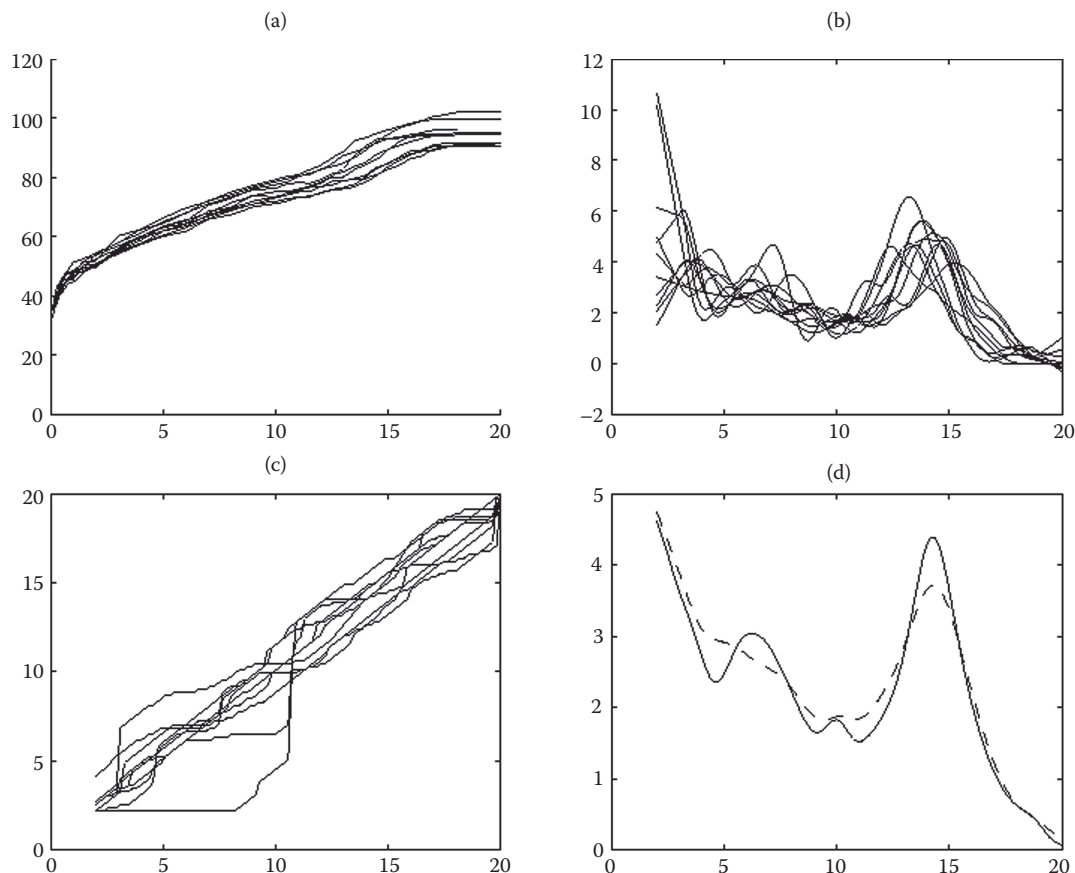


Fig. 3 Growth speed of shoulder width of boys. (a) shoulder width; (b) estimated individual growth speed; (c) estimated warping functions; (d) estimated average growth speed (solid line–aligned average; dashed line–average without alignment)

occurs at about age 13 for girls and at about age 15 for boys. These results agree with the findings of Gasser et al. ^[33,34]

APPLICATION TO BRAIN POTENTIALS

It is well known that event-related potentials obtained in repeated trials experience variable latencies of the components (for example, P3 and N1). The distortion is non-linear in the sense that the timing of features (say P1 N1 P2 N2 P3) in one trial cannot be linearly mapped to the timing of another trial. The conventional procedure for estimating the event-related potential is to average the repeated trials. The amplitude and latency variations among the single trials may lead to underestimating the amplitudes of the components (e.g., P3 and N1 of the average). The latency of a component of the average could be dictated by few trials with highest amplitudes.

To accurately estimate the event-related potential from these trials, the latency fluctuations have to be estimated first. Then the average of the aligned single trials is a good estimate of the event-related potential.^[35] An application to VP3 data is described below.

In a visual P3 (VP3) experiment, subjects are situated in front of a computer monitor and are presented with 280 visual stimuli with a uniform inter-stimulus interval of 1.6 seconds. There are 210 non-target stimuli in the shape of an outline of a square, 35 target stimuli in the shape of an X, and 35 novel stimuli, each a different colored polygon or other geometrical figure. The different types of stimuli are presented in random order. Subjects are instructed to respond to target stimuli by pressing a button. The VP3 data are recorded from 61 channels. After amplification by a factor of 10000, artifact threshold is set at 73.3 microvolts. Any trial with a value above the threshold is rejected. The number of the recorded single trials is different for different subjects and experimental conditions due to the artifact thresholding.

Fig. 5 plots the data from a subject and from one channel. There are 25 clean (no artifact) target trials recorded from the subject. The iterative algorithm of the section “Dynamic Time Warping” was applied to the target trials to obtain an estimate (warp-average) of the ERP from this channel. The result is plotted in Fig. 5. After the registration step, the structure of the brain potential comes out much more clearly.

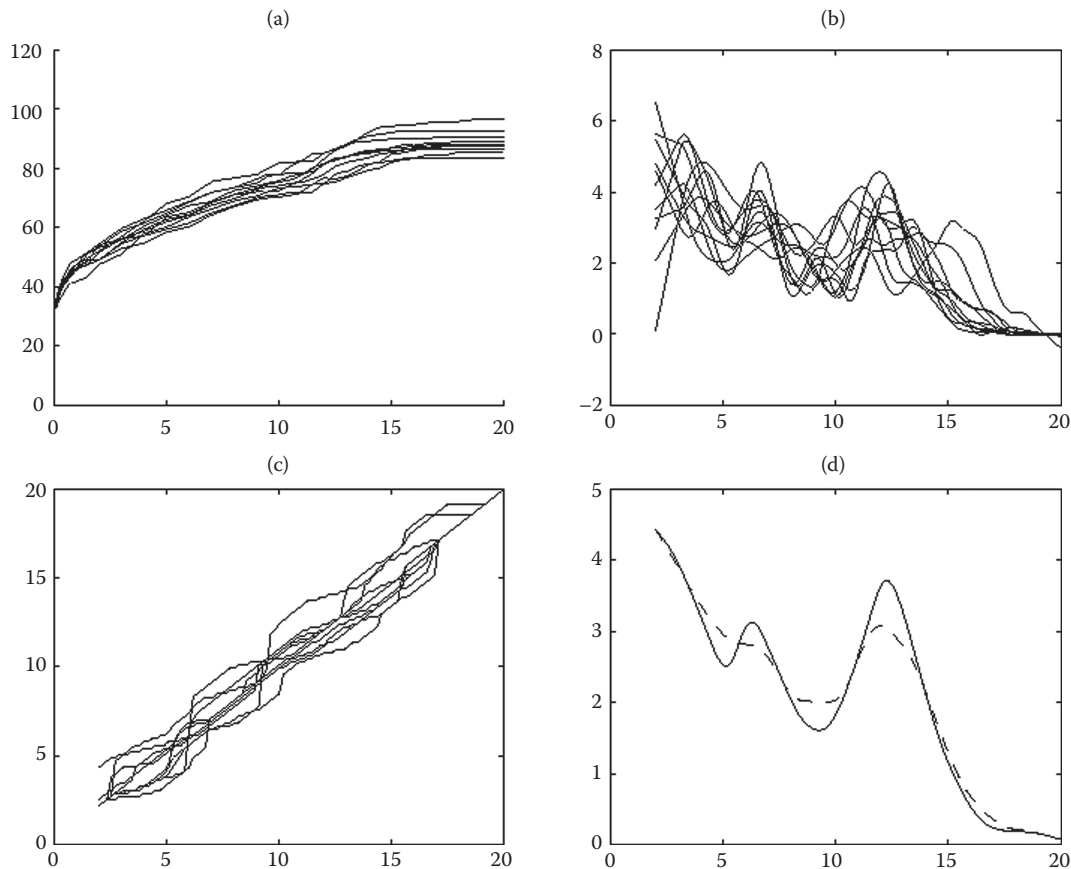


Fig. 4 Growth speed of shoulder width of girls. (a) shoulder width; (b) estimated individual growth speed; (c) estimated warping functions; (d) estimated average growth speed (solid line—aligned average; dashed line—average without alignment)

KERNEL DENSITY ESTIMATION AND MULTIVARIATE REGRESSION

We will discuss briefly non-parametric regression in the setting of density estimation and high-dimensional regression.

Density Estimation

Based on independent observations x_i , $i = 1, \dots, n$, from an unknown distribution, we want to estimate the underlying density function f . The kernel estimator^[36] is given by

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

where h is a smoothing parameter and K is a non-negative kernel that integrates to 1. The properties of the density estimator are similar to those of the kernel estimator in regression case. For example, $\hat{f}$ is as smooth as K and the smoothing parameter can be chosen via cross-validation or plug-in method. Simulation by Hanson^[37] showed that kernel density estimation is valid and as good as a semi-parametric model for modeling censored lifetime data with a mixture of Gammas baseline survival density.

Multivariate Regression

Multivariate regression problem can be stated as

$$y_i = r(x_{i1}, \dots, x_{id}) + \sigma \varepsilon_i$$

for d predictors and some unknown surface r . The random errors ε_i are assumed to be independently distributed with mean $E(\varepsilon_i) = 0$ and variance $V(\varepsilon_i) = 1$. The regression methods discussed in section “Linear Estimators” can be easily extended to multivariate case (e.g., see Muller^[38] for kernel method in multivariate case). Fan et al.^[15] have shown that the optimal kernel is the spherical Epanechnikov kernel

$$K(x_1, \dots, x_d) = \frac{d(d+2)}{2S_d} (1 - x_1^2 - \dots - x_d^2)_+$$

where $S_d = 2\pi^{d/2} / \Gamma(d/2)$ denotes the area of the surface of the d -dimensional unit ball.

The difficulty in multivariate non-parametric regression is that data are sparse (“urge of dimensionality”). Several approaches have been proposed to reduce the dimensionality. The additive modeling^[39] approximates a d -dimensional surface by sum of one-dimensional functions. The one-dimensional functions can be estimated by the non-parametric methods of the section “Linear

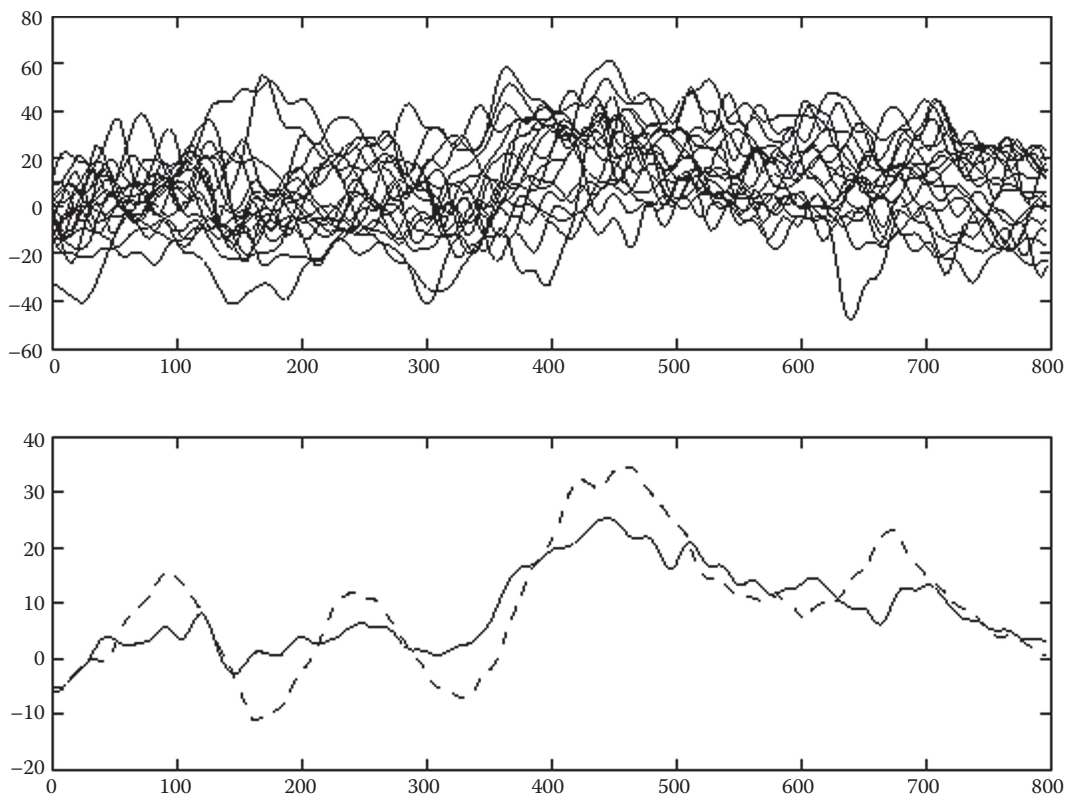


Fig. 5 Top panel—25 VP3 trials from a channel; Bottom panel—average with (dash) and without (solid) alignment. Vertical unit is in microvolts and horizontal unit in milliseconds

Estimators.” The projection pursuit method^[40] projects the d -dimensional predictive space into a space with smaller dimension (linear transformations of the predictors) and then uses additive modeling. The linear transformations and the one-dimensional functions of the additive model are estimated with an iterative algorithm.

SOFTWARE

Software for non-parametric regression is widely available. Any statistical package likely includes programs for non-parametric regression. In addition to the “official” software, one can go to a package’s website (SAS, MATLAB, Splus, R, etc.) for many user contributed software offerings.

The standard software in pharmaceutical industry is SAS, which includes several non-parametric procedures for functional estimation. SAS/INSIGHT software provides non-parametric curve-fitting estimates from smoothing spline, kernel, loess, and fixed bandwidth local polynomial estimators that are alternatives to fitting polynomials. Some specific SAS procedures are PROC LIFETEST for estimating survival curves with censoring data, PROC KDE for density estimation, PROC LOESS for estimating regression surfaces, and PROC TPSPLINE for using thin-plate smoothing splines to approximate smooth multivariate functions.

For functional data analysis, Ramsay and Silverman have developed software in R, Splus, and MATLAB, which are available at <ftp://ego.psych.mcgill.ca/pub/ramsay/FDAfuns/>. For computing a structural average from a sample of curves, Fortran codes for the iterative algorithm of the section “Dynamic Time Warping” are available at the website <http://www.unizh.ch/biostat/Software/>. The kernel smoothing software that we used is also available from this website.

CONCLUSION

Non-parametric methods allow more flexibility for the functional dependence of a response variable on predictors than a typical parametric model does, and these methods are well suited for situations where little is known about the process under study. The linear methods discussed in the section “Linear Estimators” perform well and have similar asymptotic properties. A comparison of kernel smoothing and wavelet shrinkage by Wang et al.^[41] has shown that both methods work well.

The crucial step in applying the non-parametric regression methods is to choose the smoothing parameter. Local bandwidth selection based on data (plug-in, cross-validation) is able to preserve fine structures of the regression function. Another major step is boundary correction for kernel estimators. Simulations by Hess et al.^[1] have confirmed the advantage of local bandwidth selection and boundary correction in estimating hazard functions.

We have applied the iterative algorithm based on dynamic time warping to two applications in order to construct a timing-aligned average curve from a sample of curves. The iterative algorithm either brings out new feature of the growth curves or enhances features of the brain potentials.

REFERENCES

1. Hess, K.; Serachitopol, D.; Brown, B. Hazard function estimation: a simulation study. *Stat. Med.* **1999**, *18*, 3075–3088.
2. Nadaraya, E. On estimating regression. *Theory Probab. Appl.* **1964**, *9*, 141–142.
3. Watson, G. Smooth regression analysis. *Sankhya Indian J. Stat. Ser. A* **1964**, *26*, 359–372.
4. Gasser, T.; Müller, H. Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*; Gasser, Th.; Rosenblatt, M.; Eds.; Lecture Notes in Mathematics, No. 757, Springer Verlag: Berlin, 1979, 23–68.
5. Gasser, T.; Müller, H. Estimating regression functions and their derivatives by the kernel method. *Scand. J. Stat.* **1984**, *11*, 171–185.
6. Gasser, T.; Müller, H.; Mammitzsch, V. Kernels for non-parametric curve estimation. *J. R. Stat. Soc. B* **1985**, *47*, 238–252.
7. Brockmann, M.; Gasser, T.; Herrmann, E. Locally adaptive bandwidth choice for kernel regression estimators. *J. Am. Stat. Assoc.* **1993**, *88*, 1302–1309.
8. Gasser, T.; Sroka, L.; Jennen-Steinmetz, C. Residual variance and residual pattern in nonlinear regression. *Biometrika* **1986**, *73*, 625–633.
9. Wang, K.; Gasser, T. Optimal rate for estimating local bandwidth in kernel estimators of regression functions. *Scand. J. Stat.* **1996**, *23*, 303–312.
10. Reinsch, C.M. Smoothing by spline functions. *Numer. Math.* **1967**, *10*, 177–183.
11. Silverman, B. Spline smoothing: the equivalent variable kernel method. *Ann. Stat.* **1984**, *12*, 898–916.
12. Kelly, C.; Rice, J. Monotone smoothing and its applications to dose-response curves and the assessment of synergy. *Biometrics* **1990**, *46*, 1071–1085.
13. DeBoor, C. *A Practical Guide to Splines*; Springer-Verlag: New York, 1978.
14. Fan, J. Local linear regression smoothers and their minimax efficiency. *Ann. Stat.* **1993**, *21*, 196–216.
15. Fan, J.; Gasser, T.; Gijbels, I.; Brockmann, M.; Engel, J. Local polynomial regression: optimal kernels and asymptotic minimax efficiency. *Ann. Inst. Stat. Math.* **1997**, *49* (1), 79–99.
16. Seifert, B.; Gasser, T. Variance properties of local polynomials and ensuing modifications. In *Statistical Theory and Computational Aspects of Smoothing, Contributions to Statistics*; Hardle, W.; Schimek, M.G.; Eds.; Physica-Verlag: Heidelberg, 1996, 50–79.
17. Daubechies, I. *Ten Lectures on Wavelets*; SIAM: Philadelphia, PA, 1992.
18. Meyer, Y. *Wavelets Algorithms & Applications*; SIAM: Philadelphia, PA, 1993.
19. Mallat, S. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern. Anal. Mach. Intell.* **1989**, *7*, 674–693.

20. Antoniadis, A.; Bigot, J.; Sapatinas, T. Wavelet estimators in non-parametric regression: a comparative simulation study. *J. Stat. Softw.* **2001**, *6*, 1–83.
21. Donoho, D.; Johnstone, I. Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **1994**, *81*, 425–455.
22. Gasser, T.; Gervini, D.; Molinari, L. Kernel estimation, shape-invariant modeling and structural analysis. In *Methods in Human Growth Research*; Hauspie, R.C.; Cameron, N.; Molinari, L., Eds.; Cambridge University Press: Cambridge, 2004.
23. Ramsay, J.; Silverman, B. *Functional Data Analysis*; Springer: New York, 2005.
24. Makela, T.; Clarysse, P.; Sipila, O.; Pauna, N.; Pham, C.; Katila, T.; Magnin, I. A review of cardiac image registration methods. *IEEE Trans. Med. Imaging* **2002**, *21*(9), 1011–1021.
25. Zitova, B.; Flusser, J. Image registration methods: a survey. *Image Version Comput.* **2003**, *21*, 977–1000.
26. Kneip, A.; Gasser, T. Statistical tools to analyze data representing a sample of curves. *Ann. Stat.* **1992**, *20*, 1266–1305.
27. Parsons, T. *Voice and Speech Processing*; McGraw-Hill: New York, 1986.
28. Rabiner, L.; Schmidt, C. Application of dynamic time warping to connected digit recognition. *IEEE Trans. ASSP* **1980**, *28*(4), 377–388.
29. Qi, Y. Time normalization in voice analysis. *J. Acoust. Soc. Am.* **1992**, *92*(5), 2569–2576.
30. Wang, K.; Gasser, T. Alignment of curves by dynamic time warping. *Ann. Stat.* **1997**, *25*, 1251–1276.
31. Wang, K.; Gasser, T. Synchronizing sample curves non-parametrically. *Ann. Stat.* **1999**, *27*, 439–460.
32. Wang, K.; Gasser, T. Asymptotic and bootstrap confidence bounds for the structural average of curves. *Ann. Stat.* **1998**, *26*, 972–991.
33. Gasser, T.; Kneip, A.; Ziegler, P.; Largo, R.; Molinari, L.; Prader, A. The dynamics of growth of width in distance, velocity and acceleration. *Ann. Hum. Biol.* **1991**, *18*, 449–461.
34. Gasser, T.; Ziegler, P.; Seifert, B.; Prader, A.; Molinari, L.; Largo, R. Measures of body mass and of obesity from infancy to adulthood and their appropriate transformation. *Ann. Hum. Biol.* **1994**, *21*, 111–125.
35. Wang, K.; Begleiter, H.; Projesz, B. Warp-averaging event-related potentials. *Clin. Neurophysiol.* **2001**, *112*, 1917–1924.
36. Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076.
37. Hanson, T. Modeling censored lifetime data using a mixture of gammas baseline. *Bayesian Anal.* **2006**, *1*, 575–594.
38. Muller, H.G. *Non-parametric Regression Analysis of Longitudinal Data. Lecture Notes in Statist*; Vol. 46; Springer: Berlin, 1988.
39. Stone, C.J. Additive regression and other non-parametric models. *Ann. Stat.* **1985**, *13*, 689–705.
40. Friedman, J.; Stutzle, W. Projection pursuit regression. *J. Am. Stat. Assoc.* **1981**, *76*, 817–823.
41. Wang, K.; Seifert, B.; Gasser, T. Discussion of “Wavelet shrinkage: asymptopia?” by Donoho et al. *J. R. Stat. Soc. B* **1995**, *57*, 341–343.

Odds Ratio

Dongsheng Tu

NCIC Clinical Trials Group, Queen's University, Kingston, Ontario, Canada

Abstract

The objective of this entry is to introduce the definition of the odds ratio and its statistical inferences based on data from randomized clinical trials. Some other topics related to odds ratio will also be discussed.

Keywords: Inference; Odds ratio; Regression.

INTRODUCTION

In epidemiological literature, odds ratio has been frequently used to assess the strength of the association between a binary exposure variable and a binary disease outcome. It was first introduced by Cornfield^[1] as a measure of relative risk in retrospective studies, where the other more intuitive measure, the rate ratio, cannot be estimated. When the baseline risk of the disease is low, the odds ratio provides a good approximation to the risk ratio. The odds ratio has also some mathematical advantages: It takes values between 0 and infinity, which is not the case for rate ratio; if the outcome in the analysis is reversed from positive to negative, we get a reciprocal odds ratio. Its other statistical advantages include that it derives naturally from the logistical regression models and that it arises as a natural parameter in a conditional distribution that is the base for the exact inference.^[2,3]

For a clinical trial with a binary outcome as the primary endpoint, such as whether the symptom of a patient has disappeared after a certain period of treatment, the difference between two treatment arms is usually measured by the difference in the rates for the outcome of interest. The ratio of these two rates is also often calculated, which provides a measure for the relative effect of the treatments. Recently, odds ratio has also appeared in clinical trial literature as an alternative measure of relative treatment effects (e.g., see Ref. [4]). This is especially true when a meta-analysis of several clinical trials is involved. The objective of this entry is to introduce the definition of the odds ratio and its statistical inferences based on data from randomized clinical trials. Some other topics related to odds ratio will also be discussed.

DEFINITION

Let P_T be the probability of having an outcome of interest, which can be either positive (e.g., the patient responded to a treatment) or negative (e.g., the disease of the patient progressed), for a patient treated by a test treatment (a new drug or procedure) and let P_C be the probability for

a patient treated by a control, which may be a standard treatment or a placebo. The odds that a patient will have an outcome of interest instead of not have an outcome of interest when one has received a test treatment is $O_T = P_T/(1 - P_T)$. The corresponding odds can be similarly defined for the patients treated by a control. Then the ratio of these two odds is defined as the odds ratio for the test group vs. control group, which can be written as: $\psi = [P_T(1 - P_C)]/[P_C(1 - P_T)]$. The odds ratio is always positive and has a range from 0 to infinity. $\psi = 1$ implies that there is no difference between two treatment groups in terms of the outcome of interest (i.e., $P_T = P_C$). When $\psi > 1$, patients in the test group will be more likely to have the outcome of interest than those in the control group. The converse is true when $0 < \psi < 1$. When the outcome of interest is adverse (e.g., disease progress or death), $1 - \psi$ is often defined as relative odds reduction.

Suppose, in a randomized clinical trial, that n_T patients were assigned to receive the test treatment and that n_C patients were assigned as the control. Let X_T and X_C be the observed numbers of patients in the test group and control group, respectively, who had outcome of interest. Because X_T and X_C are independent, the likelihood of these observations is a product binomial:

$$L(P_T, P_C | X_T = x_T, X_C = x_C) = \binom{n_T}{x_T} P_T^{x_T} (1 - P_T)^{n_T - x_T} \binom{n_C}{x_C} P_C^{x_C} (1 - P_C)^{n_C - x_C}$$

The natural estimates, which are also maximum likelihood estimates, for P_T and P_C , respectively, are:

$$\hat{P}_T = \frac{X_T}{n_T} \quad \text{and} \quad \hat{P}_C = \frac{X_C}{n_C}$$

Therefore a point estimate for ψ , which is also a maximum likelihood estimate and called also just as odds ratio, can be defined as:

$$\hat{\psi} = \frac{\hat{P}_T(1 - \hat{P}_C)}{(1 - \hat{P}_T)\hat{P}_C}$$

ASYMPTOTIC INFERENCES

Haldane^[5] derived the asymptotic variance of $\hat{\psi}$, which would be used to construct an asymptotic $100(1 - \alpha)\%$ confidence interval for ψ . But because the estimate of the odds ratio is always positive, its distribution is skewed and therefore a normal approximation might be very inaccurate. In this case, it is a common statistical practice to make first a transformation, such as logarithm transformation, to the estimate and then to derive a confidence interval based on transformed estimators. Form the results of Haldane,^[5] the asymptotic variance for $\ln(\hat{\psi})$ can be derived as:

$$v[\ln(\hat{\psi})] = \frac{1}{n_T P_T (1 - P_T)} + \frac{1}{n_C P_C (1 - P_C)}$$

It can be estimated by:

$$\hat{v}[\ln(\hat{\psi})] = \frac{1}{n_T \hat{P}_T (1 - \hat{P}_T)} + \frac{1}{n_C \hat{P}_C (1 - \hat{P}_C)}$$

Therefore an asymptotic $100(1 - \alpha)\%$ confidence interval based on a logarithm transformation for ψ is: $\exp(\ln(\hat{\psi}) \pm z_{1-\alpha/2} \sqrt{\hat{v}[\ln(\hat{\psi})]})$ where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ percentile of a standard normal distribution. This confidence interval is sometimes called the logit confidence interval after it was first derived by Woolf.^[6] It is widely recommended as a satisfactory approximation except when any one of X_T and X_C is too small or too close to n_T and n_C , respectively. In the later case, exact methods that will be introduced in “Exact Inferences” would be used. In the extreme cases when one of X_T , X_C , $n_T - X_T$, $n_C - X_C$ is 0, Haldane^[5] and Anscombe^[7] suggested to add 0.5 to each of these numbers and to use the following estimator, which is sometimes called the Haldane-Anscombe 1/2 correction estimator of ψ :

$$\hat{\psi}_1 = \frac{\left(\hat{P}_T + \frac{0.5}{n_T} \right) \left(1 - \hat{P}_C + \frac{0.5}{n_C} \right)}{\left(1 - \hat{P}_T + \frac{0.5}{n_T} \right) \left(\hat{P}_C + \frac{0.5}{n_C} \right)}$$

The associated $100(1 - \alpha)\%$ confidence interval for ψ is: $\exp(\ln(\hat{\psi}_1) \pm z_{1-\alpha/2} \sqrt{\hat{v}[\ln(\hat{\psi}_1)]})$ where

$$\begin{aligned} \hat{v}[\ln(\hat{\psi}_1)] = & \frac{1 + \frac{1}{n_T}}{n_T \left(\hat{P}_T + \frac{0.5}{n_T} \right) \left(1 - \hat{P}_T + \frac{0.5}{n_T} \right)} \\ & + \frac{1 + \frac{1}{n_C}}{n_C \left(\hat{P}_C + \frac{0.5}{n_C} \right) \left(1 - \hat{P}_C + \frac{0.5}{n_C} \right)} \end{aligned}$$

Gart and Zweifel^[8] showed that $\hat{\psi}_1$ and its variance estimator have smaller asymptotic bias and mean square error in comparison to $\hat{\psi}$ and its variance estimator even when none of X_T , X_C , $n_T - X_T$, $n_C - X_C$ is 0. Some authors,

such as Mantel,^[9] prefer not to add the 0.5 correction factor. Instead, if $\hat{\psi}$ is infinite for a particular dataset, they suggested that a lower confidence limit for ψ be calculated using an exact method, which will be introduced in “Exact Inferences.”

The above methods can also be used to test hypotheses about ψ . Because $H_0: \psi = 1$ is equivalent to $H_0: P_T = P_C$ and $H_a: \psi > (<)1$ is equivalent to $H_a: P_T > (<)P_C$, conventional statistical tests, such as Pearson's chi-square test or Mantel-Haenszel test, for the association in a 2×2 table can also be applied to test hypotheses about P . The details can be found in standard textbooks such as that of Lachin.^[3]

It should be noted that these conventional tests might yield different results from that derived from the confidence intervals introduced before. For example, a 95% confidence interval for ψ may cover 1 when the p value of the Person's chi-square test is less than 0.05. If it is desired to include both results from the hypothesis testing and confidence interval, a test-based confidence interval may be preferred. Let T be a test statistic for the hypotheses $H_0: \psi = 1$ vs. $H_a: \psi \neq 1$. Then an asymptotic $100(1 - \alpha)\%$ test-based confidence interval for ψ is:

$$\exp\left(\ln(\hat{\psi}) \pm \frac{z_{1-\alpha/2} \ln(\hat{\psi})}{\sqrt{T}}\right)$$

A short summary of the results involving comparisons between the logit and the test-based confidence interval can be found in Section 3.1.4 of Sahai and Khurshid.^[10]

EXACT INFERENCES

The exact confidence interval for ψ is based on the conditional distribution of X_T and X_C given $m = X_T + X_C$, which can be written as:

$$\begin{aligned} P(X_T = x_T | m = x_T + x_C, \psi) \\ = \frac{\binom{n_T}{x_T} \binom{n_C}{m - x_T} \psi^{x_T}}{\sum_{i=\alpha_1}^{\alpha_u} \binom{n_T}{i} \binom{n_C}{m - i} \psi^i} \end{aligned}$$

where $\alpha_1 = \max(0, m - n_C)$ and $\alpha_u = \min(m, n_T)$. This conditional distribution depends only on ψ . Therefore an exact $100(1 - \alpha)\%$ confidence interval $(\hat{\psi}_l, \hat{\psi}_u)$ for ψ can be obtained by solving the equations:

$$\frac{\alpha}{2} = \sum_{x=x_T}^{\alpha_u} P(x | m, \hat{\psi}_l) \quad \text{and} \quad \frac{\alpha}{2} = \sum_{x=\alpha_1}^{x_T} P(x | m, \hat{\psi}_u)$$

It may be noted that when we use this confidence interval to describe the range of the difference between two treatment groups, for consistency, ψ should also be estimated based on the conditional distribution, that is, by maximizing the conditional likelihood defined by

$L(\psi|x_T, m) = p(x_T|m, \psi)$. This conditional maximum likelihood estimator may be different from $\hat{\psi}$, which is the maximum likelihood estimator based on product binomial distribution.

We can also construct an exact test based on this conditional distribution. This is usually called the Fisher's exact test. For one-sided hypotheses $H_0: \psi = 1$ vs. $H_a: \psi > 1$ or $H_a: \psi < 1$, the p values, respectively, of Fisher's exact test is defined as:

$$P_u = \sum_{x=x_T}^{a_U} P(x|m, \psi=1) \quad \text{or} \\ p_l = \sum_{x=a_L}^{x_T} P(x|m, \psi=1)$$

The p value of the Fisher's exact test for two-sided hypotheses $H_0: \psi = 1$ vs. $H_a: \psi \neq 1$ is defined as:

$$p = \sum_{x=a_L}^{a_U} \left\{ I \left[P(x|m, \psi=1) \leq P(x_T|m, \psi=1) \right] \times P(x|m, \psi=1) \right\}$$

where $I(\cdot)$ is the indicator function.

For randomized clinical trials, some authors argued that it might be artificial to assume that m is fixed and to use the test based on the conditional distribution. Others argued that conditioning on a sufficient statistic to eliminate nuisance parameters is an appropriate statistical practice no matter if m can be assumed fixed. Besides these philosophic differences, Agresti^[11] pointed out that the root of most objections to conditional approach is their conservatism as a result of discreteness of the distribution. Some unconditional exact tests were proposed. The p values of these tests are defined as:

$$p = \sup_{0 \leq \rho \leq 1} P(T \geq t | P_T = P_C = \rho)$$

where T is a given test statistic and t is the observed value of T . For example, Suissa and Shuster^[12] derived an unconditional exact test by taking T as the unpooled z test statistic. There are still debates on which method is more suitable. Some compromises have been proposed such as to adjust exact methods based on mid p values (see a review by Berry and Armitage^[13]). More discussions about this topic can be found in Agresti.^[11]

NUMERICAL EXAMPLE

A randomized clinical trial has been conducted by the AIDS Clinical Trials Group to study the safety and efficacy of thalidomide for the treatment of esophageal aphthous ulcers in patients infected with HIV.^[4] Eleven patients were randomized to receive thalidomide (test treatment) and 13 were randomized to receive placebo (control).

The primary outcome of interest was a complete healing of aphthous ulcers at the 4-week endoscopic evaluation. Eight patients in the thalidomide group and three patients in the placebo group had outcome of interest. The odds ratio of test vs. control group is 8.89. The associated 95% legit confidence interval is (1.40, 57.57). The Haldane-Anscombe 1/2 correction estimate of the odds ratio is 7.28 with a 95% confidence interval from 1.28 to 41.34. The chi-square test statistic for $H_0: \psi = 1$ vs. $H_a: \psi \neq 1$ is 5.92 with a p value of 0.015. A 95% confidence interval based on the chi-square test is (1.53, 51.66), which is farther away from 1 than the legit interval.

Because the numbers of patients in both groups in this example are small, an exact inference may be preferred. The odds ratio estimate based on the conditional distribution is 7.95. An exact 95% confidence interval based on the conditional distribution is (1.06, 84.35). The p value of the two-sided Fisher's exact test is 0.038, which is more than double that of chi-square test.

RELATIONSHIP WITH OTHER MEASURES

The relationship of odds ratio with other measures, such as the difference or the ratio of the probabilities of having an outcome of interest between two groups, is nonlinear. Let $D = P_T - P_C$ be the difference of these probabilities. It is easy to see when P_C is known:

$$\psi = \frac{(P_C + D)(1 - P_C)}{P_C(1 - P_C - D)} = 1 + \frac{1}{\frac{P_C(1 - P_C)}{D} - P_C}$$

It is an increased function of D . For the ratio of the probabilities, denoted by R , we have:

$$\psi = \frac{(P_C R)(1 - P_C)}{P_C(1 - P_C R)} = R \left(\frac{1 - P_C}{1 - P_C R} \right) \text{ or } R = \frac{\psi}{1 + P_C(\psi - 1)}$$

ψ is close to R whenever P_C and P_T are small (i.e., the probability to have an outcome of interest is rare in both treatment groups). But when $\psi > 1$, we have always $\psi > R$. The difference between ψ and R can be very large when P_C is very large. When $\psi < 1$, we have $\psi < R$. In this case, ψ is farther away from 1 than R when P_C becomes very large.

In the example introduced in "Numerical Example," the estimates of D and R are 0.50 and 3.15, respectively. Because both the estimates of P_C and P_T are large (0.23 and 0.73, respectively), there is a big difference between the estimates of ψ and R .

For a single study, all three measures would be a choice for the measurement of the difference between groups when it exists. Sackett and Cook^[14] recommended that one absolute measure of efficacy (such as rate difference or its reciprocal, which is called the number of patients needed to be treated) and some relative measure of efficacy be included in the description of efficacy

results from a randomized clinical trial. Both rate ratio and odds ratio would be used as a measure of relative efficacy. Sinclair and Bracken^[15] argued that rate ratio might be more appealing to clinicians. As pointed out before, the odds ratio has some mathematical and statistical advantages, which may be the reasons they continue to be quoted in clinical literature.^[2] For a single study with moderate or large sample size, all three measures will likely reach similar conclusions. However, when there are multiple 2×2 tables such as from multiple strata when we perform a stratified analysis in a single study or a meta-analysis of many different studies, the conclusions reached may depend on the measure chosen to summarize the results.^[3]

OTHER RELATED TOPICS

Sample Size Calculations

Wang et al.^[16] presented sample size calculation formulas for clinical trials with odds ratio as primary endpoint under various designs and types of hypotheses. Specifically, assume that one is interested in testing hypotheses given by $H_0: \psi = 1$ vs. $H_a: \psi \neq 1$ in a two group parallel design with 1 to 1 randomization. Then the sample size required in each group to detect, with power $1 - \beta$ and two-sided significance level α , a difference between two groups as measured by an odds ratio ψ is:

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\ln^2(\psi)} \left[\frac{1}{P_T(1-P_T)} + \frac{1}{P_C(1-P_C)} \right].$$

Adjusted Odds Ratio in Logistic Regression Models

Although there are still some controversies, the treatment effect in clinical trials is often adjusted for some prognostic factors or other baseline variables that are not balanced by randomization. When the endpoint of a trial is binary, a logistic regression model is often used to perform this adjustment. Assume that Y ($= 1$ or 0) is the binary outcome and $Y = 1$ implies the outcome of interest. Let $P = P(Y = 1)$ be the probability of having an outcome of interest for a patient in the study (no matter which group one was assigned to) and Z is an indicator variable that takes a value of 1 if this patient is assigned to be treated by the new treatment and 0 otherwise. Assume that $X_1, X_2, \dots, X_K$ are the K prognostic factors or covariates for which we would like to adjust for. A logistic regression model is defined as:

$$\ln\left(\frac{P}{1-P}\right) = \alpha_0 + \beta Z + \delta_1 X_1 + \dots + \delta_K X_K$$

where $\alpha_0, \beta, \delta_1, \dots, \delta_K$ are unknown parameters. It is easy to show that:

$$\exp(\beta) = \frac{\frac{P(Y=1|Z=1, X_1, \dots, X_K)}{P(Y=0|Z=1, X_1, \dots, X_K)}}{\frac{P(Y=1|Z=0, X_1, \dots, X_K)}{P(Y=0|Z=0, X_1, \dots, X_K)}}$$

which is the ratio of the odds of having an outcome of interest instead of not having an outcome of interest when treated by a test procedure to that when treated by a control after controlling for the effects of all other covariates. $\exp(\beta)$ is usually called the adjusted odds ratio. It can be estimated by using maximum likelihood methods or some exact methods based on conditional distributions. The details can be found in Chapter 4 of Agresti.^[17]

Use of the Odds Ratio in Therapeutic Equivalence Clinical Trials

In a therapeutic equivalence trial with a binary end point, we are interested in testing whether two treatments have the same efficacy or the test treatment is not inferior to a control. We consider here only the former case. Tu^[18,19] compared some statistical procedures when the difference of outcome rates is used as a measurement of efficacy difference between two treatments. In this case, there is some difficulty in determining the acceptance limits (e.g., see Weng and Liu^[20] and Bristol^[21] for a detailed discussion). Because of this, Weng and Liu^[20] suggested that some other statistical procedures, such as those based on the odds ratio, might be more appropriate. Tu^[22] discussed some issues related to the use of the odds ratio in establishing therapeutic equivalence with binary end points. In this case, the interval hypotheses for the therapeutic equivalence of test and control can be formulated as follows: $H_0: \psi \leq \xi_L$ or $\psi \geq \xi_U$ vs. $H_1: \xi_L < \psi < \xi_U$ where ξ_L and ξ_U are prespecified lower and upper acceptance limits, respectively. A one-sided logit procedure for the testing of these interval hypotheses is to reject H_0 at the α level if $T_{LG} > z_\alpha$ and $T_{UG} < -z_\alpha$, where:

$$T_{LG} = \frac{\ln(\hat{\psi}_1) - \ln(\xi_L)}{\sqrt{\hat{v}[\ln(\hat{\psi}_1)]}} \quad \text{and} \quad T_{UG} = \frac{\ln(\hat{\psi}_1) - \ln(\xi_U)}{\sqrt{\hat{v}[\ln(\hat{\psi}_1)]}}$$

Tu^[22] evaluated some statistical properties of this and some more complicated procedures. It was found out that if we set up the acceptance interval as (0.8, 1.25), which was used for the assessment of bioequivalence based on the ratio of the means of two normal distributions, the power of the logit procedures will be very small even

when $\psi = 1$. This means that a huge sample size (on the order of thousands) would be required if we want the test to have an 80% power. By comparing the acceptance interval with that suggested for the rate difference, it was found that the interval (0.43, 2.33) might be a reasonable alternative choice. According to the guidelines for the interpretation of the values of odds ratio in Greenberg,^[23] when the true odds ratio is covered by this interval, the difference between two groups is moderate at best while true odds ratio falling into (0.8, 1.25) implies no difference. When this acceptance interval is used, the procedures based on the odds ratio have the following advantage over the procedures based on the difference: They automatically make the acceptance interval narrow when the cure rates of the drugs are large. For example, when $P_R = 0.5$, $0.43 \leq \psi \leq 2.33$ implies $-0.2 \leq P_T - P_R \leq 0.2$; and when $P_R = 0.9$, $0.43 \leq \psi \leq 2.33$ implies $-0.1 \leq P_T - P_R \leq 0.05$.

Combining Odds Ratios in Stratified or Meta-Analyses

A study with a binary endpoint may be stratified based on some baseline patient characteristics and a stratified analysis by combining treatment comparisons in the different strata may be desired. Or there were several studies that used the same binary endpoint to evaluate the same treatments and we would like to perform a meta-analysis to combine evidences about the treatment effect from these different studies. Assume that treatment differences in all strata or studies are assessed by odds ratios that are the same across all the strata or studies. An asymptotic method by Mantel and Haenszel^[24] would be used to test whether this common odds ratio is one or, together with a variance estimator proposed by Robins et al.,^[25] to construct an asymptotic confidence interval for the common odds ratio. An exact confidence interval can be calculated based on a method proposed by Mehta et al.^[26] In meta-analyses, the homogeneity of the odds ratios would be tested by an exact test proposed by Zelen^[27] or an asymptotic test proposed by Breslow and Day.^[28] Some other methods for the assessment of odds ratio homogeneity or the inference on the common odds ratio can be found in Chapter 4 of Lachin.^[3] When there is some heterogeneity in odds ratio, a random effect model method by DerSimonian and Laird^[29] would be used to obtain an overall assessment of the treatment effect.

Local and Cumulative Odds Ratio in $2 \times K$ Tables

In some clinical trials, the endpoint may have more than two categories. For example, in an FDA draft guideline for the performance of a bioequivalence study for tretinoin products, it was recommended that the overall improvement in the patient's acne be evaluated globally on the following scale: 1 = *cleared or marked improvement*: between 76% and 100% clearance of acne lesions;

2 = *moderate improvement*: between 50% and 75% clearance of acne lesions; 3 = *slight improvement*: less than 50% clearance of acne lesions; and 4 = *no change or worse*.^[30] In general, assume that the endpoint of a trial has K ordered categories and let π_{Tj} be the probability that a patient treated by the test treatment has a response in category j and let π_{Cj} be the corresponding probability for a patient treated by the control. The relative difference between the test treatment and the control can be described by $k - 1$ local odds ratio $\phi_j = (\pi_{Rj}\pi_{T(j+1)})/(B_{Tj}\pi_{R(j+1)})$, $j = 1, 2, \dots, k - 1$. Whether these two treatments are different would be assessed by testing the null hypothesis $H_0: \phi_1 = \dots = \phi_{k-1} = 1$. When the categories are not ordered, Pearson's chi-square test or Fisher's exact test can be used to test H_0 . If there is a natural ordering, the asymptotic or exact distribution of a Wilcoxon midrank test statistic can be used to perform an asymptotic or an exact test, respectively, for the above hypothesis.^[31,32] Another way to measure the difference between two treatments when endpoint is ordinal categorically is to use the so-called cumulative odds ratios $\zeta_j = [Q_{Tj}(1 - Q_{Cj})]/[(1 - Q_{Tj})Q_{Cj}]$, where $Q_{Tj} = \pi_{T1} + \dots + \pi_{Tj}$, $j = 2, \dots, k$, and Q_{Cj} is defined similarly. Assume that all ζ_j are the same (so-called proportional odds model). Jones and Whitehead^[33] derived a score test for the null hypothesis that the common cumulative odds ratio is 1. Both measures were used by Tu^[34] to test an interval hypothesis in therapeutic equivalence trials with ordinal categorical endpoints.

CONCLUSION

Although there has been continuous debate on its interpretations in clinical trials,^[35] the odds ratio is a useful measure of relative treatment effects in clinical trials with binary or categorical end points because of its many attractive mathematical and statistical properties. It should not, however, be confused with the rate ratio—the other measure of relative treatment effect. They may be quite different when the rate of the control group is not too small, as is the case in many clinical trials.

REFERENCES

1. Cornfield, J. A Statistical Problem Arising from Retrospective Studies. In *Proceedings of the Third Berkeley Symposium on Math. Statist. and Probab*; Neyman, J.; Ed.; University of California: Berkeley, 1956, 135–148.
2. Deeks, J. Swots corner: what is an odds ratio? *Bandolier* **1996**, 25, 6–9.
3. Lachin, J.M. *Biostatistical Methods: The Assessment of Relative Risk*; John Wiley & Sons: New York, 2000.
4. Jacobson, J.M.; Spritzler, J.; Fox, L.; Fahey, J.L.; Jackson, J.B.; Chernoff, M.; Wohl, D.A.; Wu, A.W.; Hooton, T.M.; Sha, B.E.; Shikuma, C.M.; MacPhail, L.A.; Simpson, D.M.; Trapnell, C.B.; Basgoz, N. Thalidomide for the

- treatment of esophageal aphthous ulcers in patients with human immunodeficiency virus infection. *J. Infect. Dis.* **1999**, *180*, 61–67.
5. Haldane, J.B.S. The estimation and significance of the logarithm of a ratio of frequencies. *Ann. Hum. Genet.* **1995**, *20*, 309–311.
 6. Woolf, B. On estimating the relation between blood group and disease. *Ann. Hum. Genet. (London)* **1955**, *19*, 251–253.
 7. Anscombe, F.J. On estimating binomial response relations. *Biometrika* **1956**, *43*, 461–464.
 8. Gart, J.J.; Zweifel, J.R. In the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika* **1967**, *54*, 181–187.
 9. Mantel, N. Biased inferences in tests for average partial association. *Am. Stat.* **1980**, *34*, 190–191.
 10. Sahai, H.; Khurshid, A. *Statistics in Epidemiology: Methods, Techniques and Applications*; CRC Press: Boca Raton, 1996.
 11. Agresti, A. Exact inference for categorical data: recent advances and continuing controversies. *Stat. Med.* **2001**, *20*, 2709–2722.
 12. Suissa, S.; Shuster, J.J. Exact unconditional sample sizes for the 2 by 2 binomial trial. *J. R. Stat. Soc., Ser A* **1985**, *148*, 317–327.
 13. Berry, G.; Armitage, P. Mid-p confidence intervals: a brief review. *Statistician* **1995**, *44*, 417–423.
 14. Sackett, D.L.; Cook, R.J. Understanding clinical trials. *BMJ* **1994**, *309*, 755–756.
 15. Sinclair, J.C.; Bracken, M.B. Clinically useful measures of effect in binary analyses of randomized trials. *J. Clin. Epidemiol.* **1994**, *87*, 881–889.
 16. Wang, H.; Chow, S.C.; Li, G. On sample size calculation based on odds ratio in clinical trials. *J. Biopharm. Stat.* **2002**, *12*, 471–483.
 17. Agresti, A. *Categorical Data Analysis*; Wiley: New York, 1990.
 18. Tu, D. A comparative study of some statistical procedures in establishing therapeutic equivalence of non-systemic drugs. *Drug Inf. J.* **1997**, *31*, 291–1300.
 19. Tu, D. Two one-sided tests procedures in establishing therapeutic equivalence of non-systemic drugs with binary clinical endpoints: fixed sample performances and sample size determination. *J. Stat. Comput. Simul.* **1997**, *59*, 190–271.
 20. Weng, C.S.; Liu, J.P. Some Pitfalls in Sample Size Estimation for Anti-Infective Study. In *American Statistical Association, Proceedings of Biopharmaceutical Section of the American Statistical Association*, Alexandria, VA 1994, 56–60.
 21. Bristol, D.R. Determining equivalence and impact of sample size in anti-infective studies: a point to consider. *J. Biopharm. Stat.* **1996**, *6*, 319–326.
 22. Tu, D. On the use of the ratio or the odds ratio of cure rates in establishing therapeutic equivalence of non-systemic drugs with binary clinical endpoints. *J. Biopharm. Stat.* **1998**, *8*, 263–282.
 23. Greenberg, R.S. Prospective Studies. In *Encyclopedia of Statistical Sciences*; Wiley: New York, 1986, Vol. 7, 315–319.
 24. Mantel, N.; Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **1959**, *22*, 719–748.
 25. Robins, J.; Breslow, N.; Greenland, S. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* **1986**, *42*, 311–323.
 26. Mehta, C.R.; Patel, N.R.; Gray, R. Computing an exact confidence interval for the common odds ratio in several 2 by 2 contingency tables. *J. Am. Stat. Assoc.* **1985**, *80*, 969–973.
 27. Zelen, M. Exact Significance Tests for Contingency Tables Embedded in a 2nd Classification. In *Proceedings of the Sixth Berkeley Symposium on Math. Statist. Probab.*; LeCam, L.; Neyman, J.; Scott, E.L., Eds.; University of California Press: Berkeley, 1972, 737–757.
 28. Breslow, N.E.; Day, N.E. *Statistical Methods in Cancer Research: Vol. I. The Analysis of Case-Control Studies*; International Agency on Cancer: Lyon, 1980.
 29. DerSimonian, R.; Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* **1986**, *7*, 177–188.
 30. FDA. *Proposed Draft Guidance for the Performance of a Bioequivalence Study for Tretinoin Products*; Centre for Drug Evaluation and Research, Food and Drug Administration: Rockville, MD, 1990.
 31. Lehmann, E.L. *Nonparametrics: Statistical Methods Based on Ranks*; Holden-Day, Inc.: San Francisco, CA, 1975.
 32. Mehta, C.R.; Patel, N.R.; Tsiatis, A.A. Exact significance testing to establish treatment equivalence with ordered categorical data. *Biometrics* **1984**, *40*, 819–825.
 33. Jones, D.R.; Whitehead, J. Sequential forms of the log rank and modified Wilcoxon tests for censored data. *Biometrika* **1979**, *66*, 105–113.
 34. Tu, D. Statistical procedures in therapeutic equivalence clinical trials with ordinal categorical endpoints. *J. Stat. Res.* **2001**, *35*, 65–77.
 35. Attig, R.B.; Clabaugh, A. Clinical trails and statistical tribulations. *Appl. Clin. Trails* **2008**, *17* (2), 42–48.

Oncology Clinical Trials: Efficient Dose-Finding Designs

Erik Rasmussen, Qi Jiang, and Chunlei Ke

Amgen, Thousand Oaks, California, U.S.A.

Abstract

Dose finding for oncology therapeutics remains a challenge in drug development. The past decades have seen many methodologic advances for dose finding in oncology drug development, but implementation of these new methods has been less than ideal, as the historic 3+3 design remains the most commonly implemented design. This chapter provides an overview of dose-finding methodologies in oncology trials. We focus on the commonly implemented designs and designs that if implemented more frequently would likely improve oncology dose finding. Designs summarized in detail include the 3+3 design, Bayesian logistic regression model, and modified toxicity probability interval design; each of these designs identifies the maximum tolerated dose. Descriptions are also provided for the Phase I/II design to use both toxicity and efficacy to identify a dose with maximum benefit–risk and the partial order continual reassessment method design to identify a maximum tolerated combination of two or more oncology therapeutics. Additionally, we provide practical recommendations on when to implement these methodologies to possibly improve oncology dose finding. Finally, recent contributions to dose-finding methodologies in oncology clinical trials are summarized.

Keywords: Bayesian logistic regression; BLRM; Continual reassessment method; CRM; Dose finding; Maximum tolerated dose; MTD; Phase I.

INTRODUCTION

Researchers and developers of oncology therapeutics have had many successes over the years. Five-year survival rates have improved meaningfully for many tumor types. Compared to the estimated 5-year survival rate in 1975–1977, the 5-year survival rate from 2003 to 2009 increased as follows^[1]:

- *Multiple myeloma*: It improved 5-year survival rate from 25% to 45%.
- *Non-Hodgkin's lymphoma*: It improved 5-year survival rate from 47% to 71%.
- *Prostate*: It improved 5-year survival rate from 68% to 100%.

Despite these successes, oncology drug development has seen many drug development failures. The Biotechnology Innovation Organization estimates that between 2006 and 2015 only 40% of oncology Phase III studies were positive.^[2] Suboptimal dosing is one possible cause of Phase III failures. Dose finding involves testing different dose levels of a drug in order to characterize the benefit–risk profile with the goal of identifying a dose with an optimal benefit–risk profile. The ideal benefit–risk profile would be

a 100% cure rate with no safety issues. In reality, the ideal scenario never occurs and dose finding is done to balance the trade-off between efficacy and toxicity. Many effective dose-finding methods have been developed that, if implemented more frequently, would likely improve the success rate in oncology drug development.

The primary purpose of this article is to provide an overview of dose-finding methodologies in oncology trials and to provide recommendations for clinicians and statisticians working in dose-finding oncology trials. The overview of dose-finding methodologies is summarized by dose-finding objective and class of design. For illustration purposes, the most frequently used design in a class, if applicable, is described in detail with other designs in the class referenced. Additionally, recent contributions to dose-finding methodologies in oncology clinical trials are introduced and referenced. The remainder of this article is structured as follows: designs to identify monotherapy maximum tolerated dose are reviewed in Section “Designs to Identify Monotherapy Maximum Tolerated Dose,” designs to identify the recommend Phase II dose in Section “Designs to Identify the Recommended Phase II Dose,” and designs to identify the combination maximum tolerated dose in Section “Designs to Identify the Combination MTD.” Recommendations regarding the use of these

designs will also be provided for each section. Finally, recent trends in dose finding are discussed in Section “Recent Trends in Dose Finding.”

DESIGNS TO IDENTIFY MONOTHERAPY MAXIMUM TOLERATED DOSE

An investigational new drug needs to obtain safety data and demonstrate safety as a monotherapy treatment in a first-in-human (FIH) study before subjects in additional clinical trials can be enrolled. The primary objective of the FIH study is usually to estimate the MTD. Various clinical trial endpoints are used to identify the MTD. The most commonly used endpoint is dose-limiting toxicity (DLT); DLT encompasses a range of adverse events that are determined to be severe enough to warrant discontinuation of dosing. Though infrequently used, ordinal toxicity and time to toxicity event are clinical endpoints that incorporate additional clinically relevant information into the dose-finding process. Occasionally, both the dose level and the dosing schedule are varied in order to identify the MTD. Sections “Designs Based on DLT” through “Designs That Optimize Dose Level and Schedule” summarize such designs to identify the MTD. Section “Discussion and Recommendations” provides recommendations regarding the use of these designs to improve dose finding.

Designs Based on DLT

Multiple methods and approaches have been proposed and evaluated for dose finding using DLT as the primary endpoint. We classify these methods into three categories. First, rule-based designs use dose escalation/de-escalation rules to treat subjects at various dose levels with the goal of converging to treat subjects at the MTD. Second, model-based designs estimate the MTD by estimating the dose–toxicity relationship. Finally, interval designs combine the aspects of rule-based and model-based designs and dose-find using toxicity probability intervals (TPIs).

Rule-Based Designs

Rules-based designs use dose escalation/de-escalation rules to treat subjects at various dose levels with the goal of converging to treat subjects at the MTD. The most commonly used rule-based approach is the 3+3 design.^[3] The 3+3 design is described next in detail; additional rule-based designs are referenced.

3+3 Design The 3+3 design uses enrollment and dosing rules to identify the MTD. The MTD is defined as the highest dose level with at least six subjects treated and fewer than 33% of subjects experiencing a DLT. In designing the 3+3 study, researchers specify the dose levels of interest. The tolerability of the lowest planned dose is

evaluated followed by a sequential evaluation of escalated dose levels until MTD identification. The following rules guide subject enrollment and dosing:

- A cohort of three subjects is enrolled and treated at a dose level. Limiting enrollment to an initial three subjects reduces the number of subjects exposed to overly toxic dose levels and reduces the number of subjects treated at sub-therapeutic dose levels.
If no DLTs, the next cohort of three subjects is enrolled and treated at an escalated dose level.
If one DLT occurs, the current cohort is expanded and three additional subjects are enrolled.
If two or more subjects experience a DLT, the dose is de-escalated for the next cohort.
- A dose level is considered to be not tolerated whenever two or more subjects experience a DLT, regardless of the cohort size ($n = 3$ or 6).
- At least six subjects must be treated at a dose level for the dose level to be considered the MTD.

The sample size for a 3+3 design is adaptive and depends on the dose level where toxicity is observed in accordance with planned dose levels. Designs that aggressively start at a high dose level will likely observe early toxicity, and therefore identify the MTD with a minimal number of subjects (e.g., 9–12 subjects). Designs that start at a low dose with many planned dose levels due to cautious dose escalation will have larger sample sizes (e.g., 30 subjects).

Other Rule-Based Designs A common modification of the 3+3 design is called an accelerated titration design in which single subject cohorts are enrolled until the initial DLT is observed. Simon et al.^[4] describe multiple variations of accelerated titration designs and evaluate the operating characteristics of these designs. Up and down designs are another class of rule-based designs; Storer^[3] describes a number of these up and down designs. Ivanova et al.^[5] evaluate the properties of several up and down designs, and they recommend alterations to improve the operating characteristics of these designs.

Model-Based Designs

Model-based designs determine the MTD by estimating the dose–toxicity relationship. A one- or two-parameter model is typically used. A model-based approach to oncology dose finding was initially proposed by O’Quigley et al. in 1990^[6] using a method called the continual reassessment method (CRM). CRM is designed to treat each subject near the optimal dose level, where the optimal level is estimated from a dose–toxicity model. As data accrue, the model is updated and optimal dose levels are reassessed.

Due to practical concerns, the originally proposed CRM is not implemented in practice. The CRM concept greatly influenced additional research on model-based dose-finding

designs with many of these designs being implemented in practice. A commonly used model-based approach is the Bayesian logistic regression model (BLRM). This design is described next in detail; additional model-based designs are referenced.

Bayesian Logistic Regression Model As described by Neuenschwander et al.,^[7] BLRM makes model-based dose-level recommendations with the model being reassessed throughout the study. Additionally, BLRM implements the concept of escalation with overdose control.^[8]

Model BLRM models the probability of toxicity (p_i) using the logistic transformation:

$$\begin{aligned}\log\left[p_i/(1-p_i)\right] &= \text{logit}(p_i) \\ &= \log(a) + \exp[\log(b)] \log(d_i/d_{\max})\end{aligned}$$

The value d_i represents the dose level of interest and $d_{\max}$ represents the maximum planned dose. The values a and b are model parameters. Since the Bayesian framework is used, the parameters a and b are random variables with the pair $\log(a)$ and $\log(b)$ assumed to follow a bivariate normal distribution. The prior for $\log(a)$ and $\log(b)$ is calibrated using at least two quantiles for each dose level (e.g., 95% credible interval) where these quantiles are specified by the study team. These team-specified quantiles are compared to model-derived quantiles in order to select the prior. The bivariate normal distribution that minimizes the discrepancy between model-derived quantiles and team-specified quantiles is selected as the prior for $\log(a)$ and $\log(b)$.

Overdose Control BLRM makes dose recommendation based on the intervals from the Bayesian posterior distribution. The probability of DLT is described in four intervals:

- Underdosing: (0, 0.20]
- Targeted toxicity: (0.20, 0.35]
- Excessive toxicity: (0.35, 0.60]
- Unacceptable toxicity: (0.60, 1.00]

The recommended dose is the dose level that maximizes the targeted toxicity, while controlling at acceptable levels the risk of excessive or unacceptable toxicity (e.g., <25% probability).

Practical Notes Practical implementation of the BLRM requires a list of planned dose levels with dosing beginning at the lowest planned dose level. Dose-level decisions will be made after a cohort of safety data are observed; a common cohort size is $n = 2$ subjects. Though the BLRM uses a model-based method to control for overdose, practical safety constraints can be added (e.g., never dose-escalate more than twofold above the current dose level). The maximum

sample size (e.g., $n = 30$ subjects) and other design features such as the early stopping rule are chosen a priori for the BLRM design. Simulations should be done to evaluate the impact of these design features on operating characteristics such as accuracy in identifying the MTD, number of DLTs, and trial duration. BLRM may perform poorly when toxicity is observed early in the study. BLRM uses a two-parameter model which can lead to erratic dose recommendations at small sample sizes (see the work of Shu and O'Quigley^[9]). When it is anticipated that toxicity will be observed early in the study at small sample sizes, a design using a one-parameter model likely would have better operating characteristics than BLRM (see Section "Other Model-Based Designs" for a description of one-parameter designs). Despite this concern, BLRM performs well in common drug development scenarios where toxicity is not observed until later in the study, and the two-parameter model has sufficient data to make reasonable dose-level recommendations.

Other Model-Based Designs Additional model-based designs build on the original CRM design with accommodations for practical concerns. Addressing the overdose risk from making dose-level decisions using data from a single subject, Goodman et al.^[10] propose a group CRM in which cohorts of $n = 3$ subjects are treated and evaluated prior to making a dose-level decision. Addressing the overdose risk using a model-based projection of toxicity risk, Moller^[11] proposes capping the level of dose escalation regardless of the model-based recommendation. Additional practical modifications to the original CRM design have been proposed (see the works of Goodman et al.^[10] and Faries and Whitehead et al.^[12,13]).

Researchers have investigated the specification of a Bayesian prior in model-based dose finding. Generally, either one- or two-parameter models have been used for the prior. The one-parameter model requires prespecified skeleton probabilities in a working model where under the Bayesian framework the skeletons can be interpreted as the prior understanding of the toxicity probability for each dose level under investigation. O'Quigley and Zohar, Lee and Cheung, and Daimon et al.^[14–16] generally show that the one-parameter model is robust and efficient in estimating MTD, if an appropriate skeleton is chosen. Iasonos and O'Quigley^[17] discuss the aspects of choosing the skeleton in the setting of a two-stage design. Whitehead and Williamson^[18] describe two-parameter models. Neuenschwander et al.^[7] describe a two-parameter model that they generally prefer to the one-parameter model as it realistically can better fit the dose-toxicity curve. Shu and O'Quigley^[9] point out that the two-parameter CRM estimates can be inconsistent and they recommend using a one-parameter CRM model.

Researchers have proposed and investigated early stopping rules for a model-based design. O'Quigley and Reiner, Zohar and Chevret, and Whitehead and Williamson^[19–20,18] describe stopping rules for Bayesian dose-finding studies.

Interval Designs

Interval designs are developed to have straightforward dose-level decision rules while incorporating model-based concepts. These designs combine the aspects of rule-based and model-based designs and dose-find using TPIs. Like rule-based designs, the interval designs dose-find by making up and down dose-level decisions. Like model-based designs, the aspects of the decision rules are based on Bayesian modeling. A useful interval design is the modified TPI (mTPI) design. This design is described next in detail; additional interval designs are referenced.

Modified Toxicity Probability Interval Ji et al.^[21] propose the mTPI, a modified design from the method described by Ji et al.^[22] The goal of the mTPI is to select a dose level with a high probability of having a toxicity rate within the target toxicity probability interval (TPI). The target TPI is defined during study design. A common target TPI is 0.20–0.35, representing a 20%–35% probability of toxicity.

The mTPI design recommends dose-level decisions using three predefined intervals and a statistic called the unit probability mass (UPM). The UPM is estimated using a Bayesian model.

Interval Definitions

- **Target TPI:** the target range on toxicity rate for a dose level (e.g., 0.20–0.35)
- **Underdosing TPI:** The range on toxicity rate for a dose level that may indicate suboptimal efficacy (e.g., toxicity rate from 0 to 0.20)
- **Overdosing TPI:** The range on toxicity rate for a dose level that indicates excessive toxicity (e.g., 0.35–1)

Unit Probability Mass

- For a dose level of interest, the UPM is the probability that the toxicity rate is within the interval of interest divided by the length of the interval.

Bayesian Model

- p_i represents the probability of toxicity at dose level i : A uniform prior distribution is assumed for each p_i . Accumulating data at dose level i are used to update posterior distribution for p_i .
- Define penalties for each possible decision and derive expected loss:
Decisions: escalate dose level, expand dose level, and de-escalate dose level
Ji et al. propose interval-based penalties in which minimization of expected loss corresponds to the highest UPM.

- Create a separate model for each p_i ; only data from the corresponding dose level i are used to update model of p_i . The study design lists planned dose levels. The initial cohort of subjects is treated at the lowest planned dose level. The mTPI dose-finding algorithm proceeds as follows:
- For the current dose level, update the Bayesian model. The updated Bayesian model will include data from the most recent cohort, applicable data from prior cohorts where subjects are treated at the current dose level and the Bayesian prior.
- For the current dose level, estimate the UPM for each interval and make dose-level recommendations for the next cohort.
De-escalate dose level: Overdose TPI has the highest UPM.
Repeat current dose level: Target TPI has the highest UPM.
Escalate dose level: Underdose TPI has the highest UPM.
- Determine if the current dose level is unacceptable. If overdose TPI has the highest UPM, the current dose level and any higher dose levels are excluded from future consideration.
- The study continues until one of the following exists:
Maximum enrollment is achieved.
All dose levels are unacceptable.

The sample size for an mTPI design is adaptive and depends on the dose level where toxicity is observed in accordance with planned dose levels. Designs that aggressively start at a high dose level will likely observe early toxicity and, therefore, identify the MTD with a minimal number of subjects (e.g., 9–12 subjects). Designs that start at a low dose with many planned dose levels due to cautious dose escalation will have larger sample sizes (e.g., 30 subjects).

Other Interval Designs Additional interval designs have also been proposed. Liu and Yuan^[23] propose the Bayesian optimal interval (BOIN) design. BOIN minimizes decision error in dose-finding trials with decision errors by optimizing the width of the target, and underdosing and overdosing TPIs. Also see the works of Ivanova et al.^[24] and Oron et al.^[24,25]

Designs Based on Ordinal Toxicity

Dose-finding study designs describe algorithms that optimize dose level using specifically defined safety endpoints, such as DLT. In practice, clinicians evaluate all available safety information including multiple safety endpoints when determining whether to follow the dose-finding algorithm. For example, the severity or grade of a toxicity is an important factor that the clinicians evaluate during review of safety data. To incorporate toxicity grade in

dose-finding designs, researchers have proposed methods to utilize ordinal toxicity information.

Meter et al.^[26] describe an extension of the CRM method for dose finding using ordinal toxicity. They propose a proportional odds (PO) model for the relationship between dose and ordinal toxicity. In this model, the intercept differs for each toxicity grade, but the relationship between increasing dose and increasing incidence of toxicity is the same for each toxicity grade. Maximum likelihood is used to estimate the parameters of interest with pseudo data used to stabilize the model. Meter et al. suggest that the pseudo data may be down-weighted or eliminated as enough actual data are collected to allow model stabilization. Additionally, Meter et al.^[27] describe an extension to the CRM method using a continuation ratio (CR) model. Doussau et al.^[28] propose a dose-finding model for longitudinal graded toxicity. Nebiyu Bekele et al.^[29] propose a dose-finding model using average toxicity score.

Designs Based on Time to First Toxicity

Dose-finding studies in oncology usually prespecify the observation period for toxicity. Often, toxicity that occurs after this observation period is not formally evaluated in dose-finding algorithms. Practically, clinicians do take into account late-onset toxicity when making dose recommendations. Researchers have developed dose-finding methodologies using time to first toxicity, an endpoint that can formally include information from subjects with incomplete follow-up and information from late-onset toxicity.

Cheung and Chappell^[30] extend the CRM method to incorporate time-to-event endpoints, a methodology that they refer to as time-to-event CRM (TTE-CRM). TTE-CRM utilizes information from early dropouts through weighted estimation. Subjects who complete the required observation period receive full weight in estimating the dose-toxicity relationship, whereas early dropouts have reduced weight. This weighting function assumes uniform distribution of time to toxicity over the observation period. Additional weight functions could also be used. Braun^[31] proposes a weight function in TTE-CRM that is a function of dose and allows timing of toxicity to vary with dose intensity. Yuan and Yin^[32] propose a methodology to jointly model time-to-event efficacy and toxicity. Polley^[33] proposes practical modifications to TTE-CRM. Wages et al.^[34] evaluate TTE-CRM in the presence of partial orders.

Designs That Optimize Dose Level and Schedule

The dose cycle frequency and the number of dose cycles can impact toxicity and efficacy. Oncology therapeutics typically have had breaks from dosing to allow patients to recover but the optimal length of this recovery period is not always obvious. A number of researchers have proposed methods to dose-find using both the dose level and the dose schedule.

Braun et al.^[35] propose a method to optimize dose level and dosing schedule. They propose modeling the hazard of toxicity so the overall probability of toxicity for an individual subject is a function of the dose level and the number and timing of doses received. They propose a triangular hazard that accounts for both dose level and follow-up time. Li et al.^[36] propose a methodology to optimize dose level and dose schedule based on toxicity and efficacy endpoints. They propose an algorithm to identify the dose level and dosing schedule with high probability of having both acceptable efficacy and acceptable toxicity. Thall et al.^[37] develop a methodology that includes two stages to adaptively optimize dose level and dosing schedule with the first stage randomizing subjects to a dosing schedule and the second stage using accumulating data to optimize dose and dosing schedule. Lee et al.^[38] develop a methodology to modify an individual's cycle 2 of treatment based on the results of cycle 1 dose level and outcome; decisions use the individual subject's cycle 1 data and data from other subjects on study. Some additional research in optimizing dose level and dose schedule can be found in the works of Zhang and Braun^[39] and Guo et al.^[40]

Discussion and Recommendations

The merits and limitations between algorithm-based designs (e.g., 3+3 design) and model-based designs are summarized in [Tables 1a](#) and [1b](#). The 3+3 design remains the most common design utilized to identify the MTD. Ji and Wang^[41] point out the popularity of the 3+3 design results from "...its simplicity (no computer program is required), transparency, and costless implementation in practice." Despite its popularity, the 3+3 design has some important limitations. The 3+3 design generally underperforms model-based designs in terms of operating characteristics such as accuracy in identifying the MTD and the percentage of Phase I subjects treated near the MTD.^[42–44] The 3+3 design also has a fixed toxicity limit of 33% which may not be ideal for certain drugs or indications. For example, the acceptable toxicity rate may be well below 33% in an indication with many treatment options and a long-expected survival. In addition, the 3+3 design is memoryless and does not utilize prior dose-level data when making dose recommendations. Finally, the inflexible decision rules for the 3+3 design can lead to operational inefficiency as decisions must be made using an exact number of evaluable subjects; these inflexible rules can slow enrollment and lead to study delays due to replacement of subjects. Many designs have been developed to overcome these deficiencies of the 3+3 design.

When resources are available to build a model, a model-based design such as the BLRM should be considered. As mentioned in Section "Model-Based Designs," researchers have shown good operating characteristics for model-based designs especially when compared to the 3+3 design. Model-based designs allow flexibility in targeting

Table 1a Merits of algorithm-based designs (e.g., 3+3 design) compared to potential limitations of model-based designs

Merits of algorithm-based design	Limitations of model-based design
1. Study executed without review of statistical output 2. Common use reduces up-front efforts to justify design features to review bodies (e.g., regulatory agencies) 3. Memoryless feature allows design to overcome misclassification of disease-related adverse event as a DLT 4. Insensitive to deviations from assumed dose–toxicity model	1. Study execution requires regular statistical support to generate statistical output 2. Operating characteristics required to justify design features to review bodies 3. Misclassification of disease-related adverse events as DLT difficult to remedy 4. Deviations from assumed dose–toxicity model can lead to bias in estimation of MTD

Table 1b Merits of model-based designs compared to potential limitations of algorithm-based designs (e.g., 3+3 design)

Merits of Model-Based Design	Limitation of Algorithm-Based Design
1. Flexibility in targeting toxicity rate 2. Flexibility in cohort sample size and open enrollment (e.g., two to four subjects per cohort) can reduce study execution delays due to subject replacement 3. Compared to 3+3 design, superior accuracy in identifying MTD^[42–44] 4. Facilitates up-front discussion regarding specifics of risk–benefit of a treatment in a particular disease indication. Increased scientific exchange allows for customization of study design.	1. 3+3 design does not have flexibility in targeting toxicity rate (less common algorithm-based designs can modify the target toxicity rate) 2. Strict decision rules based on fixed sample sizes often result in study execution delays due to subject replacement 3. Compared to model-based designs, inferior accuracy in identifying MTD^[42–44] 4. Standard approach can limit scientific exchange

a toxicity rate that allows customization of dose finding to the drug and indication. Model-based designs also incorporate data from prior dose levels when making dose-level recommendations. Additionally, model-based designs may improve dose-finding decisions by aiding communication between key decision makers. Finally, an important benefit of the model-based design is operational efficiency. In particular, model-based designs easily incorporate flexible cohort sizes (e.g., two to four subjects per cohort), which can speed up enrollment and reduce study delays due to replacement of subjects.

When resources are not available to build a model, there are good design options that overcome some of the shortcomings of the 3+3 design. For example, the mTPI interval design is simple and transparent. The mTPI design can be planned using an easy-to-use Excel tool. The mTPI has flexibility in modifying the target TPI which allows customization that meets the specific needs of an oncology drug and indication. The rule-based, up and down designs described by Ivanova et al.^[5] are also flexible enough to modify the target toxicity rate.

Model-based designs may also have dose-finding performance issues. When the actual dose–toxicity relationship deviates from the assumed dose–toxicity model, model-based approaches can make illogical dose-level recommendation (e.g., recommend dose escalation even though significant toxicity is observed at current dose level). As discussed previously, the two-parameter model can lead to erratic dose recommendations early on in the study and at small sample sizes. The choice of prior can impact early dose recommendations when actual observed data are minimal; Oron and Hoff^[45] describe

examples of small-sample behavior in Phase I cancer trials. Iasonos and O’Quigley^[17] note the circumstances where the one-parameter CRM model will underestimate the dose–toxicity relationship at higher dose levels when there is a run-in of safe lower dose. When considering a model-based approach, the researcher should execute clinical trial simulations in part to identify potential dose-finding performance issues.

DESIGNS TO IDENTIFY THE RECOMMENDED PHASE II DOSE

It is common in Phase I studies to have a primary objective of identifying the maximum tolerated dose and/or recommended Phase II dose (RP2D). This objective acknowledges that the maximum tolerated dose may or may not be the recommended future dose. The perfect drug and dose level would cure the disease in 100% of patients with no adverse effects of concern. Recognizing that this is not likely the case, the researcher seeks a RP2D that appropriately balances the trade-off between efficacy and safety. Additionally, efficiency in dose finding can be obtained using all relevant safety, efficacy, biomarker, and pharmacodynamic information in Phase I clinical trials to recommend a dose level for Phase II clinical trials. Multiple methods and approaches have been proposed and evaluated for identifying the RP2D. We classify these methods into two categories: designs that balance the trade-off between toxicity and efficacy, and designs that incorporate safety, biomarker, and pharmacokinetic/pharmacodynamic data.

Sections “Designs That Balance Toxicity and Efficacy” and “Designs That Incorporate Biomarker/Pharmacodynamic Data” summarize designs to identify the RP2D. Section “Discussion and Recommendations” provides recommendations regarding the use of these designs to improve dose finding.

Designs That Balance Toxicity and Efficacy

Designs that balance the trade-off between toxicity and efficacy seek to identify a dose level that maximizes efficacy while maintaining an appropriate safety profile. A useful trade-off design is proposed by Thall and Cook (TC).^[46] The TC design is described next in detail; additional trade-off designs are referenced.

TC Trade-Off Design

TC design is a Bayesian approach to dose finding that aims to optimize the trade-off between toxicity and efficacy. A contour C defines the acceptable trade-off between toxicity and efficacy. Contour C is constructed using the following three inputs from the investigating clinicians:

1. The drug’s lowest level of efficacy allowable if the drug has no toxicity
Example: 0% DLT rate and 10% objective response rate (ORR)
2. The drug’s highest incidence of adverse events allowable if the drug is 100% effective
Example: 50% DLT rate and 100% ORR
3. One example of intermediate safety and efficacy
Example: 25% DLT rate and 35% ORR

A smoothing procedure identifies contour C where C includes each paired safety/efficacy value described in inputs 1–3. A dose level is acceptable if the safety/efficacy profile is at or better than the contour C. TC scores the safety/efficacy profile of a dose level of interest using the distance from contour C to the dose’s predicted safety/efficacy. Dose-level profiles on contour C are scored as zero. If a dose-level profile is better than profiles on contour C, then a larger distance from contour C leads to a higher positive profile score. If a dose-level profile is worse than profiles on contour C, then a larger distance from contour C leads to a more negative score. TC uses an algorithm to identify a dose level with the maximum score, where the maximum score indicates the maximum risk/benefit. Fig. 1 shows an example contour C.

Key features of the dose-finding algorithm are as follows:

- Jointly model the efficacy/safety profile using Bayesian model.
- Prespecify dose levels of interest.

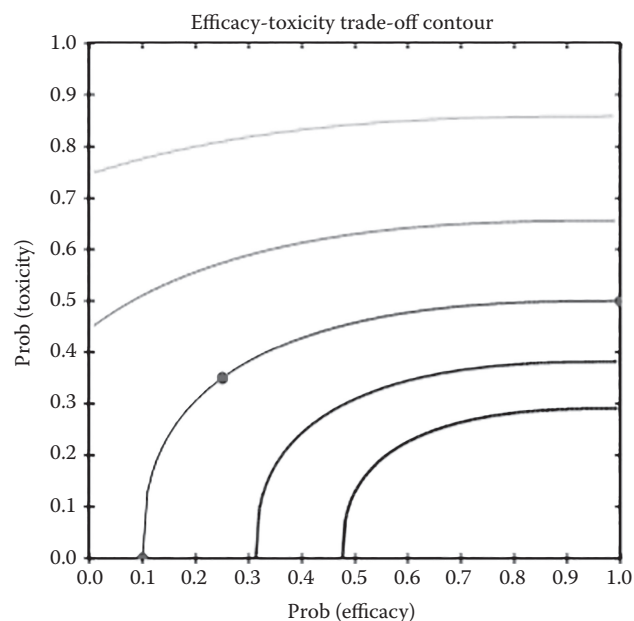


Fig. 1 Example of contour C in the TC trade-off design. Contour C defines the acceptable trade-off between toxicity and efficacy. Contour C has the red dots with the dots representing the three inputs used to derive C. The other contours represent safety/efficacy profiles which are inferior (closer to upper left) or superior (closer to bottom right) to contour C

- After each cohort, update Bayesian model and predict the efficacy and safety for each dose level of interest.
- Stop study if no dose levels are acceptable (or maximum sample size is reached).
- If at least one dose level is acceptable, treat the next cohort of subjects at the dose level with the maximum score.

The maximum sample size (e.g., $n = 60$ subjects) and other design features are chosen a priori for the trade-off design. Simulations should be done to evaluate the impact of these design features on operating characteristics. Thall et al.^[47] provide additional practical advice on how to implement this trade-off contour design.

Other Trade-Off Designs

Thall and Russell (TR)^[48] propose a Bayesian adaptive design with the goal of dose finding using trinomial outcomes (“no efficacy and no toxicity,” “efficacy and no toxicity,” “no efficacy and toxicity”), using a PO model. Braun^[49] extends the CRM to incorporate bivariate endpoints, called the bivariate CRM. Fan and Chaloner^[50] use a CR model that is more flexible than the PO model. Zhang et al.^[51] propose an approach similar to TR but using the CR model and making modifications to the dose selection criteria and stopping rules. Yin et al.^[52] develop an approach of dose finding using efficacy and toxicity that does not require specifying the parametric functional form of the dose–response curve.

Mandrekar et al.^[53] review and offer practical solutions to challenges that arise with executing a dose-finding study incorporating both efficacy/toxicity. Zhong et al.^[54] develop an approach that also incorporates a surrogate for efficacy. Yuan and Yin^[55] propose a method using time to toxicity and efficacy endpoints. Lee et al.^[56] propose a methodology with the goal to adaptively select dose to maximize risk–benefit for every subject treated on study. Dragalin and Fedorov^[57] propose a penalized D-optimal design with the penalty functioning to help balance out the interests of the individual with the interest of future patients. Dragalin et al.^[58] propose a two-stage design with information from stage 1 informing dose allocation of stage 2; stage 2 allocates patients to dose levels chosen to minimize the variance of the estimator for optimal dose.

Designs That Incorporate Biomarker/Pharmacodynamic Data

Continuous measures of efficacy and safety may be incorporated to improve dose finding. Biomarker and pharmacokinetic/pharmacodynamic data may inform dosing.^[59] Researchers have proposed methods to incorporate this information in dose finding.

Nebiyou Bekele and Shen^[60] propose jointly modeling toxicity and biomarker expression in a dose-finding trial. Ivanova and Kim^[61] generalize a methodology to dose-find using various combinations of dichotomous, continuous, or ordinal endpoints. Hirakawa,^[62] Padmanabhan et al.,^[63] and Yeung et al.^[64] propose a methodology to dose-find using correlated binary and continuous responses. Thall et al.^[65] propose a methodology to optimize the concentration and bolus drug delivery. Fedorov et al.^[66] discuss optimal dose-finding designs with correlated continuous and discrete responses. Zhong et al.^[67] develop a methodology to dose-find using toxicity, efficacy, and surrogate efficacy.

Discussion and Recommendations

Restricting dose finding in oncology to only Phase I clinical trials is a practice that can lead to suboptimal dosing of patients. We recommend that dose-finding activities continue past Phase I and that a Phase I/II design such as the TC design be strongly considered. Historically, Phase I trial identifies the MTD and Phase II accumulates additional safety data and screens for efficacy; Phase II data have not typically been used to reevaluate the MTD. The limited data in Phase I leads to uncertainty in estimation of MTD regardless of study design.^[42] Safety data from Phase II can be used to improve the estimate of the MTD. The TC design can utilize Phase II data to improve MTD estimation as well as utilize Phase I data along with Phase II data to improve efficacy estimation. Additionally, the convention of dosing at the MTD can ignore the important benefit–risk trade-offs. The level of acceptable toxicity may depend on the level of efficacy, the severity

and manageability of the toxicity, the disease prognosis, and the availability of alternative therapies. Continuing to dose-find past Phase I allows improved refinement in the understanding of a drug's benefit–risk.

Dose finding can be improved by more frequently modeling pharmacokinetics and/or dose level with pharmacodynamics, adverse events, safety labs, and efficacy endpoints. Typical dose-finding studies simplify the dose-finding problem by simplifying the dose-finding endpoints. In practice, the deciding clinicians evaluate a totality of data when making decisions. The grade of adverse events and the timing/duration of the event are considered. Safety labs such as platelet counts over time and in relation to pharmacokinetics are considered. Pharmacodynamics and efficacy data may be important. Developing exploratory models of these relationships and reviewing these models with clinicians will aid communication and decision-making.

DESIGNS TO IDENTIFY THE COMBINATION MTD

Treatment regimens with two or more active agents often show synergy in efficacy. An effective drug in monotherapy is usually evaluated in combination with other agents. With two agents, the identification of the maximum tolerated dose is a two-dimensional problem. A useful design to identify the combination MTD is the partial order CRM (POCRM) proposed by Wages et al.^[68] This design is described next in detail; additional combination MTD designs are referenced.

Partial Order CRM

POCRM reduces the two-dimensional dose finding into one-dimensional problem by using the concept of partial orders and model selection. Using partial orders, POCRM categorizes plausible models that order the treatment combinations from least toxic to most toxic. To illustrate, here is a simple example of partial orders.

Example of Partial Orders

The following treatment combinations are under investigation:

- Combination 1: Drug A at 100mg and Drug B at 100mg
- Combination 2: Drug A at 100mg and Drug B at 200mg
- Combination 3: Drug A at 200mg and Drug B at 100mg
- Combination 4: Drug A at 200mg and Drug B at 200mg

The following are partial orders:

- Combination 1 is less toxic than Combinations 2, 3, and 4.
- Combination 4 is more toxic than Combinations 1, 2, and 3.

Given the known partial orders and the unknown relative toxicity between Combinations 2 and 3, there are two plausible models ordered from least toxic to most toxic:

- Model 1: 1, 2, 3, 4
- Model 2: 1, 3, 2, 4

POCRM assigns a prior probability to each model. The model with the highest prior probability is selected as the working model for dose finding. In the absence of data or with limited data, the working model may be randomly selected. Using the selected working model, a one-dimension CRM is used for dose finding. As data accumulate, a different working model may be selected based on posterior model probabilities.

The maximum sample size (e.g., $n = 36$ subjects) and other design features are chosen a priori for the POCRM design. Simulations should be done to evaluate the impact of these design features on operating characteristics. Wages and Conaway^[69] give additional practical guidance on implementing POCRM.

OTHER COMBINATION MTD DESIGNS

Thall et al.^[70] propose a six-parameter power-logit function for the joint distribution. Yin and Yuan^[71] propose a joint model that uses a copula. Bailey et al.^[72] propose a methodology to use logistic regression with covariates. Ivanova and Wang^[73] propose an “up-and-down” method. Additional methodologies are described by Conaway et al.,^[74] Huang et al.,^[75] Yin and Yuan,^[76] Braun and Wang,^[77] Wages and Conaway,^[78] Hirakawa et al.,^[79] Mander and Sweeting,^[80] Riviere and Karelle,^[81] and Jin et al.^[82]

Discussion and Recommendations

We recommend the POCRM design for combination studies. In particular, POCRM should be considered to dose-find when a new drug is combined with an established drug. Since the established drug has a well-described efficacy profile, lowering the dose level of the established drug is often not considered. This practice may miss out on finding a dosing combination that optimizes benefit–risk. For example, the optimal benefit–risk for drug combinations may be at moderate dose levels for both drugs including using a dose level for the established drug that is below the label prescribed level. POCRM can easily interrogate multiple dose levels, even for the established drugs. POCRM leverages the well-understood CRM methodology and facilitates communication among the statistician, the clinician, and other scientists. POCRM may get complicated if partial orders lead to a large number of possible orderings. Practically, the number of orderings likely can be reduced to a smaller set of plausible orderings. The drugs of interest will have existing preclinical and clinical data to aid planning of

the combination study. Utilizing these data and good communication among the statistician, the clinician, and other scientists should allow the identification of a reasonable number of plausible orders to investigate in POCRM.

RECENT TRENDS IN DOSE FINDING

Utilization of dose expansion data in Phase I has received attention recently. It has become common practice in FIH oncology studies to add a dose expansion part with an additional 10–20 subjects. After the Phase I study identifies an MTD, the purpose of dose expansion is to gain additional safety data and preliminary efficacy data at the MTD. Iasonos and O’Quigley^[83–85] propose a methodology to utilize dose expansion data to update the estimate of the MTD. Additionally, they present a methodology to continue monitoring safety while assessing efficacy and exploring patient subsets with enhanced responsiveness to treatment.

The concept of the maximum tolerated contour (MTC) has been proposed for recent combination dose-finding methodologies. The contour represents the pairs of dose levels between Agent A and Agent B that have equal probability of toxicity, with MTC being the pairs of dose levels with the target toxicity probability. Researchers publishing methodologies for dose finding in combinations studies using MTC are Wages,^[86] Zhang and Yuan,^[87] and Mander and Sweeting.^[88]

Dose finding utilizing information from different subsets (e.g., ethnic population or tumor types) has been the topic of recent methodologies. Liu et al.^[89] discuss the bridging information across ethnic populations. Broglio et al.^[90] discuss the sharing of information between different patient populations such as different tumor types. Quintana^[91] discusses optimizing dose and schedule using data from different tumor types. Conaway and Wages^[92] use ordered groups to identify the MTDs where groups are ordered by drug tolerability. Wages et al.^[93] and Salter et al.^[94] discuss dose finding in heterogeneous groups.

Personalized dose finding has recently been further investigated. Chen et al.^[95] seek to optimize the individual dosing rule using weighted learning. Zhou et al.^[96] propose a residual weighted learning to estimate individual treatment rules. Guo and Yuan^[97] investigate a biomarker-driven personalized dose finding. Mao and Cheung^[98] propose a sequential design to aid individualized dose finding.

REFERENCES

1. Howlader, N.; Noone, A.M.; Krapcho, M.; et al. (Eds). SEER Cancer Statistics Review, 1975–2010, National Cancer Institute; Bethesda, MD; http://seer.cancer.gov/csr/1975_2010/, based on November 2012 SEER data submission, posted to the SEER Web site, April 2013.

2. Biotechnology Innovation Organization Report Online. <https://www.bio.org/sites/default/files/Clinical%20Development%20Success%20Rates%202006-2015%20-%20BIO,%20Biomedtracker,%20Amplion%202016.pdf>, (accessed November 2016).
3. Storer, B.E. Design and analysis of phase I clinical trials. *Biometrics* **1989**, *45* (3), 925–937.
4. Simon, R.; Rubinstein, L.; Arbuck, S.G.; Christian, M.C.; Freidlin, B.; Collins, J. Accelerated titration designs for phase I clinical trials in oncology. *J. Natl. Cancer Inst.* **1997**, *89* (15), 1138–1147.
5. Ivanova, A.; Montazer-Haghighi, A.; Mohanty, S.G.; Durham, S.D. Improved up-and-down designs for phase I trials. *Stat. Med.* **2003**, *22* (1), 69–82.
6. O'Quigley, J.; Pepe, M.; Fisher, L. Continual reassessment method: A practical design for phase I clinical trials in cancer. *Biometrics* **1990**, *46* (1), 33–48.
7. Neuenschwander, B.; Branson, M.; Gsponer, T. Critical aspects of the Bayesian approach to phase I cancer trials. *Stat. Med.* **2008**, *27* (13), 2420–2439.
8. Babb, J.; Rogatko, A.; Zacks, S. Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Stat. Med.* **1998**, *17* (10), 1103–1120.
9. Shu, J.; O'Quigley, J. Dose-escalation designs in oncology: ADEPT and the CRM. *Stat. Med.* **2008**, *27* (26), 5345–5353.
10. Goodman, S.N.; Zahurak, M.L.; Piantadosi, S. Some practical improvements in the continual reassessment method for phase I studies. *Stat. Med.* **1995**, *14* (11), 1149–1161.
11. Møller, S. An extension of the continual reassessment methods using a preliminary up-and-down design in a dose finding study in cancer patients, in order to investigate a greater range of doses. *Stat. Med.* **1995**, *14* (9), 911–922.
12. Faries, D. Practical modifications of the continual reassessment method for phase I cancer clinical trials. *J. Biopharm. Stat.* **1994**, *4* (2), 147–164.
13. Whitehead, J.; Brunier, H. Bayesian decision procedures for dose determining experiments. *Stat. Med.* **1995**, *14* (9), 885–893.
14. O'Quigley, J.; Zohar, S. Retrospective robustness of the continual reassessment method. *J. Biopharm. Stat.* **2010**, *20* (5), 1013–1025.
15. Lee, S.M.; Cheung, Y.K. Model calibration in the continual reassessment method. *Clin. Trials* **2009**, *6* (3), 227–238.
16. Daimon, T.; Zohar, S.; O'Quigley, J. Posterior maximization and averaging for Bayesian working model choice in the continual reassessment method. *Stat. Med.* **2011**, *30* (13), 1563–1573.
17. Iasonos, A.; O'Quigley, J. Interplay of priors and skeletons in two-stage continual reassessment method. *Stat. Med.* **2012**, *31* (30), 4321–4336.
18. Whitehead, J.; Williamson, D. Bayesian decision procedures based on logistic regression models for dose-finding studies. *J. Biopharm. Stat.* **1998**, *8* (3), 445–467.
19. O'Quigley, J.O.; Reiner, E. A stopping rule for the continual reassessment method. *Biometrika* **1998**, *85* (3), 741–748.
20. Zohar, S.; Chevret, S. The continual reassessment method: Comparison of Bayesian stopping rules for dose-ranging studies. *Stat. Med.* **2001**, *20* (19), 2827–2843.
21. Ji, Y.; Liu, P.; Li, Y.; Nebiyu Bekele, B.N. A modified toxicity probability interval method for dose-finding trials. *Clin. Trials* **2010**. doi: 10.1177/1740774510382799.
22. Ji, Y.; Li, Y.; Yin, G. Bayesian dose finding in phase I clinical trials based on a new statistical framework. *Stat. Sin.* **2007**, *17*, 531–547.
23. Liu, S.; Yuan, Y. Bayesian optimal interval designs for phase I clinical trials. *J. R. Stat. Soc.: Ser. C, Appl. Stat.* **2015**, *64* (3), 507–523.
24. Ivanova, A.; Flournoy, N.; Chung, Y. Cumulative cohort design for dose-finding. *J. Stat. Plan. Inference* **2007**, *137* (7), 2316–2327.
25. Oron, A.P.; Azriel, D.; Hoff, P.D. Dose-finding designs: The role of convergence properties. *Int. J. Biostat.* **2011**, *7* (1), 1–7.
26. Meter, E.M.; Garrett-Mayer, E.; Bandyopadhyay, D. Proportional odds model for dose-finding clinical trial designs with ordinal toxicity grading. *Stat. Med.* **2011**, *30* (17), 2070–2080.
27. Van Meter, E.M.; Garrett-Mayer, E.; Bandyopadhyay, D. Dose-finding clinical trial design for ordinal toxicity grades using the continuation ratio model: An extension of the continual reassessment method. *Clin. Trials* **2012**, *9* (3), 303–313.
28. Doussau, A.; Thiébaud, R.; Paoletti, X. Dose-finding design using mixed-effect proportional odds model for longitudinal graded toxicity data in phase I oncology clinical trials. *Stat. Med.* **2013**, *32* (30), 5430–5447.
29. Nebiyu Bekele, B.N.; Li, Y.; Ji, Y. Risk-group-specific dose finding based on an average toxicity score. *Biometrics* **2010**, *66* (2), 541–548.
30. Cheung, Y.K.; Chappell, R. Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics* **2000**, *56* (4), 1177–1182.
31. Braun, T.M. Generalizing the TITE-CRM to adapt for early- and late-onset toxicities. *Stat. Med.* **2006**, *25* (12), 2071–2083.
32. Yuan, Y.; Yin, G. Bayesian dose finding by jointly modeling toxicity and efficacy as time-to-event outcomes. *J. R. Stat. Soc.: Ser. C, Appl. Stat.* **2009**, *58* (5), 719–736.
33. Polley, M.Y. Practical modifications to the time-to-event continual reassessment method for phase I cancer trials with fast patient accrual and late-onset toxicities. *Stat. Med.* **2011**, *30* (17), 2130–2143.
34. Wages, N.A.; Conaway, M.R.; O'Quigley, J. Using the time-to-event continual reassessment method in the presence of partial orders. *Stat. Med.* **2013**, *32* (1), 131–141.
35. Braun, T.M.; Thall, P.F.; Nguyen, H.; de Lima, M. Simultaneously optimizing dose and schedule of a new cytotoxic agent. *Clin. Trials* **2007**, *4* (2), 113–124.
36. Li, Y.; Nebiyu Bekele, B.N.; Ji, Y.; Cook, J.D. Dose-schedule finding in phase I/II clinical trials using a Bayesian isotonic transformation. *Stat. Med.* **2008**, *27* (24), 4895–4913.
37. Thall, P.F.; Nguyen, H.Q.; Braun, T.M.; Qazilbash, M.H. Using joint utilities of the times to response and toxicity to adaptively optimize schedule-dose regimes. *Biometrics* **2013**, *69* (3), 673–682.
38. Lee, J.; Thall, P.F.; Ji, Y.; Müller, P. Bayesian dose-finding in two treatment cycles based on the joint utility of efficacy and toxicity. *J. Am. Stat. Assoc.* **2015**, *110* (510), 711–722.

39. Zhang, J.; Braun, T.M. A phase I Bayesian adaptive design to simultaneously optimize dose and schedule assignments both between and within patients. *J. Am. Stat. Assoc.* **2013**, *108* (503), 892–901.
40. Guo, B.; Li, Y.; Yuan, Y. A dose–schedule finding design for phase I–II clinical trials. *J. R. Stat. Soc.: Ser. C, Appl. Stat.* **2016**, *65* (2), 259–272.
41. Ji, Y.; Wang, S.J. Modified toxicity probability interval design: A safer and more reliable method than the 3 + 3 design for practical phase I trials. *J. Clin. Oncol.* **2013**, *31* (14), 1785–1791.
42. Iasonos, A.; Wilton, A.S.; Riedel, E.R.; Seshan, V.E.; Spriggs, D.R. A comprehensive comparison of the continual reassessment method to the standard 3 + 3 dose escalation scheme in Phase I dose-finding studies. *Clin. Trials* **2008**, *5* (5), 465–477.
43. Berry, S.M.; Carlin, B.P.; Lee, J.J.; Muller, P. *Bayesian adaptive methods for clinical trials*. CRC Press: New York, 2010.
44. Cheung, Y.K. *Dose finding by the continual reassessment method*. CRC Press: Boca Raton, FL, 2011.
45. Oron, A.P.; Hoff, P.D. Small-sample behavior of novel phase I cancer trial designs. *Clin. Trials* **2013**, *10* (1), 63–80.
46. Thall, P.F.; Cook, J.D. Dose-finding based on efficacy–toxicity trade-offs. *Biometrics* **2004**, *60* (3), 684–693.
47. Thall, P.F.; Cook, J.D.; Estey, E.H. Adaptive dose selection using efficacy–toxicity trade-offs: Illustrations and practical considerations. *J. Biopharm. Stat.* **2006**, *16* (5), 623–638.
48. Thall, P.F.; Russell, K.E. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics* **1998**, *54* (1), 251–264.
49. Brau, T.M. The bivariate continual reassessment method: Extending the CRM to phase I trials of two competing outcomes. *Control. Clin. Trials* **2002**, *23* (3), 240–256.
50. Fan, S.K.; Chaloner, K. Optimal designs and limiting optimal designs for a trinomial response. *J. Stat. Plan. Inference* **2004**, *126* (1), 347–360.
51. Zhang, W.; Sargent, D.J.; Mandrekar, S. An adaptive dose-finding design incorporating both toxicity and efficacy. *Stat. Med.* **2006**, *25* (14), 2365–2383.
52. Yin, G.; Li, Y.; Ji, Y. Bayesian dose-finding in phase I/II clinical trials using toxicity and efficacy odds ratios. *Biometrics* **2006**, *62* (3), 777–787.
53. Mandrekar, S.J.; Qin, R.; Sargent, D.J. Model-based phase I designs incorporating toxicity and efficacy for single and dual agent drug combinations: Methods and challenges. *Stat. Med.* **2010**, *29* (10), 1077.
54. Zhong, W.; Koopmeiners, J.S.; Carlin, B.P. A trivariate continual reassessment method for phase I/II trials of toxicity, efficacy, and surrogate efficacy. *Stat. Med.* **2012**, *31* (29), 3885–3895.
55. Yuan, Y.; Yin, G. Bayesian dose finding by jointly modeling toxicity and efficacy as time-to-event outcomes. *J. R. Stat. Soc.: Ser. C, Appl. Stat.* **2009**, *58* (5), 719–736.
56. Lee, J.; Thall, P.F.; Ji, Y.; Müller, P. Bayesian dose-finding in two treatment cycles based on the joint utility of efficacy and toxicity. *J. Am. Stat. Assoc.* **2015**, *110* (510), 711–722.
57. Dragalin, V.; Fedorov, V. Adaptive designs for dose-finding based on efficacy–toxicity response. *J. Stat. Plan. Inference* **2006**, *136* (6), 1800–1823.
58. Dragalin, V.; Fedorov, V.V.; Wu, Y. Two-stage design for dose-finding that accounts for both efficacy and safety. *Stat. Med.* **2008**, *27* (25), 5156–5176.
59. Rajman, I. PK/PD modelling and simulations: Utility in drug development. *Drug Discov. Today* **2008**, *13* (7), 341–346.
60. Nebiyu Bekele, B.; Shen, Y. A Bayesian approach to jointly modeling toxicity and biomarker expression in a phase I/II dose-finding trial. *Biometrics* **2005**, *61* (2), 343–354.
61. Ivanova, A.; Kim, S.H. Dose finding for continuous and ordinal outcomes with a monotone objective function: A unified approach. *Biometrics* **2009**, *65* (1), 307–315.
62. Hirakawa, A. An adaptive dose-finding approach for correlated bivariate binary and continuous outcomes in phase I oncology trials. *Stat. Med.* **2012**, *31* (6), 516–532.
63. Padmanabhan, S.K.; Hsuan, F.; Dragalin, V. Adaptive penalized D-optimal designs for dose finding based on continuous efficacy and toxicity. *Stat. Biopharm. Res.* **2010**, *2* (2), 182–198.
64. Yeung, W.Y.; Whitehead, J.; Reigner, B.; Beyer, U.; Diack, C.; Jaki, T. Bayesian adaptive dose-escalation procedures for binary and continuous responses utilizing a gain function. *Pharm. Stat.* **2015**, *14* (6), 479–487.
65. Thall, P.F.; Szabo, A.; Nguyen, H.Q.; Amlie-Lefond, C.M.; Zaidat, O.O. Optimizing the concentration and bolus of a drug delivered by continuous infusion. *Biometrics* **2011**, *67* (4), 1638–1646.
66. Fedorov, V.; Wu, Y.; Zhang, R. Optimal dose-finding designs with correlated continuous and discrete responses. *Stat. Med.* **2012**, *31* (3), 217–234.
67. Zhong, W.; Koopmeiners, J.S.; Carlin, B.P. A trivariate continual reassessment method for phase I/II trials of toxicity, efficacy, and surrogate efficacy. *Stat. Med.* **2012**, *31* (29), 3885–3895.
68. Wages, N.A.; Conaway, M.R.; O’Quigley, J. Continual reassessment method for partial ordering. *Biometrics* **2011**, *67* (4), 1555–1563.
69. Wages, N.A.; Conaway, M.R. Specifications of a continual reassessment method design for phase I trials of combined drugs. *Pharm. Stat.* **2013**, *12* (4), 217–224.
70. Thall, P.F.; Millikan, R.E.; Mueller, P.; Lee, S.J. Dose-finding with two agents in phase I oncology trials. *Biometrics* **2003**, *59* (3), 487–496.
71. Yin, G.; Yuan, Y. Bayesian dose finding in oncology for drug combinations by copula regression. *J. R. Stat. Soc.: Ser. C, Appl. Stat.* **2009**, *58* (2), 211–224.
72. Bailey, S.; Neuenschwander, B.; Laird, G.; Branson, M. A Bayesian case study in oncology phase I combination dose-finding using logistic regression with covariates. *J. Biopharm. Stat.* **2009**, *19* (3), 469–484.
73. Ivanova, A.; Wang, K. A non-parametric approach to the design and analysis of two-dimensional dose-finding trials. *Stat. Med.* **2004**, *23* (12), 1861–1870.
74. Conaway, M.R.; Dunbar, S.; Peddada, S.D. Designs for single- or multiple-agent phase I trials. *Biometrics* **2004**, *60* (3), 661–669.
75. Huang, X.; Biswas, S.; Oki, Y.; Issa, J.P.; Berry, D.A. A parallel phase I/II clinical trial design for combination therapies. *Biometrics* **2007**, *63* (2), 429–436.
76. Yuan, Y.; Yin, G. Sequential continual reassessment method for two-dimensional dose finding. *Stat. Med.* **2008**, *27* (27), 5664–5678.

77. Braun, T.M.; Wang, S. A hierarchical Bayesian design for phase I trials of novel combinations of cancer therapeutic agents. *Biometrics* **2010**, *66* (3), 805–812.
78. Wages, N.A.; Conaway, M.R. Phase I/II adaptive design for drug combination oncology trials. *Stat. Med.* **2014**, *33* (12), 1990–2003.
79. Hirakawa, A.; Hamada, C.; Matsui, S. A dose-finding approach based on shrunken predictive probability for combinations of two agents in phase I trials. *Stat. Med.* **2013**, *32* (26), 4515–4525.
80. Mander, A.P.; Sweeting, M.J. A product of independent beta probabilities dose escalation design for dual-agent phase I trials. *Stat. Med.* **2015**, *34* (8), 1261–1276.
81. Riviere, M.K.; Yuan, Y.; Dubois, F.; Zohar, S. A Bayesian dose-finding design for drug combination clinical trials based on the logistic model. *Pharm. Stat.* **2014**, *13* (4), 247–257.
82. Jin, I.H.; Huo, L.; Yin, G.; Yuan, Y. Phase I trial design for drug combinations with Bayesian model averaging. *Pharm. Stat.* **2015**, *14* (2), 108–119.
83. Iasonos, A.; O’Quigley, J. Dose expansion cohorts in phase I trials. *Statistics in Biopharmaceutical Research* **2016**, *8* (2), 161–170.
84. Iasonos, A.; O’Quigley, J. Integrating the escalation and dose expansion studies into a unified Phase I clinical trial. *Contemp. Clin. Trials* **2016**, *50*, 124–134.
85. Iasonos, A.; O’Quigley, J. Sequential monitoring of Phase I dose expansion cohorts. *Stat. Med.* **2017**, *36*, 204–214. doi: 10.1002/sim.6894.
86. Wages, N.A. Identifying a maximum tolerated contour in two-dimensional dose finding. *Stat. Med.* **2017**, *36*, 242–253. doi: 10.1002/sim.6918.
87. Zhang, L.; Yuan, Y. A practical Bayesian design to identify the maximum tolerated dose contour for drug combination trials. *Stat. Med.* **2016**, *35* (27), 4924–4936.
88. Mander, A.P.; Sweeting, M.J. A product of independent beta probabilities dose escalation design for dual-agent phase I trials. *Stat. Med.* **2015**, *34* (8), 1261–76.
89. Liu, S.; Pan, H.; Xia, J.; Huang, Q.; Yuan, Y. Bridging continual reassessment method for phase I clinical trials in different ethnic populations. *Stat. Med.* **2015**, *34* (10), 1681–1694.
90. Broglio, K.R.; Sandalic, L.; Albertson, T.; Berry, S.M. Bayesian dose escalation in oncology with sharing of information between patient populations. *Contemp. Clin. Trials* **2015**, *44*, 56–63.
91. Quintana, M.; Li, D.H.; Albertson, T.M.; Connor, J.T. A Bayesian adaptive phase I design to determine the optimal dose and schedule of an adoptive T-cell therapy in a mixed patient population. *Contemp. Clin. Trials* **2016**, *48*, 153–165.
92. Conaway, M.R.; Wages, N.A. Designs for phase I trials in ordered groups. *Stat. Med.* **2017**, *36*, 254–265. doi: 10.1002/sim.7133.
93. Wages, N.A.; Read, P.W.; Petroni, G.R. A phase I/II adaptive design for heterogeneous groups with application to a stereotactic body radiation therapy trial. *Pharm. Stat.* **2015**, *14* (4), 302–310.
94. Salter, A.; O’Quigley, J.; Cutter, G.R.; Aban, I.B. Two-group time-to-event continual reassessment method using likelihood estimation. *Contemp. Clin. Trials* **2015**, *45*, 340–345.
95. Chen, G.; Zeng, D.; Kosorok, M.R. Personalized dose finding using outcome weighted learning. *J. Am. Stat. Assoc.* **2016**, *111* (516), 1509–1521.
96. Zhou, X.; Mayer-Hamblett, N.; Khan, U.; Kosorok, M.R. Residual weighted learning for estimating individualized treatment rules. *J. Am. Stat. Assoc.* **2017**, *112* (517), 169–187.
97. Guo, B.; Yuan, Y. Bayesian phase I/II biomarker-based dose finding for precision medicine with molecularly targeted agents. *J. Am. Stat. Assoc.* **2017**, *112* (518), 508–520.
98. Mao, X.; Cheung, Y.K. Sequential designs for individualized dosing in phase I cancer clinical trials. *Contemp. Clin. Trials* **2017**, *63*, 51–58.

Oncology Trials: Development Strategy

Zhaoyang Teng and Yi Liu

Takeda Pharmaceuticals Inc., Cambridge, Massachusetts, U.S.A.

Jing Yang

Gilead Sciences, Foster City, California, U.S.A.

Zhaowei Hua

Takeda Pharmaceuticals Inc., Cambridge, Massachusetts, U.S.A.

Mark Chang

Veristat LLC, Southborough, Massachusetts, U.S.A.

Abstract

Clinical development of new therapies is a complex process that requires considerable time and resources. The phase 3 trial failure rate is particularly high in oncology (oncology new drug application (NDA) approval rate is only 40%). In order to streamline and expedite the drug development for oncology and more efficiently identify the right population with right regimen for the patients around the world, innovated study designs should be implemented for drug developments including oncology trials. Adaptive design provides a more flexible framework for clinical development that can help increase the efficiency for use of available resources. Increasingly, drug makers are interested in using adaptive design trials to accelerate clinical development and speed new medicines to patients in need. Clinical trials using a biomarker strategy may result in smaller study sizes and higher probability of trial success with the right target patient population which in turn reduces development risk and generates higher return on investment. Multi-Regional Clinical Trials (MRCT) is a more efficient strategy to expedite global clinical development and facilitate drug registration globally. We provide some commonly used design strategies for oncology drug development, with main focus on adaptive design, biomarker-driven study design, and study design considerations for MRCT. We believe that adaptive design, biomarker utilization, and MRCT are three integral parts of a clinical development strategy for oncology drug.

Keywords: Oncology trial; Adaptive design; Biomarker driven design; Multi-Regional Clinical Trials (MRCT); Type I error rate.

INTRODUCTION

Clinical development of new therapies is a complex process that requires considerable time and resources. The phase 3 trial failure rate is particularly high in oncology (oncology new drug application (NDA) approval rate is only 40%). Gan et al.^[1] reviewed 235 published phase 3 randomized clinical trials (RCTs) for oncology and reported 62.5% of trials did not achieve statistical significant results. There is clearly a need for more innovated study designs to improve the success rate of oncology drugs.

Adaptive design provides a more flexible framework for clinical development that can help increase the efficiency for use of available resources. Adaptive design can be used to improve drug development process across different stages. For exploratory trials, the use of adaptive design can help the sponsor make better decisions to reduce phase 3 failure rate, potentially expose fewer patients to less effective treatment for unnecessary toxicity, and shorten

clinical development timeline by combining trials from different stages. Adaptive design in the pivotal trial setting allows the sponsor to further improve the design using interim trial results without losing statistical validity and integrity of the trial.

Traditional RCTs focus on the average treatment effect in the overall population by testing the treatment in an unselected or untargeted population (i.e., all-comer design). However, many new anticancer agents are molecularly targeted and might only benefit a subgroup of patients. Efficient development of new anticancer drugs might, therefore, need to use clinical trial designs that take into account response heterogeneity between different subgroups. In oncology, predicted biomarkers are used to develop classifier to identify patients who are more likely to respond to the investigational therapies. For a well-established biomarker, a common strategy is to enroll and demonstrate treatment benefit in marker-positive (denoted as g+) patients only (i.e., enrichment design). However,

this study design will lose the opportunity for studying marker-negative (denoted as g-) patients and thus overall population in case the drug works for the entire population. In addition to all-comer design and enrichment design for clinical trial with a well-established biomarker, some other more efficient biomarker-driven designs are available to enhance the power of selecting the right therapy for the right patient. Thus, different type of biomarker-driven designs and enrollment strategies could be considered based on different degrees of confidence in biomarkers and disease prevalence.

With the increase of globalization of drug development, more and more pharmaceutical industries are conducting Multi-Regional clinical trials (MRCT) to support a new drug's safety and efficacy across different populations. Commonly, it takes a few years or even more than a decade to develop a new oncology drug and make it available to patients. If the population is not included in the original study, "bridging study" is usually required in order to gain approval in the emerging market, which causes the "drug lag" problem. A carefully designed MRCT could be used to support the new drug's approval in different regions/countries simultaneously. By conducting MRCT for oncology drug development, we can bring new effective drug to patients globally as fast as possible. It also provides a pathway for a regulatory agency to evaluate the drug's safety and efficacy profile based not only on the overall patient population but also the local region.

In this entry, we provide some commonly used design strategies for oncology drug development, with main focus on adaptive design, biomarker-driven study design, and study design considerations for MRCT.

ADAPTIVE STRATEGIES FOR ONCOLOGY TRIALS

Sample size reestimation (SSR) design is one of the most commonly used adaptive designs. The idea is to use interim analysis results to modify the final sample size of the trial. Besides SSR adaptive design, group sequential design and drop-the-loser design are increasingly used in oncology trials. In group sequential design, a trial can be stopped prematurely if there are safety or efficacy issues and depending upon the results of the interim analysis, additional modifications can be made. In drop-the-loser design, the subjects detected to have received inferior treatments at the interim analysis can be dropped out. Based on the findings of the interim analysis, additional treatment arms can also be added at this stage (Chang and Wang^[2]). This design is useful in phase II or phase II/III clinical development with same study objective and endpoint between phase II and phase III, especially when there are uncertainties regarding the dose levels. Typically, drop-the-loser design is a two-stage design. At the end of the first stage, the inferior arms will be dropped based on

some prespecified criteria. The winners will then proceed to the next stage. Some prefer to term these designs as pick-the-winner designs (Mahajan and Gupta^[3]). Chow and Chang^[4] summarized most of the commonly used adaptive designs for clinical trials. In this section, we mainly focus on the SSR design; adaptive design strategies with established biomarker will be introduced and discussed in next section.

SSR Design

If only blinded results (based on aggregate data without unblinding treatment arms) are used, there is generally no impact on the integrity of the final study results. For a typical time-to-event endpoint in oncology trials, increasing the number of patients enrolled in order to achieve the prespecified event driven analysis is generally accepted by regulatory agencies. This is well reflected in their guidances (Food and Drug Administration, FDA, Guidance for Industry:^[5] Adaptive design clinical trials for drugs and biologics; FDA Guidance for Industry and Food and Drug Administration Staff:^[6] Adaptive design for medical device clinical studies; European Medicines Agency (EMA) Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design^[7]).

On the other hand, using unblinded interim results for sample size, reestimation has gained popularity in practice due to several attractive features, especially when there is a greater uncertainty about the treatment effect. It generally requires smaller upfront investment compared to group sequential design or fixed design. With group sequential design or fixed design, treatment effect size is often conservatively assumed which leads to a larger study size to begin with. With this design, the original study size is generally smaller with an optimistic assumption on the treatment effect size.^[8] If the optimistic assumption turns out to be true, the new drug can be brought to patients quicker with a smaller trial. On the other hand, if the original assumption is too optimistic, this design allows for sample size increase to improve the probability of success.

The downside of the design is the potential type I error inflation. As shown by Proschan and Hunsberger,^[9] certain adaptation rules can inflate the type I error for this type of design by two times or more. Multiple papers have proposed ways to control the type I error in this setting. Three main approaches have been observed in the literature.

- One is to construct a combination test statistic by combining the test statistics (or p values) from separate stages: Bauer and Kohne^[10] proposed Fisher's product; Lehmacher and Wassmer^[11] and Cui, Hung, and Wang,^[12] proposed weighted sum of test statistics with fixed weights. This type of approach is very attractive because there is no restriction on the adaptations as long as the conditional invariance principle (Brannath,

Koenig, and Bauer^[13]) is followed. However, it is criticized for not satisfying one patient one vote rule.

- The second way is called the conditional error function approach (Proschan and Hunsberger^[9]) where this function is predefined and integrates (with respect to the interim results) to the prespecified level of type I error rate. The key idea is that as long as the adapted procedure yields the same value as the conditional error probability given interim results, the overall type I error is controlled by definition. A natural choice of the conditional error function is the one based on the planned design (Schäfer and Müller,^[14] Müller and Schäfer,^[15] and Müller and Schäfer^[16]). Once the conditional error probability is determined based on the interim results, a new critical value can be calculated to preserve the same conditional error probability (or equivalently one can use the conditional error probability as the new “alpha” for the next stage of the trial). A potential interpretation problem can rise if the calculated new critical value is more liberal than the conventional cutoff.
- The combination test and conditional error approach appear to be different in terms of implementation. However, if one focuses on a typical rejection rule comparing a test statistic with a critical value, it is easy to observe that each method is modifying one side of the inequality that should be equivalent mathematically as shown in the study by Proschan.^[17]
- The third approach is to leave the conventional test statistic and critical value unchanged but to modify the adaptation rule to maintain the type I error at the desired level (Brannath, Koenig, and Bauer^[13]). This approach is practically attractive as it does not require modification of the rejection rule. Two of the most commonly used adaptation rule is shown in Fig. 1, where x-axis is the conditional power at interim analysis and y-axis is the ratio of final adapted sample size to the original planned final sample size. For ways to modify

the adaptation rule in order to maintain overall type I error, the studies by Mehta and Pocock^[18] and Liu and Hu^[19] are referred.

Specific Issues with Time-to-Event Endpoint

Time-to-event endpoint is commonly used for oncology trials, and there are some specific issues that need to be considered. In a typical 2 stage SSR design, analysis timing is typically driven by the total number of events. If we calculate 1st stage test statistic based on patients enrolled before interim analysis and 2nd stage test statistic based on patients who are enrolled after the interim analysis, then data on those patients who are enrolled before interim and experience event between interim and final analysis will be lost. Fortunately, the increments between the 1st stage test statistic and the test statistic based on all enrolled patients at final analysis are stochastically independent as was shown for the classical group sequential designs. Specifically, assume d_1 is the total interim events and d_2^* is the final adapted total events. Let LR_k denotes the log-rank test statistic at stage k based on the cumulative data up to that stage. Then the incremental test statistic can be calculated by:

$$\tilde{Z}_2^* = \frac{\sqrt{d_2^*} LR_2 - \sqrt{d_1} LR_1}{\sqrt{d_2^* - d_1}} \quad (1)$$

Wassmer^[20] proposed an inverse normal combination test statistic that uses the independent increments structure of the log-rank test statistic. This method allows for many kinds of design modifications during the trial. Lehmacher and Wassmer^[11] also proposed a repeated confidence interval based on the combination of the independent increments of the log-rank test statistic.

It was shown that combination test and conditional error approaches are ways to handle adaptations at interim

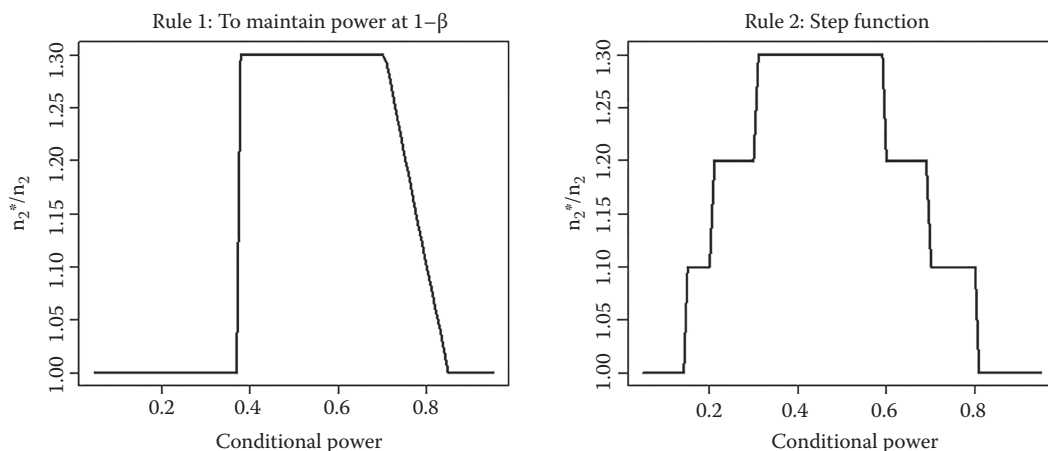


Fig. 1 Different adaptation rules for SSR design, n_2^* and n_2 are the final adaptive sample size and the original planned sample size, respectively

without alpha penalty. However, there are pitfalls in the time-to-event endpoint setting where simple application of these methods may still result in type I error inflation. As pointed out by Bauer and Posch,^[21] if the interim decision uses a surrogate endpoint such as progression free survival (PFS) that is correlated with survival, then for patients who have progressed but still alive at the interim and died before the final analysis, they will contribute to the 2nd stage overall survival (OS) event analysis (not merely determining the sample size). This violates the independence condition (between the first data and second stage data), thus also the conditional invariance principle for most methods (such as the combination test) to control type I error rate.

In order to address this issue, Jenkins, Stone, and Jennison^[22] proposed several alternative testing strategies for OS at final analysis in the more complicated subgroup selection setting with PFS as the surrogate endpoint. One preferred strategy that is proven to control overall type I error is to use the weighted test statistics for OS at final analysis (FA). The key is to determine what data for which set of patients go into the 1st and 2nd stage test statistic, respectively. The patients will be separated into two stages depending on whether they were used for PFS analysis (including patients that did not experience a PFS event) at interim. The log-rank test statistics for OS data at FA among the 1st (2nd) stage patients then forms the 1st (2nd) stage test statistic. This way of separating patients is to ensure independence between the two stage test statistics.

BIOMARKER-DRIVEN DESIGN WITH ESTABLISHED BIOMARKERS

We assume that a biomarker has been established through advances in the scientific and medical field or suggested by early phase data. Adopting the biomarker into the phase III trial design can potentially save a trial when treatment does not show expected efficacy in the overall population. Prospectively planning the biomarker testing strategy can potentially reduce the number of trials required (as retrospective findings usually requires another confirmatory trial) and thus shorten the development timeline and bring the effective treatment to the right subset of patients sooner.

We will first discuss “Phase III Fixed Design Strategy” where only final analysis is performed and then “Phase III Adaptive Design Strategy” where an interim analysis is planned to guide the design for the 2nd stage of the trial. These two biomarker strategies are based on a baseline biomarker where the marker status is obtained before receiving treatment.

Phase III Fixed Design Strategy

The fixed design strategy refers to a design with one final analysis. The main idea behind all the methods proposed in this section is to formally test both overall population

and biomarker-defined subgroup(s). The initial thinking in the literature (e.g., the study by Wang, O’Neill, and Hung^[23]) was to include the testing of g+ patients in addition to the overall population as g+ subgroup is assumed to have a much larger treatment effect. In terms of multiplicity adjustment, methods such as fixed sequence testing (either overall test followed by g+ subgroup or testing g+ subgroup followed by the overall test), alpha split, Hochberg method, and stepdown method incorporating correlation between the subgroup and the overall population are applicable in this setting. Which method to use depends on the degree of confidence in the biomarker (in terms of the differential treatment effect between g+ and g– subgroups) as well as the prevalence of the subgroup.

The following illustrates a few commonly seen designs with different enrollment strategies and testing procedures.

Stratified Designs: Enroll Overall with Biomarker Stratification

A biomarker has strong credentials when evidence for the predictive ability of the biomarker is convincing enough to assume that the treatment is more likely to be effective (and is probably more effective) in the g+ than in the g– subgroup, but the evidence is not sufficiently compelling to rule out a clinically meaningful effect in g– patients. In this scenario, a biomarker-stratified trial design is a preferred option.

Biomarker-stratified designs randomly assign both g+ and g– patients to the treatments under investigation (Fig. 2). Similar to enrichment designs, the biomarker-stratified designs must include a definitive assessment of the treatment in the g+ subgroup. Several possible statistical testing strategies can be used in biomarker-stratified designs.

Testing Strategy 1: Test g+ and Overall Sequentially

First, the treatment effect is tested in g+ patients using significance level α (Fig. 3); if the results of this hypothesis test are significant, then the treatment effect is also tested in overall patients, using the same α value.

An alternative analysis approach is to test treatment effect in overall patients using significance level α first,

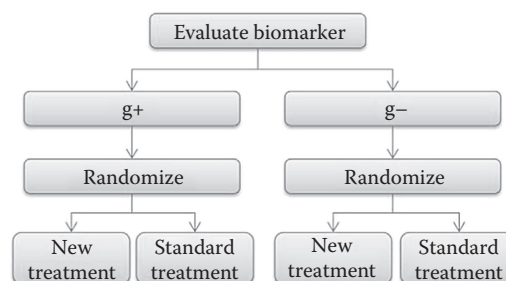


Fig. 2 Enrollment strategy for biomarker-stratified design

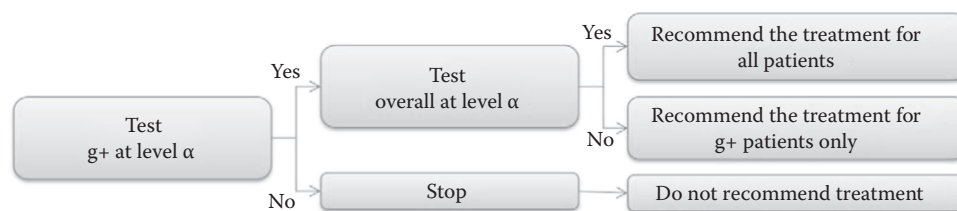


Fig. 3 Testing procedure for sequential g+ and overall design

if the results of this hypothesis test are significant, then the treatment effect is also tested in g+ patients, using the same α value.

Testing Strategy 2: Test g+ and Overall in Parallel

A parallel testing strategy with the biomarker subgroup-specific trial design is also possible; in the parallel version, testing of g+ subgroup and overall population is done simultaneously. However, to control the overall type I error rate of the design at level α , the parallel version requires allocating the overall α between the two tests, which makes statistical significance harder to achieve in the g+ subgroup.

Testing Strategy 3: Test g+ and g–

First, the treatment effect is tested in g+ patients using significance level α (Fig. 4); if the results of this hypothesis test are significant, then the treatment effect is also tested in g– patients, using the same α value.

A parallel testing strategy with the biomarker subgroup-specific trial design is also possible; in the parallel version, testing of the biomarker-positive and biomarker-negative subgroups is done simultaneously. Most of the multiple comparison procedure, e.g. Hochberg, Holm, Fall back, exhaustive Fall back, etc., needed to be used to control the overall type I error rate.

Usually, it is easier to claim statistical significant to test g+ and overall comparing to test g+ and g–. However, a key concern with the g+ and overall strategies is that when the benefit of the new therapy is restricted to the g+ patients, the design might inappropriately recommend the treatment for the g– subgroup, even though the treatment is ineffective in these patients. The overall efficacy could be highly driven by g+ subgroup, a statistically significant effect can still be observed in the overall population if its effect in the g+ patients is large enough even with no treatment effect in

the g– subgroup. So it seems logical to assess the treatment effect in this g– population after seeing a significant benefit in the g+ population. However, it is very challenging to achieve statistical significant in both g+ and g– subgroup due to reduced sample size in subgroup and smaller effect size in the g– subgroup by assumption in practice. The adaptive design described in the next section might be one of the solutions to conquer these challenges in which the ineffective subgroup will be dropped at interim and only the effective subgroup will continue to next stage and tested at the final analysis.

Testing Strategy 4: Test g+ and g– and Overall

The main reason for using a g+/overall population assessment strategy is that the test in the overall population allows implicit borrowing of information from the g+ subgroup in providing a treatment recommendation for the g– patients. In principle, this could be an attractive approach when the treatment effect is homogeneous across the biomarker-defined subgroups, as long as steps are taken to control the probability of incorrectly recommending the treatment for the g– patients if it does not work for them.

Unlike the g+/overall population strategy, this control is achieved by the Marker Sequential Test (MaST) design (Freidlin, Korn, and Gray^[24]). The MaST design (Fig. 5) incorporates analyses of g+ and g– subgroups as well as the overall population. First, the g+ subgroup is tested at a reduced significance level (α_1), which can be any value less than the desired overall significance level α of the trial. If the results of this test are significant, the g– subgroup is tested at significance level α . If the results in the g+ subgroup are not significant, the overall population is tested at significance level $\alpha - \alpha_1$. If this overall population test is significant, the treatment is recommended for the whole population; if this overall population test is not significant, the treatment is not recommended for any patients regardless of biomarker status.

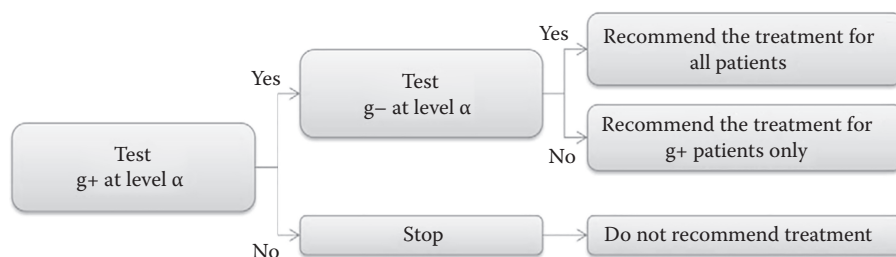


Fig. 4 Testing procedure for sequential g+ and g– design

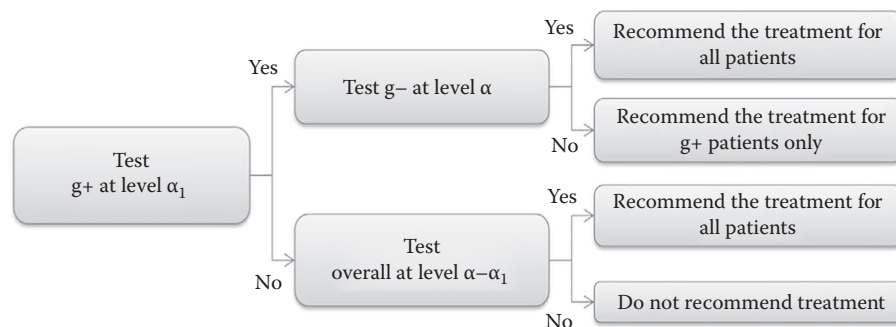


Fig. 5 Testing procedure for MaST design

This strategy only tests the overall population when no significant effect is detected in the g+ subgroup; it reduces the risk of erroneously concluding that the treatment is beneficial in the overall population when positive findings in the overall group are driven by a benefit limited to g+ patients.

Modified Stratified Design: Selectively Enroll g+ Patients

As FDA guidance has explicitly requested to include g- patients in enrichment design, or the potential of overall treatment effect in all-comer population, an enrichment design that randomizes g+ patients only may not be sufficient. On the other hand, if an all-comer population is randomized, the efficiency associated with the enrichment trial design feature is then lost. It would become a challenge, particularly when g+ is a relatively small subpopulation that a large number of g- patients have to be enrolled into such a trial in order to be able to recruit a sufficient number of g+ patients.

In order to conquer these challenges, especially when the disease prevalence is low, modified stratified designs (Yang et al.^[25]) might be an option to solve these problems, which enroll all g+ patients, but selectively enroll g- patients as illustrated in Fig. 6.

Since the patients enrolled in this study design do not represent the real disease prevalence in overall population, the following weighted estimator could be used to estimate the treatment effect for overall population

$$\hat{\delta} = \hat{p}\hat{\delta}_+ + (1 - \hat{p})\hat{\delta}_- \quad (2)$$

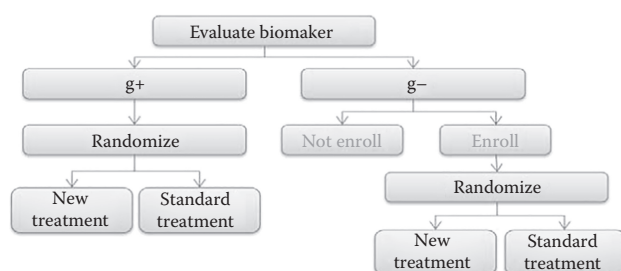


Fig. 6 Enrollment strategy for modified stratified design

where $\hat{\delta}_+$ and $\hat{\delta}_-$ are estimators of treatment effects in g+ and g- subgroups, respectively; $\hat{p}$ is the proportion of g+ patients obtained at screening phase. Theoretically, most of the testing procedures described in the section of stratified designed can be applied to modified stratified design, e.g. sequential test (g+, overall), Hochberg, Holm, fall back, exhaustive fall back, etc.

Phase III Adaptive Design Strategy

For the illustration purpose, we mainly focus on two-stage adaptive design when the best possible patient population to study is uncertain. The adaptive design strategy incorporates two stages where an interim look is prospectively planned. At the end of stage 1 (i.e., interim analysis), the prespecified decision is to maintain the original study plan or to modify the plan in terms of recruitment and testing strategy at the final analysis. All issues discussed in the previous section on fixed design strategy still apply to this section. However, we will focus more on the differences in the decision rules at interim to guide design for the second stage of the trial.

General setting for the adaptive design is to evaluate efficacy in overall and predefined g+/- subgroup with interim adaptation in two aspects:

1. enroll overall or only g+
2. test overall with/without g+ or test g+ only at the FA

Several papers fall in this setting and have proposed different interim decision rules (Wang, O'Neill, Hung;^[23] Brannath et al.^[26] Jenkins, Stone, and Jennison;^[22] and Stallard et al.^[27]).

Two sets of interim decision rules are built to facilitate patient selection in stage 2 and determine the hypotheses to be tested at the final analysis. One rule, decision rule 1, is to evaluate the overall population, the predefined g+/- subgroups separately in terms of futility and efficacy. The other rule, decision rule 2, is to compare the relative difference in efficacy between the overall population and the subgroups to determine patient selection in stage 2, while using the same futility rule as decision rule 1. Both of these two decision rules can be rooted in treatment effect, condition power, and predictive power. These three

decision parameters have equivalent performance in making interim decisions.

For illustration purpose, treatment effect hazard ratio (HR) is used to describe the decision rules. The fundamental idea of decision rule 1 is similar to Brannath et al.^[26] Decision rule 1 directs a pivotal trial to proceed into stage 2 and the final analysis as below:

1. If $HR_o \geq c_1$ and $HR+ \geq c_2$ (overall bad, g+ bad), then stop for futility
2. If $HR_o \geq c_1$ and $HR+ < c_2$ (overall bad, g+ good), then enrich to g+ patients in stage 2 and test g+ only at final analyses
3. If $HR_o < c_1$ and $HR+ \geq c_2$ (overall good, g+ bad), then continue to enroll overall population in stage 2 and test overall at final analyses
4. If $HR_o < c_1$ and $HR+ < c_2$ and $HR- \geq c_3$ (overall and g+ good, g- bad), then enrich to g+ patients only in stage 2 and test g+ only at final analyses
5. If $HR_o < c_1$ and $HR+ < c_2$ and $HR- < c_3$ (consistent, overall, and g+/g- all good), then continue to enroll overall population in stage 2 and test both overall and g+ at final analyses

Decision rule 2 adopts the similar logic as Stallard et al.,^[27] which measures the absolute difference in efficacy effect size between the overall population and the g+ subgroup. Even though decision rule 2 does not explicitly evaluate consistency among the overall population and the g+/g- subgroups, the absolute difference implicitly indicate the consistency between the overall population and the g+ subgroup. A disadvantage of Stallard et al.^[27] is not to evaluate futility simultaneously for the overall population and the g+ subgroup. Decision rule 2 circumvents this problem by using the same futility boundary from decision rule 1. Decision rule 2 will drive the following logics for interim decision making:

1. If $HR_o \geq c_1$ and $HR+ \geq c_2$ (overall bad, g+ bad), then stop for futility
2. If $|HR_o - HR+| > \varepsilon$ and $HR_o \geq HR+$ (overall and g+ not close, g+ better), enrich to g+ in stage 2 and test g+ population only at final analyses
3. If $|HR_o - HR+| > \varepsilon$ and $HR_o \leq HR+$ (overall and g+ not close, overall better), then continue to enroll overall in stage 2 and test overall population at final analyses
4. If $|HR_o - HR+| < \varepsilon$ (overall and g+ close), then continue to enroll overall in stage 2 and test both overall and g+ population at final analyses

An essential element of SSR for adaptive design is also built into the internal adaptive design strategy. Conditional power could be used for SSR.

Enrichment to the g+ subgroup, interim analysis, and sample size reassessment all contribute to type I error inflation. Inverse normal combination test under closed

test procedure is used to strongly control family-wise type I error rate. Multiplicity control procedures such as Hochberg, Simes, Bonferroni, etc., can be used to obtain adjusted p value for the intersection hypothesis.

MULTI-REGIONAL CLINICAL TRIALS

There is an increasing trend to conduct MRCT for drug development in pharmaceuticals industry. A carefully designed MRCT could be used in supporting the new drug's approval in different regions simultaneously. The primary objective of an MRCT is to investigate the drug's overall efficacy across regions while also assessing the drug's performance in some specific regions. One of the statistical challenges in conducting MRCT is to ensure that overall efficacy can be adequately preserved in regions of interest. To evaluate the possibility of applying the overall results in a MRCT to the specific regions of interest, the Japanese Ministry of Health, Labour and Welfare (MHLW)^[28] proposed two methods for determining the needed number of Japanese subjects for establishing consistency of the treatment effect between the Japanese patients and others:

Method 1: The sample size needed for the Japanese patients in a MRCT was to satisfy:

$$P(D_i/D > \pi_i) > 1 - \beta' \quad (3)$$

where D_i and D are the estimated treatment effects from the group of Japanese patients and the entire patient population, respectively, π_i is the effect retention rate ≥ 0.5 , and β' is the type II error rate ≤ 0.2 .

Method 2: The planned sample size for the Japanese group in a MRCT was to satisfy:

$$P(D_1 > 0, D_2 > 0, \dots, D_s > 0) > 1 - \beta' \quad (4)$$

where D_i represents the observed treatment effect for region i , $i = 1, \dots, s$. Here, the s is to denote the number of regions.

Consistency Requirement for Regional Approval

Based on the Japanese MHLW guidance, a couple of statistical methods were proposed to apply the overall results of the MRCT to the specific region. As mentioned in the study by Chen et al.,^[29] two qualitative methods were usually considered to assess the consistency in treatment effect between each region and the entire trial. The first is to evaluate if the estimated regional treatment effect preserves some proportion of overall treatment effect, i.e., $\hat{\delta}_i > \pi_i \hat{\delta}$, which utilizes the spirit of method 1 in Japanese guidance. Here, the value of π_i should be prespecified. The second is to test if the treatment effect based on the samples from local region is statistically significant at level α_i , i.e., $\hat{\delta}_i > z_{1-\alpha_i} \sqrt{2\sigma^2/N_i}$. Tsong et al.^[30] proposed to determine

the regional type I error rate α_i adjusted for the regional sample size. Teng and Chang^[31] proposed a unified consistency requirement that generalizes the two above consistency requirements. This unified consistency requirement is to test whether the “true” regional treatment effect preserves some proportion of the “true” overall treatment effect at significance level α_i , we can use hypothesis test:

$$H_0: \delta_i \leq \pi_i \delta \text{ versus } H_a: \delta_i > \pi_i \delta \quad (5)$$

Denote Z_i as the test statistics of the hypothesis test above, where $Z_i = (\hat{\delta}_i > \pi_i \hat{\delta}) / \text{std}(\hat{\delta}_i > \pi_i \hat{\delta})$, and then $Z_i > z_{1-\alpha_i}$ is the unified consistency requirement. When $\alpha_i = 0.5$, $Z_i > z_{1-\alpha_i}$ is reduced to $\hat{\delta}_i > \pi_i \hat{\delta}$ and when $\pi_i = 0$, $Z_i > z_{1-\alpha_i}$ is reduced to $\hat{\delta}_i > z_{1-\alpha_i} \sqrt{2\sigma^2 / N_i}$, two parameters π_i and α_i are involved in this consistency requirement, which make it more commonly feasible.

The commonly used primary endpoints in oncology trials are PFS, OS, and response rate, e.g., overall response rate, complete response rate. Thus, binary and survival endpoints are major endpoints for oncology trials. Teng, Lin, and Zhang^[32] compared the different consistency requirements of each type of endpoint: continuous, binary, and survival and provided the general recommendations for each endpoint based on the practical consideration.

Binary Endpoint

Suppose $n_i^l \sim B(N_i, p_i^l)$ is the number of events from the l^{th} treatment group in region i , and $\hat{p}_i^l = \frac{n_i^l}{N_i}$ is the estimated of the event rate p_i^l , $i = 1, \dots, s$ and $l = t, c$ which represent treatment group and control group, respectively. We assume the higher event rate is, the better the result is, e.g., response rate in oncology trials. There are three major measurements of treatment effect for binary endpoint, i.e.:

1. Risk difference: $\widehat{rd}_i = \hat{p}_i^t - \hat{p}_i^c$
2. Relative risk: $\widehat{rr}_i = \hat{p}_i^t / \hat{p}_i^c$
3. Odds ratio: $\widehat{or}_i = \frac{\hat{p}_i^t(1 - \hat{p}_i^c)}{\hat{p}_i^c(1 - \hat{p}_i^t)}$

Similar to continuous endpoint, the overall treatment effect is estimated from the weighted average of regional treatment, i.e., $\hat{b} = \sum_{i=1}^s (f_i \hat{b}_i)$, where $\hat{b} = \widehat{rd}, \widehat{rr}, \widehat{or}$ and $\hat{b}_i = \widehat{rd}_i, \widehat{rr}_i, \widehat{or}_i$.

If the first type of consistency requirement is imposed for regional approval, similar to continuous endpoint, the criterion below could be considered as consistency requirement for risk difference:

$$\widehat{rd}_i > \pi_i \widehat{rd} \quad (6)$$

Two consistency requirements could be considered for relative risk, the first one is the risk reduction; the second one is based on the log scale of relative risk.

1. Risk reduction: $(\widehat{rr}_i - 1) > \pi_i (\widehat{rr} - 1)$
2. Log scale of relative risk: $\log(\widehat{rr}_i) > \pi_i \log(\widehat{rr}) \Leftrightarrow \widehat{rr}_i > \widehat{rr}^{\pi_i}$

When comparing the two consistency requirements above, it is easy to prove that the consistency requirement of risk reduction is more stringent than log scale of relative risk with the same value of π_i between 0 and 1 when the overall result is positive, i.e., $\widehat{rr} > 1$. Thus, using the log scale of relative risk as consistency requirement is preferred since it is more powerful to preserve the consistency of treatment between the local region and overall result. In terms of odds ratio, the consistency requirement proposed for relative risk could be applied to odds ratio directly and the property of the two consistency requirement still holds.

Time-to-Event Endpoint

As mentioned in the study by Quan et al.,^[33] the proportional hazards model below is usually considered for survival endpoint.

$$\lambda_i(t) = \lambda_0(t)e^{\gamma} \quad (7)$$

where $\lambda_i(t)$ and $\lambda_0(t)$ are the hazard functions for treatment and control group, respectively; e^{γ} is the HR between treatment and control group. The distribution of regional treatment effect $\hat{\gamma}_i$ for balanced two-group design is asymptotic:

$$\hat{\gamma}_i \sim N\left(\gamma_i, \frac{4}{E_i}\right) \quad (8)$$

where E_i is the total number of events in region i . Assuming that the overall treatment effect for the entire MRCT is the weighted combination of regional treatment effect and the event proportion is considered as weight for survival endpoint, the overall treatment effect is $\hat{\gamma} = \sum_{i=1}^s \left(\frac{E_i}{E}\right) \hat{\gamma}_i$. The distribution of the overall treatment effect is asymptotic:

$$\hat{\gamma} \sim N\left(\sum_{i=1}^s \frac{E_i}{E} \gamma_i, \frac{4}{E}\right) \quad (9)$$

Two consistency requirements could be considered for survival endpoint, the first one is hazard reduction; the second one is based on the log scale of HR.

1. Hazard reduction: $(1 - e^{\hat{\gamma}_i}) > \pi_i (1 - e^{\hat{\gamma}}) \Leftrightarrow (1 - \widehat{HR}_i) > \pi_i (1 - \widehat{HR})$
2. Log scale of HR: $\hat{\gamma}_i < \pi_i \hat{\gamma} \Leftrightarrow \widehat{HR}_i < \widehat{HR}^{\pi_i}$

When comparing the two consistency requirements above, it can be proven that the consistency requirement of log scale of HR is more stringent than hazard reduction with the same value of π_i between 0 and 1 when the overall result is positive, i.e., $\widehat{HR} < 1$.

One or more than one consistency requirements are available for the three measurements of binary endpoint and survival endpoint. We summarize the consistency requirement for different endpoints in Table 1 and also indicate which consistency requirement is preferable in general for each endpoint/measurement. One should have better chance to demonstrate the consistency in treatment effect by imposing the recommended consistency requirement when the true underlying treatment effects are in fact consistent between the local region and overall population.

Most of the proposed consistency requirements utilize the spirit of method 1. However, Ikeda and Bretz^[34] point out an issue with the consistency requirement defined by method 1 as illustrated in the left-hand side of Fig. 7. Shaded area represents a consistent treatment effect for the region ($\hat{\delta}_i > \pi_i \hat{\delta}$) and a significant overall treatment ($\hat{\delta} > \delta_i$). However, area A has a larger overall treatment effect and a larger regional treatment effect compared to area B, but it fails to meet the consistency defined by method 1, i.e., $\hat{\delta}_i > \pi_i \hat{\delta}$. Thus, it is not reasonable to approve a drug in area B but reject a drug in area A.

Hu^[35] also mentioned that a drug should be approved in the region if the regional treatment effect in the sub-population is so large that it is significant at the normal level of one-sided 0.025 regardless of whether the overall treatment effect is significant. Therefore Hu^[35] proposed a new regional requirement which is to satisfy any one of the following three conditions:

1. Significant overall effect $\hat{\delta} > \delta_i$ and consistent regional effect $\hat{\delta}_i > \pi_i \hat{\delta}$
2. Significant overall effect $\hat{\delta} > \delta_i$ and regional effect above a certain threshold $\hat{\delta}_i > \delta_{i1}$
3. Significant regional effect $\hat{\delta}_i$ at the normal level of one-sided 0.025.

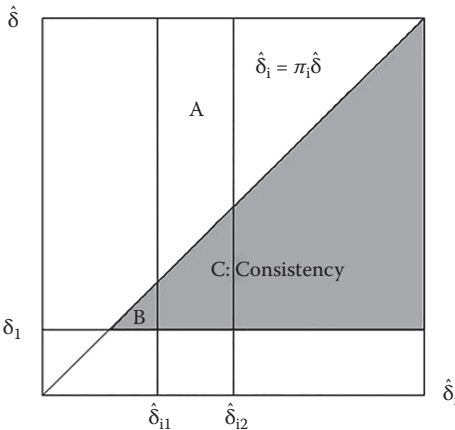


Table 1 Summary of consistency requirement for different endpoint

Endpoint	Regional requirement	Preferred in general
Continues:	Preserve π_i proportion of overall efficacy: $D_i > \pi_i D$ Regional significance at α_i : $D_i / \text{std}(D_i) > z_{1-\alpha_i}$ Unified consistency requirement: $(D_i - \pi_i D) / \text{std}(D_i - \pi_i D) > z_{1-\alpha_i}$	Yes
Binary: Risk difference	Preserve π_i proportion of overall efficacy: $rd_i > \pi_i rd$	Yes
Binary: Relative risk	Risk reduction: $(rr_i - 1) > \pi_i (rr - 1)$ Log scale of relative risk: $\log(rr_i) > \pi_i \log(rr)$	Yes
Binary: Odds ratio	Similar to relative risk	
Survival:	Hazard reduction: $(1 - e^{\hat{\gamma}_i}) > \pi_i (1 - e^{\hat{\gamma}})$ Log scale of HR: $\hat{\gamma}_i < \pi_i \hat{\gamma}$	Yes

Shaded area in right-hand side of Fig. 7 represents the reasonable approval area based on Hu's^[35] proposal.

Optimal Design for MRCT

In an MRCT, overall power is the probability of success for entire trial under the alternative hypothesis. In terms of regional success, the concept of assurance probability is usually used to measure the probability of regional approval when consistency requirement is imposed to a local region. If we consider $Z_i > z_{1-\alpha_i}$ as regional

Fig. 7 $\hat{\delta}_i$ and $\hat{\delta}$ are the estimated treatment effect and the overall treatment effect, respectively; δ_{i2} and δ_i are the statistical significant thresholds (at one-sided 0.025 level) for the regional effect and overall treatment effect, respectively; δ_{i1} is the threshold of a positive trend for the regional effect

consistency requirement, the assurance probability of region i is defined as follows:

$$AP_i = P_{\mu} (Z_i > z_{1-\alpha_i} | Z)_{z_{1-\alpha}} \quad (10)$$

The optimal goal of an MRCT is to demonstrate the treatment efficacy in all regions of interest simultaneously with adequate power when the global treatment efficacy is already proven. Thus, in order to achieve this optimal goal, the total sample size increase is sometimes ineluctable. However, different sample size allocations will lead to different total sample size increases. The optimal designs for an MRCT require appropriate allocation of the total samples to each region so that we can ensure certain overall power and assurance probabilities for all regions of interest with as smaller sample size as we can.

Two optimal designs for MRCT were proposed by Teng and Chang:^[31] 1) Minimal total sample size (MTSS) design and 2) Maximal utility (MU) design.

MTSS Design

The MTSS design is to find the minimal sample size required and the corresponding proportion for each region to ensure both desired overall power and assurance probabilities for all regions of interest, i.e.:

$$\text{Minimize: } N(\text{with the corresponding } f_i, i = 1, \dots, s) \quad (11)$$

Subject to:

$$\text{Power} \geq 1 - \beta \quad (12)$$

$$AP_i \geq 1 - \beta_i, i = 1, \dots, j \quad (13)$$

where $1 - \beta$ is the desired overall power; $1 - \beta_i$ is the desired assurance probability for region i ; $j(\leq s)$ is the number of regions of interest.

There is no constraint on total sample size for the MTSS design. However, in practice, the determination of total sample size may highly depend on the budget for the study; therefore, even though we can calculate the MTSS with the corresponding sample size proportion for each region to achieve all design requirements, the MTSS may end up to be huge, making it unfeasible for sponsors to implement. The MU design introduced below could be used with fixed/limited sample size.

MU Design

The goal of the MU design is to find the optimal sample size allocation that can maximize the global utility of the regions of interest on the premise of guaranteeing the desired overall power with the fixed total sample size, i.e.:

$$\text{Maximize: } U = \sum_{i=1}^j M_i AP_i \times (\text{with the corresponding } f_i, i = 1, \dots, s) \quad (14)$$

Subject to:

$$\text{Power} \geq 1 - \beta \quad (15)$$

$$N = N_0 \quad (16)$$

where $1 - \beta$ is the desired overall power and N_0 is the fixed total sample size that could be the maximal sample size the sponsor can afford. Here, M_i is the utility weight for region i which could be related to the number of patients or commercial viability of this region, and $\sum_{i=1}^j M_i = 1$, $j \leq s$ is the number of regions of interest.

Under the fixed total sample size, the overall power is not a good criterion to assess the performance of different sample size allocations for MRCT. As the chosen sample size allocation that can maximize the overall power may adversely affect the ability to obtain regional approval for some regions of interest, the proposed MU design by Teng and Chang^[31] is to maximize the global benefits of the drug by considering regional approval.

Adaptive Design for MRCT

The adaptive design for MRCT is more complex than that of the conventional adaptive clinical trial. During the interim analysis, we not only need to reestimate the total sample size to guarantee the overall power but also need to determine the new sample size allocation of the next stage to ensure a certain level of regional success based on the observed data. For the conventional adaptive design, the interim decision is only determined by one observed effect size; but for adaptive MRCT, we need to incorporate the observed effect size of each region in an MRCT to make decisions for both entire MRCT study and each individual region. In general, three options are available for the conventional adaptive design at interim: stop for efficacy if the observed effect size is in favorable zone; continue to the next stage if the observed effect size is in promising zone; stop for futility if the observed effect size is in unfavorable zone. In terms of the adaptive design for MRCT, the three options are applicable to both entire MRCT study and each individual region. Thus, we not only need to determine whether to stop the entire MRCT or continue to the next stage, but we also need to make the decision for each individual region whether we can stop either for efficacy or futility or continue to the next stage. Teng and Chang^[36] proposed a region-level adaptive design for MRCT including 3 steps as follows. The flow chart of this adaptive design is illustrated in Fig. 8. The study by Teng and Chang^[36] is referred for more details with regard to the statistical framework of the proposed adaptive design including the definition and the calculation of conditional power, conditional assurance probability, conditional success rate, and how to control type I error appropriately.

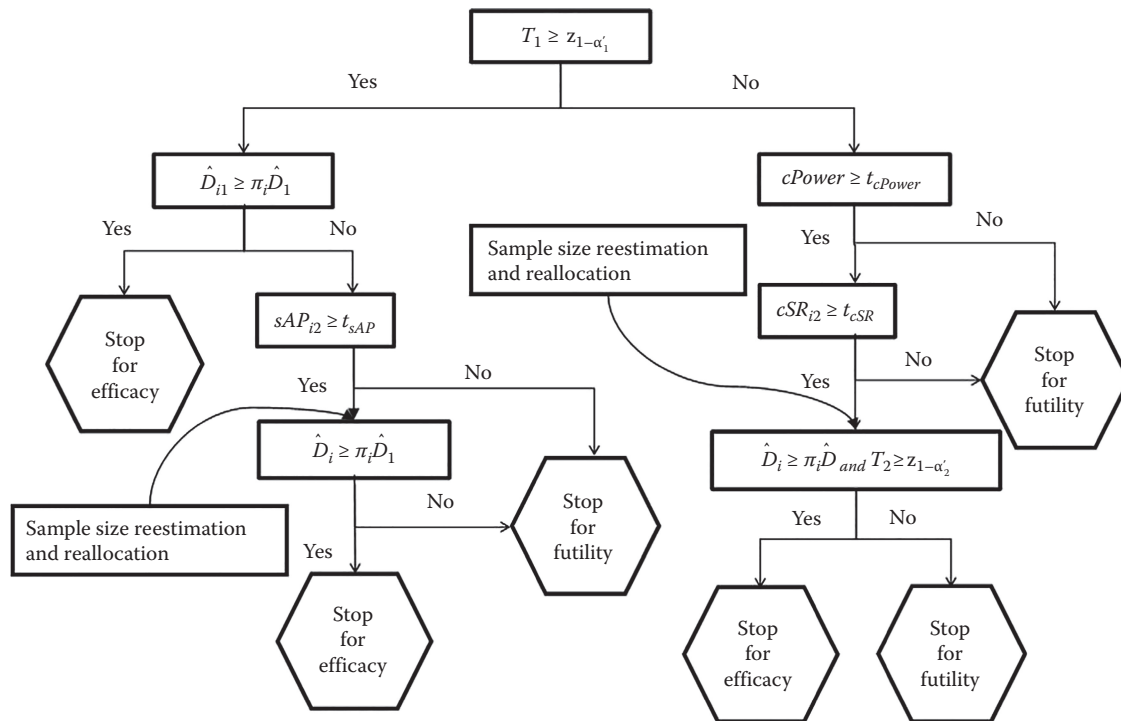


Fig. 8 The flow chart of the region-level adaptive design

Step 1: Initial sample size planning

The two optimal sample size allocation designs proposed in the previous section could be used for the initial sample size planning.

Step 2: Decision-making and sample size adaptation at interim

Scenario 1: If the test statistic at stage 1 is rejected at the significance level of this stage and all regions satisfy the prespecified consistency requirement, then the entire MRCT by claiming regional efficacy for all regions is stopped.

Scenario 2: If the test statistic at stage 1 is rejected at the significance level of this stage and some regions satisfy the consistency requirement, then these regions for efficacy are stopped. For the other regions in which the consistency requirement is not satisfied, for each region, either to stop at stage 1 without pursuing a claim in this region or to enroll more patients to ensure that this region satisfies the consistency requirement at next stage is decided.

Scenario 3: If the entire MRCT shows little efficacy at interim, then the entire MRCT for futility is stopped. It is not necessary to test the consistency requirements for each region in this case.

Scenario 4: If the overall result is in the promising zone, then whether it is necessary to continue to the next stage for each region based on conditional success rate is evaluated. If the conditional success rate is smaller than a given threshold, then this region will be dropped at stage

1 and the sample size planning for stage 2 will not involve this region anymore. Second, SSR and reallocation will be performed to determine the total sample size and the corresponding sample size proportions for the remaining regions to achieve both the desired conditional power and desired conditional assurance probabilities.

Step 3: Final decision

At the end of MRCT, claim the regional efficacy each as long as the overall result is significant, and the consistency in treatment effect between this region and the overall efficacy is also proven.

CONCLUSION

In order to streamline and expedite the drug development for oncology and more efficiently identify the right population with right regimen for the patients around the world, innovated study designs should be implemented for drug developments including oncology trials. We believe that adaptive design, biomarker utilization, and MRCT are three integral parts of a clinical development strategy for oncology drug.

Adaptive designs improve the chance and shorten the time that a safe and effective drug actually reaches the patient. They can also terminate an ineffective drug earlier—either in the whole, a specific patient population, or at a certain dose. Increasingly, drug makers are

interested in using adaptive design trials to accelerate clinical development and speed new medicines to patients in need. Clinical trials using a biomarker strategy may result in smaller study sizes, higher probability of trial success with the right target patient population which in turn reduces development risk and generates higher return on investment. MRCT is a more efficient strategy to expedite global clinical development and facilitate drug registration globally compared to bridging study and multiple regional clinical trials. The ultimate goal of MRCT is to bring new drug to patients globally as fast as possible and reduce the drug lag. MRCT also helps in expansion of clinical research into developing countries bringing medical care options to subjects who otherwise may not have access to them.

Finally, careful planning, regulatory agreements, and practical implementation of these innovative designs are the key to make the speedy and successful drug development a reality.

REFERENCES

- Gan, H.K.; You, B.; Pond, G.R.; Chen, E.X. Assumptions of expected benefits in randomized phase III trials evaluating systemic treatments for cancer. *J. Natl. Cancer Inst.* **2012**, *104*, 590–598.
- Chang, M.; Wang, J. The add-arm design for unimodal response curve with unknown mode. *J. Biopharm. Stat.* **2015**, *25* (5), 1039–1064.
- Mahajan, R.; Gupta, K. Adaptive design clinical trials: Methodology, challenges and prospect. *Indian J. Pharmacol.* **2010**, *42*, (4), 201–207.
- Chow, S.C.; Chang, M. *Adaptive Design Methods in Clinical Trials*, 2nd Ed.; Chapman and Hall/CRC Press: New York, 2008.
- FDA Draft Guidance. Adaptive design clinical trials for drugs and biologics. *Biotechnol. Law Rep.* **2010**, *29* (2), 173.
- FDA. *Guidance for Industry and Food and Drug Administration Staff. Adaptive Design for Medical Device Clinical Studies*. 2016. Available at <https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm446729.pdf> (accessed March 13, 2017).
- Committee for Medicinal Products for Human Use. *Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design*; EMEA: London, 2007.
- Jennison, C.; Turnbull, B.W. Mid-course sample size modification in clinical trials. *Stat. Med.* **2003**, *22*, 971–993.
- Proschan, M.A.; Hunsberger, S.A. Designed extension of studies based on conditional power. *Biometrics* **1995**, *51*, 1315–1324.
- Bauer, P.; Kohne, K. Evaluation of experiments with adaptive interim analyses. *Biometrics* **1994**, *50*, 1029–1041.
- Lehmacher, W.; Wassmer, G. Adaptive sample size calculations in group sequential trials. *Biometrics* **1999**, *55*, 1286–1290.
- Cui, L.; Hung, J.H.M.; Wang, S.J. Modification of sample size in group sequential clinical trials. *Biometrics* **1999**, *55*, 853–857.
- Brannath, W.; Koenig, F.; Bauer, P. Multiplicity and flexibility in clinical trials. *Pharm. Stat.* **2007**, *6*, 205–216.
- Schäfer, H.; Müller, H.H. Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Stat. Med.* **2001**, *20*, 3741–3751.
- Müller, H.H.; Schäfer, H. A general statistical principle for changing a design any time during the course of a trial. *Stat. Med.* **2004**, *23*, 2497–2508.
- Müller, H.H.; Schäfer, H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **2001**, *57*, 886–891.
- Proschan, M.A. Sample size re-estimation in clinical trials. *Biometrical J.* **2009**, *51*, 348–357.
- Mehta, C.R.; Pocock, S.J. Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Stat. Med.* **2011**, *30* (28), 3267–3284.
- Liu, Y.; Hu, M.X. Testing multiple primary endpoints in clinical trials with sample size adaptation. *Pharm. Stat.* **2016**, *15*, 37–45.
- Wassmer, G. Planning and analyzing adaptive group sequential survival trials. *Biometrical J.* **2006**, *48*, 714–729.
- Bauer, P.; Posch, M. Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections (Letter to the Editor). *Stat. Med.* **2004**, *23*, 1333–1335.
- Jenkins, M.; Stone, A.; Jennison, C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharm. Stat.* **2011**, *10*, 347–356.
- Wang, S.J.; O'Neill, R.T.; Hung, H.M.J. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharm. Stat.* **2007**, *6* (3), 227–244.
- Freidlin, B.; Korn, E.L.; Gray, R. Marker sequential test (MaST) design. *Clin. Trials* **2014**, *11* (1), 19–27.
- Yang, B.; Zhou, Y.; Zhang, L.; Cui, L. Enrichment design with patient population augmentation. *Contemp. Clin. Trials* **2015**, *42*, 60–67.
- Brannath, W.; Zuber, E.; Branson, M.; Bretz, F.; Gallo, P.; Posch, M.; Racine-Poon, A. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Stat. Med.* **2009**, *28* (10), 1445–1463.
- Stallard, N.; Hamborg, T.; Parsons, N.; Friede, T. Adaptive designs for confirmatory clinical trials with subgroup selection. *J. Biopharm. Stat.* **2014**, *24* (1), 168–187.
- Ministry of Health, Labor, and Welfare. *Basic Principles on Global Clinical Trials*; Ministry of Health, Labor, and Welfare: Tokyo, Japan, 2007.
- Chen, X.; Lu, N.; Nair, R.; Xu, Y.; Kang, C.; Huang, Q.; Li, N.; Chen, H. Decision rules and associated sample size planning for regional approval utilizing multiregional clinical trials. *J. Biopharm. Stat.* **2012**, *22* (5), 1001–1018.
- Tsong, Y.; Chang, W.J.; Dong, X.Y.; Tsou, H.H. Assessment of regional treatment effect in a multiregional clinical trial. *J. Biopharm. Stat.* **2012**, *22*, 1019–1036.
- Teng, Z.; Chang, M. *Optimal Multiregional Clinical Trials. Chapter 11. Multi-regional Clinical Trials for*

- Simultaneous Global New Drug Development*; CRC Press: Boca Raton, FL, 2016.
32. Teng, Z.; Lin, J.C.; Zhang, B. *Evaluation of Consistency Requirements in Multi-Regional Clinical Trials with different Endpoints. Part IV Innovative Clinical Trial Designs and Analysis. Statistical Applications from Clinical Trials and Personalized Medicine to Finance and Business Analytics*; Springer: Gewerbestrasse, 2016.
 33. Quan, H.; Li, M.; Chen, J.; Gallo, P.; Binkowitz, B.; Ibia, E.; Tanaka, Y.; Ouyang, S.P.; Luo, X.; Li, G.; Menjoge, S. Assessment of consistency of treatment effects in multi-regional clinical trials. *Drug Inf. J.* **2010**, *44* (5), 617–32.
 34. Ikeda, K.; Bretz, F. Sample size and proportion of Japanese patients in multi-regional trials. *Pharm. Stat.* **2010**, *9* (3), 207–216.
 35. Hu, M.X. Multiregional clinical trials in oncology drug development. In *Multi-regional Clinical Trials for Simultaneous Global New Drug Development*; CRC Press: Boca Raton, FL, 2016.
 36. Teng, Z.; Chang, M. Adaptive multiregional clinical trials. In *Multi-regional Clinical Trials for Simultaneous Global New Drug Development*; CRC Press: Boca Raton, FL, 2016.

Onset of Action

Thomas R. Stiger and Christy Chuang-Stein

Pfizer Global Research and Development, Groton, Connecticut, U.S.A.

Abstract

In this entry, we discuss four approaches that have been used to assess onset of action. These four approaches address the onset question very differently. Some tackle the question directly with the help of a definition of onset at the individual level; others compare treatment effect at specific time points.

Keywords: Fast acting; Time to response; Survival analysis; Dynamic response; Mixture models.

INTRODUCTION

Fast onset of beneficial effect is often a very desirable attribute for a pharmacological agent. The time frame that would constitute fast onset depends on the disorder and the expectation of an effective therapy for the disorder. For analgesics, for example, fast onset of pain relief is gauged in terms of minutes after administration of drug, whereas for most psychiatric disorders, time to substantial improvement for nearly all therapies takes much longer and requires multiple dosing over several days. Although many psychiatric disorders—such as depression—are chronic in nature, these conditions are often associated with acute exacerbations in which there are significant impairment and possible life-threatening (e.g., suicide) behavior. This impairment to social and occupational functioning can be great, resulting in disruption to family and social networks, as well as lost productivity. Costly treatment interventions, such as admission to an acute psychiatric ward for stabilization, are often required to ensure the safety of the patient or of those around him. The latter contributes to the economic burden of the syndrome. As a result, pharmacological interventions that could lead to rapid relief of the most distressing symptoms, or shorten the time to response or remission of the overall syndrome, could bring great value to the management of patients afflicted with these disorders. This is especially so in a resource-constrained environment.

In this entry, we discuss four approaches that have been used to assess onset of action. These four approaches address the onset question very differently. Some tackle the question directly with the help of a definition of onset at the individual level; others compare treatment effect at specific time points. While not dismissing any approach as inappropriate, we point out what type of questions these approaches can and can't answer, along with their advantages and disadvantages. To facilitate comparisons

of approaches, we provide a high-level summary of these approaches side by side in [Table 1](#).

STATISTICAL CONSIDERATIONS AND APPROACHES

Leon^[1] argued that four critical questions must be addressed in any valid analysis of comparative onset of therapeutic effect among antidepressant drugs. They are (1) how is onset defined, (2) how is onset measured, (3) how the treatment groups can be compared statistically with regard to onset of effect, and (4) how these issues affect protocol designs. Even though Leon discussed the above questions in the context of antidepressants, these four questions are often equally applicable when comparing onset of action among treatments for many other disorders.

In our discussion, we use the terms “response” and “action” interchangeably, even though “response” is used to describe the experience of a subject, while “action” is used to describe the effect of an intervention (e.g., treatment). Because a treatment's action is ultimately established through the patient's response to the treatment, we feel comfortable treating these two terms interchangeably without making an effort to distinguish between them. On those occasions when we use “onset of action” to represent the point when the treatment begins to have a clinically perceptible benefit, and “response” to represent a clear and meaningful improvement, we will explicitly state that in the text. In the latter case, a response, if confirmed, occurs after the onset of action.

Analysis at a Prespecified Early Time Point

In many clinical trials, repeated measurements of the primary endpoints are collected at earlier times despite the fact that the primary analysis is a landmark analysis at a

Table 1 A High-Level Summary of Four Approaches As well As Their Advantages and Disadvantages for Analyzing Onset of Action

Approach	When to Use	Advantages	Disadvantages
Approach #1: Analysis at a prespecified early time point	When one is interested in a different treatment effect at specific time points	It is simple	Does not define onset and does not address the distribution of onset directly
	A single time point is defined and accepted as being early	It is applicable to both a continuous and categorical endpoint (e.g., responder status)	Requires defining what time point is early and of interest
	Multiple time points are specified	It is capable of addressing sustained response in a responder-like analysis	Need to adjust for multiplicity if there are multiple “early times” unless one is willing to test sequentially When sustainability for a continuous outcome is defined by statistical significance at all subsequent time points, this procedure will suffer from a reduced power The approach is not robust to missing data unless missing can be assumed to occur completely at random
Approach #2: Repeated measures models	When one is interested in a different treatment effect at specific time points	Can handle continuous and categorical endpoints	Does not define onset and does not address the distribution of onset directly
	A single time point is defined and accepted as being early	Uses all the available data	Requires defining what time point is early and of interest
	Multiple time points are specified	Is robust to missing data as long as one can assume “missing at random”	Hard to interpret when treatment effect changes from positive to negative over visits
	It is desirable to see how the comparative treatment effect might vary over time	Sensitive for detecting nonparallel profiles	Using a slope to describe treatment effect might over-simplify the effect profile
	When one is interested in the rate of improvement and equates a higher rate with an earlier onset		Can’t address the missing data situation when using this approach for sustained response based on a binary outcome
Approach #3: Survival analysis	Onset of action could be clearly defined and time to onset can be ascertained for each subject even if it is only known to occur within an interval determined by the study	One can estimate the distribution of the onset time for each treatment	Requires a definition for “onset of action”
	When testing whether one treatment has a faster onset than another	At the population level, one can compare the onset distributions and demonstrate that time to onset is shorter in one treatment compared to another	Dichotomizing a continuous variable may be arbitrary and will result in a loss of sensitivity
		Makes use of all data up to the point of dropout for individuals who drop out	Assumption of independence of censoring and response may be questionable, especially if “sustainability” is required Inference is sensitive to combined effect of time to onset and proportion reaching onset

(Continued)

Table 1 A High-Level Summary of Four Approaches As well As Their Advantages and Disadvantages for Analyzing Onset of Action (Continued)

Approach	When to Use	Advantages	Disadvantages
Approach #4: Survival analysis using mixture models	Onset of action could be clearly defined and time to onset can be ascertained for each subject even if it is only known to occur within an interval determined by the study	Can simultaneously estimate the probability of ever reaching onset and time to onset among those reaching onset	Requires a definition for “onset of action”
	When testing whether one treatment has a faster onset than the other between two comparably efficacious treatments (in terms of overall response)	Allows one to make inferences on both questions of interest	Dichotomizing a continuous variable may be arbitrary and will result in a loss of sensitivity
	A substantial number of subjects will never experience onset if they haven’t reached onset by trial end	Makes use of all data up to the point of dropout for individuals who drop out	Assumption of independence of censoring and response may be questionable, especially if “sustainability” is required
		Considerably more sensitive than ordinary survival analysis when there is a comparable yet non-trivial proportion of non-responders	Trial may not be long enough to definitively rule out future response for those who did not respond before the end of the trial

chosen time point at or near the end of the treatment period. It is not unusual for these repeated measurements to be taken at closer intervals earlier in the trials. When a time point signaling early onset can be defined, one can compare treatment groups at the prespecified time point. A statistically significant difference between treatment groups at the early time point suggests differential treatment effect as early as the time point identified.

If the outcome variable of interest is continuous or nearly so (e.g., an ordinal scale with at least five categories), a simple linear model [e.g., one- or two-way analysis of variance (ANOVA) or analysis of covariance (ANCOVA)] can be used to compare treatment means at the specific time point. These linear models are efficient and robust to moderate deviations from the usual assumptions of normality and homogeneity of variance (i.e., maintains the Type I error rate at or below the specified level). If the outcome variable of interest is dichotomous (e.g., response or no response, event or no event), one can compare the proportion of responders between treatments via a chi-square type of test or by logistic regression. The latter allows for the adjustment with covariates considered to be associated with the response variable.

There is an increasing trend to conduct a responder analysis under which a dichotomous variable is created from a continuous one, using a cutoff value to define a response. Although dichotomization often decreases our ability

to detect a difference in the treatment effect because of the loss of information,^[2] the practice of defining a target value for treatment benefit at the individual patient level has made the responder analysis a popular one from the clinical perspective in recent years.^[3]

When a significant difference is identified at an early time point specified a priori, it is often felt that the difference needs to be sustained (at subsequent time points) to be clinically meaningful and to reduce the chance of a spurious finding. Although this requirement seems reasonable from the clinical perspective, it creates a statistical multiplicity problem in that the probability of reaching statistical significance at all subsequent time points could be substantially less than that at a single time point. This is especially so if the correlations among the test statistics at different time points are low. Besides, the analysis at each subsequent time point will be affected by dropouts and how dropouts are handled. As a result, one needs to address these issues when employing this approach.

When the primary endpoint is binary, another option is to compare the proportion of subjects who achieved a responder status at the early time point and remained so for the remainder of the study. A less stringent criterion might be to look at the proportion of individuals who achieved “onset”—for example, a 30% improvement (from baseline) at the early time point and maintained this improvement throughout the study, even though a true response required

a 50% reduction. The choice of the percentage improvement in this case might be based on what constitutes a minimal clinically meaningful improvement.

One prerequisite of the time-point analysis is our ability to define in advance the early time point of interest. If multiple time points are specified with the hope that one of them will represent the beginning of a differential treatment effect, adjustments for multiple comparisons need to be considered if the objective is to claim, via hypothesis testing, differential treatment effects for at least one of the chosen early time points. Alternatively, if one begins by testing at the latest time point and proceeds to the earlier time points backward only if statistical significance is declared for the former, one can conduct each comparison at the desirable level (e.g., 5%, as this is a special case of a step-down approach). Implicit in this approach is the assumption that one is interested in sustained response.

It should be pointed out that a statistically significant difference at an early time point implies that the drug has a different effect at that time point from the comparator (e.g., the placebo). Because *p*-values are highly affected by sample sizes, one should follow a statistically significant difference by examining the clinical relevance of the observed difference. Comparing mean changes at prespecified time points does not necessarily provide information on the actual time of onset of action at the individual level. For this reason, findings from this approach are typically interpreted at the population level.

Repeated Measures Analysis

Repeated measures analysis includes all observed data in a single model. A popular one is a likelihood-based, mixed-effects repeated measures model.^[4] The fixed effects typically include terms such as treatment, investigator, time (visit), treatment by time interaction, baseline response, and possibly baseline by time interaction. Under this model, comparisons at specific time points can be conducted using contrasts in treatment group means. One special feature of this model is the inclusion of a treatment by time interaction term. This interaction term, when significant, suggests that the mean response profiles for the treatment groups are not parallel over time. This plus point estimates of the treatment group means at each time point can be used as evidence for differential treatment effect at specific time points.

The repeated measures models have a within-subject variance component that is usually smaller than the between-subject variance component. The former can be used to test the time by treatment interaction, resulting in a more sensitive test for the interaction. The likelihood-based, mixed-effects repeated measures model is an effective and robust method for dealing with dropouts and missing data when the mechanism for missingness is dependent only on the observed values (i.e., not dependent on the values of the missing measurements).^[5] Repeated

measures analysis may also be more appropriate relative to other methods (e.g., survival analysis) when assessment visits are unevenly spaced and far apart in time.

The repeated measures model is simplified considerably when time is considered as a quantitative variable and the treatment by time interaction is reduced to fitting a separate slope parameter for each treatment group over time, suggesting a linearly monotonic response over time within each treatment group (i.e., a separate linear regression line is fitted for each treatment group). Because response to treatment often plateaus after a period of time, the model containing a slope parameter for each treatment group would be the most sensible when all the time points included in the analysis are during the period when the response continues to change approximately in a linear fashion.^[1] The latter is not an issue if there is only one post-baseline evaluation. When fitting a separate slope to each treatment group is appropriate, one can infer that the rate of change differs between treatment groups starting early on, suggesting that the group with a steeper rate of change has an earlier onset of action.

The concept of repeated measures models is also applicable to categorical data. Leon^[1] and Brannan et al.^[6] used a categorical repeated measures model to assess the probability of onset at each visit, with onset defined as achieving a certain degree of improvement from the baseline. These authors have also used the repeated measures models to look at the probability of being in a sustained improvement state at each visit. Because sustained improvement requires a subject to stay in the study for the entire treatment duration, repeated measures analysis that incorporates the sustained improvement requirement mitigates the utility that the usual repeated measures model has for handling missing data.

Similar to the analysis at a prespecified early time point, repeated measures analysis does not address the question of onset of action directly. The latter is a drawback of all analyses that test for differences in the mean changes without comparing the change amount with that required to qualify as an “action” on the part of the treatment. This drawback is rectified by the approaches included in the next two sections.

Survival Analysis

Survival analysis is by far the most commonly used analysis to address the onset of action. To use this analysis, one needs to define what qualifies as an “action” of the treatment or a “response” on the patient receiving the treatment. When a definition is given, we can look at the time required to achieve the response. Each individual has a time measurement that corresponds to the actual onset of response if he or she achieved a response, or the last assessment time if he/she failed to achieve a response. In the latter case, the onset of action is said to be censored. Because the onset time can vary among individuals, the distribution

of the onset time among a group of subjects becomes the point of interest in this analysis. Survival analysis is commonly used and is still appropriate when the actual onset of response is not known precisely, such as when measurements are taken only at prespecified time intervals. This analysis can benefit, however, from multiple measurements obtained at reasonably selected time intervals.

The Kaplan–Meier (K–M) Product-Limit method^[7] is the chosen method for estimating the distribution of the time to onset. This distribution is usually denoted as $S(t)$ and is referred to as the survival function: $S(t) = \text{Prob}\{T > t\}$. More explicitly, $S(t)$ is the probability that the time-to-onset random variable T for any individual given a certain treatment is greater than some constant real number t . For individuals who dropped out of the study, this method includes information from such individuals until the point of dropout. Using the K–M method, one can estimate $S(t)$ or, equivalently, the proportion of subjects in each treatment that will have achieved onset by any time t . In Fig. 1, for example, the K–M curve for treatment “A” indicates that it is estimated that 30% of the subjects treated with “A” will experience onset of action by Day 10. It also shows that Day 29 is the estimate in which 50% of the patients treated with “A” will have experienced onset of action. If a greater than $z\%$ response rate by time t qualifies for an early onset of action at the population level, the survival analysis could be used to decide whether the criterion is met.

When applying the K–M approach, one typically looks at the time when a subject achieved his/her first response without requiring such response to be sustaining. There are situations when sustaining response is required to qualify for a treatment action. In such a case, time to onset could be defined as the time to the first response among those with a sustained response and will be censored at the last observation time point for those who did not. There are also situations in which the onset could be defined by a lesser improvement as long as the required improvement was obtained later in the trial. For example, it is

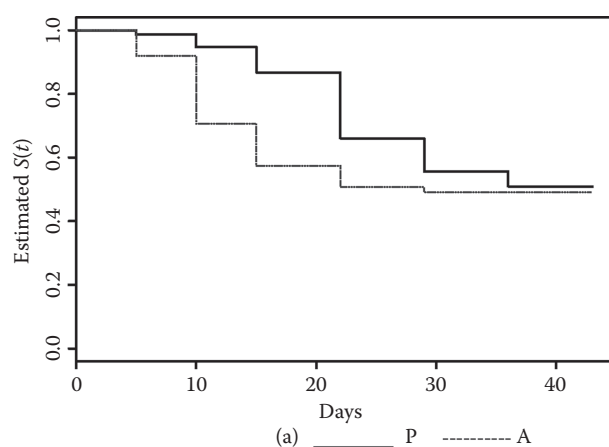


Fig. 1 K–M estimates of $S(t)$ for two treatments

not unusual to require a 50% reduction in the Hamilton Depression (HAM-D) score (from baseline) to qualify as a responder in a depression trial. However, one can define the first time a 30% improvement is realized as the onset of action for those who eventually obtained a 50% reduction in the HAM-D score. Brannan et al.^[6] focused their time-to-onset analysis on a 30% sustained improvement in a depression trial of duloxetine using the survival type of analysis.

The approach of using different thresholds for the initial and the eventual improvement described above is similar to the stopwatch approach typically used in assessing the onset of action of analgesics in treating acute pain. For individuals who eventually received meaningful pain relief, the first time they experienced a noticeable decrease in pain (as registered using a stopwatch) is used as the time of onset for the analgesic under testing at the individual level.

To compare two treatment groups with respect to their distribution of onset of action under the survival analysis, two nonparametric tests (i.e., logrank and Wilcoxon test) are often used. Between these two, the Wilcoxon test is more sensitive to differences occurring earlier along the time scale, whereas the logrank test is more sensitive to later differences. To take into account important covariates that are thought to be associated with the onset time, semiparametric methods (e.g., Cox’s proportional hazards model) and parametric models (e.g., the Weibull distribution) can be used for both estimation and inference.

When p -values from the above comparisons are significant, one can conclude that one treatment has a faster onset than the other. This information alone does not convey information on important characteristics of the distribution of onset of action. Using the estimated time-to-onset curve, however, one can estimate the percent of individuals who have achieved onset of action by a given time. Conversely, one can estimate the time by which a given percentage of patients will have achieved onset of action. One can estimate, for example, the time by which at least 50% of the patients will have achieved onset of action if the curve has crossed the 50% mark. Depending on the disorder, this estimated curve might qualify the treatment as one with an early onset of action for the underlying disorder.

When comparing the onset of action between treatment groups, a concern with a less frequent assessment schedule is that one might miss a small but meaningful improvement in onset of action for a new drug due to heavier interval censoring. Intuitively, more frequent assessments could increase the power of a trial to detect early onset if a drug in fact has the ability to produce a more rapid response than its comparator early on. The potential increase in power from utilizing a daily assessment schedule instead of a weekly schedule was evaluated in a simulation study by Tamura et al.^[8] using distributions that fit historical time-to-response data based on the HAM-D. Tamura et al. reported only an incremental increase in power when assessments were

conducted daily instead of weekly. Because of the small increase in power, Tamura et al. questioned whether the small increase was worth the added cost and complexity of daily assessments.

Survival Analysis Using Mixture Models

Assuming comparable dropout patterns in different treatment groups, comparing the distributions of time to onset between groups using the survival analysis described above is influenced by two factors: the difference in onset rates and the difference in time to onset among those who will experience onset. When two treatment groups have comparable onset rates and similar dropout patterns, comparing time to onset using the survival analysis is similar to comparing time to onset among those who will experience onset. On the other hand, when a significant difference in the onset distributions is concluded, this difference could be due to a difference in onset rates, a difference in the time to onset among those who will achieve onset, or a combination of both. In other words, the survival approach, when applied to all subjects in the group, does not address the question of time to onset exclusively. For this reason, the method in this subsection is the most appropriate for comparing a new treatment with an active control when the two treatments are expected to produce a high percentage (but not 100%) of patients who experience onset of action.

The usual application of survival analysis treats nonresponders as censored at the last assessment time point and assumes that the nonresponders would eventually have responded had the trial or follow-up period been longer. Further, the usual analysis assumes that nonresponders are in the tail portion (which is beyond the end of the trial) of the same time-to-response distribution as the responders. Such assumptions are often inappropriate when analyzing clinical trials of central nervous system (CNS) disorders, because some individuals will never respond to a treatment no matter how long they receive it. Taking depression as an example, it is generally believed that if a subject hasn't experienced a response by Week 6 or Week 8, he or she is not likely ever to experience the desirable treatment effect even if the treatment were continued beyond Week 8 (Tamura et al.).^[8]

Because of the above considerations, some researchers have felt that one ought to consider two properties when characterizing the efficacy of a treatment in many therapeutic areas. The two properties involve the proportion who will respond and the time to response conditional on achieving a response. Or if the interest is in onset instead of a more complete response, the two properties are the proportion of subjects who experience onset and the time to onset conditional on achieving onset. Because the population of interest consists of a mixture of responders and nonresponders, models that address these two-in-one analyses are often called mixture models.

In terms of nonparametric approaches, Laska and Siegel^[9] outline how one can use K-M estimators to estimate the proportion of responders (p) and the distribution of time to response conditional on achieving a response, denoted as $S^*(t)$. The K-M estimates of $S^*(t)$ is based on the following relationship: $S^*(t) = [S(t) - (1-p)]/p$. Fig. 2 is a plot of the two estimated $S^*(t)$ s (one for each treatment) vs t using the estimated $S(t)$ s contained in Fig. 1, where $1-p$ for each treatment is estimated by the estimated $S(t)$ s at Day 42. If one assumes or concludes that the proportions of responders for the two treatments are equal, one can use the logrank or Wilcoxon test for comparing the conditional time-to-response distributions. Unfortunately, neither test is very powerful—especially the logrank test when there is a substantial proportion of nonresponders. Orazem^[10] has proposed a weighted logrank test that is more powerful in the mixture setting. This test was constructed to have maximum sensitivity when the conditional hazards are proportional. When the nonresponse rate is 40%, for example, this test may have as much as 90% power when the logrank test has a power of only 20%.^[11]

When the proportions of responders are not the same, a generalized logrank test has been proposed.^[10] Its derivation, however, assumes no ties in the onset times, and its distributional assumptions are suspect for sample sizes smaller than 100 per group. A Cramer-von Mises statistic has been proposed^[8] for comparing two conditional time-to-response distributions regardless of whether the response proportions are the same between the treatment groups. The power of this method is comparable to Orazem's logrank test^[11] when there are proportional conditional hazards, and is likely to be superior under other circumstances. The methodology is computationally intensive because it relies on bootstrapping methods.

One can use a parametric approach to combine two models.^[12] One model estimates the probability of achieving a response, often done through logistic regression. The second model describes the distribution of time to response conditional on experiencing a response. The choice of

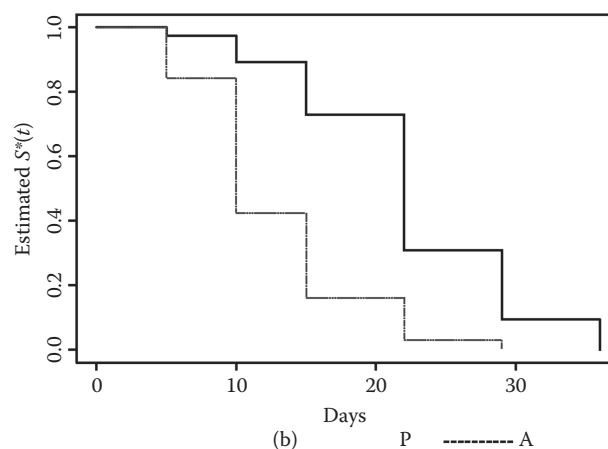


Fig. 2 K-M type estimates of $S^*(t)$ for two treatments

distribution depends on how well the characteristics of various distributions fit the data. Two common choices are the Weibull and the log-logistic, both of which are accelerated “failure” time models. The latter means that the onset time for one treatment is a multiple of the onset time of another. The Weibull model imparts the assumption of proportional hazards, whereas the log-logistic imparts the assumption of proportional odds. These various assumptions may not be tenable, but they can be checked to some degree using the data. When the assumptions are reasonable and the models are appropriate, the parameters often provide clinically useful interpretations. To test whether the conditional time-to-onset distribution is less in one treatment compared with another, one can use likelihood ratio tests. These models can also accommodate important covariates considered to be associated with the response.

Semiparametric mixture models that are an extension of the Cox proportional hazards model have been proposed and evaluated (Peng and Dear;^[13] Sy and Taylor^[14]). Although these models appear to have some nice properties with fewer distributional assumptions, the power of such models for testing the equality of the conditional time-to-onset distributions is often inferior to that of parametric approaches.

CONCLUSIONS

In this entry, we reviewed several popular statistical and methodological approaches to analyzing onset of action. In general, approaches that look only at the first occurrence of a response without addressing subsequent responses do not take into account the dynamic nature of a patient’s response over time. If the latter is important, one should consider using the repeated measures approach or applying survival analysis to the onset of a sustained response. Whether the ensuing analysis leads to a conclusion of an early onset of action depends on the definition of early onset, as well as on the therapeutic efficacy expected of the medications used to treat the underlying disorders.

As demonstrated in this entry, many definitions can be used to define onset of action. The definition should be based on what makes sense clinically, considering the disorder under investigation and the mechanism of action of the treatment. When the definition of an onset of action is determined, statistical methods can be used to characterize the distribution of the onset time among a group of subjects.

Statistical approaches and methods for assessing the onset of action should be dictated by the specific objectives of the drug program, as well as by the underlying questions. If the product profile specifically mandates efficacy at Week 1, for example, a comparison with placebo at Week 1 in a randomized study generally would be warranted, with the hope of demonstrating superiority at that time point. On the other hand, if the product profile dictates a faster efficacy than a standard treatment with (ideally)

a comparable final response rate [e.g., new antidepressant vs generic selective serotonin reuptake inhibitors (SSRI)], a time-to-response methodology would be appropriate. In the latter case, a definition of a response is necessary to conduct the analysis.

In addition to the above two considerations, the choice of statistical approaches and methods for evaluating the onset of action is likely to depend on where the candidate is in the development program. If the goal is to obtain some favorable statements in the label (e.g., sustained efficacy over 8 weeks beginning at Week 1), comparisons with the comparator are likely to be based on conservative methodologies agreed upon in advance with the regulatory agencies. On the other hand, if the intent is to provide some credible evidence in support of early onset for a publication or to assist internal decisions, there should be more flexibility in choosing the analytic approaches.

With a clear onset definition and a prespecified analytical approach, one’s ability to differentiate between treatment groups with respect to the onset of action is likely to be a function of the dose, as well as a function of how the treatment is given. In a trial in which the study drug is given in an escalating fashion to minimize adverse drug reactions, for example, one can expect a slower onset. If one wants to know how early the onset of action could be if the drug is given at its fully intended dose(s) from the beginning, one should employ a fixed-dose design wherein patients are randomized to the full dose(s) under investigation.

If one is interested in drawing inference on onset for each individual dose in a confirmatory trial with multiple doses, one will need to employ a multiple comparison procedure to control the overall false-positive rate. One of the most popular multiple comparison procedures for this application probably is the gatekeeping procedure, in which we compare each dose with the comparator according to a prespecified order. If we think that onset is a monotone function of the dose, for example, we would compare the highest dose with the comparator first. If this yields a significant result, we proceed to compare the next-highest dose with the comparator. If the overall Type I error rate is set at the 5% level, each comparison will be conducted at the 5% level, and the comparisons will stop the instant we fail to conclude a significant difference. If we are not sure of the onset-dose relationship, we should consider using other multiple comparison procedures that do not depend on a priori assumption of the onset-dose relationship.

In summary, it is important that one have a clear definition of onset of action and a thorough understanding of how the onset will be measured. The appropriate analytic approach can be selected only after the definition, the data type, and the study objectives have been determined. Given the historical difficulty companies have had in demonstrating rapid onset of action, it is essential that great care be placed on study design and on the selection of the appropriate statistical methodology.

ARTICLES OF FURTHER INTEREST

Analysis of variance; Failure-time model; Mixed effects models; Multiplicity in clinical trials; Survival analysis.

REFERENCES

1. Leon, A. Measuring onset of antidepressant action in clinical trials: an overview of definitions and methodology. *J. Clin. Psychiatry* **2001**, *62* (suppl. 4), 12–16.
2. Senn, S. Disappointing dichotomies. *Pharm. Stat.* **2003**, *2*, 239–240.
3. Lewis, J.A. In defence of the dichotomy. *Pharm. Stat.* **2004**, *3*, 77–79.
4. Laird, N.M.; Ware, J.H. Random-effects models for longitudinal data. *Biometrics* **1982**, *38*, 963–974.
5. Mallinckrodt, C.H.; Kaiser, C.J.; Watkin, J.G.; Molenberghs, G.; Carroll, R.J. Type I error rates from likelihood-based repeated measures analyses of incomplete longitudinal data. *Pharm. Stat.* **2004**, *3*, 171–186.
6. Brannan, S.K.; Mallinckrodt, C.H.; Detke, M.J.; Watkin, J.G.; Tollefson, G.D. Onset of action for duloxetine 60 mg once daily: double-blind, placebo-controlled studies. *J. Psychiatric Res.* **2005**, *39*, 161–172.
7. Kaplan, E.L.; Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **1958**, *53*, 457–481.
8. Tamura, R.N.; Faries, D.E.; Feng, J. Comparing time to onset of response in antidepressant clinical trials using the cure model and the Cramer-von Mises test. *Stat. Med.* **2000**, *19*, 2169–2184.
9. Laska, E.M.; Siegel, C. Characterizing onset in psychopharmacological clinical trials. *Psychopharmacol. Bull.* **1995**, *31*, 29–35.
10. Orazem, J. Rank test to test equality of two survival distributions in cure model Ph.D. thesis, New York, Columbia University, 1990.
11. Stiger, T.R.; Tian, H. Mixture models in the analysis of time to response in antidepressant trials; 2006.
12. Farewell, V.T. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **1982**, *38*, 1041–1046.
13. Peng, Y.; Dear, K.B.G. A nonparametric mixture model for cure rate estimation. *Biometrics* **2000**, *56*, 237–243.
14. Sy, J.P.; Taylor, J.M.G. Estimation in a cox proportional hazards cure model. *Biometrics* **2000**, *56*, 227–236.

Optimal Futility Interim Design

Zhongwen Tang

Novartis, East Hanover, New Jersey, U.S.A.

Abstract

This entry covers two major paradigms for optimal futility interim design—frequentist operational characteristics based approaches and probability of success based approaches. The design paradigms are demonstrated in three different types of trials—single arm trial with a binary end point, randomized trial with a normal end point, and randomized trial with a time-to-event end point. The entry concludes with a comparison of the two-design paradigms and a discussion.

Keywords: Predictive probability of success; Conditional probability of success; Credible interval; Decision rule; Optimal design; Utility function; Bayesian design; Design paradigm.

INTRODUCTION

Now-a-days many clinical trials are designed in a sequential way. An interim analysis (IA) is planned to allow the trial to stop early. Based on the interim data, if a trial is doomed to fail, early stopping due to futility can help to save valuable resource for more meaningful things and spare patients of lesser treatment.

The U.S. Food and Drug Administration (FDA) guidance defines futility as “when the likelihood that the treatment effect being sought, based on the interim data, is very unlikely to be established.”^[1]

The treatment effect being sought depends on the safety profile of the treatment. If the treatment has a clean safety profile, establishing statistical significant efficacy may be enough for registration. However if there are safety concerns, health authorities (HA) will require the magnitude of treatment effect to be larger than statistical significance, i.e., clinically meaningful result is needed to support the registration.

How to design futility interims in an optimal way is the focus of this entry. Since 1 futility IA is commonly used, the discussion will focus on 1 IA scenario. But the paradigms can be extended to more than 1 IA situation.

Based on the design paradigms, current optimal futility interim design can be divided into two categories.

1. Frequentist operational characteristics (FOC) based approaches—optimality is achieved by maximizing/minimizing some utility function with respect to the design parameters, while maintaining the type I error rate and power at a specific treatment effect size.
2. Probability of success (POS) based approaches—optimality is achieved by satisfying some decision rules in terms of POS in an optimal way with respect to the design parameters.

3. The design paradigms are demonstrated in three different types of trials—single arm trial with a binary end point, randomized trial with a normal end point, and randomized trial with a time-to-event end point.

METHODS

FOC-Based Approaches

In Frequentist setting, the treatment effect equals a specific value. Power is the probability of observing statistical significance in the trial. In FOC design paradigm, a utility function, based on an important design dimension such as sample size, is preselected. The optimal design is found via maximizing/minimizing the utility function with respect to the design parameters while maintaining the type I error rate and the power at a specific treatment effect parameter value. This design paradigm can be dated back to several decades to the work done by Weiss^[2] and Lai.^[3] It got popular after the work done by Simon^[4] for futility IA with binary end point to minimize sample size while holding the type I and type II error rates in two stage setting (with only 1 IA). The method can be extended to many slightly different scenarios including different end points (e.g., Chen et al.,^[5] Wason and Mander^[6]), different utility functions (e.g., Chen et al.^[5]), different parameter value (e.g., Mander and Thompson^[7]), more than 1 IA (e.g., Wason et al.^[8]), allowing stopping for efficacy at IA (e.g., Manders and Thompson^[7]), etc.

The FOC design paradigm will be demonstrated in three different types of trials in the following sub sections.

Single Arm Study with Binary End Point

Consider a single arm trial of two stages with sample size n_1 and n_2 in stages 1 and 2, respectively. If the number

of responses at the first stage is r_1 or fewer, the trial is terminated due to futility.

Simon (1989) proposed to optimize the design by minimizing the expected sample size, denoted as EN, under null hypothesis response rate p_0 holding the type I error rate (α) and type II error rate (β). No stopping due to efficacy is allowed at IA.

$EN = n_1 + (1 - PET)n_2$, where PET is the probability of early termination after the first stage. PET depends on the true response rate p . $PET = B(r_1; p, n_1)$, where $B(r_1; p, n_1)$ is the cumulative probability function of a binomial distribution with response rate p , i.e., the probability of observing r_1 or fewer responses out of n_1 observations. If there are total r or fewer responses observed at the end of the trial, the trial fails. Hence the probability of failing is:

$$B(r_1; p, n_1) + \sum_{x=r_1+1}^{\min(n_1, r_2)} b(x; p, n_1)B(r - x; p, n_2) \quad (1)$$

Type I and type II error rates can be computed using this formula with p from H_0 (null hypothesis) and H_1 (alternative hypothesis), respectively.

With prespecified parameters p_0 , p_1 (response rate at H_0 and H_1) α , β , the optimal design is found by minimizing the expected sample size with respect to the design parameters—total sample size n , IA sample size n_1 , cutoff response numbers at the IA (r_1), and final analysis (r).

Simon also proposed to use the maximal sample size as the utility function. The maximal sample size refers to the sample size at the final analysis, i.e., no stopping at IA. This approach leads to the so-called min-max design.

Randomized Two-Arm Study with Normal Outcome

Consider a two-arm randomized trial with a normal end point comparing an experimental treatment and a control treatment. Assume the individual response from control treatment has normal distribution with mean μ_c and variance σ^2 , i.e., $N(\mu_c, \sigma^2)$ and from experimental treatment σ^2 , i.e., $N(\mu_e, \sigma^2)$. $\delta = \mu_e - \mu_c$ is the true difference in the treatment effect between the experimental and control treatments. Bigger δ indicates bigger treatment effect. The stagewise test statistic is a two sample t statistic, denoted as T_1 and T_2 for stages 1 and 2, respectively. T_1 and T_2 are independent of each other. The trial can be stopped due to either, efficacy or futility at IA. If $t_1 < f$, the trial-stops for futility. If $t_1 > e_1$, the trial stops for efficacy. f is the futility boundary in the t test statistic scale, e_1 is the efficacy boundary at the IA in the t -test statistic scale.

Wason and Mander^[6] proposed to find the optimal design by minimizing the expected sample size EN holding the type I and type II error rates.

$$EN = n_1 + (1 - PET)n_2 \quad (2)$$

where $PET = P(T_1 < f) + P(T_1 > e_1)$, and n_1 is the number of patients in the first stage and n_2 is the number of patients in the second stage, if the trial does not stop at IA.

T_1 has noncentral t distribution with noncentrality parameter $\frac{\sqrt{r(1-r)n_1}\delta}{\sigma}$ and degrees of freedom $2n_1 - 2$.

σ is assumed to be known or can be estimated from existing data at design stage. r is the randomization ratio = number of patients in the control arm/total number of patients in both arms.

The test statistic in the final analysis is $T = \frac{\sqrt{n_1 T_1 + n_2 T_2}}{\sqrt{n_1 + n_2}}$. The trial will declare success if $t > e_2$ in the final analysis. The probability of failing to reject H_0 is:

$$\begin{aligned} P(\text{fail to reject } H_0) &= P(T_1 < f) + P(e_1 \geq T_1 \geq f, T \leq e_2) \\ &= P(e_1 \geq T_1 \geq f, T_2 \leq \frac{\sqrt{n_1 + n_2}e_2 - \sqrt{n_1}T_1}{\sqrt{n_2}}) \\ &= \int_{-\infty}^{(\sqrt{n_1 + n_2}e_2 - \sqrt{n_1}f)/\sqrt{n_2}} \int_f^{e_1} g(t_1)h(t_2)dt_1 dt_2 \end{aligned} \quad (3)$$

where $g(t_1)$ and $h(t_2)$ are the density functions of T_1 and T_2 , respectively. Type I and type II error rates can be calculated using this formula.

Wason et al.^[8] also proposed to use the maximum expected sample size parameter with respect to the treatment effect size as the utility function. This leads to the so-called δ -min-max design.

Randomized Trial with Time to Event End Point

Consider a two-arm randomized trial with time-to-event end point comparing an experimental treatment and a control treatment. Chen et al.^[5] proposed to optimize the interim design by maximizing the following benefit/cost ratio R with respect to the futility and efficacy boundaries at IA while maintaining some operational characteristics. The trial can be stopped for either efficacy or futility at IA.

$$R = \frac{Bp(1-\beta)}{C_t + (C - C_t)[p(1-\beta_1) + (1-p)\alpha_1]} \quad (4)$$

where C_t is the cost spent at IA, C is the cost of the trial conducted until the final analysis, B is the benefit of the treatment, p is the POS (the probability of the study drug with treatment effect size at the alternative hypothesis value), assumed to be known and can be estimated from historical data, α_1 is efficacy boundary at IA in p value scale, β_1 is the type II error rate spent at IA, and β is the overall type II error rate. B is assumed to be fixed and hence does not have any impact on the design selection.

Denote Z_t as the standard normal (under null hypothesis) test statistic at the IA and Z_1 as the test statistic in the final analysis. Chen et al.^[5] proposed to find the optimal design by maximizing R while maintaining the overall power and holding efficacy boundary in p value (not POS) scale to be α_1 and α_2 , and probability of significance to be

β_1 and β_2 in IA and final analysis, respectively, which is equivalent to the following two constraints:

$$\begin{cases} P(Z_t > Z_{\beta_1} \text{ or } Z_1 > Z_{\beta_2}) = 1 - \beta \\ Z_{1-\alpha_1} + Z_{1-\beta_1} = \sqrt{t} (Z_{1-\alpha_2} + Z_{1-\beta_2}) \end{cases} \quad (5)$$

The second equation comes from the fact that the event size of a trial without IAs is inversely proportional to $(Z_{1-\alpha} + Z_{1-\beta})^2$, where α and β are the type I and II error rate, respectively.

POS Based Approaches

Efficacy boundaries based on operational characteristics do not have intuitive interpretation and hence are difficult to be communicated with non-statistician colleagues. An alternative way to design IA resorts to a statistic called POS.

This approach is based on prediction of success in the final analysis using data collected at IA. This kind of predictive approach has been used for designing sequential trials for a while. One approach is based on the so-called stochastic curtailment (Lan, Simon, and Halperin^[9] Halperin et al.,^[10] Lan and Wittes,^[11] Davis and Hardy,^[12] Dignam et al.^[13]) using the statistic called conditional power. Conditional power has been criticized for assuming that the treatment effect size equals a specific value which is not known to be true. To address this issue, conditional power is averaged with respect to the posterior distribution of the parameter to get a statistic called predictive power (Spiegelhalter and Freeman,^[14] Dignam et al.^[13]). In common practice, PP alone, is used to support decision making. As explained later, PP is a random variable. Using PP alone, without quantifying its uncertainty to support decision making can be misleading. As shown below, CP and PP are special cases of POS.

POS based optimal futility interim design was first proposed by Tang^[15] using the concept of predictive probability of success (PPOS) in a clinical trial with a time-to-event end point. This paradigm can be extended to trials with other end points.

Probability of Success

POS is closely related to CP and PP. CP is the probability of observing statistical significance in the final analysis, given the observed data and assuming the treatment effect size equals to a specific value.

$$CP = P(\hat{\theta} < \delta | \theta, F_{\text{obs}}) \quad (6)$$

where $\hat{\theta}$ is the observed treatment effect (here, smaller $\hat{\theta}$ means bigger treatment effect), δ is the statistical significance cutoff value for $\hat{\theta}$, θ is the treatment effect parameter value, and F_{obs} is the observed data.

There are two different interpretation of CP. In Frequentist setting, CP is a fixed value given that the treatment effect parameter equals to a specific value. CP in Frequentist setting is often criticized for assuming that the treatment effect parameter equals a specific value which is not known to be true. To address this issue, CP can be interpreted in the Bayesian setting. In the Bayesian setting, the treatment effect parameter is a random variable, so is CP. PP is the expected value of CP with respect to the posterior distribution of the parameter.

$$PP = \int CP f(\theta | F_{\text{obs}}) d\theta \quad (7)$$

where $f(\theta | F_{\text{obs}})$ is the posterior density of θ . It is a kind of center of CP. Summarizing CP using PP alone is not sufficient and may be misleading. Besides the center of the variable, the amount of uncertainty associated with the random variable is needed to understood completely.

Both CP and PP use statistical significance to define success. While statistical significance is required, clinical meaningfulness of the observed treatment effect may be more relevant and critical. For example, HA often require the observed treatment effect size to be bigger than statistical significance to support registration decision. To address this issue, the power concept can be extended to the concept of POS. CP can be extended to conditional probability of success (CPOS) and PP to PPOS. CPOS is the probability of observing success given the parameter equals to a specific value.

$$CPOS = P(\hat{\theta} < \delta | \theta, F_{\text{obs}}) \quad (8)$$

where $\hat{\theta}$ is the observed the treatment effect (here, smaller $\hat{\theta}$ means bigger treatment effect), δ is the success cutoff value for $\hat{\theta}$, θ is the treatment effect parameter value, and F_{obs} is the observed data.

PPOS is the expected value of CPOS with respect to the posterior distribution of the parameter.

$$PPOS = \int CPOS f(\theta | F_{\text{obs}}) d\theta \quad (9)$$

where $f(\theta | F_{\text{obs}})$ is the posterior density of θ .

The success criteria of CPOS and PPOS are not restricted to statistical significance. It can be something else such as clinically meaningful result. Note that if success is defined as statistical significance, then PPOS reduces to PP and CPOS reduces to CP. PP is a special case of PPOS, and CP is a special case of CPOS. CP is a random variable in Bayesian setting, and so is CPOS. Summarizing CPOS using its center (such as PPOS) alone is not enough. An important part is the variability associated with CPOS.

PPOS is a conditional probability conditioned on randomly observed data. When the trial is repeated, the observed data will be different. Hence the PPOS will be

different if the trial is repeated. Therefore, PPOS is also a random variable. The interpretation of PPOS is not complete without quantifying its uncertainty.

POS concepts can be used to set up decision rules with intuitive interpretation. Both PPOS and CPOS can be used for this purpose.

Decision Rule Based on PPOS

To declare futility, PPOS needs to be small and the credible interval of PPOS needs to be relatively tight.

$$\begin{cases} \text{PPOS} < c_1 \\ \text{uPPOS} < c_2 \end{cases} \quad (10)$$

where $c_1 < c_2$ and uPPOS is a high percentile of PPOS. Both c_1 and $c_2 - c_1$ are relatively small values. The first criterion mandates that the PPOS is small. The second criterion mandates the credible interval of PPOS is relatively tight. Therefore, its calculation is supported by sufficient information. Hence PPOS is not small just to chance.

Many designs satisfy the decision criteria. The optimal design, consistent with the decision rules, satisfies the following criteria. Finding the optimal design is equivalent to solving the two equations with respect to the design parameters.

$$\begin{cases} \text{PPOS} = c_1 \\ \text{uPPOS} = c_2 \end{cases} \quad (11)$$

The first equation ensures PPOS is small and so not too many trials will be prevented from entering the next stage to guard against false negative. The first equation also makes sure that the PPOS is not too small so that not too many trials will enter next stage to guard against false positive. The second equation makes sure that the PPOS calculation is supported by enough information hence, the PPOS is not small just due to chance. The second equation also ensures that the PPOS calculation does not demand too much information, and the IA can be done at relatively early to achieve the goal of saving resources.

PPOS can be computed from the predictive distribution. In Bayesian setting, the predicted distribution can be derived using the following framework if the data at IA (denoted as x) are independent of the data after IA (denoted as y), given the treatment effect parameter θ .

$$\begin{aligned} p(y|x) &= \int p(y, \theta|x) d\theta = \int p(y|\theta, x) p(\theta|x) d\theta \\ &= \int p(y|\theta) p(\theta|x) d\theta \end{aligned} \quad (12)$$

where $p(y|\theta)$ is the likelihood of future data and $p(\theta|x)$ is the posterior distribution given interim data.

Decision Rule Based on CPOS

Tang^[16] proposed to use another POS criteria to support decision making.

To declare futility, the median CPOS (mCPOS) needs to be small and the credible interval of CPOS to be relatively tight.

$$\begin{cases} \text{mCPOS} < c_1 \\ \text{uCPOS} < c_2 \end{cases} \quad (13)$$

where $c_1 < c_2$ and uCPOS is a high percentile of CPOS. Both c_1 and $c_2 - c_1$ are relatively small values. The first criteria mandate that the mCPOS is small. The second criteria mandate that the credible interval of CPOS is relatively tight. Therefore mCPOS calculation is supported by sufficient information. Hence mCPOS is not small by chance.

Many designs satisfy the decision criteria. The optimal design, consistent with the decision rules, satisfies the following criteria. Finding the optimal design is equivalent to solving the two equations with respect to the design parameters.

$$\begin{cases} \text{mCPOS} = c_1 \\ \text{uCPOS} = c_2 \end{cases} \quad (14)$$

The POS design paradigm will be demonstrated in three different trials in the following sub-sections.

Randomized Trial with Time-to-Event End Point

Consider a two-arm randomized trial with a time-to-event end point comparing the experimental treatment and the control treatment. The parameter of interest is log hazard ratio (HR) (experimental over control), denoted as θ (smaller θ means bigger treatment effect). Assume proportional hazard, i.e., θ is a constant at any time point of the survival course. The number of events at the IA is d . The total number of events in the final analysis is $d_{\max}$. Let $t = d/d_{\max}$ be the time of IA in the information scale. The time in the final analysis is 1. The test statistics is the Wald statistic based on partial likelihood from Cox regression model:

$$Z(t) = \frac{\hat{\theta}(t)}{\hat{\sigma}(\hat{\theta}(t))} \quad (15)$$

where $\hat{\theta}(t)$ is the maximum partial likelihood estimator of θ based on data at IA, $\hat{\sigma}(\hat{\theta}(t))$ is the standard error of the estimator.

$\hat{\theta}(t)$ has asymptotic normal distribution with mean equal to the true parameter and variance inversely proportional to the number of events at the IA. $\hat{\theta}(t) | \theta \sim N(\theta, \sigma_1^2 = 1/(r(1-r)d))$

where $r = \frac{\text{number of patients in control arm}}{\text{total number of patients in both arms}}$ is the randomization ratio.

POS calculation involves two components—the posterior distribution of θ and CPOS value at a specific θ value.

Posterior distribution of θ . Deriving posterior distribution of θ starts with some prior belief of the parameter. For simplicity, the prior can be assumed to be conjugate (with respect to the likelihood) normal with some mean and some variance. That is, $\theta \sim N(\theta_0, \sigma_0^2 = 1/(r(1-r)d_0))$. The variance is written in the same format as the parameter estimate variance, and hence the amount of information in the prior distribution can be quantified in terms of number of events d_0 .

The prior distribution can be updated using the interim data to obtain the posterior distribution of the parameter. The posterior distribution is also normal.

$$\theta | \hat{\theta}(t) \sim N(\phi \hat{\theta}(t) + (1-\phi)\theta_0, \sigma_0^2(1-\phi))$$

where,

$$\phi = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2} \quad (16)$$

The posterior mean is a weighted average of the interim estimate and the prior mean. The weights depend on the relative amount of information in the interim data and the prior distribution. The more the information from the interim data, bigger the contribution from the interim estimate. The posterior variance is a portion of the prior variance. The reduction of the variance depends on the relative amount of information in the interim data and the prior distribution. The more the information from the interim data, the bigger is the reduction. When there is no information on the prior distribution (prior number of events equals to 0), the prior information disappears from the posterior distribution. The posterior mean becomes the interim parameter estimate. The posterior variance becomes the interim parameter estimate variance.

When evaluating the final analysis at IA, the data upto the IA is fixed. The only uncertainty comes from the data between the IA and the final analysis. Using the following statistic Z^* , $Z(1)$ can be partitioned into two independent components. One component $Z(t)$ depends on the data upto IA only. The other component $\frac{\sqrt{d_{\max}}Z(1) - \sqrt{d}Z(t)}{\sqrt{d_{\max} - d}}$ is essentially the independent increment from IA to final analysis. According to Tang,^[15] Z^* and $Z(1)$ have the same distribution.

$$Z^* = \omega_1 Z(t) + \omega_2 \frac{\sqrt{d_{\max}}Z(1) - \sqrt{d}Z(t)}{\sqrt{d_{\max} - d}} \quad (17)$$

where, $\omega_1 = \sqrt{t}$, $\omega_2 = \sqrt{1-t}$.

Define success in the final analysis as $Z(1) < \delta$. This is equivalent to $Z^* < \delta$. Hence CPOS at a specific parameter value can be expressed as:

$$\text{CPOS} = P\left(\omega_1 Z(t) + \omega_2 \frac{\sqrt{d_{\max}}Z(1) - \sqrt{d}Z(t)}{\sqrt{d_{\max} - d}} < \delta \mid \theta\right) \quad (18)$$

After some rearrangements, CPOS can be expressed as:

$$\text{CPOS} = P\left(\frac{\sqrt{d_{\max}}Z(1) - \sqrt{d}Z(t)}{\sqrt{d_{\max} - d}} < \frac{\delta - \omega_1 Z(t)}{\omega_2} \mid \theta\right) \quad (19)$$

According to Tang,^[15]

$$\frac{\sqrt{d_{\max}}Z(1) - \sqrt{d}Z(t)}{\sqrt{d_{\max} - d}} \sim N\left(\theta \sqrt{r(1-r)(d_{\max} - d)}, 1\right) \quad (20)$$

hence,

$$\text{CPOS} = \Phi\left(\frac{\delta - \omega_1 Z(t)}{\omega_2} - \theta \sqrt{r(1-r)(d_{\max} - d)}\right) = f(\theta) \quad (21)$$

Bigger treatment effect (smaller θ) means bigger CPOS, hence $f(\theta)$ is a monotone decreasing function of θ . The percentiles of CPOS and θ have one-to-one relationship. $(1-\alpha)*100$ percentile of CPOS is the $f(\cdot)$ transformation of $\alpha*100$ percentile of θ . Denote $(1-\alpha)*100$ percentile of CPOS as uCPOS, then,

$$\text{uCPOS} = f\left(\text{qnorm}\left(\alpha, \phi \hat{\theta}(t) + (1-\phi)\theta_0, \sigma_0^2(1-\phi)\right)\right) \quad (22)$$

where $\text{qnorm}(p, \mu, \sigma^2)$ is the $p*100$ percentile of normal distribution with mean μ and variance σ^2 . The uCPOS $(1-\alpha)$ is the width of the $(1-\alpha)*100$ percent one-sided credible interval of CPOS.

PPOS PPOS can be analytically calculated using the predictive distribution. According to Tang, the predictive distribution is

$$\hat{\theta}(1-t) \mid \hat{\theta}(t) \sim N\left(\phi \hat{\theta}(t) + (1-\phi)\theta_0, \sigma_0^2(1-\phi) + \sigma_2^2\right) \quad (23)$$

where $\sigma_2^2 = 1/(r(1-r)(d_{\max} - d))$. The predictive distribution has the same mean as the posterior distribution and added variance due to added uncertainty of estimating the parameter using data.

Define success in the final analysis as $Z(1) < \delta$. This is equivalent to $Z^* < \delta$. Hence PPOS can be expressed as:

$$\text{PPOS} = P\left(\omega_1 Z(t) + \omega_2 \frac{\sqrt{d_{\max}}Z(1) - \sqrt{d}Z(t)}{\sqrt{d_{\max} - d}} < \delta \mid \hat{\theta}(t)\right) \quad (24)$$

Based on the predictive distribution,

$$\text{PPOS} = \Phi \left(\frac{\frac{\delta - \omega_1 z(t)}{\omega_2}}{\frac{\sqrt{r(1-r)(d_{\max} - d)} - [\varphi \hat{\theta}(t) + (1 - \varphi)\theta_0]}{\sqrt{\sigma_0^2(1 - \varphi) + \sigma_2^2}}} \right) \quad (25)$$

where $\sigma_2^2 = 1/[r(1-r)(d_{\max} - d)]$ and $\Phi(\cdot)$ denotes the cumulative density function of standard normal distribution.

$(1 - \alpha)*100$ percentile of PPOS is obtained by replacing the posterior mean $\varphi \hat{\theta}(t) + (1 - \varphi)\theta_0$ in PPOS formula by the $\alpha*100$ percentile of the posterior distribution of $\theta | \hat{\theta}(t)$. The prior information is also considered as observed data.

An example Consider a big randomized oncology phase III trial comparing the experimental treatment versus the standard care. The primary end point is progression free survival (PFS). The median PFS of the standard care is estimated to be nine months. The experimental treatment is expected to extend the median PFS for three months (corresponding to a clinically meaningful HR of 0). Patients are randomly assigned to the two arms in a 1:1 ratio ($r = 0.5$). The success is defined as observing clinically meaningful result $\text{HR} < 0.75$ after observing 500 events in the final analysis. For simplicity, a non-informative prior with zero prior events is used. When the following two conditions are satisfied, futility will be declared.

1. The predictive probability of observing a clinically meaningful HR in the final analysis is less than 20% ($\text{PPOS} < 0.2$).
2. The 80th percentile of PPOS is less than 40% ($\text{uPPOS} < 0.4$).

Finding the optimal futility interim design is equivalent to finding the solution of the following equations with respect to the design parameters—timing of the analysis t and cutoff HR at IA (HR.IA).

$$\begin{cases} \text{PPOS}(t, \text{HR.IA}) = 0.2 \\ \text{80th percentile of PPOS}(t, \text{HR.IA}) = 0.4 \end{cases} \quad (26)$$

The optimal futility interim design is to implement the IA at information fraction of $0.511 (t = 0.511, d_1 = 256)$ with a cutoff HR of 0.807.

The optimal design can be visualized using the contour plot of PPOS and uPPOS versus information fraction t and cutoff HR at IA (HR.IA) (Fig. 1). First, focus on the PPOS lines. On the top of the graph, PPOS decreases as the information increases. At the bottom of the graph, PPOS increases as the information increases. The separation is at $\text{HR.IA} = 0.75$. This is because the success is defined as a HR of 0.75. When an unfavorable HR is observed, it is more likely to be an unsuccessful trial if the PPOS calculation

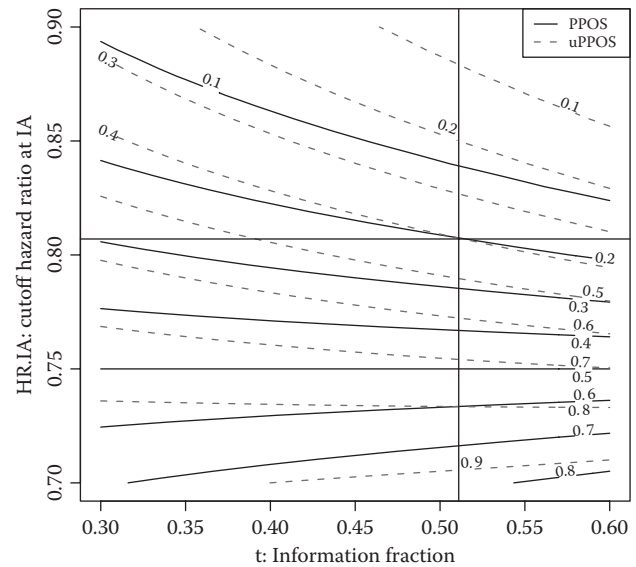


Fig. 1 Contour plot of PPOS and uPPOS vs. information fraction t and cutoff hazard ratio at IA (HR.IA). The black contour lines represent PPOS. The red contour lines represent the 80th percentiles of PPOS. The success is defined as a clinically meaningful HR of 0.75 or less. The optimal design with respect to the POS decision rule is highlighted with a pair of horizontal and vertical lines crossing it

is supported by more information. If a favorable HR is observed, it is more likely to be a successful trial if the PPOS calculation is supported by more information.

The design at the intersection of PPOS 0.2 and uPPOS 0.4 line is the optimal design with respect to the POS decision rule. This can be shown using elimination method. For the designs above the PPOS = 0.2 line, the PPOS cutoff is too small, hence too many trials will enter next stage and cause too much false positive. For designs below the PPOS = 0.2 line, the PPOS cutoff is too big, hence too many trials will be prevented from entering next stage and cause too much false negative. This shows the designs above and below the PPOS = 0.2 line are inferior to the intersection design. Now the designs in the left are at the PPOS 0.2 line. For designs to the left of the intersection, PPOS credible interval is too wide and hence PPOS calculation is supported by too little information, PPOS may be small by chance. For designs to the right of the intersection, PPOS credible interval is too tight and the calculation demands too much information and it is not efficient in achieving the goal of saving resource if the treatment is doomed to fail. This shows the designs on both sides of the PPOS = 0.2 line are inferior to the intersection design. Therefore the intersection is the optimal design corresponding to the PPOS decision criteria.

Randomized Trial with Normal End Point

Consider a two-arm randomized trial with a normal end point comparing the experimental treatment and the

control treatment. The total sample size is n . The parameter of interest is treatment effect difference θ (bigger θ means bigger treatment effect). The randomization ratio is $r = \frac{\text{number of patients in the control arm}}{\text{total number of patients}}$. Assume the two treatments share a common known variance σ^2 . The sample mean difference is denoted as $\hat{\theta}$ which is normally distributed with unknown mean treatment effect θ and variance $\frac{\sigma^2}{r(1-r)n}$. Let t be the time of the IA in the information scale. The final analysis is done at time 1 in the information scale. The sample size at the IA is $n_1 = nt$. The sample size between IA and final analysis is $n_2 = n(1-t)$.

Denote the estimate at IA as $\hat{\theta}(t)$ and the estimate using the data between IA and final analysis as $\hat{\theta}(1-t)$. Denote the estimate based on all data as $\hat{\theta}(1)$.

It can be shown that $\hat{\theta}(1)$ approximately equals to $t\hat{\theta}(t) + (1-t)\hat{\theta}(1-t)$, i.e. all data estimator is an information weighted average of the stagewise estimators.

The statistics $\hat{\theta}(t)$ and $\hat{\theta}(1-t)$ are normally distributed and conditionally independent given θ , with variance $\sigma_1^2 = \frac{\sigma^2}{r(1-r)n_1}$ and $\sigma_2^2 = \frac{\sigma^2}{r(1-r)n_2}$ respectively.

For simplicity, the prior distribution of θ can be assumed to be conjugate (with respect to the likelihood) normal $N\left(\theta_0, \sigma_0^2 = \frac{\sigma^2}{r(1-r)n_0}\right)$. The prior variance is written in the same format as the parameter estimate variance, hence the prior information can be quantified in terms of the number of observations. The interim data can be used to update the prior distribution. The resulted posterior distribution is:

$$\theta | \hat{\theta}(t) \sim N\left((1-\phi)\theta_0 + \phi\hat{\theta}(t), (1-\phi)\sigma_0^2\right)$$

where,

$$\phi = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2} \quad (27)$$

The posterior mean is a weighted average of the prior mean and interim estimate. The weights depend on the relative amount of information between the prior distribution and the interim data. The more the information in the interim data, the bigger the contribution from the interim estimate. The posterior variance is a portion of the prior variance. The reduction of variance depends on the relative amount of information between the prior information and the interim data. The more the information from the interim data, the bigger the reduction in posterior variance.

The predictive distribution is:

$$\hat{\theta}(1-t) | \hat{\theta}(t) \sim N\left((1-\phi)\theta_0 + t\hat{\theta}(t), (1-\phi)\sigma_0^2 + \sigma_2^2\right) \quad (28)$$

It has the same mean as the posterior distribution but bigger variance due to added uncertainty for estimating parameters using data.

Define success in the final analysis as $\hat{\theta}(1) > \delta$. This is equivalent to $\hat{\theta}(1-t) > \frac{n\delta - n_1\hat{\theta}(t)}{n_2}$.

Because

$$\hat{\theta}(1-t) \sim N(\theta, \sigma_2^2), \quad (29)$$

hence,

$$\text{CPOS} = \Phi\left(\frac{\frac{n\delta - n_1\hat{\theta}(t)}{n_2} - \theta}{\sigma_2}\right)$$

Percentiles of CPOS can be obtained by replacing the θ in the equation by the corresponding percentiles of θ from its posterior distribution.

Based on the predictive distribution

$$\text{PPOS} = 1 - \Phi\left(\frac{\frac{n\delta - n_1\hat{\theta}(t)}{n_2} - (1-\phi)\theta_0 - \phi\hat{\theta}(t)}{\sqrt{(1-\phi)\sigma_0^2 + \sigma_2^2}}\right) \quad (30)$$

$(1-\alpha)*100$ percentile of PPOS, uPPOS can be obtained by replacing the mean of the predictive distribution by the $(1-\alpha)*100$ percentile of posterior distribution.

Single Arm Study with Binary End Point

Consider a single arm study with a binary end point whose response rate is p . Assume the prior distribution of P to be conjugate beta (with respect to the binomial likelihood), i.e. $P \sim \text{beta}(a, b)$. The prior mean of P is $\frac{a}{a+b}$. The amount of information is measured by the sum of $a + b$. The numbers a and b can be considered as the number of responders and nonresponders respectively out of total of $a + b$ observations.

Suppose at IA, there are n observations. The number of responders at IA is x . The posterior distribution of P is then:

$$P | X = x \sim \text{beta}(a+x, b+n-x) \quad (31)$$

The predictive distribution for the number of responders Y out of m observations in the remaining of the trial is beta binomial with the following probability mass function:

$$P(Y = y | X = x) = \binom{m}{y} \frac{B(a+x+y, b+n-x+m-y)}{B(a+x, b+n-x)}$$

where B

(x, y) is a beta function; $\binom{m}{y}$ is the m choose y function.

Define success in terms of the observed response rate in the final analysis $\frac{x+Y}{m+n} > c$. This is equivalent to $Y > (m+n)c - x$.

$$\text{CPOS}(p) = 1 - \text{pbinom}(\text{floor}((m+n)c - x), m, p) \quad (32)$$

where pbinom is the cumulative binomial density function; floor is the biggest integer no more than x .

Percentiles of CPOS can be obtained by replacing p in CPOS formula by the corresponding percentiles of P from the posterior distribution.

Based on the predictive distribution, $PPOS = P(Y > (m+n)c - x \mid X = x)$.

Percentiles of PPOS can be obtained by replacing x in the predictive distribution by the corresponding percentiles of X . The percentiles of X can be obtained by multiplying n by the corresponding percentiles of P from its posterior distribution.

CONCLUSIONS AND DISCUSSIONS

In this entry, two different paradigms for optimal futility interim design, FOC versus POS based approaches, are summarized and demonstrated in three different trial scenarios.

FOC in sequential trials does not have intuitive interpretation and hence it is difficult to communicate with non-statistician colleagues. POS have intuitive interpretation and hence can facilitate team communication and interaction with HA.

FOC based approaches need to specify a specific treatment effect size in order to calculate power. POS approaches use data driven decision without specifying a specific treatment effect size which is not known to be true.

FOC based approaches maintain power at a specific treatment effect size. Power is probability of observing statistical significance. Statistical significance is often not enough to define success. Success criteria in POS are not restricted to statistical significance. POS approaches are more useful due to its flexibility. Statistically significant, but nonclinically significant results may not translate to benefit. It may not be worthwhile saving nonclinically significant power. Evaluating POS approaches using FOC criteria may be misleading.

In FOC based methods, the futility IA is designed together with the final analysis to maintain the type I and type II error rate. In contrast, in POS based approaches, the design of the futility IA is independent of the design of the final analysis. Hence its type I error rate control is conservative. This conservativeness is desirable since futility IA is generally nonbinding to allow the flexibility to continue the trial even if the futility boundary is crossed. In FOC based methods, type I error rate control assumes a binding futility IA. This may not be practical. For example, the primary end point is a short term end point. The futility criteria for short term end point may be met at IA. But the trial may continue if substantial long term benefit is observed.

FOC based approaches achieve optimality via maximizing/minimizing some utility function. The optimality is achieved in one specific dimension defined by the utility function but may suffer in the other dimensions. In practice, it is rare to find a single dimension which is important while other dimensions are trivial and ignorable. In reality, multiple factors should be taken to consideration in trial design. One dimensional optimal design can not accommodate the complexity of the reality. In contrast,

POS based approaches can select the decision rule cutoff values by incorporating information from multiple dimensions, hence the resulted optimal design achieves the best balance among multiple design dimensions.

POS approaches involve several subjective decisions including success criteria selection, POS cutoff values selection, etc. All these decisions should be extensively discussed and well justified to minimize the potential bias. Clinical trial design team with diversified (relevant) backgrounds and smooth communication can benefit more from POS based methods.

One common issue in POS practice is using PPOS alone to support decision making. The inference about treatment effect size, say θ , depends on both the point estimate $\hat{\theta}$ and the variance of $\hat{\theta}$. PPOS and uPPOS are 1:1 transformation of $\hat{\theta}$ and the variance of $\hat{\theta}$. Hence PPOS and uPPOS are equivalent statistic to $\hat{\theta}$ and the variance of $\hat{\theta}$. Using PPOS and uPPOS for decision making is equivalent to using $\hat{\theta}$ and the variance of $\hat{\theta}$. Reducing $\hat{\theta}$ and the variance of $\hat{\theta}$ to PPOS alone will cause information loss. PPOS is a statistic calculated from data. Using PPOS alone for decision making is using a statistic for statistical inference without knowing its distribution. This doesn't make sense. By the same token, mCPOS should be used together with its credible interval to support decision making as well.

REFERENCES

1. Guidance for clinical trial sponsors. Establishment and operation of clinical trial data monitoring committees. In FDA, 2006.
2. Weiss, L. On sequential tests which minimize the maximum expected sample size. *J. Amer. Stat. Asso.* **1962**, 57, 551–566.
3. Lai, T. Optimal stopping and sequential tests which minimize the maximum expected sample size. *Annal. Stat.* **1973**, 1, 659–673.
4. Simon, R. Optimal two-stage designs for phase II clinical trials. *Control. Clin. Trials* **1989**, 10, 1–10.
5. Chen, C.; Beckman, R.A.; Sun, L.Z. Optimal cost-effective Go-No Go decision in clinical development. In *Practical Considerations for Adaptive Trial Design and Implementation*; He, W., Pinhero, J., Kuznetsova, O.M., Eds.; Springer Science + Business Media: New York, 2014, 101.
6. Wason, J.; Mander, A. Minimising the maximum expected sample size in two-stage phase II clinical trials with continuous outcomes. *J. Biopharm. Stat.* **2012**, 22 (4), 836–852.
7. Mander, A.; Thompson, S. Two-stage designs optimal under the alternative hypothesis for phase II cancer clinical trials. *Contem. Clin. Trials* **2010**, 31, 572–578.
8. Wason, J.; Mander, A.; Thompson, S. Optimal multistage designs for randomized clinical trials with continuous outcomes. *Stat. Med.* **2011**, 31, 301–312.
9. Lan, K.K.G.; Simon, R.; Halperin, M. Stochastic curtailed tests in long term clinical trials. *Commun. Stat. Sequential Anal.* **1982**, 1, 207–219.

10. Halperin, M.; Lan, K.K.G.; Ware, J.H.; Johnson, N.J.; DeMets, D.L. An aid to data monitoring in long term clinical trials. *Control Clinical Trials* **1982**, *3*, 311–323.
11. Lan, K.K.G.; Witts, J. The B-value: A tool for monitoring data. *Biometrics* **1988**, *44*, 579–585.
12. Davis, B.R.; Hardy, R.J. Upper bounds for type I and type II error rates in conditional power calculations. *Commun. Stat. Theory Meth.* **1990**, *19*, 3571–3584.
13. Dignam, J.J.; Bryant, J.; Wieand, H.S. Early stopping of cancer clinical trials. In *Handbook of Statistics in Clinical Oncology*, 2nd Ed.; Ankerst, J., Ankerst, D.P., Eds.; Chapman & Hall/CRC: Boca Raton, FL, 2006; 227–246.
14. Spiegelhalter, D.J.; Freeman, L.S. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Stat. Med.* **1986**, *5*, 1–13.
15. Tang, Z. Optimal futility interim design: A predictive probability of success approach with time-to-event endpoint. *J. Biopharm. Stat.* **2015**, *25* (6), 1312–1319.
16. Tang, Z. Defensive efficacy interim design: Dynamic benefit/risk ratio view using probability of success. *J. Biopharm. Stat.* **2016**. DOI: 10.1080/10543406.2016.1198370.

Ordered Categorical Data: Test for

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Siu-Keung Tse

Department of Management Sciences, City University of Hong Kong, Hong Kong, China

Chunyan Yang

Hong Kong Baptist University, Hong Kong; and Yunnan University, Kunming, Yunnan, China

Abstract

The primary objective of this entry is to use real data sets to demonstrate the application of testing methods and provide numerical results from some of these tests.

INTRODUCTION

In clinical trials for comparing an active treatment and a standard therapy, the scale for the primary clinical endpoint is recommended to be a continuous measure in order to provide an accurate and reliable assessment. In practice, however, it is impossible, or can be extremely expensive, to measure responses continuously. Thus, in many applications, the intensity of some meaningful and well-defined event such as infection or cure of a certain disease is used as a surrogate endpoint for some unobserved latent continuous variables in clinical trials. Responses to treatments from patients are being classified into some ordered categories according to the intensity of the event. For example, based on some pre-specified evaluability criteria, the investigator may classify the responses into three categories of “worsening,” “no change,” “and improving.”

Thus, the comparison of the two treatments is conducted based on the proportions of responses for different categories. A simple approach to analyzing these ordered categorical data is to create dichotomies among the levels of the response variable by combining observations from adjacent categories. For example, the two categories “worsening” and “no change” are amalgamated into a single category as “not improving.” The resulting data would be considered as binary responses classified as “improving” and “not improving,” which can be analyzed using a standard logistic regression model. This approach of collapsing (or ignoring) some categories of responses into a single category would certainly result in a loss of information and may lead to a considerable loss of statistical power. In addition, it is a concern that a pattern of results where the active treatment increased both the proportion of failures (“worsening”) as well as the proportion of successes (“improvement”) would be statistically significant and difficult to interpret.

Testing of the one-sided alternative hypothesis that the active treatment is better than the standard therapy for ordered categorical data can be done by using an exact conditional test when the sample size is moderate. The idea is to use a multinomial model to describe the probabilities of the responses falling into the various categories. Then, the comparison of the effectiveness of the two treatments is conducted by a conditional test, which requires the evaluation of all sample points to determine the conditional p -value. More details can be found in Agresti.^[1] More recent discussion of this approach can be found in Cohen and Sackrowitz.^[2] Although this approach does not rely on any asymptotic approximation to derive the distribution of the corresponding test statistic, it is not very efficient in terms of computing time for large sample size. When the sample size is large, several methods utilizing chi-square statistic have been proposed in the literature.^[3–5] In particular, Silvapulle and Sen^[6] gave a comprehensive review and discussion of the related ideas.

Recently, Chow et al.^[7] proposed to model the response probabilities via a parametric form and the comparison of these probabilities is translated into the comparison of the associated model parameters. Based on their approach, the maximum likelihood estimates (MLE) of the parameters can be obtained. Furthermore, a likelihood ratio test can be used to compare the effectiveness of the two treatments. Based on the asymptotic normality of the MLE of the parameters, the required sample size to achieve a given level of power to assess the difference between two treatments can be derived.

The major objective of this entry is to summarize the ideas discussed in Chow et al.^[7] and demonstrate the application of their methods through the analysis of three real data sets. The next section briefly outlines, the proposed model and the corresponding assumptions. Under the proposed model, results of a likelihood ratio test to assess treatment effect for ordered categorical data

and the corresponding required sample size to achieve a given level of power are derived. The section “Numerical Results” summarizes the results of a numerical study conducted for the determination of the required sample sizes under different combinations of response probabilities for the ordered categories. Examples concerning clinical trials for comparing a test treatment with an active control are presented in the section “Examples” to illustrate the use and interpretation of the proposed test. Some concluding remarks are given in the last section.

MODEL AND ASSUMPTIONS

Suppose that there are two types of treatment administered to the subjects and the corresponding response of the subject after receiving the treatment can be classified into k different categories. Denote the probability that the response falls into the j th category after receiving the i th treatment by π_{ij} . Consider the following model

$$\log \frac{\pi_{ij}}{\pi_{ik}} = \alpha_j + \beta_i, \quad i = 0, 1; j = 1, 2, \dots, k-1, \quad (1)$$

where α_j denotes the effect of the j th category relative to the k th category and β_i is the effect of the i th treatment. Note that $\pi_{ik} = 1 - \sum_{j=1}^{k-1} \pi_{ij}$. Model (1) can be expressed as $\frac{\pi_{ij}}{\pi_{ik}} = e^{\alpha_j + \beta_i}$. This gives

$$\pi_{ik} = \frac{1}{1 + e^{\alpha_1 + \beta_i} + e^{\alpha_2 + \beta_i} + \dots + e^{\alpha_{k-1} + \beta_i}} \quad (2)$$

and

$$\pi_{ij} = \frac{e^{\alpha_j + \beta_i}}{1 + e^{\alpha_1 + \beta_i} + e^{\alpha_2 + \beta_i} + \dots + e^{\alpha_{k-1} + \beta_i}}. \quad (3)$$

The logit model defined in Eq. (1) for the modeling of response probabilities of ordered categorical responses has been discussed extensively in the literature. Related details and further references can be found in Agresti.^[1] Under model (1), the difference between the two treatments is described by the difference between the parameters β_0 and β_1 . Thus, testing of no treatment difference can be assessed by the testing the hypotheses $H_0: \beta_0 = \beta_1$ vs $H_a: \beta_0 \neq \beta_1$. Let $\beta = \beta_1 - \beta_0$. Then, testing the above hypotheses is equivalent to testing the following hypotheses

$$H_0: \beta = 0 \text{ vs } H_a: \beta \neq 0. \quad (4)$$

The alternative hypothesis can be divided into the following two one-sided hypotheses:

$$H_{a1}: \beta < 0, \text{ and } H_{a2}: \beta > 0.$$

In general, if the responses are classified into k ordered categories, it can be shown that H_{a1} and H_{a2} are equivalent to the hypotheses $H'_{a1}: \pi_{0j} > \pi_{1j}, j = 1, \dots, k-1$ and $\pi_{0k} < \pi_{1k}$, and $H'_{a2}: \pi_{0j} < \pi_{1j}, j = 1, \dots, k-1$ and $\pi_{0k} > \pi_{1k}$ respectively. However, in many applications, the objective is to assess whether a test treatment is more effective than the existing treatment. Thus, the alternative hypothesis is often one-sided rather than two-sided. In other words, the hypotheses under testing are $H_0: \beta = 0$ vs $H_a: \beta < 0$.

Likelihood Ratio Test

Suppose that the two treatments are administered to N_0 and N_1 randomly selected subjects respectively. Let $n_{01}, n_{02}, \dots, n_{0k}$ and $n_{11}, n_{12}, \dots, n_{1k}$ be the corresponding observed frequencies falling into the k categories. Note that $\sum_{j=1}^k n_{ij} = N_i, i = 0, 1$. Denote the parameter vector $\theta = (\beta, \alpha_1, \alpha_2, \dots, \alpha_{k-1})$. The likelihood function is given by $L(\theta) = \prod_{i=0}^1 \prod_{j=1}^k \pi_{ij}^{n_{ij}}$ and the corresponding log-likelihood function is

$$\log L(\theta) = \sum_{i=0}^1 \sum_{j=1}^k n_{ij} \log \pi_{ij}. \quad (5)$$

Then, the MLE $\hat{\theta} = (\hat{\beta}, \hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_{k-1})$ of θ can be found by maximizing the log-likelihood function given in Eq. (5). The likelihood equations are obtained by setting the partial derivative of the log-likelihood function defined in Eq. (5) with respect to the parameters to 0. In particular,

$$\frac{\partial \log L(\theta)}{\partial \beta} = N_1 \pi_{1k} - n_{1k} = 0 \quad (6)$$

$$\frac{\partial \log L(\theta)}{\partial \alpha_j} = \sum_{i=0}^1 (n_{ij} - N_i \pi_{ij}) = 0; j = 1, 2, \dots, k-1.$$

Then, the MLE $\hat{\theta} = (\hat{\beta}, \hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_{k-1})$ of θ can be found by solving the k likelihood equations given in (6).

Alternatively, under $H_0: \beta = 0$, the likelihood function is $L_0(\alpha_1, \alpha_2, \dots, \alpha_{k-1}) = \prod_{j=1}^k \pi_j^{(n_{0j} + n_{1j})}$, or equivalently, the log-likelihood function is given by

$$\log L_0(\alpha_1, \alpha_2, \dots, \alpha_{k-1}) = \sum_{j=1}^k (n_{0j} + n_{1j}) \log \pi_j, \quad (7)$$

where $\pi_j = e^{\alpha_j} / (1 + e^{\alpha_1} + e^{\alpha_2} + \dots + e^{\alpha_{k-1}})$ for $j = 1, 2, \dots, k-1$ and $\pi_k = 1 - \sum_{j=1}^{k-1} \pi_j$. Taking partial derivatives of the log-likelihood function given in Eq. (7) with respect to the unknown parameters $\alpha_j, j = 1, \dots, k-1$, the corresponding likelihood equations are given as

$$\frac{\partial \log L_0(\alpha_1, \alpha_2, \dots, \alpha_{k-1})}{\partial \alpha_j} = \sum_{i=0}^1 (n_{ij} - N_i \pi_j) = 0; j = 1, 2, \dots, k-1. \quad (8)$$

Denote the MLE of $\theta = (\beta, \alpha_1, \alpha_2, \dots, \alpha_{k-1})$ under H_0 by $\hat{\theta}_0 = (0, \hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_{k-1})$. Then $\hat{\alpha}_j$ can be found by solving the likelihood equations given in Eq. (8).

Based on these results, a likelihood ratio test for the hypotheses Eq. (4) can be developed by taking the difference of the log-likelihood functions under the two hypotheses. In particular, $-2[\log L_0(\hat{\theta}_0) - \log L(\hat{\theta})]$ is asymptotically chi-square distributed with one degree of freedom when N_0 and N_1 are sufficiently large.^[8]

Determination of Sample Size

In practice, it would be of interest to determine the number of subjects required in a clinical study so that the test for the hypotheses given in Eq. (4) is with a given level of power $(1-\xi)$. To achieve this, the asymptotic distribution of the MLE of the parameter β is considered. Note

$$E\left(\frac{\partial^2 \log L}{\partial \alpha_j^2}\right) = -\sum_{i=0}^1 N_i \pi_{ij} (1 - \pi_{ij}); j = 1, 2, \dots, k-1,$$

$$E\left(\frac{\partial^2 \log L}{\partial \beta \partial \alpha_j}\right) = -N_1 \pi_{1j} \pi_{1k}; j = 1, 2, \dots, k-1,$$

$$E\left(\frac{\partial^2 \log L}{\partial \alpha_l \partial \alpha_j}\right) = \sum_{i=0}^1 N_i \pi_{il} \pi_{ij}; l, j = 1, 2, \dots, k-1 \text{ and } l \neq j.$$

Thus, the Fisher information matrix is $I(\theta) = -E(\partial^2 \log L / \partial \theta \partial \theta^T)$. Partition $I(\theta)$ into $\begin{pmatrix} a & I_{12} \\ I_{12}^T & I_{22} \end{pmatrix}$, where $a = N_1 \pi_{1k} (1 - \pi_{1k})$, $I_{12} = (N_1 \pi_{11} \pi_{1k} \ N_1 \pi_{12} \pi_{1k} \dots N_1 \pi_{1,k-1} \pi_{1k})_{1 \times (k-1)}$ and

$$I_{22} = \begin{pmatrix} \sum_{i=0}^1 N_i \pi_{i1} (1 - \pi_{i1}) & -\sum_{i=0}^1 N_i \pi_{i1} \pi_{i2} & \dots & -\sum_{i=0}^1 N_i \pi_{i1} \pi_{i,k-1} \\ -\sum_{i=0}^1 N_i \pi_{i1} \pi_{i2} & \sum_{i=0}^1 N_i \pi_{i2} (1 - \pi_{i2}) & \dots & -\sum_{i=0}^1 N_i \pi_{i2} \pi_{i,k-1} \\ \vdots & \vdots & \ddots & \vdots \\ -\sum_{i=0}^1 N_i \pi_{i,k-1} \pi_{i1} & -\sum_{i=0}^1 N_i \pi_{i,k-1} \pi_{i2} & \dots & \sum_{i=0}^1 N_i \pi_{i,k-1} (1 - \pi_{i,k-1}) \end{pmatrix}_{(k-1) \times (k-1)}.$$

that the Fisher information matrix can be found by taking the expected values of the negative of the second order partial derivatives of the log-likelihood function $\log L(\theta)$ with respect to the model parameters. In particular, from Eqs. (2) and (3), it is easy to show that

$$\frac{\partial \pi_{0j}}{\partial \beta} = 0, \quad \frac{\partial \pi_{1j}}{\partial \beta} = \pi_{1j} \pi_{1k}, j = 1, 2, \dots, k-1; \text{ and}$$

$$\frac{\partial \pi_{1k}}{\partial \beta} = -\pi_{1k} (1 - \pi_{1k});$$

$$\frac{\partial \pi_{ij}}{\partial \alpha_j} = \pi_{ij} (1 - \pi_{ij}), i = 0, 1; j = 1, 2, \dots, k-1;$$

$$\frac{\partial \pi_{il}}{\partial \alpha_j} = -\pi_{il} \pi_{ij}, i = 1, 0; j = 1, 2, \dots, k-1;$$

$$l = 1, 2, \dots, k \text{ and } l \neq j.$$

Observe that $E(n_{ij}) = N_i \pi_{ij}, i = 0, 1; j = 1, 2, \dots, k$. Using these results, the expected values of the second order partial derivatives of the log-likelihood function with respect to the parameters are:

$$E\left(\frac{\partial^2 \log L}{\partial \beta^2}\right) = -N_1 \pi_{1k} (1 - \pi_{1k}),$$

The asymptotic variance of the MLE $\hat{\beta}$ of β is given by

$$V = \frac{1}{a} \left[1 + \frac{1}{a} I_{12} \left(I_{22} - \frac{1}{a} I_{12}^T I_{12} \right)^{-1} I_{12}^T \right]. \quad (9)$$

It can be shown that $\hat{\beta}$ is asymptotically normally distributed with mean β and variance V when N_0 and N_1 are sufficiently large.^[9] Let V_0 and V_1 be the value of V given in Eq. (9) evaluated at $\hat{\theta}_0 = (0, \hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_{k-1})$ and $\hat{\theta} = (\hat{\beta}, \hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_{k-1})$ respectively. In particular, if the numbers of subjects assigned to the two treatments are the same, i.e., $N_0 = N_1 = N$. Denote $D_0 = NV_0$ and $D_1 = NV_1$. Based on the asymptotic normality of the MLE $\hat{\beta}$, the required sample size N to enable the test to have power $(1-\xi)$ to detect a difference $\Delta\beta$ in is given by the following expression,

$$N \geq \frac{(\sqrt{D_0} Z_\alpha - \sqrt{D_1} Z_{1-\xi})^2}{(\Delta\beta)^2}. \quad (10)$$

NUMERICAL RESULTS

In order to provide some insight on the pattern of required sample size to achieve a given level of power, seven sets of various combinations of π_{ij} are considered which

represent various cases occurring in the comparison of two medical treatments. More specifically, the cases $(\pi_{01}, \pi_{02}, \pi_{03}) = (0.1, 0.2, 0.7)$ and $(0.1, 0.4, 0.5)$ indicate that the effect of the existing treatment is significant (i.e. π_{03} is much larger than π_{01}) with the probability of the category “no change” is set to be relatively small and large respectively. Similarly $(\pi_{01}, \pi_{02}, \pi_{03}) = (0.4, 0.1, 0.5)$ and $(0.3, 0.4, 0.3)$ indicate that the effect of the existing treatment is not significant while the probability of the category “no change” is set to be relatively small and large respectively. The case $(\pi_{01}, \pi_{02}, \pi_{03}) = (0.33, 0.34, 0.33)$ indicates that the effect of the existing treatment is not significant. The cases $(\pi_{01}, \pi_{02}, \pi_{03}) = (0.6, 0.3, 0.1)$ and $(0.1, 0.3, 0.6)$ indicate that strong negative effect and strong positive effect of the existing treatment respectively.

Furthermore, for a given set $(\pi_{01}, \pi_{02}, \pi_{03})$, the effect of the test treatment is determined by considering $\pi_{01} > \pi_{11}$ and $\pi_{03} < \pi_{13}$, which indicate that the effect of the test treatment is better. Also, three different power levels are considered, namely, 0.65, 0.8, and 0.9, which represents the low, the medium and the high levels of power respectively. Based on Eqs. 9 and 10, the sample size N to achieve the desired level of power is computed. The results are presented in Table 1.

In general, for each given case, the sample size increases as the level of power increases. Moreover, the sample sizes for the first two cases are much smaller than those for the last two cases. The first two cases represent that the effect of the test treatment is significantly better than the existing treatment. That is, $\Delta\beta$ is large. Thus, even a sample of size 80

and 60 can achieve power 0.9 respectively. However, for the next two cases, a sample of size 818 and 1057 respectively is needed to achieve the same power since the difference between two treatments is very small. Thus, a much larger sample size is needed to detect the difference. Table 1 also lists the required sample sizes for different combinations of response probability in the four categories case. It can be seen that similar patterns emerge in the four categories case. Note that the sample sizes listed in Table 1 are derived based on asymptotic approximation. Chow et al^[7] conducted a simulation study to assess the performance of these results and the simulated powers are close to, and larger than in most cases, the corresponding nominal power.

EXAMPLES

In this section, three data sets are selected and analyzed for the purpose of demonstrating the testing results given in the previous sections.

Example 1

A study was conducted to compare two treatments for ulcer and a total of 64 subjects were enrolled in the study.^[10] Qualified subjects were randomized in a 1:1 ratio into the test treatment group and a control group. Subjects were classified into three ordered categories, namely “worsening,” “no change,” “improvement,” according to the change in the size of ulcer crater. A subject was considered

Table 1 Sample Size Required to Achieve the Nominal Level of Power

Probability of categories		Nominal power		
		0.65	0.80	0.90
i. Three categories				
$(\pi_{01}, \pi_{02}, \pi_{03}) = (0.10, 0.20, 0.70)$	$(\pi_{11}, \pi_{12}, \pi_{13}) = (0.04, 0.08, 0.88)$	35	55	80
$(\pi_{01}, \pi_{02}, \pi_{03}) = (0.10, 0.40, 0.50)$	$(\pi_{11}, \pi_{12}, \pi_{13}) = (0.05, 0.20, 0.75)$	28	43	60
$(\pi_{01}, \pi_{02}, \pi_{03}) = (0.40, 0.10, 0.50)$	$(\pi_{11}, \pi_{12}, \pi_{13}) = (0.34, 0.09, 0.57)$	392	590	818
$(\pi_{01}, \pi_{02}, \pi_{03}) = (0.30, 0.45, 0.25)$	$(\pi_{11}, \pi_{12}, \pi_{13}) = (0.28, 0.42, 0.31)$	516	768	1057
$(\pi_{01}, \pi_{02}, \pi_{03}) = (0.33, 0.34, 0.33)$	$(\pi_{11}, \pi_{12}, \pi_{13}) = (0.20, 0.20, 0.60)$	30	45	62
$(\pi_{01}, \pi_{02}, \pi_{03}) = (0.60, 0.30, 0.10)$	$(\pi_{11}, \pi_{12}, \pi_{13}) = (0.50, 0.25, 0.25)$	71	102	138
$(\pi_{01}, \pi_{02}, \pi_{03}) = (0.10, 0.30, 0.60)$	$(\pi_{11}, \pi_{12}, \pi_{13}) = (0.06, 0.19, 0.75)$	73	113	158
ii. Four categories				
$(\pi_{01}, \pi_{02}, \pi_{03}, \pi_{04}) = (0.12, 0.36, 0.40, 0.12)$	$(\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14}) = (0.10, 0.30, 0.33, 0.28)$	67	97	131
$(\pi_{01}, \pi_{02}, \pi_{03}, \pi_{04}) = (0.12, 0.36, 0.40, 0.12)$	$(\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14}) = (0.10, 0.31, 0.34, 0.24)$	103	149	202
$(\pi_{01}, \pi_{02}, \pi_{03}, \pi_{04}) = (0.12, 0.36, 0.40, 0.12)$	$(\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14}) = (0.11, 0.34, 0.37, 0.18)$	374	550	751
$(\pi_{01}, \pi_{02}, \pi_{03}, \pi_{04}) = (0.23, 0.26, 0.28, 0.23)$	$(\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14}) = (0.20, 0.23, 0.25, 0.32)$	224	332	456
$(\pi_{01}, \pi_{02}, \pi_{03}, \pi_{04}) = (0.23, 0.26, 0.28, 0.23)$	$(\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14}) = (0.13, 0.15, 0.16, 0.56)$	21	31	42
$(\pi_{01}, \pi_{02}, \pi_{03}, \pi_{04}) = (0.18, 0.06, 0.59, 0.18)$	$(\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14}) = (0.11, 0.04, 0.37, 0.48)$	25	37	50
$(\pi_{01}, \pi_{02}, \pi_{03}, \pi_{04}) = (0.18, 0.06, 0.59, 0.18)$	$(\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14}) = (0.16, 0.05, 0.53, 0.25)$	237	402	551
$(\pi_{01}, \pi_{02}, \pi_{03}, \pi_{04}) = (0.56, 0.11, 0.12, 0.21)$	$(\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14}) = (0.64, 0.13, 0.14, 0.09)$	60	96	139

improvement if the size of ulcer crater is greater than 2/3 healed or completely healed. A subject was considered worsening if the size of ulcer crater is larger than before. A subject was considered no change if the size of ulcer crater is less than 1/3 healed. The numbers of subjects observed in the test treatment group for the three categories are 5, 8, and 19, respectively, while the proportions of subjects observed in the control group for the three categories are 12, 10, and 10, respectively. The first category indicated “worsening” whilst the second and third categories indicated “no change” and “improvement” respectively.

Based on these observed frequencies from the two groups, the hypotheses $H_0: \beta = 0$ vs $H_a: \beta < 0$ are tested to assess whether the test treatment is more effective in treating the disease. In particular, the MLE of $\theta = (\beta, \alpha_1, \alpha_2)$ under H_0 is found to be $\hat{\theta}_0 = (0, \tilde{\alpha}_1, \tilde{\alpha}_2) = (0, -0.5341, -0.4769)$, while the MLE of $\theta = (\beta, \alpha_1, \alpha_2)$ under H_1 is equal to $\hat{\theta} = (\tilde{\beta}, \tilde{\alpha}_1, \tilde{\alpha}_2) = (-1.1679, 0.0663, 0.1235)$. The test statistic $-2[\log L_0(\hat{\theta}_0) - \log L(\hat{\theta})]$ is evaluated and is found to be 5.18, which is significant when compared to the lower 5% value of a chi-square distribution with one degree of freedom. In other words, there is enough evidence to show that the test treatment is more effective in treating the disease, i.e., the response probability of falling into category “improving” is higher than that of the other treatment.

Example 2

A clinical trial was conducted for the evaluation of efficacy of a test treatment (intravenous infusion) to treat subjects who experienced moderate to severe Crohn’s disease that was refractory to steroids and immunosuppressant. A total of 120 subjects were enrolled in this study. Qualified subjects were randomized in a 1:1 ratio into the test treatment group and a control group. Subjects received treatment on days 0, 14, and 28. The primary endpoint considered in this clinical study was the crohns disease activity index (CDAI) score at Day 42. Subjects were classified into three ordered categories where the first category indicated “worsening” whilst the second and third categories indicated “no change” and “improvement” respectively, depending upon the CDAI score at Day 42. A subject was considered a responder (improvement) if there is a decrease from baseline of at least 70 points in CDAI at Day 42. A subject was considered worsening if his/her CDAI score increases at Day 42. A subject was considered no change if his/her CDAI score is not considered worsened or improved. The numbers of subjects observed in the test treatment group for the three categories are 9, 36, and 15, respectively, while the proportions of subjects observed in the control group for the three categories are 12, 39, and 9, respectively.

Suppose that model (1) is used to model the response frequencies of the various categories. Based on these observed frequencies from the two groups, the hypotheses $H_0: \beta = 0$ vs $H_a: \beta < 0$ are tested to assess whether

the test treatment is more effective in treating the disease. In particular, the MLE of $\theta = (\beta, \alpha_1, \alpha_2)$ under H_0 is found to be $\hat{\theta}_0 = (0, \tilde{\alpha}_1, \tilde{\alpha}_2) = (0, -0.1335, 1.1394)$ whereas the MLE of $\theta = (\beta, \alpha_1, \alpha_2)$ under H_1 is equal to $\hat{\theta} = (\tilde{\beta}, \tilde{\alpha}_1, \tilde{\alpha}_2) = (-0.636, 0.2148, 1.4877)$. The test statistic $-2[\log L_0(\hat{\theta}_0) - \log L(\hat{\theta})]$ is evaluated and is found to be 1.89, which is insignificant when compared to the lower 5% value of a chi-square distribution with one degree of freedom. In other words, there does not exist significant evidence to conclude that the test treatment is more effective in treating the disease, i.e., the response probability of falling into category “improving” is not higher than that of the other treatment.

Example 3

This study was to compare four treatments on the extent of trauma due to subarachnoid hemorrhage.^[11] For illustration purpose, two out of the four treatments were selected for comparison, namely, “Placebo” (control) and “Low dose” (treatment). A total of 400 subjects were enrolled in this study. 190 and 210 qualified subjects were randomly allocated into the test treatment group and a control group, respectively. Subjects were classified into four ordered categories, namely “Death,” “Vegetative state,” “Major disability,” “Recovery” (including minor disability and good recovery), according to the outcome after the treatments. The numbers of subjects observed in the four ordered categories for the test treatment group were 48, 21, 44, and 77, respectively, whereas the proportions of subjects observed in the four categories for the control group were 59, 25, 46, and 80, respectively.

Based on these observed frequencies from the two groups, the hypotheses $H_0: \beta = 0$ vs $H_a: \beta < 0$ were tested to assess whether the test treatment is more effective in treating the disease. In particular, the MLE of $\theta = (\beta, \alpha_1, \alpha_2, \alpha_3)$ under H_0 were found to be $\hat{\theta}_0 = (0, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3) = (0, 0.1387, -0.5053, -0.5911)$, where as the MLE of $\theta = (\beta, \alpha_1, \alpha_2, \alpha_3)$ under H_1 is equal to $\hat{\theta} = (\tilde{\beta}, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3) = (-0.1019, -0.3347, -1.1789, -0.5077)$. The test statistic $-2[\log L_0(\hat{\theta}_0) - \log L(\hat{\theta})]$ was evaluated and found to be 36.67, which is significant when compared to the lower 5% value of a chi-square distribution with one degree of freedom. In other words, there is enough evidence to show that the test treatment is more effective in treating the disease, i.e., the response probability of falling into category “Recovery” is higher than that of the other treatment.

CONCLUSION

This entry has considered the problem of comparing the effectiveness of two treatments based on categorical responses. Instead of using a multinomial model to describe the probabilities of the responses falling into the various categories and comparing the effectiveness of the

two treatments by a conditional test, we have presented an alternative approach to model the response probabilities via a parametric form and the comparison of these probabilities is translated into the comparison of the associated model parameters. Maximum likelihood estimates (MLE) of the parameters are derived and the comparison is based on a chi-squared test. Furthermore, the required sample size to achieve a given level of power to detect the difference is also given. Furthermore, three data sets are selected and analyzed based on the derived results as an illustration. Although the discussion of the example and the numerical results given are based on only three or four categories, it can be easily extended to cases with more than four categories. In general, this approach is more efficient in computing time and provides a viable tool for practitioners to assess whether a new treatment is more effective than another treatment when the responses are categorical.

REFERENCES

1. Agresti, A. *Categorical Data Analysis*; Wiley: New York, 1990.
2. Cohen, A.; Sackrowitz, H.B. Testing whether treatment is "better" than control with ordered categorical data: definitions and complete class theorem. *Stat. Decis.* **2000**, *68*, 1–25.
3. McCullagh, P. Regression models for ordinal data. *J. R. Stat. Soc. B* **1980**, *42*, 109–142.
4. Ananth, C.V.; Kleinbaum, D.G. Regression models for ordinal responses: a review of methods and applications. *Int. J. Epidemiol.* **1997**, *26*, 1323–1333.
5. Cohen, A.; Madigan, D.; Scakrowitz, H.B. Effective directed tests for models with ordered categorical data. *Aust. N. Z. J. Stat.* **2003**, *45*, 285–300.
6. Silvapulle, M.J.; Sen, P.K. *Constrained Statistical Inference*; Wiley: New York, 2011.
7. Chow, C. S.; Tse, S. K.; Yang, C.Y.; Cosmatos, D.; Chi, E. Statistical test for ordered categorical data in clinical trials. *Drug Inf. J.* **2008**, *42*, 617–624.
8. Cox, D.R.; Hinkley, D.V. *Theoretical Statistics*; Chapman and Hall: London, 1974.
9. Serfling, R.J. *Approximation Theorems of Mathematical Statistics*; Wiley: New York, 2008.
10. Doll, R.; Pygott, F. Factors influencing the rate of healing gastric ulcers. *Lancet* **1952**, *259*, 171–175.
11. Chuang–Stein, C.; Agresti, A. A review of tests for detecting a monotone dose-response relationship with ordinal response data. *Stat. Med.* **1997**, *16*, 2599–2618.

Ordered Multiple Class Receiver Operating Characteristic (ROC) Analysis

Christos T. Nakas

Laboratory of Biometry, University of Thessaly, N. Ionia, Magnesia, Greece

Constantin T. Yiannoutsos

Division of Biostatistics, Indiana University School of Medicine, Indianapolis, Indiana, U.S.A.

Abstract

Ordered Multiple Class Receiver Operating Characteristic (ROC) Analysis The receiver operating characteristic curve (ROC) is a graphical representation resulting from plotting the proportion of diseased subjects with a positive test result vs. the proportion of healthy subjects with a positive test that result from considering all possible values of a diagnostic test used as a cutoff. This entry provides a detailed definition of a ROC curve and supplies a sample application.

Keywords: Bootstrap; Classification; Diagnostic testing; ROC analysis; ROC curves; True class rate; U-statistics.

INTRODUCTION

Assessment of diagnostic markers for classification and prediction in two-class classification problems is commonly performed with the use of receiver operating characteristic (ROC) curves. The two classes under study are commonly referred to as the diseased and healthy populations. The ROC curve is a graphical representation resulting from plotting the proportion of diseased subjects with a positive test result (TPR, true positive rate) vs. the proportion of healthy subjects with a positive test (FPR, false positive rate) that result from considering all possible values of a diagnostic test used as a cutoff. The area under the ROC curve (AUC) is used as a global measure of the test's discriminatory accuracy between the two groups. Estimation of the AUC and its standard error is performed by non-parametric methods and parametric means.^[1] In three-class diagnostic problems three true class rates (TCR) are defined using an appropriate classification criterion. Extension to multiple-class problems is straightforward. In the three-class case, an ROC surface is generated in three dimensions by considering all possible diagnostic test values. The volume under the ROC surface (VUS) can be used as a global measure of the three-class discriminatory ability of the test under consideration. Estimation and inference involving the VUS are also done in a manner similar to the estimation of the two-dimensional AUC. The VUS can be used whenever the diagnostic marker measurements are continuous or comprised of ordered categories (ordinal data).

ROC SURFACE DEFINITION AND CONSTRUCTION

Roc Curve Definition and Construction

ROC curves are commonly used for the evaluation of diagnostic markers in two-class testing. Let $X_{11}, \dots, X_{1m}$ be m measurements obtained from the first class and $X_{21}, \dots, X_{2n}$ be n measurements from the second class. For example, these two classes may involve healthy patients or patients having a certain disease. Without loss of generality, suppose that subjects from the first class tend to have lower measurements than subjects from the second class. The ROC curve is defined as the set of points $(FPR(c), TPR(c))$ in the unit square, for all possible values of c , where $TPR(c) = P(X_2 > c)$ and $FPR(c) = P(X_1 > c)$ signify the probability that a diseased and a healthy patient, respectively, will have a positive test defined at the threshold point c . The TPR is the familiar concept of sensitivity of a test. An equivalent construction of the ROC curve is defined by the set of points $(TNR(c), TPR(c))$ in the unit square, where $TNR(c) = P(X_1 < c)$ is the probability that a healthy subject will have a negative test. This is the specificity of a test. The AUC is used as an overall performance index of the diagnostic accuracy of a marker, and it is equal to $P(X_1 < X_2)$. That is, the AUC is the average chance that the test score of a randomly selected subject from the first (healthy patient) group will be lower than the score of a randomly selected subject from the second (diseased patient) group. The AUC derived from a noninformative marker is 0.5 (implying that

the test resulting from this marker is no better than the flip of a coin). Conversely, the AUC resulting from a perfectly discriminating marker is 1.0 (that is, a randomly selected pair of observations from the healthy and diseased groups will always have the correct ordering of marker values). The interested reader should review Refs. [1–3], some recent textbooks on ROC curve methodology.

The Roc Surface And Hypersurface

Efforts to generalize the utility of the ROC curve for the assessment of diagnostic markers in multigroup classification problems have led to a number of approaches for the extension of ROC curve theory.^[4–10] The difficulty in generalizing the ROC curve to more than two classes results from the fact that a decision rule for a K -group classification will produce K true classification rates (TCR) and $K \times (K - 1)$ false classification rates (FCR), where $TCR_k = P(C = k|D = k)$, $k = 1, \dots, K$ and $FCR_{ij} = P(C = i|D = j)$, $i, j = 1, \dots, K$, $i \neq j$, D is the true class membership of an individual while C is the class assigned by the testing procedure. A concise graphical representation of such a decision rule is the K -dimensional ROC hypersurface that can be used for the assessment of the diagnostic capacity of the marker under study.

For the sake of simplicity we illustrate first how an ROC surface is generated from a three-group classification problem based on a diagnostic marker X . Fig. 1 depicts a hypothetical situation for the distributions of marker measurements X_1 , X_2 , and X_3 obtained from three classes where, in general, measurements from class 1 are lower than those from class 2, and measurements from class 2 are lower than those from class 3. A decision rule that classifies subjects in one of these three classes can be defined as follows: For two ordered threshold points $c_1 < c_2$ (Fig. 1).

IF marker value $X < c_1$ THEN assign subject to class 1
ELSE IF $c_1 < X < c_2$ THEN assign to class 2
ELSE assign to class 3

In practice ties between measurements and threshold values may occur. In that case, the less than signs “<” above can arbitrarily be replaced with less than or equal signs “≤” as

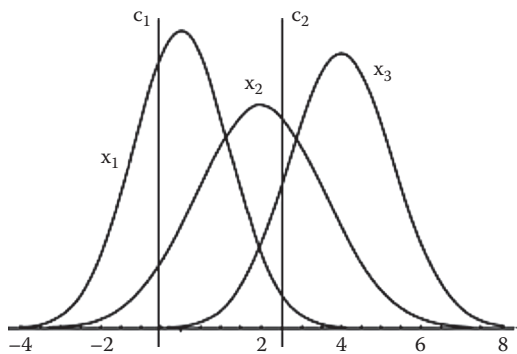


Fig. 1 Hypothetical overlapping distributions of marker measurements in a three-class experiment

long as the sets defined by the decision rule remain disjoint (i.e., no subject can be classified as being a member of more than one group). Varying the ordered decision thresholds $c_1 < c_2$ over all possible marker measurements, TCR are defined. These can be plotted in three dimensions to produce an ROC surface in the unit cube $[0,1] \times [0,1] \times [0,1]$. The TCR take values with corner coordinates $\{(1,0,0), (0,1,0), (0,0,1)\}$. This construction is a direct generalization of the ROC curve in three dimensions because ignoring c_2 and the subrules that this implies a conventional ROC curve between classes 1 and 2 is constructed. Similarly, ignoring c_1 results in an ROC curve between classes 2 and 3.

In the case of continuous or ordered categorical (ordinal) data, if the distributions of the marker values for the three groups are $X_1 \sim F_1$, $X_2 \sim F_2$, and $X_3 \sim F_3$, the theoretical TCR are defined by $F_1(c_1)$, $F_2(c_2) - F_2(c_1)$, and $1 - F_3(c_2)$, respectively, for all possible values of c_1, c_2 , with $c_1 < c_2$. The ROC surface is constructed by plotting the points $(F_1(c_1), F_2(c_2) - F_2(c_1), 1 - F_3(c_2))$ in three dimensions. In practice, the empirical cumulative distribution functions are used. These are defined from the data as the proportion of subjects in each group that are below the threshold c . For example, based on n_1 observations from the first class, the empirical cumulative distribution estimate of F_1 is defined as $\hat{F}_1(c) = \sum_{i=1}^{n_1} I(X_i < c) / n_1$ where $I(X_i < c)$ is the indicator function and is equal to 1 if observation $X_i < c$, and 0 otherwise. Fig. 2 shows an empirical ROC surface with marker measurements simulated from three different, overlapping normal distributions. The theoretical VUS constructed by the decision rule described above is equal to the probability that three random measurements, one from each class, are classified in the correct order $X_1 < X_2 < X_3$. That is, $VUS = P(X_1 < X_2 < X_3)$. This is a straightforward generalization of the two-dimensional ROC curve. An unbiased non-parametric estimator of VUS is given by

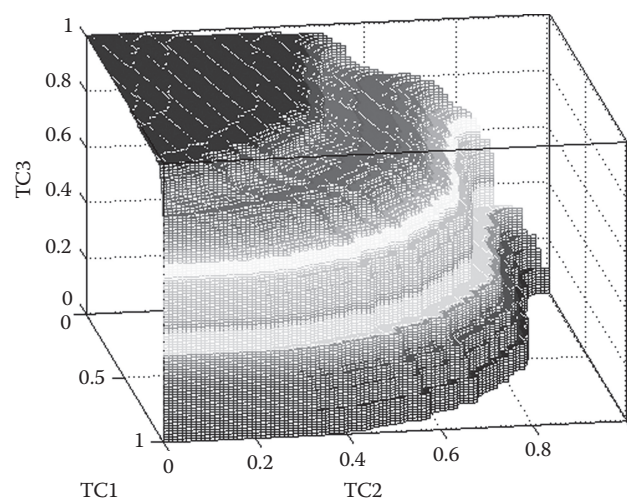


Fig. 2 ROC surface for measurements sampled from the three hypothetical distributions of Fig. 1

$$VUS = \frac{1}{n_1 n_2 n_3} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} I(X_{1i}, X_{2j}, X_{3k})$$

where n_1 , n_2 , and n_3 are the sample sizes from classes 1, 2, and 3, respectively, and $I(X_1, X_2, X_3)$ equals 1 if X_1, X_2 , and X_3 are in the correct order, and 0 otherwise. In practice, ties between measurements may occur. In that case we define

$$I(X_1, X_2, X_3) = \begin{cases} 1 & X_1 < X_2 = X_3 \\ 1/2 & X_1 < X_2 = X_3 \text{ or } X_1 = X_2 < X_3 \\ 1/6 & X_1 = X_2 = X_3 \\ 0 & \text{otherwise} \end{cases}$$

When the diagnostic marker is completely uninformative, i.e., when there is perfect overlap between the distributions of the three classes, VUS takes the value 1/6 (that is, classification into three groups using that marker is no better than random chance). Conversely, perfect separation between the three classes with $X_1 < X_2 < X_3$ for all measurements yields $VUS = 1$. Parametric modeling of the ROC surface and estimation of the respective VUS has also been developed.^[5,11]

Generalization to K -class ($K > 3$) is straightforward. An ROC hypersurface can be constructed (but not visualized) by defining a diagnostic rule with $K - 1$ ordered decision thresholds. The K TCR thus defined should be plotted in a K -dimensional space. The ROC hypersurface is produced by varying the $K - 1$ ordered decision thresholds and calculating the respective TCR. A non-parametric unbiased estimate of the volume under the hypersurface is given by

$$VUS = \frac{1}{n_1 n_2 \dots n_k} \times \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \dots \sum_{i_k=1}^{n_k} I(X_{1i_1}, X_{2i_2}, \dots, X_{ki_k})$$

where n_i , $i = 1, 2, \dots, K$ are the sample sizes from classes 1, 2, ..., K , respectively, and $I(X_{1i_1}, X_{2i_2}, \dots, X_{ki_k})$ is defined in analogy to the three-class case. The VUS under this hypersurface takes the value 1 when there is perfect separation between the K -classes in the desired order, and it takes the value $1/(k!)$ when the K -class classification test is uninformative. The variance of the VUS can be calculated based on U-statistics theory or bootstrap techniques.^[10] This can be used to produce $(1 - \alpha)\%$ confidence intervals and perform statistical inference.^[10,12]

Construction of non-parametric umbrella ROC graphs and estimation of the respective umbrella volume (UV), appropriate in studies where the ordering of interest is $X_1 < X_3 > X_2$, or $X_2 > X_1 < X_3$ is discussed in Ref. [13].

AN APPLICATION

Glial cells are found in the central nervous system. They provide structural support and protection for neurons. When neurons are injured, glial cells form a fibrous network

in the area of injury. The resulting “gliosis” is prominent in a number of conditions that damage the central nervous system such as multiple sclerosis and stroke. In particular, gliosis has been found in neuropathology studies of injury to the central nervous system resulting from infection with the human immunodeficiency virus (HIV).^[14] Myoinositol (MI) is a marker of glial cells in the brain.^[15] Its levels can be measured by an imaging technique called magnetic resonance spectroscopy (MRS). Increased levels of MI, or its ratio over another marker creatine (Cr) MI/Cr, particularly in the basal ganglia have been associated with neurological impairment resulting from HIV infection.^[16,17] MI/Cr measurements were assessed here as an illustration of a diagnostic tool for the discrimination of three subject groups: HIV-negative patients (NEG), HIV-positive patients neurologically asymptomatic (NAS), and HIV-positive patients with neurological impairment [AIDS Dementia Complex (ADC)]. Details of this study have been described in detail elsewhere.^[18,19] There are six possible orderings for MI/Cr measurements of the three classes, i.e., $NEG < NAS < ADC$, $NAS < NEG < ADC$, $NEG < ADC < NAS$, $NAS < ADC < NEG$, $ADC < NEG < NAS$, and $ADC < NAS < NEG$. The ROC surfaces corresponding to these orderings were constructed, and the volumes under these ROC surfaces were estimated to be 0.4299, 0.2090, 0.1542, 0.0867, 0.0647, and 0.0556, respectively. These results indicate that the strongest evidence supports an increasing trend in the MI/Cr measurements of the HIV-negative subjects, HIV-positive asymptomatic patients, and HIV-positive patients with neurological impairment. Fig. 3 shows the corresponding ROC surface. A 95% confidence interval for the corresponding VUS is (0.328, 0.5318) based on the U-statistics methodology described in Refs. [10,12]. Because the uninformative value of 1/6 for a three-class VUS is excluded from the 95% confidence interval, the implication is that MI/Cr measured via MRS in the basal ganglia is informative

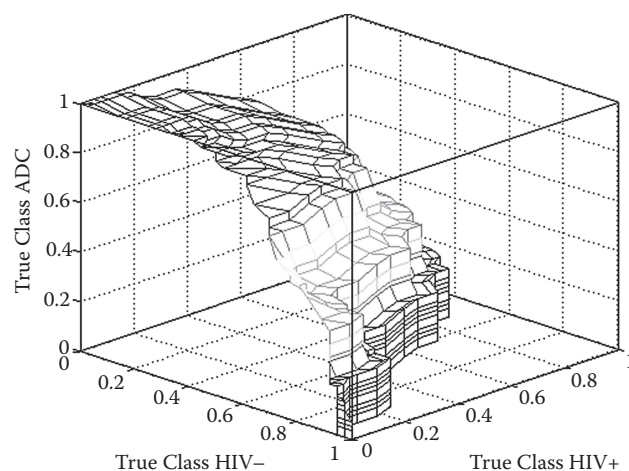


Fig. 3 ROC surface for MI/Cr in the basal ganglia measurements for the classification $NEG < NAS < ADC$

for the classification of subjects in the NEG, NAS, and ADC classes. In addition, when three measurements are given, one from each class, the probability that the MI/Cr level obtained from an HIV-negative subject is lower than that of the one obtained from an HIV-positive person and that both are lower than the MI/Cr levels measured in an individual suffering from HIV-related neurological impairment is estimated to be about 43% of the time.

CONCLUSIONS

Medical diagnostic tests often result in more than two classes. Traditional techniques such as ROC curve analysis can be generalized to three or more group classification problems by appealing to ROC surfaces in three or higher dimensions. In a similar manner, the area under the two-dimensional ROC curve, a global index of test accuracy, can be extended to three or higher dimensions by use of the VUS or hypersurface, respectively. We describe here non-parametric estimation of the VUS and use well-developed U-statistics methodology to derive estimates of the VUS along with confidence intervals. This formulation is useful especially when an inherent ordering exists in the marker measurements between the classes under study.

Because the relative size, rather than the absolute size, of the measurements is used, monotonic transformations of the measurement scale will result in the same estimated ROC surface and VUS in direct extension of this useful property of the two-dimensional ROC curve and AUC. In addition, the VUS defined above is the probability that the marker measurements from k ($k > 2$) subjects, one from each class, will be correctly ordered. This is a helpful interpretation, identical to the two-dimensional AUC. The VUS therefore maintains a number of useful properties of the two-dimensional ROC curve.

Thus, the described methodology addresses directly the question of the marker discriminatory performance over all the classes under study. Pairwise analysis of ROC curves can be performed in a post hoc manner as a detailed approach in multiple-class classification problems. In the example provided, measuring the VUS produced by MRS-measured MI/Cr levels shows that a strong trend exists between HIV-negative subjects, HIV-positive neurologically asymptomatic patients, and HIV-positive patients suffering from neurological complications associated with the infection. This observation has profound implications for the use of this marker in clinical diagnosis of HIV-associated neurological impairment as well as clinical consequences for our understanding of HIV-related neurological injury. Because MI/Cr levels of HIV-positive unimpaired patients were between those of HIV-negative and neurologically impaired subjects, the suggestion is that neurological injury occurs among

HIV-infected individuals well before clinical symptoms (leading to an ADC diagnosis) are observed.^[18,19] A three-class ROC surface analysis of these data therefore is a useful means of illuminating these relationships.

REFERENCES

1. Pepe, M.S. *The Statistical Evaluation of Medical Tests for Classification and Prediction*; Oxford University Press: Oxford, 2003.
2. Zhou, X.H.; Obuchowski, N.A.; McClish, D.K. *Statistical Methods in Diagnostic Medicine*; John Wiley & Sons, Inc: New York, 2002.
3. Krzanowski, W.J.; Hand, D.J. *ROC Curves for Continuous Data*; Chapman & Hall/CRC: Boca Raton, FL, 2009.
4. Scurfield, B.K. Multiple-event forced-choice tasks in the theory of signal detectability. *J. Math. Psychol.* **1996**, *40*, 253–269.
5. Mossman, D. Three-way ROCs. *Med. Decis. Mak.* **1999**, *19*, 78–89.
6. Hand, D.J.; Till, R.J. A simple generalization of the area under the ROC curve for multiple class classification problems. *Mach. Learning.* **2001**, *45*, 171–186.
7. Fawcett, T. *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Technical report HPL-2003-4*; HP Laboratories: Palo Alto, CA, 2003.
8. Lachiche, N.; Flach, P.A. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves. *Proceedings of the 20th International Conference on Machine Learning (ICML'03)*; AAAI Press: 2003, 416–423.
9. Ferri, C.; Hernandez-Orallo, J.; Salido, M.A. Volume under the ROC surface for multi-class problems. *Lecture Notes in Artificial Intelligence, Machine Learning: ECML 2003*; Springer: New York, 2003, 2837, 108–120.
10. Nakas, C.T.; Yiannoutsos, C.T. Ordered multiple-class ROC analysis with continuous measurements. *Stat. Med.* **2004**, *23*, 3437–3449.
11. Xiong, C.; van Belle, G.; Miller, P.J.; Morris, J.C. Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. *Stat. Med.* **2006**, *25*, 1251–1273.
12. Dreiseitl, S.; Ohno-Machado, L.; Binder, M. Comparing three-class diagnostic tests by three-way ROC analysis. *Med. Decis. Mak.* **2000**, *20*, 323–331.
13. Nakas, C.T.; Alonzo, T.A. ROC graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering. *Biometrics* **2007**, *63*, 603–609.
14. Navia, B.A.; Cho, E.S.; Petito, C.K.; Price, R.W. The AIDS dementia complex: II. Neuropathology. *Ann. Neurol.* **1986**, *19*, 525–535.
15. Brand, A.; Richter-Landsberg, C.; Leibfritz, D. Multinuclear NMR studies on the energy metabolism of glial and neuronal cells. *Develop. Neurosci.* **1993**, *15*, 289–298.
16. Lopez-Villegas, D.; Lenkinski, R.E.; Frank, I. Biochemical changes in the frontal lobe of HIV-infected individuals detected by magnetic resonance spectroscopy. *Proc. Natl. Acad. Sci.* **1997**, *94*, 9854–9859.

17. Yiannoutsos, C.T.; Ernst, T.; Chang, L.; Lee, P.L.; Richards, T.; Marra, C.M.; Meyerhoff, D.J.; Jarvik, J.G.; Richards, T.; Kolson, D.; Schifitto, G.; Ellis, R.J.; Swindells, S.; Simpson, D.M.; Miller, E.N.; Gonzalez, R.G.; Navia, B.A. Regional patterns of brain metabolism in AIDS Dementia Complex. *Neuroimage* **2004**, *23*, 928–935.
18. Chang, L.; Lee, P.L.; Yiannoutsos, C.T.; Ernst, T.; Marra, C.M.; Richards, T.; Kolson, D.; Schifitto, G.; Jarvik, J.G.; Miller, E.N.; Lenkinski, R.; Gonzalez, G.; Navia, B.A. HIV MRS Consortium. Amulticenter in vivo proton-MRS study of HIV-associated dementia and its relationship to age. *Neuroimage* **2004**, *23*, 1336–1347.
19. Yiannoutsos, C.T.; Nakas, C.T.; Navia, B.A. Assessing multiple-group diagnostic problems with multi-dimensional receiver operating characteristic surfaces: Application to proton MR Spectroscopy (MRS) in HIV-related neurological injury. *Neuroimage* **2008**, *40*, 248–255.

Ordinal Data: Crossover Design Analysis

Kung-Jong Lui

Department of Mathematics and Statistics, San Diego State University, San Diego, California, U.S.A.

Abstract

We may commonly encounter, in a crossover trial, the underlying patient response of interest on an ordinal scale—worse, same, slightly improved, and better. Since these ordinal responses are not continuous or interval data, the ordinal responses are generally not appropriate for arithmetic operation. We discuss the approaches to analyze ordinal data under a crossover design. These include the normal random effects proportional odds model, the simple change score method, and the conditional approach with using the generalized odds ratio to measure the relative treatment effects. We focus our discussion on the most frequently used AB/BA (or simple crossover) design. We present both asymptotic and exact test procedures, as well as asymptotic and exact interval estimators. We further extend the discussion to a three-treatment three-period crossover trial. We note that one can easily apply the ideas presented here to derive the corresponding test procedures and estimators when using a Latin squares to reduce the large number of groups for comparing four or more treatments in a crossover trial with a complete set of treatment-receipt sequences, or when employing an incomplete block crossover design to decrease the risk of being lost to follow-up in a crossover trial with long duration.

Keywords: Proportional odds model; Simple change score method; Generalized odds ratio; Ordinal data; Latin square; Incomplete block design.

INTRODUCTION

To save the number of patients or improve power of test procedures for a parallel groups design, the crossover design has been often employed to study non-curative treatments for chronic diseases. In randomized clinical trials, it is quite common to encounter the patient response on an ordinal scale. Because the relative distance between successive ordered categories are not truly comparable, ordinal data are generally inappropriate for arithmetic operations.^[1] For example, consider the AB/BA crossover trial conducted by 3 M Health Care Ltd to compare the suitability of two inhalation devices (A and B) among asthma patients using a standard inhaler device to deliver salbutamol.^[1] Patients were randomly assigned to either group 1 using device A for one week followed by device B for another week or group 2 using device B for one week followed by device A for another week. Patients were then requested to assess the clarity of leaflet instructions regarding devices on a four-point ordinal scale—“easy,” “only clear after rereading,” “not very clear,” and “confusing.” Since these ordinal levels of responses are not on an interval scale, the relative distance between “easy” and “only clear after rereading” is not, for example, comparable to that between “only clear after rereading” and “not every clear.” Although one may often arbitrarily assign scores, such as 0, 1, 2, 3, to

these ordinal categories and apply the t-test to test equality between treatments, providing a meaningful and easily understood summary measure based on these artificial scores to quantify the relative effect between treatments can be difficult.

Because of its simplicity, the use of the AB/BA design (or simple carry-over design) has accounted for a large proportion of crossover trials employed in practice.^[2–5] Thus, we concentrate our discussion on the development of analysis in ordinal data under the AB/BA trial. These include the normal random effects proportional odds model,^[1] the simple change score method,^[6] and the conditional approach^[7,8] with using the generalized odds ratio (GOR)^[9] to measure the relative treatment effects. Because it can be sometimes more appealing and economical to compare more than one experimental treatment with a placebo in a single trial instead of several trials, each having its own experimental and placebo arms, we extend the discussion to a three-treatment three-period crossover design.^[8] Also, because a crossover trial, in which each patient receives all treatments under comparison, can take a long time to implement especially when there are three or more treatments, we note that one can employ similar ideas as focused here to derive the test procedures and estimators under an incomplete block design.^[3] For comparing four or more treatments, we further note that it is

straightforward to develop the corresponding test procedures and estimators presented here when one uses the Latin square^[3] in a crossover trial to reduce the number of groups with different treatment-receipt sequences.^[8] Other methods to analyze ordinal data, such as the use of the log-linear model^[2] or Wilcoxon test, appear elsewhere.^[6,10]

We assume that there are no carry-over effects due to the treatment administered at an earlier period with an adequate wash-out period. If we cannot ensure this assumption of no carry-over effects with an adequate wash-out period to hold onto the basis of our subjective knowledge, as noted elsewhere,^[3,11,12] we may not want to use the crossover design. We consider the situations in which the underlying patient response is on an ordinal scale with J -ordered categories in ascending order.

AB/BA CROSSOVER DESIGN

Suppose that we compare an experimental treatment B with an active control treatment (or a placebo) A in a crossover trial. We randomly assign n_1 patients to group ($g =$) 1 with A-then-B sequence, in which patients receive treatment A at period 1 and then cross over to receive treatment B at period 2, and n_2 patients to group ($g =$) 2 with B-then-A sequence, in which patients receive treatment B at period 1 and then cross over to receive treatment A at period 2.

Normal Random Effects Proportional Odds Model

To account for the intraclass correlation between responses within patients and the period effect, Ezzet and Whitehead^[1] first proposed a normal random effects proportional odds model^[13,14] to analyze ordinal data. For patient i ($= 1, 2, \dots, n_g$) from group g ($= 1, 2$) at period z ($= 1, 2$), let $X_{iz}^{(g)} (= 1$ for the patient receiving the experimental treatment B at period z , and $= 0$, otherwise) and $Z_{iz}^{(g)} (= 1$ for period $z = 2$, and $= 0$, otherwise) denote the treatment and period covariates of this patient, respectively. We let $\mu_i^{(g)}$ denote the random effect due to the i th subject in group g , and assume $\mu_i^{(g)}$'s to be independent and identically distributed as a normal distribution with mean 0 and unknown variance σ^2 . We further let $Y_{iz}^{(g)} = j$ ($= 1, 2, \dots, J$) denote in group g the response of the i th patient at period z falling in the j th ordered category. We assume that the logit of the cumulative probability for a given patient response $Y_{iz}^{(g)}$ with $\mu_i^{(g)}$ fixed falling below (and including) the j th ($= 1, 2, 3, \dots, J - 1$) ordered category is:^[1,14]

$$\log \left(\frac{P(Y_{iz}^{(g)} \leq j | \mu_i^{(g)})}{P(Y_{iz}^{(g)} > j | \mu_i^{(g)})} \right) = \alpha_j + \beta_{BA} X_{iz}^{(g)} + \gamma Z_{iz}^{(g)} + \mu_i^{(g)} \quad (1)$$

where β_{BA} denotes the relative effect of treatment B to treatment A, and γ denotes the effect of period 2 vs. period 1. When there is no difference in effects between treatments

A and B, $\beta_{BA} = 0$. If the patient receiving treatment B tends, as compared with he/she receiving treatment A, to fall in lower categories for a given fixed period, then $\beta_{BA} > 0$. If the patient receiving treatment B tends, as compared with he/she receiving treatment A, to fall in higher categories for a fixed period, then $\beta_{BA} < 0$. Note that for a given fixed patient under model (Eq. 1), we have,

$$P(Y_{iz}^{(g)} \leq j | \mu_i^{(g)}) = \frac{\exp(\alpha_j + \beta_{BA} X_{iz}^{(g)} + \gamma Z_{iz}^{(g)} + \mu_i^{(g)})}{1 + \exp(\alpha_j + \beta_{BA} X_{iz}^{(g)} + \gamma Z_{iz}^{(g)} + \mu_i^{(g)})} \quad (2)$$

This implies that the probability for the patient response $Y_{iz}^{(g)}$ falling in the j th ($= 1, 2, 3, \dots, J$) ordered category is

$$P(Y_{iz}^{(g)} = j | \mu_i^{(g)}) = P(Y_{iz}^{(g)} \leq j | \mu_i^{(g)}) - P(Y_{iz}^{(g)} \leq j-1 | \mu_i^{(g)}) \quad (3)$$

where $P(Y_{iz}^{(g)} \leq 0 | \mu_i^{(g)}) = 0$ and $P(Y_{iz}^{(g)} \leq J | \mu_i^{(g)}) = 1$. Thus, the joint probability of $(Y_{i1}^{(g)} = j, Y_{i2}^{(g)} = j')$ for a randomly selected patient from group g is given by^[1]

$$P(Y_{i1}^{(g)} = j, Y_{i2}^{(g)} = j') = \int P(Y_{i1}^{(g)} = j | u) P(Y_{i2}^{(g)} = j' | u) f(u) du \quad (4)$$

where $f(u)$ is the probability density function of the assumed normal distribution for $\mu_i^{(g)}$. For simplicity, we let $\pi_{jj'}^{(g)}$ denote the probability $P(Y_{i1}^{(g)} = j, Y_{i2}^{(g)} = j')$. We define $n_{jj'}^{(g)}$ as the number of patients in group g ($= 1, 2$) falling in the cell with the response vector $(Y_{i1}^{(g)} = j, Y_{i2}^{(g)} = j')$ among n_g patients. The random frequencies $\{n_{jj'}^{(g)} | j = 1, 2, 3, \dots, J, j' = 1, 2, 3, \dots, J\}$ then follow the multinomial distribution with parameters n_g and $\{\pi_{jj'}^{(g)} | j = 1, 2, 3, \dots, J, j' = 1, 2, 3, \dots, J\}$. The likelihood is then proportional to:

$$\prod_{g=1}^2 \prod_{j=1}^J \prod_{j'=1}^J [P(Y_{i1}^{(g)} = j, Y_{i2}^{(g)} = j')]^{n_{jj'}^{(g)}} \quad (5)$$

which is involved in non-closed form of integrals (Eq. 4) and is required to use a sophisticated iterative numerical procedure. We can employ PROC GLIMMIX^[2,15] to obtain the maximum likelihood estimator (MLE) of parameters and the estimated asymptotic standard error (SE) of the MLE based on likelihood (Eq. 5). As noted elsewhere,^[9,16,17] however, the implicit assumption of the proportional odds model can be badly violated by many bivariate distributions. Also, we need to assume the random effects $\mu_i^{(g)}$ due to patients to follow a normal distribution for mathematical convenience when using PROC GLIMMIX. In practice, it is difficult to justify any distribution for the unobservable patient random effects. Furthermore, the number of patients is frequently small in a crossover trial and thereby the MLE can be inappropriate for use in small-sample cases. On the other hand, using the above normal random effects proportional odds model can

allow us to study whether there is a treatment-by-period interaction, and accommodate modeling a high-order or other crossover designs easily.

Simple Change Score Method

To alleviate the burden of employing a numerical iterative procedure to solve the likelihood equations and make the calculation be tractable, Senn^[6] proposed a simple change score approach by converting the base scores into a five-point scale. To illustrate Senn's idea, consider the above example regarding the assessment of the clarity of leaflet instructions between devices A and B on a four-point base scores, 0, 1, 2, 3, corresponding to the four-ordered categories "easy," "only clear after rereading," "not very clear," and "confusing." Using the difference in these base scores between periods within patients, Senn proposed classifying the score difference $\{-3\}$, $\{-2, -1\}$, $\{0\}$, $\{1, 2\}$, and $\{3\}$ as "much worse," "worse," "same," "better," and "much better," respectively. Based on these five-point change score scales, Senn then proposed that one could use various statistical methods for detecting trend effect to test equality of treatments. As noted by Ezzet and Whitehead,^[18] one of the main concerns in use of the simple change score method is to group patients with response changes that are not comparable into the same category. For example, all patients with base scores at two periods $\{(0, 1), (0, 2), (1, 2), (1, 3), (2, 3), (2, 4), (3, 4)\}$ would be grouped into the same "better" category. However, note that patients with responses changing from "easy" (coded as 0) to "not very clear" (coded as 2) is clearly different from patients with responses changing from "clear only after rereading" (coded as 1) to "confusing" (coded as 3). The above concern of grouping patients with incomparable extent of changes into the same category can exacerbate as the number of possible base scores for the original ordinal categories increases. In practice, we may wish to quantify the magnitude of the relative treatment effect. The development of a simple change score scheme to reflect the relative distance between different ordered levels and the creation of an easily appreciated summary measure of the relative treatment effect based on the change score method can be challenging.

Distribution-Free Random Effects Model with Using a Conditional Approach

We assume that the joint probability of $(Y_{i1}^{(g)}, Y_{i2}^{(g)})$ in group g for a given patient i ($=1, 2, \dots, n_g$) with $\mu_i^{(g)}$ fixed satisfies:^[19]

$$P(Y_{i1}^{(g)} < Y_{i2}^{(g)} | \mu_i^{(g)}) = \frac{1}{1 + \exp(\mu_i^{(g)} + \eta_{BA} X_{i1}^{(g)} + \gamma Z_{i1}^{(g)})} \times \frac{\exp(\mu_i^{(g)} + \eta_{BA} X_{i2}^{(g)} + \gamma Z_{i2}^{(g)})}{1 + \exp(\mu_i^{(g)} + \eta_{BA} X_{i2}^{(g)} + \gamma Z_{i2}^{(g)})}$$

$$P(Y_{i1}^{(g)} > Y_{i2}^{(g)} | \mu_i^{(g)}) = \frac{1}{1 + \exp(\mu_i^{(g)} + \eta_{BA} X_{i2}^{(g)} + \gamma Z_{i2}^{(g)})} \times \frac{\exp(\mu_i^{(g)} + \eta_{BA} X_{i1}^{(g)} + \gamma Z_{i1}^{(g)})}{1 + \exp(\mu_i^{(g)} + \eta_{BA} X_{i1}^{(g)} + \gamma Z_{i1}^{(g)})}$$

and

$$P(Y_{i1}^{(g)} = Y_{i2}^{(g)} | \mu_i^{(g)}) = 1 - P(Y_{i1}^{(g)} > Y_{i2}^{(g)} | \mu_i^{(g)}) - P(Y_{i1}^{(g)} < Y_{i2}^{(g)} | \mu_i^{(g)}) \quad (6)$$

where $\mu_i^{(g)}$'s are assumed to independently follow an unspecified probability density function $f_g(\mu)$, η_{BA} denotes the relative effect of treatment B to treatment A, and γ denotes the effect of period 2 vs. period 1. The GOR for paired-samples^[9,20,21] or Agresti's α' of responses on a given patient i in group g ($= 1, 2$) when the patient has covariates $(X_{i2}^{(g)}, Z_{i2}^{(g)})$ at period 2 vs. when he/she has covariates $(X_{i1}^{(g)}, Z_{i1}^{(g)})$ at period 1 under model (Eq. 6) is:

$$\frac{P(Y_{i1}^{(g)} < Y_{i2}^{(g)} | \mu_i^{(g)})}{P(Y_{i1}^{(g)} > Y_{i2}^{(g)} | \mu_i^{(g)})} = \exp(\eta_{BA} (X_{i2}^{(g)} - X_{i1}^{(g)}) + \gamma (Z_{i2}^{(g)} - Z_{i1}^{(g)})) \quad (7)$$

When there is no difference in effects between treatments A and B, $\eta_{BA} = 0$. When the response of the patient receiving treatment B tends to fall, for a given fixed period, in a category with higher than that of the patient receiving treatment A, $\eta_{BA} > 0$. When the response of the patient receiving treatment B tends to fall, for a given fixed period, in a category lower than that of the patient receiving treatment A, $\eta_{BA} < 0$. We define the GOR of responses for treatment B vs. treatment A as $\text{GOR}_{BA} = \exp(\eta_{BA})$. We let $\Pi_C^{(g)}$ and $\Pi_D^{(g)}$ denote for a randomly selected patient i from group g the probabilities $P(Y_{i1}^{(g)} < Y_{i2}^{(g)}) = \left(\int P(Y_{i1}^{(g)} < Y_{i2}^{(g)} | \mu) f_g(\mu) d\mu \right)$ and $P(Y_{i1}^{(g)} > Y_{i2}^{(g)}) = \left(\int P(Y_{i1}^{(g)} > Y_{i2}^{(g)} | \mu) f_g(\mu) d\mu \right)$, respectively. We can easily show that the GOR of responses between periods 2 and 1 for a randomly selected patient i from group g ($= 1, 2$) under model (Eq. 6) is also equal to:^[7,19]

$$\text{GOR}^{(g)} = \frac{\Pi_C^{(g)}}{\Pi_D^{(g)}} = \exp(\eta_{BA} (X_{i2}^{(g)} - X_{i1}^{(g)}) + \gamma (Z_{i2}^{(g)} - Z_{i1}^{(g)})) \quad (8)$$

which is neither a function of the assumed probability density function $f_g(\mu)$ nor a function of the random effects $\mu_i^{(g)}$. On the basis of Eq. 8, we can express the GOR of patient responses for treatment B vs. treatment A (for a fixed period) on a randomly selected patient i as:^[7]

$$\text{GOR}_{BA} = \exp(\eta_{BA}) = \left(\frac{\text{GOR}^{(1)}}{\text{GOR}^{(2)}} \right)^{1/2} \quad (9)$$

Note that $\Pi_C^{(g)}$ and $\Pi_D^{(g)}$ are, by definition, equal to $\sum_{j=1}^{J-1} \sum_{j'=j+1}^J \pi_{jj'}^{(g)}$ and $\sum_{j=2}^J \sum_{j'=1}^{j-1} \pi_{jj'}^{(g)}$, where $\pi_{jj'}^{(g)} = P(Y_{i1}^{(g)} = j, Y_{i2}^{(g)} = j')$, for $j = 1, 2, \dots, J$ and $j' = 1, 2, \dots, J$. As defined previously, we let

$n_{jj}^{(g)}$ denote the number of patients in group g ($= 1, 2$) falling in the cell with the response vector ($Y_{11}^{(g)} = j, Y_{12}^{(g)} = j'$) among n_g patients. We can estimate $\pi_{jj}^{(g)}$ by the sample proportion $\hat{\pi}_{jj}^{(g)} = n_{jj}^{(g)} / n_g$. Thus, we obtain a consistent estimator for $G\hat{O}R^{(g)}$ in group g ($= 1, 2$) as:

$$G\hat{O}R^{(g)} = \frac{\hat{\Pi}_C^{(g)}}{\hat{\Pi}_D^{(g)}} \quad (10)$$

where $\hat{\Pi}_C^{(g)} = \sum_{j=1}^{J-1} \sum_{j'=j+1}^J \hat{\pi}_{jj'}^{(g)}$ and $\hat{\Pi}_D^{(g)} = \sum_{j=2}^J \sum_{j'=1}^{j-1} \hat{\pi}_{jj'}^{(g)}$. This leads us to obtain a consistent estimator for $G\hat{O}R_{BA}$ (Eq. 9) as:

$$G\hat{O}R_{BA} = \left(\frac{G\hat{O}R^{(1)}}{G\hat{O}R^{(2)}} \right)^{1/2} \quad (11)$$

Using the delta method,^[14] we may obtain an estimated asymptotic variance of $\log(G\hat{O}R_{BA})$ as:^[7]

$$\text{Var}(\log(G\hat{O}R_{BA})) = \frac{\sum_{g=1}^2 \left(\frac{\hat{\Pi}_C^{(g)} + \hat{\Pi}_D^{(g)}}{n_g \hat{\Pi}_C^{(g)} \hat{\Pi}_D^{(g)}} \right)}{4} \quad (12)$$

Define $n_{dis}^{(g)} = n_C^{(g)} + n_D^{(g)}$, where $n_C^{(g)} = \sum_{j=1}^{J-1} \sum_{j'=j+1}^J n_{jj'}^{(g)}$ and $n_D^{(g)} = \sum_{j=2}^J \sum_{j'=1}^{j-1} n_{jj'}^{(g)}$ for $g = 1, 2$. We can show that the conditional probability mass function of $n_C^{(1)} = x_1$, given $n_C^{(+)} = n_C^{(1)} + n_C^{(2)}$ fixed, is the noncentral hypergeometric distribution^[14,20,22] given by:

$$P(n_C^{(1)} = x_1 | n_C^{(+)}, n_D^{(1)}, n_D^{(2)}, Q_{BA}) = \frac{\binom{n_D^{(1)}}{x_1} \binom{n_D^{(2)}}{n_C^{(+)} - x_1} Q_{BA}^{x_1}}{\sum_x \binom{n_D^{(1)}}{x} \binom{n_D^{(2)}}{n_C^{(+)} - x} Q_{BA}^x} \quad (13)$$

where $Q_{BA} = (\Pi_C^{(1)} \Pi_D^{(2)}) / (\Pi_D^{(1)} \Pi_C^{(2)}) = G\hat{O}R_{BA}^2$, and the summation for x in the denominator is over the range $\max\{0, n_C^{(+)} - n_D^{(2)}\} \leq x_1 \leq \min\{n_D^{(1)}, n_C^{(+)}\}$.

When $G\hat{O}R_{BA} = 1$ (or $\eta_{BA} = 0$), the noncentral hypergeometric distribution (Eq. 13) reduces to the central hypergeometric distribution:

$$P(n_C^{(1)} = x_1 | n_C^{(+)}, n_D^{(1)}, n_D^{(2)}, Q_{BA} = 1) = \frac{\binom{n_D^{(1)}}{x_1} \binom{n_D^{(2)}}{n_C^{(+)} - x_1}}{\binom{n_D^{(1)} + n_D^{(2)}}{n_C^{(+)}}} \quad (14)$$

where $\max\{0, n_C^{(+)} - n_D^{(2)}\} \leq x_1 \leq \min\{n_D^{(1)}, n_C^{(+)}\}$.

When studying whether there is a difference in effects between treatments A and B, we consider testing $H_0: G\hat{O}R_{BA} = 1$ (or $\eta_{BA} = 0$) vs. $H_a: G\hat{O}R_{BA} \neq 1$ (or $\eta_{BA} \neq 0$). As both $n_C^{(g)}$ and $n_D^{(g)}$ are large, we will reject the null

hypothesis $H_0: G\hat{O}R_{BA} = 1$ on the basis of $G\hat{O}R_{BA}$ (Eq. 11) and $\text{Var}(\log(G\hat{O}R_{BA}))$ (Eq. 12) at the α -level^[7] if:

$$\frac{|\log(G\hat{O}R_{BA})|}{\sqrt{\text{Var}(\log(G\hat{O}R_{BA}))}} > Z_{\alpha/2} \quad (15)$$

where Z_{α} is the upper 100(α)th percentile of the standard normal distribution.

When $n_C^{(g)}$ or $n_D^{(g)}$ is small, we may consider use of Fisher's exact test on the basis of the exact conditional distribution (Eq. 14). We will reject $H_0: G\hat{O}R_{BA} = 1$ at the α -level if we observe $n_C^{(1)} = x_1$ such that:^[7]

$$\min \left\{ P(x \geq x_1 | n_C^{(+)}, n_D^{(1)}, n_D^{(2)}, Q_{BA} = 1), P(x \leq x_1 | n_C^{(+)}, n_D^{(1)}, n_D^{(2)}, Q_{BA} = 1) \right\} \leq \frac{\alpha}{2} \quad (16)$$

where, $P(x \leq x_1 | n_C^{(+)}, n_D^{(1)}, n_D^{(2)}, Q_{BA} = 1) = \sum_{x \geq x_1} P(n_C^{(1)} = x | n_C^{(+)}, n_D^{(1)}, n_D^{(2)}, Q_{BA} = 1)$, and $P(x \geq x_1 | n_C^{(+)}, n_D^{(1)}, n_D^{(2)}, Q_{BA} = 1) = \sum_{x \leq x_1} P(n_C^{(1)} = x | n_C^{(+)}, n_D^{(1)}, n_D^{(2)}, Q_{BA} = 1)$.

Note that the magnitude of a patient change in the ordered category from $j = 1$ (at period 1) to $j = 3$ (at period 2) is larger than that from $j = 1$ (at period 1) to $j = 2$ (at period 2). However, it is not clear whether the magnitude of a patient change from category $j = 1$ (at period 1) to $j = 2$ (at period 2) is equal to that of a patient change from category $j = 2$ (at period 1) to category $j = 3$ (at period 2). When most cell frequencies $n_{jj}^{(g)}$ are reasonably large, we may use the ordinal levels of responses at the first period to form strata and employ a summary test (across strata) with use of Agresti's α accounting for the information on this partial order of patient response changes to improve power.^[8,9] Note that we base posttreatment responses at the first period to form strata, and hence the proposed summary test procedure accounting for the partial order based on Agresti's α can be appropriate only for testing equality of treatments. This is because one can reasonably assume that the patients with the same ordinal level of responses at the first period are comparable between groups under the null hypothesis of equal treatment effects. It can be generally inadequate to do stratified analysis if we use the posttreatment responses to form strata, especially when these patient responses are affected by the treatments in data analysis.

When wishing to produce an interval estimator for $G\hat{O}R_{BA}$, we can obtain an asymptotic 100(1 - α)% confidence interval for $G\hat{O}R_{BA}$ as:^[7]

$$\left[G\hat{O}R_{BA} \exp(-Z_{\alpha/2} \sqrt{\text{Var}(\log(G\hat{O}R_{BA}))}), G\hat{O}R_{BA} \exp(Z_{\alpha/2} \sqrt{\text{Var}(\log(G\hat{O}R_{BA}))}) \right] \quad (17)$$

When $n_C^{(g)}$ or $n_D^{(g)}$ is small, interval estimator (Eq. 17) may lose accuracy. In this case, we may consider use of the

interval estimator on the basis of the exact conditional distribution (Eq. 13). Given $n_C^{(1)} = x_1$, we obtain an exact $100(1 - \alpha)\%$ confidence interval for GOR_{BA} as:^[7]

$$[GOR_{BAL}, GOR_{BAU}] \quad (18)$$

where GOR_{BAL} and GOR_{BAU} are found by solving the following two equations:

$$P(x \geq x_1 | n_C^{(+)}, n_{dis}^{(1)}, n_{dis}^{(2)}, GOR_{BAL}^2) = \frac{\alpha}{2} \text{ and}$$

$$P(x \leq x_1 | n_C^{(+)}, n_{dis}^{(1)}, n_{dis}^{(2)}, GOR_{BAU}^2) = \frac{\alpha}{2}$$

An Example

Consider the crossover trial^[1] that was conducted by 3M-Riker to compare the suitability of two new inhalation devices (A and B) in asthma patients who were using a standard inhaler device delivering Salbutamol. Group 1 used device A for one week followed by device B for another week, while group 2 used these devices in reverse order. No wash-out period was felt necessary. Patients gave their assessment on the clarity of leaflet instructions accompanying the devices on an ordinal scale: “easy,” “only clear after reading,” “not very clear,” and “confusing.” Note that in this example, the higher the ordinal category, the worse is the clarity of the leaflet instruction. When assuming the proportional odds model (Eq. 2) with normal random effects and applying PROC GLIMMIX,^[15] we obtain the MLE and its estimated SE as $\beta_{BA} = -1.283$ and $SE = 0.213$. This gives the p-value < 0.0001 . Thus, there is strong evidence that the clarities of leaflet instructions between devices A and B are different. Furthermore, because $\beta_{BA} < 0$, the leaflet instruction for device B is less clear than that for device A.

When employing test procedures (Eq. 15) and (Eq. 16) to test $H_0: GOR_{BA} = 1$, we obtain both the p-values to be less than 0.001. These results strongly suggest that the clarity of leaflet instruction should be different between devices A and B as well. To estimate the GOR of patient opinions regarding the clarity of leaflet instruction between devices A and B, we obtain $\hat{GOR}_{BA} = 3.637$. This estimate suggests that approximately 3.6 times of the probability that a randomly selected patient felt more unclear of the leaflet instruction for device B than device A vs. that a randomly selected patient felt more unclear of the leaflet instruction for device A than device B. When employing interval estimators (Eq. 17) and (Eq. 18), we obtain 95% confidence intervals to be [2.356, 5.615] and [2.267, 5.891], respectively. Because both the lower limits of these resulting interval estimates fall above 1, there is significant evidence that the clarity of leaflet instructions for device B is worse than that of device A at the 5% level. This finding is again consistent with that found on the basis of the proportional odds model with assuming normal random effects.

THREE-TREATMENT THREE-PERIOD CROSSOVER DESIGN

Suppose that we compare two experimental treatments B and C with a standard treatment (or placebo) A in a three-period crossover design. We let X-Y-Z denote the treatment-receipt sequence of receiving treatments X, Y, and Z at periods 1, 2, and 3, respectively. Say, we randomly assign n_g patients to group g, where $g = 1$ denotes the group with A-B-C treatment-receipt sequence; $g = 2$ denotes the group with A-C-B treatment-receipt sequence; $g = 3$ denotes the group with B-A-C treatment-receipt sequence; $g = 4$ denotes the group with B-C-A treatment-receipt sequence; $g = 5$ denotes the group with C-A-B treatment-receipt sequence; and $g = 6$ denotes the group with C-B-A treatment-receipt sequence. For patient i ($= 1, 2, \dots, n_g$) assigned to group g ($= 1, 2, 3, 4, 5, 6$), we let $Y_{iz}^{(g)}$ denote the ordinal outcome of the patient at period z ($= 1, 2, 3$), and $Y_{iz}^{(g)} = j$ when the patient response falls in the j th ordered category. We let $X_{iz1}^{(g)}$ and $X_{iz2}^{(g)}$ denote the indicator functions of treatment-receipt for treatments B and C, respectively, setting $X_{iz1}^{(g)} = 1$ for the patient receiving treatment B at period z , and $= 0$, otherwise; and $X_{iz2}^{(g)} = 1$ for the patient receiving treatment C at period z , and $= 0$, otherwise. Furthermore, we define $Z_{iz1}^{(g)}$ and $Z_{iz2}^{(g)}$ as the indicator functions for periods 2 and 3.

Normal Random Effects Proportional Odds Model

When using the normal random effects proportional odds model, we can extend model (Eq. 2) to accommodate a three-treatment three-period crossover design by assuming that

$$P(Y_{iz}^{(g)} \leq j | \mu_i^{(g)}) = \frac{\exp(\alpha_j + \beta_{BA} X_{iz1}^{(g)} + \beta_{CA} X_{iz2}^{(g)} + \gamma_{21} Z_{iz1}^{(g)} + \gamma_{31} Z_{iz2}^{(g)} + \mu_i^{(g)})}{1 + \exp(\alpha_j + \beta_{BA} X_{iz1}^{(g)} + \beta_{CA} X_{iz2}^{(g)} + \gamma_{21} Z_{iz1}^{(g)} + \gamma_{31} Z_{iz2}^{(g)} + \mu_i^{(g)})} \quad (19)$$

where $\mu_i^{(g)}$'s denote the random effects due to patients and are assumed to be independent and identically distributed as a normal distribution with mean 0 and unknown variance σ^2 , β_{BA} and β_{CA} denote the relative effects of treatments B and C vs. treatment A, as well as γ_{21} and γ_{31} denote the relative effects of periods 2 and 3 vs. period 1. Note that we can measure the relative effect of treatment C to treatment B by $\beta_{CB} (= \beta_{CA} - \beta_{BA})$. Following similar ideas and arguments as those for equations (3)–(5), we can derive the corresponding likelihood and employ PROC GLIMMIX to obtain the MLEs and confidence intervals, as well as do hypothesis testing based on Wald's or the likelihood ratio test.

Simple Change Score Method

For brevity, we concentrate our discussion on testing equality of treatments B and A. All the following discussions can be similarly done for testing equality of treatments C and A, as well as for testing equality of treatments C and B.

When comparing treatment B with A, we first form strata according to the two periods in which treatments A and B are received. Thus, we form stratum 1 consisting of groups $g = 1$ and 3; stratum 2 consisting of groups $g = 2$ and 4; and stratum 3 consisting of groups $g = 5$ and 6. Within each stratum, we calculate the difference in the base scores between the two periods (in which treatments A and B are received) within patients and reassign patients into the five-point change score scales according to Senn's score scheme. We can then employ a summary trend test, such as a weighted sum of Wilcoxon rank sum tests across strata,^[3,10] based on these resulting five-point change score scales to test equality of treatments A and B.

Distribution-Free Random Effects Model with Using a Conditional Approach

For a given patient i ($= 1, 2, \dots, n_g$) with $\mu_i^{(g)}$ fixed from group g , we assume that the joint probability of $(Y_{iz_1}^{(g)}, Y_{iz_2}^{(g)})$, where z_1 & $z_2 = 1, 2, 3$ and $z_1 < z_2$, satisfies:^[8,19]

$$P(Y_{iz_1}^{(g)} < Y_{iz_2}^{(g)} | \mu_i^{(g)}) = \frac{1}{1 + \exp(\mu_i^{(g)} + \eta_{BA} X_{iz_1 1}^{(g)} + \eta_{CA} X_{iz_1 2}^{(g)} + \gamma_{21} Z_{iz_1 1}^{(g)} + \gamma_{31} Z_{iz_1 2}^{(g)})} \times \frac{\exp(\mu_i^{(g)} + \eta_{BA} X_{iz_2 1}^{(g)} + \eta_{CA} X_{iz_2 2}^{(g)} + \gamma_{21} Z_{iz_2 1}^{(g)} + \gamma_{31} Z_{iz_2 2}^{(g)})}{1 + \exp(\mu_i^{(g)} + \eta_{BA} X_{iz_2 1}^{(g)} + \eta_{CA} X_{iz_2 2}^{(g)} + \gamma_{21} Z_{iz_2 1}^{(g)} + \gamma_{31} Z_{iz_2 2}^{(g)})}$$

$$P(Y_{iz_1}^{(g)} > Y_{iz_2}^{(g)} | \mu_i^{(g)}) = \frac{1}{1 + \exp(\mu_i^{(g)} + \eta_{BA} X_{iz_2 1}^{(g)} + \eta_{CA} X_{iz_2 2}^{(g)} + \gamma_{21} Z_{iz_2 1}^{(g)} + \gamma_{31} Z_{iz_2 2}^{(g)})} \times \frac{\exp(\mu_i^{(g)} + \eta_{BA} X_{iz_1 1}^{(g)} + \eta_{CA} X_{iz_1 2}^{(g)} + \gamma_{21} Z_{iz_1 1}^{(g)} + \gamma_{31} Z_{iz_1 2}^{(g)})}{1 + \exp(\mu_i^{(g)} + \eta_{BA} X_{iz_1 1}^{(g)} + \eta_{CA} X_{iz_1 2}^{(g)} + \gamma_{21} Z_{iz_1 1}^{(g)} + \gamma_{31} Z_{iz_1 2}^{(g)})}$$

and

$$P(Y_{iz_1}^{(g)} = Y_{iz_2}^{(g)} | \mu_i^{(g)}) = 1 - P(Y_{iz_1}^{(g)} > Y_{iz_2}^{(g)} | \mu_i^{(g)}) - P(Y_{iz_1}^{(g)} < Y_{iz_2}^{(g)} | \mu_i^{(g)}) \quad (20)$$

where η_{BA} and η_{CA} denote the respective effect of treatments B and C relative to treatment A, as well as γ_{21} and γ_{31} denote the respective effect of periods 2 and 3 vs. period 1. The GORs of responses for treatment B vs. treatment A and for treatment C vs. treatment A are defined as $\text{GOR}_{BA} = \exp(\eta_{BA})$ and $\text{GOR}_{CA} = \exp(\eta_{CA})$, respectively. The GOR of responses for treatment C vs. treatment B is equal to $\text{GOR}_{CB} = \exp(\eta_{CA} - \eta_{BA})$. For brevity, we discuss only GOR_{BA} in the following. The discussion on hypothesis testing and interval estimation for GOR_{CA} and GOR_{CB} can be done similarly.

We define $n_C^{(g)}(z_1, z_2)$ as the number of patients with their responses at period z_1 lower than those at period z_2 among

n_g patients in group g . Similarly, we define $n_D^{(g)}(z_1, z_2)$ as the number of patients with their responses at period z_1 higher than those at period z_2 among n_g patients in group g . We define the vector of frequencies $(f_{11k}, f_{12k}, f_{21k}, f_{22k})$ in a 2×2 table for table $k = 1, 2, 3$ as:

$$\begin{aligned} (f_{111} = n_C^{(1)}(1, 2), f_{121} = n_D^{(3)}(1, 2), \\ f_{211} = n_D^{(1)}(1, 2), f_{221} = n_D^{(3)}(1, 2)), \\ (f_{112} = n_C^{(2)}(1, 3), f_{122} = n_C^{(4)}(1, 3), \\ f_{212} = n_D^{(2)}(1, 3), f_{222} = n_D^{(4)}(1, 3)), \\ \text{and } (f_{113} = n_C^{(5)}(2, 3), f_{123} = n_C^{(6)}(2, 3), \\ f_{213} = n_D^{(5)}(2, 3), f_{223} = n_D^{(6)}(2, 3)) \end{aligned} \quad (21)$$

We can show that the underlying common odds ratio (OR) of the three 2×2 tables defined in Eq. 21 is equal to $\text{GOR}_{BA}^{2, [8,19]}$. Thus, one can employ all statistical methods^[14,21] developed for a series of 2×2 tables to study the relative treatment effect GOR_{BA} . For example, consider testing $H_0: \text{GOR}_{BA} = 1$ vs. $H_a: \text{GOR}_{BA} \neq 1$ we may use the weighted-least-squares (WLS) summary test procedure.^[23] We will reject $H_0: \text{GOR}_{BA} = 1$ at the α -level^[8] if:

$$\left(\frac{\sum_{k=1}^3 W_k \text{LGOR}_{BA}^{(k)}}{\sum_{k=1}^3 W_k} \right)^2 \left(\sum_{k=1}^3 W_k \right) > Z_{\alpha/2}^2 \quad (22)$$

where $W_k = 4 / (1/f_{11k} + 1/f_{12k} + 1/f_{21k} + 1/f_{22k})$ and $\text{LGOR}_{BA}^{(k)} = \log((f_{11k} f_{22k}) / (f_{12k} f_{21k}))^{1/2}$, and Z_{α} is the upper 100(α)th percentile of the standard normal distribution. Note that if $f_{ijk} = 0$ for some observed frequencies in a 2×2 table k , we cannot calculate the test statistic (Eq. 22). We may apply the commonly used ad hoc arbitrary adjustment procedure for sparse data by adding 0.50 to each observed frequency f_{ijk} in this particular table k . We may also consider use of the Mantel-Haenszel (MH) summary test procedure^[14,19,24] for testing non-equality of treatments when one fails to employ the test procedure (Eq. 22) due to $f_{ijk} = 0$ for some cells.^[8]

When $n_C^{(g)}(z_1, z_2)$ and $n_D^{(g)}(z_1, z_2)$ are small, both the WLS and MH test procedures are theoretically invalid. In this case, we may consider use of the test procedure based on the exact conditional distribution. Given $f_{11+}^0 (= f_{111}^0 + f_{112}^0 + f_{113}^0)$, we will reject $H_0: \text{GOR}_{BA} = 1$ at the α -level^[8] if:

$$\min \left\{ P(f_{11+} \leq f_{11+}^0 | f_{+1}, f_{+2}, f_{-1}, f_{-2}), \right. \\ \left. P(f_{11+} \geq f_{11+}^0 | f_{+1}, f_{+2}, f_{-1}, f_{-2}) \right\} \leq \frac{\alpha}{2}$$

where $P(f_{11+} \leq f_{11+}^0 | f_{+1}, f_{+2}, f_{-1}, f_{-2})$

$$= \sum_{f_{11+} \leq f_{11+}^0} \prod_{k=1}^3 \frac{\binom{f_{+1k}}{f_{11k}} \binom{f_{+2k}}{f_{1+k} - f_{11k}}}{\binom{f_{++k}}{f_{1+k}}}$$

$$\text{and } P(f_{1+} \geq f_{1+}^0 \mid \underline{f}_{+1}, \underline{f}_{+2}, \underline{f}_{+1+}, \underline{f}_{+2+})$$

$$= \sum_{i_{1+} \geq f_{1+}^0} \prod_{k=1}^3 \frac{\binom{f_{+1k}}{f_{1+k}} \binom{f_{+2k}}{f_{1+k} - f_{1+k}}}{\binom{f_{+k}}{f_{1+k}}} \quad (23)$$

where $a_k \leq f_{1+k} \leq b_k$, $a_k = \max\{0, f_{1+k} - f_{+2k}\}$, $b_k = \min\{f_{+1k}, f_{1+k}\}$ (for $k = 1, 2, 3$), as well as $\underline{f}_{-u+} = (f_{u+1}, f_{u+2}, f_{u+3})'$ for $u = 1, 2$, and $\underline{f}_{-v+} = (f_{+v1}, f_{+v2}, f_{+v3})'$ for $v = 1, 2$.

Also, we can derive asymptotic interval estimators for GOR_{BA} based on WLS and MH method.^[8] Furthermore, because the underlying common OR of the three 2×2 tables (Eq. 21) is equal to GOR_{BA}^2 , we may derive the exact confidence interval for the underlying OR^[14,19,21] based on the product of the noncentral hypergeometric distributions across strata. We then take the square root of these confidence limits to obtain the exact confidence interval for GOR_{BA} .^[8]

An Example

Consider the crossover trial^[25] comparing an analgesic at low (L) and high (H) doses with a placebo (P) for relief of pain in primary dysmenorrhea patients. Here, we regard placebo as treatment A, as well as the low and high doses as treatments B and C. There were 86 patients randomly assigned to one of six groups distinguished by treatment-receipt sequences: P-L-H ($g = 1$), P-H-L ($g = 2$), L-P-H ($g = 3$), L-H-P ($g = 4$), H-P-L ($g = 5$), and H-L-P ($g = 6$). At the end of each treatment period, each patient was assessed the extent of relief on the ordinal scale: none, moderate, and complete. We code these as 1, 2, and 3, respectively. When employing PROC GLIMMIX, we obtain the MLEs (and their estimated SEs) as $\hat{\beta}_{\text{BA}} = -2.041$ (SE = 0.327), $\hat{\beta}_{\text{CA}} = -2.416$ (SE = 0.333), and $\hat{\beta}_{\text{CB}} = -0.374$ (SE = 0.284). On the basis of these estimates (of which the signs are all < 0), we may conclude that both the low and high doses can significantly improve, as compared with the placebo, the relief extent of pain at the 5% level. We also note that although the high dose can improve the outcome of patients using the low dose, this improvement is not significant at the 5% level.

When applying the WLS procedure (Eq. 22) and the exact testing procedure (Eq. 23) to test $H_0: \text{GOR}_{\text{BA}} = 1$ (or $H_0: \eta_{\text{BA}} = 0$), as well as the corresponding procedures to test $H_0: \text{GOR}_{\text{CA}} = 1$ (or $H_0: \eta_{\text{CA}} = 0$), we obtain all p-values < 0.0001 .^[8] These suggest that either the low or high dose of the analgesic vs. the placebo can affect the relief of pain in primary dysmenorrhea. When studying the relative effect of high dose vs. the low dose on the relief extent of pain, we apply the WLS and the exact procedure to test $H_0: \text{GOR}_{\text{CB}} = 1$. We obtain p-values as 0.433 and 0.649, respectively. Thus, there is no significant evidence that we have a difference in the relief extent of pain between the high and low doses. The results and other estimations for the GOR appear elsewhere.^[8,19]

INCOMPLETE BLOCK CROSSEVER DESIGN

When employing a crossover design to compare three or more treatments, the duration of a crossover trial, in which every patient receives each of the treatments under comparison, can be much longer than that of a parallel groups design. This can cause difficulty in recruiting patients into a trial and the burden of ensuring patients to follow the protocol closely. To alleviate these concerns, we may consider employing an incomplete block design in which each patient is to receive only a subset of treatments rather than all treatments under comparison.^[3,26] When using the normal random effects proportional odds model (Eq. 19) excluding the term containing γ_{31} for period 3 vs. period 1, we can still apply PROC GLIMMIX to do hypothesis testing and interval estimators for the relative treatment effect without any extra effort. Using the same set of strata as defined in the previous section, for example, we can easily derive a summary test based on the simple change score method for testing equality of treatments A and B. We can also use similar ideas as those suggested elsewhere^[27,28] and modify test procedures and interval estimators in binary data to accommodate the situations in which the patient responses are in ordinal data under the incomplete block design when using the GOR to measure the relative treatment effect.

FOUR-TREATMENT FOUR-PERIOD CROSSEVER DESIGN

When we employ the crossover design to compare four or more treatments, the number of groups with different treatment-receipt sequences can be too large to be of use. For example, when comparing four treatments over a four-period crossover design, there are 24 ($=4!$) groups with different treatment-receipt sequences. In this case, we may employ a Latin square^[3] to reduce the number of groups with different treatment-receipt sequences. It is straightforward to extend the model (Eq. 19) by including more indicator functions of treatment-received covariates and periods to accounting for a four-treatment four-period crossover design and apply PROC GLIMMIX despite of using a Latin square. Also, all the results based on the WLS method, the MH approach, and the exact conditional distribution can be, as shown elsewhere,^[8] modified to accommodate a four-treatment four-period crossover design under a 4×4 Latin square.

CONCLUSION

We need to assume the random effects in the model (Eq. 1) [or model (Eq. 19)] to be normal in use of PROC GLIMMIX to obtain the MLEs, but the normal random effects proportional odds model is ready to use for various

crossover designs. The parameters in this model are meaningful and easily understood. These parameters can also provide a summary measure to quantify the relative treatment effect. Furthermore, when the normal random effects proportional odds model assumptions are satisfied, the model-based test procedures and estimators are more efficient than the other approaches. When the number of patients in a trial is small, however, the MLEs and their relevant test procedures can be invalid in sparse data.^[29] By contrast, use of the simple change score method does not need to assume a parametric model for the underlying data and the calculation of this method is tractable. Also, since the change score method collapses ordinal data into a few categories, the use of change score test procedure can be subject to less concern of sparse data than Wald's or the likelihood tests. However, reassigning patients according to the difference in the base scores to categories on a change score scale seems to be arbitrary. It can also be difficult to provide an easily interpreted summary measure of the relative treatment effect based on the change score method. Using the GOR to measure the relative treatment effect and the conditional approach, we can avoid assuming normality for random effects. The GOR of responses can be easily appreciated and interpreted. We can derive the exact test procedure and interval estimator for the GOR in small sample cases. Furthermore, all asymptotic test procedures and estimators based on the conditional approach can be expressed in closed forms and be calculated by a hand calculator. Since it excludes the information on patients with identical levels of responses between periods, however, the use of the conditional approach with the GOR can lose efficiency. Finally, we note that test procedures and estimators using the above three approaches can all be easily modified to study the period effect.

REFERENCES

- Ezzet, F.; Whitehead, J. A random effects model for ordinal responses from a crossover trial. *Stat. Med.* **1991**, *10*, 901–907.
- Jones, B.; Kenward, M.G. *Design and Analysis of Cross-Over Trials*; 3rd Ed.; Chapman & Hall/CRC, Taylor and Francis: Boca Raton, FL, 2014.
- Senn, S.J. *Cross-Over Trials in Clinical Research*; 2nd Ed.; Wiley: Chichester, 2002.
- Senn, S.J. Cross-over trials in statistics in medicine: The first “25” years. *Stat. Med.* **2006**, *25* (20), 3430–3442.
- Hills, M.; Armitage, P. The two-period cross-over clinical trial. *Br. J. Clin. Pharmacol.* **1979**, *8* (1), 7–20.
- Senn, S.J. A random effects model for ordinal responses from a crossover trial. *Stat. Med. (Letters to the Editor)* **1993**, *12*, 2147–2150.
- Lui, K.-J.; Chang, K.-C. Hypothesis testing and estimation in ordinal data under a simple crossover design. *J. Biopharm. Stat.* **2012**, *22*, 1137–1147.
- Lui, K.-J.; Chang, K.-C.; Lin, C.-D. Testing equality and interval estimation of the generalized odds ratio in ordinal data under a three-period crossover design. *Stat. Meth. Med. Res.* **2015**. DOI:10.1177/0962280215569623.
- Agresti, A. Generalized odds ratios for ordinal data. *Biometrics* **1980**, *36* (1), 59–67.
- Koch, G.G. The use of non-parametric methods in the statistical analysis of the two-period change-over design. *Biometrics* **1972**, *28* (2), 577–584.
- Fleiss, J.L. *The Design and Analysis of Clinical Experiments*; Wiley: New York, 1986.
- Fleiss, J.L. A critique of recent research on the two-treatment crossover design. *Control. Clin. Trials* **1989**, *10* (3), 237–243.
- Clayton, D.; Cuzick, J. Multivariate generalizations of the proportional hazards model. *J. Royal Stat. Soc., Series A.* **1985**, *148* (2), 82–117.
- Agresti, A. *Categorical Data Analysis*; Wiley: New York, 1990.
- SAS Institute Inc. *SAS/STAT 9.2 User's Guide*; 2nd Ed.; SAS Institute: Cary, NC, 2009.
- Fleiss, J.L. On the asserted invariance of the odds ratio. *Br. J. Preven. Social Med.* **1970**, *24* (1), 45–46.
- Mosteller, F. Association and estimation in contingency tables. *J. Am. Stat. Assoc.* **1968**, *63*, 1–28.
- Ezzet, F.; Whitehead, J. A random effects model for ordinal responses from a crossover clinical trials. *Stat. Med. (Authors' Reply)* **1993**, *12*, 2150–2151.
- Lui, K.-J. *Crossover Designs Testing, Estimation and Sample Size*; Wiley: Chichester, 2016.
- Lui, K.-J. Notes on estimation of the general odds ratio and the general risk difference for paired-sample data. *Biometrical J.* **2002**, *44*, 957–968.
- Lui, K.-J. *Statistical Estimation of Epidemiological Risk*; Wiley: New York, 2004.
- Gart, J.J.; Thomas, D.G. Numerical results on approximate confidence limits for the odds ratio. *J. Royal Stat. Soc. B.* **1972**, *34*, 441–447.
- Fleiss, J.L. *Statistical Methods for Rates and Proportions*; 2nd Ed.; Wiley: New York, 1981.
- Robins, J.; Breslow, N.; Greenland, S. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* **1986**, *42* (2), 311–323.
- Kenward, M.G.; Jones, B. The analysis of categorical data from cross-over trials using a latent variable model. *Stat. Med.* **1991**, *10*, 1607–1619.
- Koch, G.G.; Amara, I.A.; Brown, B.W., Jr.; Colton, T; Gillings, D.B. A two-period crossover design for the comparison of two active treatments and placebo. *Stat. Med.* **1989**, *8*, 487–504.
- Lui, K.-J. Estimation of the treatment effect under an incomplete block crossover design in binary data—A conditional likelihood approach. *Stat. Meth. Med. Res.* **2015**. DOI: 10.1177/0962280215595434.
- Lui, K.-J.; Chang, K.-C. Exact tests in binary data under an incomplete block crossover design. *Stat. Meth. Med. Res.* **2016**. DOI: 10.1177/0962280216638382.
- Breslow, N.E.; Day, N.E. *Statistical Methods in Cancer Research, Volume 1. The Analysis of Case-Control Studies*; International Agency for Research on Cancer: Lyon, 1980.

Parallel Design

Mani Y. Lakshminarayanan

Biostatistics and Research Decision Sciences, Merck & Co, Inc., Rahway, New Jersey, U.S.A.

Abstract

Parallel design is probably the simplest and the most commonly used design employed in clinical trials for evaluation of the effectiveness and safety of an experimental drug against a control. This entry describes the basic principles of parallel design and provides multiple examples of various types of parallel design.

Keywords: Analysis of covariance; Endpoints; Multi-center trials; Multiple comparisons; Protocols; Randomized block design; Sample size.

INTRODUCTION

Statisticians involved in clinical research in the pharmaceutical industry encounter design of experiments almost on a daily basis while writing protocols for their clinical studies. At the outset, the primary objective(s) described in the protocols involve comparing different dose groups; to accomplish this, patients are assigned randomly to the dose groups. This general idea of assigning patients to different groups is similar to those agricultural experiments conducted by Sir Ronald Fisher, who is considered to be the father of many fundamental statistical concepts, including design of experiments, at the Rothamsted Agricultural Experiment Station near London. Sir Austin Bradford Hill was primarily instrumental in the development of controlled clinical trials in Britain. His publications in the early 1950s^[1] included fundamental concepts such as random allocation, concurrent controls, eligibility of patients, treatment schedule, and statistical analysis of the clinical data.

OVERVIEW

The term *experimental design* comes from these agricultural experiments, where the experimenters sow varieties in or apply treatments to plots, and often these plots were arranged in blocks. At this time, readers are asked to take note of terms like *randomly* and *blocks*. Variations of these experiments are carried out in clinical trials when patients are randomly allocated across various treatments within each investigational site, also known as a *center*. These basic ideas will be discussed further in this entry.

As an example, imagine a clinical trial that is planned to compare the effectiveness of an experimental drug against placebo in the treatment of patients with schizophrenia, or schizoaffective disorder. In order to accomplish

the primary objective, it is decided to include more than one dose of the experimental drug. As we think about this clinical trial, a number of important questions arise:

1. Are these dose groups different in efficacy?
2. Are there other characteristics of the patients that might affect the efficacy that should be controlled or investigated?
3. How many patients should be assigned to each dose group?
4. How should the patients be assigned to each dose group, and how the data should be collected?
5. What analysis method should be used on the collected data?
6. If there are differences between the dose groups, which ones are different?

It is important to answer all of these questions before concluding that a given trial is positive or not. Some of these issues will have to be specified a priori in the protocol, as required by the International Conference on Harmonization (ICH) Guidelines (ICH E-9);^[2] readers are referred to the entry titled *Protocol Development* for more details on the protocol specification.

Although the final conclusions in any clinical trial may depend largely on the conduct of the trial itself, it is absolutely crucial that the protocol team (including clinicians and statisticians) spend enough time on the design part of the trial. This entry is intended to discuss one of the commonly used designs in clinical trials.

DEFINITION

The simplest and the most commonly used designs employed in clinical trials are *parallel designs*. In a trial

with parallel designs, a patient is assigned randomly to one and only one treatment, as opposed to crossover designs, in which the same patient is given two or more treatments within a trial. Differences between these two types of designs will not be discussed further in this entry.

In parallel designs, there are usually at least two treatment groups: experimental and control. In our earlier example of a study involving schizophrenia, there are several dose groups of the experimental drug and a placebo as the concurrent control. Controls in clinical trials are used as standard treatments to compare and to evaluate the effectiveness of an experimental drug. When there are no standard marketed treatments available, and if it is ethical, then placebo controls are used as standard comparators. If a standard therapy is available (e.g., anti-infectives or antihypertensives), then it may be required to use it as a control to evaluate quantitatively the efficacy and safety of an experimental drug; thus a relative risk-benefit can be specified. The importance of proper controls as a part of well-designed clinical trials is stressed in ICH Guidelines (ICH E-9), especially in their requirements for sponsors to provide at least two well-controlled trials as pivotal for the new drug approvals (NDAs). There are other types of control, such as no control and historical control, that are used less frequently in clinical trials.

It is noted in this simplest definition of parallel designs, where we simply assign patients to one and only one treatment, that no specific patient characteristics that may have a bearing on the response are taken into consideration. These patient characteristics or other trial-related characteristics are typically known as *prognostic variables*. Designs that incorporate these prognostic variables (e.g., randomized block designs, analysis of covariance) will be discussed later in this entry.

THREE BASIC CONCEPTS IN EXPERIMENTAL DESIGNS

The three basic principles that serve as keystones to the development of experimental designs are: *randomization*, *replication*, and *blocking*.

There is enough literature discussing the need for using randomization in clinical trials before treatments can be compared, but there are also references in the literature that argue against the use of randomization. It was the work of pioneers such as Sir R.A. Fisher that influenced the use of randomization as the cornerstone application of statistical theory in experimental design. A classic paper by Peto et al.^[3] presents a convincing argument for using randomization in clinical trials.

In a trial with multiple doses of the experimental drug and a placebo (such as the trial described in the previous section on patients with schizophrenia), patients are assigned to each group at random using a predetermined scheme. These randomization schemes can always be

generated using computer algorithms. In general, there are several benefits associated with using random assignment of patients to treatments: avoidance of any bias in the results, balancing of prognostic factors, ensurance of proper conduct of trials, and validation of assumptions needed to apply certain statistical techniques.

The use of *replication*, as emphasized by Fisher, is critical to reduce variability or to get a proper estimate of the error variance. In general, *replication* means repetition of an entire experiment, or portions thereof, under similar experimental conditions. Although similar experimental conditions are desirable, it is not useful to obtain identical analytical results. To accomplish this, experiments can be repeated under two or more sets of conditions, for example, within a time window that is considered in a trial. For example, analysis of clinical trials data where measurements are collected across time points (e.g., blood pressure measurements in trials involving patients with hypertension) involves “change from baseline to final time point” as the response variable. In this case, a time window can be created for both baseline and final time point so that measurements can be repeated during the time window (e.g., three blood pressure measurements can be taken both at baseline and at final). Replicated measurements allow the experimenter to estimate accurately the experimental error or to estimate precisely the error variance if a sample mean is used in the estimation of a factor effect.

Blocking is used as a technique for increasing the precision of an experiment. A *block* is a subset of experimental units that are more homogeneous than the entire experiment itself. For example, in many clinical trials, responses to treatments are different for various groups of patients based on their age, sex, race, or initial severity of disease, which are some commonly cited prognostic variables that may have an impact on the response. Blocking (also known as *matching* in some literature) is a technique used to control the variability of the response caused by these prognostic variables. Randomized block designs are used to take blocks into account in the design. An example of a randomized block design is a paired comparison design, in which two patients who have similar characteristics are paired and the two treatments (drug and placebo in a two-group trial) are assigned to each member of the pair.

A sensible alternative to blocking as a method to control for the prognostic variables is *stratification*. A common stratification variable is center. Other prognostic variables, such as age, sex, race, and initial severity, can also be considered as stratification variables. For example, age could be grouped into, say, $\text{age} < 18$, $18 \leq \text{age} \leq 64$, and $\text{age} \geq 65$ as different strata. With this definition of strata for age, it is noted that patients who are homogeneous based on their age category are grouped. With regard to randomization, patients would then be randomized within each stratum separately. In clinical trials, it is very common to produce

an independent set of randomization schedules for each investigational center.

Blocking and stratification are quite often used within the same data. For example, within each stratum, randomizations could be carried out using simple random samples, or blocking could be used in the randomizations. Thus, blocked randomization is a special kind of stratification. The smallest block size is the sum of the integers defined by the allocation ratio (for a 1:1 allocation, block size is 2). Use of a large number of strata may increase the chance of departure from the specified allocation ratio, unless small block sizes are used within each stratum.

EXAMPLES OF PARALLEL DESIGNS

Completely Randomized Design

The simplest version of a completely randomized design is a one-way classification model, which is a single-factor experiment with two or more levels of a design factor, such as clinical trial designed to compare several doses of an experimental drug and a placebo where all patient data are obtained from one clinic or hospital. Mathematically, this model can be written as

$$y_{ij} = \mu_i + \varepsilon_{ij} \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, n_i \end{cases} \quad (1)$$

where y_{ij} is the (ij) th observation, a is the number of treatments (number of doses and placebo), and n_i is the number of observations in the i th treatment.

For the underlying model presented in Eq. 1, our interest is to test for the equality of the treatment means; that

is, $H_0: \mu_1 = \mu_2 = \dots = \mu_a$ If the null hypothesis is true, then it is well known that the ratio $F = MSA/MSE$ is distributed as F with $(a - 1)$ and $(N - a)$ degrees of freedom. Definitions of the terms used are as follows:

$$MSA = \frac{SSA}{a - 1} \quad MSE = \frac{SSE}{N - a} \quad (2)$$

$$SSA = \sum_i n_i (\bar{y}_i - \bar{y}_{..})^2 \quad SSE = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$$

$$N = \sum_i n_i \bar{y}_i = \sum_j \frac{y_{ij}}{n_i} \bar{y}_{..} = \sum_i \sum_j \frac{y_{ij}}{N}$$

Although these mathematical formulations would guide us in determining whether there is any signal for a difference between the treatment means, a similar analysis in clinical trials would be considered incomplete until the next step is taken, viz., responding to the primary objective specified in the protocol of showing whether or not the experimental therapy is efficacious. In other words, it is important for us to know which treatments differ from what and the

magnitude of those differences. In order to answer this primary objective in a protocol, the statistical analysis will have to involve the method of *multiple comparisons*, which will be briefly described at the end of this entry and will be covered in detail in another entry.

Designs with Prognostic Factors

One of the most widely used designs for controlling known sources of variability due to the prognostic factors is a *randomized block design*. In the case of a large phase II/III clinical trial, the study team may be required to carry out this trial across several investigational centers. Studies like this are known as *multicenter trials*. Some of the important reasons and rationale for conducting a multicenter trial include the ability to generalize the results across different set of patients and treatment settings, and expediting the enrollment of patients in a large trial. If center is considered as a stratification variable, then individual randomizations are carried out within each center. A balanced treatment assignment within each center is generally desired.

With treatment and center (as a blocking variable) in a randomized block design with complete blocks (that is, one observation per treatment in each center), the statistical model can be written as

$$y_{ij} = \mu_i + \beta_j + \varepsilon_{ij} \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, b \end{cases} \quad (3)$$

The model presented in Eq. 3 has no term included for a possible interaction between the treatments and centers. An interaction between treatment and center may exist if the directions of the treatment difference are different across centers. Because most of the analysis in a multicenter trial is based on pooled data from various centers, guidelines from most regulatory agencies suggest that a preliminary test for the treatment by center be carried out (ICH E-9 Guidelines). Thus, it is almost standard practice to consider the following revised model to Eq. 2:

$$y_{ij} = \mu_i + \beta_j + \mu\beta_{ij} + \varepsilon_{ij} \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, b \end{cases} \quad (4)$$

The interaction model in Eq. 4 is also popularly known as *cell means model*.

Hypotheses involving the equality of treatment means (μ_i), center means (β_j), and existence of interaction ($\mu\beta_{ij}$) can be tested using various sums of squares terms, similar to the ones shown in Eq. 2, once the analysis of variance table is set up.^[4] It is common to test for the interaction first. If this test is nonsignificant, it is often taken to be a sufficient (but not necessary) justification for pooling data across centers and for concluding that the treatment differences are sufficiently consistent. If a significant interaction

is identified, it is expected that the sources of the interaction, such as dropout rates and eligibility criteria, are explored and explained in the final analysis. Although it is partly based on one's judgment and what is prespecified in the statistical analysis plan in the protocol, the primary hypothesis is usually based on the two main-effects terms.

The *analysis of covariance* is another design that can be used to control the effects of prognostic variables. Imagine an experiment with a response variable y and with another variable x , where y is linearly related to x . Further, x cannot be controlled but can be observed along with y . The variable x is called a *covariate* or a *concomitant* variable. As indicated in the previous section "Three Basic Concepts in Experimental Designs," some examples of covariates are demographic variables such as age, sex, race, and baseline severity of the disease. The use of covariates is restricted to quantitative variables, and it is purely a statistical method for controlling the effects of prognostic variables; no special effort is needed for any prestudy arrangements such as randomization.

For a completely randomized design with a single factor (e.g., treatments) and a covariate (e.g., baseline severity), a statistical model can be written as

$$y_{ij} = \mu + \tau_i + \beta_i(x_{ij} - \bar{x}_{..}) + \varepsilon_{ij} \quad \left\{ \begin{array}{l} i = 1, 2, \dots, a \\ j = 1, 2, \dots, n_i \end{array} \right. \quad (5)$$

Note that the model in Eq. 5 assumes that the regression lines (number = a) have the same slope; that is, the regression lines are parallel. If the regression lines are not parallel, it is reasonable to expect an interaction between the treatments and the covariate. In order to accommodate this, the model in Eq. 5 can be generalized as

$$y_{ij} = \mu + \tau_i + \beta_i(x_{ij} - \bar{x}_{..}) + \varepsilon_{ij} \quad \left\{ \begin{array}{l} i = 1, 2, \dots, a \\ j = 1, 2, \dots, n_i \end{array} \right. \quad (6)$$

For the models specified in Eqs. 5 and 6, analyses for testing various hypotheses can be carried out using the classical ANOVA procedure. A general regression approach can also be developed for this purpose.^[4,5]

Note that in the models referred to so far, an intrinsic assumption is made that the response is measured at one time point, which is not a reality in most clinical trials. Most clinical trials involve measuring responses from a patient across several time periods (e.g., blood pressure measurements taken at several time points in a hypertension trial). Although there are repeated measurements made on subjects, one of the models already explained is used at each time point as a standard analysis. A most widely used time point is the last observation carried forward (LOCF), with the response considered as the change from baseline to this time point.

In repeated-measurements designs, there are two sizes of experimental units, subject and subject's time. Thus,

there are two parts in the model and the analysis, comparison of treatments based on subject's response and comparison of time and treatment by time interaction. Although the structure and the model considered when there are repeated measurements are similar to those of a split-plot design, the two designs operate under different assumptions. One of the main differences is that the subjects are randomly assigned to the levels of treatment, but the time intervals are not. The experimental design used for this type of data is known as *repeated-measures design*. For the analysis of such a design, one can use either a classical ANOVA approach or a multivariate approach.^[4]

ANALYSIS ISSUES

Baseline Information

Measurements are often collected longitudinally in parallel group designs in order to study the time effect and long-term safety data. In such cases, baseline measurements can be considered as valuable covariates in determining the treatment effect. Comparison of treatments are typically studied in terms of mean change from baseline using an Analysis of Covariance (ANCOVA) model with baseline value as a covariate and either the post-baseline value or the calculated change from baseline value as the dependent variable. Estimates and inference based on an ANCOVA model conditional on baseline measurements are unbiased under normality assumption in Ref. [6]. When the baseline measurement is correlated with measurements that are observed post-baseline, any adjustment made for baseline using ANCOVA has been shown to remove the conditional bias in treatment group comparisons due to chance imbalances and to improve efficiency over unadjusted comparisons in Ref. [7].

When several post-randomization measurements (repeated measures over time) are collected in a trial, a longitudinal data analysis (LDA) model can be used. In such a model, the change from baseline is calculated at each post-randomization time point and the baseline measurement is included as a covariate in the model. This model is referred to as the longitudinal ANCOVA model. An alternative approach, a full likelihood approach, initially proposed by Liang and Zeger^[8] can also be suggested in this scenario. This full likelihood LDA model includes baseline values as dependent variable in the response vector along with the post-baseline values. It is worth pointing out that the baseline mean responses for each treatment group are implicitly assumed equal in the longitudinal ANCOVA model and this assumption is a reasonable one due to randomization. A similar "constraint" can also be imposed in the full likelihood LDA model, which can be called as the constrained full likelihood LDA model (cLDA).

There are several advantages of the constrained LDA model over the longitudinal ANCOVA model: it provides

more efficient between group comparison with unbiased standard errors for within-group treatment effects, coverage at the appropriate $100(1 - \alpha)\%$ level under normality, includes baseline measurements of subjects who have missing post-baseline measurements, provides flexibility for stratification and does not make restrictive assumption regarding correlation between baseline and post-baseline measurements.

When SAS PROC MIXED procedure is used to fit a longitudinal ANCOVA model with repeated measures and missing data, the default LSMEANS estimates may not be appropriate. To fix this problem, the baseline mean should be calculated outside the SAS PROC MIXED procedure, using one baseline observation per subject. This calculated mean can be used in the LSMEANS statement with the AT= option or in the ESTIMATE statement. When the model is more complicated (e.g., adjustment for other factors), then the ESTIMATE statement provides more flexibility.

Multiple Comparison

In the “Completely Randomized Design” section, while discussing the test of significance for the equality of treatment means, it was indicated that a significant overall F -test would signal a difference in treatment means. In most clinical trials, showing the existence of an overall experimental effect is of little interest to the primary objective. In a trial involving several doses of an experimental drug and a placebo, one’s primary interest is to see how many doses are different from placebo, i.e., pairwise comparisons between placebo and all doses. Methods that are used to find these pairwise differences are known as *multiple comparisons*. Multiple comparison procedures are employed with a primary goal of controlling either a family-wise error rate (FWER) or more recently false discovery rate (FDR). Preference of choosing FDR over FWER for controlling multiplicity depends on the experimental hypotheses. A great number of procedures are available in the literature in this regard. A detailed discussion of these methods and their rationale will be covered in another entry.

Multiple Time Points

As discussed in “Designs with Prognostic Factors,” if we repeatedly measure patient’s responses across various time points and perform an analysis at each time point, then it can create a multiplicity issue. A repeated-measures design with follow-up multiple comparisons^[4] can be used to handle this issue. As explained in the previous section, a common and effective method of handling multiple time points is to express the repeated data in terms of a single time point, for example, LOCF time point. This time point is often considered the primary time point for analyses of the primary endpoints. Other useful approaches include using the slope for each patient across time or the area under the curve (AUC) as primary endpoints. Diggle et al.^[9] discuss

other approaches that are useful for analyzing data with repeated measures some of which have described in the section “Baseline Information” above.

Multiple Endpoints

It is very common that more than one endpoint (measurement) is specified as primary in phase II/III protocols. For example, in trials designed to study the efficacy and safety of an antipsychotic drug on patients with schizophrenia, one or more derived measurements from the Positive and Negative Symptoms Scale (PANSS) and the Clinical Global Assessment scales are considered primary. For these different endpoints, a separate analysis is carried out and p -values are reported separately. A problem with this type of approach is when to declare the drug efficacious. For instance, if there are five primary endpoints and only three of them show statistical significance, can we still claim efficacy? To avoid this, some protocols tend to specify only one endpoint as primary and the rest as secondary. In the analysis section, an attempt will then be made to explore the consistency of various results. Other scenarios include: declaring a set of primary endpoints as “co-primary endpoints,” multiple primary endpoints in multiple treatment arms, and clinically ordered primary endpoints. In the case of “co-primary endpoints,” no adjustment for multiple endpoints is necessary as it can be shown that the overall FWER is at most equal to the nominal α -level; see Sankoh, D’Agostino and Huque^[10] for more discussion on co-primary endpoints and some general strategies for dealing with multiple co-primary endpoints.

One approach is to consider a composite measure that can be constructed based on the primary endpoints.^[11] Although p -values are derivable for these composite measures, they are not as easy to interpret as changes or percentage changes. Multivariate approaches provide alternatives for handling multiple endpoints.

Combination Treatments

Combination treatments are very common in cancer trials. Traditionally, for combination therapies, a Phase I trial for toxicity and a separate Phase II trial for efficacy are conducted. Huang et al.^[12] have proposed a method of integrating these two phases into one. Their approach involve using a Phase I/II clinical trial to evaluate simultaneously the safety and efficacy of combination dose levels, and then select the optimal combination dose. The underlying approach involves using Bayesian posterior probabilities in the randomization to adaptively assign more patients to doses with higher efficacy levels. Combination doses with lower efficacy are temporarily closed and those with intolerable toxicity are eliminated from the trial. The trial is stopped if the posterior probability for safety, efficacy, or futility crosses a prespecified boundary.

Sample Size

If there are two groups in the design, then sample size determination is straightforward. For normally distributed variables with a common variance σ^2 , the sample size formula for each treatment group is given as $n = 2\sigma^2(z_{\alpha/2} + z_\beta)^2/(\mu_1 - \mu_2)^2$ where μ_1 and μ_2 are the true means, probability of type I error is α , and the power of the test is $1 - \beta$.

The determination of sample sizes for clinical studies with several groups (more than two) involves critical values from a *noncentral F-distribution*.^[13,14] Statisticians faced with this problem in clinical trials tend to use a practical solution; viz., they convert this into a two-group problem, where the groups chosen have the smallest difference. For this smallest difference ($\mu_1 - \mu_2$), the sample size is chosen and is considered for the entire protocol. Other issues, such as adjusting the significance level for multiple comparisons and dose-response designs, are also considered for determining sample size.

CONCLUSION

Parallel design is probably the simplest and the most commonly used design employed in clinical trials for evaluation of the effectiveness and safety of an experimental drug against a control (e.g., a placebo control or an active control). It is suggested that statistical analysis be performed in compliance with statistical principles as specified in the ICH E-9 guideline for good clinical practice.

REFERENCES

1. Hill, A.B. *Statistical Methods in Clinical and Preventive Medicine*; Livingstone: Edinburgh, 1962.
2. *International Conference on Harmonization (ICH) Guidelines; E-9: Statistical Principles for Clinical Trials*, 1997.
3. Peto, R. et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient: I. Introduction and design. *Br. J. Cancer* **1997**, *34*, 585–612.
4. Fleiss, J.L. *The Design and Analysis of Clinical Experiments*; Wiley: New York, 1986.
5. Mason, R.L.; Gunst, R.F.; Hess, J.L. *Statistical Design and Analysis of Experiments with Applications to Engineering and Science*; Wiley: New York, 1989.
6. Crager, R.M. Analysis of covariance in parallel-group clinical trials with pretreatment baselines. *Biometrics* **1987**, *43*, 895–901.
7. Frison, L.; Pocock, S.J. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Stat. Med.* **1992**, *11*, 1685–1704.
8. Liang, K.; Zeger, S. Longitudinal data analysis of continuous and discrete responses for pre-post designs. *Sankhya Indian J. Stat.* **2000**, *62* (Series B), 134–148.
9. Diggle, P.J.; Liang, K.-Y.; Zeger, S.L. *Analysis of Longitudinal Data*; Oxford University Press: Oxford, 1994.
10. Sankoh, A.J.; D'Agostino, R.B.; Huque, M.F. Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Stat. Med.* **2003**, *22*, 3133–3150.
11. O'Brien, P.C. The appropriateness of analysis of variance and multiple comparison procedures. *Biometrics* **1983**, *39*, 787–788.
12. Huang, X.; Biswas, S.; Oki, Y.; Issa, J.-P.; Berry, D.A. A parallel Phase I/II clinical trial design for combination therapies. *Biometrics* **2007**, *63*, 429–436.
13. Winer, B.J. *Statistical Principles in Experimental Design*; McGraw-Hill: New York, 1962.
14. Chow, S.C.; Shao, J.; Wang, H. *Sample Size Calculation in Clinical Research*; Marcel Dekker, Inc.: New York, 2003.

Patient Compliance

Kenneth B. Schechtman

Washington University School of Medicine, St. Louis, Missouri, U.S.A.

Abstract

Poor compliance to prescribed medication is a common problem that can have a major impact on the success of routine patient care and on the conduct and conclusions of clinical trials. The focus of this entry is on compliance to medication regimens and ways in which to measure the effect of non-compliance on a clinical trial.

INTRODUCTION

Poor compliance to prescribed medication is a common problem that can have a major impact on the success of routine patient care and on the conduct and conclusions of clinical trials. Reported clinical correlates of poor compliance include increased hospitalization in patients with hypertension,^[1] diminished improvements in low-density lipoprotein cholesterol levels in hypercholesteremic patients,^[2] and increased mortality in patients with acute myocardial infarction.^[3] The potential effects of inadequate compliance in clinical trials include underestimation of true therapeutic efficacy because medication is not taken as intended and large increases in sample-size requirements because between-group differences are reduced.^[4–7] If the therapeutic impact of poor compliance is of sufficient magnitude, the associated reduction in statistical power can produce negative conclusions in a study that might otherwise have been positive.

Broadly speaking, if a patient's behavior is consistent with the recommendations of the physician or other health-care provider, the patient is thought of as being compliant. Poor compliance can take many forms, e.g., inattention to dietary or exercise recommendations, not taking the prescribed number of pills or taking them at irregular or otherwise nontherapeutic intervals, not refilling prescriptions, and not showing up at follow-up clinic visits. The focus of this review is on compliance to medication regimens.

Patients may be noncompliant for many reasons. The most obvious cause is simple forgetfulness, a more likely circumstance if a medication is being used for an asymptomatic condition such as hypertension. Alternatively, poor compliance may be a conscious decision based on undesired treatment side effects, on the patient's perception that the illness being treated is not serious, or on patient-specific beliefs about the potential impact of individual behaviors on health status. For example, one study of elderly

diabetics^[8] evaluated compliance to measures aimed at controlling diabetes using fasting blood sugar levels and diabetes-associated hospitalization rates as surrogates. Results indicated that improved control of diabetes was significantly associated with the belief that specific behaviors were likely to improve the status of one's disease and with the general perception that an individual's behavior can have a positive impact on health status. Additional reasons for poor compliance include the cost of the medication, the cognitive status of the patient, and the complexity of the drug-taking regimen. The latter two problems are particularly relevant in elderly patients who may be unable to follow the details of a series of medication schedules for multiple chronic conditions.

MEASURING COMPLIANCE

The traditional approaches to measuring compliance include biological markers such as plasma concentrations and urine assays, requesting the information directly from the patient, and using pill counts that reflect the content of returned pill bottles. Each of these approaches can provide useful information. At the same time, each has been shown to be inadequate in key respects: (1) by studies that yield different compliance rate estimates when assessments are carried out using two or more techniques; (2) by comparisons using the newest approach to quantifying compliance—electronic monitors.

The limitations of pill counts have been emphasized by several authors.^[9,10] For example, Pullar et al.^[9] assessed compliance by using both returned-pill counts and plasma concentrations of phenobarbital in 225 subjects. Among the 204 patients who returned for a scheduled follow-up visit with their pill bottles, 161 (78.9%) were classified as compliant because pill counts indicated that between 90% and 109% of prescribed medication had been taken. But 51 of these 161 patients (31.7%) had

doses and body-weight-corrected plasma phenobarbital concentrations that, according to predefined criteria, were indicative of poor compliance. Based on this large discrepancy between the two methods, the authors conclude that “return tablet count grossly overestimates compliance” and that the continued use of pill counts to measure compliance in clinical trials “cannot be justified.” The inadequacy of patient interviews in measuring compliance is highlighted by reports such as that of Norell,^[11] who summarizes data from four studies^[12–15] that measured compliance using both patient interviews and other more objective assessment techniques. The result was that interview noncompliance rates ranged from 12% to 31%, as compared with a range of 32–82% using the more objective approach. The interview produced the lower estimate of noncompliance by amounts that ranged from 20% to 65%. In another study, Caron^[16] reported that the average patient took 47% of the prescribed dose but claimed to have taken 89%.

In recent years, the development of microcircuitry has facilitated the use of miniaturized computer chips as electronic monitors that measure compliance by recording the date and time that medication is dispensed. These devices are now broadly accepted as providing the most accurate available measures of compliance. They have been attached to pill bottles,^[17,18] eye droppers,^[19] blister packs,^[20] and aerosol containers.^[21] They have also confirmed the observation that traditional assessment techniques often yield substantial overestimates of compliance. And, as we discuss next, they have provided important new insights into patterns of noncompliance.

MAGNITUDE AND PATTERN OF POOR COMPLIANCE

While some studies have reported excellent compliance rates as high as 90%,^[20,22] it is well documented that compliance to medication can often be quite poor.^[5,17,18,19] Good compliance tends to be associated with medications that have few side effects, with acute and symptomatic diseases, and with recent diagnoses. In the Physicians' Health Study,^[22] good compliance was likely a reflection of the motivation of an interested and highly educated sample. However, when compliance is poor, it can sometimes be in the range of 50% or less.^[5,19] Given that the standard methods for measuring compliance tend to yield overestimates, it is reasonable to expect that many reported noncompliance percentages are biased underestimates of the true magnitude of the problem.

While the number of doses taken may provide the best single measure of compliance to a prescribed medication regimen, it is not the only such measure. Indeed, it has become increasingly apparent that an accurate pill count may provide an inaccurate measure of patient compliance because drugs taken with the correct total dose may

simultaneously be taken at irregular and nontherapeutic intervals. Similarly, pill counts that are indicative of less than ideal compliance may understate the problem because of the same kind of irregular drug-taking patterns.

The dimensions and characteristics of irregular dosing can best be evaluated using electronic monitoring devices, which have demonstrated, for example, that compliance tends to improve in the days immediately preceding an office visit to the physician. Thus Cramer et al.^[18] found that compliance percentages were 88% during the 5 days immediately before a clinic visit but only 73% when no clinic visit was either imminent or had just taken place. Kass et al.^[19] found similarly that compliance rates were much higher during the 24-hr period preceding an appointment than during the rest of the observation period. One implication of these observations is that even if a biological marker provides an accurate estimate of compliance immediately prior to the assay, the marker may substantially overstate the general pattern of compliance because of the tendency of patients to be more compliant when they are about to see the doctor.

A further contribution of electronic monitors to an understanding of compliance patterns lies in the observation that “drug holidays,” sustained periods during which no medication is taken, are common.^[17–19] Thus De Klerk et al.^[17] studied the drug-taking patterns of 65 patients with ankylosing spondylitis who were prescribed once-a-day medication. The mean period of study was 225 ± 90 (SD) days, for a total of 14,607 monitored days. While 80% of prescribed doses were taken, only 61% were taken during the required 24 ± 6 hr. Extra doses were taken on 3.4% of days, while no dose was taken on 19% of monitored days. Among the nine subjects who had a high pill-taking rate of between 81% and 90%, four had a drug holiday of 13 or more consecutive days. In the Kass et al. study^[19] of 184 glaucoma patients being treated with four-per-day pilocarpine eye drops, the electronic monitor found that 24.5% of patients had at least 1 day per month with no administration of drug, while 52.7% of patients took five or more doses during at least one 24-hr period. Because the nighttime dose was omitted, 50% of the time in a study whose overall compliance rate was $76.0 \pm 24.3\%$ highlights the importance of patient-friendly drug-taking schedules as a compliance-enhancing strategy. The potential clinical impact of irregular drug-taking behaviors was demonstrated by Cramer et al.^[23] in an electronically monitored study of epilepsy patients that found a strong tendency for seizures to occur during the period after doses were missed.

One further aspect of drug-taking patterns is that there is an apparent tendency for compliance to decrease over time.^[5,22,24,25] Thus in the Physicians Health Study,^[22,24] compliance rates of 95.3% during a prerandomization run-in period were reduced to 83.8% during the study itself. The investigators in the Studies of Left Ventricular Dysfunction (SOLVD^[25]) reported that compliance rates were

80% at 1 yr, 74% after 2 yr, and 69% after 3 yr. The results of the 10-week follow-up period of the Vaginal Infections and Prematurity (VIP) Study^[5] are even more pronounced. In VIP, pill counts produced an estimated compliance rate of 85% during the first week of follow-up. But these rates decreased steadily to 47% during week 9 and only 43% during the final study week.

EFFECT OF POOR COMPLIANCE ON SAMPLE SIZE REQUIREMENTS IN CLINICAL TRIALS

When patients do not comply with a prescribed therapeutic regimen, a likely consequence is reduced treatment efficacy. But in the context of a clinical trial, reduced efficacy will yield a corresponding reduction in between-group differences and an associated increase in the required sample size. What is surprising in all of this is the magnitude of the potential sample-size increase.^[4-7]

To quantify the impact that partial compliance may have on sample-size requirements in clinical trials, Schechtman and Gordon^[4] employ both a hypothetical example and the detailed compliance data provided by the Lipid Research Clinics (LRC) Coronary Primary Prevention Trial.^[2,26] In the hypothetical example, the goal is to compare success rates in a clinical trial where the placebo-group success rate is 40% and the treatment-group success rate is 60% in patients who take all of their medication as prescribed. A compliant patient is one that takes at least 80% of the medication, and it is assumed that the success rate in the average compliant patients is 90% of the success rate in patients who take all of their pills. Using data from the LRC trial that translate into noncompliant patients having a cholesterol reduction of only 35.2% of the reduction in compliant patients, the hypothetical example assumes that noncompliers will receive 40% of the therapeutic benefit that is realized in compliers. Based on these assumptions, Table 1 provides a summary of Schechtman and Gordon's conclusions about the association between compliance rates and sample size. The results indicate, for example, that if the compliance rate is 70%, the sample size requirement is 1.84 times as great as it would be if all patients took all of their pills. The corresponding sample size ratio is 2.14 for a compliance rate of 60% and 2.52 if the compliance rate is 50%.

In the LRC trial,^[2,26] a treatment group that was prescribed six daily packets of cholestyramine resin, a cholesterol-lowering agent, was compared to a placebo group in a randomized double-blind trial involving 3806 hypercholesteremic men. The primary outcome measure was coronary heart disease (CHD) mortality. If the definition of compliance requires that the subject takes at least five of the six daily packets, then the reported compliance rate in the placebo group was 67.3%, as compared with 50.8% in the treatment group.^[2,4] Based on these compliance data, on the association between compliance

Table 1 Effect of Poor Compliance on Sample-Size Requirements

Percentage of patients who comply	Sample-size ratio
90%	1.40
80%	1.59
70%	1.84
60%	2.14
50%	2.52
40%	3.01

Data are from a clinical trial where the success rate is 40% among the placebo-group patients and 60% in the treatment-group patients who take all of their pills. It is assumed that the average complier receives 90% of the benefit of subjects who take all of their pills and that the average noncomplier receives 40% of the benefit of subjects who take all prescribed medication. The sample-size ratio is the ratio of the required sample size for the tabulated percentage of patients who are compliant to the corresponding sample size if all patients take all of their pills.

rates and cholesterol reduction in the LRC trial, and on the association between cholesterol reduction and CHD mortality, it was demonstrated^[4] that if the compliance rate in the treatment group (50.8%) had been as high as it was in the placebo group (67.3%), the sample-size requirement in the LRC trial would have been reduced by about 41%. Moreover, the required sample size with the observed compliance rate was 2.97 times as great as it would have been if all patients had been compliers.

EFFECT OF POOR COMPLIANCE ON THE CLINICAL STATUS OF PATIENTS

Poor compliance is a problem only if there is a causative relationship between not taking pills and the clinical status of patients. On the one hand, it seems intuitively likely that partial compliance to an effective medication will increase the probability of negative clinical consequences. But at the same time, the magnitude of this causative relationship is likely to heavily depend on case-specific considerations, such as the disease in question, the half-life of the drug, and the frequency with which medication is prescribed. Moreover, because compliance may be strongly associated with other behavioral factors that influence clinical outcomes, it may be difficult to determine the degree to which poor compliance, as opposed to the behavioral correlates of poor compliance, are the cause of negative patient outcomes. In the Coronary Drug Project,^[27] for example, compliers to the lipid-lowering agent clofibrate had a much lower mortality rate than noncompliers to clofibrate ($P = 0.0001$). The initial suggestion is that there is a strong causal relationship between clofibrate compliance and reduced mortality. But a further look at the data suggests caution because the same association between compliance and survival exists when placebo-group data are evaluated. Apparently, the reason for the reduced mortality among

Table 2 Data from the LRC Trial on Association Between Compliance to Medication and Cholesterol Reduction in the Cholestyramine Group and in the Placebo Group

Mean daily packet count	Placebo group		Cholestyramine group		Treatment effect
	N	Percentage change in cholesterol	N	Percentage change in cholesterol	
0–1	133	–3.2	294	–3.9	–0.7
1–2	79	–1.7	145	–5.4	–3.7
2–3	88	–4.0	135	–8.2	–4.2
3–4	105	–3.5	156	–11.1	–7.6
4–5	214	–4.1	205	–14.0	–9.9
5–6	1274	–5.4	965	–19.0	–13.6

The treatment effect column is the difference between the percentage change in cholesterol in the two groups.

compliers was not the drug. Instead, the real explanation may be that medication compliers in the Coronary Drug Project also tended to engage in routine daily activities that improve cardiovascular health.

The message in the preceding example is that when disease status is heavily influenced by patient behaviors other than compliance, it may be difficult to determine whether poor compliance or correlated behavioral factors are the cause of negative outcomes. With this caveat in mind, we illustrate the degree to which poor compliance can indeed have serious negative clinical consequences.

The LRC trial data on the association between compliance to cholestyramine and cholesterol reduction^[2] are summarized in Table 2. While the table exhibits a small dose–response relationship in the placebo group, the association between cholesterol reduction and the number of packets taken is far more pronounced in the cholestyramine group. This is highlighted by the treatment effect column of Table 2, which measures treatment benefit after subtracting off the cholesterol reduction in the placebo group. Thus in contrast to the Coronary Drug Project, the association between poor compliance and poor outcome can be overwhelmingly attributed to the direct causative effects of not taking the drug as prescribed. In combination with LRC data on the association between CHD events and cholesterol reduction, the treatment-effect column of Table 2 has been used to estimate the expected number of CHD events that would have been prevented if LRC compliance rates had been greater than they actually were.^[4] The observed number of CHD events in the LRC trial was 155 in the cholestyramine group and 187 in the placebo group, suggesting that the study medication prevented about 32 events. But because of the side effects of the drug, compliance to cholestyramine (50.8%) was much lower than compliance to placebo (67.3%), where compliance is defined as taking at least five of six prescribed packets. If compliance in the treatment group had been as high as it was in the placebo group, the expected number of treatment group events would have been 145, so the number of events prevented would have increased by an expected total of 10, from 32 to

42. And if one assumes that all cholestyramine patients are in the highest compliance category of Table 2, the expected number of events prevented is increased by 24, from 32 to 56. That is, full compliance to cholestyramine is associated with an expected 75% (24/32) increase in the number of CHD events prevented by the drug, in comparison to what actually happened in the LRC trial.

The literature contains numerous other studies that associate poor compliance with negative clinical consequences, such as increased hospitalization^[1] and reduced control^[20] in patients with hypertension, seizures in patients with epilepsy,^[23] increased mortality and organ rejection in transplant recipients,^[28] and increased coronary heart disease event rates when beta blockers are not taken.^[29] The magnitude of these negative consequences is highlighted by the Joint National Committee on Evaluation, Detection, and Treatment of High Blood Pressure, which asserts that noncompliance is the most important reason for uncontrolled high blood pressure;^[30] by a 1984 symposium on the therapeutic consequences of noncompliance, which suggested that 125,000 annual deaths result from noncompliance in patients with cardiovascular diseases;^[31] and by the U.S. Chamber of Commerce estimates that treating noncompliance-induced medical problems is associated with an annual cost of \$13–15 billion.^[32]

IMPROVING COMPLIANCE IN CLINICAL TRIALS

Since poor compliance to medication has the potential to substantially reduce treatment efficacy, it is not surprising that compliance-enhancing strategies are often a major component of clinical research. Broadly speaking, such strategies can be divided into two general categories—those that precede the initiation of study therapy^[2,5,22,24–26,33] and those that are implemented after randomization.^[2,26,34]

Compliance-enhancing strategies that precede randomization are usually focused on maximizing the likelihood that patients who enter the study will be good compliers. This is ordinarily accomplished using either a formal or

an informal run-in strategy. A run-in strategy is carried out during a prerandomization period during which the likely future compliance of the patient is assessed. This may be accomplished informally, as it was in the LRC trial,^[2,26] using a screening period during which patients are expected to attend prerandomization visits aimed primarily at determining eligibility and doing preliminary assessments. If the patient does not attend these sessions or is otherwise uncooperative, the conclusion may be that the patient should not be randomized because early behavior suggests future noncompliance. Alternatively, a formal run-in strategy may be carried out.^[5,22,24,25] A formal run-in involves a prerandomization period during which potential study subjects are asked to participate as if they had already been randomized. This specifically includes the taking of placebo medication and attendance at clinic visits that would be standard if the subject were to be randomized. Patients who do not comply up to some prespecified level are then excluded from the trial prior to randomization. It has been demonstrated that run-in strategies have increased the power of several major clinical trials.^[5,24] In general, run-in strategies are likely to be of benefit when poor compliance is associated with a substantial reduction in response, when poor compliance is common early in the study, when most early poor compliers could be expected to continue that pattern if randomized, when the run-in itself is inexpensive, or when the pool of eligible patients is sufficient to ensure that run-in exclusions will not compromise recruitment goals.^[35]

Compliance-enhancing strategies that follow randomization cover a broad range of potential activities and have been discussed by many investigators.^[36–40] These strategies include: educating patients about their disease and about the importance of compliance; financial and other incentives;^[40] simplified and flexible dosing regimens that are adjusted to fit individual patient schedules; establishing positive clinician–patient relationships; self-administered approaches, such as providing calendars to be checked off when doses are taken; specialized drug packaging, such as blister packs^[20] and caps for prescription containers that sound an alarm when it is time to take the next dose;^[38] and encouraging family members to provide support.

If there is any unifying theme in the investigation of these procedures, it is that no single approach can be successfully applied in all settings. Compliance-enhancing strategies must be tailored to the patient and to the particular circumstance. Their success depends heavily upon the ingenuity and the commitment of the health care provider.

CONCLUSION

Poor compliance prescribed medication is a common problem that can have negative consequences regarding both patient outcome and the conduct of clinical trials. The

magnitude of compliance can best be evaluated using electronic monitors which have demonstrated that compliance rates as low as 50% are not uncommon. In such a setting, treatment efficacy and between group differences are reduced, a circumstance which can lead to a doubling or even tripling of the required sample size in a clinical trial. More serious are the clinical consequences of poor compliance which have been demonstrated in a variety of studies to include increased mortality and hospitalization rates, reduced control of hypertension, increased seizure frequency in epileptics, and increased rejection rates in transplant patients. Because of these negative consequences, it is incumbent upon both clinicians and researchers to use all available compliance-enhancing strategies to encourage patients to take their medication in accordance with the prescribed regimen.

REFERENCES

1. Maronde, R.F.; Chan, L.S.; Larsen, F.; Strandberg, L.R.; Laventurier, M.F.; Sullivan, S.R. Underutilization of anti-hypertensive drugs and associated hospitalization. *Med. Care* **1989**, *27*, 1159–1166.
2. Lipid Research Clinic Program, The lipid research clinics coronary primary prevention trial results II. Relationship of reduction in incidence of coronary heart disease to cholesterol lowering. *JAMA* **1984**, *251*, 365–374.
3. Beta-Blocker Heart Attack Trial Research Group, A randomized trial of propranolol in patients with acute myocardial infarction, 1: Mortality results. *JAMA* **1982**, *247*, 1701–1714.
4. Schechtman, K.B.; Gordon, M.E. The effect of poor compliance and treatment side effects on sample size requirements in randomized clinical trials. *J. Biopharm. Stat.* **1994**, *4*, 223–232.
5. Blackwelder, W.C.; Hastings, B.K.; Lee, M.L.F.; Deloria, M.A. Value of a run-in period in a drug trial during pregnancy. *Control. Clin. Trials* **1990**, *11*, 187–198.
6. Freedman, L.S. The effect of partial noncompliance on the power of a clinical trial. *Control. Clin. Trials* **1990**, *11*, 157–168.
7. Lachin, J.M.; Foulkes, M.A. Evaluation of sample size and power for analyses of survival with allowance for non-uniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics* **1986**, *42*, 507–519.
8. Hopper, S.V.; Schechtman, K.B. Factors associated with diabetic control and utilization patterns in a low-income, older adult population. *Patient Educ. Couns.* **1985**, *7*, 275–288.
9. Pullar, T.; Kumar, S.; Tindall, H.; Feely, M. Time to stop counting tablets? *Clin. Pharmacol. Ther.* **1989**, *46*, 163–168.
10. Rudd, P.; Byyny, R.L.; Zachary, V.; LoVerde, M.E.; Titus, C.; Mitchell, W.D.; Marshall, G. The natural history of medication compliance in a drug trial: Limitations of pill counts. *Clin. Pharmacol. Ther.* **1989**, *46*, 169–176.
11. Norell, S.E. Methods in assessing drug compliance. *Acta Med. Scand.* **1984**, *683*, 35–40 (suppl).

12. Feinstein, A.R.; Wood, H.F.; Epstein, J.A.; Taranta, A.; Simpson, R.; Tursky, E. A controlled study of three methods of prophylaxis against streptococcal infection in a population of rheumatic children. *N. Engl. J. Med.* **1959**, *260*, 697–702.
13. Bergman, A.B.; Werner, R.J. Failure of children to receive penicillin by mouth. *N. Engl. J. Med.* **1963**, *268*, 1334–1338.
14. Gordis, L.; Markowitz, M.; Lilienfeld, A.M. The inaccuracy in using interviews to estimate patient reliability in taking medications at home. *Med. Care* **1969**, *7*, 49–54.
15. Hecht, A.B. Improving medication compliance by teaching outpatients. *Nurs. Forum* **1974**, *13*, 112–129.
16. Caron, H.S. Compliance and the case of objective measurement. *J. Hypertens.* **1985**, *3* (Suppl I), 11–17.
17. De Klerk, E.; Van der Linden, S.; Van der Heijde, D.; Urquart, J. Facilitated analysis of data on drug regimen compliance. *Stat. Med.* **1997**, *16*, 1643–1664.
18. Cramer, J.A.; Scheyer, R.D.; Mattson, R.H. Compliance declines between clinic visits. *Arch. Intern. Med.* **1990**, *150*, 1509–1510.
19. Kass, M.A.; Meltzer, D.W.; Gordon, M.; Cooper, D.; Goldberg, J. Compliance with topical pilocarpine treatment. *Am. J. Ophthalmol.* **1986**, *101*, 515–523.
20. Eisen, S.A.; Woodward, R.S.; Miller, D.; Spitznagel, E.; Windham, C.A. The effect of medication compliance on the control of hypertension. *J. Gen. Intern. Med.* **1987**, *2*, 298–305.
21. Specter, S.L.; Kinsman, R.; Mawhinney, H.; Siegel, S.C.; Rachelefsky, G.S.; Katz, R.M.; Rohr, A.S. Compliance of patients with asthma with an experimental aerosolized medication: Implications for controlled clinical trials. *J. Allergy Clin. Immunol.* **1986**, *77*, 65–70.
22. Steering Committee of the Physicians' Health Study Research Group. Final report on the aspirin component of the ongoing Physicians' health study. *N. Engl. J. Med.* **1989**, *321*, 129–135.
23. Cramer, J.A.; Mattson, R.H.; Prevey, M.L.; Ouellette, V. How often is medication taken as prescribed? A novel assessment technique. *JAMA* **1989**, *161*, 3273–3277.
24. Lang, J.M.; Buring, J.E.; Rosner, B.; Cook, N.; Hennekens, C.H. Estimating the effect of the run-in on the power of the physicians health study. *Stat. Med.* **1991**, *10*, 1585–1593.
25. The SOLVD Investigators. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *N. Engl. J. Med.* **1991**, *325*, 293–302.
26. Lipid Research Clinics Program. The Lipid research Clinics coronary primary prevention trial results I. Reduction in incidence of coronary heart disease. *JAMA* **1984**, *251*, 351–364.
27. Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project. *N. Engl. J. Med.* **1980**, *303*, 1038–1041.
28. Schweizer, R.T.; Rovelli, M.; Palmeri, D.; Vossler, E.; Hull, D.; Bartus, S. Noncompliance in organ transplant recipients. *Transplantation* **1990**, *49*, 374–377.
29. Psaty, B.M.; Koespell, T.D.; Wagner, E.H.; LoGerfo, J.P.; Inui, T.S. The relative risk of incident coronary heart disease associated with recently stopping the use of beta-blockers. *JAMA* **1990**, *263*, 1653–1657.
30. Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure. The 1984 report of the Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure. *Arch. Intern. Med.* **1984**, *144*, 1045–1057.
31. Burrell, C.D.; Levy, R.A. Therapeutic Consequences of Noncompliance. In *Improving Medication Compliance: Proceedings of a Symposium*; National Pharmaceutical Council: Washington, DC, 1984; 7–16.
32. Yenney, S.L.; Behrens, R.A. *Health Promotion and Business Coalitions*; U.S. Chamber of Commerce: Washington DC, 1984.
33. The CAST Investigators. Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N. Engl. J. Med.* **1989**, *321*, 406–412.
34. Black, D.M.; Brand, R.J.; Greenlick, M.; Hughes, G.; Smith, J. Compliance to treatment for hypertension in elderly patients: The SHEP pilot study. Systolic hypertension in the elderly program. *J. Gerontol.* **1987**, *42*, 552–557.
35. Schectman, K.B.; Gordon, M.E. A comprehensive algorithm for determining whether a run-in strategy is a cost-effective design modification in a randomized clinical trial. *Stat. Med.* **1993**, *12*, 111–128.
36. Peck, C.L.; King, J.N. Increasing patient compliance with prescriptions. *JAMA* **1982**, *248*, 2874–2877.
37. Eraker, S.A.; Kirscht, J.P.; Becker, M.H. Understanding and improving patient compliance. *Ann. Intern. Med.* **1984**, *100*, 258–268.
38. Bond, W.S.; Hussar, D.A. Detection methods and strategies for improving medication compliance. *Am. J. Hosp. Pharm.* **1991**, *48*, 1978–1988.
39. Melnikow, J.; Kiefe, C. Patient compliance and medical research: Issues in methodology. *J. Gen. Intern. Med.* **1994**, *9*, 96–105.
40. Giuffrida, A.; Torgerson, D.J. Should we pay the patient? Review of financial incentives to enhance patient compliance. *Br. Med. J.* **1997**, *315*, 703–707.

Patient-Reported Outcomes

Amylou C. Dueck, PhD

Mayo Clinic, Scottsdale, Arizona, U.S.A.

Joseph C. Cappelleri, PhD

Pfizer Inc., Groton, Connecticut, U.S.A.

Abstract

A patient-reported outcome (PRO) is any report directly from the patient about his or her health status or condition without interpretation by a clinician or anyone else. PRO data may be collected using patient-completed questionnaires or by interviews where the patient's response is directly recorded without modification. PROs have been developed to validly measure a variety of concepts (e.g., symptoms, well-being, health-related quality of life [HRQOL], and treatment satisfaction) for purposes such as assessing disease impacts as well as treatment benefits and harms. PROs are inherently important in this era of patient-centered healthcare and are common endpoints in randomized clinical trials in a variety of disease settings (e.g., asthma, irritable bowel syndrome, pain, arthritis, and cancer). The purpose of this entry is to provide a summary of statistical methods commonly employed for the analysis of PRO data in randomized clinical trials. The general approach to statistical analysis should mirror that applied to any other endpoint in clinical research. Specifically, the analysis plan should be based on the selection of an appropriate analytical method based on the stated hypothesis and distribution of the endpoint being analyzed. This entry outlines statistical approaches for cross-sectional and longitudinal comparisons. An overview of approaches for handling multiplicity and missing data is also presented.

Keywords: Quality of life; Health-related quality of life; Questionnaires; Surveys; Patient experience; Symptoms; Self-report.

INTRODUCTION

A patient-reported outcome (PRO) is any report directly from the patient about his or her health status or condition without interpretation by a clinician or anyone else. PRO data may be collected using patient-completed questionnaires or by interviews where the patient's response is directly recorded without modification. PROs have been developed to validly measure a variety of concepts (e.g., symptoms, well-being, health-related quality of life [HRQOL], and treatment satisfaction) for purposes such as assessing disease impacts as well as treatment benefits and harms. PROs are inherently important in this era of patient-centered healthcare and are common endpoints in randomized clinical trials in a variety of disease settings (e.g., asthma, irritable bowel syndrome, pain, arthritis, and cancer).

The purpose of this entry is to provide a summary of statistical methods commonly employed for the analysis of PRO data in randomized clinical trials. The general approach to statistical analysis should mirror that applied to any other endpoint in clinical research. Specifically, the analysis plan should be based on the selection of an appropriate analytical method based on the stated hypothesis

and distribution of the endpoint being analyzed, with careful considerations for handling multiple testing as well as missing data.

This entry outlines statistical approaches for cross-sectional and longitudinal comparisons. An overview of approaches for handling multiplicity and missing data is also presented. This entry is intended to expose readers to the myriad of appropriate statistical approaches. Key references for additional reading are given to further help readers to execute sound statistical analysis of PRO data in a randomized clinical trial.

PROs IN THE REGULATORY SETTING

The term “patient-reported outcome” or “PRO” is defined by the United States (US) Food and Drug Administration (FDA) as “a measurement based on a report that comes directly from the patient (i.e., study subject) about the status of a patient's health condition without amendment or interpretation of the patient's response by a clinician or anyone else.”^[1] The FDA identifies PROs as one of the four clinical outcome assessments (COAs) that directly or indirectly measure treatment

benefits from patients, clinicians (clinician-reported outcomes [ClinROs]), non-clinician observers (observer-reported outcomes [ObsROs]), or performance assessments (performance outcomes [PerfOs]).

PROs are unique from other outcomes because they can measure concepts that are unobservable to individuals other than the patient, concepts such as some symptoms (e.g., pain and fatigue) or HRQOL. PROs can also measure concepts such as physical functioning, which may be observable by others. The FDA's guidance document^[2] on the use of PROs for product labeling remains the gold standard for the statistical analysis of PRO data for product labeling purposes in the United States. The European Medicines Agency (EMA) released a reflection paper in 2005 detailing statistical recommendations for use of HRQOL in the drug approval process in Europe.^[3] The statistical methods presented herein represent a broader collection of analytical techniques not limited to those for the specific purpose of achieving a product label claim. Thus, the applicable regulatory recommendations as well as regulatory officials should be consulted to identify the specific recommended approaches if a product label is the intended goal of analysis.

DEFINING THE ENDPOINT

PRO data are typically captured using questionnaires that undergo rigorous development processes to ensure acceptable psychometric (i.e., measurement) properties in the given target patient population (see work by Reeve et al.^[4] for recommended minimum PRO standards). In addition to adequate measurement properties, other considerations when selecting among available questionnaires may include the proposed use of the questionnaire (i.e., hypothesis to be tested), the concept to be measured, patient and administrative burden, interpretability of scores, appropriateness for the given patient population (taking into consideration such things as age, disabilities, and comprehension level), availability of language translations, and costs. Formats of individual items or questions across questionnaires vary in terms of item wording, response options, and recall period (i.e., the period of time for which patients are asked to remember when responding to the item). Each questionnaire may contain one or more questions. Typically, each questionnaire has a scoring algorithm established during its development, which should be employed to calculate scores for statistical analysis when using the questionnaire in a trial. Depending on the questionnaire, scoring may result in a single overall score and/or scores for one or more individual items or subscales (or domains). When multiple domain scores are computable, each defined endpoint in the statistical analysis plan (SAP) should clearly specify the score to be used.

Selected questionnaires are often administered at multiple time points including prior to treatment, during, and/

or after treatment. Time points are often selected based on when changes or differences are expected to occur and typically synchronous with collection time points of other clinical data. See Fairclough^[5] for information regarding selection of appropriate time points for administering PRO questionnaires. Patient burden should be a key consideration in both selecting PRO questionnaires and time points in a trial.

With the PRO questionnaire, specific score of interest (if applicable) and time points identified, there are several options for defining an endpoint. For example, in a fictitious trial testing a new eight-week treatment for fatigue versus a placebo treatment in which the PRO questionnaire (e.g., a validated questionnaire that results in a single fatigue score on a 0–10 scale with higher scores representing worse fatigue) is administered prior to treatment and after two, four, six, and eight weeks of treatment, endpoints can be defined as the fatigue score at a fixed time point (e.g., at two, four, six, or eight weeks), a summary statistic which is computed across time points (e.g., best or average fatigue score observed across two, four, six, and eight weeks), or as the collection of longitudinal measurements over time. In trials testing treatments given for a limited duration, the appropriate endpoint typically reflects the final outcome of the trial (e.g., the final time point in which PROs are assessed). However, assessments during treatment may be useful to characterize the trajectory of responses within and between treatment arms.

In trials testing extended-duration treatments, such as advanced cancer treatments which are administered indefinitely until the patient's cancer progresses, an appropriate endpoint may be defined at a fixed time point which is both early enough in the trial so that most patients are still expected to be receiving treatment and late enough so that the treatment's impact is detectable. Often alternative endpoints employing summary statistics across time points or longitudinal measurements over time are also appropriate. In general, the selected endpoint in any trial should reflect the objective of the given analysis (i.e., answer the question being asked). Once an endpoint is defined, then an appropriate statistical analysis strategy can be selected.

CROSS-SECTIONAL ANALYSIS

For an ordinal or continuous PRO score at a fixed time point, comparison between two or more arms may employ standard parametric (e.g., two-sample t-test or analysis of variance) or non-parametric (e.g., Wilcoxon rank-sum test or Kruskal–Wallis test) analyses. Note, the central limit theorem insures appropriateness of parametric approaches, even for ordinal PRO scores with limited response options and with floor or ceiling effects for sample sizes as small as five per arm.^[6,7]

Other popular approaches to account for potential imbalances in baseline scores include computation of the

change from baseline or the percentage change from baseline for each patient with subsequent comparison between arms using the same parametric or non-parametric approaches. However, Vickers^[8] points out that these methods theoretically fail to protect from bias introduced by imbalanced baseline scores and that the percentage change from baseline approach can lead to non-normally distributed data. Vickers also showed that an analysis of covariance (ANCOVA) approach comparing the PRO score at the fixed time point between arms and including the baseline score as a covariate provided the highest power, making this a favorable approach for the comparison of an ordinal or continuous PRO score at a fixed time point (when no tangible bias from missing data is expected).

Multivariable linear regression can also be considered to model an ordinal or continuous PRO score with treatment arm, baseline score, and other baseline patient characteristics expected to be related to the postbaseline PRO score as predictor variables. Sullivan and D’Agostino^[9] depict that linear regression performs similarly to ordinal logistic regression for ordinal PRO scores even with relatively small sample sizes. For a binary PRO score at a fixed time point, arm comparisons may employ chi-squared or similar tests, or a logistic regression-based approach, including treatment arm and baseline patient characteristics expected to be related to the postbaseline PRO score as predictor variables.

SUMMARY MEASURES AND STATISTICS

One of the methods for inferential purposes is to use summary measures or summary statistics. For instance, a single measure can be constructed by aggregating data across different domains on the same questionnaire if such aggregation is justified clinically and statistically. Such a summary measure can be used as the primary endpoint for hypothesis testing.

Summary statistics can be constructed on a particular subscale or domain of an instrument to summarize its repeated observations over time within an individual and then across individuals in the same treatment arm. Examples include, for each treatment arm, the average of the within-subject average, maximum, minimum, or last observed postbaseline score; the within-subject least-squares estimated slope across postbaseline scores; within-subject area under the curve (AUC); and within-subject time to reach a peak or a prespecified value.

Use of summary measures and statistics can serve several purposes. These endpoints can aid in interpretation as well as handle multiple testing by combining data across scales and/or time points into a single score. They can also be an analysis strategy for minimizing missing data because a patient can often be evaluable even if the patient missed an assessment time point. However, caution should be taken to consider any bias introduced by missing data on

these endpoints. Caution should also be taken when using summary measures and statistics to ensure that high-level summarization does not obfuscate important differences in individual scales and/or time points.

Area Under the Curve

One clinically intuitive summary statistic is the AUC. An AUC can be computed for each individual patient by connecting observed scores (Table 1) longitudinally over a fixed period of time, computing the area of each resulting trapezoid, and summing the areas of these sequential trapezoids (Fig. 1). AUCs (computed for each individual

Table 1 Sample dataset (first two patients are shown)

PatientID	WEEK	ARM	FATIGUE
1	0	1	6
1	2	1	7
1	4	1	5
1	6	1	4
1	8	1	5
2	0	1	7
2	2	1	6
2	4	1	5
2	6	1	4
2	8	1	2
...	...	...	...

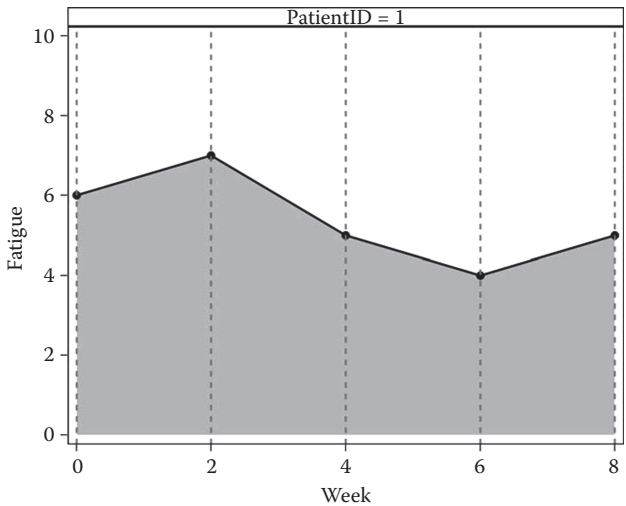


Fig. 1 Example of patient-level area under the curve calculation. Area under the curve is calculated for PatientID = 1 in Table 1. The areas of the trapezoids from left to right are 13, 12, 9, and 9 for a total area under the curve of 43. In this example, the minimum and maximum possible AUCs are 0 and 80, respectively. Therefore, the AUC of 43 can be interpreted relative to this 0 (no fatigue over the entire eight weeks) to 80 (worst possible fatigue over the entire eight weeks) scale

patient) can then be compared between arms using the previously described cross-sectional approaches. While intermittent missing data are typically handled by connecting only observed scores, missing data at the end of the fixed time period due to dropout have been handled in various ways including single imputation strategies (see Bell et al.^[10] for a summary of such methods). Bell et al.^[10] suggest that an arm-level AUC approach (i.e., computing the AUC for each arm using parameter estimates from a mixed model and comparing these between arms using a contrast) is less biased than comparing within-subject AUCs computed for each individual patient in the presence of dropout.

Responder Analysis

Another popular summarization approach is a responder analysis where a threshold is a priori defined as a clinically meaningful change (i.e., improvement) from baseline in a PRO score. Let us assume that higher scores are more favorable. Then a patient can be classified as a responder if the patient's change score meets or exceeds the selected threshold. Typically, all other patients are considered as non-responders including patients without PRO data available for analysis. This binary outcome is then compared between arms using a chi-squared or similar test. The clinically meaningful improvement is typically specific to the given PRO score, patient population, and trial context (see Wyrwich et al.^[11] for an overview of anchor-based and distribution-based methods for establishing clinically significant change thresholds).

Patients for whom the change threshold is not analytically possible can be excluded from responder analysis. For example, say that response is defined as a 2-point symptom improvement on a 0 = none, 1 = mild, 2 = moderate, 3 = severe, or 4 = very severe scale. Responder analysis would exclude patients reporting no or mild symptom level at baseline. If responder analysis is the primary (or key secondary) analytical approach for the trial, study inclusion criteria can be defined so all subjects have the opportunity to reach the prespecified clinically defined improvement. For instance, all subjects in pain trials are typically required to have at least a baseline pain score of 4 or 5 on an 11-point scale (0 = no pain, 10 = worst imaginable pain).^[12] If study inclusion criteria do not limit enrollment in such a way, power may be compromised if a substantial proportion of patients are not eligible for the planned responder analysis.

The main advantage of responder analysis is ensuring that a statistically significant result represents a clinically meaningful benefit. The disadvantages include the possibility of an increased sample size due to the loss of information associated with reducing a continuous outcome to a binary outcome. Threshold selection can also be somewhat arbitrary. Finally, interpretation can be problematic if response is only a temporary phenomenon.

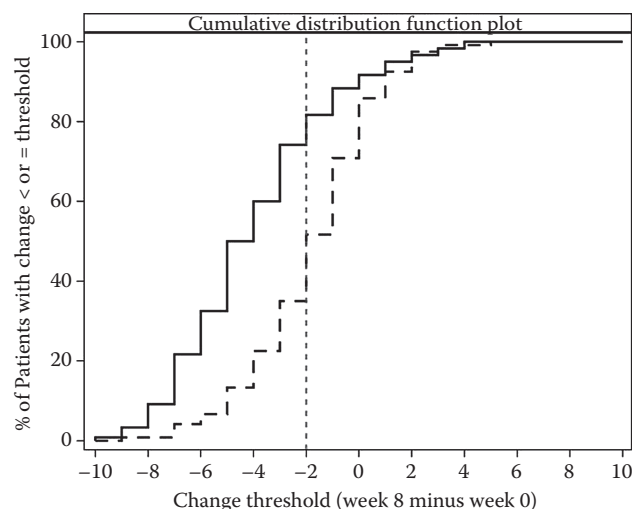


Fig. 2 Example of a cumulative distribution function plot. Negative change scores represent improvement. The solid curve represents ARM = 1 (active treatment arm) and the dashed curve represents ARM = 2 (placebo arm). At a change score of -2 (i.e., a 2-point improvement in fatigue on a 0–10 scale where higher scores represent worse fatigue), 82% of patients on active treatment versus 52% of patients on placebo experienced at least a 2-point improvement (i.e., a change score less than or equal to -2)

Supplemental analysis following any responder analysis should include assessing the impact of the selected responder definition on trial results. A graphical approach is a cumulative distribution function (CDF) plot by arm achieved by plotting the change thresholds on the x-axis and cumulative percentage of patients meeting or exceeding each threshold on the y-axis (Fig. 2). Patients who do not have PRO data available for analysis can be considered to have the worse change possible or simply considered as missing if justified. The CDF plot allows comparison between arms at every possible threshold simultaneously. Horizontal separation between CDFs of two arms supports the trial's results as being consistent across the entire range of thresholds. When CDFs overlap or crisscross, interpretation can be challenging but may be aided by a focus on clinically plausible thresholds. Proper statistical analysis of the full original scale is generally recommended with responder analysis considered as supplementary analysis.

Time to Significant Worsening or Improvement

Another summary statistic transforms longitudinal PRO data into time-to-event endpoints. Time from baseline until a clinically meaningful worsening (or improvement) can be computed for each patient. Patients without clinically meaningful worsening are typically censored at the last completed PRO assessment. Time-to-event data can be analyzed using Kaplan–Meier plots with log-rank tests or Cox regression. Like responder analysis, a threshold

representing clinically meaningful worsening or improvement should be specified a priori.

In some palliative care and/or cancer settings, disease progression and/or death are included along with PRO worsening as events (e.g., see Benzo et al.^[13]). In interpreting such an analysis, researchers should consider whether differences between arms are driven by clinically meaningful worsening of PRO scores, progressions, or deaths. Considering each endpoint separately in addition to the combined endpoint (PRO worsening, progression, or death) may aid in interpretation.

LONGITUDINAL ANALYSIS

Regression-based analyses for longitudinal PRO data are a popular and well-described analytical strategy to analyze both within-arm and between-arm changes in PRO scores over time.^[15] Such modeling approaches do not differ from those described and standardly used for other longitudinal quantitative data.^[14] Some have argued that longitudinal analysis is the preferred analysis strategy over previously described fixed time point analyses in trials with longitudinal measurements.^[15] Prior to embarking on longitudinal modeling, researchers should carefully inspect descriptive statistics such as means over time by arm, within-time point variance, and between-time point covariance to aid in proper selection and design of a longitudinal model. Longitudinal models for PRO data generally fall into two broad categories, event-driven and time-driven models where time is an ordered categorical and continuous predictor variable, respectively.

A common event-driven model is the repeated measures model. This model requires each PRO score be assigned to a discrete time point with only one PRO score per patient assigned to each event. In its simplest form, the PRO score is modeled using an intercept, fixed time effect, and fixed treatment (arm) effect with a random error term which accounts for repeated measurements within subject. A common time-driven model is the random coefficient model or mixed effects growth curve model. Time-driven models can accommodate all assessments regardless of time and can incorporate the actual timing of the assessment. Models typically include a random individual-specific intercept term with or without a random individual-specific slope term, as well as a random error term.

An unstructured covariance matrix does not assume any pattern among errors and does not assume equally spaced time points like other covariance structures such as first-order autoregressive structure. Therefore, it is a flexible and reasonable default selection for event-driven and time-driven models if sample size allows for such. However, it may be desirable to reduce the number of parameters being estimated. Alternative covariance structures from among theoretically plausible structures can be selected based on optimization of model fit criteria such as

Table 2 SAS v9 code for generating simple forms of longitudinal models using data in Table 1. Many variations on these models are possible (e.g., see Fairclough^[5]), such as inclusion of WEEK*ARM interaction terms to allow the impact of ARM to change over repeated measures or time and modeling postbaseline scores only while including the corresponding baseline score as a covariate, to name just two. CONTRAST and ESTIMATE statements can be used within PROC MIXED to set up a variety of hypothesis tests assessing between-arm differences or within-arm changes at specific time points or across time points or even model-based calculations of area under the curve

Description	SAS Code
Repeated measures model (time is a categorical variable)	proc mixed data=dataset1; class WEEK ARM PatientID; model FATIGUE = WEEK ARM / solution ddfm=kr; repeated WEEK / subject=PatientID type=UN r corr; run;
Random intercepts model (time is a continuous variable)	proc mixed data=dataset1; class ARM PatientID; model FATIGUE = WEEK ARM / solution ddfm=kr; random intercept / subject=PatientID type=UN; run;
Random intercepts and slopes model (time is a continuous variable)	proc mixed data=dataset1; class ARM PatientID; model FATIGUE = WEEK ARM / solution ddfm=kr; random intercept WEEK / subject=PatientID type=UN; run;

Akaike information criterion when the same model effects other than the covariance structure remain constant across models (e.g., see Littell et al.^[16]). Longitudinal models can be implemented in SAS, SPSS, Stata, and R (see Table 2 for SAS v9 code). Modifications to models can account for non-linear changes over time such as using polynomial terms or piecewise linear splines. Finally, use of the general estimating equations (GEE) approach, which is an extension of generalized linear models, is commonly employed when modeling binary or count-based PRO data as well as continuous data (see Fitzmaurice et al.^[14] for further information regarding GEE and how it differs from repeated effects and mixed effects models).

MULTIPLE TESTING

Like all other statistical analyses in clinical research, performing multiple PRO hypothesis tests inflates type I error. Multiplicity is a common concern in trials with PROs due to questionnaires producing multiple scale scores, use of multiple PRO questionnaires within the same study, and PRO measurement at more than one time

point during treatment. Multiplicity may also arise from multiple arms within a trial or multiple planned subgroup analyses. One should always consider the study objectives/hypotheses before making a decision to apply a method for handling multiplicity (i.e., though many possible PRO hypotheses could be tested in a study, only one or some may be of interest). If a method is needed, standard methods for handling multiplicity are readily available for use in the analysis of PROs.^[17] The selected approach may be influenced by the trial setting. For example, within a trial intending to support a label claim, strict control of type I error may be required across all PRO and other clinical endpoints that are measuring treatment benefits. Outside of the label claim setting, multiplicity may be considered for PRO endpoints separately from other endpoints. In early-stage clinical trials, relaxed control of type I error may be appropriate as subsequent confirmation is expected.

Arguably, the most common approach for handling multiplicity in trials with PRO endpoints is a specification of a single hierarchy consisting of primary, secondary, and exploratory endpoints. Gatekeeping, fixed sequence, and fallback approaches perform hypothesis tests sequentially in this protocol-specified hierarchy, the simplest of which consists of each hypothesis test being performed as long as the previous hypothesis test reached statistical significance (i.e., $p\text{-value} < 0.05$). Once a hypothesis test fails to reach statistical significance, then no formal subsequent tests are performed. In practice, all subsequent statistical tests should still be carried out and reported with the gatekeeping approach incorporated for interpretation purposes.

Other methods for handling multiple endpoints include alpha adjustment methods (e.g., Bonferroni, Hochberg's step-up, and Bonferroni–Holm's step-down methods) and resampling methods. The Bonferroni approach is considered the most conservative, while Hochberg's method is recommended though still conservative like alpha adjustment methods in general. Resampling methods take into account relationships among test statistics and are, therefore, considered less conservative. Another alternative is to allocate different type I error rates to more than one endpoint (e.g., allocate 0.03 to one endpoint and 0.01 to each of two other less important endpoints for a total maximal type I error rate of 0.05).

As previously discussed, summary measures and statistics that combine data across scales and time points, respectively, may also be an appropriate method for handling multiple testing by reducing the number of statistical tests down to a single comparison of the summary measure or statistic. Finally, a variety of multivariate methods perform an overall hypothesis test (i.e., produce a single $p\text{-value}$) across a collection of endpoints to avoid multiple testing. O'Brien's global test is a non-parametric approach for testing whether multiple outcomes consistently differ between arms in the same direction. Hotelling's T^2 or multivariate analysis of variance are parametric alternatives; however,

these approaches test for differences between arms in any direction leading to the possibility that a significant test may be driven by differences in opposing directions (i.e., favoring one arm for one endpoint and the other arm for a different endpoint). For this reason, O'Brien's test is preferred. Global tests are further described by Wassmer et al.^[18]

MISSING DATA

Missing data in clinical trials are ubiquitous for most (if not all) endpoints including PROs. Missing data are a problem for two main reasons. First, a reduced sample size impacts power for planned statistical tests. Second, missingness is often informative and can bias arm comparisons even if the rate of dropout between arms is similar. Much methodological work has been conducted to develop methods for handling missing PRO data.^[5,19] Patients miss assessments due to a variety of reasons including, but not limited to, missed appointments, hospitalizations, staff forgetting to administer the questionnaire, patient dropout due to no longer being ill or being too ill, and death. Patients may also miss individual questions within a questionnaire inadvertently or on purpose for reasons such as feeling too ill to complete all questions or perceiving items to be not applicable, difficult to understand, or uncomfortable to answer.

The best approach to handling missing data is to minimize missing data prospectively. This approach is far superior to all analytical methods for handling missing data “after the fact,” i.e., during the analysis phase. Approaches to minimize missing data prospectively include careful design, staff training, and real-time monitoring of data. PROs should be treated in the protocol with as much attention and importance as other clinical trial endpoints, including integrating PRO administration time points into the main study calendar. Also, individuals at each site should be identified to be responsible for all PRO aspects of the study. Administration should be standardized including training of site staff to check for skipped pages in paper-based questionnaires and having a back-up data collection method in the event of a missed questionnaire such as following up by telephone, mailing a back-up questionnaire, or administering a questionnaire at the next clinical visit. Ensuring professional-looking format, convenient timing, understandable and applicable questions, and feasible length of PROs can also improve compliance. Electronic PRO administration can also enhance compliance through automated patient reminders, allowing response at a time which is convenient to the patient, and real-time site notifications for missed assessments. Finally, real-time monitoring of PRO compliance can identify site-specific issues, which can be remedied to improve compliance rates moving forward. Strategies and design considerations for minimizing missing data are detailed by Fairclough^[5] and Little et al.^[20]

Data elements that will be required to understand or handle missing data should be identified. Reasons for missed PRO assessments should be collected in order to better understand the missing data mechanism. Other data elements include the date/time of the assessment, mode of administration (if questionnaires are being administered through more than one mode), other administration details including whether a proxy was used to capture the data, and auxiliary data if you plan to use it in your SAP to address missing data. Auxiliary data are variables that are associated with the outcome being modeled and are available even when the outcome is missing.^[21] For example, a physician's rating of a patient's Eastern Cooperative Oncology Group (ECOG) performance status, which is systematically collected at every visit in many cancer trials, could be considered as "auxiliary data" in modeling patient-reported physical functioning. A caregiver's or other proxy's rating for a given scale could also be used as auxiliary data.

For missing individual items within a questionnaire, scoring algorithms often allow for some missingness. Some algorithms require at least half the items in a scale be completed by the patient and rely on simple mean imputation. If the scoring algorithm does not specify how to handle individual missing items, the previously described approach may be appropriate depending on the nature of the scale. Alternative imputation strategies depending on the scale may include logical imputation. For example, in the Patient-Reported Outcomes version of the Common Terminology for Adverse Events (PRO-CTCAE, www.healthcaresdelivery.cancer.gov/pro-ctcae/), different items assess the frequency, severity, and interference of the symptoms. If the patient responds as not having the symptom of interest in the first question and omits responses to the follow-up questions about the same symptom, a reasonable approach would be to impute a value of 0 (i.e., no severity and no interference) for the missing items. See Fayers et al.^[22] for a description of other approaches as well as a framework for handling missing data at the item level.

For missing data at a scale level, a scale that is missing completely at random (MCAR) is one in which the probability of missingness is unrelated to its value. A scale that is missing at random (MAR) is one in which the probability of missingness only depends on past covariates or its past values. Finally, a scale that is missing not at random (MNAR) is one in which the probability of missingness is related to its own value even after accounting for past covariates and its past data values.

The first step in analysis is to understand the extent and patterns of missing data (Fig. 3) and the reasons for missingness. Baseline covariates can also be compared using standard statistical methods such as two-sample t-tests and chi-squared tests between patients with and without missing data for a given analysis to assess for bias. Likelihood of a missing assessment and time to dropout can also be modeled using logistic regression and time-to-event analysis, respectively.

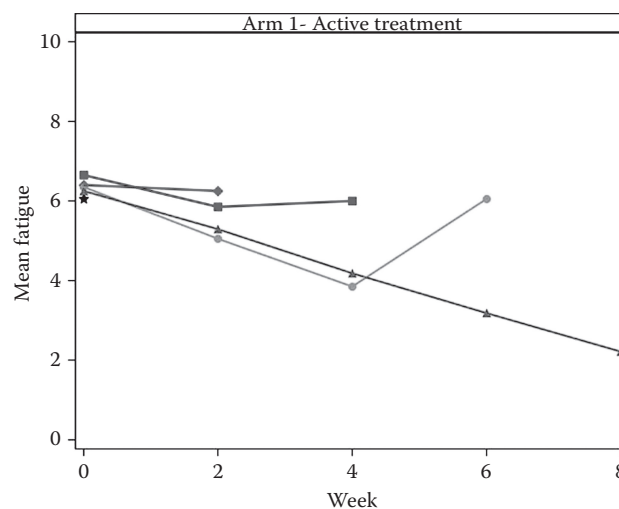


Fig. 3 Plot of mean fatigue scores for ARM = 1 stratified by last visit. A higher score represents worse fatigue. The triangle markers represent the subgroup of patients who completed the trial (N = 120). The circle, square, diamond, and star markers represent the subgroups (N = 20 each) of patients who completed six, four, two, and zero weeks of the trial, respectively. Based on this plot, patients who dropped out early tended to experience worsening after initial improvement or lack of improvement prior to dropout, whereas the patients who completed the trial on average had continual improvement over the course of the trial

The recommended analytical approach to handle missing data is to incorporate all patient data that is available and use a method which assumes MAR in primary analysis, followed by various sensitivity analyses which assume other missing data mechanisms. Results across sensitivity analyses should be descriptively compared to assess the robustness of trial results across varying missing data assumptions. Methods that assume MCAR should only be used when justified. These methods include complete case analyses as well as repeated univariate analyses, both of which include only patients with complete data. Similarly, excluding large numbers of patients (e.g., requiring a certain number of assessments during treatment) should be avoided.^[7]

Common methods that assume MAR include maximum likelihood modeling (e.g., previously described repeated measures or mixed effects growth curve models) and multiple imputation. Multiple imputation is a general framework where missing data values are filled in with likely values multiple times using a regression model and then results from these multiple imputed datasets are combined using Rubin's method.^[23] Multiple imputation is preferred to single imputation (impute a single value instead of multiple values) to ensure that variance is properly inflated to account for the added error due to imputation. However, imputation methods may decrease power and may not control the overall type I error rate to the desired level due to the variability associated with the imputed data. While MAR methods may result in some bias when data are not

missing at random, MAR methods are often robust when estimating differences between arms if observed covariates or earlier scores reflect the likelihood of subsequent missingness.

While missing data in PRO studies are often MNAR, several methods convert the MNAR problem into a lesser MAR problem. These methods include imputation approaches which incorporate auxiliary data (e.g., caregiver or proxy data) so that conditional on this auxiliary data the missing data are MAR. Pattern mixture models assume MAR within each pattern. Joint models assume MAR conditional on jointly modeled outcomes. An alternative approach, which is relatively less dependent on assumptions, involves searching for a tipping point that reverses the study conclusion.^[24]

OTHER ANALYTICAL CONSIDERATIONS

Stratification and clustering can be incorporated through appropriate model-based methods. Like other endpoints in clinical trials, the intent-to-treat approach (i.e., inclusion of all enrolled subjects with treatment analyzed according to the randomized assignment) should be considered the gold standard as analyses are unbiased due to randomization and estimate the true effectiveness of the treatment being tested. However, in some situations, per-protocol analysis (i.e., inclusion of enrolled subjects who receive treatment per protocol specifications) may be warranted, but is subject to selection bias so should be interpreted with extreme caution.

Additional PRO analyses incorporated into some trials include concurrent validation analyses (e.g., Mesa et al.^[25]). These analyses intend to assess or confirm PRO measurement properties using clinical trial data, e.g. when limited data exist within the given trial context. Typically, these analyses require supplemental data collection (e.g., patient global impression of severity or change scales).

Analysis of quality-adjusted life years (QALYs) is another analysis performed commonly as part of cost-utility or decision analysis and requires survival data and patient-reported health utilities such as those measured using EuroQOL's EQ-5D questionnaire. Health utility values tend to range from -1 to 1 (perfect health), with 0 representing death and negative values representing health states worse than death. QALY analysis weights the quantity of survival by the quality of survival to compute a single value for analysis across health states. The patient-level QALY can be computed per patient as the weighted sum of times spent at each utility level (see Billingham et al.^[26] for information about incorporating discounting of later benefits and/or handling change in utility measures between assessments).

When there is no censoring in the survival outcome, then statistical analysis can be based on usual methods such as a two-sample t-test. If there is censoring, then either QALYs

can be computed for each patient over the same fixed time period prior to any censoring or a model-based approach can be used to compute arm-level QALYs. Arm-level QALYs can be computed per arm using information from the Kaplan–Meier survival curve and repeated measures model of utility scores and then be compared between arms using a bootstrap or similar approach.^[27] Quality-adjusted time without symptoms or toxicity (Q-TWiST) method is a related method, which partitions survival into three health states.^[28] Beyond the PRO analyses described herein, many other correlative and supplemental analysis methods exist for assessing the relationships among PROs; mediating effects of PROs on other outcomes;^[7] or jointly modeling PRO data with other outcomes such as survival.^[29]

REPORTING OF RESULTS

Once analysis is complete, then results should be properly reported. The CONSORT-PRO extension expands on required elements in manuscripts of randomized trials to include key details pertaining to PROs.^[30] There are five additional checklist items specific to PROs beyond the standard CONSORT recommendations. These include that PROs be identified as a primary or secondary outcome in the abstract; PRO hypotheses be stated; measurement properties of the PRO questionnaire be provided or cited; statistical approach for dealing with missing data be explicitly stated; and PRO-specific limitations and generalizability be discussed. The modified CONSORT diagram (see supplemental content associated with the work of Calvert et al.^[30]) also includes specifying particular reasons why PRO assessments were not completed and the samples sizes for PRO analyses.

CONCLUSIONS

The vast majority (if not all) statistical methodologies described throughout this entry are not unique to PRO analyses, supporting the premise that development of a SAP for a PRO endpoint should be similar to that of any other clinical trial endpoint. Specifically, the plan should be based on the selection of an appropriate analytical method based on the stated hypothesis and distribution of the endpoint being analyzed with careful considerations for handling multiple testing and missing data. PRO questionnaires should be scored using the validated scoring algorithm and preliminary analysis should include descriptive statistics to understand trajectories of scores over time by arm; extent of missing data, its patterns, and causes; and potential selection biases in those patients who are providing PRO data.

For analysis of an ordinal or continuous PRO endpoint at a fixed time point, a recommended approach is to consider ANCOVA adjusting for baseline scores (when missing data

are not a real factor) or longitudinal modeling (with the estimate made at the fixed time point of interest). Parametric approaches are generally acceptable for ordinal PRO data even when sample size is modest. Summary measures and statistics, while seemingly useful due to allowing simplified analysis and interpretation of a single score, should be used judiciously as such measures may be impacted by missing data “behind the scenes.” In general, responder analysis should be considered as supplementary analysis to augment analysis of the full original scale. Regression-based analyses for longitudinal PRO data are effective and flexible strategies to analyze both within-arm and between-arm changes over time with models incorporating time as either a discrete or continuous event. Unstructured covariance is a reasonable approach in most cases.

Multiplicity is most commonly addressed through specification of a hierarchy of endpoints, either within all clinical endpoints of the trial for strict control or within the subset of PRO endpoints for less strict control depending on the setting. The singly best approach to handle missing data is prospective minimization of missing data. The recommended method to handle missing data once reaching the analysis phase is to use all available patient data along with a primary analysis method, typically one that assumes MAR, like repeated measures or random coefficient models, with sensitivity analyses performed. CONSORT-PRO extension should be followed for judicious reporting from clinical trials.

REFERENCES

1. U.S. Food and Drug Administration, U.S. National Institutes of Health. *BEST (Biomarkers, EndpointS, and other Tools) Resource*; Silver Spring, MD, 2016.
2. U.S. Food and Drug Administration. *Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims [Internet]*; Silver Spring, MD, 2009.
3. European Medicines Agency. *Reflection Paper on the Regulatory Guidance for the Use of Health-Related Quality of Life (HRQL) Measures in the Evaluation of Medicinal Products*; 2005.
4. Reeve, B.B.; Wyrwich, K.W.; Wu, A.W.; Velikova, G.; Terwee, C.B.; Snyder, C.F.; Schwartz, C.; Revicki, D.A.; Moinpour, C.M.; McLeod, L.D.; Lyons, J.C.; Lenderking, W.R.; Hinds, P.S.; Hays, R.D.; Greenhalgh, J.; Gershon, R.; Feeny, D.; Fayers, P.M.; Cella, D.; Brundage, M.; Ahmed, S.; Aaronson, N.K.; Butt, Z. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Qual. Life Res.* **2013**, *22*, 1889–1905.
5. Fairclough, D.L. *Design and analysis of quality of life studies in clinical trials*. 2nd Ed.; Chapman & Hall/CRC: Boca Raton, FL, 2010.
6. Norman, G. Likert scales, levels of measurement and the “laws” of statistics. *Adv. Health Sci. Educ. Theory Pract.* **2010**, *15*, 625–632.
7. Cappelleri, J.C.; Zou, K.H.; Bushmakina, A.G.; Alvir, J.M.J.; Alemayehu, D.; Symonds, T. *Patient-Reported Outcomes: Measurement, Implementation and Interpretation*; Chapman & Hall/CRC Press: Boca Raton, FL, 2013.
8. Vickers, A.J. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: A simulation study. *BMC Med Res. Methodol.* **2001**, *1*, 6.
9. Sullivan, L.M.; D’Agostino, R.B. SrRobustness and power of analysis of covariance applied to ordinal scaled data as arising in randomized controlled trials. *Stat. Med.* **2003**, *22*, 1317–1334.
10. Bell, M.L.; King, M.T.; Fairclough, D.L. Bias in area under the curve for longitudinal clinical trials with missing patient reported outcome data: Summary measures versus summary statistics. *SAGE Open* **2014**, *4*.
11. Wyrwich, K.W.; Norquist, J.M.; Lenderking, W.R.; Acaster, S. Methods for interpreting change over time in patient-reported outcome measures. *Qual. Life Res.* **2013**, *22*, 475–483.
12. Dworkin, R.H.; Turk, D.C.; Peirce-Sandner, S.; Baron, R.; Bellamy, N.; Burke, L.B.; Chappell, A.; Chartier, K.; Cleeland, C.S.; Costello, A.; Cowan, P.; Dimitrova, R.; Ellenberg, S.; Farrar, J.T.; French, J.A.; Gilron, I.; Hertz, S.; Jadad, A.R.; Jay, G.W.; Kalliomaki, J.; Katz, N.P.; Kerns, R.D.; Manning, D.C.; McDermott, M.P.; McGrath, P.J.; Narayana, A.; Porter, L.; Quessy, S.; Rappaport, B.A.; Rauschkolb, C.; Reeve, B.B.; Rhodes, T.; Sampaio, C.; Simpson, D.M.; Stauffer, J.W.; Stucki, G.; Tobias, J.; White, R.E.; Witter, J. Research design considerations for confirmatory chronic pain clinical trials: IMMPACT recommendations. *Pain* **2010**, *149*, 177–193.
13. Benzo, R.; Farrell, M.H.; Chang, C.C.; Martinez, F.J.; Kaplan, R.; Reilly, J.; Criner, G.; Wise, R.; Make, B.; Luke-tich, J.; Fishman, A.P.; Sciurba, F.C. Integrating health status and survival data: The palliative effect of lung volume reduction surgery. *Am. J. Respir. Crit. Care Med.* **2009**, *180*, 239–246.
14. Fitzmaurice, G.M.; Laird, N.M.; Ware, J.H. *Applied Longitudinal Analysis*. 2nd Ed.; John Wiley & Sons: Hoboken, NJ, 2011.
15. Mallinckrodt, C.H.; Lane, P.W.; Schnell, D.; Peng, Y.; Mancuso, J.P. Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Inform. J.* **2008**, *42*, 303–319.
16. Littell, R.C.; Pendergast, J.; Natarajan, R. Modelling covariance structure in the analysis of repeated measures data. *Stat. Med.* **2000**, *19*, 1793–1819.
17. Dmitrienko, A.; Tamhane, A.C.; Bretz, F. *Multiple Testing Problems in Pharmaceutical Statistics*; Chapman & Hall/CRC: Boca Raton, FL, 2010.
18. Wassmer, G.; Reitmeir, P.; Kieser, M.; Lehmacher, W. Procedures for testing multiple endpoints in clinical trials: An overview. *J. Stat. Plan. Infer.* **1999**, *82*, 69–81.
19. Bell, M.L.; Fairclough, D.L. Practical and statistical issues in missing data for longitudinal patient-reported outcomes. *Stat. Methods Med. Res.* **2014**, *23*, 440–459.
20. Little, R.J.; D’Agostino, R.; Cohen, M.L.; Dickersin, K.; Emerson, S.S.; Farrar, J.T.; Frangakis, C.; Hogan, J.W.; Molenberghs, G.; Murphy, S.A.; Neaton, J.D.; Rotnitzky, A.; Scharfstein, D.; Shih, W.J.; Siegel, J.P.; Stern, H. The

- prevention and treatment of missing data in clinical trials. *New Engl. J. Med.* **2012**, *367*, 1355–1360.
21. Wang, C.; Hall, C.B. Correction of bias from non-random missing longitudinal data using auxiliary information. *Stat. Med.* **2010**, *29*, 671–679.
 22. Fayers, P.M.; Curran, D.; Machin, D. Incomplete quality of life data in randomized trials: Missing items. *Stat. Med.* **1998**, *17*, 679–696.
 23. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; John Wiley & Sons: New York, 1987.
 24. Ratitch, B.; O'Kelly, M.; Tosiello, R. Missing data in clinical trials: From clinical assumptions to statistical analysis using pattern mixture models. *Pharm. Stat.* **2013**, *12*, 337–347.
 25. Mesa, R.A.; Gotlib, J.; Gupta, V.; Catalano, J.V.; Deininger, M.W.; Shields, A.L.; Miller, C.B.; Silver, R.T.; Talpaz, M.; Winton, E.F.; Harvey, J.H.; Hare, T.; Erickson-Viitanen, S.; Sun, W.; Sandor, V.; Levy, R.S.; Kantarjian, H.M.; Verstovsek, S. Effect of ruxolitinib therapy on myelofibrosis-related symptoms and other patient-reported outcomes in COMFORT-I: A randomized, double-blind, placebo-controlled trial. *J. Clin. Oncol.* **2013**, *31*, 1285–1292.
 26. Billingham, L.J.; Abrams, K.R.; Jones, D.R. Methods for the analysis of quality-of-life and survival data in health technology assessment. *Health Technol. Assess.* **1999**, *3*, 1–152.
 27. Khan, I. *Design & Analysis of Clinical Trials for Economic Evaluation & Reimbursement: An Applied Approach Using SAS & STATA*; Chapman & Hall/CRC Biostatistics Series: Boca Raton, FL, 2016.
 28. Glasziou, P.P.; Cole, B.F.; Gelber, R.D.; Hilden, J.; Simes, R.J. Quality adjusted survival analysis with repeated quality of life measures. *Stat. Med.* **1998**, *17*, 1215–1229.
 29. Ibrahim, J.G.; Chu, H.; Chen, L.M. Basic concepts and methods for joint models of longitudinal and survival data. *J. Clin. Oncol.* **2010**, *28*, 2796–2801.
 30. Calvert, M.; Blazeby, J.; Altman, D.G.; Revicki, D.A.; Moher, D.; Brundage, M.D. Reporting of patient-reported outcomes in randomized trials: The CONSORT PRO extension. *J. Am. Med. Assoc.* **2013**, *309*, 814–822.

Percentile Charts on Correlated Measures

Xuming He

Department of Statistics, University of Illinois, Champaign, Illinois, U.S.A.

Kai W. Ng

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

Abstract

This entry reviews two major approaches for estimating percentile curves of single variables, discusses the issue of joint estimation with two or more correlated variables (e.g., blood pressures and measures on four Serum Immunoglobulin G subclasses), and points the readers to more recent work on conditional growth charts.

Keywords: Box-Cox transformation; Growth charts; Percentile curves; Quantile regression.

INTRODUCTION

Percentile (or centile) charts are often used in health monitoring, medical screening, and clinical studies. In typical applications, a health professional compares certain physical or medical measurements taken from a particular subject with age-specific reference curves to detect extreme readings. The construction of reference curves is generally made from a sample of relevant subjects. When a very large sample is readily available, a purely nonparametric method may be employed by calculating various percentiles for each age group of interest. In most applications, however, we are limited in sample sizes and cannot obtain good estimates of the extreme percentiles at all ages without additional assumptions. In this paper, we review two major approaches for estimating percentile curves of single variables, discuss the issue of joint estimation with two or more correlated variables (e.g., blood pressures and measures on four Serum Immunoglobulin G subclasses), and point the readers to more recent work on conditional growth charts.

BOX-COX TRANSFORMATION

The Box-Cox transformation to normality is one approach commonly used in estimating conditional quantile curves. Suppose that y is the variable of interest. The percentile charts often consist of several percentile curves of y as a function of time. In general, we consider $Q_\tau(x)$, the τ -th conditional quantile of y given a set of covariates x , where τ is a value between 0 and 1, indicating a desired percentile level. The Box-Cox transformation operates under the assumption that $y(\hat{a}) = (y^{\hat{a}} - 1)/\hat{a}$ is normally distributed with mean μ_x and variance σ^2 for some unknown power

parameter $\hat{a}$. The special case of $\hat{a} = 0$ is taken to mean a log transformation. It is then easy to obtain that

$$Q_\tau(x) = ((\mu_x + \sigma z_\tau)\hat{a} + 1)^{1/\hat{a}},$$

where z_τ is the τ -th quantile of a standard normal distribution. If $\hat{a} = 0$, we simply have $Q_\tau(x) = \exp(\mu_x + \sigma z_\tau)$.

The literature is rich on the Box-Cox transformation to normality. Theoretically, $y(\lambda)$ cannot be exactly normal except in the special cases of $\lambda = 0$ or 1. In practice, we expect the transformed variable to be approximately normal. Maximizing the normal likelihood leads to a transformation that renders $U(\hat{a})$ as close to normal as possible under the Kullback-Leibler information number, see, [1–3]

For example, consider serum IgG (Immunoglobulin G) subclass concentrations for children. Recent developments on radial immunodiffusion using monoclonal antibodies have made possible rather accurate measures on four known subclasses of IgG (assigned IgG 1–4). Deficiencies of one or more IgG subclasses are associated with recurrent infections in children, see, [4–6] As the scope of IgG subclass measurements as a diagnostic tool has increased since the early 1980s, health professionals in various countries have found it useful to establish the percentile ranges for their own populations, see, [7–10] for example. All these researchers employed Box-Cox transformations in their work with $\mu_x = x'\beta$ for some unknown parameter β .

The power transformation approach has been generalized by Cole, [11] and Cole and Green, [12] allowing the transformation parameters to be a function of the covariate (e.g., age). It is based on the assumption that $(y^{l(x)} - m(x))/s(x)$ is approximately $N(0, 1)$ for three unknown but smooth functions $l(x)$, $m(x)$, and $s(x)$, which can be estimated by the penalized likelihood method. This is called the LMS

method, with the three letters of the name referring to the functions $l(x)$, $m(x)$, and $s(x)$. One attraction of this LMS method is that it allows local transformations that are age-dependent. Wright and Royston^[13] compared several statistical methods. They also extended the transformation method by using fractional polynomials in Ref. [14].

QUANTILE REGRESSION

Quantile regression of Koenker and Bassett^[15] provides another way to estimate percentile curves without any distributional assumptions on the response. For each level τ we assume that $Q\tau(x)$ is a parametric or non-parametric function of x . Quantile curve-fitting techniques can be found in Refs. [16,17]. This approach can be especially useful if a power transformation towards normality is unrealistic. A recent demonstration of this approach using spline approximations to smooth curves can be found in Ref. [18], where it was shown that features such as bimodality in the data will be better captured with quantile regression.

A full-scale comparison between the transformation approach and the quantile regression approach is lacking; however, both methods have been quite well studied so that the following points can be made. A transformation-based method leads to lower variance in estimating extreme percentiles. The quantile regression method does not assume any distributional information globally, so it has higher variability when τ is close to 0 or 1. On the other hand, it may be

difficult to judge whether a transformation of y has successfully made it close to being normally distributed, especially in the tails, so the lower variability in the extreme quantiles may have come at the cost of larger bias. The quantile regression approach has the advantage that the estimated τ -th quantile will divide the sample into the desired τ : $(1 - \tau)$ ratio. Since the transformation method estimates certain global parameters from the sample, it is less robust against outliers. A combination of transformation and linear quantile models has been studied in Ref. [19], but this method has not been used in the construction of percentile charts.

CORRELATED RESPONSE VARIABLES

All the work cited above focused on percentile curve estimation for one variable at a time ignoring information possibly contained in available variables of equal interest. In the IgG literature such as those ^[7-10] cited above, marginal estimation was used for each IgG subclass without consideration of the correlations among those subclass measurements. In other applications involving, say, weight and height or systolic and diastolic readings in blood pressure, the correlation can be quite significant. He, Ng, and Shi^[20] investigated whether the common practice of separate percentile curve construction, referred to as the *marginal* method, can be improved upon by *joint* estimation that takes possible correlation among variables into account. Figure 1 is a graph of the root mean

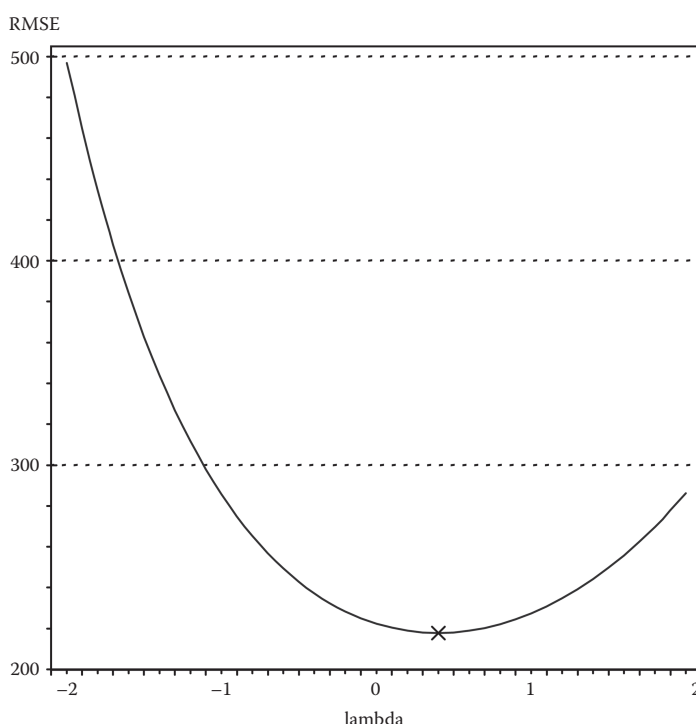


Fig. 1 Box-Cox transforms for dependent variable IgG1 in a quadratic regression of IgG1 on age: root mean square error against λ . The smallest value of RMSE is achieved at $\lambda = 0.40$

square error (RMSE) of the quadratic regression of IgG1 on age against different value of λ for the Example 1 of.^[20] The optimal λ in the figure is $\lambda = 0.40$, which was adopted in univariate Box-Cox transform in Ref. [20].

When there are several variables of interest, percentile contours may be defined in the multivariate context, and indeed there has been active research in this direction.^[21,22] Still, screening by individual variables (e.g., IgG subclasses) remains relevant and percentile regions of common forms do not directly reduce to a valid percentile interval. In this section, we follow^[20] to focus on percentile charts of individual variables based on the Box-Cox transformation to joint (multivariate) normality and compare this with the marginal method. Other tools will be briefly discussed in the concluding section.

For multivariate data without covariates, several authors have addressed this issue with partial conclusions. Johnson and Wichern^[23] represented an extended belief that it is not so worthwhile to maximize the joint normal likelihood given the additional cost of computing, which was certainly an important factor in those earlier years. In a somewhat related context, Koenker and Portnoy^[24] reached a similar conclusion for estimating regression medians. On the other hand, Velilla^[25,26] demonstrated that the joint approach leads to better efficiency and better diagnostics. For the bivariate case, the asymptotic efficiency of a joint estimate relative to a marginal estimate can be as high as 1.5.

To provide a clearer understanding of how the joint and the marginal methods compare, we follow^[20] to present in a general set-up for the multivariate Box-Cox transformation model in regression.

We consider p response variables $y_1, \dots, y_p$ that satisfy the following modeling assumptions:

$$y_i(\hat{\alpha}_i) = x' \beta_i + e_i, \quad i = 1, \dots, p \quad (1)$$

where x is the m -dimensional covariate, $\beta_i \in R^m$ are unknown parameters, and $(e_1, \dots, e_p)$ are expected to be close to multivariate normal with mean zero and variance covariance matrix Σ . The parameters $\theta_i = (\hat{\alpha}_i, \beta_i, \sigma_i)$ have to be estimated based on a sample of n observations $(x_j, y_{ij})(i = 1, \dots, p, j = 1, \dots, n)$.

Let Y_j denote the column vector of $(y_{1j}, \dots, y_{pj})$, $B = (\beta_1, \dots, \beta_p)$, and θ denote the vector of all parameters concerned. The joint log-likelihood (up to a constant) is

$$l_j(\theta) = -\frac{1}{2} \sum_{j=1}^n (Y_j(\hat{\alpha}_i)' - x_j' B) \sum_{i=1}^p (Y_j(\hat{\alpha}_i)' - x_j' B)' - \frac{n}{2} \log \left| \sum \right| + \sum_{i=1}^p \sum_{j=1}^n (\hat{\alpha}_i - 1) \log y_{ij}. \quad (2)$$

By forcing Σ to be a diagonal matrix in (2), we can decompose the log-likelihood as a sum (over i) of the marginal log-likelihood for the i th variable

$$l_j(\theta_i) = -\frac{1}{2\sigma_i^2} \sum_{j=1}^n (Y_{ij}(\hat{\alpha}_i) - x_j' \beta_i)^2 - n \log(\sigma_i) + (\hat{\alpha}_i - 1) \sum_{j=1}^n \log y_{ij}. \quad (3)$$

The marginal method maximizes $l_i(\theta_i)$ for each i to find the transformation. For both methods, the estimators of θ are asymptotically normal, but the asymptotic efficiencies differ. For example, we consider a simple logarithmic model with $p = 2$, $m = 1$, $E(x) = 0$, $Var(x) = 1$, $(\hat{\alpha}_1, \hat{\alpha}_2, \beta_1, \beta_2, \sigma_1, \sigma_2, \rho) = (0, 0, 0, 0, 1, 1, \rho)$, and $(\log y_1, \log y_2) = (e_1, e_2)$ being bivariate normal. The asymptotic efficiency of the marginal estimate relative to the joint estimate is equal to $7(7 - 4\rho^2)/(49 - 8\rho^2 + 4\rho^2)$ for $\hat{\alpha}$, and $(2 - \rho^2)$ for σ . As ρ^2 approaches 1, the asymptotic relative efficiency tends to $7/15 \approx 0.47$ for the λ parameters and $1/2 = 0.50$ for the σ parameters. There is no improvement in estimating the β parameters in this case.

In general, the information matrix of the model parameters, including the transformation power, depends on the model in a rather complicated fashion but the comparison of statistical efficiency can be made case by case. Under the special case of a simple logarithmic model, it can be shown that the efficiency gain from the joint method in estimating the transformation parameters can be quite noticeable when the correlation between the transformed responses is significant. The efficiency gain is negligible for percentile curve estimation around the median but becomes more pronounced for more extreme percentiles. A simulation study in section “Correlated Response Variables” of^[20] demonstrated this general pattern.

AN EXAMPLE

Blood pressure is one of the important health indicators for adults. In developed countries, hypertension affects one in four human subjects. Screening of blood pressure is routinely done.

Since blood is carried from the heart to all of the body's tissue and organs in vessels called arteries, blood pressure is the force of the blood pushing against the walls of those arteries. Blood pressure is at its greatest when the heart contracts and is pumping the blood. This is called systolic pressure. When the heart is at rest, in between beats, blood pressure falls. This is the diastolic pressure. It is well known that the systolic and diastolic pressure readings are quite highly correlated. In addition, there is a higher chance for older people to suffer from high blood pressure.

The example used in^[20] is based on data from 4695 male adults (from 30–80 years old) from the 1998 Health Survey of England (<http://www.data-archive.ac.uk>). The authors use their systolic and diastolic pressures to construct percentile charts as a tool to understand blood pressure changes in the general male population (including

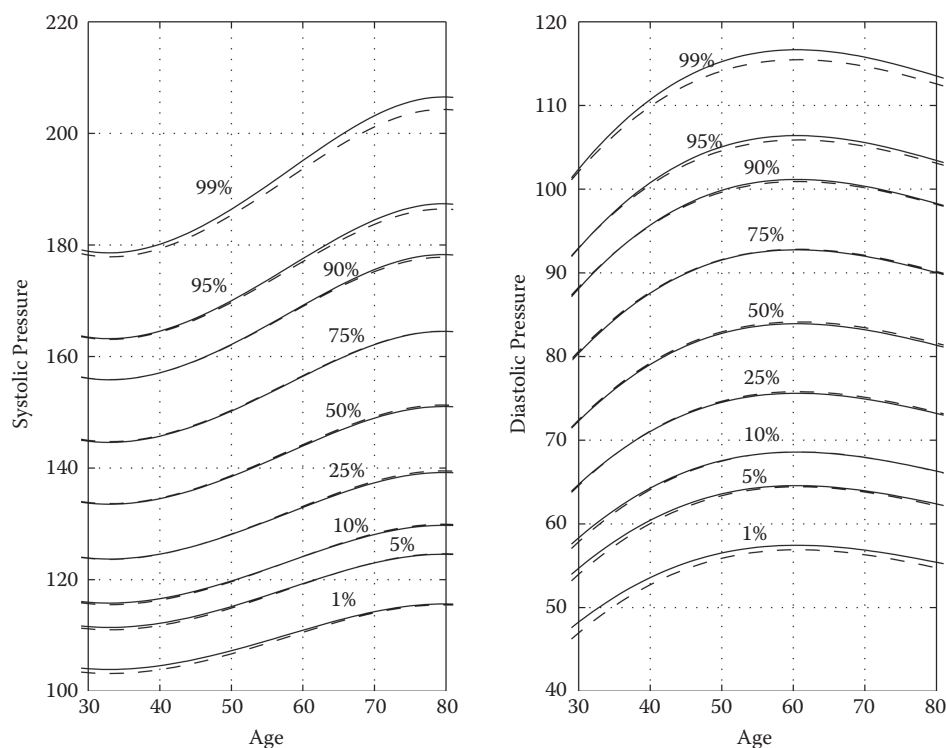


Fig. 2 Percentile curves for the blood pressure of English male adults. Solid curves are obtained from the marginal method, and the dashed ones are from the joint method

those with hypertension). A joint power transformation to normality can be considered for those two correlated blood pressure readings.

Using cubic polynomials as percentile curves, the marginal method gave a transformation power $\hat{a}_1 = -0.55$ (with standard error, $SE = 0.12$) for the systolic pressure and $\hat{a}_2 = 0.40$ (with $SE = 0.09$) for the other. The joint method however led to $\hat{a}_1 = -0.37$ ($SE = 0.10$) and $\hat{a}_2 = 0.60$ ($SE = 0.07$). The correlation between the transformed variables is about 0.7 in this case. We see that the λ estimates from the two methods are more than two SEs apart. The resulting percentile charts (displaying curves at 1%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, 99% levels) from both methods are shown in Fig. 2. Note that for extreme quantiles, the joint method produces slightly lower curves with the maximum difference of 2 mm of mercury. A medical doctor can use the curves to say that a male patient at age of 60 with a diastolic blood pressure reading of 107 is at the 95 percentile of the general male population.

CONCLUSION

Age-specific percentile curves can be computed one at a time even when there are two or more related variables of concern. For the Box-Cox transformation (towards normality), we have the option of the joint normal

likelihood method in addition to the marginal method. Albeit these two methods often produce similar charts, the joint method can lead to a noticeable efficiency gain for extreme quantile estimation, especially when the correlation among the variables is greater than half. In addition, there is an added diagnostic value for using both the marginal and the joint methods for the same problem. A substantial difference in the resulting percentile curves is indicative of severe model inadequacy or failure of power transformations to the intended normality. In such cases, other methods such as those reviewed earlier in the paper may be called for.

The comparison here between the joint and the marginal estimation was limited to the use of the Box-Cox transformation. However, our findings may serve as a reference point for users who may be interested in studying joint percentile curve estimation using a more flexible or advanced statistical method.

Recently, it has been shown in^[27] that conditional growth charts are valuable in screening current status based on individual history. The regression quantile approach can be naturally extended to percentile charts conditional on past measurements or some other covariates. Bivariate and multivariate quantile contours can be constructed using a method recently developed in Ref. [22], and research in this direction might accelerate in the future.

ACKNOWLEDGMENTS

This research was support in part by NSF award DMS-0604229 (United States), and NNSF of China Grant 10828102.

REFERENCES

- Atkinson, A.C. Testing transformations to normality. *J. R. Statist. Soc. B* **1973**, *35*, 473–479.
- Hernandez, F.; Johnson, R.A. The large-sample behavior of transformations to normality. *J. Am. Stat. Assoc.* **1980**, *75*, 855–861.
- Chen, G.; Lockhart, R.A. Box-Cox transformed linear models: A parameter-based asymptotic approach. *Can. J. Stat.* **1997**, *25*, 517–529.
- Beck, C.S.; Heiner, D.C. Selective IgG4 deficiency and recurrent infections of respiratory tract. *Ann. Rev. Respir. Dis.* **1981**, *124*, 84–93.
- Plebani, A.; Duse, M.; Monafò, V. Recurrent infections with IgG2 deficiency. *Arch. Dis. Child.* **1985**, *60*, 670–672.
- Umetsu, D.T.; Ambrosino, D.M.; Quinti, I.; Siber, G.R.; Geha, R.S. Recurrent sinopulmonary infection and impaired antibody response to bacterial capsular polysaccharide antigen in children with selective IgG-subclass deficiency. *N. Engl. J. Med.* **1985**, *313*, 1247–1251.
- Beard, L.J.; Ferrante, A.; Hangdorn, J.F.; Leppard, P.; Kiroff, G. Percentile ranges for IgG subclass concentrations in healthy Australian children. *Pediatr. Infect. Dis. J.* **1990**, *9*, S9–S15.
- Bird, D.; Duffy, S.; Isaacs, D.; Webster, A.D.B. Reference ranges for IgG subclasses in preschool children. *Arch. Dis. Child.* **1985**, *60*, 204–207.
- Plebani, A.; Ugazio, A.G.; Avanzini, M.A.; Massimi, P.; Zonta, L.; Monafò, V.; Burgio, G.R. Serum IgG subclasses concentration in healthy subjects at different age: age normal percentile charts. *Eur. J. Pediatr.* **1989**, *149*, 164–167.
- Lau, Y.L.; Jones, B.M.; Ng, K.W.; Yeung, C.Y. Percentile ranges for serum IgG subclass concentrations in healthy Chinese children. *Clin. Exp. Immunol.* **1993**, *91*, 337–341.
- Cole, T.J. Fitting smoothed centile curves to reference data (with discussion). *J. R. Stat. Soc. A* **1988**, *151*, 385–418.
- Cole, T.J.; Green, P.J. Smoothing reference centile curves. *Stat. Med.* **1992**, *11*, 1305–1319.
- Wright, E.M.; Royston, P. A comparison of statistical methods for age-related reference intervals. *J. R. Stat. Soc. A* **1997**, *160*, 47–69.
- Wright, E.M.; Royston, P. Simplified estimation of age-specific reference intervals for skewed data. *Stat. Med.* **1997**, *16*, 2785–2803.
- Koenker, R.; Bassett, G. Regression quantiles. *Econometrica* **1978**, *46*, 33–50.
- Koenker, R.; Ng, P.; Portnoy, S. Quantile smoothing splines. *Biometrika* **1994**, *94*, 673–680.
- He, X. Quantile curves without crossing. *Am. Statist.* **1997**, *51*, 186–192.
- Wei, Y.; Pere, A.; Koenker, R.; He, X. Quantile regression methods for reference growth curves. *Stat. Med.* **2006**, *25*, 1369–1382.
- Mu, Y.; He, X. Power transformation towards a linear regression quantile. *J. Amer. Statist. Assoc.* **2007**, *102*, 269–279.
- He, X.; Ng, K.; Shi, J. Marginal versus joint Box-Cox transformation with applications to percentile curve construction for IgG subclasses and blood pressures. *Stat. Med.* **2003**, *22*, 397–408.
- Chaudhuri, P. On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.* **1996**, *91*, 862–872.
- Wei, Y. An approach to multivariate covariate-dependent quantile contours with application to bivariate conditional growth charts. *J. Amer. Statist. Assoc.* **2008**, *103*, 397–409.
- Johnson, R.A.; Wichern, D.W. *Applied Multivariate Statistical Analysis*; Prentice-Hall: Englewood Cliffs, NJ, 1982.
- Koenker, R.; Portnoy, S. M estimation of multivariate regressions. *J. Amer. Statist. Assoc.* **1990**, *85*, 1060–1068.
- Velilla, S. A note on the multivariate Box-Cox transformation to normality. *Stat. Prob. Lett.* **1993**, *17*, 259–263.
- Velilla, S. Diagnostics and robust estimation in multivariate data transformations. *J. Amer. Statist. Assoc.* **1995**, *90*, 945–951.
- Wei, Y.; He, X. Conditional growth charts (with discussions). *Ann. Statist.* **2006**, *34*, 2069–2097.

Personalized Medicine: Validation of Diagnostic Medical Devices

Meijuan Li, PhD, Gene Pennello, PhD, Jincao Wu, PhD, and Laura Yee, PhD

Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Zhiwei Zhang, PhD

Department of Statistics, University of California Riverside, Riverside, California, U.S.A.

Gerry Gray, PhD

Data-Fi LLC, Silver Spring, Maryland, U.S.A.

Yuying Jin, PhD, Yaji Xu, PhD

Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Jingjing Ye, PhD

Division of Biometrics V, Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Zhiheng Xu, PhD

Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

This entry provides general considerations for the validation of diagnostic medical devices whose results have direct or indirect potential to “personalize” patient care. In particular, the clinical validation of companion diagnostic devices and corresponding therapeutic products, risk prediction devices, and monitoring devices are discussed. In addition, the often neglected topic of analytically validating the biomarker measurement(s) made by a diagnostic device, including measurement accuracy and reproducibility is discussed.

Keywords: Precision medicine; Biomarker; Measurement performance; Clinical validation; Treatment selection; Risk prediction; Monitoring; Diagnostic device.

INTRODUCTION

“Personalized medicine” or “precision medicine” has been described in grand terms as treating “the right patient with the right drug at the right dose at the right time.”^[1] The National Academy of Sciences defines precision medicine as “the use of genomic, epigenomic exposure and other data to define individual patterns of disease, potentially leading to better individual treatment.”^[2] The fervor over the potential of precision medicine to optimize patient care is warranted, due to not only advances in high-throughput molecular profiling (e.g., next generation genomic sequencing) but also to corresponding regulatory initiatives.^[3] Operationally, medical care is “personalized” by applying a medical model to stratify patients into different subpopulations, in efforts to tailor or “individualize” patient care.^[1,2] To define the patient strata, a predictive model can be used, for example, likely response

or adverse reaction to a treatment, risk of incident disease in a non-diseased subject, or risk of hospitalization, disease progression, relapse, or recurrence in a diseased patient.^[4] Predictions of future event (risk predictions), treatment response, or treatment adverse reaction are typically based on one or more biomarkers,^[5] measured or detected by diagnostic devices (tests) or by some other measurement procedure.

An early example of personalized medicine is the breast cancer treatment *Trastuzumab* (Herceptin), indicated for patients whose breast tumor tissue overexpresses human epidermal growth factor receptor 2 (HER-2). In 1998, the Food and Drug Administration (FDA) approved Herceptin together with an immunohistochemical (IHC) In Vitro diagnostic (IVD) HercepTest™ for measuring HER-2 expression.^[6] HercepTest™ is considered one of the first in vitro “companion diagnostic” devices, that is, an IVD test that is essential for the safe and effective use

of the corresponding therapeutic product.^[7] Since the FDA approvals of Herceptin and HercepTest™, interest in identifying candidate biomarkers in personalized medicine has grown tremendously. Other successful examples of therapeutic products with accompanying IVD companion diagnostics include Venclexta® (venetoclax; AbbVie Inc.) and Vysis Cll Fish Probe kit (Abbott Molecular Inc.), Erbitux (cetuximab; ImClone Systems) and *therascreen* KRAS RGQ PCR kit (Qiagen Manchester Ltd.).^[6] FDA has released draft guidance on principles for codevelopment of an in vitro companion diagnostic device (CDx) with a therapeutic product.^[8]

While the term “precision medicine” or “personalized medicine” is often interpreted as selecting the treatment best suited for an individual patient, the term is used more broadly in this entry. Specifically, personalized medicine is referred to as encompassing genetic and other available information that could be used to influence individual medical decision-making, including safety or effectiveness of a treatment or intervention, risk of incident disease, prognosis of a diseased patient either untreated or on a standard treatment, and the value of a clinically meaningful biomarker monitored over time.

Biomarkers can be classified according to their intended use (IU). Predictive biomarkers are used to predict treatment response or to inform treatment selection, usually based on a differential treatment effect between the biomarker-negative and biomarker-positive populations. Ideally, the biomarker by treatment interaction is nearly qualitative, which may occur when the biomarker is the target of the treatment. An in vitro companion diagnostic device (or test) is an IVD device and could be essential for the safe and effective use of a corresponding therapeutic product to identify: 1) patients who are most likely/likely to benefit from a particular therapeutic product or patients who are eligible for a particular therapeutic product; 2) patients likely to be at increased risk of serious adverse reactions as a result of treatment with a particular therapeutic product; and 3) response to treatment for the purpose of adjusting treatment to achieve improved safety or effectiveness.^[7] Based on how it is validated clinically, there are two main types of indications for use or claims based for a companion diagnostic device^[7] which detects or measures predictive biomarker, namely, predictive test and selection test. Predictive tests are clinically validated by demonstrating the differential treatment effect between marker-positive and marker-negative patients defined by the device by treatment interaction.^[8] This type of clinical evidence is generally demonstrated from the clinical trials in which both test-positive and test-negative subjects are enrolled. Selection tests are clinically validated by demonstrating the treatment effect in marker-positive (or marker-negative) patients defined by the device. Often for selection tests, only test-positive (or test-negative) subjects are selected for enrollment in a clinical trial. For a selection test, if the treatment efficacy

trial demonstrates adequate safety and effectiveness of the therapeutic product within the population selected by the test, then the test is considered to be “clinically validated”. In that, it selected a population that benefits from therapy.^[7]

Prognostic biomarkers are used for predicting a future outcome for patients either untreated or on a standard of care therapy. Many biomarkers, such as hormone-receptor status in breast cancer, are both prognostic and predictive. Many types of biomarkers exist, including proteins, nucleic acids, antibodies, and peptides. A biomarker can also be a collection of alterations, such as gene expression, proteomic, and metabolomic signatures. Biomarkers can be detected in the circulation (whole blood, serum, or plasma) or excretions or secretions (stool, urine, sputum, saliva, or nipple discharge) or can be tissue-derived such as formalin-fixed, paraffin-embedded biopsied tissue. Genetic biomarkers can be inherited, e.g., sequence variations in germ line DNA isolated from blood, or can be somatic, e.g., somatic DNA alterations in human cancer.

In this entry, we distinguish between the biomarker and the biomarker assay, i.e., the reagents, instruments, systems, apparatuses, or platforms used to measure or detect the biomarker, which may be referred to as the diagnostic device or test. To ensure that a new biomarker introduced into clinical practice will appropriately direct patient management, the diagnostic device measuring/detecting the biomarker should undergo rigorous evaluation, including analytical and clinical validation, prior to its incorporation into patient management. Here, validation of a candidate device refers to an evaluation of its analytical and clinical performance in the study independent from its development. A candidate device is said to be validated if it meets a set of analytical and clinical performance goals (usually with statistical significance). The performance goals are chosen such that, if met, and the device if used per instructions for use, should provide clinically significant results in a significant portion of its target population. Broadly, validation of a diagnostic device includes an assessment of its safety and effectiveness and whether its probable benefits outweigh any probable risks.

As discussed above, the IU of predictive, prognostic, and monitoring biomarker devices are quite different, implying different requirements for clinical validation. In the section “Measurement Performance Validation of Diagnostic Devices,” the validation of the measurement(s) made by diagnostic devices, e.g., their accuracy and reproducibility is discussed in general. In the sections “Treatment Selection: Clinical Validation of a Companion Diagnostic and the Corresponding Therapeutic Product” and “Risk Prediction”, the clinical validation of diagnostic devices for treatment selection, risk prediction, and monitoring are specifically discussed. In section the “Conclusion,” some concluding remarks are provided.

MEASUREMENT PERFORMANCE VALIDATION OF DIAGNOSTIC DEVICES

Results from a diagnostic device are often essential in patient treatment and management. Inadequate performance of a diagnostic device could have severe therapeutic consequences. For example, for a CDx, erroneous diagnostic device results could lead to withholding appropriate therapy or to administering inappropriate therapy. When a new biomarker is being developed, it is important to validate measurement performance of the assay detecting the biomarker. A device's measurement performance can be impacted by preanalytical, analytical, and postanalytical factors. Preanalytical factors are related to specimen collection as well as handling and storage. Postanalytical factors are related to data entry and calculations, interpretation of the results, data transfer and the methods used to report the results, etc. Analytical factors are described by characteristics, such as measurement bias, measurement precision (repeatability, intermediate precision, and reproducibility), limits of detection, limits of blank, stability (of the device and sample), interferences, sample commutability, and others. The definitions of the terminology in this section and other terms are found in literatures.^[9–13]

The types of analytical validation studies required for a diagnostic device depend on the device's IU, but two aspects of measurement validation commonly assessed across most types of tests are accuracy and precision. Accuracy is the closeness of the agreement between the result of a measurement and a true value of the measurand.^[9] Precision is the closeness of agreement between replicate measurements on the same object (e.g., sample) under specified testing conditions. The specified testing conditions can be repeatability conditions—replicate measurements on the same or similar objects under the same or similar conditions; reproducibility conditions—replicate measurements on the same or similar objects using different operating conditions; or some other set of intermediate conditions. Some factors which may impact the device's precision are reagent lot, operator, instrument, day, device run, replicate, and testing site. When the device results are visually interpreted by readers, it is important to evaluate reader-to-reader variability. For example, for IHC tests, site, lot, instrument, operator, day, run, replicate, and readers are often considered in the precision study. For continuous measurements, imprecision is typically expressed as a standard deviation and a percent coefficient of variation. For categorical or binary measurements, precision is typically expressed as percentage agreement among the replicates.

The device should be “locked down” with regard to its design parameters and fully prespecified with regard to its indications for use and IU before the initiation of measurement validation studies or trials, although carefully designed adaptive design that allows intratrial cutoff

alterations may be acceptable in some situations.^[7] For many diagnostic devices in precision medicine, the final output is a categorical result that classifies subjects into two or more groups (e.g., mutation present or absent). Devices with final categorical results often have an underlying continuous measurement that is then categorized based on a prespecified cutoff(s). The cutoff value is the point for defining test-positive or test-negative results. For companion diagnostic tests, subjects with positive test results are those who are suitable for a treatment, choose the appropriate dose, or make other clinical decisions. Prespecified cutoff values are essential for the validation of a diagnostic device in a clinical trial, since changing the cutoff values implies re-classifying the patient groups. It is recognized that the optimal cutoff may be unknown prior to the availability of clinical data. In such cases, an independent clinical trial confirming the results with the predetermined cutoff, or an adaptive design that allows intratrial cutoff alterations, would be necessary to validate the device and corresponding targeted therapy.^[7]

TREATMENT SELECTION: CLINICAL VALIDATION OF A COMPANION DIAGNOSTIC AND THE CORRESPONDING THERAPEUTIC PRODUCT

When a clinical trial is properly designed to establish the efficacy and safety of a therapeutic product in a biomarker-defined patient population, the results of the clinical trial are also used to establish the clinical validity of the CDx which measures or detects the biomarker(s). There are a variety of clinical trial designs that may be used to validate a CDx and the corresponding therapeutic product. The appropriate clinical trial design depends on the IU of device, proposed objectives, hypothesis of interest, and what has already been established during the development stage of the device and therapeutic product. Biomarker-stratified and targeted designs are the two most commonly used clinical trial designs that evaluate an investigational therapy along with a CDx (Fig. 1); however, other designs could also be considered appropriate. Without loss of generality, in this entry, we assume that the biomarker-positive patients are the targeted patients of the therapeutic product and patients in a clinical trial will be randomized to either therapeutic treatment or control arms.

The success of a clinical trial depends on many factors including, but not limited to, the following: 1) the characteristics of the biomarker; 2) the corresponding therapeutic product's safety and efficacy; 3) the ability of the CDx to predict differential effects of the treatment on outcome or the ability of the CDx to properly select patients who respond well to the treatment; and 4) the device measurement performance as discussed in the section “Measurement Performance Validation of Diagnostic Devices.”

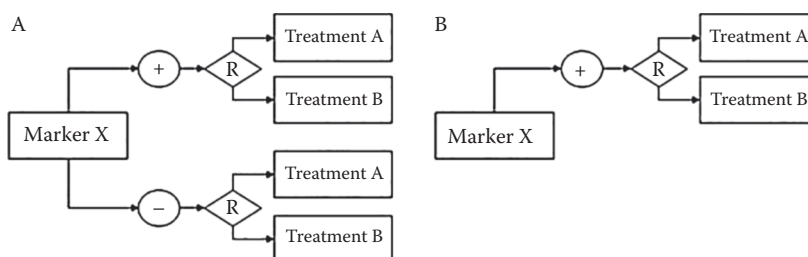


Fig. 1 Clinical trial designs. Trial A is called biomarker-stratified design^[4,14] used to evaluate treatment effects, biomarker effects, and their interaction, by stratifying randomization based on biomarker status, as determined by a companion diagnostic. Trial B is called targeted design^[4,14] used to evaluate treatment effects in a targeted population of biomarker-positive patients. Note: + = biomarker-positive; - = biomarker-negative; R = randomize. Treatment A is the experimental arm and treatment B is the control arm, typically standard of care or placebo

Biomarker-Stratified Design

In the clinical trial design depicted in Fig. 1A, both biomarker-positive and at least a proportion of biomarker-negatives are enrolled. This design is useful as it is desirable to obtain information about the safety and efficacy of the therapeutic product for both marker-positive and marker-negative patients in the clinical trial to ascertain the appropriateness of restricting the therapy to the biomarker-positive group. This approach may be particularly valuable when the biological plausibility or medical relevance of the biomarker is not well understood. When biomarker positivity (defined as the probability of patients being biomarker-positive) population is not low (e.g., >10%), an all-comers design could be used, i.e., all eligible subjects, regardless of biomarker status, are enrolled into the trial. When biomarker positivity is low, marker-positive patients may be enriched, as an all-comers trial design would require a large sample size. Specifically, all eligible biomarker-positive patients and a randomly selected proportion of eligible biomarker-negative patients will be enrolled to the trial. Therefore, biomarker-positive patients are enriched in the study population, relative to an all-comers trial.

The primary objective of the clinical trial under this design is to determine whether the biomarker of interest is truly a predictive biomarker. The most convincing evidence would be a clinically and statistically significant interaction between the treatment and the biomarker status defined by the device (denoted as γ). Therefore, the primary analysis for biomarker-stratified design is treatment by device interaction (biomarker predictive value), i.e., the hypothesis of interest:

$$H_0: \gamma \leq \delta \text{ vs. } H_a: \gamma > \delta \quad (1)$$

where δ is a predefined clinically meaningful value.

In addition to the predictive value, a biomarker-stratified design is useful for assessing a marker's main effect (marker prognostic value), drug efficacy in biomarker-positive patients (denoted as γ_1), drug efficacy in biomarker-negative patients, and overall drug efficacy

which can also be the primary analysis (denoted as μ). Note that under the marker-positive patients are enriched, if the drug efficacy is significantly different between biomarker-positive and biomarker-negative patients; the estimate of overall drug efficacy from such a trial will usually need to be adjusted for the enrichment in order to make a proper inference about the device IU population. When assessing overall drug efficacy, one also needs to determine whether μ is driven by the drug effect in biomarker-positive patients defined by the device.

Targeted Design

In the design shown in Fig. 1B, only biomarker-positive subjects are enrolled into the clinical trial. This design could be used under certain circumstances, e.g., when biological plausibility or medical relevance of the biomarker is well understood or if there is persuasive evidence indicating harm or lack of efficacy in the biomarker-negative subpopulation.^[7] Unlike the biomarker-stratified design, with this design, the predictive value of the device (and biomarker) cannot be determined because there is no information on the treatment efficacy in the biomarker-negative population. Likewise, there is no information about the prognostic value of the device. Under this design, the primary objective of the trial is to evaluate the therapeutic product efficacy in patients defined as biomarker positive by the CDx and it is not to validate the predictive value of the CDx, i.e., the hypothesis of interest:

$$H_0: \gamma_1 \leq \delta_1 \text{ vs. } H_a: \gamma_1 > \delta_1 \quad (2)$$

where δ_1 is a predefined clinically meaningful value and γ_1 is the treatment effect in device-positive patients. Several successful examples of therapeutic products with an accompanying IVD companion diagnostic including Gleevec (Imatinib; Novartis), DAKO C-KIT pharmDx (Dako North America, Inc.), Erbitux (cetuximab; ImClone Systems), and *therascreen* KRAS RGQ PCR kit (Qiagen Manchester Ltd.) were validated using a targeted design.^[6]

Bridging Study

It is less complicated when the biomarker status of patients enrolled in to the clinical trial is defined by a well-characterized investigational device, e.g., a device with analytical characteristics of the proposed “to be marketed CDx,” so that appropriate clinical and analytical validation studies apply directly to the CDx and the corresponding therapeutic product. However, for various reasons, this may be difficult or impractical to meet. Instead of the investigational CDx, a prototype of the CDx, a clinical trial assay (CTA), may be used to determine patient biomarker status in the clinical trial. One reason that the CTA might be used is that the investigational CDx is not available at the time of the clinical trial, and the CTA is available and used for patient enrollment in the trial and for the analysis of samples. In such cases, a bridging study, which is a concordance study of the proposed CDx and the CTA using the samples from the CTA-drug pivotal clinical trial, would be warranted to evaluate the clinical utility of the proposed CDx for the corresponding therapeutic product. The statistical issues, study design, and data analysis for the bridging study are discussed in detail in literature.^[15]

Prescreening Bias

When the prevalence of biomarker positivity is low, a large number of patients will usually need to be screened by the investigational CDx in order to enroll a sufficient number of biomarker-positive patients to properly power the clinical trial. In order to accelerate the enrollment process, patients could be prescreened first by prescreening tests and the results “confirmed” by the investigational CDx. Prescreening could be performed by one or several tests which could be a local developed test(s) or a prototype CDx. Prescreening may result in a biased clinical trial population that does not represent the device/therapeutic product IU population in real-world testing^[16] and consequently could lead to biased estimate of device clinical performance or drug effect.

We use the targeted design as an example to illustrate the issue. If patients are prescreened by a prescreening test, only patients who are prescreened-positive and investigational CDx-positive (denoted as double positives) will be enrolled to the trial. Patients who are prescreened negative will not be tested by the investigational CDx and will not be enrolled to the trial. Some of these patients could have been investigational CDx positive and may have been eligible to be enrolled into the trial if there were no prescreening. Thus, the trial could have been subject to selection bias due to prescreening (also called prescreening bias) because the study population is double positive which is different from the drug IU population, i.e., investigational CDx-positive patients. Prescreen negatives might represent populations with characteristics that make it more difficult

to test, etc., and so the investigational CDx would not be assessed under those conditions.

A statistical model may be used to correct a prescreening bias if the following information is available: the drug efficacy in prescreening-negative and investigational CDx-positive patients, the concordance between the prescreening test and the investigational CDx, and the percentage of investigational CDx positivity, etc. However, such information is often not available. Therefore, prescreening can create technical problems for sponsors attempting to evaluate a novel therapeutic product's safety and efficacy in its IU population as well as for CDx manufacturers attempting to provide an unbiased estimation of the investigational CDx performance. Thus, planning to enroll subjects into a trial based on the confirmation of a local test result can be problematic. When prescreening is unavoidable, it may useful to be aware of the potential bias, take precautionary steps to evaluate whether the expected marker positivity may be skewed by prescreening, and develop approaches to adequately address potential selection bias.

Adaptive Signature

The increasing awareness of heterogeneity in treatment effect has motivated clinical trial designs that prospectively allow for demonstration of treatment efficacy in a subpopulation of patients. If a promising subpopulation is available at the trial design stage, this knowledge can be used to plan a prespecified subgroup analysis in a traditional broad eligibility trial or to restrict enrollment in a targeted trial, depending on how confident the investigator is about the subpopulation. Such a well-defined subpopulation is often unavailable in the design stage; instead, the investigator may only have a preliminary idea that a collection of baseline covariates (biomarkers) could be useful in identifying a sensitive subpopulation. In that case, a target subpopulation can be developed at the end of a broad eligibility trial under an adaptive signature design or at an interim analysis (with the possibility of restricting subsequent enrollment) under an adaptive enrichment design.^[17] The various designs have been discussed and compared by Chen et al.^[18]

There has also been intensive research on estimating optimal treatment regimens from clinical trials or observational data. A treatment regimen may involve a single treatment decision based on baseline characteristics^[19] or a sequence of treatment decisions based on time-dependent covariates.^[20] These methods tend to be exploratory in nature and have to be applied in a confirmatory fashion.

RISK PREDICTION

Absolute risk is the cumulative probability that a patient experiences an adverse health event or clinical outcome of

interest by a particular time in the future.^[21] For example, healthy people may be assessed for their lifetime risk of developing breast cancer^[22] or experiencing a cardiovascular event,^[23] according to risk factors for these events. Patients with disease may be assessed for prognosis, e.g., risk of hospitalization, disease progression, mortality, relapse, or recurrence. Risk is an important factor in the patient's medical management and in fact is often the basis by which accredited organizations form guidelines on the medical management, treatment, or screening intervals of specific populations. For example, during chemotherapy treatment of patients with breast cancer, prophylactic use of granulocyte colony-stimulating factor has been recommended if the risk of febrile neutropenia is greater than 20%.^[24]

The utility of a risk prediction varies with the application. In many clinical settings, risk informs medical care professionals by helping them to determine the frequency with which patients should be screened for an event of interest. A risk prediction may also assist physicians in the decision to prescribe or not to prescribe an intervention or treatment designed to mitigate the risk.

Predictive models are increasingly being developed in attempts to estimate the risk of an event from a model that combines information on predictor variables, including patient characteristics (e.g., age, body mass index, and disease stage) and biomarkers (biological characteristics associated with the likelihood of disease). Unfortunately, most prediction models do not successfully proceed from bench (research) to bedside (clinical practice). A common reason for failure is attempting to estimate the model performance on the same dataset on which it was developed. This re-substitution can result in huge performance biases, leading to an overly high expectation of success in an independent validation study. This bias can be mitigated with rigorous cross-validation of all model development steps, especially selection of the predictor variables from potential candidates.^[25]

Another reason that many predictive models fail is poor experimental design of the development study. Examining a biomarker for association with a clinical outcome on a retrospectively collected, archived set of specimens conveniently available from disparate sources, can often result in the association being confounded by biomarker assay measurement variation due to preanalytical and analytical factors. Studies of such convenience samples are generally discouraged to mitigate the confounding of biomarker associations with unmeasured variables related to the availability of the samples. One way to ensure that uniform policies are in place for specimen collection, archival, and centralized processing is to obtain specimens from one or more adequate, well-conducted, well-controlled studies of a population appropriate for the prediction goal.^[26]

Similarly, when designing a prospective developmental study, application of good experimental design principles

is also essential. A prospectively designed development study would ideally be balanced on all controllable confounding factors such as those known to produce technical variations in the measurements from biomarker assays.

In summary, planning the intention and design in predictive model development study is as important as validating the model itself. Careful annotation of research steps can be extremely helpful to ensure that the research is reproducible.

A fully developed risk prediction model may be embedded in a medical device that is being evaluated by a regulatory authority for premarket approval. To support approval, the risk prediction device would generally be validated with data, independent of developmental data set. To ensure that the performance of the device in the validation study is generalizable to new subjects, statistical rigor necessitates that the device is “locked down” with regard to its design parameters. It is also important to ensure that subjects enrolled in the study are fully prespecified with regard to its indications for use and of the device IU population, and that the use of the device in the study is consistent with its instructions for use.

The successful validation strategies for some prognostic tests have been reviewed for statistical lessons learned.^[27]

If a risk prediction model reports a risk category (e.g., low, moderate, or high risk), differences in risk between categories can be helpful in assessing clinical significance of the model and can be estimated from Kaplan–Meier curves for the categories (assuming a prospective all-comers study design). Comparing risk categories using relative risks or hazard ratios are generally insufficient for assessing clinical significance and making decisions on patient management. For example, a relative risk of three could indicate an increase in risk from 0.1% to 0.3% or from 8% to 24%. Depending on the medical setting, the former risk difference of 0.2% could be considered clinically insignificant, while the latter risk difference of 16% could be considered clinically important, suggesting that absolute risk and absolute risk differences are generally more clinically useful to consider than just relative risk. Similarly, measures of discrimination such as sensitivity, specificity, and odds ratio are often less useful for assessing clinical significance of the risk categorization than are differences in absolute risk between the categories.

If a model reports a probabilistic value $r(x)$ of the risk for a subject based on predictor value x , then the clinical usefulness of this risk prediction can be evaluated for how well it is calibrated to the actual risk $\pi(x)$ for that subject. If $r(x)$ is not well-calibrated, then a particular subject may be mismanaged according to guidelines for management predicated on risk. The Hosmer–Lemeshow goodness-of-fit test is popular for evaluating whether probabilistic risk predictions are well-calibrated with actual risk based on grouping subjects according to their risk prediction.^[21]

In other evaluation risk prediction, predictors (or the risk prediction itself) may be included in a regression model such as a logistic regression of disease status or a Cox regression of time to disease onset. Such models can be useful for assessing the proportion of variation in risk that is explained by the predictor.^[21,28] Such models can also be used to obtain the predictiveness curve, a plot of the distribution of risk among subjects based on the particular model.^[29]

Discrimination ability or classification accuracy of the risk prediction can be assessed overall with the receiver operating characteristic (ROC) plot of pairs of sensitivity and specificity conferred by cutoffs in the risk prediction. Interestingly, models that are perfectly calibrated for risk may not necessarily achieve perfect discrimination (i.e., area under the ROC curve of 1).^[21]

To be clinically useful, predictive models ought to provide some advantage or added value in the assessment of risk compared with the armamentarium of risk factors that are often already available and is in the routine clinical use. The incremental value of a risk prediction can be assessed with a multivariable logistic or Cox regression analysis that includes known risk factors in addition to the biomarker risk prediction being evaluated. Ultimately, the clinical value of a risk prediction model may need to be assessed according to assumed clinical consequence of misclassifying risk or with a clinical validity study to evaluate whether use of the model for patient management leads to improved clinical outcomes.

CONCLUSION

This entry discusses three major IU of diagnostic devices for precision medicine, namely, treatment selection, risk prediction, and disease or treatment monitoring. General considerations for clinical validations for each IU are discussed. This entry does not discuss other IU of precision medicine, e.g., complementary diagnostics. Biomarkers play a critical role in precision medicine, with applications including prognosis, monitoring, and selection of targeted therapies. Once a candidate assay detecting a marker has been developed, the devices detecting/measuring biomarker must be studied to determine if it has clinical validity. Observation of apparent clinical validity needs to be reproduced in an independent clinical trial in order to confirm clinical validity.

ACKNOWLEDGMENTS

The authors gratefully thank Drs. Eunice Lee, Elizabeth Mansfield, and Marina Kondratovich from Center for Devices and Radiological Health; Dr. Tom Gwise from Center for Drug Evaluation and Research, U.S. Food and Drug Administration; and Dr. Richard Simon from National Institutes of Health.

REFERENCES

1. Sadee, W.; Dai, Z. Pharmacogenetics/genomics and personalized medicine. *Hum. Mol. Genet.* **2005**, *14* (2), R207–R214.
2. National Research Council: Committee on a Framework for Developing a New Taxonomy of Disease. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*; The National Academies Press: Washington, DC, 2011.
3. The Precision Medicine Initiative Cohort Program—Building a Research Foundation for 21st Century Medicine, 2015. Available at: <https://www.nih.gov/sites/default/files/research-training/initiatives/pmi/pmi-working-group-report-20150917-2.pdf> (accessed November 2016).
4. Buyse, M.; Michiels, S.; Sargent, D.J.; Grothey, A.; Matheson, A.; de Gramont, A. Integrating biomarkers in clinical trials. *Expert Rev. Mol. Diagn.* **2011**, *11* (2), 171–182.
5. Biomarkers Definition Working Group. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.* **2001**, *69*, 89–95.
6. List of Cleared or Approved Companion Diagnostic Devices (In Vitro and Imaging Tools), last updated October 28, 2016. Available at <http://www.fda.gov/companiondiagnostics> (accessed November 2016).
7. The U.S. Food and Drug Administration. In Vitro Companion Diagnostic Devices Guidance for Industry and Food and Drug Administration Staff, 2014. Available at <http://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm262327.pdf> (accessed November 2016).
8. Principles for Codevelopment of an In Vitro Companion Diagnostic Device with a Therapeutic Product. Draft Guidance for Industry and Food and Drug Administration Staff. 2016. Available at <http://www.fda.gov/ucm/groups/fdagovpublic/@fdagov-meddev-gen/documents/document/ucm510824.pdf> (accessed November 2016).
9. CLSI. *User Protocol for Evaluation of Qualitative Test Performance EP12-A*; Clinical and Laboratory Standards Institute: Wayne, PA, 2002.
10. CLSI. *Evaluation of Precision Performance of Quantitative Measurement Procedures; Approved Guideline—Third Edition, CLSI document EP05-A3*; Clinical and Laboratory Standards Institute: Wayne, PA, 2002.
11. CLSI. *Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures; Approved Guideline—Second Edition, CLSI document EP17-A2*; Clinical and Laboratory Standards Institute: Wayne, PA, 2012.
12. CLSI. *Evaluation of the Linearity of Quantitative Measurement Procedures: A Statistical Approach: Approved Guideline, CLSI document EP6-A*; Clinical and Laboratory Standards Institute: Wayne, PA, 2003.
13. CLSI. *Interference Testing in Clinical Chemistry; Approved Guideline—Second Edition, CLSI Document EP7-A2*; Clinical and Laboratory Standards Institute: Wayne, PA, 2005.
14. Freidlin, B.; McShane, L.M.; Korn, E.L. Randomized clinical trials with biomarkers: Design issues. *J. Natl. Cancer Inst.* **2010**, *102*, 152–160.
15. Li, M. Statistical consideration and challenges in bridging study in personalized medicine. *J. Biopharm. Stat.* **2015**, *25*, 397–407.

16. Li, M.; Yu, T.; Hu, Y. The impact of companion diagnostic device measurement performance on clinical validation of personalized medicine. *Stat. Med.* **2015**, *34*, 2222–2234.
17. Simon, N.; Simon, R. Adaptive enrichment designs for clinical trials. *Biostatistics* **2013**, *14*, 613–625.
18. Chen, J.J.; Lu, T.P.; Chen, D.T.; Wang, S.J. Biomarker adaptive designs in clinical trials. *Transl. Cancer Res.* **2014**, *3*, 279–292.
19. Zhao, Y.; Zeng, D.; Rush, J.; Kosorok, M.R. Estimating individualized treatment rules using outcome weighted learning. *J. Am. Stat. Assoc.* **2012**, *107*, 1106–1118.
20. Zhang, B.; Tsiatis, A.A.; Laber, E.B.; Davidian, M. Robust estimation of optimal dynamic treatment regimens for sequential treatment decisions. *Biometrika* **2013**, *100*, 681–694.
21. Gail, M.H.; Pfeiffer, R.M. On criteria for evaluating models of absolute risk. *Biostatistics* **2005**, *6*, 227–239.
22. Gail, M.H.; Brinton, L.A.; Byar, D.P.; Corle, D.K.; Green, S.B.; Schairer, C.; Mulvihill, J.J. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* **1989**, *81* (24), 1879–1886.
23. Wilson, P.W.; D'Agostino, R.B.; Levy, D.; Belanger, A.M.; Silbershatz, H.; Kannel, W.B. Prediction of coronary heart disease using risk factor categories. *Circulation* **1998**, *97*, 1837–1847.
24. Aapro, M.S.; Cameron, D.A.; Pettengell, R.; Bohlius, J.; Crawford, J.; Ellis, M.; Kearney, N.; Lyman, G.H.; Tjan-Heijnen, V.C.; Walewski, J.; Weber, D.C.; Zielinski, C. European Organisation for Research and Treatment of Cancer (EORTC) Granulocyte Colony-Stimulating Factor (G-CSF) Guidelines Working Party. EORTC guidelines for the use of granulocyte-colony stimulating factor to reduce the incidence of chemotherapy-induced febrile neutropenia in adult patients with lymphomas and solid tumours. *Eur. J. Cancer* **2006**, *42*, 2433–2453.
25. Simon, R.; Radmacher, M.D.; Dobbin, K.; McShane, L.M. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl. Cancer Inst.* **2003**, *95*, 14–18.
26. Simon, R.; Paik, S.; Hayes, D.F. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J. Natl. Cancer Inst.* **2009**, *101*, 1446–1452.
27. Tang, R.; Pennello, G. Validation of prognostic marker tests: Statistical lessons learned from regulatory experience. *Ther. Inno. Reg. Sci.* **2016**, *50* (2), 241–252.
28. Mittlbock, M.; Schemper, M. Explained variation for logistic regression. *Stat. Med.* **1996**, *15*, 1987–1997.
29. Pepe, M.S.; Feng, Z.; Huang, Y.; Longton, G.; Prentice, R.; Thompson, I.M.; Zheng, Y. Integrating the predictive-ness of a marker with its performance as a classifier. *Am. J. Epidemiol.* **2008**, *167* (3), 362–368.

Pharmaceutical Development: Statistical Designs

Annpey Pong

Merck Research Laboratories, Rahway, New Jersey, U.S.A.

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

A pharmaceutical development process consists of non-clinical (e.g., assay and process validation and stability testing), pre-clinical (e.g., animal and bioavailability bioequivalence studies), and clinical (e.g., phases 1–3 clinical trials) development. In this entry, various statistical designs that are commonly considered for achieving desired good drug characteristics as described in the United States Pharmacopeia and National Formulary (USP/NF) at various stages are reviewed.

Keywords: Pharmaceutical development process; Quality by design; Assay validation; Process validation; Stability analysis; Adaptive design methods in clinical trials.

INTRODUCTION

The pharmaceutical development process, involving drug discovery, formulation, laboratory development, animal studies, clinical development, and regulatory registration, is a continual, length, and costly process. Essentially, a pharmaceutical development process consists of different phases of development, including non-clinical development (e.g., assay development/validation in laboratory development, manufacturing process validation, and stability testing and analysis), pre-clinical development (e.g., animal studies and bioavailability/bioequivalence studies), and clinical development (e.g., phase 1–3 clinical development).^[1] These phases may occur in sequential order or be overlapped during development process. At different phases of pharmaceutical development, valid statistical designs are usually employed to ensure that the drug product possesses good drug characteristics such as identity, strength, quality, purity, safety, efficacy, and stability before and post-approval of the drug product.

In this entry, we will not only introduce the concept of quality by design suggested by the FDA, but also provide an overview of statistical designs that are commonly employed at different stages of a pharmaceutical development.

QUALITY BY DESIGN (QbD)

In many practical cases, the pharmaceutical manufacturing process is not only unable to meet pre-specified quality standard, but also inability to predict effects scale-up on final product. Quality by design (QbD) is a concept that quality could be planned, and that most quality crises and

problems relate to the way in which quality was planned. The FDA has implemented the concepts of QbD into its pre-market processes.^[2] The focus of this concept is that quality should be built into a product with an understanding of the product and process by which it is developed and manufactured along with a knowledge of the risks involved in manufacturing the product and how best to mitigate those risks. Winkle provides a comprehensive comparison of traditional approach with the systematic QbD approach (see [Table 1](#)).^[3]

The use of quality by design cannot only reduce number of manufacturing supplements required for post market changes and at the same time, allow for implementation of new technology to improve manufacturing without regulatory scrutiny. From the perspectives of FDA, the implementation of QbD will not only enhance scientific foundation for review, but also provides better coordination across review, compliance and inspection divisions within the FDA. Additionally it not only improves information in regulatory submissions and quality of review, but also provides for better consistency and for more flexibility in decision making.

NONCLINICAL APPLICATION

Assay Validation

The major objective of the validation of an assay method is to ensure that the assay method can produce unbiased and precise assay results for the active ingredient of a compound. The USP/NF indicates that the assay method needs to be validated in terms of the primary

Table 1 A Comparison between Traditional Approach and QbD Systematic Approach

Aspects	Traditional	Quality by Design
Pharmaceutical Development	Empirical; Univariate experiments	Systematic; Multivariate experiments
Manufacturing Process	Fixed	Adjustable
Process Control	In-process testing for go/no-go; Offline analysis w/ slow Response	Process Analytical Technology utilized for feedback and feed forward at real time
Product Specification	Primary means of quality control; Based on batch data	Overall quality control strategy; Based on desired product performance (safety and efficacy)
Control Strategy	Based on intermediate and end product testing	Risk-based; Controls shifted upstream; Real-time release upstream
Lifecycle Management	Reactive to problems; Scale-up and post-approval changes	Continual improvement enabled within design space

validation parameters given in Table 2.^[4] To meet the USP/NF standards for each of these parameters, an appropriate statistical design is necessarily chosen to provide sound statistical inferences for these parameters and their variances. Several commonly employed designs in assay validation are briefly described below.

Randomized block design – This is the most commonly used design in assay validation. Suppose that there are J levels of potency (say [L, M, H] where L/M/H indicate low/median/high of the level of 100% of label claim) in a recovery study. Samples are assayed on different days with and/or without the same number of replicates (say 3 replicates). This design provides independent estimates of day-to-day variability and within-day variability.

Latin square type of design – In the randomized block design, assays of different levels on the same day are usually performed in sequential order, which may introduce bias due to testing order. To avoid the bias that may be introduced by testing order, a Latin square type of design is used to balance the potential bias. In other words, the number of days should be equal to the number of levels. The Latin square design is then applied to the test sequence of the levels of potency.

Incomplete block design – In practice, the number of assays performed on each day may not be able to cover all

the levels of potency under study due to limited resources available. In addition, in some cases, the assay can only be conducted on a certain number of days, which are fewer than the number of levels of potency. In these situations, an incomplete block design may be used to randomize test sequences on each day.

Process Validation

A manufacturing process is a continuous process that involves a number of critical stages for quality assurance. At each critical stage, problems may occur during the process. Thus, process validation is essential not only to ensure that the process does what it purports to do, but also to ensure that the drug product will conform to USP/NF specifications of good drug characteristics.

Regulatory requirement – In its recent guidance on process validation, FDA underlines the concept of QbD in process validation to continue benefiting from knowledge gained, and continually improve throughout the process lifecycle. For a prospective validation, FDA requires that at least three batches be evaluated. A validation design that clearly outlines sampling plan, testing plan, and acceptance criteria for validation are briefly described below.

Sampling plan, testing plan, and acceptance criteria – The validation of a manufacturing process requires satisfactory results at each critical stage for three batches. For each batch, sampling plan and testing procedures may be different from stage to stage. Acceptance limits for the validation of a manufacturing process are usually designed to be more stringent to assure that the final product will pass USP/NF specifications. Acceptance limits are a set of sample statistics from which a confidence interval for a given parameter is constructed to meet a pre-specified lower probability bound for passing a particular USP test.

Table 2 Performance Characteristics in Assay Validation

Specificity
Linearity
Accuracy
Precision (repeatability, intermediate precision, and reproducibility)
Range
Quantitation limit
Detection limit

Scale-up

Scale-up program plays an important role to scale up a laboratory batch to a commercial (production) batch. The purpose of a scale-up program is not only to identify, evaluate, and optimize critical formulation and/or manufacturing process factors of the drug product but also to maximize or minimize excipient ranges. A successful scale-up program can result in an improvement in formulation/process or at least a recommendation on a revised procedure for formulation/process of the drug product. Some commonly employed designs in scale-up experiments are briefly described below.

Factorial design – A full factorial design is a design that consists of all possible different combinations of one level from each factor. If there are l_k levels for the k th factor X_k , the corresponding full factorial design is called a general $l_1 l_2 \dots l_K$ factorial design. For example, when $l_i=2$ for all i , the general factorial design is called a 2^K factorial design. A full factorial design provides estimates not only for main effects but also for interactions with maximum precision. The main effects and interaction effects can easily be obtained using a table of contrast coefficients and/or Yate's algorithm.^[5–6]

Fractional factorial design – This design consists of a fraction of a full factorial experiment. For example, a $(\frac{1}{2})^P$ fraction of a 2^K factorial design is called a 2^{K-P} fractional factorial design. When $P=1$, a full factorial design reduces to a one-half factorial design. For a 2^{4-1} fractional factorial design, the eight observations cannot provide independent estimates for the 16 effects alone but for some confounding effects, such as the sum of a main effect and a three-factor interaction that are confounded with each other. In practice, however, the three-factor or higher-factor interactions are usually negligible. In this case, a fractional factorial design is useful in estimating the main effects. In practice, a fractional factorial design is useful when there are many factors to be studied because it is almost impossible to perform a full factorial design even at two levels.

Central composite design – This design is a full factorial design or a fractional factorial design augmented by a $\pm \alpha$ level at each of the K factors and n central points. It consists of 1 center point, 8 points on the cube (a 2^3 factorial arrangement), and 6 star points (Table 3). It should be noted that a central composite design with $K=2$, $\alpha=1$, and $n=1$ reduces to a 3^2 factorial design. For a full 2^K factorial design, although the design provides independent estimates for the 2^K-1 effects, it does not give an estimate of the experimental error unless some runs are repeated. Unlike the full 2^K factorial design, the central composite design provides an estimate of the experimental error. The experimental error is usually estimated based on n observations at the central point.

Table 3 Central Composite Design for $K=3$ and $n=1$

Run	X_1	X_2	X_3
1	–1	–1	–1
2	1	–1	–1
3	–1	1	–1
4	1	1	–1
5	–1	–1	1
6	1	–1	1
7	–1	1	1
8	1	1	1
9	0	0	0
10	α	0	0
11	$-\alpha$	0	0
12	0	α	0
13	0	$-\alpha$	0
14	0	0	α
15	0	0	$-\alpha$

Stability Analysis

For study design and sample selection criteria, the ICH Q5C guideline recommends that a bracketing design or a matrixing design be used.^[7–12] In other words, samples for the stability program can be selected on the basis of a matrixing and/or bracketing design.

A bracketing design is a design that only samples on the extremes of certain design factors which are tested at all-time points. Bracketing can be applied to studies with multiple strengths of identical or closely related formulation, or it can also be applied to studies with the same container closure system with the fill volume and/or the container size varied.

A matrixing design is a statistical design of a stability study that allows different fractions of samples to be tested at different sampling time points. Each subset of samples represents the stability of all samples at a given time point. Differences in the samples should be identified as covering different batches, different strengths, and different sizes of the same container closure system. A matrixing design should be balanced such that each combination of a factor is tested to the same extent over the duration of the studies. It should be noted that all samples should be tested at the last time point before the submission of application.

Storage Conditions – Stability studies are usually done at several different time points. In reality, there is no unique best design. The choice of design must use the fact that analyses will be done after data are collected. Nordbrock^[8] introduced several designs that are commonly considered in stability studies which are briefly described below.

Basic Matrix 2/3 on Time Design – A complete long-term study for one strength of a dosage form in one package has three batches, with all three tested every

3 months in the first year, every 6 months in the second year, and annually thereafter. The basic matrix 2/3 on time design has only two of the three batches tested at intermediate time points as presented in Table 4. If an analysis is to be done after 18 months for a registration application, the basic matrix 2/3 on time design can be modified by testing all batches at 18 months.

Matrix 2/3 on Time Design with Multiple Packages – The basic matrix 2/3 on time design is applied to each package in a balanced fashion, as presented in Table 5. *Balance* is defined as testing each batch twice at each intermediate time point, and each package twice at each intermediate time point.

Matrix 2/3 on Time Design with Multiple Packages and Multiple Strengths – Assuming there are three packages for each strength, the basic matrix 2/3 on time design can be applied to each of the 9 sub-batches in a balanced fashion (Table 6). In this design, each sub-batch is tested twice at each intermediate time point, each package is tested twice at each intermediate time point for each batch, each batch is tested six times at each intermediate time point, and each package is tested six times at each intermediate time point. If an analysis will be done after 18 months (e.g., for a registration application), this design can be modified by testing all batch-by-strength-by-package combinations at 18 months.

Matrix 1/3 on Time Design – A further reduction in the amount of testing is accomplished by reducing the testing

Table 4 Basic matrix 2/3 on time design

Batch	Test times (months)
A	0, 3, 9, 12, 24, 36
B	0, 3, 6, 12, 18, 36
C	0, 6, 9, 18, 24, 36

Table 5 Matrix 2/3 on Time Design with Multiple Packages

Batch	Package 1	Package 2	Package 3
A	T1	T2	T3
B	T2	T3	T1
C	T3	T1	T2

Table 6 Matrix 2/3 on time design with multiple packages and multiple strengths

Batch	Strength	Package 1	Package 2	Package 3
A	10	T1	T2	T3
A	20	T2	T3	T1
A	30	T3	T1	T2
B	10	T2	T3	T1
B	20	T3	T1	T2
B	30	T1	T2	T3
C	10	T3	T1	T2
C	20	T1	T2	T3
C	30	T2	T3	T1

Table 7 Basic Matrix 1/3 on Time Design

Batch	Test times (months)
A	0, 3, 12, 36
B	0, 6, 18, 36
C	0, 9, 24, 36

in each of the preceding designs from 2/3 to 1/3. For example, the basic 1/3 on time design has one of the three batches tested at each intermediate time point, as presented in Table 7.

Matrix on Batch-by-Strength-by-Package Combinations – For multiple strengths and multiple packages, one could also choose to test only on a portion of the batch-by-strength-by-package combinations. An example of matrix design on batch-by-strength-by-package combinations is presented in Table 8, with two packages selected for each of the 6 sub-batches, and where time is also matrixed by the factor 1/2. This design is approximately balanced.

Uniform Matrix Design – This design is for which the same time protocol is used for all combinations of the other design factors.^[12] The strategy is to delete certain times at 3-month, 6-month, 9-month, and 18-month; therefore testing is done only at 12, 24, and 36 months. This design has the advantages of eliminating time points that add little to reducing the variability of the slope of the regression line. The disadvantage is that if there are major problems with the stability, there is no early warning because early testing is not done. Further, it may not be possible to determine if the linear model is appropriate.

Comparison of Designs – Nordbrock compared designs based on the power approach.^[8] Other methods of comparing designs are given in Ju and Chow^[13] and Pong and Raghavarao,^[10] where the criterion is the precision for estimating shelf life. In addition, Nordbrock^[14] compared matrix designs to full designs using the probability of achieving the desired shelf life.

PRE-CLINICAL APPLICATION

Animal Studies

The primary focus of pre-clinical drug development is to evaluate the safety of the drug product through in vitro

Table 8 Matrix 1/2 on Time and Matrix on Batch-by-Strength-by-Package

Batch	Strength	Package 1	Package 2	Package 3
A	10	T1	T2	—
A	20	T2	—	T1
B	10	T2	—	T1
B	20	—	T1	T2
C	10	—	T1	T2
C	20	T1	T2	—

assays and animal studies. In vitro assay and animal testing are often considered as surrogate for human testing, under the assumption that they can be predictive of results in humans. Pre-clinical testing involves dose selection, toxicological testing for toxicity and carcinogenicity, and animal pharmacokinetics. For selection of an appropriate dose, dose-response studies are usually conducted to determine the effective dose such as the median effective dose ED_{50} in animals. The objective of toxicological testing in animal studies is to explore not only realistic safety extrapolation, but also to evaluate new methods to test for toxicity.

In pre-clinical development, the primary focus of toxicity testing in animals includes long-term carcinogenicity studies and reproductive toxicology studies. The major objective of the long-term carcinogenicity study is to identify those compounds that are probable human carcinogens, i.e., they have abilities to cause (i) an increased incidence of tumor types, (ii) an earlier appearance of tumors, and (iii) an increased multiplicity of tumors in individuals. The purpose of reproductive toxicology is to identify the effect of a xenobiotic on mammalian reproduction.

Gad and Weil^[15] indicated that there are four basic experimental designs that are commonly used in toxicology. These designs are completely randomized design, randomized block design, Latin square design, and nested design which are briefly described below.

Completely Randomized Design – This is the most commonly used design in toxicological testing. In this design, the treatments are completely assigned to the experimental units at random. The design imposes no restrictions on the allocation of treatments to the experimental units, although a balance on the number of experimental units used per treatment group is preferred. If there is evidence of heterogeneity on experimental units, it is suggested that blocking should be used to increase the design efficiency.

Randomized Block Design – In this design, the experimental units are allocated to blocks such that the experimental units within a block are relatively homogenous and the number of experimental units within a block is equal to the number of treatments being investigated. The treatments are assigned at random to the experimental units within each block.

Latin Square Design – A Latin square design is an extension of the randomized block design. Latin square design permits the investigator to assess treatment effects when a double-blocking restriction is used on the experimental units, controlling for two sources of variation. In the Latin square design, the number of row blocks, number of column blocks, and number of treatment groups are equal.

Nested Design – A nested design is a multi-factor experiment. As an example, suppose there are two factors (say factor A and Factor B). A nested design is a design in which levels of one factor (say Factor B) are hierarchically subsumed under (or nested within) levels of another

factor (say Factor A). Thus, it is not possible to assess the complete combination of A and B levels in a nested design.

Bioequivalence Studies

When a drug product is going off patent, pharmaceutical or generic companies may file an abbreviated new drug application (ANDA) for generic approval. Generic drug products are defined as drug products that are *identical* to a drug which is the subject of an approved NDA with regard to active ingredient(s), route of administration, dosage form, strength, and conditions of use. For approval of generic drug products, the FDA requires that evidence of average bioequivalence in drug absorption be provided through the conduct of bioavailability and bioequivalence studies. Bioequivalence assessment is considered as a surrogate for clinical evaluation of the therapeutic equivalence of drug products.

As indicated in the Federal Register (Vol. 42, No. 5, Sec. 320.26(b) and Sec. 320.27(b), 1977), a bioavailability study should be crossover in design, unless a parallel or other design is more appropriate for valid scientific reasons. A typical 2×2 crossover design can be expressed as (TR, RT), where TR is the first sequence of treatments and RT denotes the second sequence of treatments. Under the (TR, RT) design, qualified subjects who are randomly assigned to either sequence 1 (TR) or sequence 2 (RT), where TR represents the test product (T) first and then cross-over to receive the reference product (R) after a sufficient length of wash-out period. Similarly, RT sequence starts with product R first, and then product T after the wash-out period.

One of the limitations of the standard 2×2 crossover design is that it does not provide independent estimates of intra-subject variabilities since each subject receives the same treatment only once. Therefore, the following alternatives are often considered:

- Design 1: Balaam's design – i.e., (TT, RR, RT, TR);
- Design 2: Two-sequence, three-period dual design – i.e., (TTRR, RTT);
- Design 3: Four-period design with two sequences – i.e., (TRRT, RTTR);
- Design 4: Four-period design with four sequences – i.e., (TTRR, RRTT, TRTR, RTTR).

The above study designs are also referred to as higher-order crossover designs. A higher-order crossover design is defined as a design with the number of sequences or the number of periods greater than the number of treatments to be compared.

For comparing more than two drug products, a Williams' design is often considered. Williams' design is a variance stabilizing design. More information regarding the construction and good design characteristics of Williams' designs can be found in Chow and Liu.^[16]

CLINICAL APPLICATION

Selecting an appropriate statistical design is critical in clinical development during the process of drug development. Ideally an optimal design will account for considerations from different perspectives. The commonly statistical designs used in clinical application are summarized below.

Parallel versus Crossover

Crossover Design – A crossover design is a modified randomized block design in which each block receives more than one treatment at different dosing periods. Patients in each block receive different sequences of treatments. For a crossover design it is not necessary that the number of treatments in each sequence be greater than or equal to the number of treatments to be compared. We will refer to a crossover design as a $p \times q$ crossover design for p sequences of treatments administered at q different dosing periods. Basically a crossover design has the following advantages: (i) it allows a within-patient comparison between treatments, since each patient serves as his or her own control. (ii) It removes the inter-patient variability from the comparison between treatments. (iii) With a proper randomization of patients to the treatment sequences, it provides the best unbiased estimates for the differences between treatments.^[16–17]

Parallel Group Design – A parallel group design is a complete randomized design in which each patient receives one and only one treatment in a random fashion. Basically there are two types of parallel group design for comparative clinical trials, namely, group comparison (or parallel-group) designs and matched pairs parallel designs. The simplest group comparison parallel group design is the two-group parallel design which compares two treatments (e.g., a treatment group vs. a control group). The ICH E9 guideline “*Statistical Principles for Clinical Trials*” indicates that the parallel group design is the most common trial design for confirmatory trials.^[18]

Cluster Randomized Design

In clinical trials, the randomization units could be some social intact units such as family, school, worksites, athletic teams, hospitals, or communities. These social units are called *clusters*. The randomized design with clusters is referred to as cluster randomized design.

One of the major considerations for design and analysis of cluster randomized trials is the control of the intra-cluster and inter-cluster variations. Note that the subjects within the same cluster might share the same traits or have similar characteristics. In other words, the subjects within the same cluster are more similar than are those between clusters. One statistical measure to quantify this similarity is the intra-class correlation coefficient (ICC). If the intra-class correlation coefficient, denoted by ρ , is positive, then

the intra-cluster variation is smaller than the inter-cluster variation. The ICC plays a very important role in analysis of cluster randomized trials using subjects as the unit of inference.

Titration Design

For phase I safety and tolerance studies, Rodda et al.^[19] classify traditional designs as (i) rising single-dose design, (ii) rising single-dose crossover design, (iii) alternative-panel rising single-dose design, (iv) alternative-panel rising single-dose crossover design, (v) parallel-panel rising multiple-dose design, and (iv) alternative-panel rising multiple-dose design.^[20–21]

Phase I studies are usually conducted in young, healthy male volunteers. The purpose of phase I studies is to obtain initial appraisal of drug safety through the evaluation of vital signs, physical health, and adverse events and frequent assessments of hematology, blood chemistry, and urine samples.

In medical practice, if the study medicine is intended for cancer or some life-threatening diseases, it may not be ethical to conduct phase I safety and tolerance studies on normal volunteers due to potential toxic or fatal effects.

Adaptive Design

In February 2010, the FDA circulated a draft guidance on adaptive design clinical trials, which defines an adaptive design as a study that includes a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of (usually interim) data from subjects in the study.^[22] Analyses of the accumulating study data are performed at prospectively planned time points within the study, with or without formal statistical hypothesis testing. The benefits and limitation of adaptive design are outlined in [Table 9](#).

Depending upon the adaptations employed, adaptive designs can be classified into several types which are briefly described below.^[23]

Adaptive Randomization Design – This design allows modification of randomization schedules based on varied and/or unequal probabilities of treatment assignment both prospectively and after the review of the response of previously assigned subjects. It assigns more subjects to promising test treatment under investigation and potentially to increase the probability of success of the intended trial. The commonly used adaptive randomization procedures include treatment-adaptive randomization, covariate-adaptive randomization, and response-adaptive randomization. In practice, an adaptive randomization design may be valuable in trials with a relatively small sample size or a trial with short-term outcomes. The issue regarding the balance of patient characteristics between the treatment groups is a concern for this design. The imbalance

Table 9 Possible Benefits and Limitation for Utilizing Adaptive Designs in Clinical Trials

Possible benefits	Limitations
Flexibility for identifying optimal clinical benefits in a more efficient way	Validity and quality/integrity are the major concerns due to possible optional bias caused by adaptations applied
Correct wrong assumption early	Statistical methods for some specific adaptive designs (e.g., less well—understood designs) are not well-established
Select the most promising option early	Criteria for decision-making may not scientifically justifiable
Make use of emerging external information	Statistical inference (e.g., p-value and CI) may not be reliable and the overall type I error rate may not be controlled
React to surprise (positive or negative) early	The role of Data Monitoring Committee (DMC) is not clear
Speed development process	There are some obstacles for implementation

in important characteristics is problematic especially for confirmatory studies.

Adaptive Group Sequential Design – For the classical group sequential design, statistical methods and various stopping boundaries based on different boundary functions for controlling an overall type I error rate are available in the literature.^[23] With adaptations, such as sample size re-estimation based on unblinded interim analyses, the standard methods for the classical group sequential design may not be appropriate, as it may not be able to control the overall type I error rate at the desired level of 5% due to potential issues pertaining to adaptations of a given study design. Appropriate statistical procedures are necessary to avoid the potential increase in the study-wide type I error rate.^[22]

Flexible Sample Size Re-estimation Design – This design allows for sample size adjustment or re-estimation based on the observed data at interim. In general, sample size is determined before the trial formally starts based on pilot estimates of efficacy endpoints and their variability, or a best guess for the lowest clinically meaningful effect size between the treatment and control groups. Practically, parameter misspecification may be inevitable, which can lead to an underpowered design if the true variability is much larger than the initial specification of the variability.^[22] Thus, it is of interest to adjust sample sizes adaptively based on accrued data from the ongoing trial. However, sample size re-estimation suffers from the same disadvantage as the original power analysis for sample size calculation prior to the conduct of the study because it is performed by treating estimates of the study parameters, which are obtained based on data observed at interim, as true values. Sample size adjustment or re-estimation could be done in either a blinded^[24] which is based on overall data or unblinded fashion^[25] which is based on the criteria of treatment effect-size, variability, conditional power, and/or reproducibility probability.

Drop-the-losers Design – This design contains multiple stages which allow (i) dropping the inferior treatment groups, (ii) modifying treatment arms, and/or (iii) adding additional arms after the review of accumulated data at interim. Drop-the-losers design is also known as selection

design or pick-up-the-winner design. A drop-the-losers design is useful in phase II trials with the goal of finding the appropriate dose and frequency of dosing for later phase of clinical development. Typically, drop-the-losers design is the first stage of a two-stage design.^[26] At the end of the first stage, the inferior arms will be dropped based on some pre-specified criteria. The winners will then proceed to the next stage. In practice, the study is often powered for achieving a desired power at the end of the study. In other words, there may not be any statistical power for the analysis at the end of the first stage for dropping the losers. It, however, should be noted the selection criteria and decision rules play important roles for drop-the-losers designs.

Adaptive Dose Finding Design – This design is often used in early phase clinical development to identify the maximum tolerable dose (MTD), which is usually considered as the optimal dose for later phase of clinical development. In practice, it is undesirable to have too many subjects exposed to the dose limiting toxicity (DLT), while it is desirable to have a high probability for achieving the MTD with a limited number of subjects. Thus, the selection of initial dose, dose range, and criteria for dose escalation and/or dose de-escalation is important to the success of the adaptive dose finding design. In recent years, several useful adaptive dose finding designs are proposed in the literature.^[27] Among those methods, continual re-assessment method (CRM) in conjunction with Bayesian approach introduced is usually considered.^[28] For the method of CRM, the dose-response relationship is continually reassessed based on accumulative data collected from the trial. The next patient who enters the trial is then assigned to the potential MTD level. However, a potential disadvantage is that the model may predict a higher MTD due to delayed response and/or a constraint on dose-jump.^[29] It is not uncommon that the sponsors propose CRM-based approaches in their regulatory submissions. In general, CRM-based approaches are considered to be more efficient than that of the commonly used “3+3” rule with respect to accuracy and the allocation of the MTD except when the true dose is among the lower levels.

Biomarker-Adaptive Design – This design allows for adaptations based on the response of biomarkers. This design is useful in the following ways: (i) identify patient population which is most likely to respond to the test treatment under study, (ii) identify natural course of disease, (iii) early detection of disease, and (iv) help in developing personalized medicine.^[30] It should be noted that correlation between biomarker and true clinical endpoint regardless of the treatment given makes a prognostic marker. A prognostic biomarker informs distinct expected clinical outcomes, which is independent of treatment. They provide information about the natural course of the disease in individuals who have or have not received the treatment under study. However, prognostic markers cannot be used to guide to choosing a particular therapy.^[31] In other words, correlation between biomarker and true clinical endpoint does not make a predictive biomarker. A predictive biomarker informs the treatment effect on the clinical endpoint, i.e. it identifies patients who are sensitive or non-sensitive to a given agent. Biomarker-adaptive designs are typically used in exploratory studies which are important in selecting patient population for subsequent trials. This design is imbedded in a confirmatory trial to modify patient eligibility criteria after the interim look. In such a situation, statistical adjustment is needed to avoid increasing the type I error rate.^[32]

Adaptive Treatment-Switching Design – This design allows the investigator to switch a patient's treatment from an initial assignment to an alternative treatment if there is evidence of lack of efficacy, disease progression, or safety of the initial treatment. It is commonly seen in oncology clinical trials due to ethical consideration.^[33] In practice, it is not uncommon that up to 80% of patients may switch from one treatment to another in cancer research. Such a high percentage of subjects who switched due to disease progression could lead to change in hypotheses to be tested and cause further challenges in result interpretation.^[34]

Adaptive-Hypothesis Design – This design allows modifications or changes in hypotheses based on interim analysis results. Modifications of hypotheses of on-going clinical trials based on accrued data can certainly have an impact on the type I error rate, statistical power for testing the hypotheses with the pre-selected sample size for achieving the desired power. The examples include pre-planned switch from a single hypothesis to a composite hypothesis or multiple hypotheses, pre-planned switch the null hypothesis and the alternative hypothesis, and pre-planned change due to switch between the primary study endpoint and the secondary endpoints.

Phase I/II (or Phase II/III) Adaptive Seamless Design – This design combines the study objectives, which are traditionally addressed in separate trials, into one single study. Most commonly used adaptive seamless designs include (i) adaptive seamless phase I/II design, and (ii) adaptive seamless phase II/III design. An adaptive seamless phase I/II design is referred to a design that combines a phase I

trial which usually aims to find MTD for an investigational drug, and a phase II trial which examines the efficacy of the drug at the identified MTD. An adaptive seamless phase II/III design is a two-stage design consisting of a so-called learning or exploratory stage (phase II) and a confirmatory stage (phase III). Thus, an adaptive seamless design may reduce study sample size as compared to traditional designs. In addition, an adaptive seamless design is more efficient because there is no lead time between the two stages. For an adaptive seamless phase II/III design, a typical approach is to power the study for the phase III confirmatory phase and obtain valuable information with certain assurance at the phase II learning stage.

Multiple Adaptive Design – This design includes any combination of the above mentioned adaptive designs. A multiple adaptive design is more flexible but more problematic. Although a multiple adaptive design is more attractive, it makes good statistics practice (GSP) more difficult and challenging in practice.

CONCLUSION

In pharmaceutical development, each design has its own merits and limitations under different circumstances at various stages. How to select an appropriate design for pharmaceutical development when planning a clinical trial is an important question. The answer to this question depends on many factors which include (i) number of treatments to be compared, (ii) characteristics of the treatments, (iii) study objectives, (iv) availability of experimental units (e.g., subjects), (v) inter-subject and intra-subject variability, (vi) duration of the study, and (vii) dropout rates. In conclusion, there is no best design, and the valid statistical designs play an important role to optimize the quality of drug product in pharmaceutical development.

REFERENCES

1. Chow, S.C. and Liu, J.P. Statistical Design and Analysis in Pharmaceutical Science. Marcel Dekker: New York, 1995.
2. FDA. *Guidance for Industry – Process Validation: General Principles and Practices*. CDER/CBER/CVM, US Food and Drug Administration, Rockville, MD, January, 2011.
3. Winkle, H.N. Implementing quality by design. Presented at PDA/FDA Joint Regulatory Conference Evolution of the Global Regulatory Environment: A Practical Approach to Change. Washington, DC, September 24, 2007.
4. FDA. Analytical procedure and methods validation for drugs and biologics, Rockville, MD, 2015.
5. Myers, R.H. Response Surface Methodology. Allyn and Bacon: Boston, MA, 1976
6. Hicks, C.R. Fundamental Concepts in the Design of Experiments. 3rd Edition, CBS College Publishing: New York, 1982.

7. Helboe, P. New designs for stability testing programs: Matrix or factorial designs. Authorities' viewpoint on the predictive value of such studies, *Drug Information Journal* **1992**, 26, 629–634.
8. Nordbrock, E. Statistical comparison of stability study designs. *Journal of Biopharmaceutical Statistics* **1992**, 2, 91–113.
9. DeWoody, K.; Raghavarao, D. Some optimal matrix designs in stability studies. *Journal of Biopharmaceutical Statistics* **1997**, 7, 205–213.
10. Pong, A.; Raghavarao, D. Comparison of bracketing and matrixing designs for a two-year stability study. *Journal of Biopharmaceutical Statistics* **2000**, 10, 217–228.
11. Chow, S. *Statistical Design and Analysis of Stability Studies*. Chapman and Hall/CRC Press: New York, 2007.
12. Murphy, J.R. Uniform matrix stability study designs. *Journal of Biopharmaceutical Statistics* **1996**, 6, 477–494.
13. Ju, H.L.; Chow, S.C. On stability designs in drug shelf-life estimation. *Journal of Biopharmaceutical Statistics* **1995**, 5, 201–214.
14. Nordbrock, E. Use of statistics to establish a stability trend: Matrixing. *Pharmaceutical Stability Testing to Support Global Markets*. 203–209, **2010**.
15. Gad, S.; Weil, C.S. *Statistics and Experimental Design for Toxicologists*. Galdwell: Telford Press, NJ, 1988.
16. Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*; 3rd Ed.; Chapman and Hall/CRC Press, Taylor & Francis: New York, 2009.
17. Jones, B.; Kenward, M.G. *Design and Analysis of Cross-over Trials*; 2nd Ed., Chapman and Hall, London, 2003.
18. ICH E9. International Conference on Harmonization Guideline E9, Statistical Principles for Clinical Trials. Geneva, Switzerland, 1998.
19. Rodda, B.E.; Tsianco, M.C.; Bolognese, J. A.; Kersten, M.K. Clinical development. In *Biopharmaceutical Statistics for Drug Development*, Peace, K.E., Ed.; Dekker: New York, 1998.
20. Ting, N. *Dose Finding in Drug Development*. Springer: New York, 2006.
21. Chow, S.C.; Liu, J.P. *Design and Analysis of Clinical Trial*; 3rd Ed.; John Wiley & Sons: New York, 2013.
22. FDA. Draft Guidance for Industry—Adaptive Design Clinical Trials for Drugs and Biologics. US Food and Drug Administration: Rockville, Maryland, 2010.
23. Chow, S.C.; Chang, M. *Adaptive Design Methods in Clinical Trials*; 2nd Ed.; Chapman and Hall/CRC Press: New York, 2011.
24. Gould, A.L. Planning and revising the sample size for a trial. *Statistics in Medicine* **1995**, 14, 1039–1051.
25. Cui, L.; Hung, H.M.J.; Wang, S.J. Modification of sample size in group sequential trials. *Biometrics* **1999**, 55, 853–857.
26. Thall, P.F.; Simon, R.; Ellenberg, S.S. A two stage design for choosing among several experimental treatments and a control in clinical trials. *Biometrics* **1989**, 45, 537–547.
27. Pong, A. Guest-Editor's Note: "Statistical Issues in Adaptive Design Methods in Clinical Trials." *Journal of Biopharmaceutical Statistics* **2007**, 17, 1133–1134.
28. Storer, B.E. Design and analysis of phase I clinical trials. *Biometrics* **1989**, 45, 925–937.
29. O'Quigley, J.; Pepe, M.; Fisher, L. Continual reassessment method: a practical design for Phase I clinical trials in cancer. *Biometrics* **1990**, 46, 33–48.
30. Freidlin, B.; Korn, E.L. Biomarker-adaptive clinical trial designs. *Pharmacogenomics* **2010**, 11, 1679–82.
31. Freidlin, B.; McShane, L.; Korn, E.L. Randomized clinical trials with biomarkers: design issues. *J. Natl. Cancer Inst.* **2010**, 102, 152–160.
32. Sargent, D.J.; Conley, B.A.; Allegra, C.; Collette, L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J. Clin. Oncol.* **2005**, 23, 2020–2027.
33. Shao, J.; Chang, M.; Chow, S.C. Statistical inference for cancer trials with treatment switching. *Statistics in Medicine* **2005**, 24, 1783–1790.
34. Branson, M.; Whitehead, J. Estimating a treatment effect in survival studies in which patients switch treatment. *Statistics in Medicine* **2002**, 21, 2449–2463.

Pharmacodynamic Issues

Cheng-Tao Chang

Novo Nordisk Inc., Princeton, New Jersey, U.S.A.

Robert L. Wong

Abbott Laboratories, Parsippany, New Jersey, U.S.A.

Abstract

This entry provides a definition for pharmacodynamics and describes a number of factors that affect the selection of statistical methods for pharmacodynamics evaluations.

Keywords: Clinical trial; Parametric approach; Pharmacodynamics.

INTRODUCTION

There are several ways of defining pharmacodynamics. Here are three examples.

- i. The branch of pharmacology dealing with the reactions between drugs and living systems.^[1]
- ii. The study of the biochemical and physiological effects of drugs and their mechanisms of action on the body.^[2]
- iii. The study of how drugs can affect the body.^[3]

After being swallowed, injected, inhaled, nasally administered, or absorbed through the skin, most drugs enter the bloodstream, circulate throughout the body, and interact with a number of target sites. Depending on its property or route of administration, a drug may act in only a specific area of the body (e.g., the action of antacids is confined largely to the stomach). When a drug interacts with the target tissues (e.g., cells or organs), a desirable therapeutic effect usually takes place, which is referred to as efficacy or effectiveness in clinical trials. On the other hand, if a drug interacts with cells, tissues, or organs that are not the desired targets, side effects may occur. In clinical trials, side effects are also known as adverse events (AEs) or adverse reactions (ARs). Hence, in a Phase III clinical trial pharmacodynamic (PD) endpoints are primarily safety or efficacy (or both) parameters. Efficacy endpoints are generally more heterogeneous than safety endpoints from one therapeutic area to another because the majority of the safety endpoints are usually restricted to a limited number of specific evaluations, including physical examinations, electrocardiogram (ECG), laboratory variables, vital signs, and adverse-event reporting; drug-level monitoring may also be used to assess safety. For this reason, the focus of the statistical methodologies in this entry is on efficacy parameters and not on safety parameters.

OVERVIEW

The major organ systems in the body include cardiovascular, respiratory, nervous, skin, musculoskeletal, blood, digestive, endocrine, urinary, and reproductive. An investigational drug may be evaluated for efficacy and/or safety for one or more of these systems. In the United States, an investigational new drug (IND) application containing chemistry, toxicity, and other pre-clinical data is submitted to the U.S. Food and Drug Administration (FDA). Other countries utilize a similar process, allowing clinical studies to begin after a review of clinical data. After the investigational drug has been evaluated in clinical trials, a new drug application (NDA) is filed by its sponsoring drug company; this is then reviewed and, hopefully, approved by the Center for Drug Evaluation and Research (CDER) of the FDA. As of 2009, CDER had organized the Office of New Drugs into six divisions: Office of Drug Evaluation I, Office of Drug Evaluation II, Office of Drug Evaluation III, Office of Antimicrobial Products, Office of Nonprescription Products, and Office of Oncology Products. Each division is responsible for evaluating new drugs in some specific therapeutic areas. It is then helpful to understand the therapeutic drug involved and which division of the CDER will be handling the NDA submission; Some meetings are usually held between the division involved and the drug sponsor before IND application and NDA submission in order to reach mutual understanding and agreement about how the investigational drug will be conducted in clinical trials, what type of relevant data will be collected, what endpoints are considered primary and/or secondary, and what statistical consideration is deemed necessary for any study-related issues, and so on.

This entry is organized as follows: The “Introduction” first defines pharmacodynamics. The section “Overview” provides a general overview of some of the primary endpoints for some of the most common diseases within

different organ systems when evaluating pharmacodynamic effects of investigational drugs used in Phase III clinical trials. The section “Factors Affecting the Selection of Statistical Methods for Pharmacodynamic Evaluations” describes a number of factors that affect the selection of statistical methods for pharmacodynamics evaluations. “Other Issues” addresses some additional issues related to Phase II/III pharmacodynamic evaluations.

These pharmacodynamic endpoints discussed in the remainder of this section are often used in clinical trials and agreed upon with regulatory health authorities for the specific indications.

CARDIOVASCULAR SYSTEM

Coronary Artery Disease/Angina

Patient mortality is the most important clinical endpoint in trials assessing the beneficial effects of drugs on coronary artery disease. Such studies often are long term (e.g., five or more years in duration) and may recruit thousands of patients. Patient morbidity (e.g., revascularization, stroke, nonfatal myocardial infarction) are also assessed in such trials. Subjective chest pain and exercise tolerance (stress test) are common endpoints in angina trials.

Hypertension

The mean change from baseline in systolic and diastolic blood pressure and cardiovascular mortality and morbidity are endpoints used in hypertension trials.

Congestive Heart Failure

Patient mortality, exercise tolerance, the number of hospitalizations, and cardiovascular morbidity are common endpoints in trials assessing the effects of drugs in congestive heart failure.

Hyperlipidemia

Low-density lipoprotein, cholesterol levels, and cardiovascular morbidity are endpoints in trials assessing the effect of lipid-lowering agents.

RESPIRATORY SYSTEM

Asthma

Asthma is a very common reaction airway disease. A clinically relevant primary endpoint used in asthma trials is a change in forced expiratory volume in 1 sec (FEV1).

Such FEV1 measurements are determined by spirometry evaluations.

NERVOUS SYSTEM

Alzheimer's Disease

Cognitive and functional scales specifically designed to assess Alzheimer's disease are used as standard endpoints in Alzheimer's trials. Such trials often require that hundreds to thousands of patients be recruited.

Parkinson's Disease

Functional scales specifically designed to assess Parkinson's disease are typically used in evaluating drugs in this disease.

SKIN

Acne

Lesion count and the nature of the lesions (e.g., papular, pustular counts) are commonly used endpoints in acne trials.

Psoriasis

Plaque size, induration, scaling, and redness are generally accepted endpoints in psoriasis trials. Other parameters to be considered include the extent and the severity of psoriatic lesions.

MUSCULOSKELETAL SYSTEM

Osteoarthritis

Tender joints and pain-functional endpoints (e.g., Western Ontario and McMaster Universities Osteoarthritis Index, or WOMAC) are assessed in osteoarthritis trials. The WOMAC is a self-administered questionnaire that is highly reliable and sensitive to changes in health status in patients with osteoarthritis.^[4]

Rheumatoid Arthritis

Endpoints used in trials assessing the safety and efficacy of drugs for rheumatoid arthritis include the number of swollen and painful joints and patient and physician's global assessment of well-being. Radiographic progression of joint destruction is being assessed in current trials with rheumatoid arthritis.

GASTROINTESTINAL SYSTEM

Peptic Ulcer Disease

Endoscopic evidence of improvement or resolution of an ulcer is used as an endpoint in ulcer trials. Symptomatic improvement alone following administration of the investigational drug is *not* sufficient proof of benefit in Phase III ulcer trials.

ENDOCRINE SYSTEM

Diabetes

Diabetes is a systemic disease that may affect multiple organs, including the skin (dermopathy), blood vessels (peripheral vasculopathy), kidney (nephropathy), eyes (retinopathy), and nervous system (neuropathy). Clinical trials in diabetes could assess these clinical endpoints, but they also require laboratory endpoints such as hemoglobin A1C and fasting blood glucose levels.

Osteoporosis

The incidence of bone fracture is a primary efficacy endpoint in osteoporosis trials; because these trials may require a very long duration of follow-up for evaluating bone fractures (e.g., up to five years), bone density has been used as a surrogate marker in osteoporosis trials.

OTHER AREAS

Infectious Diseases

Infections can involve any body system and therefore have many primary endpoints that can be used in clinical trials. The absence of an infectious agent by culture and mortality (e.g., sepsis trials) may be used. In addition, surrogate markers (e.g., viral load in acquired immunodeficiency syndrome [AIDS] studies) have been used.

Oncology

The incidence of patient mortality is an endpoint in many oncology trials. In those instances where there is an extremely high mortality rate, alternative endpoints such as the time of death and tumor burden can be used.

Transplantation

In recent years there has been an increase in the number of drugs approved for use in transplantation as primary or adjunctive immunosuppression for kidney, liver, and heart transplantation. Phase III clinical trials currently use the incidence of acute rejection episodes and patient and graft

survival as necessary endpoints for approval by health authorities.

FACTORS AFFECTING THE SELECTION OF STATISTICAL METHODS FOR PHARMACODYNAMIC EVALUATIONS

The choice of statistical method employed to analyze the pharmacodynamic parameters of interest depends on a number of factors:

- A. Study objective and design.
- B. Type and number of explanatory variables and response variables.
- C. Distribution of the PD parameter.
- D. Model assumptions.
- E. Therapeutic area of the study drug.
- F. Frequency of evaluation of the PD parameter.
- G. Measurements of the PD parameter.
- H. Regulatory concerns of the conduct of a clinical trial.

STUDY OBJECTIVE AND DESIGN

The objectives for Phase III clinical studies may include one or more of the following four objectives:^[5]

- i. Demonstrate/confirm efficacy.
- ii. Establish a safety profile.
- iii. Provide an adequate basis for assessing the benefit-risk relationship to support labeling.
- iv. Establish the dose-response relationship.

A study objective will define what PD parameter is to be considered as the primary clinical endpoint and what comparisons or investigations are deemed most clinically relevant. In Phase III clinical trials, the primary PD parameter is usually an efficacy variable because the primary objective of most Phase III trials is to provide strong scientific evidence on the efficacy of the study drug. The design of the study to evaluate the investigational drug is driven by the study objective. Study design determines how the data from the clinical study are to be collected while minimizing bias and variability from possible extraneous sources in a clinical study. Additionally, a study design that appears to be the most effective one from a statistical perspective may not be suitable or feasible for the study from a clinical or pharmacological perspective.

The most popular study design in Phase III is a parallel group design (known as completely randomized design). Other designs, such as crossover design and factorial design, are also frequently used.

Parallel group design is the simplest design of all. A parallel group design may be extended to include one additional explanatory variable that is used as a blocking

factor. Such a parallel group design with a blocking factor is known as a randomized block design. A parallel group design may include repeated evaluation of the same PD variable over time, which is known as a repeated measurements design.

Usually, there are two or more treatment arms involved in clinical trials, often with one arm assigned to placebo (a placebo-controlled study) or to an active comparator agent (an active-controlled study). Placebo may not be used in any arm because of ethical or other reasons.

Each patient typically receives only one treatment in a parallel group design. In clinical trials where all the treatments are administered to each patient so that he or she can serve as his or her control, a crossover design is the study design of choice. A factorial design is highly desirable when two or more treatments are evaluated simultaneously in the same set of patients through the use of varying combinations of the treatment or when the combinations of two or more explanatory variables (including the treatment) are of interest.

TYPE AND NUMBER OF EXPLANATORY VARIABLES AND RESPONSE VARIABLES

Explanatory variables are also known as independent variables or simply as input variables. Response variables are also called dependent variables or output variables. The response variables throughout this entry are referred to as clinical responses, that is, the PD parameter.

Generally, any data can be classified as either categorical or numerical. A categorical variable is either nominal, when it does not have a natural ordering (e.g., gender), or ordinal, when it is associated with a certain order (e.g., condition of disease, such as mild, moderate, or severe). A numerical variable can be either continuous or discrete. Theoretically, a continuous variable is not countable and can take on any possible value in a certain interval. On the other hand, a discrete variable is countable and in many cases assumes only a finite number of values. A categorical variable, in fact, can be treated as a numerical variable when we can artificially assign numerical values to the categories of a categorical variable. Of course, it is understood that the numerical values of a categorical variable with a nominal scale do not have any meaning in terms of ordering. Based on this argument, we can simply say that a PD variable can be classified as either continuous or discrete (with a nominal scale or an ordinal scale).^[6,7]

In reality, all variables collected from clinical trials should be classified as discrete variables for the following reasons: (i) they are collected from a finite number of patients; and (ii) they are countable. Therefore, the fundamental idea of determining whether a variable is continuous or discrete is dependent upon the number of different values obtained from that variable of interest.

If a variable takes on a small number of values (there is no specific rule as to how small is *small*—it should be relative to the sample size), we can conclude that the variable is discrete. In contrast, if data values encompass a relatively wide range of different values, the variable can be considered continuous.

A commonly used statistical method for evaluating a continuous variable if the data are normally distributed is analysis of variance. In contrast, Fisher's exact test or chi-square test may be applied for categorical-data analysis (i.e., a discrete variable with a nominal or ordinal scale). A continuous variable may be classified as a discrete variable with an ordinal scale. For example, we can categorize a response to antihypertensive therapy in terms of the systolic blood pressure's being either no less than 20% reduction from baseline or less than 20% reduction from baseline.

Additionally, the number of variables (either explanatory or response) or the number of levels for each variable in a clinical trial also affects the selection of certain statistical methods. Here are some examples.

- i. A two-sample *t*-test may be employed if there are only two levels for the only explanatory variable with a continuous scale (i.e., two doses for the treatment), but the same *t*-test cannot be applied if there are three or more treatment levels. In the case in which there are three or more treatment levels, an analysis of variance may be applied.
- ii. Two-way analysis of variance or analysis of covariance may be employed if two explanatory variables are involved.
- iii. A multivariate approach may be considered if there is more than one primary PD parameter to be considered simultaneously.

DISTRIBUTION OF THE PHARMACODYNAMIC PARAMETER

A parametric approach is generally preferred when the PD variable can be justified to be fairly normally distributed. When the continuous PD variable is skewed or not normally distributed, a non-parametric method may be used. In general, a parametric method is more powerful than a non-parametric method. Note that the power of a test is the probability that the null hypothesis is rejected when it is not true.

Sometimes the skewness of the data is attributed to the existence of outliers. The discussion on handling outliers will be detailed under "Regulatory Concerns of the Conduct of a Clinical Trial" because the consideration of removing outlier(s) from a statistical analysis is based not only on statistical concerns but also on regulatory concerns.

As for discrete variables, the most common distributions are binomial, multinomial, and Poisson. Logit (log of odds)

models are used with binomial and multinomial distributions. The log-linear model is used with the Poisson distribution.

MODEL ASSUMPTIONS

A parametric approach is used in a conventional statistical model along with a number of assumptions. The assumption of normality is only the first step to determine the suitability of a parametric method. There are other conditions required for a specific model, including the following:

- i. *Independence*: Most of the models require independence of the data. This means that the observations from different treatments are obtained from different patients. A different model needs to be considered when observations are dependent, such as the data from a study with a crossover design or from a study with repeated measures.
- ii. *Homogeneous variance*: Equal variability is required in the approach with conventional analysis of variance. When this requirement is not met, there are two possible solutions: (i) use an algebraic transformation such as log to reduce the observations in a much smaller scale, in hopes of achieving the equality of variance; (ii) employ a more advanced method, such as a generalized *F*-test^[8] that does not require equal variance in the model. However, the software program to accommodate such methodologies for exact tests does not exist in SAS.

THERAPEUTIC AREA OF THE STUDY DRUG

In almost all cases, time should be considered a continuous variable. Intuitively, we may think about applying analysis of variance to analyze “time” when it is the primary PD parameter, but this is not always the case. For example, in a clinical trial to assess an analgesic drug, the onset time (the time that the degree of pain starts to diminish) can be a primary PD parameter. Analysis of variance can be applied to compare the mean onset time among different treatment groups. However, in cancer trials, survival analysis may be more appropriate (e.g., in a patient with cancer, the time to survival after a treatment). In summary, the mean time can be generally compared with a *t*-test or analysis of variance, but the time to the first predefined event is usually analyzed with a survival analysis.

FREQUENCY OF EVALUATION OF THE PHARMACODYNAMIC PARAMETER

When a PD parameter is evaluated more than once throughout a clinical trial, it is more appropriate to employ

a method that takes into account the repeated measures. A method such as one-way analysis of variance with repeated measures may be appropriate for such a trial.

MEASUREMENTS OF PHARMACODYNAMIC PARAMETERS

The measurement of a primary PD parameter in a clinical trial protocol must be clearly defined. For example, it is not adequate to indicate that mortality is the primary parameter in a cancer trial. Mortality can be assessed in many ways: the proportion of patients alive at a pre-determined point in time, the overall survival curve, the mean survival time during a fixed time interval, or a combination of assessments.

When a continuous PD variable has been determined, the change from baseline or the percentage change from baseline may be of most interest. One major concern with such a measure is the direction of the change, which may be above or below baseline, so the average of such a measure from all patients could be misleading; for example, the elevation or decrease of blood glucose from baseline is an important indication for a diabetes trial assessing a hypoglycemic agent. On the other hand, if only the magnitude of the change (or percentage change) is considered, the probability distribution of this absolute value of this derived PD variable may not be easily identified and would greatly impact the conduct of the hypothesis testing for the trial.

As discussed earlier, a continuous variable can be converted into a discrete variable with an ordinal scale. When such a conversion is deemed clinically relevant, the employment of statistical methods should be adjusted accordingly.

REGULATORY CONCERNS ABOUT THE CONDUCT OF A CLINICAL TRIAL

In January 1997, the FDA published a draft guideline on statistical principles for clinical trials.^[9] It addresses statistical considerations about handling a number of potential issues for a clinical trial. They include the following.

Type of Comparison

This is closely related to the study objective and will determine how the hypothesis testing should be set up: a two-sided test, a one-sided test (a right-tailed test or a left-tailed test). The following three possible comparisons are usually demonstrated in Phase III trials.

Superiority Trials

These are meant to demonstrate that the study drug is superior to placebo in a placebo-controlled study or is superior

to an active control treatment. Generally, superiority trials are assumed, unless it is explicitly stated otherwise.

Equivalence Trials

Some examples:

- i. Pharmacokinetic bioequivalence trials to show two formulations are bioequivalent or a generic drug is bioequivalent to an innovator drug
- ii. Show equivalence between a comparator agent at one dose level and the study drug at a lower dose level
- iii. Show equivalence between two different dosing regimens of the study drug with an identical total dose, such as 150 mg twice a day (b.i.d.) vs. 100 mg three times a day (t.i.d.).

Non-inferiority Trials

These can be seen in many active-controlled studies to show that the efficacy of an investigational drug is no worse than that of the comparator drug. In some situations, a trial can be designed to add a placebo arm to an active-controlled study so that the objectives of the study are to establish superiority to the placebo and to show a similar efficacy (and/or safety) profile to the comparator agent.

Patient Population Included in the Statistical Analysis

This does not affect the decision-making in terms of statistical approaches, but it would certainly influence the results and interpretation of the results. The following two types of patient populations are generally allowed in the statistical analysis.

All Randomized Patients

An analysis based on this population is also well known as an “intent-to-treat” analysis. In theory, the analysis should include all eligible patients who were randomized to the study treatments. In practice, it often excludes patients who did not receive any dose of study treatments (including placebo) or did not have any data post-randomization. Two major problems arise in this case: First, should we include patients who may fall into one of the following problems: protocol violation, error in treatment assignment, or use of prohibited medications during the study? Second, in which treatment group should a patient be placed in the statistical analysis if he is not treated with the assigned study drug? For example, should the patient be in treatment group A or B if he actually received treatment B but was randomized to treatment A? There is no clear direction in the guidelines as to how to resolve these problems; they should be handled case by case.

Per-Protocol Patients

An analysis using this patient population is referred to as evaluable patients analysis. Evaluable patients should be defined as concisely as possible in the protocol or at latest before the unblinding of the treatment. Any change to the definition after unblinding has to be justified.

Handling Missing Values and Outliers

Missing values do occur in almost every clinical trial. Unfortunately, there is no universally applicable method for handling missing values. The best way is to predefine the methods to be used for missing values in the analysis plan, especially for the study that may involve a substantial number of missing values. As with the handling of missing values, the statistical method for detecting outliers should be well described before the treatment codes are unblinded. The removal of outliers from statistical analysis will be more convincing when the outliers can be justified clinically as well as statistically.

SUMMARY

A selection of a specific statistical method for evaluating PD parameters in a clinical trial should be carefully based on all of the factors just discussed. This will help a sponsoring pharmaceutical company to get the study drug approved by the FDA by employing an appropriate statistical method (it may be a very simple one) that is acceptable from statistical, medical, and regulatory standpoints.

OTHER ISSUES

There are a number of additional issues that have received specific attention by the FDA for evaluations of new drugs. Since there is no absolute statistical method to handle these specific topics, these will be discussed from clinical and regulatory perspectives.

Surrogate Markers

Surrogate markers are now being considered as appropriate target endpoints for various diseases in clinical trials, especially in those diseases where the surrogate marker has been validated (e.g., acute rejection episodes as a predictor of chronic organ rejection), where medical need may justify a rapid assessment (e.g., T-cell count because of the high mortality from AIDS), or where the duration of a trial (bone fracture endpoint in osteoporosis trials) or the number of patients required for enrollment (e.g., rare medical conditions) may be problematic.

A surrogate endpoint is an endpoint that is intended to relate to a clinically important outcome but does not in

itself measure a clinical benefit. Surrogate endpoints may be used as primary endpoints when appropriate (when the surrogate is reasonably likely or well known to predict a clinical outcome^[10]). The clinical conclusions may be misleading if the selected surrogate endpoints for the studies are not vigorously established. Fleming^[11] discusses details of using surrogate markers in clinical trials, and Fleming and DeMets^[12] present examples of surrogate endpoints that failed to be validated in clinical trials.

Subgroup Considerations

Particular subgroups deserve special PD consideration because of unique characteristics, thereby requiring the need to evaluate such subgroups separately in clinical trials. These subgroups include the elderly and pediatric populations, different races, gender differences, and medical conditions that could affect PD interpretation. It would therefore be appropriate in these subgroups to analyze the PD effects of a drug separate from the rest of the trial population.

Elderly Population

It is well known that the elderly population is different from the younger population with respect to sensitivity to many drugs.^[13] For example, patients treated with nonsteroidal anti-inflammatory agents have a higher incidence of gastrointestinal bleeding than do younger patients. The FDA also provides guidance on the need for collecting information on geriatric patients.^[14] There is no clear definition of a geriatric population; normally this includes patients 65 years old and older. However, the FDA encourages drug companies to include patients in clinical studies, if possible, who are 75 years old and older.

Pediatric Population

Body surface area is different in very young children, compared to adolescents or adults; this often requires dosing of a drug on a milligram per kilogram basis compared to adults, among whom body surface area is not dramatically different. Bowel length is much shorter in very young children, which is well known to affect the pharmacokinetics of a drug and therefore the PD endpoints in a clinical trial.

When pediatric patients are to be considered for participation in clinical trials, safety data from adult patients should be available.^[15] In addition, for a drug expected to be used in children, evaluation should be made in the appropriate group. When a drug is to be evaluated in children, it is usually appropriate to begin with older children before extending the trial first to younger children and then to infants.^[5]

Racial Differences

When a new drug is to be developed globally, the drug may be tested among different ethnic groups. It is highly

suggested by the FDA that such a drug be characterized as ethnically sensitive or insensitive. A new chemical entity (NCE) is said to be sensitive to ethnic factors when its pharmacokinetic, PD, or other characteristics suggest the potential for clinically significant impact by intrinsic and/or extrinsic ethnic factors on safety, efficacy, or dose response. Extrinsic ethnic factors are factors associated with the patient's environment and/or culture. Extrinsic factors tend to be less genetically and more culturally and behaviorally determined. Examples are medical practice, diet, socioeconomic status, and exposure to pollution and sunshine. Intrinsic ethnic factors are characteristics associated with the drug recipient. These are factors that help to define and identify a subpopulation and that may influence the ability to extrapolate clinical data between ethnic groups (or between regions). Examples of intrinsic ethnic factors include polymorphism, age, gender, height, weight, lean body mass, body composition, and organ dysfunction.^[16]

The following is an example of a drug showing its sensitivity to ethnic factors: African Americans have been reported to have less drug bioavailability from immunosuppressive drugs administered after transplantation, which has been attributed to lower absorption compared to the Caucasian population. This pharmacokinetic difference has a significant impact on the PD endpoint of acute rejection because African Americans who receive kidney transplants are known to have poorer graft survival.^[17]

Gender Differences

Because of unique anatomy and physiology, women are known to have complex pharmacokinetics different from men.^[18] Although pharmacokinetic differences between men and women have been extensively reported, less is known about PD differences between men and women.

Issues related to pregnancy and oral contraceptive use appear to be important.^[19] Pregnancy may increase the metabolism of anti-epileptic drugs, and oral contraceptives are well known to interfere with the metabolism of many drugs; on the other hand, certain drugs may impair contraceptive efficacy. Women have been reported to be more prone to develop torsade de pointes from drugs such as quinidine and procainamide than men, and women have been reported to have less sensitivity to insulin than men. Ongoing research in this area will delineate additional PD differences between men and women.

Patients with Impaired Renal or Liver Function

After entering the body, a drug is eliminated either by excretion or by metabolism. Whereas elimination can occur via any of several routes, most drugs are cleared either by elimination of unchanged drug by the kidney or by metabolism in the liver. For a drug eliminated primarily via renal excretory mechanisms, impaired renal function may alter its PK and PD to an extent that the

dosage regimen needs to be changed from that used in patients with normal renal function. Whereas the most obvious type of change arising from renal impairment is a decrease in renal excretion (or possibly renal metabolism) of a drug or its metabolites, renal impairment has also been associated with other changes, such as changes in hepatic metabolism, plasma protein binding, and drug distribution. These changes may be particularly prominent in patients with severely impaired renal function and have been observed even when the renal route is not the primary route of elimination of a drug.^[20]

Meta-Analysis

The results from more than one clinical trial may be combined in order to draw a more robust clinical conclusion. A meta-analysis of clinical trials may also be necessary when data exist from multiple studies that are each insufficiently powered to reach statistically significant conclusions or if each study has a limited number of enrolled patients (because of logistical or other reasons). There is a substantial amount of literature published in this area [e.g., Hedges and Olkin,^[21] Hunter and Schmidt,^[22] and Geller and Proschan^[23]].

Pharmacokinetic/Pharmacodynamic Relationships

During recent drug development and evaluation efforts, PK/PD relationships have been a focus because the combination of these two areas, through an appropriate modeling, will lead to the therapeutically most relevant relationship between pharmacological effect and time and will produce an optimal dose recommendation. Pharmacokinetic data and PD data can be linked directly or indirectly. The direct or indirect link will play an important role in determining an appropriate mathematical PK/PD model.^[24]

CONCLUSION

As drug development advances at an unprecedented pace with much more complicated technology, the challenge of encountering more and more—not-easy-to-be-resolved medical and statistical issues—is inevitable. This certainly will make the pharmaceutical industry one of the most attractive and challenging opportunities for statisticians.

REFERENCES

1. Merriam-Webster's Collegiate Dictionary, 10th Ed.; Merriam-Webster: Springfield, MA, 1994; 871.
2. Goodman, A.G.; Goodman, L.S.; Rall, T.W.; Murad, F. *The Pharmacological Basis of Therapeutics*, 7th Ed.; MacMillan: New York, 1985.
3. Merck Manual of Medical Information, Home Ed.; Merck: Pennsylvania, 1997; 31–34.
4. Bellamy, N.; Buchanan, W.W.; Goldsmith, C.H.; Campbell, J.; Stitt, L.W. Validation study of WOMAC: A health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J. Rheumatol.* **1988**, *15*, 1833–1840.
5. *International Conference on Harmonization E8: Guidance on General Considerations for Clinical Trials*, U.S. Food and Drug Administration. 1997.
6. Selwyn, M.R. *Principles of Experimental Design for the Life Science*; CRC Press: Boca Raton, FL, 1996.
7. Agresti, A. *Categorical Data Analysis*; Wiley: New York, 1990; 2–7.
8. Weerahandi, S. *Exact Statistical Methods for Data Analysis*; Springer Verlag: New York, 1994.
9. *International Conference on Harmonization E9: Note for Guidance on Statistical Principles for Clinical Trials*, 1997.
10. Sheskin, D.J. *Handbook of Parametric and Nonparametric Statistical Procedures*; CRC Press: Boca Raton, FL, 1996.
11. Fleming, T.R. Surrogate endpoints in clinical trials. *Drug Inf. J.* **1996**, *30*, 545–551.
12. Fleming, T.R.; DeMets, D.L. Endpoints in clinical trials: Are we being misled? *Ann. Intern. Med.* **1996**, *125*, 605–613.
13. Swift, C.G. Pharmacodynamics: Changes in homeostatic mechanisms, receptor and target organ sensitivity in the elderly. *Br. Med. Bull.* **1990**, *46*, 36–52.
14. *International Conference on Harmonization E7: Studies in Support of Special Populations: Geriatrics*; 1994.
15. *International Conference on Harmonization M3: Non-Clinical Safety Studies for the Conduct of Human Clinical Trials for Pharmaceuticals*; 1997.
16. *International Conference on Harmonization E5: Ethnic Factors in the Acceptability of Foreign Clinical Data*; 1997.
17. Katznelson, S.; Gjertson, D.W.; Cecka, J.M. The Effect of Race and Ethnicity on Kidney Allograft Outcome. *Clinical Transplants 1995*; UCLA Tissue Typing Laboratory: Los Angeles, CA, 1995, 379.
18. Fletcher, C.V.; Acosta, E.P.; Strykowski, J.M. Gender differences in human pharmacokinetics and pharmacodynamics. *J. Adolesc. Health* **1994**, *15*, 619–629.
19. Harris, R.Z.; Benet, L.Z.; Schwartz, J.B. Gender effects in pharmacokinetics and pharmacodynamics. *Drug* **1995**, *50*, 222–239.
20. *International Conference on Harmonization: Pharmacokinetics and Pharmacodynamics in Patients with Impaired Renal Function: Study Design, Data Analysis, and Impact on Dosing and Labeling*; 1997.
21. Hedges, L.V.; Olkin, I. *Statistical Methods for Meta-Analysis*; Academic Press: New York, 1985.
22. Hunter, J.E.; Schmidt, F.L. *Methods of Meta-analysis*; SAGE Publications: Beverly Hills, CA, 1990.
23. Geller, N.L.; Proschan, M. Meta-analysis of clinical trials: A consumer's guide. *J. Biopharm. Stat.* **1996**, *6*, 377–394.
24. Derendorf, H.; Hochhaus, G. *Handbook of Pharmacokinetic/Pharmacodynamic Correlation*; CRC Press: Boca Raton, FL, 1995, 79–120.

Pharmacodynamics with Covariates

Cheng-Tao Chang

Novo Nordisk Inc., Princeton, New Jersey, U.S.A.

Robert L. Wong

Abbott Laboratories, Parsippany, New Jersey, U.S.A.

Abstract

This entry discusses several univariate statistical methods that are used mostly for Phase III clinical studies for considering explanatory factors other than treatment variable and pharmacodynamic (PD) response.

Keywords: ANOVA; Covariate; Linear regression; Logistic regression.

INTRODUCTION

In this entry, we will discuss several univariate statistical methods that are used mostly for Phase III clinical studies for considering explanatory factors other than treatment variable and pharmacodynamic (PD) response. We specify the underlying conditions (including study design, model assumptions, and number of data points per patient) that apply to each method. Note that the “explanatory variable” refers to treatment (or dosage level) and any other variable used as the stratified variable or covariate. In this entry, we will classify the explanatory variable as a “treatment” variable or as an “other explanatory” variable. “Response variable” refers to the PD parameter, and we will consider only the clinical trials with a single primary PD parameter.

We will divide clinical trials into five groups according to the type of PD parameter and the number and type of explanatory variables. Table 1 displays the summary of these five groups.

The discussion of each statistical method will cover purpose, statistical model, hypothesis testing, test statistic, rejection region, confidence interval, two-sided p -value, description of and notes on the method, and any related SAS procedure. (SAS, the most popular statistical software package, is also accepted by the Food and Drug Administration as one of the standard packages.

The statistical methods discussed for each group are listed in Table 2.

STATISTICAL METHODS

Group 1

Two-Way Analysis of Variance

Study Design: Parallel group design with a blocking factor.

Model Assumptions: Normal distribution, independence, equal variance for all treatments.

Number of Observations per Patient for PD Parameter: 1.

Purpose:

- To compare the equality of treatment means.
- To compare the equal block effect.
- To determine the interaction effect.

Statistical Model: $y_{ijm} = \mu + \tau_i + \eta_j + \tau\eta_{ij} + \varepsilon_{ijm}$, for $i = 1, 2, \dots, k$; $j = 1, 2, \dots, s$; $m = 1, 2, \dots, n_{ij}$ where

- Y_{ijm} = PD value for patient m in treatment group i and block j
- μ = unknown overall population mean
- τ_i = unknown treatment effect for treatment group i
- η_j = unknown j th block effect
- $\tau\eta_{ij}$ = interaction effect of i th treatment with j th block
- ε_{ijm} = random error associated with patient m in treatment group i and block j

Analysis of Variance Summary: See Table 3.

Table 1 Classification for Model Grouping

	PD variable		Explanatory variables			
	Continuous (C) vs. discrete (D)	No. of values ^a	Treatment variable	Other explanatory variables	Continuous (C) vs. discrete (D)	No. of values
1	C	—	2 or more	1	D	2 or more
2	C	—	2 or more	1	C	—
3	D	2	2	1	D	2 or more
4	D	2	2 or more	1 or more	C or D	2 or more
5	C	—	2 or more	1 or more	C or D	2 or more

^aWhen “C” is specified, this column is not needed.

Table 2 Statistical Methods for Each Group

Group	Statistical methods
1	Two-way analysis of variance Crossover design One-way analysis of variance with repeated measures
2	Linear regression One-way analysis of covariance
3	Cochran-Mantel-Haenszel test
4	Logistic regression
5	Cox proportional hazards analysis

Computation for Sum of Squares:

$$\begin{aligned}
 SS(T) &= \sum_{i=1}^k \sum_{j=1}^s \sum_{k=1}^{n_{ij}} (\bar{y}_{i..} - \bar{y}_{...})^2 \\
 SS(B) &= \sum_{i=1}^k \sum_{j=1}^s \sum_{k=1}^{n_{ij}} (\bar{y}_{.j.} - \bar{y}_{...})^2 \\
 SS(TB) &= \sum_{i=1}^k \sum_{j=1}^s \sum_{k=1}^{n_{ij}} (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 \\
 SS(E) &= \sum_{i=1}^k \sum_{j=1}^s \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij.})^2 \\
 SS(TOT) &= \sum_{i=1}^k \sum_{j=1}^s \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{...})^2 \\
 \bar{y}_{i..} &= \sum_{j=1}^s \sum_{k=1}^{n_{ij}} y_{ijk} / n_{i.}, \quad n_{i.} = \sum_{j=1}^s n_{ij} \\
 \bar{y}_{.j.} &= \sum_{i=1}^k \sum_{k=1}^{n_{ij}} y_{ijk} / n_{.j}, \quad n_{.j} = \sum_{i=1}^k n_{ij} \\
 \bar{y}_{ij.} &= \sum_{k=1}^{n_{ij}} y_{ijk} / n_{ij} \\
 \bar{y}_{...} &= \sum_{i=1}^k \sum_{j=1}^s \sum_{k=1}^{n_{ij}} y_{ijk} / n, \quad n = \sum_{i=1}^k \sum_{j=1}^s n_{ij}
 \end{aligned}$$

Table 3 Analysis of Variance Summary for Parallel Group Design with a Blocking Factor

Source of variation	Degrees of freedom	Sum of squares	Mean squares
Treatment	$k - 1$	SS(T)	MS(T)
Block	$s - 1$	SS(B)	MS(B)
Treatment $\times$ block	$(k - 1)(s - 1)$	SS(TB)	MS(TB)
Error	$N - sk$	SS(E)	MS(E)
Corrected total	$N - 1$	SS(TOT)	—

Hypothesis Testing for Treatment Effect:

- $H_0: \tau_i = 0$ for all i vs. $H_a: \tau_i \neq 0$ for some i if treatment effect fixed.
- $H_0: \sigma_\tau^2 = 0$ vs. $H_a: \sigma_\tau^2 \neq 0$ if treatment effect random.

Test Statistic for Treatment Effect:

$$F_0 = \frac{MS(T)}{MS(E)} \quad \text{for fixed-treatment effect}$$

$$F_0 = \frac{MS(T)}{MS(TB)} \quad \text{for random-treatment effect}$$

Rejection Region at α Level:

- $F_0 > F_{(s-1, N-k), 1-\alpha}$ for fixed-treatment effect.
- $F_0 > F_{(s-1, (k-1)(s-1)), 1-\alpha}$ for mixed- or random-treatment effect.

Two-Sided p Value:

- $1 - \text{PROBF}(F_0, k - 1, N - sk)$ for fixed-treatment effect.
- $1 - \text{PROBF}(F_0, k - 1, (k - 1)(s - 1))$ for random-treatment effect.

Hypothesis Testing for Block Effect:

- $H_0: \eta_i = 0$ for all i vs. $H_a: \eta_i \neq 0$ for some i if treatment effect fixed.

- b. $H_0: \sigma_\eta^2 = 0$ vs. $H_a: \sigma_\eta^2 \neq 0$ if treatment effect random.

Test Statistic for Block Effect:

$$F_0 = \frac{MS(B)}{MS(E)} \quad \text{for fixed-block effect}$$

$$F_0 = \frac{MS(B)}{MS(TB)} \quad \text{for random-block effect}$$

Rejection Region at α Level:

- $F_0 > F_{(s-1, N-k), 1-\alpha}$ for fixed-block effect.
- $F_0 > F_{(s-1, (k-1)(s-1)), 1-\alpha}$ for random-block effect.

Two-Sided p Value:

- $1 - \text{PROBF}(F_0, s-1, N-sk)$ for fixed-block effect.
- $1 - \text{PROBF}(F_0, s-1, (k-1)(s-1))$ for random-block effect

Hypothesis Testing for Treatment-by-Block Interaction Effect:

- a. $H_0: \tau\eta_{ij} = 0$ for all i, j vs. $H_a: \tau\eta_{ij} \neq 0$ for some i, j for fixed model.
- b. $H_0: \sigma_{\tau\eta}^2 = 0$ vs. $H_a: \sigma_{\tau\eta}^2 \neq 0$ for mixed or random model.

Test Statistic for Interaction Effect: $F_0 = MS(TB)/MS(E)$ for any model.

Rejection Region: $F_0 > F_{((k-1)(s-1), N-sk), 1-\alpha}$.

Two-Sided p Value: $1 - \text{PROBF}(F_0, (k-1)(s-1), N-sk)$.

Related SAS Procedure: PROC GLM (or MIXED or ANOVA).

Description of and Notes on the Method:

- a. This model requires one other explanatory variable (a discrete variable) that serves as the blocking factor in the model.
- b. The model may apply to the situation in which two drugs are the treatment and blocking factors, and the combination of the two drugs is the interaction effect. In this case, the treatment randomization schedule is very likely to be generated with a full consideration of each factor and the combination, and the sample size for the study is estimated appropriately to show a sufficiently high power (≥ 0.80). For those situations where the blocking factor is a stratified variable, such as gender, race, age group, or underlying disease condition, the treatment randomization schedule in many cases does not take into account the stratification, so the near-equal number of sample size for each stratified variable may not occur. In addition, the sample size calculation, again in most cases, is based on the treatment difference only. Hence, the result from

the study may not provide strong power for the statistical significance for the blocking effect as well as for the interaction effect.

- c. The treatment randomization schedule is normally based on the number of k groups if the blocking factor is a stratified factor, and it will be based on the number of ks groups for a drug combination study.
- d. If the blocking factor represents the center effect for a multi-center study, the sample size for each center may vary to a great degree. In some cases, several small centers may be collapsed into a single “dummy” center in order to reduce the degrees of freedom for center effect and to increase the sample size for each center. In reality, the treatment effect will be difficult to conclude in a trial where the treatment by center interaction is statistically significant for a trial with a large number of centers and the treatment effect goes in different directions among centers (e.g., in a trial with 50 centers and two treatment groups, 50% of the 50 centers showed that treatment A is superior to treatment B, and the remaining 50% centers showed that treatment A is inferior to treatment B).
- e. The model is “fixed” if both treatment and block effects are fixed; the model is “random” when both treatment and block effects are random. It is a “mixed” model when one of the two effects (treatment or block) is random and the other is fixed.
- f. The model is “balanced” when all the cells have the same number of observations; otherwise, it is “unbalanced.”
- g. There are four ways of computing sum of squares in SAS for unbalanced data, known as Type I through Type IV. In general, Type III sum of squares is selected.^[1,2]
- h. The interaction term can be ignored if every cell contains only one observation or the interaction effect is insignificant. In this case, the model is reduced to a model with a randomized block design.
- i. If the treatment effect represents different dose levels (including the placebo) of a study drug, it is desirable to investigate the trend effect, such as linear or quadratic. This can also apply to the blocking factor if it is measured on an ordinal scale. In such a case, the interaction term can be decomposed into the combination of those trends, such as $T_{\text{linear}} \times B_{\text{linear}}$ or $T_{\text{linear}} \times B_{\text{quadratic}}$.
- j. PROC ANOVA in SAS can be applied to a balanced case only. PROC GLM is developed mainly for the use of the fixed-effect model. PROC MIXED is a new SAS procedure designed to handle the more complicated, so-called mixed model.
- k. Confidence intervals for each level of treatment factor (and/or blocking factor) as well as for the mean

difference between two levels for the same factor can be constructed via a T distribution, but note that the standard error used in the confidence interval depends on the type of factor (fixed or random).

- l. Multiple comparisons may be desirable to determine pairwise differences between various levels of the treatment effect if the treatment effect is deemed significant.
- m. There are a number of other methodologies proposed for combination drug trials, such as response surface method.^[3-5]
- n. This two-way analysis of variance can be extended to any higher order (≥ 3) of analysis of variance (i.e., a study with two or more “other explanatory variables”). With a higher-order analysis of variance, many different types of interaction exist.

Crossover Design (Using A 2×2 Crossover Design for Illustration)

Study Design: Crossover design.

Model Assumptions: Normal distribution, dependence, equal variance for all treatments, compound symmetry.

Number of Observations per Patient for PD Parameter: 2.

Purpose: To compare the equality of treatment means.

Statistical Model: $y_{ijm} = \mu + Q_i + S(Q)_{m(i)} + P_j + \tau_{(ij)} + \varepsilon_{ijm}$, for $i, j = 1, 2; m = 1, 2, \dots, n_i$ where

- y_{ijm} = PD value for patient m in period j in treatment sequence i
- μ = overall population mean
- Q_i = effect of i th sequence
- $S(Q)_{m(i)}$ = error term associated with m th subject in sequence i
- $\tau_{(i,j)}$ = treatment effect in period j for treatment sequence i
- P_j = effect for j th period
- ε_{ijm} = random error associated with patient m in treatment sequence i and period j

Analysis of Variance Summary: See [Table 4](#).

Computation for Sum of Squares:

Table 4 Analysis of Variance Summary for 2×2 Crossover Design

Source of variation	Degrees of freedom	Mean squares
Sequence	1	MS(Q)
Subject (sequence)	$n_1 + n_2 - 2$	MS(S(Q))
Treatment	1	MS(T)
Period	1	MS(P)
Error	$n_1 + n_2 - 2$	MS(E)
Corrected total	$2(n_1 + n_2) - 1$	MS(TOT)

$$MS(Q) = \sum_{i=1}^2 \sum_{j=1}^2 \sum_{m=1}^{n_i} (\bar{y}_{i..} - \bar{y}_{i...})^2$$

$$MS(S(Q)) = \frac{\sum_{i=1}^2 \sum_{j=1}^2 \sum_{m=1}^{n_i} (\bar{y}_{i.m} - \bar{y}_{i..})^2}{n_1 + n_2 - 2}$$

$$MS(T) = \sum_{i=1}^2 \sum_{j=1}^2 \sum_{m=1}^{n_i} (\bar{y}_{ij.} - \bar{y}_{i..})^2$$

$$MS(P) = \sum_{i=1}^2 \sum_{j=1}^2 \sum_{m=1}^{n_i} (\bar{y}_{.j.} - \bar{y}_{i...})^2$$

$$SS(TOT) = \sum_{i=1}^2 \sum_{j=1}^2 \sum_{m=1}^{n_i} (\bar{y}_{ijk} - \bar{y}_{i...})^2$$

Hypothesis Testing for Treatment Effect: $H_0: \tau_1 = \tau_2$ vs.

$$H_a: \tau_1 \neq \tau_2$$

Test Statistic for Treatment Effect: $F_0 = MS(T)/MS(E)$.

Rejection Region at α Level: $F_0 > F_{(1, n_1 + n_2 - 2, 1 - \alpha)}$.

Two-Sided p Value: $1 - \text{PROBF}(F_0, 1, n_1 + n_2 - 2, 1 - \alpha)$.

Hypothesis Testing for Period Effect: $H_0: P_1 = P_2$ vs.

$$H_a: P_1 \neq P_2$$

Test Statistic: $F_0 = MS(P)/MS(E)$.

Rejection Region at α Level: $F_0 > F_{(1, n_1 + n_2 - 2, 1 - \alpha)}$.

Two-Sided p Value: $1 - \text{PROBF}(F_0, 1, n_1 + n_2 - 2)$.

Related SAS Procedure: PROC GLM (or MIXED).

Description of and Notes on the Method:

- a. We use a 2×2 crossover design as an example for illustration.
- b. One discrete “other explanatory variable” is required. The “period” factor is the “other explanatory variable” in this model.
- c. This is used to compare two treatments with the removal of the time effect because each treatment is administered in both dosing periods to all patients. Each patient serves as his or her control.
- d. There are two major sources of variation: inter-subject variation and intra-subject variation. In this model, the inter-subject variation is composed of two components: sequence and patient within sequence; the intra-subject variation contains three components: treatment, period, and error.
- e. There are two types of variability associated with crossover design: between-patient variability and within-patient variability. MS(S(Q)) in the analysis of variance refers to the between-patient variance, and (MS(E)) refers to the within-patient variance.
- f. This model is commonly used in a Phase I PK bio-equivalence study.
- g. It is important that the study allow a sufficiently long wash-out period between the two treatment administrations to avoid the carryover effect.
- h. This model assumes no carryover effect.

- It is important that a baseline value be obtained in each period.
- The model can be extended to a higher-order number of treatments and time periods. Note that using a higher-order crossover study may potentially increase the dropout rate because the duration of study is prolonged, and when an odd number of treatments is used for a clinical study, it would be necessary to employ William's design in order to maintain the equal number of appearances for each ordered pair of treatments.
- The crossover design is not suitable for study drugs with long terminal half-lives or for study drugs with irreversible PD responses.
- In theory, repeated measures is defined as a series of the same information (data) taken under the same experimental condition from the same patient over time. Therefore, strictly speaking, this crossover design should not be classified as a case of a repeated-measures design.
- "Compound symmetry" means that the correlations between each pair of two time points are the same. Mauchly's criterion can be used for this purpose, which is also available in SAS GLM REPEATED PROCEDURE.^[6]
- When there are repeated measures for each period, the analysis proposed by Wallenstein and Fisher can be used.^[7]
- This analysis assumes a fixed-effect model.
- A confidence interval for treatment mean difference can be obtained.
- In general, patients with missing values are removed from the statistical analysis.

One-Way Analysis of Variance with Repeated Measures

Study Design: Parallel group design with a time factor.

Model Assumptions: Normal distribution, independence, equal variance, compound symmetry.

Number of Observations per Patient for PD Parameter: Multiple.

Purpose:

- To determine each treatment profile over a period of time.
- To compare the treatment effects.

Statistical Model: $y_{ijm} = \mu + \tau_i + S_{j(i)} + T_m + \tau T_{im} + \varepsilon_{ijm}$ for $i = 1, 2, \dots, k; j = 1, 2, \dots, n_i; m = 1, 2, \dots, p$ where

- y_{ijm} = PD value for patient j at time m in treatment group i
- μ = overall population mean
- τ_i = treatment effect for treatment group i
- $S_{j(i)}$ = effect for j th patient in treatment group i
- T_m = effect at m th time point

- τT_{im} = interaction effect of i th treatment with m th time point
- ε_{ijm} = random error associated with patient j in treatment group i at time j

Analysis of Variance Summary: See Table 5.

Computation for Sum of Squares:

$$\begin{aligned} SS(T) &= \sum_{i=1}^k \sum_{j=1}^{n_i} \sum_{m=1}^p (\bar{y}_{i..} - \bar{y}_{...})^2 \\ SS(BET) &= \sum_{i=1}^k \sum_{j=1}^{n_i} \sum_{m=1}^p (\bar{y}_{ij.} - \bar{y}_{i..})^2 \\ SS(TIME) &= \sum_{i=1}^k \sum_{j=1}^{n_i} \sum_{m=1}^p (\bar{y}_{.m} - \bar{y}_{...})^2 \\ SS(T \times TIME) &= \sum_{i=1}^k \sum_{j=1}^{n_i} \sum_{m=1}^p (\bar{y}_{i.m} - \bar{y}_{i..} - \bar{y}_{.m} + \bar{y}_{...})^2 \\ SS(TOT) &= \sum_{i=1}^k \sum_{j=1}^{n_i} \sum_{m=1}^p (\bar{y}_{ijm} - \bar{y}_{...})^2 \\ SS(W) &= SS(TOT) - SS(T) - SS(BET) \\ &\quad - SS(TIME) - SS(T \times TIME) \end{aligned}$$

where

$$\begin{aligned} \bar{y}_{i..} &= \sum_{j=1}^{n_i} \sum_{m=1}^p y_{ijm} / pn_i \\ \bar{y}_{.m} &= \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ijm} / N \\ \bar{y}_{ij.} &= \sum_{m=1}^p y_{ijm} / p \\ \bar{y}_{i.m} &= \sum_{j=1}^{n_i} y_{ijm} / n_i \\ \bar{y}_{...} &= \sum_{i=1}^k \sum_{j=1}^{n_i} \sum_{m=1}^p y_{ijm} / pN \text{ and } N = \sum_{i=1}^k n_i \end{aligned}$$

Table 5 Analysis of Variance Summary for Parallel Group Design with a Time Factor

Source of variation	Degrees of freedom	Sum of squares	Mean squares
Treatment	$k - 1$	SS(T)	MS(T)
Patient (treatment)	$N - k$	SS(BET)	MS(BET)
Time	$p - 1$	SS(TIME)	MS(TIME)
Treatment $\times$ Time	$(k - 1)(p - 1)$	SS(T $\times$ Time)	MS(T $\times$ Time)
Error	$(N - k)(p - 1)$	SS(W)	MS(W)
Corrected total	$Np - 1$	SS(TOT)	

Hypothesis Testing for Treatment Effect: $H_0: \tau_i = 0$ for all i vs. $H_0: \tau_i \neq 0$ for some i

Test Statistic for Treatment Effect: $F_0 = \text{MS}(T)/\text{MS}(\text{BET})$.

Rejection Region for Treatment Effect: $F_0 > F_{(k-1, N-k), 1-\alpha}$.

Two-Sided p Value for Treatment Effect: $1 - \text{PROBF}(F_0, k-1, N-k)$.

Hypothesis Testing for Time Effect: $H_0: T_m = 0$ for all m vs. $H_0: T_m \neq 0$ for some m

Test Statistic for Time Effect: $F_0 = \text{MS}(\text{TIME})/\text{MS}(W)$.

Rejection Region for Time Effect: $F_0 > F_{(p-1, (N-k)(p-1)), 1-\alpha}$.

Two-Sided p Value for Time Effect: $1 - \text{PROBF}(F_0, p-1, (N-k)(p-1))$.

Hypothesis Testing for Treatment-by-Time Interaction Effect:

$H_0: \tau T_{im} = 0$ for all i, m vs. $H_0: \tau T_{im} \neq 0$ for some i, m

Test Statistic for Interaction Effect: $F_0 = \text{MS}(T \times \text{TIME})/\text{MS}(W)$.

Rejection Region for Interaction Effect: $F_0 > F_{((k-1)(p-1), (N-k)(p-1)), 1-\alpha}$.

Related SAS Procedure: PROC GLM or PROC MIXED.

Description of and Notes on the Method:

- One discrete “other explanatory variable” is required. The “time” factor is the “other explanatory variable” in this model.
- Repeated measures can be collected in at least two ways:
 - They are over a certain period of time within the same visit (e.g., measurement of blood pressure over 24 hours post-treatment at one clinic visit). In this case, there is only one baseline.
 - They are collected at a number of visits, with one observation per visit. In this case, a baseline at each visit is highly desirable.
- Compound symmetry has to be examined to ensure the validity of the model.
- The preceding discussion points apply to the fixed model only; in other words, both treatment and time effects are fixed. This is generally what the model assumes for Phase III clinical trials because the dosage levels and time points are selected with clinical reasons.
- When a mixed or a random model is encountered, then the test statistic employed to test the time effect is $F_0 = \frac{\text{MS}(\text{TIME})}{\text{MS}(T \times \text{TIME})}$. As for the test of the treatment effect, a pseudo- F -test of the form $\frac{\text{MS}(T)}{\text{MS}(\text{BET}) + \text{MS}(T + \text{TIME}) - \text{MS}(W)}$ can be obtained, which approximates an F distribution with $(k-1, w)$ degrees of freedom, where (Ref. [6], pp. 509–512; Ref. [8])

$$w = \frac{[\text{MS}(\text{BET}) + \text{MS}(T \times \text{TIME}) - \text{MS}(W)]^2}{\frac{(\text{MS}(\text{BET}))^2}{N-k} + \frac{(\text{MS}(T \times \text{TIME}))^2}{(k-1)(p-1)} + \frac{(\text{MS}(W))^2}{(N-k)(p-1)}}$$
- There is another way to deal with a repeated measure, consider “individual time point analysis.” This means

the data are analyzed separately at each time point, with an adjustment of α level. [9,10]

- An overlaid response-time plot is useful to visualize the trend of the PD response for between- and within-treatments comparisons.
- One major drawback for repeated-measures design is a higher possibility of encountering missing values. Four “imputation” procedures are commonly used:
 - Remove patients with missing values. This allows a straightforward statistical analysis. However, those incomplete patients may provide important information for the treatment.
 - Estimate the missing values by minimizing the sum of squares due to error. This is appropriate statistically only when very few missing values are observed in the study.
 - Replace the missing values with the last available observation. This is known as the “last observation carried forward” procedure, as suggested by Gillings and Koch. [11]
 - Replace the missing values with the mean of all valid observations.
- Two major sources of variation, as with the model in the section “Crossover Design Using a 2×2 Crossover Design,” are observed: inter-subject variation and intra-subject variation. In this model, the inter-subject variation contains the variation due to treatment and due to the patient within treatment; the intra-subject variation contains three components: time, treatment by time, and error. The mean square due to error is the estimate of intra-subject variance, and the mean square due to patient (within treatment) is the estimate of inter-subject variance.
- The model can be generalized to a higher-order analysis of variance/covariance (i.e., one or more “other explanatory variables”) with repeated measures.

Group 2

Simple Linear Regression

Study Design: Parallel group design.

Model Assumptions: Normal distribution, independence.

Number of Observations per Patient for PD Parameter: 1.

Purpose:

- To determine the relationship between a PD variable and a prespecified explanatory variable.
- To compare the relationship between the two treatment groups.

Within-Treatment-Group Analysis

Statistical Model: $y_{ij} = \alpha_i + \beta_i x_{ij} + \varepsilon_{ij}$ for $i = 1, 2, \dots, k; j = 1, 2, \dots, n_i$ where

- y_{ij} = PD value for patient j in treatment group i
- α_i = unknown intercept for treatment group i
- β_i = unknown slope for treatment group i
- x_{ij} = value of covariate for patient j in treatment group i
- ε_{ij} = random error associated with patient j in treatment group i

Point Estimation of Unknown Parameters for a Fixed i :

$$\hat{\beta}_i = \frac{S_{xy}^i}{S_{xx}^i}, \quad \hat{\alpha}_i = \bar{y}_i - \hat{\beta}_i \bar{x}_i, \quad \hat{y}_{ij} = \hat{\alpha}_i + \hat{\beta}_i x_{ij}$$

$$r_i = \frac{S_{xy}^i}{\sqrt{S_{xx}^i S_{yy}^i}}$$

where

$$S_{xy}^i = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i)$$

$$S_{xx}^i = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \text{ and } S_{yy}^i = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

Hypothesis Testing for α_i : $H_0: \alpha_i = 0$ vs. $H_1: \alpha_i \neq 0$
Test Statistic:

$$T_0 = \frac{\hat{\alpha}_i}{\sqrt{\left(\frac{\sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2}{n_i - 2} \right) \left(\frac{1}{n_i} + \frac{\hat{x}_i^2}{S_{xx}^i} \right)}}$$

Rejection Region: $|T_0| > t_{(ni-2, 1-\alpha/2)}$
Two-Sided p Value: $2(1 - \text{PROBT}(|T_0|, n_i - 2))$
Hypothesis Testing for β_i : $H_0: \beta_i = 0$ vs. $H_1: \beta_i \neq 0$
Test Statistic:

$$T_0 = \frac{\hat{\beta}_i \sqrt{S_{xx}^i}}{\sqrt{\left(\frac{1}{n_i - 2} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2 \right)}}$$

Rejection Region: $|T_0| > t_{(ni-2, 1-\alpha/2)}$
Two-Sided p Value: $2(1 - \text{PROBT}(|T_0|, n_i - 2))$
Hypothesis Testing for ρ_i (Correlation Coefficient):
 $H_0: \rho_i = 0$ vs. $H_1: \rho_i \neq 0$

Test Statistic:

$$T_0 = \frac{r_i \sqrt{n_i - 2}}{\sqrt{1 - r_i^2}} \quad \text{for small } n_i (n_i < 20)$$

$$Z_0 = \frac{\frac{1}{2} \ln \frac{1 + r_i}{1 - r_i}}{\sqrt{\frac{1}{n_i - 3}}} \quad \text{for large } n_i (n_i \geq 20)$$

Rejection Region at α Level: $|T_0| > t_{(ni-2, 1-\alpha/2)}$ or $|Z_0| > Z_{(1-\alpha/2)}$

Two-Sided p Value:

$2(1 - \text{PROBT}(|T_0|, n_i - 2))$ or $2(1 - \text{PROBNORM}(|Z_0|))$

Analysis of Variance Approach (for Treatment Group i):

See Table 6.

Computation for Sum of Squares:

$$\text{SS}(X) = \frac{(S_{xy}^i)^2}{S_{xx}^i} \quad \text{and} \quad \text{SS}(E) = \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2$$

Hypothesis Testing for No Relationship Between PD Parameter and Explanatory Variable: $H_0: \beta_i = 0$ vs.

$H_a: \beta_i \neq 0$

Test Statistic for No Relationship Between PD Parameter and Explanatory Variable: $F_0 = \text{MS}(X)/\text{MS}(E)$

Rejection Region at α Level: $F_0 > F_{(1, ni-2), 1-\alpha}$

Two-Sided p Value: $\text{PROBF}(F_0, 1, n_i - 2)$

Related SAS Procedure: PROC REG (or GLM).

Between-Treatment-Groups Analysis

Hypothesis Testing for $\rho_i - \rho_m$: $H_0: \rho_i = \rho_m$ vs. $H_a: \rho_i \neq \rho_m$

Test Statistic:

$$Z_0 = \frac{\frac{1}{2} \ln \frac{1 + r_i}{1 - r_i} - \frac{1}{2} \ln \frac{1 + r_m}{1 - r_m}}{\sqrt{\frac{1}{n_i - 3} + \frac{1}{n_m - 3}}}$$

Rejection Region: $|Z_0| > z_{(1-\alpha/2)}$

Two-Sided p Value: $2(1 - \text{PROBNORM}(|Z_0|))$

Hypothesis Testing for $\beta_i - \beta_m$: $H_0: \beta_i = \beta_m$ vs. $H_a: \beta_i \neq \beta_m$

Table 6 Analysis of Variance Summary for Treatment Group n_i

Source of variation	Degrees of freedom	Sum of squares	Mean squares
Regression	1	SS(X)	MS(X)
Error	$n_i - 2$	SS(E)	MS(E)
Corrected total	$n_i - 1$	SS(TOT)	

Test Statistic for $\beta_i - \beta_m$:

- a. For small sample size,

$$T_0 = \frac{\hat{\beta}_i - \hat{\beta}_m}{S_{\text{pool}} \sqrt{\frac{1}{S_{xx}^i} + \frac{1}{S_{xx}^m}}}$$

where

$$S_{\text{pool}}^2 = \frac{\sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2 + \sum_{j=1}^{n_m} (y_{mj} - \hat{y}_{mj})^2}{n_i + n_m - 4}$$

- b. For large sample size ($n_i > 25$ and $n_m > 25$),

$$Z_0 = \frac{\hat{\beta}_i - \hat{\beta}_m}{\sqrt{\frac{1}{n_i - 2} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2 \frac{1}{S_{xx}^i} + \frac{1}{n_m - 2} \sum_{j=1}^{n_m} (y_{mj} - \hat{y}_{mj})^2 \frac{1}{S_{xx}^m}}}$$

Rejection Region at α Level: $|T_0| > t_{(ni+nm-4, 1-\alpha/2)}$ or $|Z_0| > Z_{(1-\alpha/2)}$

Two-Sided p Value: $2(1 - \text{PROBT}(|T_0|, n_i + n_m - 4))$ or $2(1 - \text{PROBNORM}(|Z_0|))$

Description of and Notes on the Method:

- The explanatory variable for this model is specified as continuous. In fact, regression analysis will work fine with discrete explanatory variables in an ordinal scale with a reasonable number of different values
- The regression analysis focuses on the relationship between the identified explanatory variable and the PD parameter within the treatment group. That is different from an analysis of covariance (discussed in the next section), which focuses more on the between-treatments comparison. In addition, this regression model does not assume that all treatment groups have equal slope.
- It can be seen that there are two approaches to deal with within-treatment regression analysis. But both methods for testing a zero slope are the same (one using a t -test, the other using an F -test).
- There are two measures of interest with regard to the relationship between the PD parameter and the explanatory variable: one is measured as a linear slope, the other is expressed as a correlation coefficient.
- Confidence intervals for intercept and slope can be constructed.

- f. It is often desirable to obtain a confidence interval for the mean response at a fixed value (say, at x_0) of the explanatory variable for treatment group i , which is given by:

$$\bar{y}_i + \hat{\beta}_i(x_0 - \bar{x}_i) \pm t_{(n_i-2, 1-\alpha/2)} \times \sqrt{\left(\frac{\sum_{j=1}^n (y_{ij} - \hat{y}_{ij})^2}{n_i - 2} \right) \left(\frac{1}{n_i} + \frac{(x_0 - \bar{x}_i)^2}{S_{xx}^i} \right)}$$

- A regression plot with the predicted line is always made for this analysis. In addition, it is usually of great interest to construct confidence bands for the regression line.
- ρ_i = unknown population correlation coefficient for treatment group i and r_i = observed correlation coefficient for treatment group i .
- When only one explanatory variable is employed, we have a simple linear regression, which is discussed in this section.
- If more than one observation occurs for most of the values for the explanatory variable, the error term in the ANOVA can be partitioned into two components: pure error (PE) and lack of fit (LOF). $SS(\text{PE})$ can be calculated by summing up the sum of squares for all PDs at the same x value, and $d.f.(\text{PE}) = u$ (where u = number of different values of x). Thus, $SS(\text{LOF}) = SS(\text{Error}) - SS(\text{PE})$, and $d.f.(\text{LOF}) = d.f.(\text{Error}) - d.f.(\text{PE})$. In such a case, a test of the adequacy of the model can be made by means of the test statistic of the form $F = MS(\text{LOF})/MS(\text{PE})$. Note that the denominator of the test statistic for X remains unchanged [i.e., $MS(\text{Error})$].
- When more than one explanatory variable is employed, we have a multiple linear regression. The degrees of freedom (d.f.) for the source of "regression" in the ANOVA table are equal to the number of explanatory variables.
Associated with the multiple regression model are multiple correlation, partial correlation, overall F -test, and partial F -test. The details can be found in any statistics book that discusses the topic of multiple regression (e.g., Ref. [12]).
- In a simple regression case, a second-order or higher term may also be needed for the model, which is known as polynomial regression.^[12]
- A non-parametric method using ranks may be applied to regression analysis.^[13]
- Linearity is the simplest situation between the two variables; it is much more complicated when the relationship is nonlinear.^[14,15]

One-Way Analysis of Covariance

Study Design: Parallel group design.

Model Assumptions: Normal distribution, independence, equal variance, equal slope for all treatment groups.

Number of Observations for PD Parameter per Patient: 1.

Purpose: To compare the mean adjusted for covariate(s) of all treatment groups are equal.

Statistical Model: $y_{ij} = \alpha_i + \beta x_{ij} + \varepsilon_{ij} = \mu + \tau_i + \beta x_{ij}$ for $i = 1, 2, \dots, k; j = 1, 2, \dots, n_i$ where

- y_{ij} = PD value for patient j in treatment group i
- α_i = unknown intercept for treatment group i
- β = unknown common slope
- μ = overall population mean
- τ_i = treatment effect for treatment group i
- x_{ij} = value of the covariate for patient j in treatment group i
- ε_{ij} = random error associated with patient j in treatment group i

Analysis of Variance Summary: See Table 7.

Computation for Sum of Squares:

$$SS(\text{TOT}) = s_{yy}$$

$$SS(\text{E}) = \frac{\left(\sum_i S_{yy}^i \right) \left(\sum_i S_{xx}^i \right) - \left(\sum_i S_{xy}^i \right)^2}{\left(\sum_i S_{xx}^i \right)}$$

where

$$SS(\text{T}) = \frac{S_{yy}S_{xx} - S_{xy}^2}{S_{xx}} - SS(\text{E})$$

$$SS(\text{X}) = \left(\sum_i S_{yy}^i \right) - SS(\text{E})$$

Table 7 Analysis of Variance Summary for Parallel Group Design

Source of variation	Degrees of freedom	Sum of squares	Mean squares
Treatment	$k - 1$	$SS(\text{T})$	$MS(\text{T})$
Covariate	1	$SS(\text{X})$	$MS(\text{X})$
Error	$N - k - 1$	$SS(\text{E})$	$MS(\text{E})$
Corrected total	$N - 1$	$SS(\text{TOT})$	

$$S_{yy} = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$

$$S_{xx} = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2$$

$$S_{xy} = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})(y_{ij} - \bar{y}_{..})$$

$$S_{yy}^i = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

$$S_{xx}^i = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2$$

$$S_{xy}^i = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})(y_{ij} - \bar{y}_{i.})$$

$$\bar{y}_{i.} = \sum_{j=1}^{n_i} y_{ij} / n_i$$

$$\bar{y}_{..} = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} / N$$

$$\bar{x}_{i.} = \sum_{j=1}^{n_i} x_{ij} / n_i, \quad \bar{x}_{..} = \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij} / N$$

$$N = \sum_{i=1}^k n_i$$

Point Estimation for the Unknown Parameters for a Fixed i :

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\alpha}_i = \bar{y}_{i.} - \hat{\beta} \bar{x}_{i.}, \quad \hat{y}_{ij} = \hat{\alpha}_i + \hat{\beta} x_{ij}$$

Hypothesis Testing for Treatment Effect:

- $H_0: \tau_i = 0$ for all i , vs. $H_0: \tau_i \neq 0$ for some i if treatment effect is fixed.
- $H_0: \sigma_\tau^2 = 0$ vs. $H_a: \sigma_\tau^2 \neq 0$ if treatment effect is random.

Test Statistic for Treatment Effect: $F = MS(\text{T}) / MS(\text{E})$.

Rejection Region at α Level: $F > F_{(k-1, N-k-1), 1-\alpha}$.

Two-Sided p Value: $1 - \text{PROBF}_{(F_0, k-1, N-k-1)}$.

100(1 - α)% Confidence Interval for the Adjusted Mean Response for Treatment Group i :

$$\hat{\alpha}_i + \hat{\beta} \bar{x}_{i.} \pm t_{(N-k-1, 1-\alpha/2)} \sqrt{MS(\text{E}) \left(\frac{1}{n_i} + \frac{(\bar{x}_{i.} - \bar{x}_{..})^2}{S_{xx}} \right)}$$

100(1 - α)% Confidence Interval for the Difference of Adjusted Mean Response Between Treatment Group i and Treatment Group j :

$$\hat{\alpha}_i - \hat{\alpha}_j \pm t_{(N-k-1, 1-\alpha/2)} \sqrt{MS(E) \left(\frac{1}{n_i} + \frac{1}{n_j} \frac{(\bar{x}_i - \bar{x}_j)^2}{S_{xx}} \right)}$$

Hypothesis Testing for Slope: $H_0: \beta = 0$ vs. $H_0: \beta \neq 0$

Test Statistic for Slope: $F = MS(X)/MS(E)$.

Rejection Region at α Level: $F > F_{(1, N-k-1), 1-\alpha}$.

Two-Sided p Value: $1 - \text{PROBF}(F_0, 1, N-k-1)$.

Related SAS Procedure: PROC GLM (or MIXED).

Detailed Descriptions:

- When a covariate (an uncontrollable variable that is highly correlated with the PD parameter but not directly affected by the treatment) is taken into consideration in a clinical trial, analysis of covariance is a way to adjust for the covariate.
- This model can be extended to any number of covariates.
- Because a requirement for this model is the common slope among all treatment groups, a preliminary test on the common slope is recommended that can be done with the following model in SAS: $y_{ij} = \mu + \eta + \beta\eta + \varepsilon_{ij}$. The analysis of covariance is not appropriate if the interaction term of $\beta\tau_i$ is significant.
- Another way of handling a covariate in a situation where the covariate is the baseline is to adjust this value from the postdose PD value first and then to apply analysis of variance or other appropriate methods.
- The model is reduced to a linear regression if the treatment effect is insignificant. On the other hand, the model is reduced to a one-way analysis of variance if the slope associated with the covariate is insignificant.

Group 3

Cochran-Mantel-Haenszel Test

Study Design: Parallel group design.

Model Assumptions: Binomial distribution, independence.

Number of Observations per Patient for PD Parameter: 1.

Purpose: To compare the proportions of the two treatment groups based on stratified samples.

Hypothesis Testing: $H_0: \pi_1 = \pi_2$ vs. $H_a: \pi_1 \neq \pi_2$

Test Statistic:

$$M_0 = \frac{\left(\sum_i \left(n_{i11} - \frac{n_{i1+}n_{i1+}}{n_{i++}} \right) - 0.5 \right)^2}{\sum_i \left(\frac{n_{i1+}n_{i1+}n_{i2+}n_{i2+}}{n_{i++}^2(n_{i++} - 1)} \right)}$$

Table 8 Observed Frequencies for Group 3

Stratum I	Treatment 1	Treatment 2	Row total
Responder	n_{111}	n_{112}	n_{11+}
Non-responder	n_{121}	n_{122}	n_{12+}
Column total	n_{1+1}	n_{1+2}	n_{1++}

Rejection Region at α Level: $M_0 > \chi_{1, 1-\alpha}^2$

Two-Sided p Value: $1 - \text{PROBCHI}(M_0, 1)$.

Observed Frequency Table: See Table 8

Related SAS Procedure: PROC FREQ.

Description of and Notes on the Method:

- The stratum may represent study centers, gender, or patients' underlying condition (e.g., diabetics vs. non-diabetics), etc.
- The quantity "0.5" in the test statistics represents a correction for continuity, which may not be used.
- An interaction may exist if the difference of the two proportions varies from one stratum to another. Under this circumstance, each stratum shall be analyzed separately.
- The number of strata is best kept to a minimum, for at least two reasons:
 - To avoid small or empty cells.
 - Because the interaction term is difficult to interpret if significant.
- This test can be generalized to cases with k treatments ($k > 2$), r responses ($r > 2$), and s strata ($s > 2$).^[16] However, from a practical point of view, all k , r , and s shall be kept as small as possible for the same two reasons described in point (d).

Group 4

Logistic Regression

Study Design: Parallel group design.

Model Assumptions: Binomial distribution, independence.

Number of Observations per Patient for PD Parameter: 1.

Purpose: To determine the correlation between the proportion of the PD parameter and explanatory variable(s).

Statistical Model: $E \left(\log \left(\frac{\pi(x)}{1 - \pi(x)} \right) \right) = \alpha_0 + \sum_i \beta_i x_i$

Hypothesis Testing for One Correlation β_i : $H_0: \beta_i = 0$ vs. $H_1: \beta_i \neq 0$

Test Statistic for β_i : $\chi_w^2 = \left(\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \right)^2$

Rejection Region for β_i at α level: $\chi_w^2 > \chi_{1, 1-\alpha/2}^2$.

Hypothesis Testing for a subset of model parameters β_q :

$H_0: \beta_q = 0$ vs. $H_a: \beta_q \neq 0$

Test Statistic for β_q : $\chi_w^2 = \mathbf{b}_q' (\text{COV}(\mathbf{b}_q))^{-1} \mathbf{b}_q$

Rejection Region for β_q at α level: $\chi_w^2 > \chi_q, 1 - \alpha/2^2$.

Related SAS Procedure: PROC LOGISTIC.

Description of and Notes on the Method:

- In this method, there are two responses for the PD parameter, with 0 indicating the non-occurrence of the event of interest, and 1 indicating non-non-occurrence.
- The notations in the model are defined as follows:
 - x_1 = treatment effect (if there are two treatments, such as active vs. control, we define $x_1=1$ for active drug and $x_1=0$ for placebo)
 - $x_2, \dots, x_k = k - 1$ other explanatory variables ($k \geq 1$), which can be continuous or discrete
 - $\pi(\mathbf{x})$ = probability of having event associated with $\mathbf{x} = (x_1, x_2, \dots, x_k)$
 - $E(Y)$ = expected value of Y
 - α_0 = intercept
 - β_i = unknown true correlation (slope) between $\pi(\mathbf{X})$ and x_i
 - $\beta_q = (\beta_1, \beta_2, \dots, \beta_q)'$ is a subset of $\beta = (\beta_1, \beta_2, \dots, \beta_k)'$, ($q \leq k$)
 - $\mathbf{b}_q = (b_1, b_2, \dots, b_q)'$ is an estimate of $\beta_q = (\beta_1, \beta_2, \dots, \beta_q)'$
 - $\text{COV}(\mathbf{b}_q)$ = variance-covariance matrix for $\mathbf{b}_q$,
 - $\text{SE}(\hat{\beta}_i)$ = standard error of $\hat{\beta}_i$
- The test statistic is called the Wald statistic, which is asymptotically chi square distributed with 1 degree of freedom for a sufficiently large sample size. The first test statistic is a special case of the second test statistic when $q=1$.
- There is another common test for logistic regression, which is well known as the maximum log-likelihood test, which is asymptotically chi square distributed with q d.f.
- We should first test β_i for any $i \geq 2$ to determine which covariates are correlated with the proportion of the PD response. Then we test β_1 for the proportion between the two treatment groups.

f. Note that
$$\pi(x) = \frac{1}{1 + \exp\left(-\left[\alpha_0 + \sum_i \beta_i x_i\right]\right)}$$

Hence, in the case of two treatments we have

$$x(x_1=1) = \frac{1}{1 + \exp\left(-\left[\alpha_0 + \sum_{i \geq 2} \beta_i x_i\right]\right)} \quad \text{and}$$

$$x(x_1=0) = \frac{1}{1 + \exp\left(-\left[\alpha_0 + \beta_1 \sum_{i \geq 2} \beta_i x_i\right]\right)}$$

- Again, in the case of two treatments, the odds for the control group is given by

$$\begin{aligned} \text{odds}(\text{control}) &= \frac{\pi(x_1=0)}{1 - \pi(x_1=0)} \\ &= \exp\left(\alpha_0 + \sum_{i \geq 2} \beta_i x_i\right) \\ \text{odds}(\text{active}) &= \frac{\pi(x_1=1)}{1 - \pi(x_1=1)} \\ &= \exp\left(\alpha_0 + \beta_1 + \sum_{i \geq 2} \beta_i x_i\right) \end{aligned}$$

Hence, the odds ratio of the active group relative to the control group is odds (active)/odds (control) = $\exp(\beta_1)$. This implies that β_1 is a measure of the odds ratio. The $1 - \alpha\%$ confidence interval for $\exp(\beta_1)$ can be obtained in a conventional way as follows: $\exp(\hat{\beta}_1 \pm \alpha/z\text{SE}(\hat{\beta}_1))$.

- Based on the discussion in point (g), the odds ratio can be generalized for any covariate i , which is equal to $\exp(\beta_i)$.
- Logistic regression is analogous to analysis of covariance. Analysis of covariance is for the comparison of mean responses adjusted for covariate(s), while logistic regression is for the comparison of proportion responses adjusted for covariate(s).

Group 5

Cox Proportional Hazards Analysis

Study Design: Parallel group design.

Model Assumptions: Independence.

Number of Observations per Patient for PD Parameter: 1.

Purpose: To determine and compare the survival time between the PD parameter and explanatory variable(s).

Statistical Model: $h(t) = \lambda(t) \exp(\sum_i \beta_i x_i)$.

Hypothesis Testing for one correlation β_i : $H_0: \beta_i = 0$ vs. $H_a: \beta_i \neq 0$

Test Statistic for β_i : $\chi_w^2 = \left(\frac{\hat{\beta}_i}{\text{SE}(\hat{\beta}_i)} \right)^2$

Rejection Region for β_i at α level: $\chi_w^2 > \chi_{1, 1 - \alpha/2^2}$.

Hypothesis Testing for a Subset of Model Parameters β_q : $H_0: q = 0$ vs. $H_a: q \neq 0$

Test Statistic for β_q : $X_w^2 = \mathbf{b}_q' (\text{COV}(\mathbf{b}_q))^{-1} \mathbf{b}_q$.

Rejection Region for β_q at α level: $\chi_w^2 > \chi_q, 1 - \alpha/2^2$.

Related SAS Procedure: PROC PHREG.

Description of and Notes on the Method:

- In this method, there are two responses for the PD parameter, with 0 indicating the non-occurrence of the event of interest and 1 indicating non-occurrence.

- b. The notations in the model are defined as follows:
- x_1 = treatment effect (in the case of two treatments, we may define $x_1 = 1$ for active drug and $x_1 = 0$ for placebo)
 - $x_2, \dots, x_k = k - 1$ other explanatory variables ($k \geq 1$), which can be continuous or discrete
 - $h(t)$ = hazard function at time t associated with $\mathbf{x} = (x_1, x_2, \dots, x_k)$
 - $\lambda(t)$ = unspecified initial hazard function at time t
 - β_i = correlation (slope) between $\pi(\mathbf{x})$ and x_i
 - $\boldsymbol{\beta}_q = (\beta_1, \beta_2, \dots, \beta_q)'$ is a subset of $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)'$
 - $\mathbf{b}_q = (b_1, b_2, \dots, b_q)'$ is an estimate of $\boldsymbol{\beta}_q = (\beta_1, \beta_2, \dots, \beta_q)'$
 - $\text{COV}(\mathbf{b}_q)$ = variance-covariance matrix for $\mathbf{b}_q$
 - $\text{SE}(\boldsymbol{\beta}_i)$ = standard error of estimate of $\boldsymbol{\beta}_i$
- c. The test statistic, called the Wald statistic, was discussed under "Group 4, Logistic Regression."
- d. The log-likelihood test is also common for this analysis, in addition to the Wald chi-square test.
- e. We should first test β_i for any $i \geq 2$ to determine which covariates are correlated with the hazard function, then we test β_1 between the two treatment groups.
- f. Cox proportional hazards analysis is analogous to logistic regression, where logistic regression is dealing with the proportion of responses for treatment groups, with adjustment for covariate(s), and Cox proportional hazards model is dealing with survival time responses for treatment groups, with adjustment for covariate(s). On the other hand, Cox proportional hazards analysis is also similar to the log-rank test because both tests are used for analyzing survival time. The log-rank test, however, does not take into account any covariate, whereas the Cox proportional hazards analysis does.
- g. It can be shown that $\exp(\hat{\beta}_i)$ is an estimated risk ratio for covariate i and is independent of time and other covariates.

CONCLUSION

Because there are potentially many other different cases that may arise in a clinical setting, this entry only a purpose

of covering mostly known cases and the most commonly used statistical methodologies.

REFERENCES

1. Searle, S.R. *Linear Models for Unbalanced Data*; Wiley: New York, 1987.
2. Littell, R.C., Freund, R.J.; Spector, P.C. *SAS System for Linear Models*; SAS Institute: Cary, NC, 1991.
3. Stewart, W.H. Application of response surface methodology and factorial designs to clinical trials for drug combination development. *J. Biopharm. Stat.* **1996**, *6*, 219–230.
4. Hung, H.M.J. On Evaluation of Multiple-Dose Combination Drugs in Factorial Clinical Trials; Amer. Stat. Assoc. *Proceedings of the Biopharm Section; American Statistical Association*, 1996, 25–32.
5. Hung, H.M.J. Global tests for combination drug studies in factorial trials. *Stat. Med.* **1996**, *15*, 233–247.
6. *SAS/STAT User's Guide. Version 6*, 4th Ed.; SAS Institute: Cary, NC, 1990, 2.
7. Wallenstein, S.W.; Fisher, A.C. The analysis of the two-period repeated measurements crossover design with application to clinical trials. *Biometrics* **1977**, *33*, 261–269.
8. Hicks, C.R. *Fundamental Concepts in the Design of Experiments*, 3rd Ed.; CBS College: New York, 1982, 210–224.
9. Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **1988**, *75*, 800–802.
10. Wright, P. Adjusted P-values for simultaneous inference. *Biometrics* **1992**, *48*, 1005–1013.
11. Gillings, D.; Koch, G. The application of intention-to-treat to the analysis of clinical trials. *Drug Inf. J.* 1991, *25*, 411–424.
12. Kleinbaum, D.G.; Kupper, L.L. *Applied Regression Analysis and Other Multivariable Methods*; Duxbury Press: Boston, MA, 1978.
13. Birkes, D.; Dodge, Y. *Alternative Methods of Regression*; Wiley: New York, 1993, 111–142.
14. Gallant, A.R. *Nonlinear Statistical Methods*; Wiley: New York, 1989.
15. Bates, D.M.; Watts, D.G. *Nonlinear Regression Analysis and Its Application*; Wiley: New York, 1988.
16. Agresti, A. *Categorical Data Analysis*; Wiley: New York, 1990, 234–235.

Pharmacodynamics with No Covariates

Cheng-Tao Chang

Novo Nordisk Inc., Princeton, New Jersey, U.S.A.

Robert L. Wong

Abbott Laboratories, Parsippany, New Jersey, U.S.A.

Abstract

This entry discusses several univariate statistical methods that are used mostly for Phase III clinical studies for considering one independent and one dependent variable.

Keywords: Covariate; Pharmacodynamics; Univariate.

INTRODUCTION

In this entry, we will discuss several univariate statistical methods that are used mostly for Phase III clinical studies for considering one independent and one dependent variable. This entry will serve as a review for biostatisticians and as a utility tool for non-statisticians. In the following discussions, we will specify the underlying conditions (including study design, model assumptions, and number of data points per patient) that apply to each method. Note that “explanatory variable” refers to treatment (or dosage level). “Response variable” refers to the PD parameter, and we will consider the clinical trials with a single, primary PD parameter.

We will divide clinical trials into four groups according to the type of PD parameter and the number and type of explanatory variables. Table 1 displays the summary of these four groups. The statistical methods for the groups in Table 1 may apply to single-center studies or to multi-center studies with no consideration of a “center” effect. However, in general it is advised that an investigation of the center effect as well as the center-by-treatment effect be performed *before* data from all centers in a multi-center clinical study are combined and analyzed. Statistical methods for considering the center effect and/or other effects, such as demographic data and baseline evaluations, are discussed in the entry titled *Pharmacodynamics with Covariates*.

A discussion of each statistical method will cover purpose, statistical model, hypothesis testing, test statistic, rejection region, confidence interval, two-sided p -value, description of and notes on the method, and any related SAS procedure. (SAS, the most popular statistical software package, is also accepted by the Food and Drug Administration as one of the standard packages.)

The statistical methods discussed for each group are listed in Table 2.

STATISTICAL METHODS

Group 1

Two-Sample t-Test

Study Design: Parallel group design (or completely randomized design).

Model Assumptions: Normal distribution, independence, and equal variance.

Number of Observations per Patient for PD parameter: 1.

Purpose: To compare the equality of two treatment means.

Hypothesis Testing: $H_0: \mu_1 = \mu_2$ vs. $H_a: \mu_1 \neq \mu_2$

$$\text{Test Statistic: } T_0 = \frac{(\bar{Y}_1 - \bar{Y}_2)}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Rejection Region at α Level: $|T_0| > t_{(n_1 + n_2 - 2), (1 - \alpha/2)}$

100(1 - α)% Confidence Interval:

$$\bar{Y}_1 - \bar{Y}_2 \pm t_{(n_1 + n_2 - 2), (1 - \alpha/2)} \left(S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

Two-Sided p Value: $2 \times (1 - \text{PROBT}(|T_0|, n_1 + n_2 - 2))$.

Related SAS PROC Procedure: PROC TTEST.

Description of and Notes on the Method:

- Each patient is assigned to one of the two treatment groups and contributes only one observation to the entire data set.
- The data collected from the two treatment groups are independent because they are from different patients.
- μ_1, μ_2 represent the unknown population means for treatment groups 1 and 2, respectively.

Table 1 Classification for Model Grouping

Group	PD variable		Treatment variable No. of values
	Continuous (C) vs. discrete (D)	No. of values ^a	
1	C	—	2
2	C	—	2 or more
3	D	2	2
4	D	2 or more	2 or more

^aWhen “C” is specified, this column is not needed.

Table 2 Statistical Methods for Each Group

Group	Statistical methods
1	Two-sample <i>t</i> -test Paired <i>t</i> -test Wilcoxon's rank-sum test Wilcoxon's signed-rank test
2	One-way analysis of variance Kruskal-Wallis test Log-rank test
3	Fisher exact test McNeMar's test
4	Chi-square test

- d. $\bar{Y}_1, \bar{Y}_2$ are the sample means for treatment groups 1 and 2, respectively; n_1 and n_2 are the sample sizes for treatment groups 1 and 2, respectively.
- e. S is the pooled standard deviation, calculated via $\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$ where S_1^2 and S_2^2 are the variances of treatment groups 1 and 2, respectively.
- f. PROBT (x , d.f.) expressed in the two-sided p value is the cumulative probability function for the t distribution in SAS.
- g. $t_{(n_1 + n_2 - 2), (1 - \alpha/2)}$ is the numerical value at which the cumulative probability is $1 - \alpha/2$ under a t distribution with $n_1 + n_2 - 2$ degrees of freedom.
- h. The hypothesis testing can be modified from a two-sided test to a one-sided test, depending on the study objective.
- i. The p -value of a one-sided test can be obtained by halving the p -value of the two-sided test if the test statistic is positive and the alternative hypothesis is a right-tailed test ($H_a: \mu_1 > \mu_2$), or the test statistic is negative and the alternative hypothesis is a left-tailed test ($H_a: \mu_1 < \mu_2$).
- j. The rejection region for the right-tailed test ($H_a: \mu_1 > \mu_2$) is $T_0 > t_{(n_1 + n_2 - 2), (1 - \alpha)}$; the rejection region for the left-tailed test ($H_a: \mu_1 < \mu_2$) is $T_0 < t_{(n_1 + n_2 - 2), (\alpha)}$.
- k. A preliminary two-sided F -test on equal variances of the two treatment groups is highly suggested. When the null hypothesis of equal variance is rejected, the

test statistic listed earlier may not be applied. Instead, you may employ an approximate t -test statistic of the form $T_0 = \frac{(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ with the approximate degrees of freedom^[1]

$$\text{d.f.} = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\left(\frac{\left(\frac{S_1^2}{n_1} \right)^2}{(n_1 - 1)} \right) + \left(\frac{\left(\frac{S_2^2}{n_2} \right)^2}{(n_2 - 1)} \right)}$$

Paired t-Test

Study Design: Parallel group design.

Model Assumptions: Normal distribution, dependence.

Number of Observations per Patient for PD Parameter: 2.

Purpose: To compare the equality of two observations within the same treatment group.

Hypothesis Testing: $H_0: \mu_d = 0$ vs. $H_a: \mu_d \neq 0$

$$\text{Test Statistic: } T_0 = \frac{\bar{Y}_d}{\frac{S_d}{\sqrt{n}}}$$

Rejection Region at α Level: $|T_0| > t_{(n-1), (1-\alpha/2)}$

100(1 - α)% Confidence Interval: $\bar{Y}_d \pm t_{(n-1), (1-\alpha/2)} S_d / \sqrt{n}$

Two-Sided p Value: $2 \times (1 - \text{PROBT}(|T_0|, n - 1))$.

Related SAS Procedure: PROC MEANS.

Description of and Notes on the Method:

- a. This test, in general, applies to the comparison of two readings under two conditions for the same treatment from the same sample of patients. The two readings are dependent and highly correlated since they are obtained from the same patient sample. Normally, this test is more useful for within-treatment comparison than for between-treatments comparison. Two examples of such an application are baseline vs. posttreatment and two posttreatment results at two different time points.
- b. One reason why this paired t -test is not an ideal test for comparing two treatments if the two evaluations are collected under two different treatments for the same patients is that one treatment is always administered before (or after) the other, so the treatment effect may be confounded with the order/time effect. It is more appropriate to design such a study in a crossover fashion, so every patient can receive both treatments. (Please refer to the entry on *Pharmacodynamics with Covariates* for a detailed discussion of crossover design.)

- c. μ_d = population mean of paired difference
 - $\bar{Y}_d$ = sample mean of μ_d
 - n = number of pairs
- d. Patients with only one observation are removed from the analysis.
- e. A one-sided paired test can also be employed in a similar way to a two-sample t -test.

Wilcoxon's Rank-Sum Test

Study Design: Parallel group design.

Model Assumptions: Independence.

Number of Observations per Patient for PD Parameter: 1.

Purpose: To compare the equality of two treatment medians.

Hypothesis Testing: $H_0 : M_1 = M_2$ vs. $H_a : M_1 \neq M_2$

$$\text{Test Statistic: } Z_0 = \frac{\text{sign}\left(R_1 - \frac{n_1(N+1)}{2} - 0.5\right)}{\sqrt{\frac{n_1 n_2}{12} \left(N+1 - \frac{\sum_i m_i(m_i^2 - 1)}{N(N-1)}\right)}}$$

Rejection Region at α Level: $|Z_0| > Z_{1-\alpha/2}$

Two-Sided p -Value: $2 \times (1 - \text{PROBNORM}(|Z_0|))$.

Related SAS Procedure: PROC NPARIWAY.

Description of and Notes on the Method:

- a. This parallels the two-sample t -test, with the following two major differences:
 - i. Normal distribution of the PD parameter is not required.
 - ii. Comparison is based on the median, not on the mean.
- b. A preliminary test of normality can be performed to determine whether the data are normally distributed. A Shapiro-Wilk or Kolmogorov can be used, which is available in the PROC UNIVARIATE procedure of SAS.^[2]
- c. In order to employ this non-parametric test, the data from both treatment groups are first ranked from lowest to highest. The average rank is assigned to tied values if applicable.
- d. R_1 represents sum of ranks associated with treatment 1, n_1 and n_2 represent the sample size for treatments 1 and 2, respectively. $N = n_1 + n_2$.
- e. m_i represents the number of ties for the i th value that has ties.
- f. $\text{sign}(|y| - a) = \begin{cases} y - a & \text{if } y > a \\ -(|y| - a) & \text{if } y < a \end{cases}$
- g. $\text{PROBNORM}(x)$ is the cumulative probability function at x for the standard normal distribution in

SAS, and $Z_{1-\alpha/2}$ is a numerical value at which the cumulative probability is $1 - \alpha/2$ under a standard normal curve.

- h. "0.5" in the test statistic is known as "correction for continuity." It is not always necessary to include it in the test statistic. When it is not used, the numerator of the test statistic can be simplified as $R_1 - n_1(N+1)/2$ without the absolute-value sign.
- i. The Wilcoxon rank-sum test is shown to be equivalent to another nonparametric test for two independent samples, i.e., the Mann-Whitney U test, when the continuity for correction is not used^[3] (Ref. [4], p. 406).
- j. A one-sided Z -test may be performed in the same manner as for a t -test.
- k. For $n > 30$, we may feel free to use a two-sample t -test instead of this nonparametric test if the data are symmetric, based on the central limit theorem that the sample mean tends to be normally distributed for large sample sizes (> 30) regardless of the underlying distribution of the data.
- l. Another approximation test procedure to employ is a two-sample t -test with the ranked data.^[5]

Wilcoxon's Signed-Rank Test

Study Design: Parallel group design.

Model Assumptions: Dependence.

Number of Observations per Patient for PD parameter: 2.

Purpose: To compare the median of two observations within the same treatment group.

Hypothesis Testing: $H_0 : M_d = 0$ vs. $H_a : M_d \neq 0$

Test Statistic:

$$T_0 = \frac{\left(\frac{R^+ - R^-}{2}\right)\sqrt{n-1}}{\left(\frac{n^2(n+1)(2n+1)}{24} - \frac{n \sum_i (m_i^3 - m_i)}{48}\right) - \left(\frac{R^+ - R^-}{2}\right)^2}$$

for small sample size,

$$Z_0 = \frac{\text{sign}\left(\min\{R^+, R^-\} - \frac{n(n+1)}{4} - 0.5\right)}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum_i (m_i^3 - m_i)}{48}}}$$

for large sample size.

Rejection Region at α Level: $|T| > T_{(n-1), (1-\alpha/2)}$ or $|Z_0| > Z_{1-\alpha/2}$.

Two-Sided p -Value: $2 \times (1 - \text{PROBT}(|T_0|, n-1))$ or $2 \times (1 - \text{PROBNORM}(|Z_0|))$.

Related SAS Procedure: PROC UNIVARIATE.

Description of and Notes on the Method:

- This test deals with paired data when the normality assumption cannot be fulfilled.
- Wilcoxon's sign-ranked test is the nonparametric version of the paired t -test.
- M_d = population mean of paired difference. n = number of non-zero pairs.
- The differences are ranked from lowest to largest with respect to the absolute value, and zeros are ignored.
- R^+ = sum of ranks related to positive differences, and R^- = sum of ranks related to negative differences.
- $\min\{R^+, R^-\}$ = smaller value of R^+ and R^- .
- m_i represents the number of ties for the i th value that has ties.
- The value of $\min\{R^+, R^-\} - n(n+1)/4$ is always non-positive.
- The "0.5" correction for continuity in the test statistic is not always needed. The absolute-value sign should be removed in the test statistic if "0.5" is not used.
- $$\text{sign}\left(\left|\min\{R^+, R^-\} - \frac{n(n+1)}{4}\right| - 0.5\right)$$
$$= \begin{cases} \left|\min\{R^+, R^-\} - \frac{n(n+1)}{4}\right| - 0.5 & \text{if } R^+ > R^- \\ -\left(\left|\min\{R^+, R^-\} - \frac{n(n+1)}{4}\right| - 0.5\right) & \text{if } R^+ < R^- \end{cases}$$
- A one-sided test can be performed in a similar manner to that for a t -test.

Group 2**One-Way Analysis of Variance**

Study Design: Parallel design.

Model Assumptions: Normal distribution, independence, equal variance for all treatments.

Number of Observations per Patient for PD Parameter: 1

Purpose: To compare the equal means of all treatment groups.

Statistical Model: $y_{ij} = \mu + \eta + \varepsilon_{ij}$ for $i = 1, 2, \dots, k$; $j = 1, 2, \dots, n_i$ where

- y_{ij} = PD value for patient j in treatment group i
- μ = unknown overall population mean
- τ_i = unknown treatment effect for treatment group i
- ε_{ij} = random error associated with patient j in treatment group i

Analysis of Variance Summary: See Table 3.

Computation for Sum of Squares:

$$SS(T) = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y}_{..})^2, \quad SS(E) = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

Table 3 Analysis of Variance Summary for Parallel Group Design

Source of variation	Degrees of freedom (d.f.)	Sum of squares (SS)	Mean squares (MS)
Treatment	$k - 1$	SS(T)	MS(T)
Error	$N - k$	SS(E)	MS(E)
Corrected total	$N - 1$	SS(TOT)	—

$$SS(TOT) = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2,$$

$$MS(\text{Source } X) = \frac{SS(\text{Source } X)}{\text{d.f.}(\text{Source } X)}$$

$$\bar{y}_i = \sum_{j=1}^{n_i} \bar{y}_{ij} / n_i, \quad \bar{y}_{..} = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} / N, \quad N = \sum_{i=1}^k n_i$$

Hypothesis Testing for Treatment Effect:

- $H_0: \tau_i = 0$ for all i , vs. $H_a: \tau_i \neq 0$ for some i if treatment effect fixed.
- $H_0: \sigma_\tau^2 = 0$ vs. $H_a: \sigma_\tau^2 \neq 0$ if treatment effect is random.

Test Statistic for Treatment Effect: $F_0 = MS(T)/MS(E)$.

Rejection Region at α Level: $F_0 > F_{(k-1, N-k), 1-\alpha}$.

Two-Sided p Value: $1 - \text{PROBF}(F_0, k-1, N-k)$.

100(1 - α)% Confidence Interval for Treatment i :

$$\bar{y}_i \pm t_{(N-k, 1-\alpha/2)} \left(\sqrt{\frac{MS(E)}{n_i}} \right)$$

100(1 - α)% Confidence Interval for the Difference Between Treatments i and i' : $\bar{y}_i - \bar{y}_{i'} \pm t_{(N-k, 1-\alpha/2)}$

$$\left(\sqrt{MS(E) \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)} \right)$$

Related SAS Procedure: PROC ANOVA or PROC GLM.

Description of and Notes on the Method:

- PROBF(x , df1, df2) is the cumulative probability function for an F distribution in SAS, and $F_{(df1, df2), \alpha}$ represents a numerical value at which the cumulative probability function is α under an F distribution with (df1, df2) degrees of freedom where $x = a$ random value.
- When the treatment effect is fixed, it implies that $\sum_i \tau_i = 0$. When it is random, it is normally distributed with a variance of σ_τ^2 .
- Typical examples in this group are a placebo-controlled clinical trial with two or more dosage levels of the investigational drug, and an active-controlled study with a study drug at one or more dosage levels and a comparator agent at one or more dosage levels.

Several important things that this approach always brings are:

- i. Departure of homogeneity of variance: First, use Bartlett's test to determine whether the variances are homogeneous.^[6] If not, log-transformed data or a generalized F -test, as discussed in Ref. [6], p. 360, may be used.
- ii. Multiple/pairwise comparison: In general, pairwise comparisons are needed when an overall F -test for treatment difference is significant and it is desirable to determine which pair(s) of the treatments contributes to the differences. Many procedures have been proposed, including Fisher's least significant difference (LSD) test, Bonferroni-Dunn test, Tukey's honestly significant difference (HSD) test, Newman-Keuls test, Scheffe test, and Waller-Duncan test. Another test, particularly useful for comparing several treatments with a control, is the Dunnett test (1955).^[7]
- iii. Dose response: It is well known that when a dose response is desirable in a clinical trial (i.e., an investigation on how the efficacy of the drug is related to the dosage given in the trial), a common design is to allow each patient to receive all or part of doses available in the study, in order to reduce variability. However, it is not uncommon to use a parallel group design for investigating the dose response. The trend of dose response can be evaluated by setting up orthogonal linear contrasts (linear, quadratic, cubic, quartic, etc.) of the treatment means (Ref. [8], pp. 117–138).
- d. The treatment effect can be fixed or random, depending on how the treatment levels are selected in the study. In either case, the test statistic remains the same, but the implication of the statistical conclusion will vary.
- e. When k (the number of treatment groups) = 2, the F -test is exactly the same as for a two-sample t -test.
- f. This F -test for analysis of variance is non-directional. More specifically, it would signal a "difference," but without showing the direction of either a "superiority" or an "inferiority." However, with the aid of multiple comparisons, a better picture of the direction may be observed.
- g. PROC ANOVA is mainly for the equal sample size case, and PROC GLM can handle unequal sample sizes.

Kruskal-Wallis Test

Study Design: Parallel group design.

Model Assumptions: Independence.

Number of Observations per Patient for PD Parameter: 1.

Purpose: To compare the equality of the median for all treatment groups.

Hypothesis Testing: $H_0: M_1 = M_2 = \dots = M_k$ vs. $H_a: M_i \neq M_j$ for some i, j .

$$\text{Statistical Method: } k_0 = \frac{\frac{12}{N(N+1)} \left(\sum_{i=1}^k \frac{R_i^2}{n_i} \right) - 3(N+1)}{1 - \frac{\sum_{j=1}^g m_j(m_j^2 - 1)}{N(N^2 - 1)}}$$

Rejection Region at α Level: $K_0 > \chi^2_{(k-1), 1-\alpha}$.

Two-Sided p Value: PROBCHI($K_0, k-1$).

Related SAS Procedure: PROC NPAR1WAY WILCOXON.

Description of and Notes on the Method:

- a. The Kruskal-Wallis test is a non-parametric version of analysis of variance to deal with PD parameters with more than two treatments. It is an extension of the Wilcoxon rank-sum test, just like analysis of variance is an extension of the two-sample t -test.
- b. The data are ranked from the lowest to highest over the combined data set.
- c. Let y_{ij} = the observation of the PD parameter for patient j in treatment group i , r_{ij} = the corresponding rank for y_{ij} , n_i = sample size for group i . Then $R_i = \sum_{j=1}^{n_i} r_{ij}$ and $N = \sum_{i=1}^k n_i$.
- d. Assume there are g categories in which m_j observations are tied in the j th category.
- e. PROBCHI(x) is a cumulative probability function at x for a chi-square distribution in SAS, and $\chi^2_{(k-1), 1-\alpha}$ is a numerical value at which the cumulative probability is $1 - \alpha$ for a chi-square distribution with $k - 1$ degrees of freedom.
- f. For k (the number of treatment groups) = 2, the Kruskal-Wallis test is identical to the Wilcoxon rank-sum test.
- g. Another approximation test procedure to employ is a one-way analysis of variance with the ranked data.
- h. Pairwise comparisons may be employed in two ways if the overall Kruskal-Wallis test is significant (Ref. [4], pp. 402–405):
 - i. Using the Wilcoxon rank-sum test for each pair of treatment groups, with a caution of adjustment for the α level.
 - ii. Using a method similar to the Bonferroni-Dunn test for analysis of variance by identifying the minimum required difference between the means of the ranks of any two groups

Log-Rank Test (Mantel-Cox Test)

Study Design: Parallel group design.

Model Assumptions: Independence.

Number of Observations per Patient for PD parameter: 1.

Purpose:

- To estimate the survival rate for each treatment group.
- To compare the survival curves over time for two or more treatment groups.

Kaplan-Meier Estimate for $S_i(t)$ for Treatment Group i :

$$\hat{S}_i(t) = \prod_{j=1}^t \left(\frac{n_{ij} - d_{ij}}{n_{ij}} \right)$$

100(1 - α)% Confidence Interval for $S_i(t)$:

$$\hat{S}_i(t) \pm Z_{1-\alpha/2} \sqrt{\frac{\hat{S}_i(t)(1 - \hat{S}_i(t))}{n_{i0} - c_{it}}}$$

Hypothesis Testing for Survival Curves:

$$H_0: S_1(t) = S_2(t) = \dots = S_k(t) \text{ for all time } t \quad \text{vs.} \\ H_a: S_i(t) \neq S_j(t) \text{ for some } i, j, \text{ and } t$$

Statistical Method: $Q_0 = \sum_{i=1}^k \frac{(d_{i+} - e_{i+})^2}{e_{i+}}$ where

$$d_{i+} = \sum_{j=1}^b d_{ij}, \quad e_{i+} = \sum_{j=1}^b e_{ij}, \quad e_{ij} = \frac{d_{+j} n_{ij}}{n_{+j}} \\ d_{+j} = \sum_{i=1}^k d_{ij}, \quad n_{+j} = \sum_{i=1}^k n_{ij}$$

Rejection Region at α Level: $Q_0 > \chi^2_{(k-1), 1-\alpha}$.

Two-Sided p Value: $1 - \text{PROBCHI}(Q_0, k - 1)$.

Related SAS Procedure: PROC LIFETEST.

Description of and Notes on the Method:

- This test applies to the PD parameter when the time to a specified event (e.g., death) is of most interest. This test allows a comparison of the Kaplan-Meier survival curve between two or more treatment groups. Patients who discontinue from the study before the first occurrence of the pre-defined event or patients who never experienced the pre-defined event during the entire course of the study are called *censored*.
- The study is divided into b time periods, denoted by t for $t = 1, 2, \dots, b$.
- $S_i(t)$ = survival rate at time t for treatment group i
 - d_{ij} = number of patients experiencing predefined event for treatment group i at time j
 - n_{ij} = number of patients for treatment group i at beginning of time j (i.e., number of patients at risk at time j)
 - n_{i0} = number of patients recruited to treatment group i
 - c_{it} = number of patients censored for treatment group i before time t

Table 4 displays information on the number surviving and not surviving for each treatment group for any given time j .

Table 4 Frequency Table for Survival and Non-survival

At time j	Survival		Row subtotal
	Yes	No	
Treatment 1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
Treatment 2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Treatment k	d_{kj}	$n_{kj} - d_{kj}$	n_{kj}
Column subtotal	d_{+j}	$n_{+j} - d_{+j}$	n_{+j}

- There are several methods for obtaining the standard deviation for the Kaplan-Meier proportion. The one shown in the confidence interval was given by Parmar and Machin (1984).^[9]
- Another common test for survival curves is known as the log maximum-likelihood test.

Group 3

Fisher Exact Test

Study Design: Parallel group design.

Model Assumptions: Binomial distribution, independence.

Number of Observations Per Patient for PD parameter: 1.

Purpose: To compare the proportion of responders between two treatment groups.

Hypothesis Testing: $H_0: \pi_1 = \pi_2$ vs. $H_a: \pi_1 < \pi_2$

$$\text{Test Statistic: } p_0 = \sum_{x \leq x_1} \frac{\binom{n_1}{x} \binom{n_2}{r-x}}{\binom{N}{r}}$$

Rejection Region at α Level: $p_0 < \alpha$.

Observed Frequency Table: See Table 5.

Related SAS Procedure: PROC FREQ

Description of and Notes on the Method:

- The x in the test statistic refers to the count in the (1,1) position of the frequency table, and it is obtained by fixing n_1 , n_2 , N , and r in the observed frequency table.
- $\binom{x}{y} = \frac{x!}{y!(x-y)!}$
- The requirement of this method is that all of the values for the PD parameter be classified into two responses, regardless of the original measurement of the PD parameter.

Table 5 Observed Frequency Table

	Treatment 1	Treatment 2	Row total
Responder	x_1	x_2	$r (= x_1 + x_2)$
Non-responder	y_1	y_2	$y_1 + y_2 = N - r$
Column total	$n_1 (= x_1 + y_1)$	$n_2 (= x_2 + y_2)$	$N (= n_1 + n_2)$

- This test generally applies to studies with small sample size (< 20).
- π_i = the true unknown rate for responders for treatment group i .
- When $H_a: \pi_1 > \pi_2$ is desirable, we simply change the summation $x \leq x_1$ to $x \geq x_1$ in the test statistic.
- When $H_a: \pi_1 \neq \pi_2$ is desirable, the p -value is obtained by first computing the probabilities for all possible combinations having the same row and column totals as the observed frequency table. Then we add the probabilities for combinations that have a probability smaller than or equal to the observed one.
- This test can be generalized into a study with more than two responses and more than two treatment groups.

McNemar's Test

Study Design: Parallel group design.

Model Assumptions: Binomial distribution, dependence.

Number of Observations per Patient for PD Parameter: 2.

Purpose: To compare the proportions of responders between two treatment groups.

Hypothesis Testing: $H_0: \pi_1 = \pi_2$ vs. $H_a: \pi_1 \neq \pi_2$

$$\text{Test Statistic: } Q_0 = \frac{(|B - C| - 1)^2}{B + C}$$

Rejection Region: $Q_0 > \chi^2_{1, 1-\alpha}$.

Two-Sided p Value: $1 - \text{PROBCHI}(Q_0, 1)$.

100(1 - α)% Confidence Interval for $\pi_1 - \pi_2$:

$$\frac{B - C}{n} \pm z_{(1-\alpha/2)} \frac{1}{n} \sqrt{B + C - \frac{(B - C)^2}{n}}$$

Observed Frequency Table: See Table 6.

Related SAS Procedure: PROC FREQ.

Description of and Notes on the Method:

- Each patient contributes two observations under two conditions, such as before and after treatment, or under two different treatments, or at two different time points.
- This assumes patients' well-being with regard to the PD parameter are the same before treatments. This assumption, in reality, may not be true.
- π_1 is estimated by $(A + B)/N$, and π_2 is estimated by $(A + C)/N$.
- The quantity "1" in the test statistic is a correction for continuity, which may not be needed for the analysis.

- The chi-square test statistic without the correction for continuity is also equivalent to a Z -test statistic of the form $Z_0 = (B - C)/\sqrt{B + C}$, which follows a standard normal distribution and is useful for a one-sided test.
- The McNemar test has been extended to a $k \times k$ ($k > 2$) case in the Stuart-Maxwell test^[10] (i.e., for studies with k treatment groups and k responses).

Group 4

(Pearson's) Chi-Square Test

Study Design: Parallel group design.

Model Assumption: Multinomial distribution, independence.

Number of Observations per Patient for PD Parameter: 1.

Purpose: To determine and compare the relationship between the PD parameter and the explanatory variable.

Hypothesis Testing: $H_0: \pi_{ij} = (\pi_i)(\pi_j)$ for all i, j vs. $H_a: \pi_{ij} \neq (\pi_i)(\pi_j)$ for some i, j

$$\text{Test Statistic: } Q_0 = \sum_{j=1}^k \sum_{i=1}^m \frac{\left(\left| n_{ij} - \frac{n_{+j}n_{i+}}{n_{++}} \right| - 0.5 \right)^2}{\frac{n_{+j}n_{i+}}{n_{++}}}$$

Rejection Region at α Level: $Q_0 > \chi^2_{(m-1)(k-1), 1-\alpha}$.

Two-Sided p Value: $1 - \text{PROBCHI}(Q_0, (m-1)(k-1))$.

Observed Frequency Table ($m \times k$ Table): See Table 7.

Related SAS Procedure: PROC FREQ.

Description of and Notes on the Method:

- π_{ij} = unknown true proportion for having response i in treatment group j
 - π_i = unknown true proportion for having response i regardless of treatment group
 - π_j = unknown true proportion for treatment group j regardless of responses
- The null hypothesis indicates the independence between treatment group and response.
- There is another type of hypothesis testing for a chi-square test concerning a test for homogeneity; it means all proportions for all cells are the same.
- The quantity "0.5" in the test statistic is the "continuity correction." An absolute value for the difference in the denominator is always needed when the "continuity correction" is considered. The "continuity correction" in the test statistic may not be needed.

Table 6 Observed Frequency Table

Trt 1	Trt 2		Row total
	Responder	Non-responder	
Responder	A	B	A + B
Non-responder	C	D	C + D
Column total	A + C	B + D	N (= A + B + C + D)

Table 7 Observed Frequency Table ($m \times k$ Table)

	Treatment 1	Treatment 2	...	Treatment k	Row total
Response 1	n_{11}	n_{12}	...	n_{1k}	n_{1+}
Response 2	n_{21}	n_{22}	...	n_{2k}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
Response m	n_{m1}	n_{m2}	...	n_{mk}	n_{m+}
Column total	n_{+1}	n_{+2}	...	n_{+k}	n_{++}

- e. There are a number of arguments about using the chi-square test (Ref. [4], p. 218):
- i. Do not use the chi-square test when $n_{++} < 20$.
 - ii. All n_{ij} s are at least 5 if $20 \leq n_{++} < 40$.
 - iii. All expected cell frequencies are at least 1 if $n_{++} \geq 40$.
 - iv. For a 2×2 table with $n_{++} < 20$, Fisher exact test is suggested.
- f. Another approach to handling an $m \times k$ table is to employ a likelihood-ratio chi-square test with the test statistic expressed as follows:

$$G^2 = 2 \sum_j \sum_i n_{ij} \log \left(\frac{\frac{n_{ij}}{n_i + n_{+j}}}{\frac{n_{++}}{n_{++}}} \right)$$

This test statistic also follows a chi-square distribution with $(m - 1)(k - 1)$ degrees of freedom.

CONCLUSION

Because there are potentially many other different cases that may arise in a clinical setting, this entry only a purpose of covering mostly known cases and the most commonly used statistical methodologies.

REFERENCES

1. Satterthwaite, F.W. An approximate distribution of estimates of variance components. *Biom. Bull.* **1946**, *2*, 110–114.
2. *SAS Procedures Guide. Version 6*, 3rd Ed.; SAS Institute: Cary, NC, 1990, 617–634.
3. Lehmann, E.L. *Nonparametrics: Statistical Methods Based on Ranks*; Holden-Day: San Francisco, CA, 1975; 12.
4. Sheskin, D.J. *Handbook of Parametric and Nonparametric Statistical Procedures*; CRC Press: Boca Raton, FL, 1996.
5. Conover, W.J.; Iman, R.L. Rank transformations as a bridge between parametric and nonparametric statistics. *Am. Stat.* **1981**, *35*, 124–129.
6. Winer, B.J.; Brown, D.R.; Michels, K.M. *Statistical Principles in Experimental Design*, 3rd Ed.; McGraw-Hill: New York, 1991, 106–107.
7. *SAS/STAT User's Guide. Version 6*, 4th Ed.; SAS Institute: Cary, NC, 1990, Vol. 2, 945.
8. Berry, D.A. *Statistical Methodology in the Pharmaceutical Sciences*; Marcel Dekker: New York, 1990, 117–138.
9. Parmar, M.K.B.; Machin, D. *Survival Analysis: A Practical Approach*; Wiley: New York, 1995, 36–40.
10. Fleiss, J.L. *Statistical Methods for Rates and Proportions*; Wiley: New York, 1973, 77–79.

Pharmacoeconomics

Joseph Heyse and John R. Cook

Merck Research Laboratories, West Point, Pennsylvania, U.S.A.

Michael F. Drummond

Centre for Health Economics, York, U.K.

Abstract

The purpose of this entry is to provide an overview of the terminology, family of technologies, and key methodologic features of health economic evaluations. The primary focus is on the statistical considerations of planning and analyzing studies and interpreting the results.

INTRODUCTION

The evaluation of pharmaceutical and biological products has traditionally focused on considerations of safety and efficacy. Federal regulation requires that studies of safety and efficacy be conducted before a product is approved and marketed, and as a result, the methods for product development have been enhanced. In recent years, the information needs of health care decision makers has expanded beyond the traditional measures of safety and efficacy to include patient outcomes, cost, and cost-effectiveness. The primary motivation for this shift is a growing concern over the cost of health care; the current availability of health-related interventions exceeds society's ability and willingness to pay. One important consequence in the changing focus of health care decision making is the need for appropriate methods and decision rules to guide choices toward interventions most likely to yield the greatest value for the level of investment. "Health economics" is a field of study that applies economic concepts to health and health-care systems. An important topic in health economics is the evaluation of health care treatments and programs on the basis of costs (inputs) and consequences (outputs). "Pharmacoeconomics" focuses on economic evaluations of the use of pharmaceutical products, programs, and services. The purpose of this article is to provide an overview of the terminology, family of technologies, and key methodologic features of health economic evaluations. The primary focus is on the statistical considerations of planning and analyzing studies and interpreting the results.

IMPORTANCE OF ECONOMIC EVALUATION

The concern over the rising costs of health care has prompted several initiatives that influence the price, availability, or use of medicines. Many hospital and managed

care pharmacies have a strict budgeting process for providing pharmaceutical products and services. Suboptimal decisions are sometimes made in this setting when the pharmacy decisions are not coordinated with the larger health care system. For example, it is possible that the use of a new, higher-priced antibiotic would cause an increase in the pharmacy budget yet may reduce costs overall if the use of the antibiotic were associated with shorter lengths of hospital stay or a reduced rate of reinfection. For this reason, many hospitals and managed care organizations have lists, called "formularies," of selected pharmaceuticals and the dosages judged to be the most appropriate or cost advantageous for patient care. Physicians who practice as part of a managed care provider organization are required or encouraged to prescribe from the formulary. Many managed care organizations also recommend specific disease-management programs of care as a guideline for practicing physicians. Many countries maintain drug formularies at the national level to determine the reimbursement level for drugs. Other initiatives are specifically directed at prices; one example used in Germany and the Netherlands is a practice called "reference-based pricing," in which product reimbursement is set at the same level for all drugs in a therapeutic class. The reference price is usually set according to the least expensive drug in the class. Reference-based pricing is becoming more common worldwide as a way to control health care costs.

Another sign of the increasing importance of economic evaluations in health care decision making has been the emergence of guidelines for the conduct of cost-effectiveness analyses. The purpose of the guidelines has been to encourage a rational diffusion and application of medical technologies. Australia^[1] and Ontario, Canada,^[2] have proposed guidelines for the provision of cost-effectiveness data before drugs are listed on the national or provincial formulary. Several European countries have also implemented guidelines.^[3] The U.S. Food and Drug Administration^[4]

applies strict standards to claims of cost-effectiveness by pharmaceutical companies. These standards include the use of adequate and well-controlled studies demonstrating evidence for the comparative claims about the drugs in question. Marketing statements about drugs and vaccines that are not drawn directly from the product label generally require the same level of evidence as claims. Major medical journals, including *The New England Journal of Medicine*^[5] and the *British Medical Journal*^[6] have issued guidelines for the design and reporting of economic evaluations considered for publication. Most recently, the Panel on Cost Effectiveness in Health and Medicine, sponsored by the U.S. Public Health Service, reported recommendations on standardized methods for conducting cost-effectiveness analyses.^[7,8] Whereas most guidelines have dealt specifically with the methodologic aspects of studies, Hillman et al.^[9] discussed guidelines for the conduct of studies sponsored by pharmaceutical companies.

These initiatives have also resulted in an increased level of competition among pharmaceutical companies to make comparative claims on the basis of pharmacoeconomic outcome variables, including resource use, cost, and cost-effectiveness. Many clinical development programs now include pharmacoeconomic objectives, which can greatly affect the overall design and implementation of clinical trials. Nonclinical studies of disease epidemiology and the costs of treating the targeted diseases are also common in clinical development programs as a way of supplementing trial data. These changes have broadened the traditional roles and responsibilities of the clinical trial statistician.

Finally, the Academy for Managed Care Pharmacy (AMCP) has proposed a standard format for the provision of cost-effectiveness data to managed care plans. Although this does not constitute a formal requirement, it might lead to more plans requesting such data.^[10]

TYPES OF ANALYSES

Economic evaluations of health care technologies involve a formal assessment of the costs and consequences of medical treatments or treatment programs. The major components of an economic evaluation include identifying, measuring, and valuing both the resources applied to the health care program (inputs) and the health improvements derived from the health care program (outputs). The theoretical foundation of the assessment is the concept of economic efficiency, in the sense of optimizing health outcomes for given fixed levels of resources consumed.^[11, 12] The basic model recognizes that all health care interventions involve costs (measured in terms of the economic value of the resources used to treat or prevent illness) and health consequences (such as improvements in health status) and may include savings in future health care costs.

One type of study often conducted at the start of a health economic research program is a "burden-of-illness study."

This form of analysis typically combines epidemiologic data on the incidence and prevalence of the disease with cost data in order to estimate the total direct and indirect cost attributable to a disease incurred by the population over a defined period. Burden-of-illness studies are useful for influencing prioritization by alerting policy makers and health care professionals to the importance of a problem. Several important issues have been raised about the conduct and use of these studies. Drummond^[13] noted that their use is limited because they are essentially non-comparative and therefore do not indicate whether more or less resource should be devoted to treatment. In addition, the methods for estimating the value of lost productivity and reduced productivity are thought to possibly exaggerate those costs. This might arise at the societal level, e.g., because lost labor can usually be replaced if the workforce is not at full employment. At the other extreme, burden of illness studies do not account for the intangible costs related to a diminished quality of life, or for the impact of the disease on others in the form of grief. They can, however, provide a benchmark against which a new intervention can be evaluated.

Comparative economic evaluations provide the most useful information for guiding health care decisions. Researchers have developed several forms of comparative economic analysis, which can broadly be classified into the following groups:^[14] cost analysis, cost-consequences analysis, cost-effectiveness analysis, cost-utility analysis, and cost-benefit analysis.

A complete evaluation of therapy costs is needed for all forms of economic analyses. In a "cost analysis," or "cost-minimization analysis," the comparison is based on an accounting of the resources used as input to the therapy or health care program. This form of analysis is used when the effectiveness of the candidate and comparative therapies can be assumed to be equal and when the comparison is made strictly on an evaluation of costs. Although therapies are rarely equal in terms of effectiveness, cost minimization can provide a reasonable approximation in some cases. For example, de Lissoy et al.^[15] performed a cost analysis of imipenem and cilastatin compared with tobramycin plus clindamycin for the treatment of patients suspected of having acute intra-abdominal infection. The study compared hospital treatment costs during an episode of infection for patients enrolled in a clinical trial comparing the effectiveness of the two therapy groups. Costs of hospital stays in regular and intensive care units, medical procedures, antibiotics, other drugs, and laboratory evaluation were included in the analysis. The mean total costs did not differ significantly between the two treatment groups (\$12,040 for imipenem and cilastatin and \$11,220 for tobramycin plus clindamycin); however, the study showed that costs were significantly affected primarily by the severity of illness, as measured by Acute Physiology and Chronic Health Evaluation acute physiology score. Use of a cost-minimization analysis for this study was

conservative for imipenem and cilastatin because the principal finding in the main clinical study was a lower incidence of treatment failure in patients in the imipenem and cilastatin group ($P = 0.043$). This is not only an important health outcome but is an additional source of costs that was not included in the economic evaluation.

A “cost-consequences analysis” extends the cost-minimization analysis by also considering health outcomes. This study displays complete but separate enumerations of the costs and health outcomes (measured in natural units) of the candidate therapies. Because the results are presented in a disaggregated form, the overall analysis of the study is left to the user; no specific value is placed on the outcomes. Cost-consequences analyses are most useful to analysts who have the technical ability to use the reported results as a basis for applying their own methods of evaluation.

A “cost-effectiveness analysis” measures the health outcomes of the candidate treatment programs in natural units. Examples include variables such as life-years gained, fractures avoided, and coronary events avoided. In the classic form of cost-effectiveness analysis, the results are expressed by the ratio cost per unit of health benefit gained as a measure of efficiency. For example, Oster and Epstein^[16] used an epidemiologic model based on the Framingham Heart Study to evaluate the cost-effectiveness of pharmacologic therapy for reducing cholesterol levels. They estimated that for a 40- to 44-year-old man with elevated serum cholesterol levels, therapy with the pharmaceutical agent cholestyramine costs \$76,500 per additional life-year gained.

A “cost-utility analysis” is a form of cost-effectiveness analysis that values the consequences of treatment programs by using health state preference scores or utility weights. For example, this method allows assessments on the quality of life-years gained, not just the crude number of years. This approach is particularly useful for evaluating treatments that may extend life expectancy (possibly by preventing fatal disease in the future) but may be associated with adverse experiences in the short term. Cost-utility analyses also provide a common metric for overall reductions in morbidity and mortality. For the broader policy perspective, this allows comparisons across treatments for different diseases or even for comparisons of treatment versus prevention strategies. In the U.S. literature, cost-utility analyses are rarely used.

A “cost-benefit analysis” evaluates health outcomes in monetary terms rather than in terms of patient outcome. This approach makes the measure of health benefit commensurate with resource costs. This may be the broadest form of analysis because it allows direct assessment of the benefits of the treatment program with respect to the costs and has the potential to capture significant nonhealth benefits. Two methods for capturing monetary value to health states are used most often.^[14,17] The “human capital method” evaluates health benefits in terms of the

economic productivity gained from the increased survival or decreased morbidity. Wage rates are typically used for the analysis. The “willingness-to-pay” method defines the maximum amount of money an individual is prepared to spend to secure a defined health state. Both methods present significant analytic challenges,^[7,14] but the willingness to pay approach is considered by most economists to be the more theoretically correct because it recognizes the impact of mortality and morbidity on aspects of life beyond employment, such as housework and leisure time.

There are still relatively few willingness-to-pay studies in the health economics literature. In one example, O'Brien et al.^[18] asked enrollees in a health maintenance organization whether they would be willing to pay to upgrade their insurance coverage to include a new supportive drug used in cancer chemotherapy.

DESIGN CONSIDERATIONS

Study Perspective

The perspective of the economic evaluation determines which costs and health outcomes are relevant and plays an important role in how they should be valued. Possible study perspectives include that of the patient, employer, health care provider, hospital, managed care organization, insurance company, or society at large. Gold et al.^[7] recommended using the “societal perspective” because it is the most comprehensive and is relevant to the broad allocation of resources. This perspective includes a complete evaluation of the health effects and costs experienced by anyone significantly affected by the intervention. Economic evaluations of pharmaceutical products often use a narrower perspective that would be relevant to specific customer segments. The study perspective is an important design feature and should be defined early in the study process. Drummond et al.^[14] recommended performing the evaluation for various sectors and combining results to allow the study to potentially address questions from several perspectives. This approach might be useful if for example the evaluation is intended to address costs that may be incurred by a managed care organization or a hospital and the benefits applied to specific patient populations.

Costing Methods

Three general categories of costs may be examined within an economic evaluation of a health care technology: direct costs, indirect costs, and intangible costs.

“Direct costs” include the value of medical resources used in the provision of the treatment and the consequences of the treatment, both in terms of beneficial effects or unwanted side effects. For pharmaceutical programs, direct costs include the costs of drugs, medical staff, medical facilities, medical procedures, diagnostic and

laboratory tests, and supplies. When appropriate, direct nonmedical costs, which are borne by the patient seeking care, may be included in the analysis. These would include the costs of transportation, child-care, housekeeping, and so forth. Several authors have discussed methods for assigning direct costs to units of health care resources;^[14,19] such discussion is beyond the scope of this paper. Most studies capture patient-specific data on the health care resources used for treatment and then conduct separate studies to estimate costs for these resources. Patient-level cost data can then be computed by combining the cost and resource data from the trial.

“Indirect costs” include the value of productivity lost because of the underlying illness or its treatment. Luce et al.^[19] referred to these costs as “productivity costs” to avoid confusion with the term as it is used by accountants for overhead costs. They include the costs associated with lost or impaired ability to work or to engage in leisure activity because of illness and lost economic productivity caused by premature death. Historically, the “human capital approach” has been used to estimate productivity costs, but this method has been criticized for its use of gross earnings that include employee benefits. It is argued that using earnings raises both ethical and equity concerns because this method does not place a value on unpaid employment (e.g., homemaking) and may undervalue the impact of the disease on women and the elderly. In addition, the method may overestimate the true costs because short-term losses in productivity may be lower than average wage rates. For longer-term losses, companies can replace laborers to minimize losses in production. In this context, Koopmanschap and Rutten^[20] proposed a “friction cost” approach based on the notion that the amount of production lost depends on the time span that organizations need to restore the initial production level. Their estimates for the Netherlands in 1988 show that the human capital method can provide estimates that exceed the friction cost by a factor of 10. Luce et al.^[19] recommended that productivity costs be estimated and reported separately in cost-effectiveness analysis to avoid double counting the beneficial effects of treatment.

Outcomes such as reduction of pain, emotional well-being, social functioning, activities of daily living, mobility, and cognitive functioning are all important considerations in the economic evaluation of a pharmaceutical treatment program. Failure to include these factors reduces the credibility of the assessment and may threaten its adoption by decision makers. The value associated with these types of outcomes has been called “intangible benefits.”^[14] These measures are often captured in clinical trials by using validated quality-of-life instruments. For purposes of economic evaluation, their impact is included as part of the effectiveness measure in a cost-utility analysis. Monetary values of intangible outcomes would be (implicitly) included in the willingness-to-pay method.

Outcome Measures

Because economic evaluation focuses on efficiency, measuring the benefit of the intervention in terms of health outcomes is as important as measuring benefit in terms of cost. Several approaches have been used, and there is no clear consensus on the end points to use for the evaluations.^[17] The choice is often driven by complexities in the measurement instruments and by possible limitations in the design of the clinical programs. Short-term “surrogate, or intermediate, measures” of response to treatment are attractive in these situations when they provide reliable indicators of long-term patient outcome. For example, in the area of cardiovascular disease, meta-analysis^[21] and long-term clinical trials^[22,23] have shown that lowering serum cholesterol levels confers the benefits of a reduced risk for coronary heart disease and increased survival and thus supports the cholesterol hypothesis. These findings established the rationale for a cost-effectiveness analysis based on intermediate endpoints by Schulman et al.^[24] These researchers reported a comparative analysis of various lipid-lowering drugs by plotting percentage reduction in cholesterol levels by the 5-year cost of patient management associated with the use of each drug. Economic evaluations based on intermediate endpoints assume that the effect of therapy on the ultimate clinical endpoint is fully captured through an established relationship with the surrogate marker. This assumption can be difficult to validate without the appropriate long-term clinical trials.

Cost-effectiveness analyses can also be based on different types of “clinical end points,” such as coronary events avoided or fractures avoided. An advantage of using end points prevented as the effectiveness measure is that they are closely related to the clinical objectives of therapy. This makes ratios such as cost per event avoided a meaningful measure for decision makers. When the medical condition is potentially life-threatening, “mortality end points” such as deaths avoided are applicable. One common measure of cost-effectiveness that has been used for medical evaluations is the “cost per life-year gained.” The advantage of using life-years gained is that this measure has a very clear definition as the difference between the treatment groups in the areas under the respective survival curves. As such, it takes into account the number of life-years during which the risk for death is reduced. The obvious disadvantage of mortality end points is that they do not apply to the evaluation of treatments for non-life-threatening illnesses and do not consider the patient’s quality of life.

The quality-adjusted life-year (QALY) method^[25] is an attempt to combine the concepts of quality and length of survival. The QALY method uses measures of health outcome that assign, for each period of time, an overall measure of a patient’s quality of life by using a weight defined from 0 (death) to 1 (perfect health). Health states with diminished quality are scored less than 1. The number of

QALYs is simply the sum of the QALY weights over the time horizon for the evaluation. The number of QALYs gained is a popular measure of effectiveness used in cost-utility analyses.

The QALY concept is not without controversy. This measure can be interpreted as the number of years of perfect health a patient values equivalently to their years with diminished quality. For example, a QALY of 1 could be derived from a single year with perfect life or from several years with diminished quality. Mehrez and Gafni^[26] argued that patients would value the same number of QALYs differently depending on the path by which the QALYs are accumulated. They proposed a healthy years equivalent (HYE) measure that is derived by using a complex two-step procedure to determine patient preferences for different paths for accumulating QALYs. At the other extreme, Cox et al.^[27] argued that the QALY approach is overly complex and that simpler measures addressing the dimensions of quality of life (such as physical or mental well being) and survival separately should be used.

The end point chosen is an important factor in economic evaluations. Drummond et al.^[28] showed that the choice of end point can have an important impact on the ranking of alternative therapies for lipid lowering. The ranking depended on the value placed on life-years after coronary heart disease developed, the effect that drug therapy had on quality of life, and even the discount rate at which future QALYs were valued. Ultimately, the selection of end points depends on the nature of the disease and treatment program being studied, and, if this can be determined, what is most relevant to the decision makers.

Discounting

Economic evaluations should include the value of all costs, adverse experiences, and benefits that are the direct consequence of the disease or treatment program under study. This may include a stream of costs and health effects over a long period of time. It is important to recognize that the costs and benefits may vary over time and may be different for the programs under study. For example, the cost of drug therapy to reduce serum cholesterol levels is incurred with the initiation of therapy, but the benefits are realized through a reduced risk for coronary heart disease in later years. In addition, comparisons of drug therapy with surgical procedures need to account for the different timing of costs and benefits.

The method of “discounting to present values” is used to account for the timing of the stream of costs and health benefits. Computationally, $PV(X) = \sum X_i / (1 + r)^i$ gives the present value for the stream of costs or health effects, denoted by X_i for year i , using a discount rate r . The value assigned to r is an important and controversial design consideration, as is the notion of discounting both future costs and future health effects at the same rate. For example,

Van Hout^[29] argued that different discount rates may be required for costs and effects. The Panel on Cost Effectiveness in Health and Medicine^[7,8] recommended discounting costs and health outcomes at the same 3% discount rate. A riskless 3% discount rate was chosen to be consistent with the real rate of return on long-term U.S. government bonds. To maintain comparability with previously published analyses, they recommend also using a 5% rate and conducting a sensitivity analysis for rates ranging from 0% to 7%. Undiscounted gains in life expectancy and quality-adjusted life expectancy can be reported separately as they are often of interest to the reader.

RESEARCH METHODS

A variety of research methods have been used for economic evaluations, including modeling designs and studies using clinical economic trials that involve prospective or retrospective data collection.

Modeling Studies

Modeling studies have employed two main types of methods to simulate patient outcomes and cost by using assumptions about the effects of therapy. Clinical “decision-analytic models” systematically lay out all probable pathways and consequences of the candidate treatment programs in a decision tree. The likelihood and expected cost of each pathway are estimated by use of data from clinical trials or epidemiologic investigations. These models have been used mostly for treatments for acute conditions, which may resolve in a short period. “Markov models” are used for diseases and treatments that are characterized by recurrent disease states or treatment programs.

“Epidemiologic models” have been used for disease areas in which the outcomes are related to sets of risk factors and the clinical outcomes may take years to develop. Several models for the treatment of elevated cholesterol levels have been developed on the basis of data from the Framingham Heart Study.^[16,30,31] These models estimate the risk for coronary heart disease according to pretreatment and posttreatment cholesterol levels by using risk equations derived from the Framingham study and by using estimates of treatment efficacy derived from short-term clinical trials. Similar types of models have been developed for osteoporosis and HIV infection on the basis of epidemiologic links between clinical outcomes and risk factors that can be altered by treatment.

Clinical-Economic Trials

The collection and analysis of economic data in clinical trial settings is becoming more common. These types of studies have been called “piggyback studies,” a term denoting that the scope of the trial was broadened to address economic

questions in addition to the clinical evaluation. Typically, they are undertaken during Phase III of the drug development process. Many of these studies have the advantages of being multicenter, randomized, double-blinded, and placebo-controlled—features that offer a high degree of internal validity to allow researchers to address specific questions about the safety and efficacy of the candidate treatment.

Studies of this type, however, have important drawbacks for economic evaluations.^[32,33] They may have insufficient sample size for addressing questions about cost and, more important, lack generalizability and external validity because of the use of a comparator (often placebo) that may be inappropriate in a real-world setting. Moreover, the studies are usually conducted in specialized medical centers with restrictive patient populations, and they define patient care by using rigorous clinical protocols in which the physician is blinded to knowledge of treatment. All of these factors limit the usefulness of these studies in providing data to answer questions about the cost and consequence of therapies as they will be used in real-world settings.

Schulman et al.^[34] and Mauskopf et al.^[35] discussed strategies for undertaking economic assessment within the Phase I through Phase IV clinical development program. Drummond and O'Brien^[36] and O'Brien and Drummond^[37] addressed issues of clinical and statistical significance in the context of socioeconomic evaluations of medicines. They recommended using confidence interval methods instead of strict accept/reject hypotheses as a way to permit the readers of the published findings to assess the magnitude and significance of the observed differences. Schulman et al.^[38] discussed their attempts to overcome these limitations in the economic evaluation of the Flolan® International Randomized Survival Trial (FIRST), a multinational Phase III clinical trial of Flolan® in the treatment of congestive heart failure. Freeman-tle and Drummond^[39] addressed the issue of blinding in clinical-economic trials, noting that blinding may lead to artificialities in treatment protocols that may affect costs and potential benefits.

An alternative to the piggyback design is the “randomized clinical economic trial” done expressly for cost-effectiveness analysis. Oster et al.^[40] discussed the design of a randomized trial intended to assess effectiveness and cost of two cholesterol treatment programs that involved stepped-care regimens with several drug choices. One treatment program was modeled after the 1988 guidelines of the National Cholesterol Education Program. This stepped-care approach recommended initiating niacin therapy and then introducing a series of other pharmacologic agents until the low-density lipoprotein cholesterol levels attained target levels. The other program was a regimen that started with the HMG-CoA reductase inhibitor lovastatin and allowed changes in therapy if the target level was not attained or if the patient preferred to switch therapy. Lovastatin was also allowed in the final step of

the stepped-care regimen. The study was conducted in a managed care setting; after randomization, patient management approximated usual clinical practice. Provider compliance with the treatment plans was encouraged but not enforced, and patients paid for medication as they customarily would. Another feature of the study is that it used the intention-to-treat principle for both design and analysis because patients in the different program groups may actually have been prescribed the same drug at some points in the trial.

“Prediction models” can also be developed using the clinical and health care utilization data collected during the main trial. This allows the analysis to consider, e.g., questions about longer time horizons or specific patient subgroups. Mark et al.^[41] supplemented data from the 1-year Global Utilization of Streptokinase and TPA for Occluded Arteries (GUSTO) study with data on projected life expectancy and hospital cost data to evaluate the cost-effectiveness of tissue plasminogen activator compared with streptokinase for acute myocardial infarction. By using cost-effectiveness analysis, the authors concluded that the added costs of tissue plasminogen activator therapy were consistent with those of other well-accepted medical technologies when evaluated in comparison to the benefits of therapy. Similarly, the Diabetes Control and Complications Trial (DCCT) group^[42] examined the cost-effectiveness of aggressive approaches to the management of insulin-dependent diabetes mellitus. They used a Monte Carlo simulation model that was based on the clinical end point results of the DCCT and was supplemented with other clinical, epidemiologic, and cost studies. Both of these studies used novel modeling techniques and secondary sources of data to overcome limitations in the designs of the main studies to answer the relevant economic questions.

METHODS OF ANALYSIS

Much attention has recently focused on statistical methods for assessing uncertainty in health economic evaluations.^[43–45] Uncertainty in the estimate of program costs and effects arises from two primary sources. The first is uncertainty in assumptions about the form of the underlying economic model, including values for the key model parameters, such as the time horizon or discount rate. In other situations, economic models use point estimates for parameters relating to patient outcomes or cost. O'Brien et al.^[46] calls this first type of evaluation a “deterministic analysis” and recommend that uncertainty be assessed by using a comprehensive “sensitivity analysis.” Briggs et al.^[47] discuss the different forms of sensitivity analysis that can be applied. These may include simple reporting of the results over a range of parameters, an analysis of extremes, or a threshold analysis. Probabilistic sensitivity analyses are also used in some studies.

The second source of uncertainty is sampling variability in model variable estimates, which are derived from studies of individual patients. This type of evaluation is called a “wholly stochastic analysis,” and formal statistical analyses can be used to estimate variability and to construct the 95% confidence interval estimates of the key economic parameters. Evaluations that involve uncertainty from both model parameters and sampling variability are called “partially stochastic analyses.” The rest of this section briefly discusses the statistical considerations involved with analyzing cost data and conducting cost-effectiveness evaluations.

Analysis of Cost Data

Analysis of cost data should first determine whether treatment programs differ in the actual resources utilized (e.g., hospitalization, procedures, and drug use). A candidate treatment may reduce one type of resource but increase another type. Computing patient cost involves combining resource data with prices in a way that is consistent with the study perspective. The statistical analysis of cost data can present several potential challenges. The distribution of costs is typically right-skewed, with a few patients incurring very large costs. Use of summary measures that are insensitive to extreme values, such as the median, are usually not desirable to organizations attempting to manage total aggregate patient costs. Patients with high resource use are substantial contributors to aggregate treatment costs but have a small influence on the median.

Cost distributions also tend to exhibit large variability that can pose problems for the provision of accurate and precise estimates of cost differences with the appropriate level of statistical power. The variability in cost data often increases with mean cost to such a degree that the usual normalizing and variance-stabilizing transformations are not helpful. For these reasons, nonparametric methods, resampling methods (e.g., bootstrap methods), and randomization tests may be preferred to parametric tests done by using analysis of variance. Heyse et al.^[48] provided an overview of statistical considerations in analyzing cost and utilization data.

Another important complication relates to patient cost data that are subject to right censoring. Censoring occurs whenever study data on patient outcomes collected over time are incomplete. Difficulties arise in the analysis because only partial data on some patients are available. This situation can arise because of study design considerations (e.g., a clinical trial uses a rolling enrollment period and a fixed study termination) or because a patient is lost to follow-up. The latter can occur for many reasons: for example, patients may transfer to other health care providers or drop out of the study because of the treatment's poor efficacy or adverse effects. Patient censoring due to death provides several challenges in the statistical analysis. Some

authors consider patient death a censored observation for cost analysis. For example, Dudley et al.^[49] suggested that for cost studies of surgical procedures, deaths should be treated as censored observations so that hospitals with high mortality rates and therefore lower costs are not rewarded. The alternative argument is that patient death presents a well-defined end point for the accumulation of costs and therefore is uncensored.

Recently, health economists have shown interest in statistical methods to estimate mean costs when data for some patients are right censored. Dudley et al.^[49] compared several analytic models, including both parametric and nonparametric survival models for estimating the effects of clinical factors on the hospital costs of coronary bypass graft surgery. Fenn et al.^[50] pointed out the biases that result when the censored observations are either excluded from the calculations (called the uncensored method) or included as complete observations (the full sample method). They suggested using the Kaplan–Meier estimators to incorporate the censoring information. However, Carides and Heyse^[51] showed that direct application of this method can also result in a substantial bias.

Lin et al.^[52] proposed a method that uses the sample mean costs for patients who die within distinct time intervals defined for the study (e.g., monthly or yearly intervals). These interval means are combined with Kaplan–Meier estimates of dying within the interval to obtain an estimate of mean cost. Carides et al.^[53] improved on the method of Lin et al. by using a two-stage estimator that first fits the uncensored patient cost data to an appropriate statistical model using linear or nonlinear regression and then applies Kaplan–Meier survival estimation to the fitted values. Both methods are statistically consistent and have only a slight bias in small samples but the two-stage method gains efficiency from the modeling or smoothing that occurs during the regression stage. A final method proposed by Zhao and Tsiatis^[54] for estimating the distribution of quality-adjusted survival times was modified by Bang and Tsiatis^[55] to estimate the survival distribution for costs. This method computes the weighted sum of the number of patients with costs exceeding a specific value y . The weights are directly related to the censoring probabilities at each time point. The distribution can then be used to compute summary statistics such as the mean. A disadvantage is that the estimator can be highly unstable for small weights.

Cost-Effectiveness Ratios

The cost-effectiveness of a candidate treatment or treatment program is typically evaluated by using the “incremental cost-effectiveness ratio,” defined by the equation

$$R = \frac{\Delta C}{\Delta E} = \frac{C_B - C_A}{E_B - E_A}$$

where C_A and C_B are the mean costs associated with treatments A and B and E_A and E_B are the mean health effects associated with these treatments. The incremental cost-effectiveness ratio R measures the additional cost per additional unit of health outcome for therapy B relative to therapy A.

Point estimates of R , denoted by $\hat{R} = \Delta\hat{C}/\Delta\hat{E}$ have typically been reported for clinical trials by using average between-treatment costs and effects. Recent reports in the literature have emphasized the need to provide 95% confidence interval estimates of R to account for the inherent variability of the estimates. Computing interval estimates of R is also important because the limited sample sizes used in clinical trials are usually not sufficient to detect differences in mean cost between therapies, or to get precise estimates of cost-effectiveness ratios.

Several methods have been proposed for estimating cost-effectiveness ratios using 95% confidence intervals. These include the use of a Taylor series approximation for the variance of $\hat{R}$ ^[46] and application of Fieller's theorem^[56–58] for estimating 95% confidence limits for the ratio of two random variables. A box method was proposed by Wakker and Klaassen^[59] and extended by Laska et al.^[58] This method provides a conservative upper-bound estimate of R by using the separate 95% confidence interval estimates of ΔC and ΔE . In addition, a bootstrap estimate of R was suggested by O'Brien et al.^[46] and further developed by Chaudhary and Stearns.^[56] The latter authors provide a nice overview of the Taylor series, Fieller's theorem, and the bootstrap method. They also give specific formulations to apply to each of those methods. Polsky et al.^[60] used simulation methods to compare the box method, Taylor series, Fieller's theorem, and bootstrap for this application. They conclude that the bootstrap and Fieller's theorem were more dependably accurate than the others because they allow for the asymmetric shape of the sampling of distribution of $\hat{R}$. Moreover, the box method does not account for the correlation between costs and effects.

Most of the difficulties encountered with the available methods for confidence interval estimates of cost-effectiveness ratios are related to situations in which ΔE is numerically small. The problem is heightened when the estimate of R is based on estimates of ΔE from clinical trials. Even if the null hypothesis that $\Delta E = 0$ is rejected, the treatment effects on the health outcome variable may be numerically small or modest. All of the available methods deal with the problem of small denominators in different ways. The accuracy of the Taylor series approximation depends on having small coefficients of variation for the separate estimates of between treatment differences in both costs and effects. The accuracy of the approximation begins to fail as the between-treatment differences approach 0. Laska et al.^[58] noted the complexity in interpreting confidence interval estimates obtained by use of Fieller's theorem because this method depends

on whether the individual confidence interval estimate of ΔE excludes 0. It is possible to have inclusion limits for R (R_L, R_U) if the confidence interval for ΔE excludes 0, and exclusion limits of the form “not (R_L, R_U)” if the confidence interval for ΔE includes 0. The condition that the entire 95% confidence interval estimate of the denominator ΔE not include 0 is equivalent to rejection of the null hypothesis that the treatments differ in mean effectiveness at the $\alpha = 0.05$ level. It is also possible to obtain confidence intervals for R that consist of the whole line $(-\infty, \infty)$ if the coefficients of variation for both ΔE and ΔC are large. Wakker and Klaassen^[59] suggest that an estimate of R is useful only when the data are sufficient to allow the conclusion that ΔE is not zero.

Cook and Heyse^[61] proposed a method for estimating R and its 95% confidence interval through an angular transformation. Fig. 1 shows the angular transformation done by using the cost-effectiveness plane.^[62] The slopes of the rays connecting points A and B to the origin are sometimes called the “average cost-effectiveness ratios.” The incremental cost-effectiveness ratio R is simply the slope of the line segment connecting the mean costs (C) and health outcomes (E) for treatments A and B. Redefining the origin of the cost-effectiveness plane to the point (E_A, C_A), θ is simply the angle connecting the $\Delta E > 0$ axis to the A–B line segment. Table 1 shows how θ can be used to facilitate the interpretation of cost-effectiveness ratios.

Values of θ in quadrants II or IV indicate that one of the treatments dominates the other in the sense that one treatment has a greater effect and lower cost. Values of θ in quadrants I or III indicate that the two treatments trade off costs and effects. Use of θ allows a specific interpretation of the relative direction of treatments A and B. The ratio R does not allow a specific interpretation. For example, the ratio cannot distinguish between quadrants II and IV when R is negative and cannot distinguish between quadrants I and III when R is positive. However, the interpretation of the cost-effectiveness of therapy B relative to A depends on the quadrant.

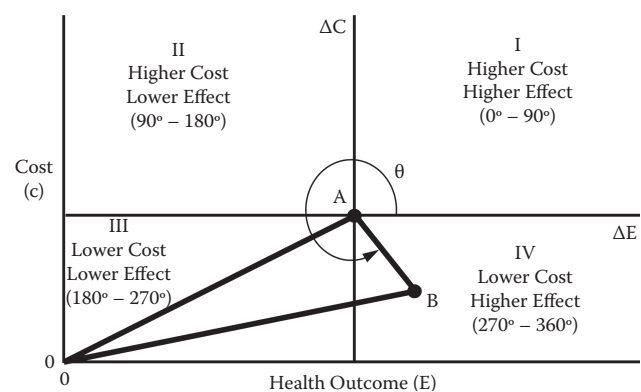


Fig. 1 Cost-effectiveness angle (θ): program B relative to program A

Table 1 Use of θ to Help Interpret Cost-Effectiveness Ratios

Direction of ΔC and ΔE	Range for angle	Quadrant
$\Delta C \geq 0; \Delta E \geq 0$	0–90°	I
$\Delta C \geq 0; \Delta E \leq 0$	90–180°	II
$\Delta C \leq 0; \Delta E \leq 0$	180–270°	III
$\Delta C \leq 0; \Delta E \geq 0$	270–360°	IV

For purposes of estimation, the method uses standardized differences in average cost and effect for the angular transformation. This is given by the equation

$$\theta = \begin{cases} \tan^{-1} \left[\Delta \hat{C} / s_{\Delta \hat{C}}, \Delta \hat{E} / s_{\Delta \hat{E}} \right] & \text{if } \Delta \hat{C} \geq 0 \\ 180 + \tan^{-1} \left[\Delta \hat{C} / s_{\Delta \hat{C}}, \Delta \hat{E} / s_{\Delta \hat{E}} \right] & \text{if } \Delta \hat{C} < 0 \end{cases} \quad (1)$$

where $s_{\Delta \hat{C}}$ and $s_{\Delta \hat{E}}$ are the standard errors of the between-group differences in mean costs and effects. The transformation θ stabilizes the variability of the ratio, especially when the effects are numerically small. The angle θ and its standard error can be estimated by using the jackknife^[63] or the bootstrap.^[64] The grouped jackknife method^[65] may offer some advantages over the bootstrap because it can easily adjust for multiple centers by doing the calculations leaving out one center at a time. Estimates of ratios can be computed by back-transforming from θ to R through Eq. 1.

Net Benefits

Stinnet and Mullahy^[66] proposed an alternate formulation for economic evaluations based on estimating the net monetary (or health) benefits associated with the candidate treatment. The approach is closely related to the incremental cost-effectiveness ratio by starting with the premise that a candidate program will be adopted if $R = (\Delta C / \Delta E)$ is less than a predefined threshold ratio λ . Net monetary benefits, as a function of λ , are expressed by

$$\text{NMB}(\lambda) = \lambda \Delta E - \Delta C$$

in units of cost, and net health benefits are defined equivalently by

$$\text{NHB}(\lambda) = \Delta E - (1 / \lambda) \Delta C$$

in units of health effects. With either formulation, a positive net benefit is supportive for adopting the candidate treatment program. The key to the net benefits approach is the ability to specify λ in order to translate health effects into costs, or vice versa. Many analysts will compute $\text{NMB}(\lambda)$ or $\text{NHB}(\lambda)$ for a broad range of threshold values in a sensitivity analysis. Note that $\text{NMB}(0) = -\Delta C$ provides a relationship with the cost-minimization analysis. Similarly, $\text{NMB}(R) = 0$ relates net monetary benefits and cost-effectiveness analysis.

Heitjan^[67] and Willan and Lin^[68] identified several advantages of the net benefit approach relative to the incremental cost-effectiveness ratio. The advantages are due to the ability to use sample means, variances, and covariances directly to estimate the model parameters of net benefits. This avoids the computational complexities inherent with ratio estimation. Additionally, individual values of NMB or NHB can be computed for each patient in a clinical trial, so that parametric or nonparametric statistical analysis can be readily applied. The net benefit analysis is different from a cost-benefit analysis because the former is derived from the cost-effectiveness ratio. The threshold cost-effectiveness ratio λ is considered the policy-defined trade-off between costs and health effects. In this sense, the net benefits method is also related to the willingness-to-pay method. The cost-benefit analysis involves a monetary valuation of the health effects.

Illustration: Scandinavian Simvastatin Survival Study

The Scandinavian Simvastatin Survival Study (4S) was a randomized, multinational, double-blind, placebo-controlled trial of simvastatin for hypercholesterolemic patients with existing heart disease.^[22] The study consisted of 4444 patients from Denmark ($n = 713$), Finland ($n = 868$), Iceland ($n = 157$), Norway ($n = 1025$), and Sweden ($n = 1681$) who previously had myocardial infarctions or stable angina pectoris. The median duration of patient follow-up was 5.4 years. The study showed that simvastatin was associated with a 30% reduction in the risk for death from any cause ($P = 0.0003$). Data on hospital admissions were prospectively collected to evaluate the effect of simvastatin on health care resource use. Pedersen et al.^[69] reported that patients randomly assigned to simvastatin had 34% fewer hospital days than patients assigned to placebo ($P < 0.0001$).

Other investigators have based formal cost-effectiveness analyses of cholesterol lowering with simvastatin on 4S.^[69,70] It is not our intention to reconsider these evaluations but rather to illustrate the statistical methods. We focus on the cost-effectiveness measure cost per death averted. Costs were assigned to simvastatin use and hospitalization by using Swedish prices converted to U.S. dollars. Simple summaries of the 4S results are given in Table 2. Simvastatin was associated with fewer patient deaths during the trial, but this positive health effect came at additional patient costs. The overall full-sample cost-effectiveness ratio was \$59,201 per additional death averted. Table 3 gives a comparison of several methods for estimating R with 95% confidence intervals. Most of the methods yield similar results because the health effect in this study was substantial. The Taylor series method appears to differ from the others because it provides symmetric limits, which are probably not appropriate for ratios of this type.

Table 2 Summary of Patient Costs and Survival in 4S

	Placebo	Simvastatin	Difference (S – P)
Sample size	2223	2221	
Proportion surviving to end of trial	0.885	0.918	0.033
Average cost/patient	\$3131	\$5098	\$1966

Table 3 Comparison of Methods to Estimate R and 95% CI

Method	Estimate of R	95% Confidence interval
<i>Full sample</i>		
Taylor series	\$59,201	(26,579, 91,824)
Fieller's	\$59,201	(37,369, 126,690)
<i>Bootstrap</i>		
Untransformed	\$63,992	(35,612, 126,819)
Using θ	\$58,146	(35,612, 126,819)
<i>Jackknife using θ</i>		
Grouping by center	\$59,364	(38,968, 116,596)
Stratifying by country	\$62,653	(36,723, 180,691)

R = incremental cost per death averted.

Specifying a threshold cost-effectiveness ratio of $\lambda = \$75,000$ per death avoided, the net monetary benefit is 524.7 with 95% confidence interval (– 826.3, 1875.7). The positive point estimate of the NMB indicates a favorable trade-off of net costs and net health effects. However, because the 95% confidence interval covers zero this estimated NMB did not attain a level of statistical significance at the $\alpha = 0.05$ level. Fig. 2 shows a graph of the NMB and the 95% confidence interval for a broad range of threshold values λ . The sample cost-effectiveness ratio

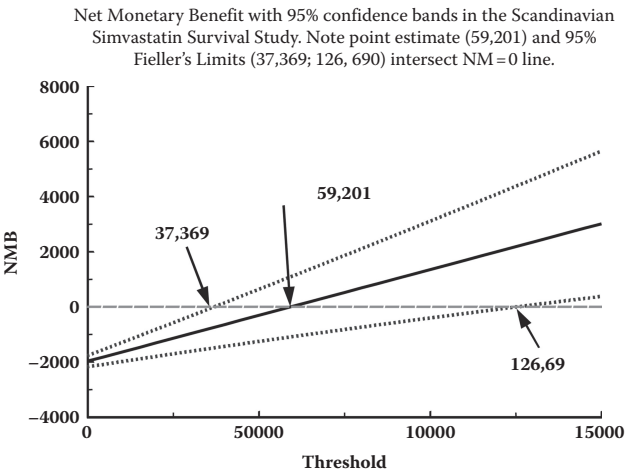


Fig. 2 Net monetary benefit (NMB) and 95% confidence intervals for a range of threshold values λ

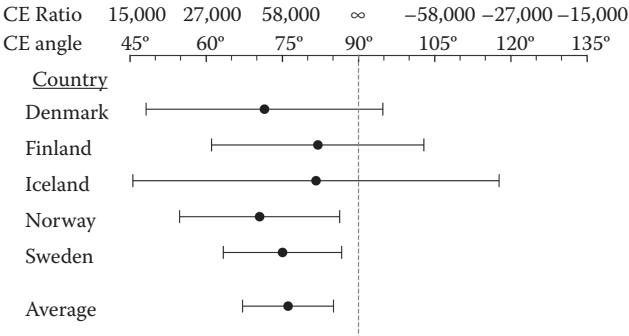


Fig. 3 Ninety-five percent confidence intervals for the incremental cost per additional survivor in the Scandinavian Simvastatin Survival Study by country and for the overall country average. CE = cost-effectiveness

can be picked off the graph as the value of λ where the NMB crosses the horizontal axis. Also note that the 95% confidence interval estimate of the cost-effectiveness ratio using Feiller's theorem can be derived as the values of λ where the upper and lower bound estimates of NMB cross the horizontal axis. This important relationship between the net benefits method and the cost-effectiveness ratio was derived by Heitjan.^[67]

Recall that 4S was a multinational study. Because of the potential for large variability among countries in the utilization and cost of health care resources, the appropriateness of combining economic data across the countries should be assessed. Cook et al.^[71] provided a test for homogeneity of the individual country ratios that can be used in this case. Fig. 3 shows the individual cost-effectiveness ratio estimates per country from 4S and the 95% confidence intervals. These ratios appear to be homogeneous ($P = 0.73$), justifying the appropriateness of the overall pooled ratio estimate given in Fig. 3 and Table 3.

INTERPRETING THE RESULTS

Drawing conclusions from the results of cost-effectiveness analyses can also be complex because doing so usually involves some evaluation of the trade-offs between costs and health outcomes. The exception would be therapies that are associated with both lower costs and greater health outcomes. These therapies are said to dominate the comparator therapy, and interpreting the results in these cases is straightforward when a cost-consequences analysis is used. However, most evaluations will need to compare programs that favor one on cost and the other on effectiveness. In these cases, the analyst needs to base a conclusion on the magnitude of the incremental cost-effectiveness ratio. One method proposed by Laupacis et al.^[72] for Canada was to classify technologies into one of five grades for recommendation on the basis of their incremental cost per QALY gained. The authors did acknowledge, however, that

cost-effectiveness should not be the sole basis on which to recommend the acceptance of a new therapy.

Another method often used in published studies compares the cost-effectiveness ratio for the therapy under study to ratios for a broad range of well-accepted medical therapies. For example, Johannesson et al.^[70] reported that the cost per life-year gained in patients with coronary heart disease who received simvastatin therapy ranged from \$3800 for 70-year-old men with total cholesterol levels of 309 mg/dl to \$27,400 for 35-year-old women with total cholesterol levels of 213 mg/dl. The authors concluded that the ratios were well within the range that was considered cost-effective in other studies. Rankings of the cost per life-year or cost per QALY gained for a broad range of therapies, referred to as “league tables,” have been published.^[73] League tables should be interpreted with caution because the methods used for the individual evaluations may be very different.

Bayesian Statistical Inference

It is widely recognized by economists and statisticians that simple point estimates of the mean costs and benefits are not adequate to describe the results of economic evaluations.^[74] Most of the available methods provide confidence interval estimates of the means, or ratios of the means, by incorporating estimates of variability into the evaluation. Other classic methods of statistical inference using formal hypothesis testing may have limitations when applied to studies with inadequate power to detect meaningful differences. Jones^[75] argued for the use of formal Bayesian methods that elicit prior beliefs by using expert opinion. Luce and Claxton^[76] recommended the use of decision-analytic models as essentially an informal application of Bayesian reasoning. Their proposal was to combine experimental and nonexperimental data with expert opinion as the basis for evaluating medical technologies.

Some evaluations are beginning to apply empirical Bayes methods to cost-effectiveness analysis. Jones^[77] recommended reporting the estimated probability distribution function for the economic variable of interest to enable the user to make a more informed interpretation about the uncertainty in the estimate. Other forms of analysis may simply estimate the probability that the cost-effectiveness ratio falls within each of the four quadrants of the cost-effectiveness plane. Heitjan et al.^[78] used simulated random draws from the estimated posterior distributions of cost and effectiveness parameters to tabulate the probabilities that the cost-effectiveness ratio falls within several important areas of the cost-effectiveness plane. One useful graphical display for summarizing the results of a cost-effectiveness analysis is the cost-effectiveness acceptability curve proposed by van Hout et al.^[79] The curve provides an estimate of the probability (plotted on the Y-axis) that the cost-effectiveness ratio for a candidate therapy is favorable with respect to a target ratio R (plotted on the X-axis). Their method assumes that costs and benefits are bivariate normal and uses point estimates of the variables to integrate over the region of the cost-effectiveness plane that gives ratios as favorable as R . A different computational method that does not make any distributional assumptions is a simple probability plot based on the number of bootstrap resampled cost-effectiveness ratios that are as favorable as R . This nonparametric method was applied to the 4S data, and the resulting cost-effectiveness acceptability curve is shown in Fig. 4. For example, Fig. 4 could be used to show that there is greater than a 90% probability that simvastatin therapy is associated with a cost-effectiveness ratio at least as favorable as \$100,000 per death avoided. Alternatively, there is greater than a 90% probability that the net monetary benefits are positive for a threshold of \$110,000 per death avoided. Results of this type are of interest to decision makers who may establish policy on the basis of the magnitude of estimated cost-effectiveness ratios.

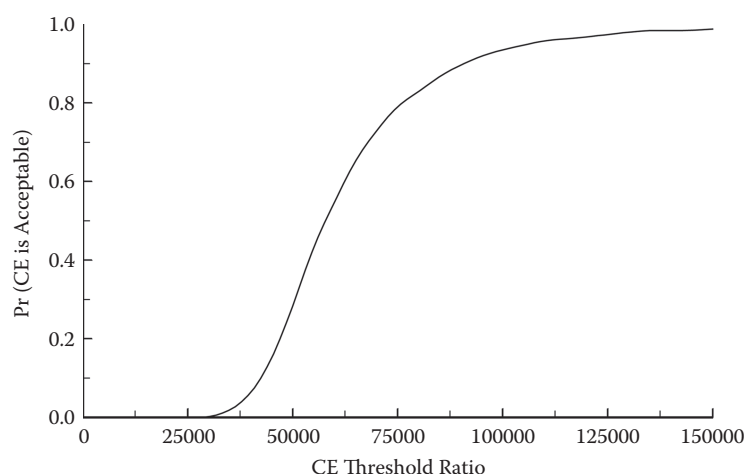


Fig. 4 Cost-effectiveness (CE) acceptability curve in the Scandinavian Simvastatin Survival Study. The graph plots the probability that for simvastatin, the cost per additional survival would be deemed acceptable for a range of cost-effectiveness ratio thresholds

CONCLUSION

The concern over the costs of health care will continue to grow as more novel health care technologies become available to a growing elderly population. In this environment, health care providers will attempt to identify and apply treatment programs that offer the greatest benefit for the fixed level of investment in health care. Formal economic evaluations will be looked on to help guide these decisions. Pharmaceutical companies will need to significantly enhance their clinical development programs to address the needs of the changing marketplace. More emphasis will be placed on studies of patient outcomes, including quality of life and cost-effectiveness, much earlier in the product development program. Moreover, as guidelines and formal requirements for cost-effectiveness analysis are put in place, there will be a growing need to assemble specific documentation of the socioeconomic impact of the candidate product. This information would also be helpful in marketing new products to sophisticated managed care organizations. Drummond and McGuire^[45] provided a contemporary textbook that describes the available methods to integrate utilization, cost, and outcome data within an appropriate theory of economic evaluation.

Statisticians working in the pharmaceutical industry will need to address the complexities that these changes will bring to the development program. There will be a shift away from formal hypothesis testing to methods of estimation, including the use of statistical models to help identify the best uses of the candidate therapies. There will also be opportunities to work on data sources outside of the clinical program to provide the most comprehensive evaluation for the product.

REFERENCES

1. Commonwealth of Australia. *Guidelines for the Pharmaceutical Industry on Preparation of Submissions to the Pharmaceutical Benefits Advisory Committee: Including Economic Analyses*; Department of Health and Community Services: Canberra, Australia, 1995.
2. Ontario Ministry of Health. *Ontario Guidelines for Economic Analysis of Pharmaceutical Products*; Ministry of Health: Toronto, Canada, 1994.
3. Harris, A.; Buxton, M.F.; O'Brien, B.J.; Rutten, F.F.H.; Drummond, M.F. Using economic evidence in reimbursement decisions for health technologies: Experience of 4 countries. *Expert Rev. Pharmacoecon. Outcomes Res.* **2001**, *1* (1), 7–12.
4. Food and Drug Administration. *Comparing Treatments: Safety, Effectiveness and Cost-Effectiveness*; Masur Auditorium, National Institutes of Health: Bethesda, MD, March, 23, 24, 1995.
5. Kassirer, J.P.; Angell, M. EDITORIAL: The Journal's policy on cost-effectiveness analysis. *N. Engl. J. Med.* **1994**, *331*, 669–670.
6. Drummond, M.F.; Jefferson, T.O. On behalf of the BMJ economic evaluation working party, guidelines for authors and peer reviewers of economic submissions to the *BMJ*. *Br. Med. J.* **1996**, *313*, 275–283.
7. *Cost-Effectiveness in Health and Medicine*; Gold, M.R., Siegel, J.E., Russell, L.B., Weinstein, M.C., Eds.; Oxford University Press: New York, 1996.
8. Siegel, J.E.; Torrance, G.W.; Russell, L.B.; Luce, B.R.; Weinstein, M.C.; Gold, M.R.; Members of the Panel on Cost-effectiveness in Health and Medicine Guidelines for pharmacoeconomic studies, recommendations from the panel on cost effectiveness in health and medicine. *Pharmacoeconomics* **1997**, *11*, 159–168.
9. Hillman, A.L.; Eisenberg, J.M.; Pauly, M.V.; Bloom, B.S.; Glick, H.; et al. Avoiding bias in the conduct and reporting of cost-effectiveness research sponsored by pharmaceutical companies. *N. Engl. J. Med.* **1991**, *324*, 1362–1365.
10. Sullivan, S.D.; Lyles, A.; Luce, B.; Gricar, J. AMCP guidance for submission of clinical and economic evaluation data to support formulary listing in the United States Health Plans and pharmacy benefit management organizations. *J. Manag. Care Pharm.* **2002**, *7* (4), 272–282.
11. Weinstein, M.C.; Stason, W.B. Foundations of cost-effectiveness analysis for health and medical practices. *N. Engl. Med.* **1977**, *296*, 716–721.
12. Garber, A.M.; Phelps, C.E. Economic foundations of cost-effectiveness analysis. *J. Health Econ.* **1997**, *16*, 1–31.
13. Drummond, M.F. Cost-of-illness studies, a major headache? *Pharmacoeconomics* **1992**, *2*, 1–4.
14. Drummond, M.F.; O'Brien, B.J.; Stoddart, G.L.; Torrance, G.W. *Methods for Economic Evaluation of Health Care Programmes*, 2nd Ed.; Oxford University Press: Oxford, 1997.
15. de Lissovoy, G.; Elixhauser, A.; Luce, B.R.; Weschler, J.; Mowery, P.; Reblando, J.; Solomkin, J. Cost analysis of imipenem–cilastatin versus clindamycin with tobramycin in the treatment of acute intra-abdominal infection. *Pharmacoeconomics* **1993**, *4*, 203–214.
16. Oster, G.; Epstein, A.M. Cost-effectiveness of antihyperlipidemic therapy in the prevention of coronary heart disease. *J. Am. Med. Assoc.* **1987**, *258*, 2381–2387.
17. Johannesson, M.; Jonsson, B.; Karlsson, G. Outcome measurement in economic evaluation. *Health Econ.* **1996**, *5*, 279–296.
18. O'Brien, B.J.; Goeree, R.; Gafni, A.; Torrance, G.W.; et al. Assessing the value of a new pharmaceutical: A feasibility study of contingent valuation in managed care. *Med. Care* **1998**, *36* (3), 370–384.
19. Luce, B.R.; Manning, W.G.; Siegel, J.E.; Lipscomb, J. Chapter Six: Estimating Costs in Cost-Effectiveness Analysis. In *Cost-effectiveness in Health and Medicine*; Gold, M.R.; Siegel, J.E.; Russell, L.B.; Weinstein, M.C., Eds.; Oxford University Press: New York, 1996. 176–213.
20. Koopmanschap, M.A.; Rutten, F.F.H. Indirect costs in economic studies, confronting the confusion. *Pharmacoeconomics* **1993**, *4*, 446–454.
21. Gould, A.L.; Rossouw, J.E.; Santanello, N.C.; Heyse, J.F.; Furberg, C.D. Cholesterol reduction yields clinical benefit: A new look at old data. *Circulation* **1995**, *91*, 2274–2282.
22. Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with

- coronary heart disease: The Scandinavian Simvastatin Survival Study (4S). *Lancet* **1994**, 344, 1383–1389.
23. Shepherd, J.; Cobbe, S.M.; Ford, I.; Isles, C.G.; Lorimer, A.R.; MacFarlane, P.W.; McKillop, J.H.; Packard, C.J. Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. *N. Engl. J. Med.* **1995**, 20, 1301–1307.
 24. Schulman, K.A.; Kinoshian, B.; Jacobson, T.A.; Glick, H.; Willian, M.K.; Koffer, H.; Eisenberg, J.M. Reducing high blood cholesterol level with drugs. *J. Am. Med. Assoc.* **1990**, 264, 3025–3033.
 25. Torrance, G.W.; Feeny, D. Utilities and quality adjusted life years. *Int. J. Technol. Assess. Health Care* **1989**, 5, 559–575.
 26. Mehrez, A.; Gafni, A. Quality-adjusted life years, utility theory, and healthy-years equivalents. *Med. Decis. Mak.* **1989**, 9, 142–149.
 27. Cox, D.R.; Fitzpatrick, R.; Fletcher, A.E.; Gore, S.M.; Spiegelhalter, D.J.; Jones, D.R. Quality of life assessment: Can we keep it simple? *J. R. Stat. Soc., Ser. A* **1992**, 155, 353–393.
 28. Drummond, M.F.; Heyse, J.F.; Cook, J.R.; McGuire, A. Selection of endpoints in economic evaluations of coronary heart disease interventions. *Med. Decis. Mak.* **1993**, 13, 184–190.
 29. Van Hout, B.A. Discounting costs and effects: A reconsideration. *Health Econ.* **1998**, 7 (7), 581–594.
 30. Goldman, L.; Weinstein, M.C.; Goldman, P.A.; Williams, L.W. Cost-effectiveness of HMG-CoA reductase inhibition for primary and secondary prevention of coronary heart disease. *J. Am. Med. Assoc.* **1991**, 265, 1145–1151.
 31. Glick, H.; Heyse, J.F.; Thompson, D.; Epstein, R.S.; Smith, M.E.; Oster, G. A model for evaluating the cost-effectiveness of cholesterol-lowering treatment. *Int. J. Technol. Assess. Health Care* **1992**, 8, 719–734.
 32. Drummond, M.F.; Davies, L.M. Economic analysis alongside clinical trials: Revisiting the methodological principles. *Int. J. Technol. Assess. Health Care* **1991**, 7, 561–573.
 33. Powe, N.R.; Griffiths, R.I. The clinical-economic trial: Promise, problems, and challenges. *Control. Clin. Trials* **1995**, 16, 377–394.
 34. Schulman, K.A.; Llana, T.; Yabroff, K.R. Economic assessment within the clinical development program. *Med. Care* **1996**, 34, DS89–DS95.
 35. Mauskopf, J.; Schulman, K.; Bell, L.; Glick, H. A strategy for collecting pharmacoeconomic data during phase II/III clinical trials. *Pharmacoeconomics* **1996**, 3, 264–277.
 36. Drummond, M.F.; O'Brien, B.J. Clinical importance, statistical significance and the assessment of economic and quality of life outcomes. *Health Econ.* **1993**, 2, 205–212.
 37. O'Brien, B.J.; Drummond, M.F. Statistical versus quantitative significance in the socioeconomic evaluation of medicines. *Pharmacoeconomics* **1994**, 5, 389–398.
 38. Schulman, K.A.; Buxton, M.; Glick, H.; Sculpher, M.; Guzman, G.; Kong, J.; Backhouse, M.; Mauskopf, J.; Bell, L.; Eisenberg, J.M.; FIRST Investigators. Results of the economic evaluation of the FIRST study, a multinational prospective economic evaluation. *Int. J. Technol. Assess. Health Care* **1996**, 4, 698–713.
 39. Freemantle, N.; Drummond, M.F. Should clinical trials with concurrent economic analyses be blinded? *J. Am. Med. Assoc.* **1997**, 277, 63, 64.
 40. Oster, G.; Borok, G.M.; Menzin, J.; Heyse, J.F.; Epstein, R.S.; Quinn, V.; Benson, V.; Dudl, R.J.; Epstein, A. A randomized trial to assess effectiveness and cost in clinical practice: Rationale and design of the Cholesterol Reduction Intervention Study (CRIS). *Control. Clin. Trials* **1995**, 16, 3–16.
 41. Mark, D.B.; Hlatky, M.A.; Califf, R.M.; Naylor, C.D.; Lee, K.L.; Armstrong, P.W.; Barbash, G.; White, H.; Simoons, M.L.; Nelson, C.L.; Clapp-Channing, N.; Knight, D.; Harrell, F.E.; Simes, J.; Topol, E.J. Cost effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction. *N. Engl. J. Med.* **1995**, 332, 1418–1424.
 42. Diabetes Control and Complications Trial Research Group. Lifetime benefits and costs of intensive therapy as practiced in the Diabetes Control and Complications Trial. *J. Am. Med. Assoc.* **1996**, 276, 1409–1415.
 43. Mullahy, J.; Manning, W. Chapter Eight: Statistical Issues in Cost-Effectiveness Analyses. In *Valuing Healthcare Costs, Benefits, and Effectiveness of Pharmaceuticals and Other Medical Technologies*; Sloan, F., Ed.; Cambridge University Press: New York, 1994, 149–241.
 44. Manning, W.G.; Fryback, D.G.; Weinstein, M.C. Chapter Eight: Reflecting Uncertainty in Cost-Effectiveness Analysis. In *Cost-Effectiveness in Health and Medicine*; Gold, M.R. Siegel, J.E., Russell, L.B., Weinstein, M.C., Eds.; Oxford University Press: New York, 1996, 247–275.
 45. Drummond, M.F.; McGuire, A. *Economic Evaluation in Health Care: Merging Theory with Practice*; Oxford University Press: Oxford, 2002.
 46. O'Brien, B.J.; Drummond, M.F.; LaBelle, R.J.; William, A. In search of power and significance: Issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Med. Care* **1994**, 32, 150–163.
 47. Briggs, A.; Schulpher, M.; Buxton, M. Uncertainty in the economic evaluation of health care technologies: The role of sensitivity analysis. *Health Econ.* **1994**, 3, 95–104.
 48. Heyse, J.F.; Cook, J.R.; Carides, G.W. *Statistical Considerations in Analysing Health Care Resource Utilization and Cost Data. Economic Evaluation in Health Care*; Drummond, M.; McGuire, A., Eds.; Oxford University Press: Oxford, 2001. 215–235, Chapter 9.
 49. Dudley, R.A.; Harrell, F.E.; Smith, L.R.; Mark, D.B.; Califf, R.M.; Pryor, D.B.; Glower, D.; Lipscomb, J.; Hlatky, M. Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery. *J. Clin. Epidemiol.* **1993**, 46, 261–271.
 50. Fenn, P.; McGuire, A.; Phillips, V.; Backhouse, M.; Jones, D. The analysis of censored treatment cost data in economic evaluation. *Med. Care* **1995**, 33, 851–863.
 51. Carides, G.W.; Heyse, J.F. Nonparametric Estimation of the Parameters of Cost Distributions in the Presence of Right-Censoring. In *1996 Proceedings of the Biopharmaceutical Section, American Statistical Association Annual Meetings*; 1996, 186–191.
 52. Lin, D.Y.; Fewer, E.J.; Etzioni, R.; Wax, Y. Estimating medical costs from incomplete follow-up data. *Biometrics* **1997**, 53, 113–128.

53. Carides, G.W.; Heyse, J.F.; Iglewicz, B. A regression-based method for estimating mean treatment cost in the presence of right-censoring. *Biostatistics* **2000**, *1*, 299–313.
54. Zhao, H.; Tsiatis, A.A. A consistent estimator for the quality-adjusted survival time. *Biometrika* **1997**, *84*, 339–348.
55. Bang, H.; Tsiatis, A.A. Estimating medical costs with censored data. *Biometrika* **2000**, *87*, 329–343.
56. Chaudhary, M.A.; Stearns, S.C. Estimating confidence intervals for cost effectiveness ratios: An example from a randomized trial. *Stat. Med.* **1996**, *15*, 1447–1458.
57. Willan, A.R.; O'Brien, B.J. Confidence intervals for cost-effectiveness ratios: An application of Fieller's theorem. *Health Econ.* **1996**, *5*, 297–305.
58. Laska, E.M.; Meisner, M.; Siegel, C. Statistical inference for cost-effectiveness ratios. *Health Econ.* **1997**, *6*, 229–242.
59. Wakker, P.; Klaassen, M.P. Confidence intervals for cost/effectiveness ratios. *Health Econ.* **1995**, *4*, 373–381.
60. Polsky, D.; Glick, H.A.; Willke, R.; Schulman, R. Confidence intervals for cost-effectiveness ratios: A comparison of four methods. *Health Econ.* **1997**, *6*, 243–252.
61. Cook, J.R.; Heyse, J.F. Use of an angular transformation for ratio estimation in cost-effectiveness analysis. *Stat. Med.* **2000**, *19*, 2989–3003.
62. Black, W.C. The CE Plane: A graphic representation of cost-effectiveness. *Med. Decis. Mak.* **1990**, *10*, 212–214.
63. Miller, R.G. The jackknife—A review. *Biometrika* **1974**, *61*, 1–15.
64. Efron, B.; Gong, G. A leisurely look at the bootstrap, jackknife and cross-validation. *Am. Stat.* **1983**, *37*, 36–48.
65. Mosteller, F.; Tukey, J.W. *Data Analysis and Regression: A Second Course in Statistics*; Addison-Wesley: Reading, MA, 1977.
66. Stinnett, A.A.; Mullahy, J. Net health benefits: A new framework for the analysis of uncertainty in cost-effectiveness analysis. *Med. Decis. Mak.* **1998**, *18*, S68–S80, (Supplement).
67. Heitjan, D.F. Feiller's method and net health benefits. *Health Econ.* **2000**, *9*, 327–335.
68. Willan, A.R.; Lin, D.Y. Incremental net benefit in randomized clinical trials. *Stat. Med.* **2001**, *20*, 1563–1574.
69. Pedersen, T.R.; Kjekshus, J.; Berg, K.; Olsson, A.G.; Wilhelmsen, L.; Wedel, H.; Pyorala, K.; Miettinen, T.; Haghefelt, T.; Faergeman, O.; Thorgeirsson, G.; Jonsson, B.; Schwartz, J.S.; Scandinavian Simvastatin Survival Study Group. Cholesterol lowering and the use of healthcare resources, results of the Scandinavian Simvastatin Survival Study. *Circulation* **1996**, *93*, 1796–1802.
70. Johannesson, M.; Jonsson, B.; Kjekshus, J.; Olsson, A.G.; Pedersen, T.R.; Wedel, H.; Scandinavian Simvastatin Survival Study Group. Cost-effectiveness of simvastatin treatment to lower cholesterol levels in patients with coronary heart disease. *N. Engl. J. Med.* **1997**, *336*, 332–336.
71. Cook, J.R.; Drummond, M.; Glick, H.; Heyse, J.F. Assessing the Appropriateness of Combining Data from Multi-national Clinical Trials. In *Technical Report, Clinical Biostatistics and Research Data Systems*; Merck Research Laboratories: New York, 2001.
72. Laupacis, A.; Feeny, D.; Detsky, A.S.; et al. How attractive does a new technology have to be to warrant adoption and utilization: Tentative guidelines for using clinical and economic evaluations. *Can. Med. Assoc. J.* **1992**, *146*, 473–481.
73. Drummond, M.F.; Torrance, G.W.; Mason, J. Cost-effectiveness league tables: More harm than good? *Soc. Sci. Med.* **1993**, *37*, 33–40.
74. Briggs, A.H. A Bayesian approach to stochastic cost-effectiveness analysis. *Health Econ.* **1999**, *8*, 257–261.
75. Jones, D.A. A Bayesian Approach to the Economic Evaluation of Health Care Technologies. In *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd Ed.; Spilker, B., Ed.; Lippincott-Raven Publishers: Philadelphia, PA, 1996; 1189–1196.
76. Luce, B.R.; Claxton, K. Redefining the analytic approach to pharmacoeconomics. *Health Econ.* **1999**, *8*, 187–189.
77. Jones, D.A. The role of probability distributions in economic evaluations. *Br. J. Med. Econ.* **1995**, *8*, 137–146.
78. Heitjan, D.F.; Moskowitz, A.J.; Whang, W. Bayesian estimation of cost-effectiveness ratios from clinical trials. *Health Econ.* **1999**, *8*, 191–201.
79. Van Hout, B.A.; Al, M.J.; Gordon, G.S.; Rutten, F.F.H. Costs, effects and C/E-ratios alongside a clinical trial. *Econ. Eval.* **1994**, *3*, 309–319.

Pharmacoeconomics: Cost-Effectiveness Analysis

Shahram Heshmat

Department of Public Health, University of Illinois at Springfield, Springfield, Illinois, U.S.A.

Abstract

The purpose of this entry is to discuss the concept of cost-effective analysis for the evaluation of health services, and to familiarize readers with how to interpret CEA results.

Keywords: Cost-effectiveness; Health services; Resource allocation.

INTRODUCTION

As resources have become scarcer and public accountability has become greater, health care providers have been forced to reexamine the costs and the benefits of its activities as a step in enhancing their abilities to allocate resources more effectively. Cost-effectiveness analysis (CEA) provides health professionals with basic tools to ensure appropriate resource allocation among competing demands. The purpose of this article is to discuss the concept of CEA for the evaluation of health services, and to familiarize readers with how to interpret CEA results.

TYPES OF ECONOMIC EVALUATION

Economic evaluation provides health professionals with basic tools to ensure appropriate resource allocation among competing demands. Economic evaluation is a subfield within economics that has explored the question of how to maximize consumer welfare within a limited budget. Given the scarcity of resources, every decision to use public health resources requires to examine the costs and the benefits of those decisions. From an economic perspective, an efficient allocation of society's limited resources requires the benefits of an intervention to exceed the costs. This means the valuation people place on those programs is greater than the cost of producing them. The basic task of economic evaluation involves a comparison of the costs and benefits of adopting program A rather program B (or simply status quo).

Economic evaluation includes cost-minimization, cost-effectiveness analysis (CEA), cost-utility analysis (CUA), and cost-benefit analysis (CBA) studies. There is substantial overlap between these four types of cost evaluation. Each characterizes costs in the same terms, but differs with respect to outcome measures (the benefits of health intervention).

In cost-minimization analysis, the effectiveness of the interventions is assumed or has been shown to be the same, and only the cost difference in the interventions is determined. A cost-minimization analysis address the question of which alternative is least cost given that they are equally effective.

Cost-effectiveness and cost-benefit analyses are formal methods for comparing the benefits and costs of a medical intervention in order to determine whether it is worth doing. The only difference between the two is that CEA measures benefits in terms of some standard of clinical outcome or effectiveness, such as mortality rates, years of added life, or quality-adjusted life years. CEA usually is based on a one-dimensional measure of health outcome (e.g., reductions in blood pressure or premature births averted). In CBA, all benefits are converted into monetary values which allow compare health-related programs with other types of programs or policies. The underlying goal of CEA/CBA analyses is to find which alternative provides maximum aggregate health benefits for a given level of resources or, equivalently, which alternative provides a given level of health benefits at the lowest cost. CEA is most suitable when comparing interventions that have similar health outcomes.

Cost-utility analysis (CUA) is a subset of CEA studies. Cost-effectiveness and cost-utility analysis are identical except that the effectiveness measure used in a cost-utility analysis is one that reflects societal or individual preferences for the outcome. Cost-utility analysis (CUA) is a broader form of analysis than cost-effectiveness analysis. CUA seeks to collapse multi-dimensional consequences of health status (i.e., degrees of disability) into a one-dimensional metric (e.g., Quality-Adjusted Life Years). The consequences of programs are measured in time units adjusted by health utility weights. That is, states of health associated with the outcomes are valued relative to one another. In general terms this means that one can assess the quality of (for example) life-years gained, not just the crude number

of years. This approach is particularly useful for those health treatments that extend life only at the expense of side effects or that produce reductions in morbidity rather than mortality. Gold et al.^[1] do not distinguish between cost-utility analysis and cost-effectiveness analysis.

Another type of economic evaluation study is the cost-of-illness (COI) study. The goal of COI study is to estimate the total societal costs of caring for persons with an illness compared with persons without the illness. The result of a cost-of-illness study can be used as input to a formal cost-effectiveness analysis.

WHAT IS COST-EFFECTIVENESS ANALYSIS?

The basic tasks of CEA are to identify, measure, value, and compare the costs and consequences of the alternatives being considered.^[2] The central purpose of CEA is to compare the relative values of different interventions in creating better health and longer life. The results of such evaluations are typically summarized in a cost-effectiveness ratio (*C/E*), where the denominator reflects the gain in health from a candidate intervention (measured, e.g., in terms of years of life gained, premature births averted) and the numerator reflects the cost of obtaining that health gains. The cost-effectiveness ratio [dollars per life-year (LY)] serves as a yardstick for measuring the relative priority of health interventions that compete for limited resources. The underlying rationale for CEA is to try and maximize the health effects for a given amount of resources.^[2]

In CEA, the consequences of programs are measured in the most appropriate natural or physical units, such as “years of life gained” or “cases correctly diagnosed.” Health outcomes can range from intermediate outcomes, such as millimeters-of-mercury blood pressure reduction or disability days averted, to more distal outcomes such as lives saved, life-years gained, or quality-adjusted life-years (QALYs) gained, or “cases correctly diagnosed.” A QALY is the equivalent of one year of life in full health. To estimate the relative cost-effectiveness of various services, CEA requires some common unit of measurement to help compare programs. QALY has been used historically in studies to assess the relative value of different interventions, with each intervention carrying a “price tag” or a rough estimate of the cost to save a comparable year of life. For example, childhood immunizations and smoking cessation cost less than \$5,000 per QALY gained. The amount here can be considered as a purchasing price for one QALY. An intervention is “cost-saving” if it reduces costs while improving health.

The QALYs (pronounced “qualy”), or analogous measure, is the most comprehensive measure of outcome used in CEA, incorporating both quality and survival information.^[1] The purpose of QALYs is to measure in a single figure the impact of an intervention that increases both

quantity and quality of life. This measure assumes that different states of illness can be assessed in terms of utility patients provide relative to perfect health over a period of one year.

Quality weights can be elicited from professionals with expertise in the condition and intervention, from patients with the condition, or from the general public.^[3] The concept of the quality-adjusted life-year is explicit recognition that there are states of health for which people are willing to take a measurable risk of a bad outcome (usually death) to avoid.^[4] A common method to elicit people’s utility is to describe medical condition (e.g., being a diabetic) and ask them to choose between ten years in that condition or some smaller number of years without it. If they respond that they would choose five years of perfect health life over ten years with diabetes, then they are assessing life with the illness as half as good as diabetic life. A simpler approach that avoids assigning utility values to ill health is to look only at the life years (LYs) saved by an interventions. Evidence suggests that in many policy interventions QALYs and LYs provide similar rankings.

The quality weights are usually measured in a scale that extends from 0 (death) to 1 (perfect health). An intervention that extends a person’s life for one year and provides perfect health generates 1 QALY. In contrast, an intervention that extends a person’s life for one year but at a diminished quality of life, with a value of 50%, produces half as much benefit, or 0.5 QALY. An intervention that does not extend a person’s life but improves quality of life from 0.50 to 0.75 produces 0.25 QALY per year. Consider, for example, an illness such as diabetes that reduces one’s quality of life by 40%, so the person’s QALY over the year would be 0.60. Thus, an intervention that would prevent such an illness would yield 0.40 QALYs. If the intervention saved a million people from having the diabetes every year this would amount to 400,000 QALYs per year. If the intervention cost \$100 million, and an alternative investment of \$100 million into, say, safer airline travel saved 500 QALYs per year, then the current illness prevention can be said to be more cost-effective.

CEA incorporate preferences for receiving health-care benefits that occur now rather than later. People value a year of good health now more highly than a year of good health in ten year’s time, so a policy to reduce an illness today might be valued more than an intervention to reduce, say, risk of heart disease, in the distant future. Thus, allowance needs to be made for the differential timing of costs and consequences. Economists call this the notion of time preference. Discounting is more controversial, however, when it is applied to health effects. The assumption implicit in discounting implies that health effects gained in the future are less valuable than health effects gained today. For example, is it less valuable to avert a heart attack ten years from now than a heart attack next year, and do they have the same impact on health-related quality of life and on life expectancy? Economists who work on CEA

have long accepted that health effects should be discounted in the same way that the dollar expenditures are and that the same discount rate should be used. Something like 3% is typical for government projects. If the interest rate is 5%, for example, the present value of \$1 to be obtained in one year is about \$.97—if you invest \$.97 today, you will have \$1 in one year, so \$.97 accurately reflects the present value if \$1 in one year. With a higher interest rate, the present value is lower because the opportunity costs involved in waiting to get the dollar are greater.

In CEA, the added costs and health outcomes associated with a program are used to calculate the incremental cost-effectiveness ratio relative to some comparator. No attempt is made to value the consequences, so it is implicitly assumed that the output concerned is in some sense “worth having.” Although in CEA, the economic valuation of health consequences is avoided, the problem of deciding in absolute terms whether a program is worth funding rests on whether a particular value of the cost-effectiveness ratio (e.g., \$50,000 per QALY) is acceptable.

Valuing human life is a necessary part of any economic evaluation. The most common approach is to calculate the value of a generic human life. The value is revealed from individual behavior. It is the answer to the question of “how much are you willing to pay to reduce the risk to your life?” People make risk-money tradeoff decisions on a daily basis. For example, how much people are willing to pay for airbags in cars and other safety features? Suppose that there is a 1 in 100,000 chance that an airbag in your car will save your life, and that you are willing to pay \$50 for an airbag to reduce the risk by 1 in 100,000. This implies that the airbag (intervention) would save, statistically, one life. This implies that the value of the one life saved is $50 \times 100,000 = \$5$ million (C/E or $\$50/0.00001$).

The incorporation of measures of quality of life into decision making about allocation of resources is a cornerstone of the outcome movement. In theory, CEA enables decision-makers to compare how much benefit patients are receiving from seemingly incomparable health services.^[2] Decision-makers may be federal, state, or local. They may be in the private sector or in the public sector. If a moderately expensive treatment brings benefits, CEA can show whether it is more cost-effective than a slightly expensive treatment that brings modest benefits. It makes allocation tradeoffs explicit; for example, if a health insurance company reimburses for Pap smears every three years instead of yearly, CEA can estimate how much money it will save and how many lives of enrollees will be lost. Cost-effectiveness analysis teaches us that after we spend money to achieve important benefits, further benefits often come at greater and greater financial cost. For example, an early CEA showed that an inexpensive screening test for colon cancer, if repeated for a sixth time in a patient whose first five tests were normal, continued to find colon cancer, but at a cost of \$26 million for every year of life gained.^[5]

KEY CONCEPTS IN PERFORMING COST-EFFECTIVENESS ANALYSIS

Cost-effectiveness analysis depends on the perspective of the decision-maker. Specifically, a program or intervention may be deemed to be cost-effective from the standpoint of the individual patient or from the standpoint of a specific inpatient service. However, the same program or intervention may not be cost-effective when viewed and assessed from the overall societal point of view. Analyses based on different perspectives cannot be compared because, reflecting the interests associated with each perspective, they count different health effects and resources and value them differently. Most published studies in the medical literature tend to evaluate cost-effectiveness from the perspective of society as a whole and thereby generally define cost-effectiveness in terms of dollars or resource spent per year of life saved. The appropriate perspective depends on the objective of the study. It should be noted that from the perspective of the comprehensive societal viewpoint, all costs and effects are considered no matter who pays the costs or who receives the effects. Such comprehensive information is necessary to support decision-makers who want to consider the impact of their decisions on all affected parties.

In order to compare the desirability of programs, the CEA ratios must be based on comparable definitions of cost and outcome. A problem in comparing programs' costs is that studies may differ in the types of cost they include in the numerator of the CEA ratio. For example, the value of time that patients spend in obtaining care or the value of the reduction in work loss because of health improvement is often excluded from CEA. Inclusion/exclusion may have a substantial effect on the cost-effectiveness ratio.

The basic approach is to measure all relevant costs and benefits and to determine the ratio between the two. All other things being equal, an alternative with a lowest cost-benefit ratio (whether benefits are measured in terms of dollars or some other metric) is preferable to an alternative with a higher cost-benefit ratio. However, in the case of mutually exclusive interventions, for example, surgery or drugs but never both, the issue is whether the additional improvement in benefits is worth the additional cost. In such cases, incremental cost-effectiveness or cost-benefit analysis is required.

The panel on CEA in Health and Medicine of the U.S. Public Service^[2] emphasized the importance of conduct of a “reference case” analysis. The reference case analysis is an analysis that is performed according to the standard set of rules. The reference case analysis allows comparison with other cost-effectiveness analyses performed using the same set of rules. This greatly enhances the usefulness of information to policymakers. In the reference case, analysis is performed from the societal perspective.

ALTERNATIVES

Cost-effectiveness analysis requires the specification of two or more alternatives (or programs). By definition, assessment of CEA is a relative phenomenon, and thus the calculation of a single cost-effectiveness measure, that is, cost-effectiveness ratio, provides little, if any, quantitative insight regarding joint clinical and economic impact. For example, if a hospital is considering an expansion of its intensive care facilities, it may perform a study that assesses the cost-effectiveness of two viable alternatives—the expansion of the current intensive care unit (ICU) from 10 to 15 beds or the construction of a new 10-bed intermediate care.

The choice of alternatives (comparators) can make an enormous difference in the cost-effectiveness of an intervention because the elements in a cost-effectiveness ratio are calculated as differences between two alternatives. The numerator of the ratio, net costs, is the difference in costs between the intervention and the comparator. The denominator, net health effects, is the difference in health effects between the two. For example, when annual Pap smears are compared to no screening, they cost about \$40,000 per life-year saved (1985 dollars). But compared with the more realistic alternative of screening every two years, the cost per life-year is more than \$1 million.^[6]

Alternatives are either independent or mutually exclusive programs. Two programs, A and B, are defined as independent if the costs and effectiveness of program A (B) are not affected by whether program B (A) is implemented or not. The two programs are viewed as applying to two different populations, for example, the treatment of ulcer patients and the treatment of arthritis patients.

Cost-effectiveness analysis may include mutually exclusive programs with different applications of the same medical intervention or drug, such as drug regimens, or different levels of a program, such as drug dosage. For example, cholesterol reduction in persons with different cholesterol levels could be regarded as different programs that compete for resources, and cost-effectiveness would be different in each group. In the case of competing mutually exclusive alternatives for the same patients, the appropriate analysis would be based on incremental cost-effectiveness analysis.

AVERAGE RATIO VERSUS INCREMENTAL RATIO

An average cost-effectiveness ratio and an incremental or marginal cost-effectiveness ratio are different concepts.^[7] An average cost-effectiveness ratio is estimated by dividing the cost of the intervention by the effectiveness of the intervention compared to “doing nothing.” An incremental or marginal cost-effectiveness ratio is an

estimate of the cost per unit of effectiveness of switching from one intervention to another. If only one program is compared with doing nothing, then the incremental cost-effectiveness ratio will be the same as the average cost-effectiveness ratio. Furthermore, explicit declaration of doing nothing as the alternative intervention helps to frame discussion of the desirability of the intervention.

In estimating an incremental or marginal cost-effectiveness ratio, both the numerator and denominator of the ratio represent differences between the alternative interventions.

$$\text{Difference in cost/Difference in effectiveness} = (C_1 - C_2) / (E_1 - E_2)$$

Where

Difference in cost = cost of intervention (C_1)—Cost of alternative (C_2)

And

Difference in effectiveness = Effectiveness of the intervention (E_1)—Effectiveness of the alternative (E_2)

For example, if a treatment A cost \$50,000, and its average effects in patients with the disease is to add 2.5 QALYs, the average cost-effectiveness ratio of the intervention A would be \$50,000/2.5, or \$20,000 per QALY. The incremental cost-effectiveness of the intervention A compared with the alternative, doing nothing, is also \$20,000 because

$$(C_1 - C_2) / (E_1 - E_2) = (\$100,000 - 0) / (3.5 \text{ QALY} - 0 \text{ QALY})$$

Cost-effectiveness analysis is based on the solution to a simple optimization problem. Given a budget constraint, an explicit objective such as QALYs, and a set of alternative programs, such as treatments, that use resources (cost C_i) and contribute to the objective (effectiveness E_i), the optimal resource allocation is to rank order the programs according to their cost-effectiveness ratios (C_i/E_i) and to select them from lowest to highest to the point where the resource budget is exhausted. This allocation can be shown to yield the maximum total effectiveness, i.e., QALYs gained, subject to the budget constraint.^{[1,8]a}

The following examples^b illustrate the decision rules based on CEA for resource allocation decisions that can be used to maximize the health effects for a fixed budget. The fixed budget used to maximize the health effects will also implicitly yield a price per unit of health effects on the margin. We start with independent programs.

^aThe optimization rule is based on the assumption of divisibility of programs with constant returns to scale, but violations of this assumption are to be expected in practice.

^bThese examples are adapted from Refs. [1,8].

Independent Programs

The decision rule for CEA of independent programs is that they should be ranked from most to least *C/E* ratios. Services that bring large benefits at small cost would be at the top of the list, and those that bring few benefits at high cost would be at the bottom. With a fixed budget, programs should then be implemented in order of their *C/E* ratios until the budget is exhausted.

Assume that the health department in a large metropolitan area has been given a budget of \$500,000 to spend on three different independent health programs, that is, three health programs for three different patient groups. Assume further that each program is the only one available for that patient group so that the three programs are compared with doing nothing for each patient group. This implies that for each program, the average and incremental cost-effectiveness ratios are the same for each patient group. The objective is to save the greatest aggregate number of life-years.

The costs and effects (in terms of life-years gained) of each program are shown in Table 1, where the programs have also been ranked according to their *C/E* ratios. The cost of program A is \$100,000 and the number of life-years gained per patient is 50, yielding a cost-effectiveness ratio of \$2,000, and so on. These cost-effectiveness ratios can be viewed as a ranking of the different programs in terms of their desirability; that is, it is possible to say that program A should be given priority over program C because it has a lower cost-effectiveness ratio. The optimal allocation of the \$500,000 budget is to fund program programs A, B, and C, up to a cutoff of \$6000/LY. The total of 130 LYs saved is the most possible.

In general, the size of budget or the choice of a critical, or cutoff, value of the cost-effectiveness ratio will determine which of the available programs should be implemented. For example, if we have a budget of \$120,000, only program A will be adopted because this program will maximize the effects for the given budget. If the budget increases to \$320,000, both programs A and B will be adopted because this maximizes the effectiveness for the budget. Finally, if the budget increases further to \$500,000, programs A, B, and C will be funded. The decision rule is thus to start implementing the program with the lowest cost-effectiveness ratio and then to add additional programs according to their cost-effectiveness ratios until the budget is exhausted.

Table 1 Example of Three Independent Programs

Programs	Cost	Life-years	<i>C/E</i>
A	\$120,000	50	\$2,400
B	\$200,000	50	\$4,000
C	\$180,000	30	\$6,000
D	\$200,000	25	\$8,000

Note also that the different budgets yield an implicit price that we are willing to pay per life-year gained on the margin. For instance, if a budget of \$120,000 is used, this implies that we are willing to pay \$2,400 per life-year gained on the margin. It is the price we are paying per life-year gained for the program with the highest cost-effectiveness ratio that is implemented. If a budget of \$500,000 is used, this implies that we are willing to pay \$6,000 per life-year gained on the margin. Thus the cost-effectiveness cutoff in this example is \$6,000/LY if the budget is \$500,000. This means that if a new program is added to the menu, it should be adopted only if its cost-effectiveness ratio is less than \$6,000. With a limited budget, the cost that appears in the numerator of the *C/E* ratio of a program is the net burden of the program on the budget.

A classic example of CEA is the experience of Oregon's attempt to ration medical benefit with its limited Medicaid budget.^[5] The plan was simple. The legislature appointed a commission to rank Medicaid services from the least cost-effective. Services that bring large benefits at small cost would be at the top of the list, and those that bring few benefits at high cost would be at the bottom. The state would then estimate the cost of providing each service to an expanded number of Medicaid beneficiaries, starting at the top of the list and moving down until the money ran out. At that point, it would draw a line and all services below the line would no longer be reimbursed by the Medicaid program. By refusing to pay for services that are not very cost-effective, the state would save money, enabling it to add extra people to Medicaid. In this case, cost-effectiveness analysis shows how to maximize health benefits within a specific health-care budget. By specifying its Medicaid budget and the number of people that would be covered, Oregon could determine which services it could afford. Thus it would, in theory, show how to buy the greatest amount of medical benefits with its Medicaid dollars.

Mutually Exclusive Programs

Often, applications of CEA in health care involve comparisons among competing alternatives for the same condition, or the appropriate intensity of an intervention or the appropriate groups to receive it, for example, alternative drug dosages for a given condition, or comparisons between drugs and surgical treatment. In such competing choice situations, the basic CEA decision rule needs to be modified to incorporate the possibility of mutually exclusive programs for the same condition. This section considers the decision rules for mutually exclusive programs. Similar to the independent programs, a fixed budget can be used as the decision rule to decide which program should be implemented.

For the mutually exclusive programs, the first step is to order the programs according to their effectiveness. The next step is to calculate the incremental *C/E* ratios. Then, dominated alternatives should be excluded. The

Table 2 Four Mutually Exclusive Programs

Programs	Cost	Benefits (LYs)	Average <i>C/E</i>	Incremental <i>C/E</i>	Without program C
A	\$50,000	10	\$5,000	\$5,000	\$5,000
B	\$200,000	20	\$10,000	\$15,000	\$15,000
C	\$900,000	30	\$30,000	\$70,000	–
D	\$1000,000	40	\$25,000	\$10,000	\$40,000

incremental *C/E* ratios should then be recalculated without the dominated alternatives being included. This process continues until a number of programs with increasing incremental *C/E* ratios remain. With a fixed budget, we then start by considering the least effective program, and then continuously switch patients to more effective programs until the budget is exhausted.

Assume that there are four different alternative programs (e.g., four different cholesterol-lowering drugs) for treating patient groups. The costs and effects are shown in Table 2. Because the four programs are mutually exclusive, it is only possible to carry out one of the programs for each patient group. The first column in Table 2 shows the costs; the second column shows the effects; the third column shows the average *C/E* ratios, which compares each program to the status quo baseline alternative. Note that the average cost-effectiveness ratios are misleading and should not be used as a decision rule.

The correct analysis examines the incremental cost of program B relative to A, and so on, as well as its incremental benefit. For program A, the incremental *C/E* ratio is equal to the costs of A compared to doing nothing divided by the effects of A compared to doing nothing, which gives an incremental cost-effectiveness ratio of \$5,000. This is the same as the average *C/E* ratio for A because it is the least effective program.

For B, the incremental *C/E* ratio is equal to the costs of B minus the costs of A (\$200,000 – \$50,000 = \$150,000) divided by the effects of B minus the effects of A (20 – 10 = 10), which gives an incremental *C/E* ratio of \$15,000 (150,000/10). For C, the incremental cost-effectiveness ratio is equal to the costs of C minus the costs of B (900,000 – 200,000 = 700,000) divided by the effects of C minus the effects of B (30 – 20 = 10), which gives an incremental cost-effectiveness ratio of \$70,000 (700,000/10). For D, the incremental *C/E* ratio is equal to the cost of D minus the costs of C (1000,000 – 900,000 = 100,000) divided by the effects of D minus the effects of C (40 – 30 = 10), which gives an incremental *C/E* ratio of \$10,000 (100,000/10).

The next step is to exclude dominated alternative. A program is dominated if the incremental *C/E* ratio decreases for the next program with higher effectiveness (it yields fewer health benefits for the same amount of money). The reason for excluding the dominated alternative is that irrespective of the size of the budget, more units of effectiveness are always gained by doing the more effective program.

In this example, program C is dominated because the incremental *C/E* ratio for D decreases. Therefore, program C should be excluded from further consideration. After C has been excluded, the incremental *C/E* ratios should be recalculated. The incremental *C/E* ratios for A and B will be the same as before, but the incremental *C/E* ratio for D will change. For D, the incremental *C/E* ratio is now equal to the costs of D minus the costs of B (1000,000 – 200,000 = 800,000) divided by the effects of D minus the effects of B (40 – 20 = 20), which gives an incremental cost-effectiveness ratio of \$40,000 (800,000/20).

The incremental *C/E* ratios show the implied price per life-year gained from implementing the different programs. The ratios cannot, however, be used to rank the programs because without knowing the size of the budget or the price per life-year gained that society is willing to pay, it is impossible to know which of the four mutually exclusive programs should be implemented.

With a fixed budget, the decision rule is to start by considering the program with the lowest incremental *C/E* ratio, and then switch patients to continuously more effective programs until the budget is exhausted. Table 3 shows the program choice for different budget sizes. Because the programs are mutually exclusive, each patient can only receive one of the programs. This means that as the budget increases, we will switch from a less effective program to a more effective program.

For a budget of \$50,000, program A will be implemented. If all patients are treated with program A, all the budget of \$50,000 will be exhausted, so it is not possible to switch any patients to more effective treatments. If the budget is above \$50,000, we can start to switch patients from treatment A to B. If the budget is \$200,000, we will be able to switch all the patients to program B. Finally, if the budget exceeds \$200,000, we can continue to switch patients from program B to program D.

Table 3 The Choice of Mutually Exclusive Program for Different Sizes of the Budget

Budget	Choice of program
50,000	A
200,000	B
1,000,000	D

If the budget is \$1 million or greater, all the patients will receive the most effective treatment, D.

The different budgets also implicitly yield the price we are willing to pay for extra life-years gained. For example, for a budget of \$200,000, the implied price per life-year gained on the margin is \$15,000, that is, the incremental *C/E* ratio for program B. This criterion for judging the acceptability of a *C/E* ratio is known as the shadow price of an explicit budget constraint. The shadow price or the *C/E* ratio of the marginal program represents the opportunity cost of diverting resources to fund a new program. Therefore any new program that competes for resources must have a *C/E* ratio of \$15,000/LY or less to allow a reallocation of resources to the new program to result in a gain of life-years. The criterion would, of course, change as the budget expands, higher *C/E* ratios become acceptable, and vice versa as the budget decreases.

To illustrate the principle of incremental cost-effectiveness analysis, consider the study by Goldman et al.^[9] They evaluated alternative doses of the drug lovastatin for cholesterol reduction in patients who have had a heart attack. The drug has been shown to reduce serum cholesterol, and the amount of cholesterol reduction increases with increasing dose. The benefits are reduced mortality and cost from heart disease. Cost-effectiveness of dosages of 20, 40, and 80 mg are considered among patients with cholesterol levels of at least 250 mg/dl who had suffered a previous heart attack. Table 4 shows a sample of their results. The model calculated present values of net medical cost and life-years for the U.S. population over the period from 1990 to 2015.

Note that one could be easily misled by using the average *C/E* for the 80-mg dose, which is \$25,800. However, the correct measure of its cost-effectiveness is \$84,300, its incremental cost per life-year compared to the lower dose.

Cost-effectiveness information often highlights the counterintuitive high cost of some seemingly inexpensive medical services. Imagine that we offer a \$1,000 screening test to a group of 2,000 people at a total cost of \$2 million. The test saves 100 lives, at a cost of \$20,000 (\$2 million/100) per life saved. This seems like a bargain. What society would not want to spend \$20,000 to save someone's life, or \$2 million to save 100 lives? But what

if an alternative screening test was available that was half as expensive and would save 99 lives? According to CEA, the existence of this less-expensive test radically alters the cost-effectiveness of the first test. As we have seen, the first test costs \$2 million; the alternative test costs only \$1 million. The first test saves 100 lives and the second saves 99. Thus the real cost-effectiveness of the first test is not \$20,000 per life saved, but \$1 million per life saved. We have to increase spending from \$1 million to \$2 million to save one additional life. The pitfall in economic evaluation is failure to acknowledge the existence of less-expensive, although somewhat less beneficial, options in evaluating the *C/E* of a program. After all, any option can be made to look cost-effective if it is compared to a sufficiently less *C/E* alternative.^[5]

CONCLUSION

Although *C/E* is but one element of public decisions about health care, gaining a clearer quantitative understanding of this important element cannot help but improve decision processes. Resource allocation decisions can never be shaped by the mechanical ranking of cost-effectiveness ratios. Ratios provide information about one type of "value," health benefit per dollar spent. But other values of society, including considerations of distributive justice and fairness (e.g., giving priority at times to the sickest of individuals), require that CEA be viewed as an informer of decision making rather than a decision maker per se.^[6] For example, sex education and the distribution of condoms in public school have raised a range of concerns reflecting conflicting social values involving students, parents, educators, and public health officials. The economic evidence must be combined with an understanding of preferences and values of the stakeholders that are not intrinsically technical.

In the face of increasingly limited resources CEA provides opportunities for more efficient delivery of health care. The challenge would be restructure the system in a fashion to create incentives that encourage the appropriate delivery of efficient interventions.^[10]

REFERENCES

1. Gold, M.R.; Siegel, J.; Russel, L.B.; Weinstein, M. *Cost-effectiveness in Health and Medicine*; Oxford University Press: New York, 1996.
2. Drummond, M. *Methods for the Economic Evaluation of Health Care Programmes*; 3rd Ed.; Oxford University Press: New York, 2005.
3. Loomes, G.; McKenzie, L. The use of QALYs in health care decision making. *Soc. Sci. Med.* **1989**, *28*, 299–308.
4. Mendelson, D. N. Outcomes and effectiveness research in the private sector. *Health Aff.* **1998**, *17* (5), 75–90.

Table 4 Cost-Effectiveness Analysis for Alternative Doses of Lovastatin

Dose (daily)	Total cost (billion of \$)	Effectiveness LYs	<i>C/E</i>	dC/dLY
<i>Men, 65–74</i>				
20 mg	3.615	348,272	10,400	10,400
40 mg	7.051	477,204	14,800	26,650
80 mg	14.657	567,468	25,800	84,300

5. Ubel, P. A. *Pricing Life*; The MIT Press: Cambridge, MA, 2000.
6. Eddy, D. M. *Clinical Decision Making: from Theory to Practice*; Jones and Bartlett: Sudbury, MA, 1996.
7. Detsky, A.S.; Naglie, I. Gary a clinician's guide to cost-effectiveness analysis. *Ann. Intern. Med.* **1990**, *113* (2), 147–154.
8. Johannesson, M. *Theory and Methods of Economic Evaluation of Health Care*; Kluwer: Dordrecht, 1996.
9. Goldman, L.; Weinstein, M.C.; Goldman, P.A.; Williams, L.W. Cost-effectiveness of HMG-CoA reductase inhibition for primary and secondary prevention of coronary heart disease. *JAMA* **1991**, *265*, 1145–1151.
10. Cohen, J.T.; Neumann, P.J.; Weinstein, M.C. Does preventive care save money? Health economics and the presidential candidates. *N Engl J Med.* **2008**, *358*, 661–663.

Phase I Cancer Clinical Trials

Seung-Ho Kang

Ewha Womans University, Seoul, Korea

Chul Ahn

University of Texas Medical School, Houston, Texas, U.S.A.

Abstract

In phase I cancer clinical trials, the subjects are extremely ill patients with cancer who will often be at very high risk of death in the short term under all standard therapies. This entry reviews some of the traditional methods and designs of phase I cancer clinical trials as well as reassessment methods.

INTRODUCTION

The phase I clinical trials represent the first time drugs are given to man and are mainly carried out to establish human toleration. In many phase I clinical trials, the subjects are healthy adult volunteers, while in phase I cancer clinical trials, the subjects are extremely ill patients with cancer who will often be at very high risk of death in the short term under all standard therapies. The design and statistical analysis of the phase I cancer clinical trials are different from those of the other phase I clinical trials. More research into the design and statistical analysis of the phase I cancer clinical trials has been conducted than that of the other phase I clinical trials. Therefore the phase I clinical trials are often used to represent the phase I cancer clinical trials among statisticians.^[1] Here we review only the phase I cancer clinical trials.

The primary aim of the phase I cancer clinical trials is to determine the maximum tolerated dose (MTD) of a new drug, which will be used as a recommended dose level for experimentation in phases II and III cancer clinical trials. In most cancer therapy, toxicity is a prerequisite for optimal antitumor activity. Hence one must endure some degree of treatment-related toxic reaction if patients are to have a reasonable chance of favorable response. Because higher doses are associated with both greater therapeutic benefits and an increased probability of severe toxic reaction, an anticancer drug should be administered at the maximum dose that cancer patients can tolerate. Consequently, the goal of a cancer phase I trial is to determine the highest dose associated with an acceptable level of toxicity, which is often called the maximum tolerated dose (MTD).

Perhaps the two most serious problems currently facing phase I cancer clinical trials are the limited sample size and the ethical concerns. Usually, the total number of patients entered is quite small (such as 20 to 30), and there may be considerable heterogeneity in the patients' health status at entry despite the establishment of entry criteria. It makes it

hard for biostatisticians to find the best design and analysis procedure which are usually evaluated by large sample theory. In contrast to other phase I trials, cancer phase I trials also have a therapeutic aim. Typically, participants in cancer phase I trials are patients at advanced disease stages who consent to participate in the trial only as a last resort in seeking cure. One should design cancer phase I trials to minimize both the number of patients treated at low, nontherapeutic doses as well as the number given severely toxic overdoses.

THE TRADITIONAL METHODS

Although the traditional designs are shown to have poor operating characteristics compared with the continual reassessment method,^[2–6] in practice, the traditional designs still predominate in phase I designs. The primary reasons seem to be that (1) the method is easy to describe and does not require a statistician to design or evaluate and (2) no modeling is required beyond an assumption that the dose–toxicity relationship is nondecreasing. The traditional designs are as follows.

A precise definition of the dose-limiting toxicity (DLT), an occurrence of unacceptable toxicity, is provided in the protocol. The recommendation for grading toxicity is given, for example, in.^[7] The selection of the starting dose depends heavily on pharmacology and toxicology from preclinical studies. Two measures useful in translating the drug potency in animals to that in humans are the dog toxic low dose (TLD), defined as the minimum dose at which any toxicity is observed, and the mouse lethal dose 10 (LD_{10}), defined as the dose that kills 1/10 of the animals. The starting dose in human trials is usually 10% of mouse LD_{10} if it was nontoxic in dog studies. If it was toxic in dogs, then 1/3 the dog TLD is often selected. Subsequent dose levels are often determined by a modified Fibonacci scheme. Let x_1 denote the starting dose. According to a

modified Fibonacci scheme, $x_2 = 2x_1$, $x_3 = 1.67x_2$, $x_4 = 1.5x_3$, $x_5 = 1.4x_4$, and thereafter, $x_{k+1} = 1.33x_k$.

The following two types of the traditional designs are popular. The traditional design without dose de-escalation is described as follows. Three patients are assigned at the first dose. We proceed to the next higher dose level with a cohort of three patients until at least one patient experiences the DLT. If at least two out of three patients experience the DLT, then the previous dose level is defined as the MTD. If one out of three patients experiences the DLT, then three more patients are treated at the same dose level. If at least one of the three additional patients experiences the DLT, then the previous dose will be considered as the MTD. Otherwise, the dose will be escalated.

The traditional design with dose de-escalation is the same as the above design except for the following things. If at least two out of six patients experience the DLT in the initial cohort and the additional cohort, or at least two out of three patients experience the DLT in the initial cohort treated at a dose level, then we have exceeded the MTD. We would treat another three patients at the previous dose level if there were only three patients treated at that level. Hence the dose may de-escalate until we either reach a dose level in which six patients are already treated or observe at most one DLT out of six patients. We define the MTD as the highest dose level (≥ 1) in which we have treated six patients with at most one patient experiencing the DLT or a dose level 0 if there were at least two patients experiencing the DLT at dose level 1. The level 0 is not a level at which we would treat patients. When the level 0 is decided as the MTD, in practice, dose levels are lowered and a new trial is conducted.

In most previous studies on phase I cancer clinical trials, a few scenarios of dose–toxicity relationships were selected and investigations were performed by simulation. Recently, Lin and Shih^[8] derived the exact formulae of key statistical properties of the traditional 3 + 3 designs both with and without dose de-escalation. Using their results, Kang and Ahn^[9,10] conducted the exact computation of several key statistical properties for hundreds of dose–toxicity curves that belong to the same family of increasing functions. The performance of the traditional designs depends on the unknown true dose–toxicity relationship and dose levels, which makes it hard to draw a general conclusion. However, investigations for the following logistic and the hyperbolic tangent functions shed some lights on the performance of the traditional designs.

$$\varphi_1(x_i : a) = \frac{\exp(1.5a + a \times x_i)}{1 + \exp(1.5a + a \times x_i)},$$

$$x_i = -4.0 + 0.5(i - 1), \quad 1 \leq i \leq 13$$

and

$$\varphi_2(x_i : b) = \left[\frac{\tanh(x_i) + 1}{2} \right]^b,$$

$$x_i = -1.4 + 0.25(i - 1), \quad 1 \leq i \leq 13$$

respectively. These two families of functions are plotted in Fig. 1. The exact computation of several key statistical properties is performed for each value of both $0.8 \leq a \leq 2.4$ for the logistic function and $0.5 \leq b \leq 2.0$ for the hyperbolic tangent function with the increment of 0.01. The expected toxicity rates at the MTD of the traditional designs without and with dose de-escalation are plotted for the logistic and the hyperbolic tangent functions in Fig. 2. The solid and dotted lines represent the traditional designs with and without dose de-, respectively. The expected toxicity rate at the MTD turns out to be between 17% and 22% for the logistic and the hyperbolic tangent dose–toxicity relationship if the toxicity rate is very small in low dose levels.^[10] The traditional designs are often considered to produce a 33% toxicity rate, but the authors are unaware of any published result to support this consideration. Fig. 3 displays the expected number of toxicity and the expected number of required patients in the logistic and the hyperbolic tangent functions. The solid and dotted lines represent the traditional designs without and with dose de-escalation, respectively. On average, five to six toxicities are expected to occur in the traditional design with dose de-escalation, while two to three toxicities are likely to occur in the design without dose de-escalation. The expected number of required patients ranges from 10 to 30, and the difference between the two designs is about three patients.

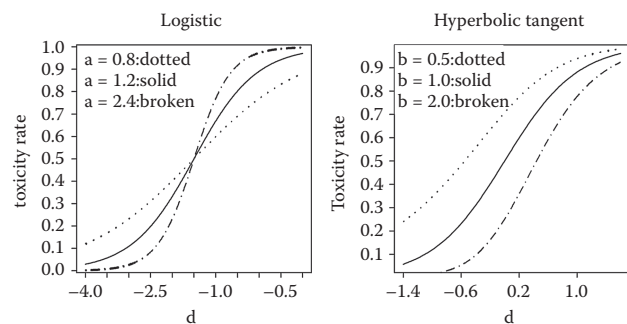


Fig. 1 Dose–toxicity curves

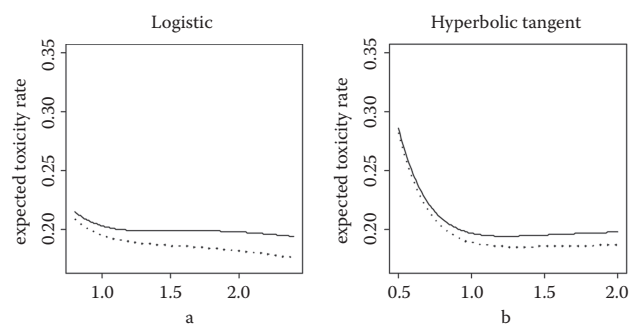


Fig. 2 Expected toxicity rate at MTD. Solid: without; dotted: with dose de-escalation

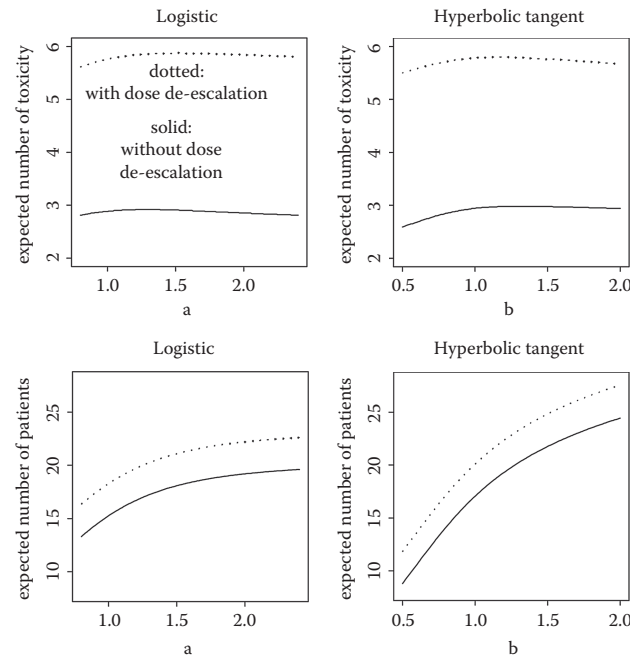


Fig. 3 Expected number of toxicity and patients

THE CONTINUAL REASSESSMENT METHOD

One problem with the standard design is that many patients are treated at low dose levels. Another problem is the large variability of the estimated MTD around the true MTD. The major criticism of the standard design is that it has no interpretation as an estimate of the dose level which yields a specified toxicity rate. The results of Kang and Ahn^[9,10] show clearly that it is difficult for the toxicity rate at the MTD determined from the traditional designs to be close to a given toxicity rate which is not between 17% and 22%.

O'Quigley et al.^[3] proposed the continual reassessment method (CRM), which reduces the number of patients treated with possibly ineffective dose levels. The starting dose of the CRM is selected to be the prior estimate of the MTD. When toxicity outcomes are known for successive patients, the dose–toxicity curve is updated and used to determine the dose at which to treat the next patient. Shen and O'Quigley^[11] established the consistency and asymptotic normality for the maximum likelihood estimates. They also showed that the recommended dose will generally converge to the target toxicity rate as the sample size increases. However, there are situations in which the recommended dose may converge to a level close to the target level, but not necessarily the closest. This contrasts with the conclusions of O'Quigley et al.^[3] Shen and O'Quigley^[11] describe this situation as well as provide the conditions under which convergence is assured.

The CRM is described as follows. Let $x_1, x_2, \dots, x_k$ denote the dose levels chosen for experiment. Let θ be the

probability of toxic response corresponding to the aimed target dose level. Let Y_i be a random variable for the response of the i th patient. The binary random variable Y_i takes on 1 for toxic response and 0 for nontoxic response. Consider some simple dose–response function for $E(Y_i)$ and denote this by $\Psi(x_i, a)$. Let $x(j)$ denote the dose level chosen for the j th patient. The CRM is performed as follows.

1. A mathematical model $\Psi(x_i, a)$ for dose–response is assumed. For example,

$$\Psi(x_i : a) = \frac{\exp(a_0 + a \times x_i)}{1 + \exp(a_0 + a \times x_i)},$$

$$\Psi(x_i : a) = \left[\frac{\tanh(x_i) + 1}{2} \right]^a$$

2. Assume a prior distribution for a , $g(a) = f(a, \Omega_1)$. For example, the exponential prior is popular.

$$g(a) = \exp(-a), \quad 0 < a < \infty$$

3. Define θ , aimed target toxicity level.
4. Assign the first patient in the dose level whose probability of toxicity is judged closest to θ .
5. After observing Y_j , the posterior distribution of a is modified. Specifically, let

$$\phi(x(j), y_j, a) = \Psi^{y_j}(x(j), a)(1 - \Psi(x(j), a))^{(1-y_j)}$$

Then, the likelihood of $(Y_1, \dots, Y_j, X_1, \dots, X_j)$ is

$$\prod_{l=1}^j \phi(x(l), y_l, a)$$

The updated posterior distribution of a after observing $(Y_1, \dots, Y_j, X_1, \dots, X_j)$ is

$$f(a, \Omega_{j+1}) = \frac{g(a) \prod_{l=1}^j \phi(x(l), y_l, a)}{\int_0^\infty g(u) \prod_{l=1}^j \phi(x(l), y_l, u) du}$$

where $\Omega_{j+1} = \{y_1, \dots, y_j\}$

6. We compute the posterior mean

$$\mu(j+1) = \int_0^\infty af(a, \Omega_{j+1}) da$$

Take $\hat{a} = \mu(j+1)$ and find the dose level for the next patient such that

$$\left(x_i - \Psi_{a=\mu(j+1)}^{-1}(\theta) \right)^2$$

is minimized.

7. We repeat steps 5 and 6 until we reach a predetermined fixed sample size, or some other trial terminating condition is satisfied. The recommended dose is the dose such that

$$\left(x_i - \Psi_{a=\mu(n+1)}^{-1}(\theta) \right)^2$$

is minimized.

The following example is from O'Quigley et al.^[3] Suppose that we use the following model

$$\Psi(x_i, a) = \left(\frac{\tanh(x_i) + 1}{2} \right)^a$$

and the exponential prior distribution.

$$g(a) = \exp(-a), \quad a > 0$$

Suppose that the true state is $a_0 = 0.5$ and the dose levels are given by

Dose level	1	2	3	4	5	6
x_i	-1.47	-1.1	-0.69	-0.42	0.0	0.42
$\Psi(x_i, a=0.5)$	0.22	0.32	0.45	0.54	0.69	0.80

Let the target toxicity rate be $\theta = 0.2$. Suppose that experimentation starts at $x(1) = x_3$. From Table 1 in O'Quigley et al.,^[3] we observe that $y_1 = 0$. Then

$$\begin{aligned} \phi(x(1), y_1, a) &= 1 - \left(\frac{\tanh(-0.69) + 1}{2} \right)^a \\ &= 1 - 0.201009^a \\ f(a, \Omega_2) &= \frac{\exp(-a)(1 - 0.201009^a)}{\int_0^\infty \exp(-u)(1 - 0.201009^u) du} \\ &= (1/0.616035) \exp(-a) (1 - 0.201009^a) \\ \mu_2 &= \int_0^\infty af(a, \Omega_2) da = 1.38 \end{aligned}$$

Solving $\Psi(x, \mu(2)) = 0.2$, we have $x = \theta - 0.3964$. Because x_4 is closest to x , x_4 is the next dose level. This is the way that $x(2) = x_4$ is found in Table 1.^[3]

Although the CRM outperforms the traditional designs, the original CRM has some difficulties to be implemented in real practice. One of them is that it takes too long to complete the trial because the CRM treats one patient at a time. Goodman et al.^[4] proposed the modified CRM that assigns more than one patient at a time to each dose level and limits the dose escalation by one level. Another modification is that the starting dose of the modified CRM is the lowest dose level, not the prior estimate of the MTD. This modification makes clinicians more comfortable, probably because treating at less effective doses is viewed less serious than treating at too toxic doses. Goodman et al.^[4] showed that the modified CRM reduces the duration of the trial by 50–67%, reduces a toxicity incidence by 20–35%, and lowers toxicity severity from the original CRM.

The modified CRM which leans to a Bayesian methodology can be changed to a classic framework leaning upon likelihood theory in which the maximum likelihood estimate is used instead of the posterior mean.^[12] Shen and O'Quigley^[11] proved the consistency and asymptotic normality for the maximum likelihood estimates under model misspecification. Cheung and Chappell^[13] interpreted the main condition under which the CRM is consistent and apply it to evaluate the model sensitivity in the CRM. With extensive simulation studies, Ahn^[6] compared nine phase I designs which include the traditional design, the two-stage modified Storer's design,^[14] the two-stage Korn's design,^[15] the modified CRM, and the two-stage modified CRM.^[5] In general, the modified CRM with two or three patients per cohort performs well compared with the other designs. Especially, the modified CRM has substantial gains in the chance of choosing the correct MTD than the traditional design.

Research on the CRM has been still active until a recent date. The CRM is modified to solve the problem of deciding the dose level of the next patient when toxicity cannot be observed immediately after treatment.^[16,17] Stopping rules based on a predetermined number of iterations^[2,4,18] and more sophisticated criteria have been studied.^[19–21] Another modification by O'Quigley et al.^[22] suggests the use of covariate information in CRM. Ishizuka and Ohashi^[23] proposed to monitor a posterior density function

of the occurrence of the DLT at each dose level instead of a which is difficult to interpret, especially for clinical investigators. Piantadosi and Liu^[24] and Legedza and Ibrahim^[25] proposed a method of incorporating pharmacokinetic data into a phase I cancer clinical trial which is a variant of the CRM. Braun^[26] extended the CRM to a bivariate trial design in which the MTD is estimated based on both toxicity and disease progression. The CRM employs a one-parameter dose–response curve such as the logistic or the hyperbolic tangent function. Gasparini and Eisele^[27] proposed a curve-free CRM in which the assumption of a specific dose–response is not needed, and Cheung^[28] modified the curve-free CRM when a low percentile of toxicity is targeted. However, O’Quigley^[29] pointed out that the curve-free method is equivalent to the usual model-based CRM.

OTHER DESIGNS

Storer^[14] proposed two two-stage designs (BC and BD), a combination of simple dose-escalation algorithms. In the first stage, a single patient is treated at each dose level. The next patient is treated at the next lower dose level if a toxic response is observed, otherwise at the next higher dose level. When the first toxicity is reached, the dose is decreased one level and the second stage begins with fixed sample size. In the second stage of the design BC, if none of two has DLT, then escalate, whereas if one or two of these have DLT, then de-escalate. In the second stage of design BD, three patients are treated in the second stage. Escalation occurs if no toxicity is observed, and de-escalation occurs if two or more patients have toxicity. If a single patient has toxicity, then the next group of three patients is treated at the same dose level. After completion of the second stage, a dose–response model is fitted to the data and the MTD is estimated by the maximum likelihood estimate.

Babb et al.^[30] proposed a Bayesian dose escalation with overdose control (EWOC) in which approaches the MTD as fast as possible subject to the constraint that the predicted proportion of patients who receive an overdose does not exceed a specified value. Instead of using a dose with toxicity deemed closest to the target toxicity at each update, as the CRM does, EWOC uses the dose with probability of higher toxicity than the current estimate of the MTD, not exceeding some critical value.

Leung and Wang^[31] proposed a new approach which requires no assumptions on the dose–toxicity relationship other than that the probability of toxicity is nondecreasing with dose. The method determines escalation and de-escalation after adjusting the toxicity estimates at each dose to ensure a nondecreasing order. The adjusted toxicity estimates are obtained from isotonic regression.

Durham et al.^[32] proposed random walk rules for phase I cancer clinical trials. It is simple to implement random walk rules, and finite and asymptotic distribution

theory is completely worked out. The small-sample properties of random walk rules determined by exact theory compare favorably to those of the CRM determined by simulation.

CONCLUSION

In summary, phase I clinical trials are challenged by reaching the MTD efficiently and quickly while offering patients safe and appropriate treatment. Thus an optimal phase I trial design must fully evaluate drug toxicity but, at the same time, must also minimize the likelihood of treating patients at both subtherapeutic and severely toxic doses. The traditional-cohort escalation trial design has many shortcomings because of small cohort sizes and its ad hoc nature. Recently, there has been increasing interests in reexamining the phase I trial design. To identify the MTD, methods have been proposed such as the CRM, Bayesian approaches, and up-and-down procedures or strategies for monitoring toxicity and efficacy. In the near future, we expect more publications of successfully conducted phase I cancer clinical trials that employed one of the designs described in this article.

REFERENCES

1. Friedman, L.M.; Furberg, C.D.; DeMets, D.L. *Fundamentals of Clinical Trials*, 3rd Ed.; Mosby-Year Book: St. Louis, MO, 1996.
2. O’Quigley, J.; Chevret, S. Methods for dose finding studies in cancer clinical trials: A review and results of a Monte Carlo study. *Stat. Med.* **1991**, *10*, 1647–1664.
3. O’Quigley, J.; Pepe, M.; Fisher, M. Continual reassessment method: A practical design for phase I clinical trials in cancer. *Biometrics* **1990**, *46*, 33–48.
4. Goodman, S.; Zahurak, M.; Piantadosi, S. Some practical improvements in the continual reassessment method for phase I studies. *Stat. Med.* **1995**, *14*, 1149–1161.
5. Moller, S. An extension of the continual reassessment methods using a preliminary up-and-down design in a dose finding study in cancer patients, in order to investigate a greater range of doses. *Stat. Med.* **1995**, *14*, 911–922.
6. Ahn, C. An evaluation of phase I cancer clinical trial designs. *Stat. Med.* **1998**, *17*, 1537–1549.
7. Miller, A.B.; Hoogstraten, F.B.; Staquet, M.; Winkler, A. Reporting results of cancer treatment. *Cancer* **1981**, *47*, 207–214.
8. Lin, Y.; Shih, W. Statistical properties of the traditional algorithm-based designs for phase I cancer clinical trials. *Biostatistics* **2001**, *2*, 203–215.
9. Kang, S.H.; Ahn, C. The expected toxicity rate at the maximum tolerated dose in the standard phase I cancer clinical trial design. *Drug Inf. J.* **2001**, *35* (4), 1189–1200.
10. Kang, S.H.; Ahn, C. An investigation of the traditional algorithm-based designs for phase I cancer clinical trials. *Drug Inf. J.* **2002**, *36*, 865–873.

11. Shen, L.; O'Quigley, J. Consistency of continual reassessment method under model misspecification. *Biometrika* **1996**, *83* (2), 395–405.
12. O'Quigley, J.; Shen, L. Continual reassessment method: A likelihood approach. *Biometrics* **1996**, *52*, 673–684.
13. Cheung, Y.K.; Chappell, R. A simple technique to evaluate model sensitivity in the continual reassessment method. *Biometrics* **2002**, *58*, 671–674.
14. Storer, B. Design and analysis of phase I clinical trials. *Biometrics* **1989**, *45*, 925–937.
15. Korn, E.; Midthune, D.; Chen, T.; Rubinstein, L.; Christian, M.; Simon, R. A comparison of two phase I trial design. *Stat. Med.* **1994**, *13*, 1799–1806.
16. Cheung, Y.K.; Chappell, R. Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics* **2000**, *56*, 1177–1182.
17. Thall, P.F.; Lee, J.J.; Tseng, C.; Estey, E.H. Accrual strategies for phase I trials with delayed patient outcome. *Stat. Med.* **1999**, *18*, 1155–1169.
18. Chevret, S. The continual reassessment method in cancer phase I clinical trials: A simulation study. *Stat. Med.* **1993**, *12*, 1093–1108.
19. O'Quigley, J.; Reiner, E. A stopping rule for the continual reassessment method. *Biometrika* **1998**, *85*, 741–748.
20. Heyd, J.M.; Carlin, B.P. Adaptive design improvements in the continual reassessment method for phase I studies. *Stat. Med.* **1999**, *18*, 1307–1321.
21. Zohar, S.; Chevret, S. The continual reassessment method: Comparing of Bayesian stopping rules for dose-ranging studies. *Stat. Med.* **2001**, *20*, 2827–2843.
22. O'Quigley, J.; Shen, Z.L.; Gamst, A. Two sample continual reassessment method. *J. Biopharm. Stat.* **1999**, *9*, 17–44.
23. Ishizuka, N.; Ohashi, Y. The continual reassessment method and its applications: A Bayesian methodology for phase I cancer clinical trials. *Stat. Med.* **2001**, *20*, 2661–2681.
24. Piantadosi, S.; Liu, G. Improved designs for dose escalation studies using pharmacokinetic measurements. *Stat. Med.* **1996**, *15*, 1605–1618.
25. Legedza, A.; Ibrahim, J. Longitudinal design for phase I clinical trials using the continual reassessment method. *Control. Clin. Trials* **2000**, *21*, 574–588.
26. Braun, T. The bivariate continual reassessment method extending the CRM to phase I trials of two competing outcomes. *Control. Clin. Trials* **2002**, *23*, 240–256.
27. Gasparini, M.; Eisele, J. A curve-free method for phase I clinical trials. *Biometrics* **2000**, *56*, 609–615.
28. Cheung, Y.K. On the use of nonparametric curves in phase I trials with low toxicity tolerance. *Biometrics* **2002**, *58*, 237–240.
29. O'Quigley, J. Curve-free and model-based continual reassessment method designs. *Biometrics* **2002**, *58*, 245–249.
30. Babb, J.; Rogatko, A.; Zacks, S. Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Stat. Med.* **1998**, *17*, 1103–1120.
31. Leung, D.H.; Wang, D. Isotonic designs for phase I trials. *Control. Clin. Trials* **2001**, *22*, 126–138.
32. Durham, S.D.; Flournoy, N.; Rosenberger, W.F. A random walk rule for phase I clinical trials. *Biometrics* **1997**, *53*, 745–760.

Phase II Clinical Development

Naitee Ting

Boehringer-Ingelheim Pharmaceuticals, Inc., Ridgefield, Connecticut, U.S.A.

Guojun Yuan

LFB USA, Inc., Framingham, Massachusetts, U.S.A.

Abstract

Most of the drugs available in pharmacy started out as a chemical or biological compound discovered in laboratories. A chemical compound is relatively a smaller molecule synthesized in a laboratory, whereas a biological compound tends to be a large molecule extracted from living bodies, such as plant or animal tissues. When first discovered, this new compound is denoted as a product candidate. Medicinal product development is a process that begins when the product candidate is first discovered and continues until it is available to be prescribed by physicians to treat patients. It is usually a new chemical or biological entity synthesized by scientists from pharmaceutical companies, universities, or other research institutes. This new compound will have to undergo the product development process and obtain regulatory approval before it can be used by the general patient population.

Keywords: Phase II; Clinical development; Proof of Concept; dose-ranging; dose-response; go/no go decision.

BACKGROUND

Most of the drugs available in pharmacy started out as a chemical or biological compound discovered in laboratories. A chemical compound is relatively a smaller molecule synthesized in a laboratory, whereas a biological compound tends to be a large molecule extracted from living bodies, such as plant or animal tissues.^[1] When first discovered, this new compound is denoted as a product candidate. Medicinal product development is a process that begins when the product candidate is first discovered and continues until it is available to be prescribed by physicians to treat patients.^[2,3] It is usually a new chemical or biological entity synthesized by scientists from pharmaceutical companies, universities, or other research institutes. This new compound will have to undergo the product development process and obtain regulatory approval before it can be used by the general patient population.

The product development process can be broadly classified into two major components—non-clinical development and clinical development. Non-clinical development includes all product testing performed outside the human body, which can be further broadly divided into pharmacology, toxicology, and product formulation. In contrast, clinical development is based on studies conducted in the human body to understand how the human body reacts to the product candidate and, in addition, how the product candidate helps patients with the disease treatment or if it leads to any potential adverse events. Clinical development can be further divided into phases I, II, III, and IV.

Phase I trials are designed to study the short-term effects; e.g., pharmacokinetics (PK, what does a human body do to the drug?), pharmacodynamics (PD, what does a drug do to the human body?), and the upper bound of dose range [maximal tolerated dose (MTD)]. Phase II trials are designed to study the efficacy of the new product in well-defined patient populations, often with dose-response relationship as one of the major objectives. Phase III studies are usually longer term, large-scale studies to confirm findings established from earlier trials. These studies are also used to detect the adverse effects caused by cumulative dosing. If a new product is found to be safe and efficacious from the first three phases of clinical testing, in the United States, a New Drug Application for a new chemical entity or a Biological License Application for a new biological entity is submitted to Food and Drug Administration (FDA) for review. Once the new product is approved by FDA, phase IV (postmarketing) studies are planned and carried out. Many of the phase IV study designs are dictated by the FDA to examine safety questions; some designs are used to establish new uses or indications.

PHASE II CLINICAL DEVELOPMENT

Among all the phases of clinical development, phase II is one of the most important components^[4] before entering into the phase III trials. Phase III pivotal studies are typically large scale with very high cost (the cost in some of phase III studies could be even more than the total cost

spent from the very beginning of a compound discovery to the end of phase II studies). Therefore, a well-designed phase II study could be very beneficial to both patients and sponsors and accordingly reduce the time-to-market launch and/or increase the probability of success. Phase II is also the last opportunity to stop a bad (or not-so-good) candidate from moving further into phase III.

Phase II trials are designed to study the preliminary efficacy, explore a range of doses for phase III, and understand the safety profile of an investigational product. Unlike phase I studies (except for life-threatening diseases), subjects recruited in phase II studies are patients with the disease for which the product is developed. Response variables considered in phase II studies are mainly clinical efficacy variables and safety variables. For example, in a trial for the evaluation of hypertension (high blood pressure), efficacy variables are blood pressure measurements. For an anti-infective trial, response variables can be the proportion of subjects cured or the time-to-cure for each subject. Phase II studies are mostly designed with parallel treatment groups (in contrast to crossover designs). Hence, if a patient is randomized to receive dose A of the test product, then this patient is planned to be treated with dose A of this product throughout the entire study.

Phase II trials are often designed to compare one or a few doses of a test product against a placebo control. These studies are usually shorter term (e.g., several weeks) and designed with a small or moderate sample size. Patients recruited for phase II trials are somewhat restrictive, that is, they tend to be with a certain disease severity (e.g., moderate severity), without other underlying diseases, and are not on background treatments. The two major objectives of phase II clinical trials are to demonstrate preliminary clinical benefit, which is typically referred to as a proof of concept (PoC) study, and to explore a range of doses for phase III confirmatory trials, which is generally called phase IIb dose-ranging study.

A typical PoC trial includes two treatment groups—a placebo group and a testing dose group. The testing dose is typically the MTD or a dose that is slightly below the MTD. A dose-ranging trial usually includes multiple doses of the testing product and placebo. One typical dose-ranging trial uses placebo, low dose, medium dose, and high dose of the testing product. In general, PoC trials and dose-ranging trials tend to be parallel designs with fixed doses. Some phase II trials can have both the PoC part and the dose-ranging part built in to one trial. In this case, PoC part will be demonstrated first, followed by the dose–response explorations. From a dose-ranging point of view, the objective of phase I should be to help exploring the upper limit of dose range (MTD), and that of phase II should be to help finding the lower limit of targeted dose range. This lower limit is also known as minimum effective dose.

Design a Phase IIa PoC Trial

One of the natural questions in phase II stage is whether or not the investigational drug works for the target patient population. In most clinical development programs for treating chronic diseases, phase II is the first time the candidate product can be tested in patients with the target disease under study. It is generally the first time the product efficacy can be observed and evaluated. In other words, before the first phase II clinical trial, the sponsor does not have an opportunity to study whether or not the test product works in patients. Under this circumstance, it is especially important that the product efficacy be established from the first phase II clinical trial. The most commonly used study design for the first phase II clinical trial is a PoC study. A PoC typically includes two treatment groups—a placebo group and a test product treatment group. In order to offer the best opportunity for the test product to demonstrate its efficacy, the dose of test product used in a PoC should be the highest possible dose. And this dose is usually the MTD, or a dose that is slightly below MTD. This design allows the project team to make a “go/no go” decision on the test product after study results are obtained.

If the results fail to demonstrate a clear efficacy, then a “no go” decision can be made for this candidate product and further development can be stopped. The sponsor does not have to make more investment on it; budget and resources can be allocated to develop other candidates that show more promise for success. This approach can also be considered as a “sledgehammer” approach, because such a design helps simplify the conclusion to be a “make it or break it” result for the product under development; that is, a clear “go/no go” decision should be made based on the analysis of data obtained from the PoC study.

Another type of study is known as a “proof of mechanism” (PoM) study. In the laboratories in which basic research and discovery are performed, it is important that a good understanding of the “mechanism of action” be available for the compound that is being discovered. For example, before the beta blockers were developed for the treatment of high systolic blood pressure (SBP), the biological theory was that the beta receptors in heart muscles may increase SBP as a response to stress stimulants. Given this understanding, the laboratory scientists searched for a chemical structure that could bind to the beta receptors and inhibits their activities. Such a chemical compound could be considered as a “beta blocker” because it blocks the activities of beta receptors in heart muscles.

As discussed earlier, product efficacy cannot be established until phase II. However, in the case of developing a beta blocker, there could be opportunities to observe whether the mechanism of action can be demonstrated in a phase I clinical trial. This can be achieved by observing the activities of beta receptors in healthy normal

volunteers. In other words, although healthy subjects with normal SBP are recruited in phase I, activities of beta receptors from these subjects can still be measured. By assessing beta receptor activities from these subjects, a PoM objective could potentially be achieved. Oftentimes, the measurement that helps establishing PoM is denoted as a biomarker. Generally, PoM is not a well-defined, widely accepted term that applies to all candidate products. Only under certain situations where the marker can be easily identified, and clearly measured, a PoM study design would be contemplated.

Impact of PoC Decisions

Most PoC studies tend to have two treatment groups—one placebo control group and another treatment group, with a high dose of the candidate product. Some PoC studies are designed with more than two treatment groups. Generally, the PoC designs are based on two key assumptions: 1) the MTD estimated from phase I is appropriate; and 2) efficacy dose–response relationship is monotonic.

Under these two assumptions, the highest possible dose (either MTD or a dose slightly below MTD) of the test product offers the best opportunity to demonstrate efficacy response. In other words, if this dose does not show efficacy response, then there is no need to invest any more resources to develop this test product. Hence, only after the concept is proven, extended development activities can be funded and resources can be allocated to develop the compound under investigation.

A positive PoC result triggers many important development activities. A long-term toxicity study is one of those activities. Drug formulation and drug supply also depend upon a positive PoC. At early phase II, there is usually just enough drug supply to handle the first few dose-ranging studies. If the PoC result is positive, there will be more study medication needed for phase III clinical development and large amount of raw materials need to be ordered or manufactured. Moreover, depending on the dosage forms that may be necessary for additional phase II and phase III trials, various formulations of the candidate product will have to be prepared. On the other hand, if the concept is not proven, there is no need to stock up a huge pile of raw materials for the making of large quantities of the study drug or study biological compound.

Furthermore, additional bioavailability or bioequivalence (or both) studies may be necessary. In some situations, additional drug–drug interaction studies may also be needed. Therefore, a clear “go/no go” decision is critical from the PoC study results. When the study results are not clear (inconclusive), many of the important development activities will be delayed.

According to the clinical development plan for a study medication, if the phase III study needs to be of one-year duration, then (before the phase III study is started) a

one-year toxicity study on animals should be completed, with no safety findings. Usually such a one-year toxicity study will not be initiated unless the concept is proven from early phase II. In practice, if the “go/no go” decision is delayed because of inconclusiveness, then all these operations could be postponed and, eventually, this problem of inconclusiveness may delay the entire product development program.

How to Communicate Risks Associated with a PoC Study

Usually a statistician is involved in a study design when there is a need for sample size calculation. Sample size calculation is one of many essential contributions provided by a trial statistician. An experienced statistician will begin by understanding the primary end point and by probing its clinically meaningful difference, δ . Sometimes this quantity is readily available. However, in many cases, this could be the beginning of long discussions.

After the study completes, clinical data are analyzed and results are presented to the project team. The project team members and upper management evaluate these results and make a decision. A “go/no go” decision is easy if both statistical significance is observed and treatment difference exceeds δ , where δ is the postulated clinically meaningful difference. These results indicate a clear “go” decision. On the other hand, when the results are not statistically significant and the observed treatment difference is too small, then the team may make, oftentimes reluctantly, a “no go” decision. However, if the observed treatment difference is not too small, but failed to reach statistical significance, then the team is in a difficult position and not clear as to whether a “go” or a “no go” decision could be made with confidence. Under this circumstance, a situation of inconclusiveness prevails.

In many cases, when the observed p value is close to statistical significance, but not strictly statistically significant, the team tends to decide to move forward. However, further development could indicate this “go” decision was a mistake because the observed treatment difference was not large enough. These difficulties occurred because, at the time of designing PoC, a δ which is larger than the true desired treatment difference was used for various reasons. One typical case could be that such a δ was selected based on the available sample size; i.e., from the sample size and the given α and β , a treatment difference can be calculated. A larger δ was chosen so that such a PoC study was feasible under the given ethical concerns or budget constraints. However, after the PoC results are ready, although the observed treatment difference was actually smaller than the hypothesized δ , the project team may still make a “go” decision simply because the p value was either statistically significant or close to be significant.

On the other hand, there were situations where the decision was “no go” even when a statistical significance is observed, yet the clinically meaningful difference δ was not achieved. One example can be found when the PoC end point is different from the phase III end point. When this is the case, a lowest δ was selected for PoC and the decision rule was that, if the PoC results cannot deliver such a δ , then there is no hope for the test product to meet the phase III primary end point.

Under certain situations, the PoC study may have to be repeated so that a clear decision can be made. Inconclusiveness can be thought of, as the time period between the study results obtained and a definitive “go” or “no go” decision is made. During this period, the team could consider a “go” decision, even when the observed p is greater than α . On the other hand, if the observed p is less than α , but the observed treatment difference is less than δ , but a “no go” decision was made incorrectly (β could be inflated). In either case, the risk of making a wrong decision is greater than the prespecified error rate.

Based on the above discussion, it is clear that when there is an observed statistical significance, and the decision is “go”, then the risk of making a wrong decision is preserved under the prespecified level alpha (α). Also, when statistical significance is not observed, and the decision is “no go”, then the risk of making a wrong decision is also preserved under the prespecified level of beta (β). However, if a “go” decision is made when there is no statistical significance, then the risk of making a wrong “go” decision is increased to be more than the prespecified risk level (α). On the other end, when there is a statistical significance and a “no go” decision is made, then the risk of making a wrong “no go” decision is increased to be more than the prespecified β . Of course, these decisions are usually made after long discussions among team members after looking at all clinical data on hand—primary and secondary end points, multiple doses if more than one dose was used, safety data, sometimes PK or outcomes research data, and so on.

The main issue we hope to address in this entry is during the time between when the PoC data is ready and the actual decision is made, there is a period of time that the PoC study was considered as “inconclusive.” This time period is difficult because there is always tight timelines for the entire project—Will there be a need for a long-term toxicity study? When should it start? What amount of drug supply will be necessary? At what time frame (phase II requirements and phase III requirements)? What dosages are to be formulated? Will there be additional PK studies such as drug–drug interaction or food effect studies to be designed? At this time, a “go/no go” decision is very critical. Yet, the data are not always clear. Regardless of all these discussions and various perspectives, in the end, a decision has to be taken—to go or not to go. This problem becomes much more difficult when the PoC study needs to be repeated.

Although the design of a PoC study is considered as a “sledgehammer” approach, in practice, there could be many difficulties and challenges in decision-making after the study is completed. It is critical to consider that in the design stage of such a study, there should be sufficient discussions among team members with the objective of minimizing the risk of potential inconclusiveness.

Design a Phase IIb Dose-Ranging Trial

Once a PoC study is positive, the next question is “what is the appropriate range of doses that could deliver desired treatment benefit to patients with acceptable safety profiles?” A dose-ranging trial typically includes a placebo (representing zero dose) and several doses of the test product. One very common design is a four group study with placebo, low dose, medium dose, and high dose groups of the drug candidate under development. In practice, usually the first dose-ranging study covers a wide dose range, and then the next dose-ranging study will be designed with a narrower dose range and tease out the target doses to be assessed in phase III.^[5]

In general, the most difficult challenge in designing the first dose-ranging trial in clinical development of a new medicinal product is the selection of doses to be included in this design. Ideally, for the first dose-ranging study design, it is more important to cover a wide dose range, instead of simply adding more doses to a narrow range of doses. This is both practical and scientific. Basically speaking, a trial with four to five test doses, plus a placebo control will deliver a good understanding of where the test medication is most active, if the dose range is wide enough, and the dose spacing is reasonable.^[5]

The dose-ranging study is designed with two very important assumptions: 1) the MTD was correctly estimated from earlier trials; and 2) efficacy response is monotonic to ascending doses. However, it is unclear as to what is the range of doses and where the efficacy activities would be. At this stage, the best guidance available is data obtained from preclinical experiments or from phase I PK/PD studies. Even under the assumptions that the MTD was correctly estimated, and there is a monotonic efficacy dose–response relationship, there can be many possible shapes of dose–response curves.

It is common that a dose–response model is used in analyzing dose-ranging trials. Sometimes, the Multiple comparison procedures and modeling (MCP-Mod) is also applied to help making inferences from clinical trial data. However, from the study design point of view, use of a model could be based on additional assumptions. It is well-known that “All models are wrong, some are useful”. Designing dose-ranging trials based on a wrong model, or wrong parameters, could potentially lead to very undesirable consequences. In data analysis, after the blind is broken, efficacy dose–response data are observed. At this point, use of a model that “fits the data well” is a reasonable thing to do.

IMPORTANCE OF PHASE II CLINICAL DEVELOPMENT

An efficient development program is to progress a good candidate as fast as possible, so that patients will benefit from it at the earliest possible date. Meanwhile, the program should allow a bad candidate to be eliminated as early as possible so that the sponsor will not have to waste additional investments or resources on it.

Based on experiences accumulated, the best implementation of an efficient development strategy would be “design to stop.” In other words, design every study with a clear objective that if the candidate is not good enough, then stop developing it. With the concept of “design to stop” for every study design, it offers the opportunity to stop developing candidates with low likelihood of being successful at the earliest stage, saving investments and resources so that the entire development portfolio can be made more efficient. Under this strategy, a natural question would be, if every study follows the “design to stop” principle, how about those very good candidates? In fact, if a candidate is really good, it will demonstrate its strength given a reasonable decent robust design. Hence, a “design to stop” strategy tends to be one of the most efficient strategies of clinical development for new medicinal products.

One example of “design to stop” is that in most of the development programs, the first phase II clinical trial is a PoC trial. Such a design allows the candidate the best opportunity, if a non-decreasing dose–response trend holds, to demonstrate its efficacy by using the highest allowable dose and to compare it with an agent known to be ineffective (placebo). If under this setting, the candidate still cannot show efficacy, the sponsor would be comfortable to stop developing it. This paradigm is a common example of the thinking behind “design to stop.” In other words, this strategy allows an unsuccessful candidate to be eliminated at the earliest stage.

In contrast to “design to stop,” another strategy would be “design to progress.” If the sponsor has much confidence on the candidate being developed, and decided to design a PoC study with an active control, to allow the experimental candidate to be compared with an active control intervention that is already approved for the indication being studied; the active agent could be a standard intervention and one of the market leaders. If the candidate is really strong, it could be superior to the active comparator, and the sponsor would then progress this candidate aggressively. The downside of such a design is that if the candidate is non-superior to the active control, then it is not known whether the candidate works in some way. Given this example, it helps to clarify what is meant by “design to progress,” which is a contrast strategy to “design to stop.” Statistically, the idea of “design to stop” is implemented by a clear primary statistical hypothesis in comparing against placebo with a strict control on the level of significance (α).

The two major types of phase II clinical trials are PoC studies and dose-ranging studies. A dose-ranging study can be thought of both as a confirmatory study and as an exploratory study. It can include statistical hypotheses with multiple comparison adjustments to ensure α protection. It can also include modeling or estimating dose–response relationship, where responses at a given dose or within a range of doses could be obtained.

One of the most important objectives of the phase II clinical development program is to help propose particular doses or a range of doses to help with the designs in the pivotal phase III trials. It is important to note that in product approval, for a given dose of the study product, if it is safe and efficacious at that dose, then that dose can be approvable. Hence, after the product is approved and prescribed for the general patient population, it will be very difficult to study a dose of this product which is lower than the doses studied in phase II or higher than a dose that is evaluated at phase III. In other words, phase II is the critical time of exploring a wide range of doses. If this dose range is not studied in phase II, then the opportunity of learning dose–response relationship for that range of doses is lost.

For a product candidate to enter phase III clinical development, it is a huge sponsor commitment to move the candidate forward. A typical phase III program includes much larger scale, longer term clinical trials that require tens or hundreds of million dollars of investments. In the United States, the FDA encourages sponsor to set an “end of phase II” meeting with FDA. End of phase II and beginning of phase III are very critical from a sponsor’s point of view. Because this is the last milestone point for a sponsor to stop developing a product that may not be deemed “successful” (here, successful is considered as both a regulatory success and a commercial success; i.e., the product can be approved and marketed so that development investments can be returned). Note that a product that is not very successful is different from a failed product. A sponsor needs to evaluate the likelihood of a candidate being successful in terms of the huge amount of investments for the entire development program. There are many examples where a candidate failed after phase III—it did not receive any approval even when some of the phase III trial results were beneficial and noteworthy. There are also many examples where products were successfully approved by regulatory agencies but failed to generate meaningful returns of investments. Therefore, for any given candidate at the stage between late phase II and early phase III, the sponsor needs to evaluate all evidence obtained from non-clinical and phase I and phase II clinical studies to project how likely this candidate can be eventually successful. Based on these critical thinking and benefit/risk evaluations, a sponsor makes the very important “go/no go” decision on the candidate being developed to determine whether the candidate can be progressed to phase III.

Given the fact that phase II is the stage where a product candidate first demonstrates its activity in patients with the disease, and helps a sponsor in making the “go/no go” decision of progressing the candidate to phase III, phase II can be thought to be the most important phase across the entire product development process.

REFERENCES

1. Morrow, T. Defining the difference: What makes biologics unique. *Biotechnol. Healthcare* **2004**, *1*, 24–29.
2. Ting, N. *Drug Development, Encyclopedia of Biopharmaceutical Statistics*. 2nd Ed.; Marcel Dekker: New York, 2003, 317–324.
3. Ting, N. *Introduction and Drug Development Process, Dose Finding in Drug Development*; Springer: New York, 2006, 1–17.
4. Ting, N. 2 clinical development in treating chronic diseases. *Drug Inf. J.* **2011**, *45*, 431–442.
5. Yuan, G.; Ting, N. *First Dose Ranging Clinical Trial: More Doses? or A Wider Range, Clinical Trial Biostatistics and Biopharmaceutical Applications*; CRC Press: London, 2015, 55–74.

Phase III Clinical Trials with Response-Adaptive Allocation

Atanu Biswas

Indian Statistical Institute, Kolkata, India

Rahul Bhattacharyya

University of Calcutta, Kolkata, India

Abstract

Response-adaptive allocation, usually considered in a phase III clinical trial, is advocated to satisfy the ethical requirement of skewing the allocation in favor of the better treatment arm. We discuss the logistics and models of the designs in two frameworks—one using response and allocation history and the other where the allocation probabilities are estimated functions of parameters of response distributions. We then discuss several important issues like covariate, longitudinal, and multivariate data; Bayesian designs; and real-life applications of response-adaptive designs.

Keywords: Delayed response; Drop-the-loser; Link function-based allocation; Longitudinal response; Optimal target; Play-the-winner; Randomized play-the-winner; Sequential maximum likelihood estimates; Skewing allocation; Urn model.

INTRODUCTION

Response-adaptive allocation, usually considered in a phase III clinical trial, is advocated to satisfy the ethical requirement of skewing the allocation in favor of the better treatment arm. We discuss the logistics and models of the designs in two frameworks—one using response and allocation history and the other where the allocation probabilities are the estimated functions of parameters of response distributions. We further discuss several important issues like covariate, longitudinal, and multivariate data; Bayesian designs; sample size calculations; and real-life applications of response-adaptive designs.

WHAT IS RESPONSE-ADAPTIVE ALLOCATION?

Clinical trials are the scientific experiments involving human beings to evaluate the performance of the new drugs or therapies (referred to as treatments) over the existing. Evaluation of new treatments requires recruitment and allocation of subjects to the treatments under consideration. But, prior to the actual experiment, the clinician is at the state of *equipose* and tends to assign equal number of subjects to each treatment arm for possible evaluation of the treatments. Again, among the treatments under consideration, there is superior as well as inferior treatments and allocation without considering the treatment effectiveness outweighs the ethical imperative of providing the best possible care for individual patients within the trial. A clinical trial usually takes months or years to recruit

the patients. However, if the responses are immediate or reasonably quick to obtain, the clinician can repeatedly evaluate the performance of the treatments based on the currently available information, i.e., response, allocation, and covariate history of the already-assigned patients together with the current patient's covariate information for the allocation of the next incoming patient. Treatment allocation procedures for the incoming subjects based on the accrued information so far are known as *adaptive allocation designs* or simply *adaptive designs*. Depending on the nature of the information used for adjustment, treatment allocation procedures are either treatment-adaptive (if only treatment allocation information is used for adjustment) or response-adaptive (if allocation and response information is used for adjustment). In a response-adaptive allocation, coupled with the current patient's covariate information, the available covariate information is used, and thus the allocation is referred to as covariate-adjusted response-adaptive (CARA).

Thus, the key element of any response-adaptive allocation is the sequentially adjusted allocation probabilities. A non-trivial (i.e., other than 0 and 1) allocation probability ensures concealment of the allocation and hence reduces selection bias, but trivial extreme allocation probabilities make the allocation deterministic.

However, the usefulness of a response-adaptive allocation can be best explained if we consider redesigning a real data set. Consider a data set comparing azidothymidine (AZT) and placebo to reduce maternal to infant HIV transmission.^[1] In the trial, following an equal allocation procedure, out of a total of 476 patients, 238 pregnant

women were assigned to each treatment arm. A total of 60 newborns were found to be HIV positive in the placebo group and only 20 were found to be HIV positive in the AZT group. Zelen and Wei^[1] illustrated that a randomized play-the-winner (PW) rule (a response-adaptive design, to be described later on) could allocate in a 300:176 ratio with a standard deviation (SD) 20. In Table 1, we provide an illustration of such different response-adaptive designs for binary responses with the aid of this data set to give an indication of the benefits out of different standard designs, where we present the expected allocation ratio, its SD, and the expected number of failures for different allocation designs. These illustrate the benefits from response-adaptive designs.

Similarly, we consider the data set from a trial of fluoxetine hydrochloride.^[2] The response is the negative of reduction in the HAM-D¹⁷ score, which is in the 52-point scale. In the trial, the patients were classified according to the rapid eye movement latency (REML) status. However, we consider only the patients with shortened REML (42 in total) for our purpose and, treating the responses to be continuous, we calculate the final mean scores, which are 10.67 (with SD = 5.50) for fluoxetine and 5.53 (with SD = 7.75) for placebo. Treating these as the true parameter values of the response distributions, we redesigned the trial using different response-adaptive designs for continuous responses. The expected allocation ratio together with its SD, and the expected total responses are provided in Table 1, along with some other response-adaptive designs (described later on) to illustrate the promise of response-adaptive designs. The figures of Table 2 tell the same story of assigning more patients to the better treatment.

Table 1 Redesigning AZT trial using response-adaptive designs

Design	Expected allocation ratio	SD	Expected/actual number of failures
Observed	238:238	—	80
PW	356:122	21	60
RPW	328:148	53	65
DL	357:119	19	64
RSIHR	249:127	11	78

Table 2 Redesigning fluoxetine trial using response-adaptive designs

Design	Expected allocation ratio	SD	Total expected/actual responses
Observed	21:21	—	356
BB	30:12	5	400
N-BB	30:12	5	400
CDL (c = 0)	25:17	3	370
BM (c = 0)	28:14	5	391

RESPONSE AND ALLOCATION HISTORY-BASED ALLOCATIONS

Binary Responses

Play-the-Winner (PW) Allocation

Consider a clinical trial involving two treatments (say treatments 1 and 2) and assume that the responses are binary (i.e., of success/failure type) and immediate. The first significant contribution on sequential designing a trial in a response-adaptive manner is the PW rule of Zelen.^[3] A PW allocation assigns an incoming subject based on the outcome of the previous patient's response; if the previous patient's response is a success, the same treatment is assigned to the next patient, but other treatment is assigned for a failure. For example, if the previous patient was assigned to treatment 1 and a failure is observed, PW assigns the next subject to the opposite treatment (i.e., treatment 2) with probability one.

Randomized PW (RPW) Allocation

For the PW procedure, given the allocation-and-response history of the previous patients, the allocation probabilities are either 0 or 1, and hence there is selection bias. Wei and Durham^[4] incorporated treatment unpredictability in the PW allocation using all available response and allocation information from all the previously allocated patients. The allocation procedure is described by an urn model with two types of balls representing two treatments. Initially, the urn contains α balls of each type. For the allocation of an incoming subject, a ball is drawn with replacement and the corresponding treatment is assigned. If the patient's response is a success, additional β balls of the same type are added to the urn; and if the patient's response is a failure, β balls of the opposite type are added. The rationale behind the allocation is to enhance the chance of assigning the treatment *doing better*. The resulting allocation is known as the randomized PW or RPW (α, β) rule. For example, in an RPW(2,1) rule, the first patient is assigned to either of the treatments with equal probability. Suppose that the first patient is assigned to treatment 1 and a success is observed. Then, a treatment 1 type ball is added to the urn keeping treatment 2 type balls unchanged at two. Thus, the urn now contains a total of $(4 + 1) = 5$ balls, and hence the chance of selecting a treatment 1 type ball is $\frac{2+1}{5} = \frac{3}{5}$, which is nothing but the assignment probability of the second subject to treatment 1. However, the assignment probability for the second subject to treatment 1 would be $\frac{2}{5}$, if a failure is observed for the first patient.

However, the number of treatments under consideration may be more than two and the corresponding extensions with mathematical details can be found in further works.^[5,6] The RPW rule is a particular case of the Generalized

Friedman's Urn.^[5] Also, *delayed response* for the RPW rule has been studied by Bandyopadhyay and Biswas.^[7]

Birth and Death Urns

If an extra ball is added to the urn for a success and nothing is done for a failure, we get randomized Pólya urn^[8] (RPU). On the other hand, if a ball is added for a success and removed for a failure, a birth and death process can be generated. However, removal of balls enhances the chance of exhausting balls of a treatment type and hence the concept of immigration balls are introduced. Specifically, apart from treatment type balls, the urn contains immigration balls with number equal to the number of treatments. If an immigration type ball is drawn, no patient is allocated, and the immigration ball is replaced together with additional balls, one ball of each treatment type. The procedure is repeated until a treatment type ball is drawn. Again, if we allow to remove a ball for failure, do nothing following a success and introduce immigration balls, we get drop-the-loser (DL) allocation.^[9] Embedding these designs into a continuous time stochastic process, different properties are investigated.

If N_{kn} assignments to treatment $k = 1, 2$ are observed out of a total of n assignments in a clinical trial, the observed allocation proportion $\frac{N_{kn}}{n}$ to treatment k is an indicative measure of the ethical capacity of the allocation. Assuming $p_k (= 1 - q_k)$ as the success probability for treatment k , we provide the long run value of $\frac{N_{kn}}{n}$ for different allocation designs in Table 3.

It is easily observed that these ratios exceed $\frac{1}{2}$, if treatment 1 is better than treatment 2 and vary from $\frac{1}{2}$ according to the degree of superiority of treatment 1 (except for RPU). Interestingly, the limiting allocation proportions for PW, RPW, and DL rules are the same and are proportional to $\frac{1}{q_i}$; however, their small sample performances differ. Thus, more allocation is expected in the long run if the treatment has the least failure rate.

Categorical Responses

Often the responses in a clinical trial are measured in an ordinal categorical scale like nil, mild, moderate, severe,

etc. Clubbing the categories, one can transform the data to binary and can run the urn-based procedures for allocation. But, clubbing the categories results in loss of information and hence allocation designs for categorical responses are more sensible. Bandyopadhyay and Biswas^[6] upgraded the two treatment RPW considering $\ell + 1$ ordered categories $0, 1, \dots, \ell$ for responses. The allocation starts with an urn containing α balls of each type. If treatment k is assigned and the response is j , an additional $j\beta$ balls of type k along with $(\ell - j)\beta$ balls of the opposite kind are added to the urn. A subsequent update within the framework of DL rule with high skewing ability and less variability is also available.^[14]

Continuous Responses

The urn model-based allocations require binary treatment responses, but the responses may be continuous in practice. Ivanova, Biswas, and Lurie^[10] developed a continuous version of the DL rule (referred to as CDL) considering discretization, which was further modified^[11] by introducing the idea of fractional balls to reduce variability of the observed allocation proportion. Another urn design for continuous responses can be found in the work of Bandyopadhyay and Biswas^[12] (referred to as N-BB, N stands for non-parametric), where a Wilcoxon–Mann–Whitney-type statistic is used to update the urn. They obtained the limiting allocation proportion to treatment 1 as $P(X_1 > X_2)$, where the continuous response variable for treatment k is denoted by X_k .

ESTIMATED PARAMETER-BASED ALLOCATION USING LINK FUNCTIONS

The central idea behind an estimation-based allocation is the specification of a target allocation function. Suppose we have t treatments and the response distribution to treatment k involves the parameter θ_k . Then a target allocation function for treatment k is a function $\rho_k(\theta_1, \dots, \theta_t) \in [0, 1]$ such that:

$$\sum_{k=1}^t \rho_k(\theta_1, \dots, \theta_t) = 1 \quad (1)$$

for all $\theta_k, k = 1, \dots, t$. The choice of ρ_k is motivated by either ethical issue or precision or a convenient combination of both.

Link Function–Based Design for Continuous Responses

For continuous responses, the idea of link function-based allocation was developed through *treatment effect mapping*^[13] in a non-parametric setup. For two treatment trials, the idea is to assign the $(i + 1)$ st incoming subject to

Table 3 Limiting allocation proportion (LAP) to treatment 1

Design	LAP to treatment 1
Equal	$\frac{1}{2}$
RPU	$I(p_1 > p_2)$
BDU/BDUI	$\frac{\frac{1}{p_1 - q_1}}{\sum_{i=1}^2 \frac{1}{p_i - q_i}}$ if $\frac{1}{2} > p_1 > p_2$
PW/RPW/DL	$\frac{\frac{1}{q_1}}{\sum_{i=1}^2 \frac{1}{q_i}}$

treatment 1 with probability $g(\hat{\Delta}_i)$, where $\hat{\Delta}_i$ is the available estimate of some treatment effect measure Δ and $g(\cdot)$ is a function on $[0,1]$ ensuring ethical norms. Rosenberger^[13] continued the development with $g(x) = \frac{1+x}{2}$ and $\hat{\Delta}_i$ as the available estimate of the centered and scaled linear rank statistic. For survival outcomes, when some observations are subject to censoring, Rosenberger and Seshaiyer^[14] developed a response-adaptive allocation considering g as above but with $\hat{\Delta}_i$ as the centered and scaled log-rank test statistic.

An important link function-based allocation can be found in Bandyopadhyay and Biswas^[15] (referred to as BB), where an allocation is developed for normal linear model-based responses. In particular, if the response distribution of treatment k is $N(\mu_k, \sigma_k^2)$, the allocation probability for the $(i+1)^{\text{st}}$ patient to treatment 1 is given by $G(\hat{\mu}_{1i} - \hat{\mu}_{2i})$, where $\hat{\mu}_{ki}$ is the estimate of μ_k based on the data from the first i patients, and $G(\cdot)$ is the cumulative distribution function (cdf) of a random variable which is symmetric about 0. Bandyopadhyay and Biswas^[15] suggested to take:

$$G(\hat{\mu}_{1i} - \hat{\mu}_{2i}) = \Phi\left(\frac{\hat{\mu}_{1i} - \hat{\mu}_{2i}}{T}\right) \quad (2)$$

for some tuning parameter $T (> 0)$, reflecting the information on response variability. In fact, one can take $T = \sqrt{\hat{\sigma}_{1i}^2 + \hat{\sigma}_{2i}^2}$ in case of unequal variances. This design is based on ethics; the allocation probability is 50% for no (estimated) treatment difference and gets skewed in favor of the treatment producing higher observed mean response (i.e., the treatment doing better) at any stage. Suppose, at any stage, the estimates of the parameters based on the available data so far are $\hat{\mu}_1 = 10.67$, $\hat{\mu}_2 = 5.53$, $\hat{\sigma}_1 = 5.50$, and $\hat{\sigma}_2 = 7.75$. Then, the probability of allocating the next patient to treatment 1 is:

$$\Phi\left(\frac{10.67 - 5.53}{\sqrt{5.50^2 + 7.75^2}}\right) = 0.7057 \quad (3)$$

In fact, this will be the allocation probability to treatment 1 for the 43rd patient in the shortened REML group.

Ethics-Based Target Allocation Functions

Ethics is the main driving force in assigning a larger fraction of subjects to the better treatment. Thus, if we maintain a higher allocation probability to the better treatment, ethical requirements are fulfilled. In particular, assume that the unknown parameters θ_k , $k = 1, \dots, t$, are related to the location of the response distributions and higher responses are desirable. Then, increasing values of θ_k indicate the increasing superiority of the k^{th} treatment and the inequality $\theta_k > \theta_s$ signifies the superiority of treatment k over treatment s . Ideally, an ethics-based target $\rho_k(\theta_1, \dots, \theta_t)$ must be increasing in θ_k for fixed θ_s , $s(\neq k) = 1, \dots, t$. Moreover, the ordering $\theta_k > \theta_s$ must imply

the corresponding ordering of target allocation functions, that is, $\rho_k(\theta_1, \dots, \theta_t) > \rho_s(\theta_1, \dots, \theta_t)$.

For example, consider binary responses with success rate θ_k for treatment k , and the odds ratio-based target:^[16]

$$\rho_k(\theta_1, \dots, \theta_t) = \frac{\frac{\theta_k}{1-\theta_k}}{\sum_{j=1}^t \frac{\theta_j}{1-\theta_j}}, k = 1, \dots, t \quad (4)$$

Then, under the equality of success rates, $\rho_k = \frac{1}{t}$ for all $k = 1, \dots, t$, and ρ_k satisfies the other desirable properties.

If a higher response is desirable, treatment k is regarded as the *most promising* if it is capable of producing larger responses more frequently than its competitors or equivalent, if X_k is stochastically larger than any X_s ($s \neq k$). Thus, the relative effectiveness measure $P(X_k > \max_{s \neq k} X_s)$ can be taken as the target allocation function ρ_k for treatment k to provide a unique target allocation function using the knowledge of the response distributions.^[17] However, if the response for the k^{th} treatment is normal with mean θ_k and variance σ_k^2 in a two treatment trial, target allocation for treatment 1 reduces to:

$$\rho_1 = P(X_1 > X_2) = \Phi\left(\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) \quad (5)$$

as the same target function developed in Eq. 2.

Target Allocation Functions Combining Ethics and Precision Measures

Ethics and precision usually act in the opposite directions and for a meaningful combination, we primarily need relevant measures of ethics and precision. Assume that n_{kv} subjects are assigned to treatment k ($= 1, \dots, t$). Then, an ethical metric can be expressed as $G_E(n_1, \dots, n_t) = n_1 \Psi_1 + \dots + n_t \Psi_t$, where Ψ_k measures the ethical capacity of treatment k . For example, in binary response trials, $\Psi_k = 1 - \theta_k$ results in $G_E(n_1, \dots, n_t)$ as the total number of failures. Again, in a continuous response trial, if a lower response is favorable, the clinician might want to minimize the total expected responses.

$$G_E(n_1, \dots, n_t) = \sum_{k=1}^t E(X_k) \quad (6)$$

On the other hand, for a measure of precision $G_V(n_1, \dots, n_t)$, one can take some suitable functions of the Fisher's information matrix like $|I^{-1}(\theta | n_1, \dots, n_t)|$ or $\text{Trace}(I^{-1}(\theta | n_1, \dots, n_t))$. Thus, the optimum proportion $\frac{n_k}{n_1 + \dots + n_t}$, $k = 1, \dots, t$, compromising the ethical and the statistical needs, could be derived from the following constrained optimization problem:

$$\text{Minimize } G_E(n_1, \dots, n_t) \quad (7)$$

subject to the restriction:

$$G_v(n_1, \dots, n_t) \leq \kappa \quad (8)$$

where κ is a constant. Depending on the situation, there may be more than one restrictions. Equivalently,^[18] the optimum proportions can be derived by minimizing weighted objective function $wG_E(n_1, \dots, n_t) + (1-w)G_v(n_1, \dots, n_t)$ for $w \in (0,1)$. However, often we are interested in some linear combinations $A^T\theta$ of the parameters. For example, in binary response trials, the parametric functions of interest are often the differences $\theta_i - \theta_t, i=1, \dots, t-1$ with treatment t as the “control” treatment. In such a situation, it is natural to take $G_v(n_1, \dots, n_t)$ as some function of $A^T\Gamma^{-1}(\theta|n_1, \dots, n_t)A$. Considering the above choice, Atkinson and Biswas^[19,20] suggested another way to combine ethics and optimality to develop response-adaptive D_A -optimal designs. However, the problem becomes complicated when multiple treatments are involved. However, an alternative formulation^[21] considers the non-centrality parameter of the large sample contrast test of homogeneity for testing $H_0: \theta_1 - \theta_t = \theta_2 - \theta_t = \dots = \theta_{t-1} - \theta_t = 0$ as a precision measure. Since the power function of such a test is increasing in the non-centrality parameter $\phi(n_1, \dots, n_t)$, it is suggested^[27] to take the constraints $\phi(n_1, \dots, n_t) \geq \kappa$ and $\frac{n_j}{\sum_{k=1}^t n_k} \geq B, j=1, \dots, t$, where $\kappa (>0)$ is a constant and $B \in [0, t^{-1}]$ instead of $G_v(n_1, \dots, n_t) \leq \kappa$.

As examples, consider a two treatment binary response trial with success probability θ_k for treatment k . If we take $\psi_k = 1 - \theta_k$ and $G_v(n_1, n_2) = \text{Trace}(\Gamma^{-1}(\theta|n_1, n_2)) = \frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}$, then the optimal proportion to treatment k is the well-known RSIHR^[22] target:

$$\rho_k = \left(\frac{n_k}{n_1 + n_2} \right)_{\text{opt}} = \frac{\sqrt{\theta_k}}{\sqrt{\theta_1} + \sqrt{\theta_2}}, \quad k=1,2 \quad (9)$$

However, these authors interpreted $G_v(n_1, n_2)$ as the variance of the estimated treatment difference.

Our next example is based on continuous responses, where we assume that a lower response is desirable. Then any response exceeding a clinically significant threshold c could be defined as a failure and hence $G_E(n_1, n_2)$ with $\psi_k = P(X_k > c)$ could measure the total number of failures for continuous response trials. Considering the variance of the estimated treatment difference or equivalently the trace of the inverse of the information matrix as the measure of precision, we can get the optimal allocation function of Biswas and Mandal^[23] (referred to as BM).

Implementing the Target Allocation Function

The target allocation functions, discussed so far, involve unknown parameters and hence cannot be applied straight a way in practice. For practical implementation, a doubly

adaptive biased coin design^[24,25] (DBCD) can be used. For two treatments, if the prefixed allocation function to treatment 1 is ρ_1 , then the allocation probability of the $(j+1)^{\text{st}}$ subject to treatment 1 is $g^{(\gamma)}\left(\frac{n_{1j}}{j}, \hat{\rho}_{1j}\right)$ with $g^{(\gamma)}(x, \rho_1) \in [0,1]$, defined by:

$$g^{(\gamma)}(x, \rho_1) = \begin{cases} 1 & \text{if } x = 0 \\ \frac{\rho_1 \left(\frac{\rho_1}{x}\right)^\gamma}{\rho_1 \left(\frac{\rho_1}{x}\right)^\gamma + (1-\rho_1) \left(\frac{1-\rho_1}{1-x}\right)^\gamma} & \text{if } 0 < x < 1 \\ 0 & \text{if } x = 1 \end{cases} \quad (10)$$

where $\gamma (\geq 0)$ is a randomness controlling parameter. Under certain restrictions^[25] on g , DBCD achieves the desired target ρ_1 in the long run. For $\gamma = 0$, DBCD reduces to the Sequential Maximum Likelihood Procedure^[22], where $g^{(\gamma)}(x, \rho_1) = \rho_1$, and at each stage maximum likelihood estimates of the unknown parameters are plugged in the expression of ρ_1 . The subsequent extension to accommodate multiple treatments is also developed.^[21]

DIFFERENT ISSUES

Covariates

Covariates are the auxiliary information associated with the study subjects, influencing their responses. For an enumeration of the importance of covariates, we consider the results from the Stroke Prevention in Atrial Fibrillation (SPAF) trial.^[26] The randomized trial compares two treatments, namely, *aspirin* and *placebo*, in the presence of a single binary covariate. The covariate is an indicator of the anticoagulation status (whether receiving or not). SPAF trial resulted in a significant difference between *aspirin* and *placebo* in reducing the number of strokes among patients receiving anticoagulation. However, no worthwhile difference was observed among the patients without receiving the anticoagulation therapy and this illustrates that the treatments perform differently for patients with different covariate profile. As a result, if we ignore the interaction between the treatments and the covariate, *aspirin* would be prescribed for every patient to reduce stroke. Therefore, the randomization probability of an incoming subject must include the covariate information, apart from the other available information.

One of the attempts can be found in the study by Bandyopadhyay and Biswas,^[27] where the RPW allocation is modified to incorporate categorical covariate information. Like the original RPW allocation, the modified version rewards the successful treatment with the addition of more balls but the number depends on the covariate category of the patient. In particular, following a success, for a patient with unfavorable covariate category, they allowed addition of more balls of the same kind and less

balls of opposite kind. However, for the same response with a favorable covariate category, number of balls to be added is lower than that for the unfavorable covariate category. Assuming the same, a covariate-adjusted DL allocation is also developed^[28] and studied in detail. For linear model-based responses, one can use the covariate-adjusted estimates $\hat{\mu}_{1i} - \hat{\mu}_{2i}$ to develop a covariate adjusted version of Eq. 2. A CARA design for a general class of continuous responses can be found in the study by Biswas and Bhattacharya.^[29]

Longitudinal Responses

Quite often, the patients in a clinical trial are monitored over a period of time and repeated responses (i.e., longitudinal responses) are obtained. For example, after treatment, the patient may be examined twice a week over a period of two weeks for the assessment of disease status, which gives a sequence of upto four correlated responses. For such responses, Biswas and Coad,^[30] Biswas and Dewanji^[31,32] extended the popular RPW allocation considering two treatments and longitudinal binary responses. Like the RPW (α, β) procedure, the allocation for longitudinal responses starts with an urn containing α balls of two types. If, at some monitoring time, recurrence (i.e., failure) is observed for a patient, β balls of the opposite treatment type are added to the urn whereas for a non-recurrence (i.e., success), β balls of the same type are added. As an example, consider the allocation of the third incoming patient. If the first patient is assigned to treatment 1 and has two successes and one failure so far, and the second patient has one failure so far on treatment 2, then the urn contains $\alpha + 3\beta$ balls of type 1, and $\alpha + \beta$ balls of type 2. Thus, the allocation probability to treatment 1 for the third patient is simply $\frac{\alpha+3\beta}{2\alpha+4\beta}$. However, the first two patients will give more responses subsequently and the urn will be updated accordingly each time. For further development in the presence of covariate information' refer the study by Biswas, Bhattacharya, and Park.^[33]

Multivariate Responses

Responses obtained from the treated subjects can be multivariate in nature. Quite often, safety and efficacy are the two components of a bivariate response. Development of a response-adaptive allocation in such cases requires the utilization of the multivariate response history for any assignment. Biswas and Coad^[30] generalized the link function-based design (Eq. 2) considering multivariate responses and multiple treatment situations.

Missing Data

Possibility of missing responses is a rarely addressed issue in the context of response-adaptive designs. One exception

can be found in the study by Biswas and Rao,^[34] where an allocation design is developed for the normal linear model-based responses assuming missing-at-random responses and complete availability of the covariate information. Specifically, they suggest to impute the missing responses at each stage and considering the available (i.e., observed and imputed) response and covariates; estimates of the treatment difference are calculated and plugged into Eq. 2 to get the allocation probability of the next incoming subject.

Bayesian Designs

Response-adaptive designs are data-driven, and hence it has similarity with the Bayesian philosophy of prior utilization. One of the response-adaptive designs developed in a Bayesian paradigm can be found in the study by Biswas and Angers,^[35] where the allocation probability at each stage is the predictive density of the allocation indicator of the incoming patient. Assuming two treatments and normal linear model-based responses, they developed an allocation function considering covariates.

Real-Life Applications

Real-life applications of response-adaptive designs are very limited. The initial RPW allocation-based Michigan extracorporeal membrane oxygenation (ECMO) trial^[36] was a disaster mostly due to the small sample size and inefficient choice of design parameter. The Boston ECMO trial^[37] was another application of RPW rule, whereas crystalloid preload trial^[38] was an application of PW rule. Eli Lilly sponsored Fluoxetine trial^[2] was again an RPW-based application. Biswas and Dewanji^[31,32] illustrated a real clinical trial with rheumatoid arthritis patients for longitudinal responses. However, the authors feel that more coordination between the practitioners and the academicians is needed in order to have more and more successful response-adaptive allocations in practice.

Power and Variability

Response-adaptive allocations usually favors the better performing treatment arm for further allocation. But such an allocation induces correlation among the treatment assignments and causes, in general, a loss in statistical power. However, under fairly general conditions and large sample sizes, it can be established that loss in power is inversely proportional to the variability of the observed treatment assignments.^[39] Therefore, among a number of response-adaptive designs targeting the same allocation proportion, the design with the minimum variability seems best. A class of such allocations, referred to as *asymptotically best*, can be found depending on the attainment of the

lower bound on the asymptotic variance of the observed allocation proportion.^[40]

Sample Size Calculation

The primary objective in most of the clinical trials is to explore the safety and effectiveness of the treatments under study. Then it is natural to have sufficient number of patients in the study to establish the effects of the treatments with certain statistical precision. Therefore, sample size calculation at the planning stage is necessary to ensure detection of a clinically significant difference with high probability. Usually sample size calculation depends on the specification of a significance level (say, α), power of a relevant test (say, $1 - \beta$), and a clinically significant difference (say, Δ). As a simple example, assume that the response variable (may be continuous, binary or time-to-event) for treatment k has unknown mean μ_k and known variance $\sigma_k^2, k = 1, 2$. For specified Δ , consider testing $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 - \mu_2 = \Delta (> 0)$ at a significance level α . If a right-tailed test is used to ensure power $1 - \beta$, then an allocation procedure targeting ρ in the limit, requires:

$$n = \left(\frac{\sigma_1^2}{\rho} + \frac{\sigma_2^2}{1-\rho} \right) \left(\frac{\tau_\alpha + \tau_\beta}{\Delta} \right)^2 \quad (11)$$

observations approximately to detect the difference Δ (referred to as type I sample size^[40]). Thus, n is inversely proportional to $|\Delta|$, that is, as the difference increases, more samples are needed to maintain the power requirement. But Eq. 11 ignores the actual randomization adopted in the trial and hence can be far from the actual number. However, a number of sample size formulae considering the actual randomization can be found in the study by Hu and Rosenberger.^[40] In practice, σ_1 and σ_2 are unknown and need to be replaced by their estimates. As an example consider the RSIHR allocation targeting $\rho = \frac{\sqrt{\mu_1}}{\sqrt{\mu_1} + \sqrt{\mu_2}}$. Then, we need approximately 201 observations for a test of $H_0: \mu_1 = \mu_2 = 0.4$ against $H_1: \mu_1 = 0.6, \mu_2 = 0.4$, when $\alpha = 0.05$ and $\beta = 0.10$.

CONCLUSION

This entry provides a very brief overview of the logistics and references of response-adaptive designs in some important directions. Naturally, many important aspects and details are not included in this entry. Booklength discussions^[40–42] on response-adaptive designs are a good source of knowledge on the topic. However, Atkinson and Biswas^[42] provided a detailed description of the development, logistics, and applications of almost all types of response-adaptive designs. Interested readers can go through any of these books for further exposure on this topic.

REFERENCES

1. Zelen, M.; Wei, L.J. Foreword. In *Adaptive Designs*; Flournoy, N.; Rosenberger, W.F., Eds.; Institute of Mathematical Statistics: Hayward, 1995, i–iii.
2. Tamura, R.; Faries, D.; Andersen, J.; Heiligenstein, J. A case study of an adaptive clinical trial in the treatment of out-patients with depressive disorder. *J. Am. Stat. Assoc.* **1994**, *89* (427), 768–776.
3. Zelen, M. Play-the-winner rule and the controlled clinical trial. *J. Am. Stat. Assoc.* **1969**, *64* (325), 131–146.
4. Wei, L.J.; Durham, S. The randomized play-the-winner rule in medical trials. *J. Am. Stat. Assoc.* **1978**, *73* (364), 838–843.
5. Wei, L.J. The generalized Polya's urn for sequential medical trials. *Ann. Stat.* **1979**, *7* (2), 291–296.
6. Bandyopadhyay, U.; Biswas, A. Selection procedures in multi-treatment clinical trials. *Metron*. **2002**, *LX* (1–2), 143–157.
7. Bandyopadhyay, U.; Biswas, A. Delayed response in randomized play-the-winner rule: A decision theoretic outlook. *Calcutta Stat. Assoc.* **1996**, *46*, 69–88.
8. Durham, S.D.; Flournoy, N.; Li, W. Sequential designs for maximizing the probability of a favorable response. *Can. J. Stat.* **1998**, *26* (3), 479–495.
9. Ivanova, A. A play-the-winner-type urn design with reduced variability. *Metrika*. **2003**, *58* (1), 1–13.
10. Ivanova, A.; Biswas, A.; Lurie, H. Response adaptive designs for continuous outcomes. *J. Stat. Plan. Infer.* **2006**, *136* (6), 1845–1852.
11. Biswas, A.; Bhattacharya, R. Reduction of variability of response-adaptive designs for continuous treatment responses in phase III clinical trials. *Drug Inf. J.* **2010**, *44* (1), 9–19.
12. Bandyopadhyay, U.; Biswas, A. An adaptive allocation for continuous response using Wilcoxon-Mann-Whitney score. *J. Stat. Plan. Infer.* **2004**, *123* (2), 207–224.
13. Rosenberger, W.F. Asymptotic inference with response-adaptive treatment allocation designs. *Ann. Stat.* **1993**, *21* (4), 2098–2107.
14. Rosenberger, W.F.; Seshaiyer, P. Adaptive survival trials. *J. Biopharm. Stat.* **1997**, *7* (4), 617–624.
15. Bandyopadhyay, U.; Biswas, A. Adaptive designs for normal responses with prognostic factors. *Biometrika*. **2001**, *88* (2), 409–419.
16. Basak, G.K.; Biswas, A.; Volkov, S. An urn model for odds ratio based adaptive phase III clinical trials. *J. Biopharm. Stat.* **2009**, *19* (5), 838–856.
17. Bandyopadhyay, U.; Bhattacharya, R. An invariant allocation function for multi-treatment clinical trials. *Stat. Interf.* **2016**, *9* (1), 1–10.
18. Cook, R.D.; Wong, W.K. On the equivalence of constrained and compound optimal designs. *J. Am. Stat. Assoc.* **1994**, *89* (426), 687–692.
19. Atkinson, A.C.; Biswas, A. Bayesian adaptive biased-coin designs for clinical trials with normal responses. *Biometrics*. **2005**, *61* (1), 118–125.
20. Atkinson, A.C.; Biswas, A. Adaptive biased-coin designs for skewing the allocation proportion in clinical trials with normal responses. *Stat. Med.* **2005**, *24* (16), 2477–2492.

21. Tymofeyev, Y.; Rosenberger, W.F.; Hu, F. Implementing optimal allocation in sequential binary response experiments. *J. Am. Stat. Assoc.* **2007**, *102* (477), 224–234.
22. Rosenberger, W.F.; Stallard, N.; Ivanova, A.; Harper, C.N.; Ricks, M.L. Optimal adaptive designs for binary response trials. *Biometrics*. **2001**, *57* (3), 173–177.
23. Biswas, A.; Mandal, S. Optimal adaptive designs in phase III clinical trials for continuous responses with covariates. In *mODa 7—Advances in Model-Oriented Design and Analysis*; Bucchianico, A.D.; Lauter, H.; Wynn, H.P., Eds.; Physica-Verlag: Heidelberg, 2004, 51–58.
24. Eisele, J.; Woodroffe, M. Central limit theorems for doubly adaptive biased coin designs. *Ann. Stat.* **1995**, *23* (1), 234–254.
25. Hu, F.; Zhang, L.X. Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *Ann. Stat.* **2004**, *32* (1), 268–301.
26. Hart, R.G.; Halperin, J.L.; Pearce, L.A.; Anderson, D.C.; Kronmal, R.A.; McBride, R.; Nasco, E.; Sherman, D.G.; Talbert, R.L.; Marler, J.R. Lessons from the stroke prevention in atrial fibrillation trials. *Ann. Intern. Med.* **2003**, *138* (10), 831–838.
27. Bandyopadhyay, U.; Biswas, A. Allocation by randomized play-the-winner rule in the presence of prognostic factors. *Sankhya Series B.* **1999**, *61* (3), 397–412.
28. Bandyopadhyay, U.; Biswas, A.; Bhattacharya, R. Drop-the-loser design in the presence of covariates. *Metrika* **2009**, *69* (1), 1–15.
29. Biswas, A.; Bhattacharya, R. A covariate adjusted response adaptive allocation for a general class of continuous responses. *J. Stat. Theory Pract.* **2016**, *10* (4), 852–863.
30. Biswas, A.; Coad, D.S. A general multi-treatment adaptive design for multivariate responses. *Sequent. Anal.* **2005**, *24* (2), 139–158.
31. Biswas, A.; Dewanji, A. A Randomized longitudinal play-the-winner design for repeated binary data. *Aust. NZ. J. Stat.* **2004**, *46* (4), 675–684.
32. Biswas, A.; Dewanji, A. Inference for a RPW-type clinical trial with repeated monitoring for the treatment of rheumatoid arthritis. *Biometrical J.* **2004**, *46* (6), 769–779.
33. Biswas, A.; Park, E.; Bhattacharya, R. Covariate-adjusted response-adaptive designs for longitudinal treatment responses: PEMF trial revisited. *Stat. Methods Med. Res.* **2012**, *21* (4), 379–392.
34. Biswas, A.; Rao, J.N.K. Missing responses in adaptive allocation design. *Stat. Probabil. Lett.* **2004**, *70* (1), 59–70.
35. Biswas, A.; Angers, J.F. A Bayesian adaptive design in clinical trials for continuous responses. *Stat. Neerl.* **2002**, *56* (4), 400–414.
36. Bartlett, R.H.; Roloff, D.W.; Cornell, R.G.; Andrews, A.F.; Dillon, P.W.; Zwischenberger, J.B. Extracorporeal circulation in neonatal respiratory failure: A prospective randomized trial. *Pediatrics*. **1985**, *76* (4), 479–487.
37. Ware, J.H. Investigating therapies of potentially great benefit: ECMO. *Stat. Sci.* **1989**, *4* (4), 298–340.
38. Rout, C.C.; Rocke, D.A.; Levin, J.; Gouws, E.; Reddy, D. A reevaluation of the role of crystalloid preload in the prevention of hypotension associated with spinal anesthesia for elective cesarean section. *Anesthesiology*. **1993**, *79* (2), 262–269.
39. Hu, F.; Rosenberger, W.F. Optimality, variability, power: Evaluating response-adaptive randomization procedures for treatment comparisons. *J. Am. Stat. Assoc.* **2003**, *98* (463), 671–678.
40. Hu, F.; Rosenberger, W.F. *The Theory of Response-Adaptive Randomization in Clinical Trials*; John Wiley & Sons, Inc.: Hoboken, NJ, 2006.
41. Baldi Antognini, A.; Giovagnoli, A. *Adaptive Designs for Sequential Treatment Allocation*; CRC Press: Boca Raton, FL, 2015.
42. Atkinson, A.C.; Biswas, A. *Randomised Response-Adaptive Designs in Clinical Trials*; CRC Press: Boca Raton, FL, 2014.

Placebo Effect

Thomas R. Stiger

Pfizer Inc., Groton, Connecticut, U.S.A.

Edward F. C. Pun

Agouron Pharmaceuticals, Inc., San Diego, California, U.S.A.

Abstract

Prior to the advent of modern medicine, many medications prescribed as a part of medical practice were subsequently found to be pharmacologically inert. This entry examines how despite this fact, people felt they were being helped by the drug, an occurrence known as the placebo effect. The entry also discusses the role of this effect in modern clinical trials.

INTRODUCTION

Prior to the advent of modern medicine, many medications prescribed as a part of medical practice were subsequently found to be pharmacologically inert. Despite this fact, such inactive medication, often referred to as a “placebo”—which literally means “I shall please”—appeared to be efficacious for a large number of medical problems. Shapiro^[1] has suggested that most medical practice prior to the 17th century was just an exploitation of this “placebo effect.” This is reflected in Mothby’s *New Medical Dictionary* written in 1785, which defined placebo as “a commonplace method or medicine.”^[2] Toward the end of the 18th century, when the medical community was starting to doubt the effectiveness of some commonly prescribed medications, the term placebo was viewed more derisively. For example, the 1811 edition of Harper’s *Medical Dictionary* defined placebo as “an epithet given to any medium adopted more to please than to benefit the patient.”^[3]

OVERVIEW

In the last 50 years or so, much has been learned and written about the construct known as the placebo effect. Much of this large body of knowledge about the effects of placebo has been gleaned from the use of placebo as a control in clinical trials. A response to placebo has been reported for diseases involving nearly all organ systems within the body. Satisfactory response to placebo has been reported to be at least 30% or more for a variety of disorders and often as high as 70%.^[4] This had led some in the medical community to recommend placebo as a legitimate form of treatment.^[5] Despite the amount of research on this phenomenon and the extensive reports of responses to placebo, there still is not a good understanding about the mechanism and causes of the placebo effect. In fact, there is still some

confusion and a lack of consensus on what constitutes the placebo and a placebo response.

Beecher^[6] defined the placebo as a pharmacologically inert substance. This is consistent with the definition given in Stedman’s *Medical Dictionary*, 26th edition,^[7] which appears as follows: (1) “an inert substance given as a medicine for its suggestive effect,” (2) “an inert compound identical in appearance to material being tested in experimental research, which may or may not be known to the physician and/or patient, administered to distinguish between drug action and suggestive effect of the material under study.” In a pharmacological setting, it is common to expand the first definition by equating an inert substance to all factors other than the active pharmacological agent. However, this definition implies that any form of specific nonpharmacological intervention, such as psychotherapy, is a form of placebo.

In order to avoid such erroneous classification, more expanded definitions of placebo attempt to differentiate between specific actions of an agent and its nonspecific effects, the latter being defined as placebos. Shapiro^[8] proposed a broad definition of placebo that includes nonpharmacological cases as well. He defines the placebo as any therapy (or component of therapy) that is deliberately or knowingly used for its nonspecific, psychological, or psychophysiological effect, or that is used unknowingly for its presumed or believed effect on a patient, symptom, or illness, but which, unknown to a patient and therapist, is without specific activity for the condition being treated. This definition encompasses the use of placebo as a control in clinical trials as well as a treatment in medical practice. However, classifying a therapy’s effect as nonspecific or specific is often difficult in practice, particularly when the therapy is nonpharmacological. This has led some authors to conclude that one cannot define placebo in a logically consistent way.^[9] Still others have proposed more elaborate and lengthy definitions that have attempted to avoid the alleged contradictions inherent in the specific and nonspecific distinction.^[10]

As a result of various definitions of placebo, there is also much lack of agreement on what a placebo effect is and what it is not. Some examples of definitions of the placebo effect are the following: (1) the improvement that occurs in a patient who is being treated with placebo;^[11] (2) a significant therapeutic response to “inert” or “placebo” substances;^[12] and (3) an effect that is produced by a medical procedure’s therapeutic intent and not its specific nature.^[13] What some authors consider factors or causes of the placebo effect others claim to be alternative effects that are not part of the placebo response. Because the focus of this entry is on pharmacological agents in clinical trials, placebo will be defined as simply an inert pill or injection masking as an active agent; correspondingly, the placebo effect will be defined as a therapeutic response to the treatment with placebo. Whereas items such as spontaneous remission (i.e., natural course of healing) and regression toward the mean would generally not be considered part of the placebo effect, we will consider such phenomena because they can play a key role in the observed improvement of placebo-treated subjects.

REGULATORY REQUIREMENTS

Before a new drug can be approved for use in nearly any country throughout the world, it must be demonstrated that it is safe and effective. For nearly 30 years, the Food and Drug Administration (FDA), through the Federal Register, has mandated that new drugs must be evaluated via randomized, blinded, controlled clinical trials. Similar requirements have been adopted by many other Western countries. To gain approval in the United States, a drug must be shown to be efficacious, usually in at least two adequate and well-controlled trials. Whereas the guidelines do not specifically mandate the use of placebo as a control, placebo usually is the standard for comparison. To demonstrate efficacy in a placebo-controlled trial, one has to show superiority to placebo; that is, the drug effect can add more to the underlying placebo effect. In many studies in which the placebo group has improved substantially, even drugs that are known to be efficacious do not show statistical superiority over placebo. For example, sertraline, a widely used efficacious drug for major depressive disorder (MDD) was found to be statistically superior to placebo in just two of the five placebo-controlled trials that were conducted in MDD as part of the approval process.^[14] In the last 17 years, there have been only two anxiolytics approved by the FDA. It is speculated that it is just too difficult to demonstrate superiority to placebo in generalized anxiety disorder, where the placebo response rate is notoriously high. Hence, it is assumed that if one could reduce the placebo response, it would be easier to demonstrate that an effective agent is superior to placebo. To that end, one first needs to consider the mechanisms that are associated with the placebo response.

FACTORS INFLUENCING THE PLACEBO RESPONSE

Factors and mechanisms associated with the placebo response can be crudely classified into two types:^[12] environmental factors and patient factors. The former includes the clinical setting, the attitude of the medical care personnel, and the expectation of the principle investigator. The latter refers to patient history, patient personality traits, and the patient’s experience and expectations. Environmental and patient factors are related, and their interaction in the form of a doctor–patient relationship can create a very important social psychological phenomenon that is thought to be highly influential on the placebo effect.

Clinical settings and medical practitioners that convey empathy, warmth, interest, and understanding are generally thought to contribute to nonspecific positive therapeutic response, especially in disorders in which the measures of response to treatment are more subjective. If the clinician has an enthusiastic attitude and a positive expectation, this is likely to increase the response to placebo. On the other hand, a practitioner or experimenter that shows aloofness, curttness, and pessimism will create an environment that will probably diminish the placebo effect.

Patient factors include personality, history, and demographic variables. Unlike the environmental factors, these items are generally not controllable. It has been stated by several researchers^[15]—with varying degrees of evidence—that patients that are more acquiescent and suggestible are more likely to respond to placebo. The patient’s expectations and attitude about the treatment, which are shaped to a large degree by a confluence of his or her personality and history, also are believed to be associated with the degree of placebo response. For example, patients who want to be in control are felt to be more likely to respond to placebo because the act of taking a pill is an action that will make them feel in charge. Intuitively, patients that have had favorable experiences with medication are thought to be more responsive to placebo. In contrast, patients that are highly skeptical of much of modern medicine or generally shun conventional medical treatment have been found to be placebo nonresponders.^[16] Conditioning, which can be influenced by experience and expectation, has also been implicated as a cause of the placebo effect. For example, airway resistance^[17] and gastric contractibility^[18] have been reported to be influenced by placebo. Other physiological variables that may be influenced by classical conditioning, and thus potentially be affected by placebo, include heart rate, blood pressure, EKG patterns, vasomotor activity, and the immune response.^[19,20] It has been reported that placebo has been found to produce endorphins in pain studies,^[21–23] although the validity of this finding has been challenged.^[4,24,25]

Bush^[16] has categorized factors that influence the placebo effect into two classes: “bound” and “manipulable.” Bound factors depend on the history and personality of the patient

and clinicians and are viewed as generally unalterable. Manipulable factors, which include most of the environmental factors discussed in the previous paragraph, are viewed as alterable to a large degree. Further examples of these factors include the shape, color, and characteristics of the active and placebo agents. Shapira et al.^[25] studied the use of oxazepam in either a green, yellow, or red tablet in 48 patients; they reported that green tablets tended to be more effective for anxiety, whereas yellow pills seemed to more effective for depressive symptoms. Blackwell et al.^[26] suggested that multiple pills and pills of larger size produced a larger placebo effect. Thomson^[27] insinuated that the “active placebo” (one that has similar side effects to the active agent it is to mimic) atropine amplified the placebo response in tricyclic-antidepressant trials.

STATISTICAL ISSUES

Factors Associated with the Placebo Response

Improvement in response to placebo does not imply that placebo or associated nonspecific effects have generated the therapeutic benefit. For example, improvement in some cases is the result of spontaneous remission. One very important factor that probably accounts for a large degree of the improvement in the placebo group in clinical trials is the phenomenon of regression toward the mean, which was first described by Galton.^[28] “Regression toward the mean” is the tendency for an extreme variable value to move closer to the mean when the variable is remeasured. McDonald and Mazzuca^[29] argued that much of the improvement observed in the placebo group in clinical trials is attributed to regression toward the mean. To explain regression toward the mean mathematically, let the random variables Y_1 and Y_2 represent some response variable for the same individual at two separate time points. Further, for the sake of simplicity, assume that the joint distribution of Y_1 and Y_2 is bivariate normal with a common mean μ , common variance σ^2 , and correlation coefficient ρ . Then

$$E(Y_2 - Y_1 | Y_1 = y_1) = (\mu - y_1)(1 - \rho)$$

Note that if $y_1 > \mu$, then $E(Y_2 - Y_1) < 0$; moreover, the expected decrease from the first measure to the second is proportional to $|\mu - y_1|$, i.e., to how extreme the first value is. This expected decrease (when $y_1 > \mu$) is also proportional to $1 - \rho$; hence one would expect a larger decrease as the correlation becomes less positive (repeated measures on the same individual are almost always positively correlated). As a result, outcome variables that are less reliable—or reproducible—will be more prone to the effects of regression toward the mean.

Because detailed information concerning μ , σ^2 , and ρ for many laboratory variables is available in the

medical literature, McDonald and Mazzuca^[29] calculated the expected change in a repeat measure when the first measure was greater than $\mu + 3\sigma$. For 15 laboratory variables, they found the expected decrease ranged from 2.5% to 26%, which agreed well with empirical data from 12,000 patients obtained via hospital medical records.

Since entrance criteria for clinical trials are usually based on the subject's presenting either a high or low value at baseline, regression toward the mean is likely to explain a large degree of the improvement observed in clinical trials, particularly when the response variable is fairly subjective and not highly reliable.

When the outcome variable is rather subjective, observer bias is another factor that could potentially contribute to the improvement attributed to placebo. This factor has been largely ignored in the plethora of literature about the placebo response. Indications of observer bias appear in the data from clinical trials that incorporate a single-blind placebo run-in phase (this design feature will be discussed further in the next section) in order to screen for “placebo responders” prior to randomization into the double-blind phase of the study.^[30] It has been the authors' experience, based on many placebo-controlled trials using placebo run-ins, that the degree of improvement on the average (or the percentage of “responders”) during this initial phase is generally far less than that experienced by the patients randomized to placebo during a comparable period of time within the double-blind portion of the study. Whereas this observation could in part be due to regression toward the mean, there may be an upward bias in the measurements (i.e., more abnormal) at the end of the single-blind placebo phase of the trial for two reasons: First, the observer or investigator knows the subjects are receiving placebo; second, there may be incentive for the investigative site to reach their recruitment goals. On the other hand, during the double-blind portion, there may be a bias toward more improved measures or ratings because the investigator may speculate that the subject is on an effective medication and hence is expecting an improvement.

Design Considerations

One design feature frequently incorporated into clinical trials (e.g., hypertension and psychopharmacology trials), as already mentioned, is the use of a single-blind placebo run-in. One of the main purposes of this run-in phase is to screen out subjects that show a substantial improvement with placebo. Despite the lack of critical review of such a design, a placebo run-in is almost always used in depression studies. The few investigations that have attempted to assess this design have generally not found the placebo run-in to be effective at reducing the placebo response rate.^[30] Trivedi and Rush^[31] conducted a meta-analysis on 101 blinded placebo-controlled depression studies that were conducted from 1959 to 1990. After cross-classifying the trials by the type of active drug (i.e., tricyclics,

monoamine oxidase inhibitors, and serotonin reuptake inhibitors) and whether they were outpatient or inpatient, they found few statistically significant differences in the placebo response rates between trials that had a run-in and those that did not.

Regression toward the mean is one reason that the placebo run-in appears not to be effective at reducing the response to placebo. Subjects are still entered into the study on the basis of having an extreme baseline value for the response of interest. In the absence of other reasons for extreme baseline values, such as observational bias, one approach to minimize the effects of regression toward the mean would be to take multiple measurements during a screening period. These measures could then be averaged to obtain a baseline value. Another approach would be to take two “baseline” measurements (that would perhaps be at least within a week of each other): The first would be used for determining eligibility (i.e., the subjects would have to meet some criterion for disease severity) and the other used as the baseline measurement for the purpose of calculating change.

As previously implied, observational bias may be another reason that single-blind placebo run-ins may not be useful at reducing the placebo effect. In studies in which the primary outcome variable is subjective, one possible design that would address observational bias would be one that had a random-length double-blind placebo run. One could perceive of many variations of such a design. One possibility for comparing one active agent to placebo is as follows: After a 1-week single-blind placebo period, subjects still meeting the eligibility criterion (i.e., an extreme value for the outcome of interest) would be randomized in double-blind fashion to placebo (95% of the subjects) or active drug (5% of the subjects). After a 1-week period (and a clinical assessment), half of the placebo subjects would be randomized to placebo while the other half continues to receive placebo. The primary analysis for efficacy would include only those subjects randomized at week 2 that still continued to meet the entrance criterion. Week 2 would be considered baseline. One could also conduct a secondary analysis that would include all randomized subjects (week 1 would be considered baseline for the 5% of the subjects randomized to active drug at week 1; week 1 might also be considered to be the baseline for all other subjects in such a secondary analysis). This design, as well as variations of a random-length placebo period, would probably be even more effective at reducing observational bias if the precise nature of the run-in and the data analysis were not completely spelled out in the protocol. Although this may cause some ethical and logistic concerns, it might be sufficient just to state in the protocol that a randomization to the two treatments will occur over several weeks. The details of the planned analysis could be contained in another document, separate from the procedures of the protocol.

There is yet another way to reduce observational bias in outcomes that are measured on some type of calibrated instrument (e.g., sphygmomanometer, spectrophotometer, scales for weight): Recalibrate in such a way that one randomly sets the “zero mark.” For instance, a diastolic blood pressure of 80 mm Hg would appear as 100 on an sphygmomanometer in which zero was set at 20. Such tactics would be most useful for reducing observational bias in single-blind studies.

ATTEMPTS AT PREDICTING PLACEBO RESPONSE IN DEPRESSION

In much of clinical research, particularly in psychopharmacology, there is a perception that the placebo response is becoming more pronounced,^[32] which has led many researchers to believe that it is becoming increasingly more difficult to demonstrate the efficacy of new medications. By the nature of clinical trials, usually it is implicitly assumed that the placebo effect and the therapeutic effect of an active medication are additive, but this may not be the case in many instances, particularly when the placebo effect is large and there is a ceiling effect on the degree of improvement. While there has been a large amount published about qualitatively describing “placebo responders,” only in the last decade or so has there been much effort at attempting to predict placebo response on an individual basis based on screening information.^[33–39] If one could predict placebo response based on demographic or medical history information, then one could exclude likely “placebo responders” prior to entry into the study. Most of these efforts at finding predictors of placebo response in psychiatric disorders have entailed retrospective, exploratory analyses of small data sets comprised of one to a few studies. As a result, there has been a smattering of reports of “statistically significant” differences between “placebo responders” and nonresponders from one investigation to another but limited agreement among the reports. Following are some of the variables that have been cited from two or more published studies as being related to placebo response in depression. The likelihood of being a placebo responder increases as (1) the duration of the current episode decreases, (2) the degree of precipitating stress increases, and (3) the number of previous treatments for depression decreases.

In unpublished research conducted by these authors and colleagues, an artificial neural network (ANN) was trained (using a data set containing patients from four depression studies) to predict which subjects—based on some of the variables listed previously as well as others—would respond to placebo (experience a 50% or more decrease in the Hamilton Depression Rating Scale). Whereas the ANN seemed to have some success for predicting placebo response, it did not appear to enhance the difference between the active drug and placebo. In fact, removing

the predicted “placebo responders” from the analysis of a validation data set decreased the average improvement of the active drug more than that of the placebo group. This effect has also been reported by Reimherr et al.^[30] In depression, this suggests that “placebo responders” also show a large response to active drug and that removal of such subjects leaves a population of subjects that are in general more treatment resistant. For a binary outcome, it has been speculated by Day^[40] that in some instances there may be an optimal level of placebo response for the purpose of detecting a difference between placebo and active drug. For example, the placebo response and the therapeutic effect due to active drug might be additive on a logistic scale. If the true odds ratio between active and placebo is 3.86, then increasing the placebo response from 0.1 to 0.2 results in an increase in the active drug response from 0.3 to 0.49. It is also possible in certain cases that the “placebo response” and the “therapeutic response” might be more than just additive on measurement scales that are commonly used for continuous outcomes.

REFERENCES

- Shapiro, A.K. *Behav. Sci.* **1960**, 5, 398–430.
- The Placebo Response in Modern Perspectives in World Psychiatry*; Howells, J.G., Ed.; Oliver and Boyd: Edinburgh, 1968, Vol. 2.
- Stanford, R.L.; Schmitz, T.H. *Stat Pharmaceutical Industry*; Buncher, C.R., Tsay, J.Y., Eds.; Marcel Dekker: New York, 1981, 231–250.
- Kienle, G.S.; Keine, H. *Altern. Ther.* **1996**, 2, 39–54.
- Brown, W.A. *Neuropsychopharmacology* **1994**, 10, 265–288.
- Beecher, H.K. *JAMA* **1955**, 159, 1602–1606.
- Stedman's Medical Dictionary*, 26th Ed.; Spaycar, M., Ed.; Williams and Wilkins: Baltimore, MD, 1995.
- Shapiro, A.K. *J. Clin. Pharm.* March–April **1970**, 73–78.
- Götzsche, P.C. *Lancet* **1994**, 344, 925–926.
- Grünbaum, A. *Psychol. Med.* **1986**, 16, 19–38.
- Schweizer, E.; Rickels, K. *J. Clin. Psychol.* **1997**, 58 (Suppl. 11), 30–38.
- Piercy, M.A.; Sramek, J.J.; Kurtz, N.M.; Cutler, N.R. *Ann. Pharmacother.* **1996**, 30, 1013–1019.
- Liberman, R. *J. Chronic Dis.* **1962**, 15, 1–761.
- Summary Basis of Approval—Zoloft Tablets*; FDA, Dec. 1991.
- Cleophas, T.J.M. *JAMA* **1995**, 273, 1–283.
- Bush, P.J. *J. Am. Pharm. Assoc.* **1974**, 14, 671–674.
- Luparello, T.; Leist, N.; Lourie, C.H. *Psychosom. Med.* **1970**, 32, 509–513.
- Sternbach, R.A. *Psychophysiology* **1964**, 1, 67–72.
- McDonald, D.; Stern, J.; Hahn, W. *J. Psychosom. Res.* **1963**, 7, 97–106.
- Straus, J.L.; Cavanaugh, S.V.A. *Psychosomatics* **1996**, 37, 315–326.
- Heylin, M. *Chem. Eng. News* **1978**, 56, 1–24.
- Levine, J.D.; Gordon, N.C.; Fields, H.L. *Lancet* **1978**, 2, 54–657.
- Goldstein, A.; Grevert, P. *Lancet* Dec 23, **1978**, 30, 1–1385.
- Wall, P.D. *Ciba Found. Symp.* **1993**, 174, 211–817.
- Shapira, K.; McClelland, H.A.; Griffiths, N.R. *BMJ* **1970**, 2, 446–449.
- Blackwell, B.; Bloomfield, S.S.; Buncher, C.R. *Lancet* **1972**, 1, 1279–1282.
- Thomson, R. *Br. J. Psychiatry* **1982**, 140, 64–68.
- Galton, F. *J. Anthropol. Inst. G.B. Irel.* **1885–1886**, 15, 246–263.
- McDonald, C.J.; Mazzuca, S.A. *Stat. Med.* **1983**, 2, 417–427.
- Reimherr, F.W.; Ward, M.F.; Byerley, W.F. *Psychol. Res.* **1989**, 30, 191–199.
- Trivedi, M.H.; Rush, J. *Neuropsychopharmacology* **1994**, 11, 33–43.
- Uhlenhuth, E.H.; Matuzas, W.; Warner, T.D.; Thompson, P.M. *Psychopharm. Bull.* **1997**, 33, 31–39.
- Brown, W.A.; Dornseif, B.E.; Wernicke, J.F. *Psychol. Res.* **1988**, 26, 259–264.
- Brown, W.A. *Psychopharm. Bull.* **1988**, 24, 14–17.
- Woodin, K.E. *Positive Placebo Response in Clinical Trials of Depressed Outpatients*; Ph.D. Dissertation. Univ. of Massachusetts: Amherst, 1990.
- Khan, A.; Brown, W.A. *Psychopharm. Bull.* **1991**, 27, 271–274.
- Brown, W.A.; Johnson, M.F.; Chen, M. *Psychol. Res.* **1992**, 41, 203–214.
- Shear, M.K.; Leon, A.C.; Pollack, M.H.; Rosenbaum, J.F.; Keller, M.B. *Psychopharm. Bull.* **1995**, 31, 273–278.
- Taiminen, T.; Syvälahti, E.; Saarijärvi, S.; Niemi, H.; Lehto, H.; Ahola, V.; Salokangas, R. *J. Nerv. Ment. Dis.* **1996**, 184, 109–113.
- Day, S.J. *Stat. Med.* **1988**, 7, 1187–1194.

Population Bioequivalence

Naitee Ting

Pfizer Global Research and Development, New London, Connecticut, U.S.A.

Greg C. G. Wei

Pfizer Global Research and Development, Groton, Connecticut, U.S.A.

William W. B. Wang

Merck Research Laboratories, West Point, Pennsylvania, U.S.A.

Abstract

As more and more generic drug products become available in the market place, there are increasing concerns over whether generic drugs approved based on the current regulation of average bioequivalence (ABE) have the same quality, safety, and efficacy profiles as brand-name drug products. This entry provides several statistical models for determining population bioequivalence.

INTRODUCTION

Bioequivalence studies are conducted to demonstrate equivalence in the bioavailability of the active ingredient in different formulations. The U.S. Food and Drug Administration (FDA) requires pharmaceutical companies to show bioequivalence between different formulations or generic companies between generic drugs and brand drugs before approval. A recent FDA guidance on bioequivalence proposes criteria for assessment of population bioequivalence (PBE) and individual bioequivalence (IBE). This article introduces the background and methods for testing PBE.

In recent years, generic medicines have played an increasingly important role in the efforts to contain the costs of medicines. As more and more generic drug products become available in the market place, there are increasing concerns over whether generic drugs approved based on the current regulation of average bioequivalence (ABE) have the same quality, safety, and efficacy profiles as brand-name drug products.

OVERVIEW

In the U.S. 21 Code for Federal Regulation (CFR) 320.1, bioequivalence is formally defined as “the absence of a significant difference in the rate and extent to which the active ingredient or active moiety in pharmaceutical equivalents or pharmaceutical alternatives becomes available at the site of drug action when administered at the same molar dose under similar conditions in an appropriately designed study.” In 1992, in order to meet these provisions, the U.S. FDA issued an official guidance titled *Statistical Procedures for Bioequivalence Studies Using a Standard Two-Treatment Crossover Design*.^[1] According to this

guidance, in order to establish that two drugs are bioequivalent, it is only necessary to demonstrate that two drugs are similar in average bioavailability in terms of pharmacokinetic parameters such as AUC or $C_{\max}$. From a statistical point of view, this similarity in average bioavailability, often termed average bioequivalence, considers only the population means, but ignores the variability of pharmacokinetic parameters across different subject populations or across multiple measurements from the same subject. Hence, there are concerns that this approach may not guarantee that the two drugs are similar in all or even most individual subjects in the intended patient populations.^[2,3]

Anderson and Hauck^[4,5] first introduced the terms “population bioequivalence” and “individual bioequivalence” to address these concerns. Since then, extensive research results are available on the criteria as well as the associated procedures for establishing these kinds of bioequivalence (see, e.g., Refs. [6–8]). In 1995, the FDA formed a bioequivalence working group to investigate these issues. As a result of this effort, the FDA issued a preliminary draft guidance titled *In Vivo Bioequivalence Studies Based on Population and Individual Bioequivalence Approaches*.^[9] Based on public comments, the FDA issued a revised draft guidance titled *Average, Population, and Individual Approaches to Establishing Bioequivalence*.^[10] This draft guidance proposed mixed-scaled aggregate criteria for population bioequivalence and individual bioequivalence, with bootstrap-based^[9] and moment-based^[10] methods for testing of these criteria. In January 2001, the FDA issued a final guidance entitled *Statistical Approaches to Establishing Bioequivalence*.^[11] This guidance replaced the previous draft guidance and the 1992 FDA bioequivalence guidance. In addition, the FDA issued another draft guidance entitled *Bioavailability and Bioequivalence Studies for Nasal Aerosols and*

Nasal Sprays for Location Reaction,^[12] which also calls for population bioequivalence.

This encyclopedia entry focuses on population bioequivalence. Topics related to individual bioequivalence are addressed in a separate entry by Chow and Liu (see *Individual Bioequivalence*). The next section reviews and discusses the FDA 2001 guidance with respect to population bioequivalence. The section “Statistical Analysis” covers the statistical analysis for population bioequivalence. The last section provides general discussions and comments.

POPULATION BIOEQUIVALENCE REGULATORY REVIEW

Population Bioequivalence Concept

One of the key concerns on average bioequivalence is whether it can guarantee drug interchangeability. Population bioequivalence and individual bioequivalence address drug interchangeability under two different clinical situations. Population bioequivalence refers to the interchangeability for a patient who has not previously received either test or reference drug, and is about to receive one or the other, whereas individual bioequivalence refers to interchangeability for a patient who switches from a reference drug to a test drug.^[4]

To guarantee population bioequivalence of test and reference drugs, the population distribution of the pharmacokinetic measurements for test and reference drugs should be similar when these drugs are administered to new or drug-naïve patients. Under the normality assumption, not only the population average but also the overall variability of the pharmacokinetic measurements for test and reference drugs should be similar. To quantify the similarity, the FDA 2001 guidance adopted the notion of population difference ratio (PDR), which measures the ratio of the expected squared difference between test and reference drugs to the expected squared differences between two administrations of reference drugs, when these drugs are given to drug-naïve patients. The PDR can be written as:

$$\text{PDR} = \sqrt{\frac{E(T-R)^2}{E(R-R')^2}} \quad (1)$$

Conceptually, a test drug (T) is population bioequivalent to a reference drug (R) if the PDR is within an acceptable limit. Based on this important concept, the FDA population bioequivalence criterion (PBC) can be formulated and tested under different statistical models and study designs.

Statistical Model

Statistical analyses of bioequivalence data are typically based on a statistical model for the logarithm of the

pharmacokinetic measures (e.g., AUC and $C_{\max}$). Most of the statistical models proposed for bioequivalence assessment are linear mixed-effects models, which usually include fixed effects such as treatment, sequence, and period, as well as random effects of subjects. Under a conventional 2×2 crossover design, the following linear mixed-effects model or some variations of this model are typically used for bioequivalence analysis:^[4,13,14]

$$Y_{ijk} = \mu + s_{ij} + \pi_j + \tau_k + \varepsilon_{ijk} \quad (2)$$

$$i = 1, \dots, n; j = T, R; k = 1, 2$$

In Eq. 2, Y_{ijk} is the log-transformed pharmacokinetic response of the i th subject on drug j in the k th sequence, μ represents the overall mean, (s_{iT}, s_{iR}) is independently distributed normal random vector for subject i with mean

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and covariance matrix

$$\begin{pmatrix} \sigma_{BT}^2 & \rho\sigma_{BT}\sigma_{BR} \\ \rho\sigma_{BT}\sigma_{BR} & \sigma_{BR}^2 \end{pmatrix}$$

π_j is the fixed effect of the drug j (where $\pi_1 + \pi_2 = 0$), τ_k is the fixed effect for sequence (where $\tau_1 + \tau_2 = 0$), and ε_{ijk} are independent normal random variables with mean 0 and variance σ_{wj}^2 ($j = T, R$).

In this model, each subject has subject-specific means of $\mu + s_{iT} + \pi_T + \tau_k$ and $\mu + s_{iR} + \pi_R + \tau_k$ for the test and reference drugs, respectively. Ignoring the sequence effect, the model assumes that these subject-specific means come from a distribution with population means μ_T and μ_R with respective between-subjects variances of σ_{BT}^2 and σ_{BR}^2 . The total population variances for the test and reference drugs (σ_{TT}^2 and σ_{TR}^2 , respectively) are summations of between-subjects variances and within-subject variances (σ_{WT}^2 , σ_{WR}^2). Similar models can be used for repeated crossover design.

Bioequivalence Criterion

The 1992 and 2001 guidances recommend the following criterion for average bioequivalence:

$$-\theta_A \leq \mu_T - \mu_R \leq \theta_A$$

where the bioequivalence limit (θ_A) is usually set at $\ln(1.25)$.

Conceptually, as described in the section “Population Bioequivalence Concept,” the PDR should be controlled by a specified limit to guarantee population bioequivalence. Based on this concept, the 2001 guidance proposes the following mixed-scaling population bioequivalence criterion:

$$\text{PBC} = \frac{(\mu_T - \mu_R)^2 + (\sigma_{TT}^2 - \sigma_{TR}^2)}{\max(\sigma_{T0}^2, \sigma_{TR}^2)} < \theta_p \quad (3)$$

where σ_{T0}^2 is a specified scaling constant for total variance and θ_p is the population bioequivalence limit. This criterion is equivalent to the following linearized criteria:

$$\begin{aligned}\eta_1 &= (\mu_T - \mu_R)^2 + (\sigma_{TT}^2 - \sigma_{TR}^2) - \theta_p \sigma_{TR}^2 < 0, \\ &\text{for } \sigma_{TR} > \sigma_{T0} \quad \text{reference - scaling} \\ \eta_2 &= (\mu_T - \mu_R)^2 + (\sigma_{TT}^2 - \sigma_{TR}^2) - \theta_p \sigma_{T0}^2 < 0, \\ &\text{for } \sigma_{TR} \leq \sigma_{T0} \quad \text{constant - scaling}\end{aligned}\quad (4)$$

The mixed-scaling strategy is proposed in consideration of different drug variability and different therapeutic window. Reference scaling can effectively loosen the population bioequivalence criteria for more variable reference products (i.e., $\sigma_{TR} > \sigma_{T0}$). Constant scaling is recommended for drug products that have a wide therapeutic window and low variability (i.e., $\sigma_{TR} \leq \sigma_{T0}$), to avoid unnecessarily tightening its population bioequivalence criteria. An exception is narrow therapeutic window drugs or drug products, where reference scaling is recommended regardless of the reference variability.

Note that under reference scaling, PBC is monotonically related to PDR: $PDR = [(PBC/2) + 1]^{1/2}$. Based on the consideration of a maximum allowable PDR of 1.25, the 2001 guidance recommends that scaling constant of 0.2 (i.e., $\sigma_{T0} = 0.2$). The guidance also suggests that the population bioequivalence limit θ_p be chosen as

$$\theta_p = \frac{(\ln 1.25)^2 + \varepsilon_p}{\sigma_{T0}^2}$$

where the value of ε_p for population bioequivalence is guided by the consideration of the variance term ($\sigma_{TT}^2 - \sigma_{TR}^2$) added to the average bioequivalence criterion.

Study Design

For the purpose of establishing bioequivalence between a test (T) and a reference (R) drug, the FDA recommends a standard two-sequence two-period in vivo bioequivalence study design (TR/RT) in its July 1992 guidance on *Statistical Procedures for the Bioequivalence Studies Using a Standard Two-Treatment Crossover Design*.^[1] In the 2001 guidance, the FDA points out that conventional 2×2 crossover design, as well as replicated crossover design, can be used for testing of ABE and PBE. For example, the 2001 FDA guidance illustrates the testing for population bioequivalence under the four-period, two-sequence, two-drug study design could be TRTR/RTRT.

Use of crossover designs for bioequivalence studies allows each subject to serve as his or her own control to improve the precision of the comparison. However, one of the assumptions underlying this principle is that carryover effects are either absent or equal for both drugs. In most cases (for both replicated and conventional crossover designs), the inequality of carryover effects is considered

unlikely in a bioequivalence study. If sponsors suspect the carryover effects might be an issue, the 2001 guidance recommends the BE study with parallel design.^[11]

Sample Size Calculation

According to the FDA guidance,^[11] a minimum number of 12 subjects should be included in any bioequivalence study. When an average bioequivalence criterion is selected using either nonreplicated or replicated designs, published formulas can be used to determine the sample size for average bioequivalence.^[14] Because analytical approaches for population bioequivalence estimation are not available, sample size for the population bioequivalence should be estimated by simulation. The simulations should be conducted using a default situation allowing the two drugs to vary as much as 5% in average bioequivalence with equal variances. The study should have 80% or 90% power to conclude bioequivalence between these two drugs.^[11]

STATISTICAL ANALYSIS

General Background

For average bioequivalence, the 1992 guidance and the 2001 guidance recommend that statistical analysis for pharmacokinetic parameters be based on the Schuirman's two one-sided test procedure.^[15] To establish bioequivalence, the calculated 90% confidence interval for the ratio of product averages in the original scale should fall within 80–125%. Equivalently, the calculated 90% confidence interval for the difference between two product averages in log scale should fall within $\ln(0.8)$ and $\ln(1.25)$. Chow and Liu^[14] give an excellent overview of average bioequivalence.

The FDA 2001 guidance^[11] proposes a method based on moment estimators of variance components to evaluate population bioequivalence. Before method of moments was introduced, population bioequivalence was evaluated using bootstrap approach with restricted maximum likelihood (REML) estimators under a mixed-effects model. However, there are some concerns in using the bootstrap to determine whether a new drug provides population bioequivalence. One concern is that bootstrap does not provide a unique solution. For example, if a sponsor submits a package and claims population bioequivalent based on the sponsor's bootstrap data, FDA may run a bootstrap using a different seed. With the new bootstrap, the FDA finds that the submitted drug is not population bioequivalent. In this dilemma, how can the final decision be made?

Therefore, nonbootstrap methods are developed recently to evaluate population bioequivalence (see, e.g., Refs. [11,16,17]). Most of these methods are based on moment estimators, and they are confidence interval approaches using the modified large sample (MLS) concept.^[18]

As discussed earlier, FDA considers the use of bioequivalence criteria in Eq. 3 to determine population bioequivalence. Bioequivalence can be concluded if the null hypothesis (H_0) below is rejected at a given Type I error (α):

$$H_0 : \text{PBC} \geq \theta_p \quad \text{vs.} \quad H_a : \text{PBC} < \theta_p$$

Under this setting, the one-sided upper $(1 - \alpha) \times 100\%$ confidence bound U can be used to test the hypothesis—that is, if $U < \theta_p$, then population bioequivalence can be concluded. On the other hand, $U \geq \theta_p$, leads to accepting H_0 and population bioequivalence is rejected.

Based on model Eq. 2 under a 2×2 crossover design, the sums of squares and mean squares can be formulated as follows:

Source of Variation	SS	df	MS	E(MS)
Drug	$SS_D = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{i..})^2$	1	S_D^2	θ_D
Intrasubject	$SS_I = \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{.j.})^2$	$2(n-1)$	S_I^2	θ_I
Treatment R	$SS_{TR} = \sum_i (Y_{i11} - \bar{Y}_{.11})^2 + \sum_i (Y_{i12} - \bar{Y}_{.12})^2$	$2(n-1)$	S_{TR}^2	θ_{TR}
Treatment T	$SS_{TT} = \sum_i (Y_{i21} - \bar{Y}_{.21})^2 + \sum_i (Y_{i22} - \bar{Y}_{.22})^2$	$2(n-1)$	S_{TT}^2	θ_{TT}

In general, these θ_i s are referred to as variance components, SS denotes sum of squares, df stands for degrees of freedom, MS represents mean square, and E(MS) denotes expected value of mean square. It can be shown that SS_I/θ_I , SS_{TT}/θ_{TT} , SS_{TR}/θ_{TR} , are all χ^2 random variables with $2(n-1)$ degrees of freedom, and SS_D/θ_I is a noncentral χ^2 random variables with noncentrality parameter λ . The population bioequivalence criteria in Eq. 3 can then be expressed in terms of variance components (θ_i s) as

$$\begin{aligned} \text{PBC} &= \frac{(\theta_D - \theta_I)/n + \theta_{TT} - \theta_{TR}}{\theta_{TR}} \quad \text{if } \sigma_{TR}^2 > \sigma_{T0}^2 \\ \text{PBC} &= \frac{(\theta_D - \theta_I)/n + \theta_{TT} - \theta_{TR}}{\sigma_{T0}^2} \quad \text{if } \sigma_{TR}^2 \leq \sigma_{T0}^2 \end{aligned} \quad (5)$$

Howe^[19] developed an extension of the Cornish–Fisher expansion to compute percentiles for the sum of two random variables. Howe's idea was later used to develop confidence intervals on various functions of variance components. For example, methods developed for sums (see, e.g., Ref. [20]), linear combinations,^[21] and ratio of variance components^[22] are introduced in the book by Burdick and Graybill.^[18] These methods are generally referred to as the MLS methods. The general idea for constructing a $(1 - \alpha) \times 100\%$ confidence interval for θ ($\theta = \sum_i c_i \theta_i$), a linear function of variance components, is to use the moment estimator $\hat{\theta}$ as the point estimate for θ , and then add and subtract the square root of a function of variance estimator for components of θ , together with quantile points F_i^*

from F distributions, to approximate the upper and lower confidence bounds such that

$$P \left[\hat{\theta} - \sqrt{\sum_i f_i(c_i^2 \hat{\theta}_i^2, F_i^*)} \leq \theta \leq \hat{\theta} + \sqrt{\sum_i g_i(c_i^2 \hat{\theta}_i^2, F_i^*)} \right] \geq 1 - \alpha$$

The FDA 2001 Method

The FDA 2001 guidance extends a method developed by Hyslop et al.^[13] for the evaluation of individual bioequivalence to solve the population bioequivalence problem. However, that method is not directly applicable to population bioequivalence testing due to the violation of a primary assumption of independence among the estimated components (as pointed out in Ref. [17]).

The method mentioned in the 2001 FDA Guidance is based on a two-sequence, four-period (2×4) crossover design. With some minor modifications, this method can be expressed in a 2×2 design. After converting Eq. 5 into a linear form, it follows that

$$\begin{aligned} \frac{(\theta_D - \theta_I)/n + \theta_{TT} - \theta_{TR}}{\theta_{TR}} &\Rightarrow \eta_1 \\ &= (\theta_D - \theta_I)/n + \theta_{TT} - (1 + \theta_p)\theta_{TR} \end{aligned}$$

and

$$\begin{aligned} \frac{(\theta_D - \theta_I)/n + \theta_{TT} - \theta_{TR}}{\sigma_{T0}^2} &\Rightarrow \eta_2 \\ &= (\theta_D - \theta_I)/n + \theta_{TT} - \theta_{TR} - \theta_p \sigma_{T0}^2 \end{aligned}$$

The population bioequivalence can be evaluated in the following steps:

1. Compute S_{TR}^2 .
2. If $S_{TR}^2 > \sigma_{T0}^2$, construct a $(1 - \alpha) \times 100\%$ upper confidence bound for the linearized function η_1 using

$$U_{R1} = S_D^2/2n + S_{TT}^2 - (1 + \theta_p)S_{TR}^2 + \sqrt{V_{R1}}$$

where

$$\begin{aligned} V_{R1} &= (H_D - E_D)^2 + (1/F_L - 1)^2 S_{TT}^4 + (1 + \theta_p)^2 \\ &\quad (1 - 1/F_U)^2 S_{TR}^4, \quad H_D = \left\{ \sqrt{E_D} + t_{1-\alpha; 2(n-1)} \sqrt{S_I^2/n} \right\}^2, \quad E_D = \\ &\quad S_D^2/n, \quad F_L = F_{1-\alpha; 2(n-1), \infty}, \quad F_U = F_{\alpha; 2(n-1), \infty} \cdot F_{\alpha; df1, df2} \end{aligned}$$

represents the F value with $df1$ as the degrees of freedom in the numerator and $df2$ the degrees of freedom in the denominator, with area α to the right. $t_{1-\alpha; 2(n-1)}$ is the t value with degrees of freedom $2(n-1)$ and area α to the right.

3. If $S_{TR}^2 \leq \sigma_{T0}^2$, construct an upper bound for the linearized function η_2 using

$$U_{C1} = S_D^2/n + S_{TT}^2 - S_{TR}^2 - \theta_p \sigma_{T0}^2 + \sqrt{V_{C1}}$$

where

- $V_{C1} = (H_D - E_D)^2 + (1/F_L - 1)^2 S_{TT}^4 + (1 - 1/F_U)^2 S_{TR}^4$.
4. Reject $H_0: \theta_{PBE} \geq \theta_p$ and claim population bioequivalent if the computed upper bound is < 0 .

The above four steps summarize the FDA method for the evaluation of population bioequivalence.

In addition to the FDA method, a few, newer methods were recently developed. Two of these new methods are introduced in the following subsections. One is an alternative MLS method that preserves the ratio form of the population bioequivalence criteria [without linearization^[16]]. The other method is proposed to deal with the correlation among the estimated variance components.^[17]

An Alternative Modified Large Sample Method

Lu et al.^[22] propose an MLS confidence interval on $(a\theta_1 - b\theta_2)/\theta_3$. In a recent paper by Quiroz et al.,^[16] a method to evaluate population bioequivalence is proposed by applying the method by Lu et al.^[22] for the referenced scale criterion and applying the method by Ting et al.^[21] for the constant scaled criterion. The approach by Quiroz et al. first converts the noncentral χ^2 variable SSD/θ_1 into a central χ^2 variable with n^* degrees of freedom, where $n^* = (1 + 2\hat{\lambda})^2/(1 + 4\hat{\lambda})$, and $\hat{\lambda} = n(\bar{Y}_R - \bar{Y}_T)^2/2S_1^2$. Then the χ^2 variables are combined to fit the form $(a\theta_1 - b\theta_2)/\theta_3$.

Let $\theta_{DT} = \theta_D/n + \theta_{TT}$ and $\theta_{IR} = \theta_I/n + \theta_{TR}$, based on Satterthwaite.^[23] Then $n_{DT} S_{DT}^2/\theta_{DT}$ has an approximate χ^2 distribution with n_{DT} degrees of freedom (df), and $n_{IR} S_{IR}^2/\theta_{IR}$ has an approximate χ^2 distribution with n_{IR} df , where $S_{DT}^2 = S_D^2/n + S_{TT}^2$, $S_{IR}^2 = S_I^2/n + S_{TR}^2$,

$$n_{DT} = \frac{S_{DT}^4}{(S_D^4/n^2 n^*) + (S_{TT}^4/n_{TT})}, \quad n_{IR} = \frac{S_{IR}^4}{(S_I^4/n^2 n_1) + (S_{TR}^4/n_{TR})}$$

and $n_{TR} = n_{TT} = n_1 = 2(n - 1)$.

Now the reference-scaled PBC can be rewritten as $(\theta_{DT} - \theta_{IR})/\theta_{TR}$. Using this approach, the resulting $1 - \alpha$ upper bound for PBC is

$$U_{R2} = \begin{cases} K_0 \hat{\rho} + \sqrt{V_{U1}}/S_{TR}^2 & \text{if } T \leq F_{L1} \\ K_0 \hat{\rho} + \sqrt{V_{U2}}/S_{TR}^2 & \text{if } T > F_{L1} \end{cases}$$

where

$K_0 = 1 - 2/n_{TR}$	$\hat{\rho} = (S_{DT}^2 - S_{IR}^2)/S_{TR}^2$	$T = S_{DT}^2/S_{TR}^2$
$V_{U1} = K_1 S_{DT}^4 + K_2 S_{IR}^4$	$V_{U2} = K_3 S_{DT}^4 + K_4 S_{IR}^4$	$K_1 = [K_0(1/F_{L1} - 1)]^2 - K_2/F_{L1}^2$
$K_2 = (K_0 - 1/F_{U1})^2$	$K_3 = (1/F_{L2} - K_0)^2$	$K_4 = [K_0(1 - F_{L1})]^2 - K_3 F_{L1}^2$
$F_{L1} = F_{1 - \alpha; n_{DT}, n_{IR}}$	$F_{L2} = F_{1 - \alpha; n_{DT}, n_{TR}}$	$F_{U1} = F_{\alpha; n_{IR}, n_{TR}}$

Again, the notation $F_{\alpha; df1, df2}$ represents the F value with degrees of freedom $df1$ in the numerator and $df2$ in the denominator, with area α to the right. For the constant

scale form PBC, a confidence interval on $\theta_{DT} - \theta_{IR}$ can be constructed using the Ting et al.^[21] method:

$$U_{C2} = \left(S_{DT}^2 - S_{IR}^2 + \sqrt{H_1^2 S_{DT}^4 + G_2^2 S_{IR}^4 + H_{12} S_{DT}^2 S_{IR}^2} \right) / \sigma_{T0}^2$$

where $H_1 = 1/F_{1 - \alpha; n_{DT}, \infty} - 1$, $G_2 = 1 - 1/F_{\alpha; n_{IR}, \infty}$, and $H_{12} = [(1 - F_{L1})^2 - H_1^2 F_{L1}^2 - G_2^2]/F_{L1}$.

The Quiroz et al.^[16] method for testing hypotheses concerning θ_{PBE} is:

1. Compute S_{TR}^2 .
2. If $S_{TR}^2 > \sigma_{T0}^2$, construct an upper bound using U_{R2} .
3. Otherwise, using U_{C2} .
4. Reject $H_0: PBC \geq \theta_p$ and claim population bioequivalence if the computed upper bound is less than θ_p .

Another New Method That Considers Correlation Among Variance Components

Chow, Shao, and Wang^[17] propose a new method to evaluate population bioequivalence. They consider the dependence among the chi-square random variables and include a covariance structure in the estimate. In their method, the upper bound for the reference scaled criterion is

$$U_{R3} = \hat{\delta}^2 + S_{TT}^2 - (1 + \theta_p) S_{TR}^2 + t_{1 - \alpha, 2(n-1)} \sqrt{V_{R3}}$$

where

$$\begin{aligned} \hat{\delta} &= \frac{1}{2} (\bar{Y}_{21} - \bar{Y}_{12} - \bar{Y}_{11} + \bar{Y}_{22}), \\ V_{R3} &= (2\hat{\delta}, 1, -(1 + \theta_p)) C (2\hat{\delta}, 1, -(1 + \theta_p))', \\ C &= \begin{pmatrix} \hat{\sigma}_{1,1}^2/2n & (0,0) \\ (0,0)' & (C_1 + C_2)/(4n - 4) \end{pmatrix} \end{aligned}$$

C_1 is the sample covariance matrix of $((Y_{i11} - \bar{Y}_{*11})^2, (Y_{i21} - \bar{Y}_{*21})^2)$, C_2 is the sample covariance matrix of $((Y_{i22} - \bar{Y}_{*22})^2, (Y_{i12} - \bar{Y}_{*12})^2)$, $i = 1, \dots, n$ and

$$\hat{\sigma}_{1,1}^2 = [1/(2n - 2)] \sum_{k=1}^2 \sum_{i=1}^n (Y_{ik1} - Y_{i2k} - \bar{Y}_{*1k} + \bar{Y}_{*2k})^2$$

The upper bound for the constant scaled criterion is:

$$U_{C3} = \hat{\delta}^2 + S_{TT}^2 - S_{TR}^2 - \theta_p \sigma_{T0}^2 + t_{1 - \alpha, 2(n-1)} \sqrt{V_{C3}}$$

where $V_{C3} = (2\hat{\delta}, 1, -1) C (2\hat{\delta}, 1, -1)'$

Regarding the choice of U_{R3} or U_{C3} , the authors suggest to perform a hypothesis test to observe if S_{TR}^2 is $> \sigma_{T0}^2$. This test is performed by comparing $S_{TR}^2/\chi_{\alpha; 2(n-1)}^2$ with σ_{T0}^2 . If $S_{TR}^2/\chi_{\alpha; 2(n-1)}^2 \geq \sigma_{T0}^2$, then the upper bound is constructed using U_{R3} . Otherwise U_{C3} is used.

One advantage of this method is that it can be applied to unbalanced cases, that is, when there are different numbers

of subjects in sequence 1 and sequence 2. Furthermore, methods proposed by Chow et al.^[17] are also available for 2×3 designs and 2×4 designs.

CONCLUSION

The methodological development of nonbootstrap methods by Hyslop et al.,^[13] Quiroz et al.,^[16] and Chow et al.^[17] have made the implementation of population bioequivalence testing more feasible. All these MLS methods avoid the problems suffered by the bootstrap method, which was originally proposed by the FDA in its early draft guidance for PBE and IBE. The MLS methods can perform population bioequivalence testing quickly with very moderate programming work. In addition, they also provide an effective mean for sample size assessment via simulation, which is almost impossible to do by the bootstrap method because of the sheer amount of computing time required. The performance of each MLS method was studied via simulation by its set of authors in terms of the Type I error or the coverage of the one-sided 95% confidence interval for the values of criteria in Eq. 4.^[13] Certain authors also performed simulation to study the power of their method and the FDA method for claiming population bioequivalence.^[16,17] In summary, all authors claimed the validity of their own methods. The interested readers should be referred to the articles listed in the references for further reading.

Although the major technical difficulties for population bioequivalence testing have been largely resolved by the development of the MLS methods. The controversies are still reverberating. The major division among the pharmaceutical industry, the regulatory agencies, and the academic community is on the necessity for amending the existing average bioequivalence criterion. Even though the concept of prescribability is laudable, the deficiency of average bioequivalence is yet evident. In other words, there are no or little evidences reported so far showing that a commercial formulation or a generic drug is either not safe or not efficacious although it was shown to be average bioequivalent to the formulation or the brand drug that was carefully studied in the clinical development. Many people strongly believe that the criterion of average bioequivalence is still adequate to protect consumers' interest for the most time. The compatibility of variability in the key pharmacokinetic parameters between two drugs is a valid concern and it is not reflected in the average bioequivalence criterion. Nonetheless, people still question whether the proposed population bioequivalence criterion is a good solution.

In the statistical community, people see the development in bioequivalence criteria as an opportunity for applying more sophisticated experimental designs and statistical techniques, but, ultimately, it depends on the consensus among medical practitioners, the regulatory agencies, and

the pharmaceutical companies. For the readers who are interested in the position of pharmaceutical industry, they should be referred to "PhRMA Perspective on Population and Individual Bioequivalence."^[24]

REFERENCES

1. FDA, *Guidance: Statistical Procedures for Bioequivalence Studies Using a Standard Two-Treatment Crossover Design*; Office of Generic Drugs, Food and Drug Administration, Public Health Service, U.S. Department of Health and Human Services, 1992.
2. Hwang, S.; Huber, P.; Hesney, M.; Kwan, K.C. Bioequivalence and interchangeability. *J. Pharm. Sci.* **1978**, *67*, 343–346.
3. Ekbohm, G.; Melander, H. The subject-by-formulation interaction as a criterion for interchangeability of drugs. *Biometrics* **1989**, *45*, 1249–1254.
4. Anderson, S.; Hauck, W.W. Consideration of individual bioequivalence. *J. Pharmacokinet. Biopharm.* **1990**, *18*, 259–273.
5. Hauck, W.W.; Anderson, S. Types of bioequivalence and related statistical considerations. *Int. J. Clin. Pharmacol. Ther.* **1992**, *30*, 181–187.
6. Sheiner, L.B. Bioequivalence revisited. *Stat. Med.* **1992**, *11*, 1777–1788.
7. Schall, R.; Luus, H.G. On population and individual bioequivalence. *Stat. Med.* **1993**, *12*, 1109–1124.
8. Schall, R. A unified view of individual, population and average bioequivalence. *Bio-Int. 2: Bioavailab. Bioequiv. Pharmacokinet. Stud.* **1995**, 91–106.
9. FDA, *Draft Guidance: In vivo Bioequivalence Studies Based on Population and Individual Bioequivalence Approaches*; Center for Drug Evaluation and Research, Food and Drug Administration, Public Health Service, U.S. Department of Health and Human Services, December 1997.
10. FDA, *Draft Guidance: Average, Population, and Individual Approaches to Establishing Bioequivalence*; Center for Drug Evaluation and Research, Food and Drug Administration, Public Health Service, U.S. Department of Health and Human Services, September 1999.
11. FDA, *Guidance for Industry: Statistical Approaches to Establishing Bioequivalence*; Center for Drug Evaluation and Research, Food and Drug Administration, Public Health Service, U.S. Department of Health and Human Services, January 2001.
12. FDA, *Draft Guidance: Bioavailability and Bioequivalence Studies for Nasal Aerosols and Nasal Sprays for Local Action*; Center for Drug Evaluation and Research, Food and Drug Administration, Public Health Service, U.S. Department of Health and Human Services, June 1999, Draft.
13. Hyslop, T.; Hsuan, F.; Holder, D.J. A small sample confidence interval approach to assess individual bioequivalence. *Stat. Med.* **2000**, *19*, 2885–2897.
14. Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*; Marcel Dekker: New York, 1999.

15. Schuirmann, D.J. A comparison of the two one-sided test procedure and the power approach for assessing the bioequivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* **1987**, *15*, 564–657.
16. Quiroz, J.; Ting, N.; Wei, G.C.G.; Burdick, R.K. A modified large sample approach in the assessment of population bioequivalence. *J. Biopharm. Stat.* **2000**, *10* (4), 527–544.
17. Chow, S.C.; Shao, J.; Wang, H. Statistical tests for population bioequivalence. *Statistica Sinica*, **2003**, 539–554.
18. Burdick, R.K.; Graybill, F.A. *Confidence Intervals on Variance Components*; Marcel Dekker: New York, 1992.
19. Howe, W.G. Approximate confidence limits on the mean of $X + Y$ where X and Y are two tabled independent random variables. *J. Am. Stat. Assoc.* **1974**, *69*, 789–794.
20. Graybill, F.A.; Wang, C.M. Confidence intervals on non-negative linear combinations of variances. *J. Am. Stat. Assoc.* **1980**, *75*, 869–873.
21. Ting, N.; Burdick, R.K.; Graybill, F.A.; Jeyaratnam, S.; Lu, T.-F.C. Confidence intervals on linear combinations of variance components that are unrestricted in sign. *J. Stat. Comput. Simul.* **1990**, *35*, 135–143.
22. Lu, T.-F.C.; Graybill, F.A.; Burdick, R.K. Confidence intervals on the ration of expected mean squares $(\theta_1 - d\theta_2)/\theta_3$. *J. Stat. Plan. Inference* **1989**, *21*, 179–190.
23. Satterthwaite, F.E. An approximate distribution of estimates of variance components. *Biom. Bull.* **1946**, *2*, 110–114.
24. Barrett, J.S.; Batra, V.; Chow, A.; Cook, J.; Gould, A.L.; Heller, A.H.; Lo, M.W.; Patterson, S.D.; Smith, B.P.; Stritar, J.A.; Vega, J.M.; Zarifta, N. PhRMA perspective on population and individual bioequivalence. *J. Clin. Pharmacol.* **2000**, *40*, 1–10.

Encyclopedia of Biopharmaceutical Statistics, 4th Edition

Volume IV

Population PK/PD–Z-Score

Pages 1747—2332

Traditional
(cont.)–Z-Score

Targeted–Traditional

Stability–Survival

Scaled–SROC

ROC–Sample Size

Regulatory–Risk

Propensity–Rank
Regression

Population
PK/PD–Profile



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Population PK/PD Analysis

D. Concordet and C. Ané

Ecole Nationale Vétérinaire de Toulouse, Toulouse, France

F. Léger

Université Paul-Sabatier and Institut Claudius-Regaud, Toulouse, France

Abstract

Pharmacokinetics and pharmacodynamics are divisions of pharmacology that study the action of the body on the drug and the action of the drug on the body, respectively. This entry focuses on population pharmacokinetics and pharmacodynamics modeling a powerful approach in the selection of adequate dosage regimen.

INTRODUCTION

Pharmacokinetics and pharmacodynamics are divisions of pharmacology that study the action of the body on the drug and the action of the drug on the body, respectively. They complete dose titration studies aimed to select rational dosage regimens.

The classical pharmacokinetic (PK)/pharmacodynamic (PD) studies are experimental. They are usually performed on healthy volunteers or highly selected patients to minimize interindividual variability. These studies allow to demonstrate a mechanism and to obtain rough quantitative information on the link between the dose and the effect (PK/PD behavior) of the drug.

Population PK/PD studies quantify the effect of the drug on a population of patients who could use the drug. They allow quantifying, explaining, and predicting how the variability of the drug plasma concentration acts on the variability of the drug effect. They enable optimization (individualization) of dosage regimen. Usually performed on a sample of patients who are representative of the target population, they can be considered as observational studies.

Population PK/PD relies on the use of models. A model can be defined as a simplified description of reality. Most models used in population PK/PD are statistical models. These models describe observations (concentrations and effects), but may also give some outlines of the underlying biological process. This article focuses mainly on the modelling aspects of PK/PD analysis.

First, we recall the definitions of pharmacokinetic, pharmacodynamic, PK/PD, and the population approach. The modelling approach is illustrated with the study of the curious toxic effect of topotecan, an anticancer drug. Simulations are then performed, by using estimates of the population parameters, to propose a dosage regimen with a controlled toxicity.

DEFINITIONS

One of the primary goals of drug development is to generate data that permit to select the appropriate starting dose of the drug and, subsequently, to adjust dosage to the needs of a particular patient or group of patients. Ideally, such knowledge should allow to increase the likelihood of achieving the intended therapeutic effect while reducing the risk of adverse events. One clinical development approach used to achieve this is multidose-level clinical parallel dose–response testing, which is represented in [Fig. 1](#).

However, as advocated by Toutain,^[1] this approach does not allow provision of information on the shape of effect for a particular patient. The different doses are compared with statistical tests; thus the selected doses depend heavily on the minimum difference that can be evidenced with a given power. In this case, a large interindividual variation makes difficult the discrimination between dose effects and is considered as a difficulty in the dose selection.

A PK/PD approach gives a better insight in understanding how the body acts on the dose to give concentrations (PK step) and how variations of these concentrations are involved in the effect (PD step) (cf. [Fig. 2](#)). Holford and Sheiner^[2,3] give more precise definitions of pharmacokinetics, pharmacodynamics, and PK/PD.

Population Pharmacokinetics

Pharmacokinetic models describe the interaction between drug input and the drug disposition. Usual functions used to describe the drug input are intravenous bolus, infusion for a given time, first-order process that approximate the oral absorption.

Compartmental models are the most common approach used to describe drug disposition. Compartments are virtual spaces within which drug distribution is assumed to

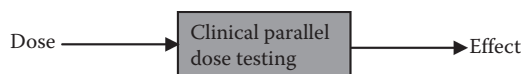


Fig. 1 A clinical design allows to study the effect of doses. It can be considered as a black box. The dose is the input; the effect is the output

be uniform. A compartment is characterized by its apparent volume (which is simply the constant of proportionality between the quantity of drug it contains at a given time and the drug concentration at this time) and clearance parameters (which measure the magnitude of the rates of flow of distribution and elimination). In multicompartment models, the clearances between two compartments are assumed to be equal in both directions.

As an example, in the PK/PD analysis presented hereafter, an anticancer drug (topotecan) was administrated to ($N = 42$) women by 30-min intravenous infusions on five consecutive days. Total topotecan plasma levels were analyzed according to a two-compartment model with linear elimination from the central compartment (cf. Fig. 3). The pharmacokinetic parameters were the clearance (CI) of the central compartment, the volume (V_c) of the central compartment, the volume (V_p) of the peripheral compartment, and the clearance Q between compartments.

The drug concentrations measured in two patients receiving the same dose at the same time after a drug administration are not equal. This variability can be decomposed into two parts: a within-patient variability (essentially explained by analytical error) and a between-patients variability, explained by different pharmacokinetic parameters.

Population pharmacokinetics studies the variability of PK parameters and the variability of concentration profiles. The main idea, which governs population pharmacokinetics, is that each patient is characterized by one's own PK parameters. Making inference on the "population" of PK parameters is then equivalent to making inferences on the population of patients.

Patients involved in population PK studies are representative of a population of patients who intend to use the drug. The "representativeness" is measured with respect to all potential sources of variations of the drug concentrations. As an example, if the concentration profiles observable in males are the same as the ones observable in females, it is not necessary to include both genders in the sample. Using a representative sample makes possible the identification of covariates that explain (or at least are correlated to) the concentration variability. For a given patient, it is then

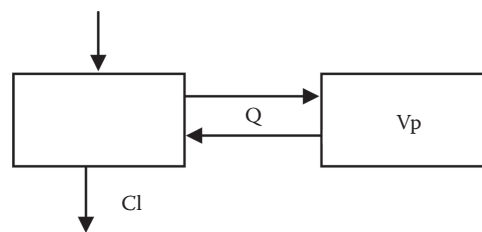


Fig. 3 A two-compartment model with linear elimination from the central compartment was used to describe the drug kinetics

possible to predict (with the knowledge of its covariates) a typical concentration profile that represents the more probable profile that can be observed on patients with the same value of covariates.

The nonlinear mixed-effects models are usually used to describe these different stages of variability. They can be written as:

$$\begin{cases} C_{ij} = f(\phi_i, t_{ij}) + g(\phi_i, t_{ij}, Y) \varepsilon_{ij}, & j = 1, \dots, n_i \\ \phi_i = A_i \mu + \eta_i, & i = 1, \dots, N \end{cases} \quad (1)$$

where C_{ij} is the concentration measured on the i th patient at time t_{ij} , ϕ_i is the p -vector of individual kinetics parameters for the i th patient, $f(\phi, t)$ is the mean concentration at time t for an individual with a kinetic parameter ϕ , and ε_{ij} are independent $N(0,1)$ random variables. The function $g(\phi, t, \beta)$ is the standard deviation of the error of a concentration measurement at time t . Usually, the different concentrations are assumed to be measured with a constant coefficient of variation β , in other terms $g(\phi, t, \beta) = \beta f(\phi, t)$.

The mean variations between the ϕ_i can be explained by covariates whose values are contained in the columns of the matrices A_i . The independent random variables η_i represent the differences between the actual individual kinetic parameters and their average $A_i \mu$. They are assumed to be distributed according to a $N(0, \Omega)$ distribution. Thus the parameters that have to be estimated are (μ, Ω, β) .

PK/PD

Pharmacodynamics allows to relate the drug effect (PD effect) to the concentration at its site of action; therefore, it is independent of time. The rationale for PD studies is a better understanding of the physiological response to a drug. The choice of PD response is often difficult and is out of the scope of this paper. Most drugs circulate in

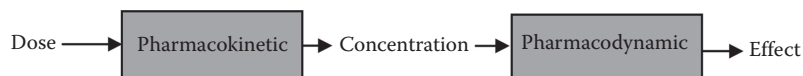


Fig. 2 The PK/PD approach splits the black box of clinical trial into two "grey boxes." The first box transforms a dose into a concentration (PK), whereas the second box links the concentration to the effect (PD)

different tissues, but the observation of concentration at the effect site is even difficult or impossible. Actually, the drug concentrations are easily accessible; thus they are measured in plasma. For this reason, modeling mainly focuses on establishing a link between the plasma (or blood) drug concentration and the PD effect. To avoid confusion with pharmacodynamics, Holford and Sheiner suggested to name PK/PD the link between the drug plasma concentration and the PD effect.

Several classes of PK/PD models exist. The simplest ones are direct models in which there is no delay between a change in plasma drug concentration and a change in effects. This occurs when the site of action has reached equilibrium with plasma.

Indirect models allow describing a delay. This delay may be explained by the fact that the concentration is measured in plasma that is not in equilibrium with the site of action. A plot of effects-vs.-plasma concentrations (ordered by time) showing a curve (hysteresis curve) that turns counterclockwise measures the degree of lag between the change in plasma concentration and that at the effect site. A usual way to model this delay is to add an effect compartment to the kinetic model. This compartment, which is linked to the plasma concentration compartment, mimics the site of action of the drug. Because this site has usually negligible volume compared to the volume of the plasma, the PK/PD analysis is performed in two steps. First, the PK models without the effect compartment are used to estimate the PK parameters. Then PD data are analyzed with the PK model with the effect compartment in which the PK parameters obtained in the first step are maintained fixed. In this second step, only the volume of the effect compartment and its rates of exchange with the central compartment of the PK models are estimated.

The last class of PK/PD models is the physiological class. In these models, the effect of the drug is mediated by some indirect mechanism that is located separately from its site of action. The example given below is an illustration of a model of this class. An anticancer drug administered to patients gives toxicity—a decrease of neutrophil cell count. Each dividing cell is likely to be killed when this drug is present. From a macroscopical point of view, the bone marrow produces progenitor stem cells that divide rapidly, and so can be killed. If these cells survive, they continue to mature without obstacle in the bone marrow. Finally, they migrate into the blood—the observed pool—as white blood cells.

In a regulatory point of view, knowledge of PK/PD relationships can contribute to the strength of evidence to support efficacy and to address issues and questions related to the safety of drug doses and dosing regimens in certain populations of patients. The ICH E4 guideline for industry, “Dose–Response Information to Support Drug Registration,”^[4] describes the purpose of exposure–response information and the uses of dose–response and/or concentration–response data in choosing doses during

the drug development process. A recent US Food and Drug Administration (FDA) guidance for industry, “Exposure–Response Relationships: Study Design, Data Analysis, and Regulatory Applications,”^[5] provides recommendations for sponsors and applicants on the use of exposure–response information in the development of drugs by focusing on human studies.

Population PK/PD

If there exists a between-subjects variability in plasma drug concentration, this variability is generally lower than the one observable in effects. This is essentially due to the variety of potential sources of variations of the PD response. Population PK/PD models allow the description of this variability. These models are obtained by adding to the population PK models (Eq. 1) two components describing how the effects vary with concentration (and eventually time) and how the parameters involved in the first component vary between individuals or/and within individuals (interoccasion variability), respectively:

$$\begin{cases} Y_{ij} = N(\psi_i, \phi_i, t_{ij}) + h(\psi_i, \phi_i, t_{ij}, Y) \varepsilon_{ij}, & j = 1, \dots, n_i \\ \psi_i = m + \eta_i, & i = 1, \dots, N \end{cases} \quad (2)$$

where Y_{ij} is the response measured on the i th patient at time t_{ij} , ψ_i is the p -vector of individual PD parameters for the i th patient, $N(\psi, \phi, t)$ is the mean PD response at time t for an individual with a PK parameter ϕ and a PD parameter ψ , and ε_{ij} are independent standard Gaussian random variables. The function $h(\psi, \phi, t, \beta)$ is the standard deviation of the error of a PD response at time t . Usually, the different responses are assumed measured with a constant coefficient of variation β , in other terms $h(\psi, \phi, t, \beta) = \beta N(\psi, \phi, t)$.

As with the PK model, covariates that explain the between-subjects variability can be incorporated within the ψ_i population mean m . The independent random variables η_i represent the differences between the actual individual PD parameters and their average m . They are assumed to be distributed according to a $N(0, D)$ distribution. The parameters that have to be estimated are (m, D, β) . The framework of analysis for standard direct and indirect models is fully developed in Refs. [2] and [3].

Simulations that use estimates of PK and PD population parameters are performed to predict the variability of effects that can be encountered within the whole population with a specific dosage regimen. Such a simulation approach is now used to optimize the design of clinical trials.

Finally, the dosage regimen can be individualized to control the effect on a patient by taking into account the different covariates involved in the explanation of PK/PD variability.

There is no explicit regulatory requirement regarding “population PK/PD,” but the US FDA guidance for industry on population pharmacokinetics^[6] opens the door to widespread application of mixed-effects predictive models and provides guidelines that can be used for population PK/PD analyses.

Software for Population PK/PD Analysis

One of the most commonly used softwares for population PK and PK/PD analysis is NONMEM (the acronym for nonlinear mixed-effects modelling).^[7] NONMEM allows to fit nonlinear mixed-effects (statistical regression-type) models. A powerful feature of mixed-effects modelling is its ability to accommodate patient data as they arise in the course of routine clinical therapy, where data are typically sparse and obtained at unstructured times. The maximum likelihood estimator of the unknown parameters of models (Eqs. 1 and 2) cannot be easily computed in a closed way. The main estimation method used in NONMEM is first order (FO). It consists in approximating the model described by Eq. 1 (Eq. 2, respectively) by performing a first-order Taylor expansion of f and g (N and h , respectively) about the mean value of the random variables ϕ_i (ψ_i , respectively). Thus the obtained models are linear with respect to the random effects, and their parameters can be estimated using standard methods.

AN EXAMPLE

The following study gives an example of the modelization process and it shows conclusions that can be expected from a population PK/PD study. The method and results presented here are fully developed in Ref. [8].

A study was performed to find an admissible regimen, which is effective and controls the toxic effect of topotecan. Topotecan is an anticancer drug given by intravenous infusion to women with ovarian cancer as second-line therapy (cf. Ref. [9], for instance). The major toxic effect of topotecan is a decrease in neutrophil count, which occurs 8–15 days after drug administration (Fig. 4). A primary index used to measure this is the time the neutrophil count remains below the fixed limit: 500 PN/mm³. Another index is the minimum neutrophil count reached. It has been shown in mice^[10] that this toxicity varies not only with the dose given, but also, importantly, with the duration of the exposure to the drug. Thus, when the total dose is given in a single infusion, a low toxicity is observed; but when the same total dose is fractionated over 5 days, toxicity is high. Surprisingly, when it is fractionated over 20 days or more, toxicity is reduced.

The analysis is presented in three sections. “The Pharmacokinetic Model” describes briefly the PK model and gives the population estimates. The analysis of a physiological

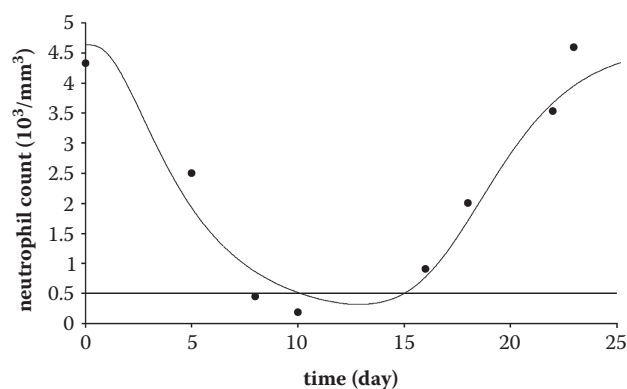


Fig. 4 An example of observed and fitted neutrophil profiles. The horizontal line at 0.5×10^3 PN/mm³ allows one to define the primary index of toxicity (time spent below)

(PD) model that explains the toxicity observed in mice and the discussion on the PK/PD link are presented in “The Pharmacodynamic Model.” Then, using this model, the occurrence and magnitude of the toxicity in-patients are predicted using simulations.

The Pharmacokinetic Model

Topotecan was administrated by 30-min intravenous infusions to $N = 42$ women on 5 consecutive days. For each patient, the three first daily doses were fixed a priori, and the two last doses were adjusted to reach a total area under the curve (AUC) within a targeted range (37,500–75,000 nM min). The PK data were analyzed by Friberg et al.^[11] on a subsample of $N_1 = 31$ women. Let us recall briefly their results. Total topotecan plasma levels were analyzed according to a two-compartment model with linear elimination from the central compartment. The resulting individual parameters were the clearance Cl of the central compartment, the volume V_c of the central compartment, the volume V_p of the peripheral compartment, and the clearance Q between compartments. This clearance Q was assumed to be fixed in the population, and estimated by $\hat{Q} = 46.6$ l hr⁻¹. The individual PK parameters $\phi_i = (\ln Cl_i; \ln V_{ci}; \ln V_{pi})$ were assumed to be independent identically distributed (i.i.d.) and drawn from a Gaussian $N(\mu, \Omega)$ distribution, and the covariance matrix Ω was assumed to be diagonal. The estimation of (μ, Ω) obtained from Ref. [11] in the subsample of size 31 is given in Table 1.

Table 1 PK Population Parameter Estimates

	$\ln Cl$	$\ln V_c$	$\ln V_p$
$\hat{\mu}$	2.99	3.66	3.35
$\sqrt{\hat{\Omega}}$	0.42	0.55	0.38

Clearance (Cl) is expressed in liters per hour; volumes V_c and V_p are expressed in liters.

Finally, the individual parameters ϕ_i were predicted by the Bayesian estimates (maximum a posteriori) for all 42 patients. In the rest of this article, the kinetic profiles are considered as known and fixed.

The Pharmacodynamic Model

The discussion concerning the pharmacodynamic model is divided into three parts. First, a physiologically based structural model is chosen. Then, we consider the properties and choice of the function describing the drug action on cells. Finally, a nonlinear mixed-effects model is set up, and its parameters are estimated.

The Choice of the Structural Model

Topotecan is a drug that acts during the replication of DNA. It binds to topoisomerase I when this enzyme is unwinding DNA. At this stage, replication is stopped, DNA is broken, and the cell dies. Thus, each dividing cell is likely to be killed when topotecan is present. From a macroscopical point of view, the bone marrow produces progenitor stem cells that divide rapidly, and so can be killed. If these cells survive, they continue to mature without obstacle in the bone marrow. Finally, they migrate into the blood—the observed pool—as white blood cells. When no drug is given, this system is at equilibrium, and can be described using the family of compartmental models given in Fig. 5.

Bone marrow constitutes a nonobserved part of the life of these cells. To get information about this part, a large number of different outputs (neutrophil profiles), obtained with a large number of different inputs (kinetics), are necessary. We have at our disposal roughly a single shape of PK profile. Remember that all the women received five consecutive daily infusions, and that the two last doses were adjusted so as to reach a target AUC. Consequently, the data are not rich enough to allow a precise description of the actual model. Therefore, we deliberately chose to use a model built on numerous data for 5-fluorouracil in rats.^[12] As this drug acts at the same stage as topotecan, it appears reasonable to take the structure of the rat model and to scale it to humans, although we will see that slight modifications are necessary. The model used in Ref. [12] contains five compartments.

Two compartments are sensitive to the drug, two are nonsensitive, and the last compartment is the blood pool. The exchanges are second-order exchanges, except for what leaves the blood, which is of order 1. All bone marrow compartments share the same second-order constant k . Recall that when the exchanges are of order 2, the behavior of a given cell depends on the size of the compartment. The more cells present in this compartment, the more rapidly this particular cell will move. This type of exchange roughly mimics the birth of new cells in the compartment, and describes adequately the observed neutrophil profiles, which quickly leave and come back to baseline. In the original model,^[12] the production rate was assumed to be constant, up to a feedback mechanism. More precisely, if N_{base} is the baseline neutrophil count, and $N(t)$ is the neutrophil count at time t , then the production rate is set to $k_{\text{in}} N_{\text{base}}/N(t)$. Finally, the action of the drug is modeled by a first-order killing rate on each of the two sensitive compartments. This killing rate is assumed to be linked directly to the drug concentration C_t at time t : $F(C_t) = k_e C_t$.

Applying this model to data on women, the individual curve fitting was not satisfactory, for two reasons. The feedback leads to a rebound, which is pronounced in rats. As it is missing in human patients, this feedback mechanism is inappropriate for our study. Without feedback, the model used in Ref. [12] appears to be too constrained to allow a rapid decrease or increase of the curve. In this model, cells have to spend the same time in the two different states (sensitive/nonsensitive). For this reason, we chose two different rates of exchange (k and k'), as shown in Fig. 5. The mean residence time of a cell in the nonsensitive region is then approximately proportional to the number of nonsensitive compartments, $2k'^{-1}$. Similarly, the mean residence time of a cell in the sensitive region is proportional to $2k^{-1}$. In other respects, we notice that increasing the number of nonsensitive compartments does not determine another dynamic behavior because an increase in k' can compensate. The same argument applies to the number of sensitive compartments and the constant k .

In summary, there are two differences between the previous model in Ref. [12] and ours: feedback is discarded and two rates of exchange are used instead of one. The differential equations driving the selected model are now detailed.

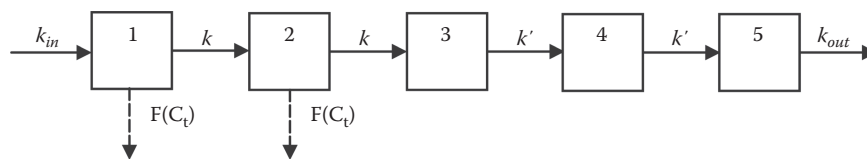


Fig. 5 General compartment model. The first four compartments are the different stages of a cell life in the bone marrow. The last compartment is the blood pool. In the original model, it was assumed that $k = k'$ and that a feedback mechanism (a link between compartments 5 and 1) was used

The neutrophil profile in the blood is the solution $N(t)$, $t \geq 0$, of the system:

$$\begin{cases} \frac{\partial X_1}{\partial t} = k_{in} - kX_1^2 - F(c_i)X_1 \\ \frac{\partial X_2}{\partial t} = kX_1^2 - kX_2^2 - F(c_i)X_2 \\ \frac{\partial X_3}{\partial t} = kX_2^2 - k'X_3^2 \\ \frac{\partial X_4}{\partial t} = k'X_3^2 - k'X_4^2 \\ \frac{\partial N}{\partial t} = k'X_4^2 - k_{out}N \end{cases} \quad (3)$$

When $t = 0$, the system is at equilibrium, i.e., $X_1(0) = X_2(0) = \sqrt{k_{in}/k}$, $X_3(0) = X_4(0) = \sqrt{k_{in}/k'}$, and $N(0) = k_{in}/k_{out}$. In the following, the neutrophil profile denoted by $N(t)$ is also denoted by $N(F)(t)$ to emphasize the dependence on F . In the final model, $F(C_i)$ is set to $k_e C_i$. “Drug Action: The PK/PD Link” deals with the rationale of this choice.

Drug Action: The PK/PD Link

The choice of the drug action, modeled by the killing function F , has not yet been discussed. The killing function $F(c)$ has some obvious properties: It cancels at $c = 0$ (no drug, no action), and it is increasing (more drug, greater effect). Intuitively, the shape of F determines the drug action: when F is convex, low concentrations give little toxicity; but when F is concave, low concentrations rapidly give toxicity. Consequently, for a fixed total dose, it seems that when F is convex, a low target toxicity can be reached with a large number of small doses; whereas when F is concave, the same target toxicity will be reached with a small number of high doses. Thus the shape of F seems to be of major importance, especially in the context of this paper.

Actually, the impact of the shape of F is reduced by the existence of a limit of toxicity. This limit is intrinsic to this family of catenary models. It is reached when the sensitive compartments are emptied (by the drug action). In that case, whatever the killing rate, the drug cannot kill more cells than those arriving in these compartments. Even if the killing rate $F(C_i)$ is very high, its effect on the system is very similar to the one that would be obtained with a smaller killing rate. Two different mechanisms drive the toxicity: when $F(C_i)$ is low (below the limit of toxicity), the shape of F determines the toxicity; when $F(C_i)$ is high, only the time it spends above the limit of toxicity governs the effect. These intuitive considerations need to be formalized to quantify the maximal effects.

To this end, let Δ be a fixed length of time and $0 < \varepsilon < 1$.

Let $T = 1/\sqrt{k_{in}/k}$ ($T' = 1/\sqrt{k_{in}/k'}$, respectively) be the mean residence time in the sensitive compartments (nonsensitive compartments, respectively) at steady state. Let $N_b = k_{in}/k_{out}$ be the neutrophil count at steady state (neutrophil count at baseline). Set:

$$K_0 = \max \left\{ \frac{1}{\Delta} \ln \left(\frac{20}{\varepsilon} \sqrt{\frac{k'}{k}} \right); \frac{4}{3\varepsilon \min\{T, T'\}} \left(6 + 4 \frac{\Delta}{T'} \right) \right\} \quad (4)$$

For all killing rate functions $F(C_i)$, the following property is true: if $F(C_i) \geq K_0$ on an interval of length Δ , say $[t_0, t_0 + \Delta]$, then decreasing F to K_0 on the interval $[t_0, t_0 + \Delta]$ does not change the effect more than εN_b .

More precisely: $\sup |N(F)(t) - N(F^{(K_0)})(t)| \leq \varepsilon N_b$, where $F^{(K_0)} = F$ outside $[t_0, t_0 + \Delta]$ and $F^{(K_0)} = K_0$ on $[t_0, t_0 + \Delta]$.

The proof of this result is deliberately omitted because it is too long and without practical interest. Notice that K_0 , as previously defined, does not depend on F . The main consequence of this property is that all shapes of F above K_0 give nearly the same toxicity. A simple way to explain this property is to consider the following example. Let us take the first patient of this study. Assume that her kinetics $(C_i)_{i \geq 0}$, as well as her structural parameters (k_{in}, k, k', k_{out}) are known. Assume now that the killing function F is a step function: $F^{(K)}(c) = 0$ if $c < c_0$, $F^{(K)}(c) = K$ elsewhere, and c_0 is known. Because $(C_i)_{i \geq 0}$ and c_0 are fixed, the lapse of time Δ during which $C_i \geq c_0$ is fixed (e.g., $\Delta = 5 \times 1$ hr). The proposition says that for all $\varepsilon < 1$, there exists K_0 such that for all $K \geq K_0$, $\|N(F^{(K)}) - N(F^{(\infty)})\|_\infty \leq \varepsilon N_b$.

This last inequality means, first, that toxicity is limited by the one given by $N(F^{(\infty)})$. Second, it implies that when K becomes large, the shape of the neutrophil curve becomes constant.

A first consequence of this property is an estimability problem of F . Even if the map $K \rightarrow N(F^{(K)})$ remains injective, its derivative tends to zero (uniformly on time) when K is large. Thus, if the actual K of the considered individual is large (more than K_0), the Fisher information matrix is nearly degenerate, which implies that the maximum likelihood estimator of such a K has too a large variance to be useful in practice.

Whatever is the shape of F above K_0 , it cannot be properly estimated with a reasonable variance. Thus whatever is the chosen parameterization for F , there exists an area of this parameter space where the estimation is difficult. The estimation quality (variance of the estimator) of the parameters governing the shape of F below K_0 depends on the quantity of information available below this limit. If this information is poor, a large variance of the estimator is expected whatever is the parameterization of F . In such a case, the simplest shape for F is to be preferred (i.e., a linear shape). As already mentioned, all women of the study have very similar kinetic profiles, so that the information of the data only concerns a narrow range of concentrations.

Thus we are confronted with two possibilities: either the drug concentrations lead to killing rates above K_0 , thus we have no information about the shape of F for small concentrations, or the drug concentrations lead to small killing rates; but because the range of these concentrations is narrow, the shape of F can be documented for only a small interval of concentrations. In both cases, the linear killing rate $F(c) = k_c c$ has to be chosen.

The second consequence drawn from this property is a qualitative explanation of the toxic behavior observed in mice.^[10] Recall that with a single dose, the observed toxicity is low: because concentrations reach high values, they lead to high killing rates (above K_0), but only for a short time. When this dose is fractionated over 5 days, concentrations are lower but are high enough to lead to killing rates above K_0 . Because the total time spent by the killing rate above K_0 is long, toxicity is high. Finally, when the total dose is fractionated over 20 days, concentrations are too low to give high killing rates and high toxicity. Thus, our model implies a high correlation between toxicity and the time spent by the killing rate above K_0 . That is exactly what is observed in Ref. [10]: a high correlation between toxicity and the time the drug concentration remains above 0.7 μM .

The Population PK/PD Analysis

As it can be seen on a patient-by-patient analysis, both the response curves and the individual PD parameter $\psi = (\ln k_{in}, \ln k, \ln k', \ln k_e)$ vary widely. The PD analysis relies on the concentration time course. Recall that the individual PK profiles are known. We assume here, as have others,^[12,13] that k_{out} does not vary among patients and is equal to $\sim 0.1 \text{ hr}^{-1}$.

The family of nonlinear mixed-effects PD models described by Eq. 2 has been used with $h(\psi_i, \phi, t_{ij}, \sigma) = \sigma N(\psi_i, \phi, t_{ij})$ and $N(\psi, \phi, t)$, with the neutrophil count defined by Eq. 3. For each patient, the neutrophil count $N(\psi_i, \phi_i, t_{ij})$ depends on the PK parameter (ϕ_i). The ψ_i are assumed to be independent of the PK individual parameters ϕ_i . Because all the parameters (k_{in}, k, k', k_e) are positive, we parameterized the model with their logarithm and assumed these logarithms to be distributed according to a normal distribution with mean m and variance D .

With regard to the number of patients involved in the trial, we chose to take D as a diagonal matrix. The estimation method used is FOCE;^[7] it provides an asymptotically Gaussian estimator $\hat{\theta} = (\hat{\sigma}^2, \hat{m}, \hat{D})$ of θ whose asymptotic variance will be denoted by $V(\theta)$.

Table 2 gives the estimation of σ^2 , m , and $\sqrt{D}$, as well as their asymptotic standard error (in parentheses).

It turns out that the optimized criterion (FOCE) has a large number of local minima, reflecting a large distance from the asymptotic framework. Therefore the asymptotic variance-covariance matrix of the estimator should be interpreted with care.

An example of an individual fitted curve is given in Fig. 4.

Table 2 PD Population Parameter Estimates

	$\ln k_{in}$	$\ln k$	$\ln k'$	$\ln k_e$
m	-1.02 (7.4×10^{-3})	-10.1 (0.14)	-7.92 (0.30)	1.55 (0.076)
$\sqrt{D}$	0.39 (0.98)	2.07 (0.82)	0.62 (0.23)	0.84 (0.37)
σ^2	0.37 (0.17)			

Simulations and Toxicity Predictions

This section is devoted to giving a whole set of dosage regimens with controlled toxicity. Recall that the primary measure of toxicity is the time spent with the neutrophil counts below 500 PN/mm^3 . When this time is longer than 7 days, toxicity is considered intolerable. First, we give an index that allows one to decide qualitatively whether, for a given regimen, the toxic behavior of the drug is rather concentration-dependent (toxicity is linked to the daily dose) or time-dependent (toxicity is linked to the time spent by the drug concentration above a limit). Finally, we give quantitative results that determine set dosage regimens with acceptable toxicity.

Simulation of a PK Toxicity Index

The limit of toxicity K_0 gives information on exposure to the drug. It can be used as a surrogate that gives some useful complementary information on the toxic behavior. Indeed, as previously seen, when drug concentrations are low, the drug toxicity is concentration-dependent. But as soon as the limit K_0 is reached, toxicity depends on the time spent above K_0 and is rather time-dependent.

Let us define a toxicity index (TI). For a chosen duration Δ and a chosen $\varepsilon > 0$ (these choices are discussed later), we set $\text{TI} = K_0/k_e$. The main advantage of this index is that it is homogeneous to a drug plasma concentration and can be compared directly to the kinetic profile. Because both K_0 and k_e vary among patients, it is possible to simulate their distribution and then to derive the 5% percentile of TI. In other respects, as has been shown,^[14] it is possible to use covariates such as creatinine clearance to predict for each patient the PK profile for a dosage regimen chosen a priori. As an example, assume that the expected concentration remains below the 5% quantile of TI. It means that the early sensitive compartments are not emptied by the action of the drug, with a probability of 95%. This implies that the toxicity is concentration-dependent with this dosage regimen. It would not be so if the concentration time course had crossed the index TI for longer than Δ .

Let us detail the rationale for the choice of ε and Δ . As can be seen on Eq. 4, K_0 depends on maturation times T and T' , whose values are about 260 and 90 hr. Therefore, K_0 is about $(120\varepsilon)^{-1}(6 + 0.044\Delta)$. Moreover, it is natural to set Δ below 24 hr—the delay between two consecutive

infusions. So the influence of the term 0.044Δ is low compared to 6. We set $\Delta = 5$ hr. It remains to choose ε , which is a proportion of the baseline neutrophil count. If the difference between two neutrophil profiles is of the same order as the critical threshold, then the resulting toxicities are similar. Recall that the critical threshold (500 PN/mm^3) is around 10 times less than the baseline neutrophil count. These considerations lead us to choose $\varepsilon = 0.05$ (5%). Even if this rough method does not tell us directly whether or not an intolerable toxicity is reached, it gives qualitative information about the toxic behavior.

Simulation of Admissible Dosage Regimens

The aim of this part is to predict the set of admissible dosage regimen. A dosage regimen is said to be admissible if an intolerable toxicity occurs for less than a fixed percentage of patients (say 5%). We limited ourselves to dosage regimens with constant daily doses (one infusion each day) and proceeded as follows. For a fixed dosage regimen (i.e., with a fixed number of infusions and a fixed daily dose), we determined the percentage of patients with intolerable toxicity by using a simulation method detailed hereafter. Then, the dosage regimen was adjusted to obtain a percentage equal to 5%. Now, let us detail the simulation method. First, a sample $\theta_j^* = (\sigma_j^{*2}, m_j^*, D_j^*)$ of size 200 was drawn from a $N(\hat{\theta}, V(\hat{\theta}))$. Next, for each θ_j^* , two samples $(\psi_{ij}^*)_{i=1,\dots,150}$ and $(\phi_{ij}^*)_{i=1,\dots,150}$ were drawn from $N(m_j^*, D_j^*)$ and from the PK population distribution $N(\hat{\mu}, \hat{\Omega})$, respectively, given in section devoted to “Population Pharmacokinetics.”

Then, $Y_{ij}^*(t) = N^*(\psi_{ij}^*, \phi_{ij}^*, t)$ was computed, as well as the time $T_{ij}^* = \int_0^\infty 1_{Y_{ij}^*(t) \leq 500 \text{ PN/mm}^3} dt$. The proportion of patients with intolerable toxicity was then estimated as the percentage $\frac{1}{200 \times 150} \sum_{ij} 1_{T_{ij}^* \leq 7 \text{ days}}$.

Fig. 6 gives the critical daily dose obtained with these simulations as a function of the number of infusions. This figure shows that the toxicity occurrence increases with the total dose. Moreover, total doses that give few toxicities do not depend on the number of infusions: it is about 0.17 mg for 5% toxicity and about 1.22 mg for 10% toxicity. In other words, if the target total dose is low, the number of infusions can be chosen as desired without any change of the toxicity occurrence. A striking fact shown in this figure is that the total admissible dose giving 30% toxicity decreases from 6.2 mg (for one infusion) to 4.9 mg (for 15 infusions), and then it increases up to 5.1 mg for 21 infusions. Because 6.2 mg for one infusion produces the same toxic effect as 4.9 mg for 15 infusions, the 1-day schedule can be considered as less toxic than the 15-day schedule (it takes more drug to produce this level of toxicity). Similarly, the 21-day schedule appears to be less toxic than the 15-day schedule. These assertions have been clinically evidenced (cf. Ref. [15] for instance). Fig. 6 is a tool for comparing schedules in a rather qualitative way.

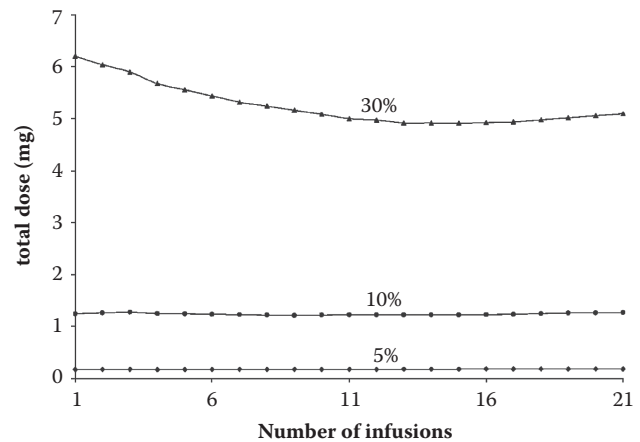


Fig. 6 Curves giving the set of dosage regimen with the same fixed occurrence of toxicity (5%, 10%, and 30%). (View this art in color at <http://www.dekker.com>.)

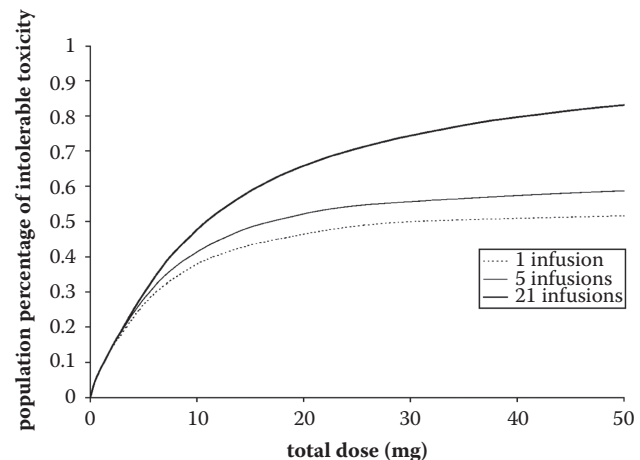


Fig. 7 Cumulative distribution function of intolerable toxicity for the 1-day, 5-day, and 21-day schedules

Figure 7 gives the full quantitative description of toxic effects for these schedules: 1, 5, and 21 infusions.

Notice that the three curves are superposed for low doses/low toxicities. This means that when the total dose is low, the occurrence of toxicity does not depend on the number of infusions, illustrating the concentration dependence and toxic behavior of the drug. When the total dose increases, the occurrence of toxicity tends to plateau at a level that depends on the number of infusions, and below 100%. This is a consequence of the phenomenon described in “Drug Action: The PK/PD Link.” When the limit of toxicity K_0 is reached (with high concentrations), the drug toxic behavior becomes time-dependent. Of course, this study deals only with neutropenia, and many other types of toxicity may occur with high drug doses.

Recall that five infusions were administered and the daily dose could vary during the cycle. The total dose varied between 4.77 and 14.3 mg. As the total dose was

not equally fractionated on the 5 days, we chose to simulate the percentage of intolerable toxicity in our sample as follows. For the i th patient, a sample $(\psi_{ij}^*)_{j=1,\dots,100}$ was drawn from the PD population distribution $N(\hat{m}, \hat{D})$. These 100 individuals were given the PK parameters of the i th patient, as well as her five daily doses. The critical times $T_{ij}^* = \int_0^\infty 1_{N(\psi_{ij}^*, \hat{\sigma}_{ij}^*, t) \leq 500 \text{ PN/mm}^3} dt$ were computed. The proportion of intolerable toxicity in the sample was then estimated as the percentage $\frac{1}{42 \times 100} \sum_{ij} 1_{T_{ij}^* \leq 7 \text{ days}}$. We obtained 39%.

To see if this prediction agrees with the data, the empirical percentage of toxicity in the sample was evaluated by interpolating linearly the observed neutrophil counts for each woman. With this method, only seven women (17%) showed an intolerable toxicity, which is far from 39%. Actually, the empirical percentage of toxicity depends on the chosen method of interpolation. As the neutrophil profiles are convex around their minimum, the linear interpolation underestimates the time spent below 500 PN/mm³, especially when certain observation times are far from each other. With a smooth interpolation, 19 women (45%) showed an intolerable toxicity, which is in agreement with the predicted percentage of toxicity.

CONCLUSION

Population PK/PD modeling provides a powerful approach in the selection of adequate dosage regimen. Even if the modelization given in this paper is only the first step of a longer modelization process (range of concentrations are too narrow to allow estimation of F ; efficacy is not studied), it shows the richness of conclusions that can be reached: the control of the effects (toxicity, efficacy, and so on) via a rational choice of individualized dosage regimen. This example also evidences the weakness of this approach.

The main impediment to the systematic implementation of this approach is the lack of methods to validate the model. Actually, the choice of a specific PK/PD model lies on a number of assumptions that are not easy to check: representativeness of the sample, normality of random effects, assumptions on the structure of variance, and so on. These assumptions have an influence on the predictions that can be drawn from a specific study (especially those obtained by simulations). Also, even when the adopted model provides a good fitting, there is no guarantee that the obtained predictions can be trusted. General statistical methods, such as tests on the variance components, would be extremely useful in the rational choice of model.

Semiparametric (nonparametric, respectively) models are probably a partial answer to some of these limitations. In these models, the distribution of the individual effects (the intraindividual random variable, respectively) is estimated. However, the properties of this family of models (at least the asymptotic ones) are not known. More work is needed in this direction.

REFERENCES

1. Toutain, P.L. Pharmacokinetic/pharmacodynamic integration in drug development and dosage-regimen optimisation for veterinary medicine. *AAPS PharmSci* **2002**, *4* (4), Article 38.
2. Holford, N.H.G.; Sheiner, L.B. Understanding the dose-effect relationship: Clinical application of pharmacokinetic-pharmacodynamic models. *Clin. Pharmacokinet* **1981**, *6*, 429–453.
3. Holford, N.H.G.; Sheiner, L.B. Kinetics of pharmacologic response. *Pharmacol. Ther* **1982**, *16*, 143–166.
4. International Conference of Harmonisation E4. *Dose-Response Information to Support Drug Registration*; Food and Drug Administration, Center for Drug Evaluation (CDER), November 1994.
5. Guidance for Industry. *Exposure-Response Relationships—Study Design, Data Analysis, and Regulatory Applications*; U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation (CDER), April 1999.
6. Guidance for Industry. *Population Pharmacokinetics*; U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation (CDER), February 1999.
7. Beal, S.L.; Sheiner, L.B. *NONMEM User's Guide. Non-linear Mixed Effects Models for Repeated Measures Data*; University of California: San Francisco, 1992.
8. Ané, C.; Concordet, D. Population pharmacokinetics/pharmacodynamics relationships of an anticancer drug. *Stat. Med* **2003**, *22*, 833–846.
9. Rowinsky, E.K.; Grochow, L.B.; Hendricks, C.B.; Ettinger, D.S.; Forastiere, A.A.; Hurowitz, L.A.; McGuire, W.P.; Sartorius, S.E.; Lubejko, B.G.; Kaufmann, S.H.; Donhoffer, R.C. Phase I and pharmacologic study of topotecan: A novel topoisomerase I inhibitor. *J. Clin. Oncol* **1992**, *10*, 647–656.
10. Guichard, S.; Montazeri, A.; Chatelut, E.; Hennebel, I.; Bugat, R.; Canal, P. Schedule-dependent activity of topotecan in OVCAR-3 ovarian carcinoma xenograft: Pharmacokinetic and pharmacodynamic evaluation. *Clin. Cancer Res* **2001**, *7*, 3222–3228.
11. Friberg, L.E.; Brindley, C.J.; Karlsson, M.O.; Devlin, A.J. Models of schedule dependent haematological toxicity of 2'-deoxy-2'-methylidenecytidine (DMDC). *Eur. J. Clin. Pharmacol* **2000**, *56*, 567–574.
12. Friberg, L.E.; Freijs, A.; Sandstrom, M.; Karlsson, M.O. Semiphysiological model for the time course of leukocytes after varying schedules of 5-fluorouracil in rats. *J. Pharmacol. Exp. Ther* **2000**, *295* (2), 734–740.
13. Minami, H.; Sasaki, Y.; Saijo, N.; Ohtsu, T.; Fujii, H.; Igarashi, T.; Itoh, K. Indirect-response model for the time course of leukopenia with anti-cancer drugs. *Clin. Pharmacol. Ther* **1998**, *64*, 511–521.
14. Bugat, R.; Canal, P.; Chatelut, E. Population pharmacokinetics of topotecan: Intraindividual variability of total drug. *Cancer Chemother. Pharmacol* **2000**, *46* (5), 375–381.
15. O'Reilly, S. Topotecan: What dose, what schedule, what route? *Clin. Cancer Res* **1999**, *5*, 3–5.

Postmarketing Adverse Drug Event Signaling

Yi Tsong

U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.

Abstract

Postmarket adverse drug reaction (ADR) surveillance is designed exactly for the purpose of identifying the low-frequency reactions that are not caught in pre-marketing studies, the high-risk group, drug–drug interaction, the longterm effect of the drug, and the increased severity and/or frequency of known reactions. This entry provides statistical analysis methods for identifying these reactions in a drug.

INTRODUCTION

A sponsor seeking to market a new drug product is required by law to have an application [new drug application (NDA)] reviewed and approved by the U.S. Food and Drug Administration (FDA). The principal requirement of the NDA is that the safety and efficacy of the drug is supported by data collected through phase I to III trials. Often, the clinical trial is composed of relatively healthy patients, with a short study duration for adverse event (AE). Once the drug is marketed, the same drug product is available to a patient population that is broader and not well studied. The drug may be used for indications of different disease states; for longer duration; patients, such as the elderly, children, and women not studied in pre-marketing studies; and in patients exposed also to other treatments, etc. Postmarket adverse drug reaction (ADR) surveillance is designed exactly for the purpose of identifying the low-frequency reactions that are not caught in pre-marketing studies, the high-risk group, drug–drug interaction, the long-term effect of the drug, and the increased severity and/or frequency of known reactions.

OVERVIEW

The FDA started the Spontaneous Reporting System (SRS) in 1968 by collecting, processing, and analyzing adverse drug event (ADE) reports. Most of them were reported on the ADE reporting form designed by the FDA (Form 3500, which was later replaced by the MedWatch Form in 1992) and submitted by manufacturers of prescription drugs as a condition for marketing in the United States,

while others are submitted voluntarily by concerned health professionals and consumers. Each report contains the patient's demographic information including age and gender; AE information, including date and outcome of event; description of event, evidence, and existing medical history; information of suspect and concomitant medicine(s), including drug name, dose and frequency, diagnosis for use, whether the event abated after drug use was stopped, and whether the event reappeared after the drug was reintroduced; and the reporter's name and address. Each report was reviewed by a trained professional in the FDA to assure that the form was complete and the information was correct before it was entered into the computer database. Together, these reports in the FDA database make up the SRS of the United States. Details of the laws and regulations on postmarketing reporting of adverse reactions have been described by Faich,^[1] Anello,^[2] and others.^[3–5] The methods for managing the SRS have also been previously described by Anello.^[6]

In 1997, the FDA overhauled the postmarket Adverse Reaction Reporting System in order to accommodate electronic submission and unify the coding with MedDRA^[7] (a dictionary of adverse drug reactions) and renamed the system to Adverse Event Reporting System (AERS). In comparison to SRS, the post-1996 AERS is more structured, has more capacity, and uses more detailed coding terminologies. The FDA now receives over 200,000 ADE reports annually and currently has over 1.6 million reports in its computerized database. It forms a major database for postmarketing adverse drug event monitoring and screening.

The information in the AERS database is rich but cannot be used for computation of incidence rates of specific AE and drug combinations. The difficulties result from the following three reasons:

1. The method of ascertainment of an AE is not uniform among the reporters and nonreporters, and the reporters have their own criteria for identifying an

The views expressed in this entry are those of the author and do not necessarily represent the official position of the U.S. Food and Drug Administration.

AE. It is generally believed that the number of reports is much smaller than the actual number of events.

2. Although all the reports were reviewed carefully in the FDA, the quality of the information collected for a particular AE varies with the reporter.
3. The population at risk is unknown, hence the number of reports is not necessarily closely related to the number of individuals taking the drug.

In spite of these limitations, AERS is a valuable database that can be used for signaling the possibility of excessive or outlying ADE and for generating hypotheses.

Epidemiological and statistical applications were often used in the signaling process. The ADE signaling procedures used in the United States have evolved in the last three decades. Several statistical approaches have been proposed to evaluate adverse experiences using the SRS database. For example, Knapp et al.^[8] and O'Neill and Harris^[9] investigated screening methods for identifying drugs with excessive rates of adverse experiences on a periodic basis. Norwood and Sampson^[10] and Tsong^[11] investigated statistical procedures for detecting increases in frequencies of adverse experience reporting. Lao et al.^[12] investigated statistical procedures for continuous monitoring of increased reports of adverse experience. Rossi et al.,^[13] Hsu,^[14] and Tsong^[15] studied the procedure to adjust for influential factors when comparing the AE reporting of drugs of the same category. Lately, data-mining techniques have been introduced into this field by DuMouchel^[16] and have received much attention from the FDA^[17–19] and the drug-safety research community. In the following sections, three types of signaling procedures are described and discussed.

MONITORING THE INCREASE OF REPORTING OF ADVERSE DRUG EVENTS

Prior to 1994, reporting of increased frequency of ADE reports of the drug product was part of the regulatory requirement of drug sponsor for the purpose of postmarketing safety monitoring. Reasons for increased frequency can be attributed to increased off-label usage, government policy changes, awareness of the AE, or publicity of reports on the AE. In addition, the sponsors' overeagerness to comply with the policy made the task of assessing the association of the increased reporting with safety concerns beyond manageable. Although the regulatory requirement was waived in 1995, the procedures are still used to signal unusual happenings in ADE reporting. The statistical methods used to detect a significant increase have evolved since the first FDA Draft Guidance for Reporting Adverse Drug Reaction issued in 1985^[4] until the last revision issued in 1990.^[5] The 1985 draft guidance recommended a type I error rate inflated normal approximation estimation method. In 1988, Norwood and Sampson^[10] proposed

a conservative exact method to control the type I error rate. In 1990, an FDA draft guidance^[4] recommended a normal approximation testing method based on a research report published later in 1992 by Tsong.^[11] Tsong^[11] compared the type I error rates of seven methods and concluded that the 1990 FDA draft guidance had the best control of the type I error rate.

Let

X_r, X_c = The random variables of the number of reported ADRs in the reporting interval and referenced interval, respectively;

x_r, x_c = The observed number of reported ADRs in the reporting interval and referenced interval, respectively;

$x_t = x_r + x_c$ = The total observed number of ADRs reported in the two intervals combined;

s_r, s_c = The estimated sales volume in the reporting interval and the referenced interval, respectively;

$s_t = s_r + s_c$ = The estimated sales volume of the two intervals combined;

R = Ratio of sales, s_r/s_c ;

PR = Proportion of estimated sales volume in the reporting interval, s_r/s_t ;

D_r, D_c = The reporting rates in the reporting interval and comparison interval, respectively, i.e., $D_r = E(X_r/s_r)$, $D_c = E(X_c/s_c)$.

The notation can be presented in a two-by-two table in Table 1. Note that x_r, s_r, s_c , and s_t are all known margins.

If there is no increase in the reporting period, then $E(X_r/x_r) = s_r/s_t$. Under such null hypothesis, the number of reports that received X_r in the reporting period would follow the following binomial distribution

$$X_r \sim \text{Bin}(x_t, p), \text{ where } p = s_r/s_t$$

In other words, testing for increased frequency means testing the following hypotheses

$$H_0 : E(X_r) = x_t p \text{ vs. } H_a : E(X_r) > x_t p$$

Under the null hypothesis, the 90% approximate confidence interval of $E(X_r)$ is

$$(Rx_c - Z_{0.95}\sqrt{(x_r + x_c)R}, Rx_c + Z_{0.95}\sqrt{(x_r + x_c)R})$$

where $R = s_r/s_c$. Let

$$C = Rx_c + Z_{0.95}\sqrt{(x_r + x_c)R}$$

A significant increase is concluded if $x_r \geq C$.

Table 1 Notation for Monitoring Increased Frequency of Reports of ADR

	Time interval		Total
	Reporting	Comparison	
AE reported	X_r	X_c	x_t
R_x usage estimates	s_r	s_c	s_t

Table 2 ADE Reports of Amnesia and Number of Prescriptions (Triazolam and Temazepam, by Calendar Year)

Calendar year	Triazolam		Temazepam	
	Number of reports	Rxs ^a	Number of reports	Rxs
1981	0	0	1	872
1982	0	0	1	3229
1983	37	1971	1	4537
1984	59	4617	0	5046
1985	32	6870	0	5438
1986	16	9017	0	5498
1987	30	10,458	0	5424
1988	93	11,021	1	5283
1989	27	8744	2	5383
1990	59	7450	2	5451
1991	3	6399	0	5259
Pool	356	66,547	8	5142

From the 1992 CDER Memorandum for Advisory Committee on Neuropharmaceutical Drug Products.

^aRxs: in 1000 prescriptions.

As an example, we consider the data presented to the 1992 CDER Advisory Committee on Neuropharmaceutical Drug Products (Table 2). In order to test whether there was an increased reporting of amnesia associated with triazolam after a concern about this ADE was publicized in 1987. The ADE reports and estimated usages of triazolam are divided into two periods (prior to 1987 and in and after 1987) as shown in Table 3.

$$\begin{aligned}
 R &= (\text{Estimated Use in Post-1987 Interval}) / \\
 &\quad (\text{Estimated Use in Pre-1987 Interval}) \\
 &= 33,614 / 32,933 = 1.021 \\
 C &= Rx_c + Z_{0.95} \sqrt{(x_r + x_c)R} \\
 &= 1.021 \times 174 + 1.645 \sqrt{(356 \times 1.021)} = 209
 \end{aligned}$$

As $x_r = 182 < 209$, it was concluded that there was no evidence of significant increase in the frequency of reports after the publicity.

Note that the estimated usage used in these data is the usage projected to the U.S. population. Variation of the estimate is not accounted for in the analysis. In fact, the analysis relies only on the precision of the sales ratio.

SIGNALING OF HIGH REPORTING RATE

Comparing AE reporting using the SRS database often involves two drugs marketed in different years, which presents a unique problem because of the characteristics of the system. The factors that may lead to bias in the comparison include differences in the number of years on the market of the drugs for the years to be compared, differences in manufacturer reporting patterns, differences in

Table 3 Publicity Effect on ADE Report of Amnesia for Triazolam Prior to Publicity—1983 to 1987, Post Publicity—1988 to 1991

	Time interval	
	Post	Prior
AE reported	182	174
Rx usage estimates	33,614	32,933
Reporting rates	182/33,614	174/32,933
	= 541.4/100,000	= 528.3/100,000

From the 1992 CDER Memorandum for Advisory Committee on Neuropharmaceutical Drug Products.

exposure populations, secular reporting trends, reporting variations, diagnosis and prescription variation, and publicity of the AEs.^[13] Rossi et al.^[13] and Hsu^[14] compared the reports of ADEs associated with piroxicam with seven other NSAIDs that entered the market in different years. They proposed to adjust the number of reports for the market share, differences in reporting patterns associated with the year of marketing, and secular trends of reporting of all ADEs of all drugs combined. They found that the adjusted number of ADE reports for piroxicam ranked first among all NSAIDs. Hsu^[14] also proposed a normal approximation for the confidence interval of the adjusted reporting rates (number of reports per 100,000 prescriptions). Tsong^[15] proposed the application of incidence density statistics to analyze the data.

Using a notation similar to that in the last section, for a given ADE, let x_{1j} , x_{2j} , and $x_j (=x_{1j}+x_{2j})$ denote the number of reports for drug 1, drug 2, and both drugs combined; let s_{1j} , s_{2j} , and s_j denote the number of prescriptions for drug 1, drug 2, and both drugs combined, respectively, in year j ; let R_{1j} , R_{2j} denote the ADE reporting rates for drug 1 and drug 2, respectively, in year j (calendar year). R_{1j} and R_{2j} are estimated by $r_{1j}=x_{1j}/s_{1j}$, $r_{2j}=x_{2j}/s_{2j}$. $RR_j=R_{1j}/R_{2j}$ is the ratio of the two drugs. $P_{ij}=s_{ij}/s_j$ is the proportion of prescriptions for drug i ($i=1,2$) out of s_j , the total prescriptions for drug 1 and drug 2 combined.

Assuming that x_j and P_{ij} are given, under the null hypothesis of equal reporting rates for drug 1 and drug 2, X_{1j} , the random variable of the number of reports for drug 1 (and x_{1j} is the realization of X_{1j}) can be assumed to be distributed as $\text{Bin}(x_j, P_{1j})$ as stated in the last session. For testing $H_0: RR_j=1$ vs. $H_a: RR_j>1$, the binomial test is the conditionally most powerful test. When x_j or x_{1j} is small, the exact binomial p values can be determined. Otherwise, the normal approximation of the binomial test is the following Z test by comparing

$$Z = [x_{1j} - (x_j P_{1j})] / \left[\sqrt{x_j P_{1j} (1 - P_{1j})} \right]$$

against $Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the $(1-\alpha/2)$ th percentile of the standard normal distribution; typically, $\alpha=0.05$. RR_j is estimated by $rr_j=(x_{1j}/s_{1j})/(x_{2j}/s_{2j})$ and its confidence interval can be calculated using Cornfield's method.^[15] The confidence

interval given below based on the simple normal approximation is close to Cornfield's method for the ratio of reporting rates for single marketing year data and pooled data in general without involving complicated computation.

Let $p(x_{ij}) = x_{ij}/x_j$; then, $p(x_{ij})$ is an unbiased estimate of $P(X_{ij})$, the proportion of reports for drug 1; then, the 95% confidence limits of $P(X_{ij})$, p_{Lj} and p_{Uj} , are

$$p_{Lj} = \left\{ x_{ij} + Z_{\alpha/2}^2 - \sqrt{x_j Z_{\alpha/2}^2 [p(x_{ij}) - p(x_{ij})^2] + Z_{\alpha/2}^4 / 4} \right\} / (x_j + Z_{\alpha/2}^2) \quad (1)$$

and

$$p_{Uj} = \left\{ x_{ij} + Z_{\alpha/2}^2 + \sqrt{x_j Z_{\alpha/2}^2 [p(x_{ij}) - p(x_{ij})^2] + Z_{\alpha/2}^4 / 4} \right\} / (x_j + Z_{\alpha/2}^2) \quad (2)$$

Hence the 95% confidence interval for RR_j , the ratio of reporting rates in year j , is

$$((p_{Lj}/P_j)/(q_{Lj}/Q_j), (p_{Uj}/P_j)/(q_{Uj}/Q_j)) \quad (3)$$

where $q_{Lj} = 1 - p_{Lj}$, $q_{Uj} = 1 - p_{Uj}$, $Q_j = 1 - P_j$.

In order to summarize the yearly comparisons, the following Mantel-Haenszel-type statistic is used.

$$Z^* = \frac{\sum_j (x_{ij} - x_j P_j)}{\sqrt{\sum_j x_j P_j (1 - P_j)}} \quad (4)$$

Z^* is approximately distributed as $N(0,1)$ under the null hypothesis of equal reporting rates. R , the overall ratio of rates, is estimated by

$$\sum_j [x_{ij}(1 - P_j)] / \sum_j (x_{2j} P_j) \quad (5)$$

Confidence limits are estimated using Cornfield's method.^[20]

Table 4 presents the amnesia reporting rate per 100,000 prescriptions of triazolam and temazepam, the ratio of reporting rates, and the corresponding 95% confidence limits for each calendar year when either one of the drugs was in the market, as well as the statistics for all years pooled and summarized. The data were presented in the 1992 CDER Memorandum for Advisory Committee Meeting on Neuropharmaceutical Drug Products.

When comparing the reporting rates of two drugs first marketed in different years, one needs to consider the year factor that may bias the comparison in the following two ways. Due to the well-established evidence that the first three years of marketing result in many more ADE reports than later years (Fig. 1),^[22] regardless of the actual incidence of such reactions, one may need to stratify the data by marketing year in order to adjust for this effect (Table 5). However, once the data are stratified by the marketing year, one has to compare the reporting rates that occurred in different calendar years. The U.S. reporting of all ADEs of all drugs combined is not at a constant rate through all the years. There is a large variation in the reporting rates during the 1968 to 1994 interval (Fig. 2). Although it is difficult to explain this fully, part of the

Table 4 ADE Reports of Amnesia for Triazolam and Temazepam, by Calendar Year

Calendar year	Triazolam			Temazepam			95% CI		
	Number of reports	Rxs ^a	Rate ^b	Number of reports	Rxs	Rate	Rate ^c ratio	Lower limit	Upper limit
1981	0	0	0.0	1	872	1.1			
1982	0	0	0.0	1	3229	0.3			
1983	37	1971	18.8	1	4537	0.2	85.2	14.8	491.6
1984	59	4617	12.8	0	5046	0.0	INF	16.8	INF
1985	32	6870	4.7	0	5438	0.0	INF	6.6	INF
1986	16	9017	1.8	0	5498	0.0	INF	0.5	INF
1987	30	10,458	2.9	0	5424	0.0	INF	4.1	INF
1988	93	11,021	8.4	1	5283	0.2	44.6	7.8	254.5
1989	27	8744	3.1	2	5383	0.4	8.3	2.2	31.6
1990	59	7450	7.9	2	5451	0.4	21.6	5.8	80.2
1991	3	6399	0.5	0	5259	0.0	INF	0.6	INF
Pool	356	66,547	5.3	8	51,420	0.2	34.4	17.3	68.4
Mantel-Haenszel*	356	66,547		6	47,319		45.6	28.9	72.0

From the 1992 CDER Memorandum for Advisory Committee on Neuropharmaceutical Drug Product.

^aRxs: in 1000 prescriptions.

^bRate: number of reports per 100,000 prescriptions.

^cRatio of reporting rates = (rate of triazolam reports)/(rate of temazepam reports).

*Excluding 1981 and 1982 when there was no usage of triazolam.

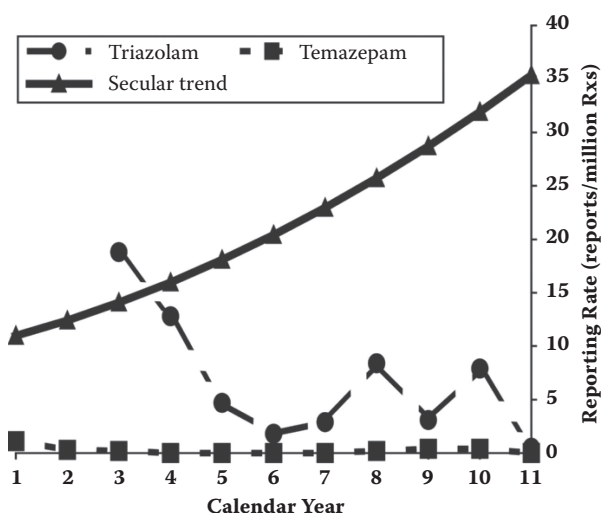


Fig. 1 Amnesia reporting rates of triazolam, temazepam, and all-drug-all-event reporting rate

decrease in all-drug-all-ADE reporting rates from 1968 to 1980 was due to the changes in policy on what to report, while the increase observed after 1980 was partially due to the promotion of the reporting system, the increased effort of the manufacturers, and the availability of more medicines. However, this increase in reports appears to be independent of manufacturer.

The trend with time changes of reports of all ADEs of all drugs combined can be fitted rather well with a quadratic regression curve for its good fit and simplicity. The data are best described with the quadratic regression curve

$$\text{Rate} = 10.96 + 1.34(\text{year} - 1981) + 0.11(\text{year} - 1981)^2$$

The expected all-drug-all-ADE reporting rates for each year are given in Table 5. In comparing the amnesia ADE reporting rates of triazolam and temazepam in their marketing years, one must need to adjust the ratio of rates according to the differences resulting from the secular trend of reporting of all ADEs of all drugs combined of the two calendar years in comparison. For example, in 1983, the first marketing year of triazolam, the reporting rate of all ADEs of all drugs combined was 1/0.78 of the rate in 1981, the first marketing year of temazepam.

Taking the reporting rate of all ADEs of all drugs combined in the corresponding marketing year of temazepam as the reference, the adjusting factor $f_{(j)}$ for the secular trend of reporting of the nine years between 1983 and 1991 can be calculated by (reporting rate of all ADEs of all drugs combined in the marketing year of temazepam)/(reporting rate of all ADEs of all drugs in the corresponding marketing year of triazolam) and given in the last column of Table 6.

Let the adjusting factor be denoted by $f_{(j)}$ for the j th marketing year of triazolam and temazepam, respectively; the ratio adjusted for the secular trend, $RR'_{(j)}$ is then defined by

$$RR'_{(j)} = r_{1(j)}f_{(j)} / r_{2(j)}$$

$P'_{(j)}$, the proportion of triazolam prescriptions adjusted for the secular trend, is $s_{1(j)} / (s_{1(j)} + s_{2(j)}f_{(j)})$. Here Z statistics and the confidence limits are calculated accordingly.

Table 5 ADE Reports of Amnesia for Triazolam and Temazepam, by Year of Marketing

Marketing year	Triazolam				Temazepam				Rate ratio ^c	Lower limit	Upper limit
	Calendar year	Number of reports	Rxs ^a	Rate ^b	Calendar year	Number of reports	Rxs ^a	Rate ^b			
1	1983	37	1971	18.8	1981	1	872	1.1	16.4	2.8	94.5
2	1984	59	4617	12.8	1982	1	3229	0.3	41.3	7.2	236.5
3	1985	32	6870	4.7	1983	1	4537	0.2	21.1	3.7	122.3
4	1986	16	9017	1.8	1984	0	5046	0.0	INF	—	—
5	1987	30	10,458	2.9	1985	0	5438	0.0	INF	—	—
6	1988	93	11,021	8.4	1986	0	5498	0.0	INF	—	—
7	1989	27	8744	3.1	1987	0	5424	0.0	INF	—	—
8	1990	59	7450	7.9	1988	1	5283	0.2	41.8	7.3	239.8
9	1991	3	6399	0.5	1989	2	5383	0.4	1.3	0.3	6.3
Pooled		356	66,547	5.3		6	40,710	0.1	36.3	16.5	79.7
Mantel-Haenszel		356	66,547			6	40,710		37.0	22.3	61.4
Reports prior to 1988											
Pooled		174	32,933	5.3		3	19,122	0.2	33.7	11.4	99.9
Mantel-Haenszel		174	32,933			3	19,122		34.2	16.8	69.4

^aRxs: in 1000 prescriptions.

^bRate: per 1 million prescriptions.

^c(Rate of triazolam)/(rate of temazepam).

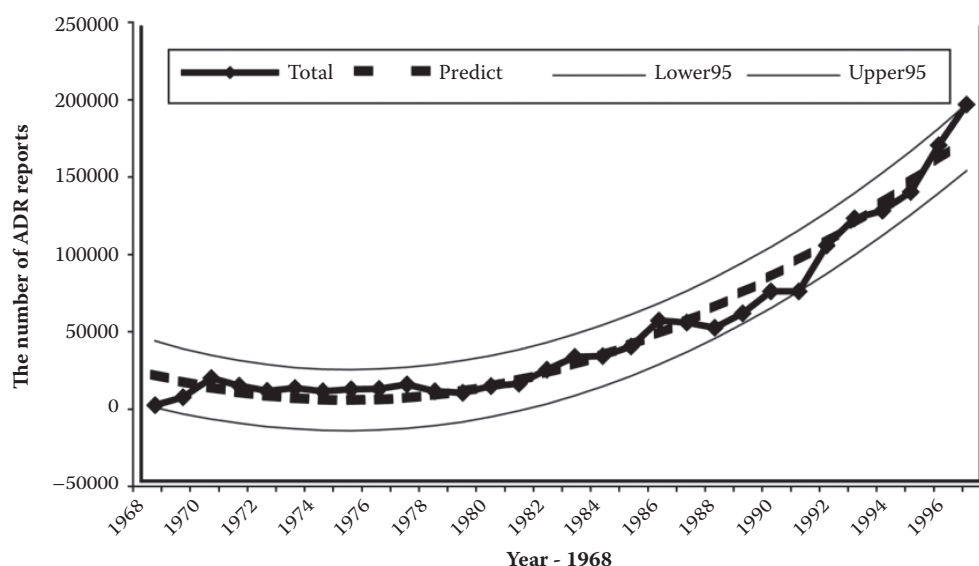


Fig. 2 All ADR reports with fitted regression and 95% confidence limits

Table 6 Expected Overall Reporting Rates and Adjusting Factor

Marketing year	Triazolam		Temazepam		Adjustment factor, $f_{(j)}^a$
	Calendar year	Overall rate	Calendar year	Overall rate	
1	1983	14.1	1981	11.0	0.78
2	1984	16.0	1982	12.4	0.78
3	1985	18.1	1983	14.1	0.78
4	1986	20.4	1984	16.0	0.78
5	1987	23.0	1985	18.1	0.79
6	1988	25.8	1986	20.4	0.79
7	1989	28.8	1987	23.0	0.80
8	1990	32.1	1988	25.8	0.80
9	1991	35.5	1989	28.8	0.81

$f_{(j)}^a$ = (Overall reporting rate for temazepam calendar year)/(overall reporting rate for triazolam calendar) for the same marketing year.

When combining all marketing years, the regular pooled ratio is calculated by

$$RR' = \frac{\left[\left(\sum_j x_{1(j)} \right) / \left(\sum_j s_{1(j)} \right) \right]}{\left[\left(\sum_j x_{2(j)} \right) / \left(\sum_j s_{2(j)} f_{(j)} \right) \right]} \quad (6)$$

The Mantel–Haenszel ratio and the Z statistics can be derived as

$$RR'_{MH} = \sum_j \left[x_{1(j)} (1 - P'_{(j)}) \right] / (x_{2(j)} P'_{(j)}) \quad (7)$$

$$Z^* = \frac{\sum_j x_{1(j)} - \sum_j x_{2(j)} P'_{(j)}}{\sqrt{\sum_j x_{2(j)} P'_{(j)} (1 - P'_{(j)})}} \quad (8)$$

Table 7 presents the reporting rates of the first nine marketing years of the two drugs, stratified by marketing year and adjusted for the secular trend.

After being adjusted for the marketing year and the secular trend of reporting, the amnesia reporting rate of triazolam was reduced from 45 folds (Mantel–Haenszel statistic in Table 3) to about 29 folds (Mantel–Haenszel statistic) of the rates of temazepam.

Note that:

1. in order to adjust for the marketing year, the data of the ADE reporting of temazepam in the marketing years beyond the latest marketing year of triazolam were not used;
2. the adjustment of secular trend is determinative without using the estimation error of the estimate;
3. as stated in the last section, usage of the drug is an estimate with standard error that was not accommodated in the analysis;

Table 7 ADE Reports of Amnesia for Triazolam and Temazepam, by Year of Marketing Adjusted for the Secular Trend of All Reporting

Marketing year	Triazolam				Temazepam				Rate ratio ^c	Lower limit	Upper limit
	Calendar year	Number of reports	Rxs ^a	Rate ^b	Calendar year	Number of reports	Rxs	Rate			
1	1983	37	1971	14.6	1981	1	872	1.1	12.7	2.2	73.6
2	1984	59	4617	9.9	1982	1	3229	0.3	32.1	5.6	183.7
3	1985	32	6870	3.6	1983	1	4537	0.2	16.4	2.8	95.2
4	1986	16	9017	1.4	1984	0	5046	0.0	INF	1.8	INF
5	1987	30	10,458	2.3	1985	0	5438	0.0	INF	3.2	INF
6	1988	93	11,021	6.7	1986	0	5498	0.0	INF	9.6	INF
7	1989	27	8744	2.5	1987	0	5424	0.0	INF	3.5	INF
8	1990	59	7450	6.4	1988	1	5283	0.2	33.7	5.9	193.0
9	1991	3	6399	0.4	1989	2	5383	0.4	1.0	0.2	5.1
Pooled		356	66,547	4.2		6	40,710	0.1	28.7	3.1	63.0
Mantel-Haenszel		356	66,547			6	40,710		29.0	20.8	65.7
Reports prior to 1988											
Pooled		174	32,933	4.1		3	19,122	0.2	26.3	8.9	78.0
Mantel-Haenszel		174	32,933			3	19,122		26.6	15.2	76.8

From the 2000 FDA Memorandum "Epidemiologic Review of PPA Safety Issues" for the FDA Advisory Committee on Over-the-Counter Drug Products.

^aRxs: in 1000 prescriptions.

^bPer 1 million prescriptions, triazolam rate adjusted for overall reporting rate difference in 2 years.

^c(Rate of triazolam)/(rate of temazepam).

- adjusting for marketing year and secular trend gives a more conservative signal for high ADE reporting than without adjustment of the drug of interest that is marketed more recently than its control;
- signaling high ADE reporting rate through comparison between drugs could be sensitive to the difference in patient population and the difference in ADE reporting pattern of the two drugs.

Generalized approaches for signaling of high ADE reporting rate in comparison with multiple control drugs were also proposed by Rossi et al.,^[13] Hsu,^[14] and Tsong.^[15]

SIGNALING REPORTING PROFILE DIFFERENCE

As noted in the last section, although the data are rich in the AERS system, there is no appropriate denominator information for incidence rate estimation. As a surrogate, the ADE reporting rate per million prescriptions is used in signaling the relative safety of a drug product with the appropriately selected comparative drug or drugs. The usage estimate used in the estimation is often obtained from data vendors. The process can be costly and time consuming. It is often unfeasible when the drugs studied had a long marketing history or when many drugs are used in the comparative group. In that case, sometimes signaling unusual reporting pattern may be informative and can be performed without using the estimation of patients or

usage. For example, safety concerns of stroke risk associated with treatment with phenylpropanolamine was signaled by the extremely large proportion of the ADE reports in a 2-by- k frequency table display of celebravascular artery (CVA) event report vs. other reports by drug product as in [Tables 8](#) and [9](#). The data were presented at the FDA Advisory Committee Meeting of Over-the-Counter Drug Products in October 2000.

Signaling of outlying reporting profile or testing for reporting profile difference between drugs is also of interest and often used in postmarketing surveillance monitoring. [Table 10](#) shows an example of the comparison of the ADE reporting profiles of triazolam and temazepam in order to assess concerns that the ADE profile of triazolam may be different from other drugs of the same class. A chi-square test was used for testing $H_0: P_{ik} = P_{2k}$, $k=1$ to K , with P_{ik} representing the proportion of reports of the k th ADE. The data were presented at the Meeting of the FDA Advisory Committee on Neuropharmaceutical Drug Products held in 1992.

Signaling of reporting profile difference was also often used in comparisons between genders, age groups, and two drugs.

DATA MINING FOR ADE SIGNALING

When a large database of spontaneous reports of adverse reactions is considered, there are other practical needs

Table 8 Phenylpropanolamine (PPA) and CVA Reports in Women Aged 10–59 Years (1977–January 1991)

Product	CVA	Non-CVA	Total
<i>Direct reports from medical professionals</i>			
PPA-diet	20 (40.8%)	29 (59.2%)	49
PPA-cough/cold	4 (5.7%)	66 (94.3%)	70
All non-PPA	78 (0.5%)	14,168 (99.5%)	14,246
Total	102 (0.7%)	14,263 (99.3%)	14,365
<i>Manufacturer report</i>			
PPA-diet	0 (0.0%)	3 (100%)	3
PPA-cough/cold	4 (5.1%)	74 (94.9%)	78
All non-PPA	643 (0.9%)	72,799 (99.1%)	73,442
Total	647 (0.9%)	72,876 (99.1%)	73,523

From the 2000 FDA Memorandum “Epidemiologic Review of PPA Safety Issues” for the FDA Advisory Committee on Over-the-Counter Drug Products.

Table 9 Direct Domestic Spontaneous Reports of CVA in Women Aged 10–59 Years Received by the FDA (1977–January 1991)

Product category	Number of CVA reports	Number of total reports
PPA-diet	19	26%
Oral contraceptive	15	20%
Thrombolytic agent	7	10%
Lactation suppressive	6	8%
Chemotherapeutic	5	7%
Radiocontrast	4	5%
Anticoagulant	3	4%
PPA-cough/cold	3	4%
Miscellaneous	11	15%
Total	73	100%

From the 2000 FDA Memorandum “Epidemiologic Review of PPA Safety Issues” for the FDA Advisory Committee on Over-the-Counter Drug Products.

for the events evaluator beyond statistics. The first is the ability to assess the patterns of the patients’ experience with the drug product. The second is the need to advance the understanding of drug-associated events. Thus techniques that will enhance the ability to address these needs, such as data mining and visualization tools, have been introduced.^[16,17,22]

The approach that the FDA began to explore in 1998 was originally developed by William DuMouchel^[16] of AT&T. In the last four years, refinements have been made to the methods, the speed of access to the database, and the performance of the computations. The approach was also enhanced by coupling the data-mining software with visual graphics software for displaying the results. The data-mining model developed by the FDA^[17,18] consists of the following components:

1. FDA’s routinely updated structured database of ADE reports.

2. A model to extract AERS data records and format information for data-mining analysis.
3. Data-mining statistical algorithms.
4. Custom-designed software to display several kinds of visual graphics.

Pairwise Association

The basic statistic consideration is to consider the complete database as a large frequency table with drug product (more than 1400 products) as the row factor and AE (more than 950 different events) as the column (Table 11). The column margins, D_i ’s, are the total number of reports of the events. The row margins, A_j , are the total number of reports of the drugs.

When there is no association between drug and event of the reports in the database, p_{ij} , the proportion of the reports of a specific event j , among all reports of the drug product i , should be equal to A_j/n , the proportion of all reports of the specific event among all reports of all drugs. In other words, under the “no-association” hypothesis, the expected number of reports of event j of a given drug i is the proportion of all reports in the selected event of all $E_{ij}=D_i A_j/n$. The association is measured by

$$(O/E)_{ij} = n_{ij}/E_{ij} = n_{ij}n/D_i A_j \quad (9)$$

This association is the measurement of information connecting the i th drug to the j th event in the reporting system.

The empirical Bayes procedure introduced by DuMouchel^[16] in ADR data mining would consider each n_{ij} as a sample observation from a Poisson distribution with unknown mean μ_{ij} . The interest in data mining is focussed on $\lambda_{ij}=\mu_{ij}/E_{ij}$. The approximately 1400·952 λ_{ij} ’s are sampled from a common prior distribution that is a mixture of two gamma distributions as follows,

$$\pi(\lambda; \alpha_1, \beta_1, \alpha_2, \beta_2, P) = Pg(\lambda; \alpha_1, \beta_1) + (1-P)g(\lambda; \alpha_2, \beta_2)$$

Table 10 Distribution of Adverse Events Reported

Event	Number of triazolam, ADR	Percent of total, ADR	Number of temazepam, ADR	Percent of total, ADR
Agitation	250	13.5	21	20.8
Amnesia	356	19.2	6	5.9
Psychosis	327	17.7	15	14.9
Confusion	400	21.6	20	19.8
Hostility	101	5.5	2	2.0
Seizure	55	3.0	1	1.0
Fatality	210	11.4	28	27.7
Depression	106	5.7	7	6.9
Psycho/depression	45	2.4	1	1.0
Statistics for table of event by drug				
Statistic	DF	Value	P value	
Chi-square	8	39.0	0.001	
Likelihood ratio chi-square	8	37.7	0.001	

From the 1992 CDER Memorandum for Advisory Committee on Neuropharmaceutical Drug Products.

Table 11 Drug-Event Report Frequency in SRS Database

Drug	Adverse event						Total
	1	2	...	J	...	952	
A	n_{a1}	n_{a2}		n_{aj}		n_{a952}	D_a
B	n_{b1}	n_{b2}					D_b
⋮							⋮
I				n_{ij}			D_i
⋮							⋮
1400							D_{1400}
⋮							⋮
	A_1	A_2	...	A_j	...	A_{952}	n

where $g(\lambda; \alpha, \beta)$ is the density of the gamma distribution with mean $= \alpha/\beta$ and variance $= \alpha^2/\beta$, and P is the proportion of mixture.

$$g(\lambda; \alpha, \beta) = \beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda} / \Gamma(\alpha)$$

The five parameters ($\alpha_1, \beta_1, \alpha_2, \beta_2, P$) are determined by the observed data. For a given λ and E , the marginal distribution of each N_{ij} is a mixture of negative binomial distributions,

$$\Pr(N_{ij} = n_{ij}) = Pf(n_{ij}; \alpha_1, \beta_1, E_{ij}) + (1-P)f(n_{ij}; \alpha_2, \beta_2, E_{ij})$$

with

$$\begin{aligned} f(n_{ij}; \alpha, \beta, E_{ij}) \\ = (1 + \beta/E_{ij})^{-n} (1 + E_{ij}/\beta)^{-\alpha} \Gamma(\alpha + n_{ij}) / \Gamma(\alpha) n_{ij}! \end{aligned}$$

In addition, the posterior distribution of each λ after observing n_{ij} is a mixture of two gamma distributions with modified parameters,

$$(\lambda | N_{ij} = n_{ij}) \sim \pi(\lambda; \alpha_1 + n_{ij}, \beta_1 + E_{ij}, \alpha_2 + n_{ij}, \beta_2 + E_{ij}, Q(n_{ij}))$$

where $Q(n_{ij})$ is the posterior probability that λ came from the first component of the mixture, given n_{ij} . It is given by

$$\begin{aligned} Q(n_{ij}) = Pf(n_{ij}; \alpha_1, \beta_1, E_{ij}) / [Pf(n_{ij}; \alpha_1, \beta_1, E_{ij}) \\ + (1-p)f(n_{ij}; \alpha_2, \beta_2, E_{ij})] \end{aligned}$$

The posterior expectation of λ and $\log(\lambda)$ given n_{ij} are given by

$$\begin{aligned} E[\lambda | N_{ij} = n_{ij}] = Q(n_{ij})(\alpha_1 + n_{ij}) / (\beta_1 + E_{ij}) \\ + (1 - Q(n_{ij}))(\alpha_2 + n_{ij}) / (\beta_2 + E_{ij}) \end{aligned}$$

and

$$\begin{aligned} E[\log(\lambda) | N_{ij} = n_{ij}] = Q(n_{ij})[\Psi(\alpha_1 + n_{ij}) \\ - \log(\beta_1 + E_{ij})] + (1 - Q(n_{ij})) \\ \times [\Psi(\alpha_2 + n_{ij}) - \log(\beta_2 + E_{ij})] \end{aligned}$$

where $\Psi(x)$ is the digamma function. The empirical Bayes measure used in data mining cell counts ranking is defined as

$$\begin{aligned} \text{EBlog} 2_{ij} = E[\log_2(\lambda_{ij}) | N_{ij} = n_{ij}] \\ = E[\log(\lambda_{ij}) | N_{ij} = n_{ij} / \log(2)] \end{aligned}$$

The quantity $\text{EBlog} 2_{ij}$ is a Bayesian version of $\log_2(O/E)_{ij}$.

For large E_{ij} and $(O/E)_{ij}$, $\Psi(x) \approx \log(x)$, and $\log(\alpha + n_{ij}) - \log(\beta + E_{ij}) \approx \log(n_{ij}/E_{ij}) = \log(O/E)_{ij}$. A quick-and-dirty 95% approximate confidence interval of λ_{ij} was proposed by DuMouchel^[18] as follows,

$$\text{EBGM}_{ij} \exp\left\{-2/\sqrt{(n_{ij}+1)}\right\} \lambda_{ij}$$

$$\text{EBGM}_{ij} \exp\left\{2/\sqrt{(n_{ij}+1)}\right\}$$

where $\text{EBGM}_{ij} = 2^{\text{EBlog}2ij}$.

Rankings of $\text{EBlog}2_{ij}$ or EB05 ($\approx \text{EBGM}_{ij} \exp\{-2/\sqrt{(n_{ij}+1)}\}$), the lower 95th confidence limit of λ_{ij} , are used for signaling purposes.

Multi-Item Association

When dealing with the reports of multiple drugs and events, the interest is on the number of reports instead of the independent drug–event combinations as used in $\text{EBlog}2_{ij}$. To address this, DuMouchel and Pregibon^[23] proposed to consider the problem within the framework of the so-called market basket problem. Assume that there is a maximum of three drugs and K ADEs mentioned in each report, under such framework, let n be the total number of reports, $i(1)$, $i'(2)$, $i'''(3)$, be the index of the first to the third drug associated with the report with the restriction that $i(1) < i'(2) < i'''(3)$, and $j_1, j_2, \dots, j_K$ be the index of the first to K th ADE mentioned in the report with the restriction that $j_1 < j_2 < \dots < j_K$, respectively. Let the frequency of the item sets of size 1 drug and 1 ADE up to four drugs and 4 ADE be denoted by $n_{i(1),j_1}$, $n_{i(1),i'(2),j_1}$, $n_{i(1),j_1,j_2}$, ..., $n_{i(1),i'(2),i'''(3),j_1,j_2,j_K}$, respectively. Note that

$$D_i = \sum_{i(1)=i, \text{ or } i'(2)=i, \text{ or } i'''(3)=i} \left[n_{i(1),i'(2),j_1} + n_{i(1),j_1,j_2} + \dots + n_{i(1),i'(2),i'''(3),j_1,j_2,j_K} \right]$$

$$A_j = \sum_{j_1=j, \text{ or } j_2=j, \text{ or } \dots, \text{ or } j_K=j} \left[n_{i(1),i'(2),j_1} + n_{i(1),j_1,j_2} + \dots + n_{i(1),i'(2),i'''(3),j_1,j_2,j_K} \right]$$

Then, proportions of a given drug i or a given ADE j , under the hypothesis that there is a null association between drug and ADE, are

$$P_i = D_i/N$$

$$P_j = A_j/N$$

respectively.

The expected frequencies for the drug-ADE item under the null association hypothesis are

$$E_{ij} = N P_i P_j, E_{iij'} = N P_i P_j P_{j'}, E_{iij'j''} = N P_i P_j P_{j'} P_{j''}, \dots$$

Hence (O/E) 's can be calculated. The multi-item empirical Bayes shrinkage of (O/E) 's can be carried out using the same procedure as described in the "Pairwise Association" section.

The drug-ADE of all cells in the database is ranked by the size of the Bayes shrinkage of (O/E) . A preliminary signal is generated when the adjusted measure association has a value greater than a predetermined threshold. A preliminary signal is then examined by reviewing the information displayed via computer graphics before a signal is issued for verification information.

As discussed in the last section, data-mining signaling is not signaling safety concerns based on incidence estimates (i.e., adjusted for patient population) or the surrogate estimates such as the reporting rate (i.e., adjusted for usage). It is rather signaling the outlying or unusual reporting of a specific drug-ADE among the ADE reporting of the specific drug in the sense that the reporting profile of this specific drug is outstanding. The difference between the results of the two types of signaling may be illustrated with the data of the reporting of the central nerve system or the psychologic ADE of triazolam and temazepam (see Table 12 and Table 13). For the nine ADEs, the usage adjusted reporting rate of triazolam ranks higher than the rate of temazepam (Table 12).

However, without usage adjustment, of three ADEs (i.e., agitation, depression, and fatality) the O/E of temazepam ranks the highest among the 18 O/E in Table 13. The ratios of the rates (Fig. 3) have the same profile as that of the ratio of O/E 's (Fig. 4) but with much smaller magnitude. In contrast to triazolam, the ratio of O/E is less than 1 in three ADEs. The difference in magnitude can be explained by the difference in usage and the difference in the reports of the two drugs. Actually, the ratio of the number of ADE reports of triazolam to temazepam, 18.32 (=1850/101), is 11.24 times the ratio of the usage, 1.63 (=66,547/40,710). The ratio equals exactly the ratio of the reporting rate ratio to the O/E ratio of any ADE shown in Figs. 3 and 4. The signaling based on O/E would be the same as the signaling based on reporting rate (which is often considered as surrogate of incidence density rate) when the ratio equals 1.

Szarfman^[18,19] reported the correlation between usage and the number of ADE reports of some selected drug categories. The correlation ranges from 40% in antineoplastic and in antiviral drug class to over 90% in each of cardiac, antihypertensive, antidiabetics, antilipemics, antihistamines, lung smooth-muscle relaxants, and anticoagulants drug class. The overall correlation is about 63%. This suggested the heterogeneity of the reporting rates among the drug classes. In addition, it is also logical to consider that the reporting profiles in terms of what to report are different among the drug classes. The heterogeneity among the drug classes suggests that refinement needs to focus on the subdatabase instead.

Table 12 Reporting Rate of Drug-ADE

Event	Triazolam			Temazepam			Rate ratio ^c
	ADRs ^a	Rate	Rank ^b	ADRs	Rate	Rank	
Agitation	250	3.76	4	21	0.52	12	7.28
Amnesia	356	5.35	2	6	0.15	15	36.30
Psychosis	327	4.91	3	15	0.37	13	13.34
Confusion	400	6.01	1	20	0.49	11	12.23
Hostility	101	1.52	6	2	0.05	16	30.89
Seizure	55	0.83	8	1	0.02	17	33.65
Fatality	210	3.16	5	28	0.69	10	4.59
Depression	106	1.59	7	7	0.17	14	9.26
Psycho/ depression	45	0.68	9	1	0.02	18	27.53
Usage	66,547			40,710		1.63	11.24

From the 1992 CDER Memorandum for Advisory Committee on Neuropharmaceutical Drug Products.

^aTotal number of reports over the first nine marketing years of triazolam.

^bRank among all rates in the table.

^cRatio of triazolam reporting rate to temazepam reporting rate.

Table 13 Ratio of Observed to Expected Number of Reports of ADE

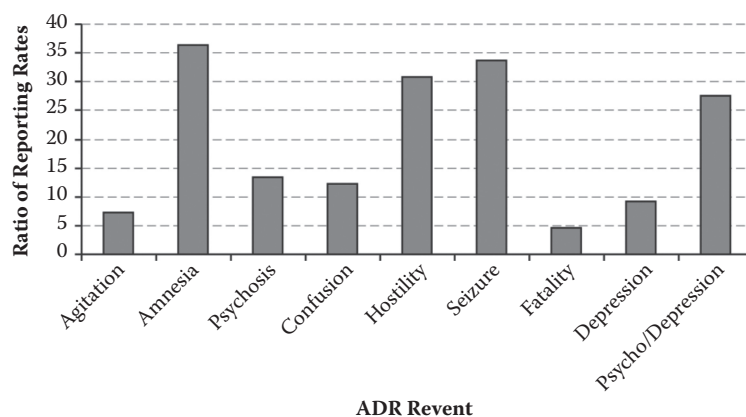
Event	Triazolam			Temazepam			Ratio of O/E ^c
	ADR ^a	O/E	Rank ^b	ADR	O/E	Rank	
Agitation	250	0.97	11	21	1.50	2	0.65
Amnesia	356	1.04	4	6	0.32	18	3.25
Psychosis	327	1.01	8	15	0.85	14	1.19
Confusion	400	1.00	9	20	0.92	13	1.09
Hostility	101	1.03	6	2	0.38	16	2.71
Seizure	55	1.04	5	1	0.34	17	3.06
Fatality	210	0.93	12	28	2.27	1	0.41
Depression	106	0.99	10	7	1.20	3	0.85
Psycho/ depression	45	1.03	7	1	0.42	15	2.45
All ADE	1850			101			18.32

From the 1992 CDER Memorandum for Advisory Committee on Neuropharmaceutical Drug Products.

^aTotal number of reports over the first nine marketing years of triazolam.

^bRank among all O/E's in the table.

^cRatio of triazolam O/E to temazepam O/E.

**Fig. 3** Ratio of ADR reporting rates of triazolam and temazepam (first nine marketing years)

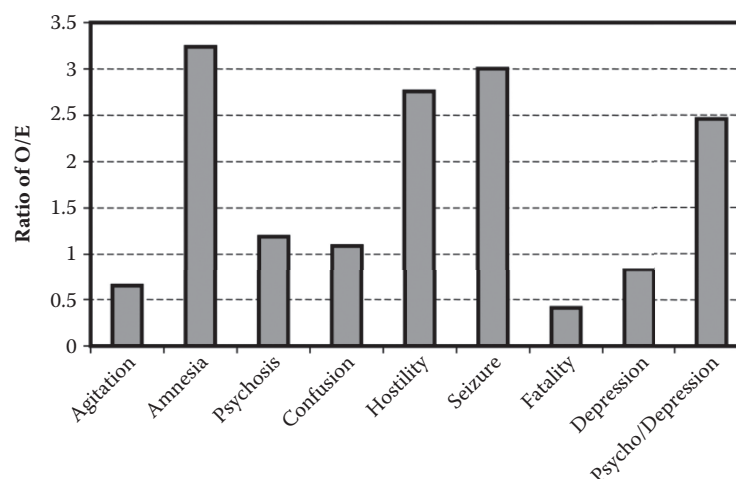


Fig. 4 Triazolam–temazepam ratio of *O/E*

CONCLUSION

The history of postmarketing adverse drug event signaling has been long and the spontaneous reporting database has been criticized, ignored, and often misused over the years. Yet because of its size of information and timeliness, it saved countless lives by the alerts generated from it and the follow-through using this imperfect system. The experienced FDA spontaneous report handlers and analysts can often generate an initial drug-event signal by simply noticing the unusual reporting pattern, unusual changes of the reporting pattern, or the high volume of reports of a drug among the drugs of a similar category or for the similar indications. For example, if a drug was approved for marketing under some stringent risk management plan, an increased frequency of reporting is a signal of potential failure of the risk management plan. There are numerous examples where a sponsor withdrew a drug from the market due to the follow-through initiated by a strong early marketing signal generated from a simple table such as Table 7 or Table 8. In the case of our example, PPA was used as the over-the-counter drug product under the grandfather law and manufacturers were not required by law to submit the report to the FDA. Most of the cases were reported directly by medical professionals. Because of the weakness of the spontaneous reporting data, the strength of the initial signal was not well received by professionals. The risk was verified by a large scale prospective case-control study conducted by a research group lead by a Yale University epidemiologist to verify risk. It took almost ten years for the study design, data collection, and analysis to be completed.

There are many versions of the methods similar to the one reported in the section “Signaling of High Reporting Rate,” for detecting the signal of a high reporting rate of a given drug-event. The one reported in this article is one of the most stringent procedures to adjust for the factors that were estimable. The results of Table 7 were interpreted

based on the assumption that the patient populations and prescription patterns of the two drugs were comparable. A calibration constant may be used when the prescription patterns are different. In fact, when the clinical trial data were reanalyzed, a better understanding of equivalent dose of the two drugs was taken into consideration and the relative risk was found much smaller and not statistically significant.

Supporters of using signaling procedure without estimate of patient population often criticize the precision and reliability of the estimate. However, what was used in the signaling procedures of increased frequency of reporting and high reporting rate are the ratio of usage estimates that serve as the estimate of ratio of patient population sizes. In practice, once a drug of interest is identified for monitoring, the FDA will take steps to collect information on the usage of the drug of interest and its comparison drug(s).

As stated earlier, the interpretation of a signal for risk requires the estimate of ratio of patient populations. Signaling procedure based on data-mining is a useful tool to generate signals for an unusual reporting profile. Basically, it is an early signaling procedure before a drug of interest is identified for monitoring. However, the capability of using a data-mining tool to study drug–drug interaction in a large database makes its utility much more promising for some advanced application in post-marketing adverse event signaling.

REFERENCES

1. Faich, G.A. Adverse drug reaction monitoring. *New Engl. J. Med.* **1986**, *314* (24), 1589–1592.
2. Anello, C. Regulatory scene, postmarketing surveillance legislation: Structure and results. *Pharm. Med.* **1987**, *2*, 11–22.
3. Pearson, K.C.; Kennedy, D.L. Adverse drug reaction and the Food and Drug Administration. *J. Pharm. Pract.* **1989**, *11* (4), 209–213.

4. *Draft Guidelines for Postmarketing Reporting of Adverse Drug Reactions*; U.S. Food and Drug Administration: Rockville, MD, 1985.
5. *Guidelines for Postmarketing Reporting of Adverse Drug Reactions*; U.S. Food and Drug Administration: Rockville, MD, 1992.
6. Anello, C. Management of ADR reports at the FDA. *Drug Inf. J.* **1985**, *19*, 291–294.
7. Medical Dictionary for Regulatory Activities, Version 4.0. International Federation of Pharmaceutical Manufacturers Association, 2001.
8. Knapp, D.E.; Zax, B.B.; Rossi, A.C.; O'Neill, R.T. A method for postmarketing screening of adverse reactions to drugs: Initial results. *Drug Intell. Clin. Pharm.* **1980**, *14*, 24–27.
9. O'Neill, R.T.; Harris, N.D. The FDA Screening of Adverse Reaction Reports—Some Statistical Considerations. A Method for Postmarketing Screening of Adverse Reactions to Drugs. *CDER Division of Biometrics Report*; FDA: Rockville, MD, 1985.
10. Norwood, P.K.; Sampson, A.R. A statistical methodology for drug surveillance of adverse drug reaction report. *Stat. Med.* **1988**, *7*, 1023–1030.
11. Tsong, Y. False alarm rates of statistical methods used in determining increased frequency of reports on adverse drug reactions. *J. Biopharm. Stat.* **1992**, *2* (1), 9–30.
12. Lao, C.S.; Anello, C.; Markovitz, B.A. Application of CUSUM Technique and Beta-Binomial Model in Monitoring Adverse Drug Reactions. *The Proceedings of Biopharmaceutical Section*; American Statistical Association, 1989.
13. Rossi, A.C.; Hsu, J.P.; Faich, G.A. Ulcerogenicity of piroxicam: Analysis of spontaneously reported data. *Br. Med. J.* **1987**, *294*, 147–150.
14. Hsu, J.P. Refinement of the Methods of Analysis for the NSAID Study. *CDER, Division of Biometrics Memorandum*; FDA: Rockville, MD, 1985.
15. Tsong, Y. Comparing reporting rates of adverse events between drugs with adjustment for year of marketing and secular trends in total reporting. *J. Biopharm. Stat.* **1995**, *5* (1), 95–114.
16. DuMouchel, W. Bayesian data mining in large frequency tables, with an application to the FDA Spontaneous Reporting System. *Am. Stat.* **1999**, *53* (3), 177–189.
17. O'Neill, R.T. Some US Food and Drug Administration perspective on data mining for pediatric safety assessment. *Curr. Ther. Res.* **2001**, *62* (9), 650–663.
18. Szafrman, A. The Application of Bayesian Data Mining and Graphic Visualization Tools to Screen FDA's Spontaneous Reporting System Database. *Proceedings of the Section on Bayesian Statistical Science*; American Statistical Association, 2000, 67–71.
19. Szafrman, A.; Talarico, L.; Levine, J. Analysis and Risk Assessment of Hematological Data from Clinical Trials. *Comprehensive Toxicology*. Sipes, I.G.; McQueen, C.A.; Gandolfi, A.J., Eds.; In *Toxicology of Hematopoietic System*; Elsevier Science, Inc.: New York, 1997, Vol. 4, 363–379.
20. Rothman, K.J.; Boice, Jr., J.D. *Epidemiologic Analysis with a Programmable Calculator*; Epidemiology Resources, Inc.: Boston, MA, 1982.
21. Webber, J.C.P. Epidemiology of Adverse Reactions to Nonsteroidal Anti-Inflammatory Drugs. In *Advances in Information Research*; Rainfold, K.D.; Velo, G.P., Eds.; Raven Press: New York, 1984, Vol. 6.
22. Jones, J.K. The role of data mining technology in the identification of signals of possible adverse drug reactions: Value and limitations. *Curr. Ther. Res.* **2001**, *62* (9), 664–673.
23. DuMouchel, W.; Pregibon, D. Empirical Bayes Screening for Multi-Item Associations. *Proceedings of the Conference on Knowledge Discovery and Data*; ACM Press: San Diego, CA, 2000, 67–76.

Postmarketing Surveillance

Patricia B. Cerrito

University of Louisville, Louisville, Kentucky, U.S.A.

Abstract

Once a drug is released on the market, it can be used for a variety of illnesses that were not in the original randomized testing process. Therefore, there are several questions that must be answered through postmarketing surveillance. This entry will focus on investigative techniques to examine these issues.

INTRODUCTION

Drugs are not approved for general use until they go through an elaborate testing period. The “gold standard” of randomized, controlled trials must be utilized for drugs to be granted approval. However, once a drug is released on the market, it can be used for a variety of illnesses that were not in the original randomized testing process. It can be prescribed to a general population that was previously excluded from the study subjects. Therefore, there are several questions that must be answered through postmarketing surveillance. Problems to be considered include (1) the effects on previously excluded subpopulations, (2) the cost effectiveness of the new drug compared to older treatments, and (3) the occurrence of previously unknown adverse events. This article will focus on investigative techniques to examine these issues.

OVERVIEW

The first such problem is to determine the effects of the drug on populations previously excluded from the randomized trials. Two populations that are virtually routine to exclude are children and pregnant women. Studies generally recruit few in minority populations.^[1] This can usually be extended to the exclusion of women who might become pregnant; that is, to exclude women of childbearing age. However, once a drug is approved for use, exclusion criteria no longer exist. A drug can be prescribed for anyone for any problem. A physician may have to assume liability when prescribing a drug for a nonapproved use, unless sufficient support exists in the medical literature for treatment. There is no liability for extending use to a segment of the population not previously studied.

Clinical trials by their nature must have an ending point. Usually, even Phase III clinical trials last a few weeks,

possibly a few months, rarely several years. However, particularly if the medication is prescribed for chronic problems, patients can take the medication for 30, 40, or 50 years. There might be possible long-term effects that are not seen in the short-term trials. Also, adverse reactions are rare; they will not always become apparent in clinical trials. The power analysis that determines the sample size for the clinical trial is defined in terms of the main effect; a much larger sample would be required to examine rare occurrences. This must be carefully monitored in postmarketing surveillance.

Another problem to be examined is that of cost-effectiveness. A new drug does not ordinarily treat a new illness; it usually has a claim to be more effective or safer than drugs currently in use. However, sometimes the effects are relatively similar. In this case, it is necessary to examine the cost of providing the benefit to the patient. This becomes particularly crucial as health management plans are devising formularies and limiting the drugs that can be used for treatment.

Each of these problems must be addressed through methodology and through data collection concerning drugs as they are used on the general population in postmarketing surveillance. For the researcher, statistical methods and data-mining techniques that can be used on observational data must be used. For the practicing clinician, there must be a way to record data from patients. For both, there is need to understand information in the medical literature that reports on postmarketing surveillance results.

DATA COLLECTION

The quality of the data generally depends upon the examination of the four parameters of hypothesis testing: significance level, power, effect size, and sample size. There are a number of ways that data can be collected in

Table 1 Quality of Evidence Collected in Postmarketing Surveillance

Level of evidence	Description	Treatment recommendation
I	Large, phase IV randomized, controlled trials patients	A. Should prescribe treatment to patients
II	Smaller randomized, controlled trials with low power so that the results are not necessarily conclusive	B. Should prescribe treatment to patients
III	Nonrandomized trial; contemporaneous controls	C. In the absence of severe adverse effects, should prescribe treatment
IV	Nonrandomized trial; historical controls	C
V	Uncontrolled studies, case series	C

From Ref. [2].

postmarketing surveillance. [Table 1](#) provides a chart of the quality of evidence.^[2] When there is any evidence to indicate that a patient might benefit without undue risk, the treatment should be tried, particularly if that treatment is not only very low risk but also of low cost.

It is important to examine any study for methodological flaws or for insufficient sample size, particularly if the results are negative. The power of a statistical test is the probability that it will lead to the rejection of the null hypothesis. It is desirable, although impossible, to have a small level of significance and a large power simultaneously without specifying a level of indifference (where two drug treatments are so close in effect that the investigator is indifferent as to outcome). Then the sample size is chosen. However, when small numbers of patients are used, the power will be low.

Observational studies (which include all studies without randomized controls) are particularly subject to confounding. This occurs when two variables examined are directly related to a third variable unexamined. The possible existence of a confounding factor can completely void the conclusions of the study. Therefore, great care must be taken to collect data on any known or suspected confounders.

POSTMARKETING QUESTIONS

Traditional methods of statistical analysis have focused on a very small number of variables^[3] utilizing a technique such as Bayesian analysis or logistic regression. More recently, data-mining techniques such as neural networks, fuzzy logic, and decision trees have been used:

Traditional model building is based on the use of rigorous statistical techniques such as multiple linear regressions. In contrast, neural networks are “adaptive computing,” in that they learn from data to build a model to explain an application or biological system. (There are no a priori assumptions of a linear relationship among variables.) A neural network is designed from the application, not the other way around. Applications may be simple or complex and linear or nonlinear such as in finance or other noisy systems (i.e., highly variable) where the relationship between variables is not in a straight line.^[4]

And stated in Ref. [5]:

The experienced statistician, perhaps the most capable of guiding the development of automated tools for data analysis, may also be the most acutely aware of all the difficulties that can arise when dealing with real data. This hesitation has bred skepticism of what automated procedures can offer and has contributed to the strong focus by the statistical community on model estimation to the neglect of the logical predecessor to this step, namely model identification.

The statistician’s tendency to avoid complete automation out of respect for the challenges of the data, and the historical emphasis on models with interpretable structure, has led that community to focus on problems with a more manageable number of variables (a dozen, say) and cases (several hundred typically) than may be encountered in statistical problems, which may be orders of magnitude larger at the outset. With increasingly huge and amorphous databases, it is clear that methods for automatically hunting down possible patterns worthy of fuller, interactive attention are required. The existence of such tools can free one up to, for instance, posit a wider range of candidate data features and basis functions (building blocks) than one would wish to deal with, if one were specifying a model structure “by hand.”

It is becoming increasingly clear that the more traditional techniques are not adequate in postmarketing surveillance. The algorithms developed through data mining examine databases with a set of built-in processes that automatically test hypotheses to generate interesting and unexpected rules and graphs that characterize the database. The automatic hypotheses-formation and -testing cycle continues until important rules and patterns emerge. It is up to the investigator to verify the significance of each rule and pattern.

Once the model has been identified, the verification process needs to begin. An outline of the process is given in [Table 2](#).^[6]

Drug Interactions

There are a number of statistical models that can be used to determine risk factors associated with medication use in the general population. The primary difficulty is to determine the approximate size of the population taking the medication. Unlike other developed nations, the United

Table 2 Process for Verifying an Identified Model

Procedure	Rationale
Developing an understanding of the application domain, the relevant prior knowledge, and the goals of the end user	What are the goals? What performance criteria are important? Will the final produce of the process be classification, visualization, summarization, prediction? Is understandability an issue? What is the trade-off between simplicity and accuracy?
Creating a target data set, selecting a data set, or focusing on a subset of variables or data samples on which discovery is to be performed	This involves consideration of homogeneity of the data, change over time, sampling strategy, etc.
Data cleaning and preprocessing	This involves the removal of noise or outliers, deciding on strategies for handling missing data fields, etc.
Data reduction and transformation	This involves finding useful features to represent the data, depending on the goal of the task; using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data; and projecting the data onto spaces in which a solution is likely to be easier to find.
Choosing the data-mining task	Deciding whether the goal is classification, regression, clustering, summarization, modeling, etc.
Data mining	The actual analysis, automated and most often done with software; includes neural network analysis.
Evaluating output	Determining what is actually knowledge and what is “fool’s gold.” Knowledge must be filtered from other outputs. This can be done by using statistical checks, visualization, or expertise.
Incorporating knowledge into the performance system	Checking and resolving potential conflicts with previously extracted knowledge.

From Ref. [6].

States does not have a national prescription registry to make it possible to track such information.

Once a population estimate has been made, the following variables need to be collected:^[7]

D = dosage and schedule taken over the time interval $[0, T]$

A = set of patient attributes

E = adverse event type

Then it is possible to estimate the conditional probability of E given the values of D , A , and T . There are a number of methods suggested for making an estimate. One that takes into consideration the fact that reported data are often dependent in various ways is that given in Cerrito and Cerrito.^[8] Define the vector (X, Y, T) , where

X = drug benefit

Y = drug risk (Table 3)

T = duration of medication use

The data collected are assumed to be from a weakly dependent set. In this way, the patient attributes needed in the model of Ref. [7] can be omitted. Then the probability density function of the vector (X, Y, T) can be estimated by

$$\hat{f}(x, y, t) = \frac{1}{na_n^3} \sum_{j=1}^n k\left(\frac{x - X_j}{a_n}, \frac{y - Y_j}{a_n}, \frac{t - T_j}{a_n}\right)$$

where

$k(\dots)$ = known multivariate function, called the kernel function

$(X_1, Y_1, T_1), \dots, (X_n, Y_n, T_n)$ = collected observations

a_n = constant smoothing parameter depending on n

Suppose it is only possible to define risks in terms of discrete categories $\{1, 2, \dots, m\}$. Suppose category 1 represents the most serious risk and category m represents the least serious so that the risks can be ordered. The standard scale is as shown in Table 3.^[9] Then a discrete estimator developed by Ahmad and Cerrito^[10] can estimate the discrete risks at each interaction level:

$$\hat{f}(x, y, t) = \frac{1}{na_n^2} \sum_{j=-\infty}^{\infty} W(\hat{s}_n, i, j) \sum_{l \in C_n(l)} k\left(\frac{x - X_l}{a_n}, \frac{t - T_l}{a_n}\right)$$

Table 3 Standard Scale for Risk

Significance rating	Severity	Documentation
1	Major: The effects are potentially life-threatening or capable of causing permanent damage	Suspected or established
2	Moderate: The effects may cause a deterioration in a patient's clinical status; additional treatment may be necessary	Suspected or established
3	Minor: The effects are usually mild; consequences may be bothersome or unnoticeable	Suspected or established
4	Major/moderate	Possible
5	Minor	Possible
	Any	Unlikely

From Ref. [9].

where $C_n(j) = \{\text{indices } l \text{ such that } (X_l, Y_l, T_l) \text{ is such that } Y_l \text{ is in category } j\}$. The function $W(s, i, j)$ is a discrete window weight function defined on $S \times J \times J$, where S is an interval on the real line and $J = \{\dots, -1, 0, 1, \dots\}$ satisfying^[11]

$$W(0, i, j) = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

It has been shown^[12] that the choice of the kernel function will not significantly alter the outcome. The choice of the smoothing parameter can have significant impact. Techniques have been developed to optimize the choice of this parameter. However, because the outcomes are discrete, various smoothing parameters can be attempted and investigated using ROC curves (for exactly two outcomes)^[13] or by using the standard training/testing/validation process routinely used in neural network analysis.^[14] The misclassification rate (m1) of the training set is compared to the misclassification rate of the testing set (m2). If $m1 < m2$ then the value of a_n is adjusted upward. The goal is to have m1 and m2 within a small interval of each other.

The commercially available NeuralWare II from Aspen-Works, Inc., does perform kernel density estimation. In addition, PROC DISCRIM and PROC KDE in SAS from SAS Institute, Inc., will perform kernel density estimation. For an adaptation of the discrete-continuous parameter, the symbolic computational software Mathematica from Wolfram, Inc., can easily be used to develop the necessary programming code to perform the required algorithm. Suggestions of such Mathematica code are given in the [Appendix](#) to this entry.

Excluded Populations

Although some populations (pregnant women and children) are routinely excluded from clinical trials because of risk and liability issues, still other groups are indirectly excluded because of small numbers of volunteers. Yet it is clear that race and ethnicity do have an impact on the pharmacokinetics of drugs.^[1] Bioavailability (F) has three determinants, $F = f_a f_g f_h$, where

f_a = fraction of drug absorbed
 f_g = fraction escaping gut metabolism
 f_h = fraction escaping hepatic first-pass metabolism

Studies that have compared drug absorption for different races have demonstrated that there is no difference when drug absorption is passive. However, when the drug undergoes active absorption, there are statistically significant differences. Studies comparing racial groups have had different outcomes for different drugs, making generalizations impossible.^[15] Differences in absorption can result in differences in effect for different populations. In this case, the model defined in the previous section ("Drug Interactions") needs to be expanded to include a vector Z of patient characteristics:

$$\hat{f}(x, y, z, t) = \frac{1}{na_n^3} \sum_{j=-\infty}^{\infty} W(\hat{s}_n i, j) \sum_{l \in C_n(j)} k\left(\frac{x - X_l}{a_n}, \frac{z - Z_l}{a_n}, \frac{t - T_l}{a_n}\right)$$

Another population generally excluded from clinical trials is the elderly. For this population, males average 13.26 different prescriptions; females average 16.25.^[16] Interaction studies typically examine two to three drugs in combination. It is necessary to develop tools that will investigate medications prescribed in such large numbers.

Cost-Effectiveness Analysis

Cost-effectiveness is a broad term that encompasses three major types of models:^[17]

- Cost-benefit studies. Both costs and outcomes are measured in dollars.
- Cost-effectiveness studies. Outcomes are quantified in terms of health status measures, such as the number of symptom-free days.
- Cost-utility studies. Effectiveness is measured in terms of utility, such as quality-adjusted life-years.

There are also different perspectives taken with regard to cost analysis. The broadest perspective is societal, where the interest is to define the greatest good for the greatest number of people (while possibly increasing risks for some individuals). An intermediate perspective belongs to an insurer or a health maintenance organization. The narrowest perspective belongs to an individual patient or an individual provider. These perspectives can sometimes be at odds.

The different perspectives will clash, for the broader viewpoints will sometimes be at the expense of individuals. Health maintenance organizations tend to deny any treatment without Level I (or possibly Level II) evidence. Oregon attempted to prioritize its Medicaid funding based on optimizing results from a societal perspective. The program failed because the societal perspective often differed considerably from the individual perspective.^[18]

In each case, an attempt is made to quantify health outcomes, usually assuming a linear relationship. In addition, most cost-effectiveness analysis depends upon rough estimates of probabilities, usually provided by the "Delphi interview method." This involves asking a group of experts for their perceptions. Real data are rarely used to develop the model. Instead, a sensitivity analysis is performed. The estimates given by the experts are altered to determine just how much deviation will still give the same outcomes (a measure of robustness). However, as in the case of peptic ulcers, if the experts were unaware of the actual cause of an illness, the sensitivity analysis would not provide an accurate response. Therefore, cost-effectiveness analysis will always support conventional belief.

The model for examining cost takes a form similar to the one described under "Drug Interactions" in this

section. There is a vector (E, C, Z) where E = effects (both benefits and risks), C = costs, and Z = patient characteristics. Then if a linear relationship is assumed, it is defined by the equation

$$e_{ij} = G_j(c_{ij}) + \varepsilon_{ij}$$

However, costs from a societal perspective are assumed to stay within 2 standard deviations of the mean. Unfortunately, health care has much to do with outliers, i.e., individuals at the extreme end of need, who use health care resources way beyond this preset norm. Therefore, it is preferable to use a nonlinear model with

$$\hat{f}(e, c, z) = \frac{1}{na_n^2} \sum_{j=-\infty}^{\infty} W(\hat{s}_n, i, j) \sum_{l \in D_n(l)} k\left(\frac{c - C_l}{a_n}, \frac{z - Z_l}{z_n}\right)$$

assuming that E takes discrete values.

EXAMPLES FROM THE MEDICAL LITERATURE

Surveillance of *Helicobacter pylori* Treatments

The problem of peptic ulcers is one of extreme interest. For years the standard treatment was to use an acid inhibitor such as omeprazole or ranitidine. It was assumed that the underlying cause of ulcers was stress. Therefore, it was also assumed that a maintenance dose of medication was required to prevent recurrence or that stress-reduction therapy would be needed to prevent additional problems. It was also assumed that bacteria could not survive in the stomach, so until recently no studies were conducted to investigate the possibility of infection.^[19] The discovery of *H. pylori* in the stomach completely changed the entire course of investigation. Therefore, it is possible to examine all aspects of postmarketing surveillance in the investigation of the treatment and outcome of peptic ulcer disease, particularly because all medications used to treat the problem were fairly common and already approved for use for other illnesses.

Drug Interactions

Once it was discovered that *H. pylori* could survive in the acidic environment of the stomach, studies were conducted to determine a treatment that could successfully eradicate it. All studies involved postmarketing surveillance, for all drugs tested were already approved for use. Decisions were made early on to use drugs in combination: various antibiotics and acid suppressors. Penston^[20] focused on the combination of omeprazole, amoxicillin, and metronidazole. Others^[21–24] examined other combinations. No study has shown that these combinations put patients at risk for serious adverse reactions. Table 4,^[2] summarizes the level of evidence for eradicating *H. pylori* for a variety of gastrointestinal complaints. Treatment is recommended for

Table 4 Level of Evidence for Eradicating *H. pylori*

Complaint	Level of evidence	Recommendation
Peptic ulcer	I	Treat
Bleeding duodenal ulcer	I	Treat
Nonulcer dyspepsia	II	Treat
Early gastric cancer	III	Treat
Gastric mucosa-associated lymphoid tissue	V	Treat
Gastroesophageal reflux disease	Below Level V	Do not treat

From Ref. [2].

levels III–V because it has low risk and low cost and there is little indication that another treatment is more effective (Table 4).

Another type of interaction to examine occurs when different problems occur simultaneously. Stress ulcers still occur, particularly in an intensive care unit. Rune^[25] explored the possibility that *H. pylori* contributed to the incidence of stress ulcers. Examining 874 patients (sufficient numbers for 90% power), the presence of *H. pylori* was found to be statistically significant.

Long-Term Follow-Up

Since acid suppressors were developed as the standard treatment for peptic ulcers, a study^[26] was conducted to determine the long-term treatment results for peptic ulcers. The results demonstrated clearly that the traditional measures of acid suppression and surgery were somewhat temporary; the majority of patients still had gastric problems. This study was not completed until years after the discovery of *H. pylori*. Yet none of the subjects followed was given a test for the presence of the bacteria; none were treated for the infection. It is very speculative that this occurred, because physicians are still generally reluctant to prescribe triple therapy^[27] or because of a reluctance to stop the study early. However, if this study had been initiated at an earlier time, there should not have been so much complacency concerning the effectiveness of acid suppressors in treatment. Patients with recurring problems should not have been so casually dismissed as having stress disorders. No matter how much confidence exists in a particular treatment, there will always be unknowns that can drastically change the perspectives.

Excluded Populations

Caraco et al.^[15] conducted a study to compare absorption rates of omeprazole between Chinese and Caucasian subjects. The study concluded that the Chinese had greater bioavailability of the drug. It is suggested that Chinese patients should routinely receive a lower dose of the drug

to compensate for the greater bioavailability. It is not suggested what that dose should be. It also remains unknown just how this absorption rate of omeprazole affects the triple therapy routinely used to eradicate *H. pylori*.

Another study^[28] was conducted to investigate the problem of ulcers in the elderly population. A sample of 121 patients ranging in age from 61 to 89 years was treated with omeprazole and with one or two antibiotics for 3 or 7 days. There was a total of six different combinations. This study has sufficient power (80%) to detect a 15% difference in outcome with a 5% level of significance. The study also resulted in 5% of the subjects with adverse reactions to the medications. The optimal treatment was omeprazole, clarithromycin, and metronidazole for 7 days.

Cost-Effectiveness Analysis

Since acid suppressors were developed to treat ulcers, a cost-effectiveness analysis was conducted to compare this treatment with the triple therapy used to eradicate *H. pylori*.^[29] This study obviously had to be done from a societal perspective, because the individual perspective would be a preference for complete eradication of the cause of the ulcers rather than to use lifelong maintenance therapy to relieve symptoms. The study did conclude that the societal perspective was identical to the individual perspective. Therefore, from all perspectives, eradication is cost-effective. Yet it is still little prescribed.^[27]

Greenberg et al.^[30] examined the need for routine endoscopy for patients with gastrointestinal problems. It is suggested that a trial of omeprazole should be made prior to routine endoscopy to reduce costs and to increase the effectiveness of endoscopy. For patients for whom omeprazole relieves symptoms but does not heal, a serology test for the presence of *H. pylori* should be done in place of endoscopy. O'Malley et al.^[31] investigated cost-effectiveness in a different way. The claim is made that many patients presenting with gastrointestinal complaints actually have psychiatric disorders. The claim is then made that screening for psychiatric disorders prior to upper endoscopy is cost-effective. However, of the patients diagnosed with a psychiatric disorder, 26% also had a medical problem. Of the patients with a disorder, 44% were diagnosed with somatoform that should not be diagnosed without first screening for a medical disorder. Out of the 116 patients in the study, none were tested for the presence of *H. pylori*.

Surveillance of Fluoxetine

There are a number of papers in the medical literature discussing postmarketing surveillance of this drug. Some^[32,33,34] discuss anecdotal case studies involving the use of fluoxetine. Without statistical evidence, it is not known whether the adverse effects reported in these editorials are actually caused by fluoxetine or are just

coincidental. Even if real, there is no way of knowing the actual risk, nor is there any way of knowing whether the risk is for a particular segment of the population. The most such letters can do is to make a practicing physician aware of the possibility of an adverse reaction.

Many of the postmarketing studies have examined segments of the population that were excluded from trials prior to FDA approval.^[35,36] Chambers et al.^[36] provide Level III evidence. Consider the method of recruiting subjects:

From 1989 through 1995, the California Teratogen Information Service and Clinical Research Program received approximately 1500 calls requesting information on the potential teratogenic effects of fluoxetine. An estimated one-third of these inquiries were made by pregnant women currently taking the drug. We selected 288 of these women for inclusion in the study on the basis of accessibility by telephone and willingness to participate. During this same period, pregnant women who called the program with questions about drugs and procedures not considered teratogenic—including acetaminophen use, dental radiography, and limited alcohol ingestion (< 1 oz [30 ml] of 100 percent alcohol per week before pregnancy was recognized)—were asked to enroll in the study as a control group. From this group, 254 women were selected as controls because their inquiries were closest in time to those of the women taking fluoxetine. The majority of women in each group were enrolled in the study during the first trimester of their pregnancies, and all were enrolled before any outcomes of the pregnancy were known, including knowledge of conditions that were diagnosed prenatally.

Each woman who enrolled in the study completed a questionnaire that included her history of previous pregnancies and family medical history, socioeconomic and demographic information for each woman and her partner, and exposures during the current pregnancy. The exposure history included dosages, dates, and indications for all medications; use of caffeine; use of supplemental vitamins; occupational exposures; infectious or chronic diseases; prenatal testing or other medical procedures; and use of recreational drugs, tobacco, and alcohol. Each woman was provided with a diary in which she was asked to keep a record of any additional exposures that might occur before delivery. We supplemented this record by calling the women throughout their pregnancies.

Only women who called for information were asked to participate. It was not determined if there were differences in the two groups of women, specifically in any socioeconomic variables; the control group was matched by length of gestation only. A power analysis is needed to determine if there is a sufficient sample size to detect differences in rare events. A severe adverse effect can be expected to occur in the population only 1–2% of the time. To detect even a 4% difference (from 10 per 1000 women to 14 per 1000 women) requires a total sample size of 1000 patients. There were only 482 women enrolled in the study. Moreover, fewer infants were evaluated for severe reactions, reducing

the sample size even more. This study was totally unable to determine if real differences in safety existed. In addition, a subgroup analysis was performed based upon trimester of exposure. Subgroup analysis should always be used to generate hypotheses, never to make inference.^[37] Yet this paper makes conclusions based upon this subgroup analysis.

CONCLUSION

There are a number of studies that must be conducted once a drug has been approved and marketed: long-term follow-up, additional diagnoses, previously excluded populations, and cost analysis. Most of these studies will involve the collection of observational data instead of randomized clinical trials. Statistical methods that can safely utilize such observational data must be used to generate a hypothesis. These methods include data mining and neural network analysis. Once the hypothesis has been generated, it must be validated by additional data.

Appendix

MATHEMATICA PROGRAM FOR UNIVARIATE KERNEL DENSITY ESTIMATION

```
s[x_]:= (0.75*(1 - 0.2*x^2)/Sqrt[5])/x^2 < 5
s[x_]:= 0/x^2 >= 5
r[x_,y_,a_]:= s[(x - y)/a]
f[x_,a_]:= Sum[r[x,arr[[i]],a], {i, 1, Length[arr]}]
(a*Length[arr])
u[x_,a_]:= f[x - Mean[tab]]/StandardDeviation[tab],a]/
standardDeviation[tab]
```

MATHEMATICA PROGRAM FOR BIVARIATE KERNEL DENSITY ESTIMATION

```
t[x_,y_]:= (3*(1 - x^2 - y^2)/Pi)/x^2 + y^2 < 1
t[x_,y_]:= 0/x^2 + y^2 >= 1
h[x_,y_,u_,v_,a_]:= t[(x - u)/a, (y - v)/a]
g[x_,y_,a_]:= Sum[h[x,y,arr1[[i]], arr2[[i]], a], {{i,1,
Length[arr1}}]/(a^2(Length[arr1])
b[x_,y_,a_]:= N[g[(x - Mean[tab 1])/Standard-Deviation
[tab1],(y - Mean[tab2])/Standard-Deviation[tab2],a]]/
(StandardDeviation[tab1]*StandardDeviation[tab2])
```

MATHEMATICA PROGRAM FOR DISCRETE-CONTINUOUS ESTIMATION

```
w[s_,i_,j_]:= .5(1 - s)^Abs[i - j]/Abs[i - j] >= .5
w[s_,i_,j_]:= 1 - s/Abs[i - j] < .5
n:= Sum[arr[[j]],{j,1,Length[arr]}]
p[s_,i_]:= Sum[w[s,i,j]*arr[[j]],{j,1,Length[arr]}/n
```

REFERENCES

1. Johnson, J.A. Influence of race or ethnicity on pharmacokinetics of drugs. *J. Pharm. Sci.* **1997**, *86*, 1328–1333.
2. Howden, C.W. For what conditions is there evidence-based justification for treatment of *Helicobacter pylori* infection? *Gastroenterologist* **1997**, *113*, S107–S112.
3. Matheus, C.J.; Piatetsky-Shapiro, G.; McNeill, D. Selecting and Reporting What is Interesting. In *Advances in Knowledge Discovery and Data Mining*; Fayyad, W.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Eds.; MIT Press: Cambridge, MA, 1996.
4. Weiss, S.I.; Kulikowski, C. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural networks, Machine Learning, and Expert Systems*; Morgan Kaufman: San Francisco, CA, 1991.
5. Elder, J.F.; Pregibon, D. A Statistical Perspective on Knowledge Discovery in Databases. In *Advances in Knowledge Discovery and Data Mining*; Fayyad, W.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Eds.; MIT Press: Cambridge, MA, 1996.
6. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. The KDD process for extracting useful knowledge from volumes of data. *Commun. ACM* **1996**, *39*, 27–34.
7. Lane, D.; Rawson, N. Inferential Problems in Postmarketing Surveillance. In *Statistical Methodology in the Pharmaceutical Sciences*; Owen, D.B., Ed.; Marcel Dekker: New York, 1990.
8. Cerrito, P.B.; Cerrito, J.C. Assessing risk in postmarketing surveillance. *J. Biopharm. Stat.* **1991**, *1*, 221–235.
9. Tatro, D.S. *Drug Interaction Facts*, 3rd Ed.; Facts and Comparisons: St. Louis, MO, 1992.
10. Ahmad, I.A.; Cerrito, P.B. Nonparametric estimation of joint discrete-continuous probability densities with econometric applications. *J. Stat. Plan.* **1993**, *41*, 349–363.
11. Wang, M.; van Ryzin, J. A class of smooth estimators for discrete distributions. *Biometrika* **1981**, *68*, 301–309.
12. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Chapman and Hall: London, 1986.
13. Knuiman, M.W.; Vu, H.T.; Segal, M.R. An empirical comparison of multivariable methods for estimating risk of death from coronary heart disease. *J. Cardiovasc. Risk* **1997**, *4*, 127–134.
14. Kasabov, N.K. *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering*; MIT Press: Cambridge, MA, 1996.
15. Caraco, Y.; Lagerstrom, P.O.; Wood, A.J.J. Ethnic and genetic determinants of omeprazole disposition and effect. *Clin. Pharmacol. Ther.* **1996**, *60*, 157–167.
16. Moeller, J.; Mathiowetz, N. Prescribed Medicines: A Summary of Use and Expenditures by Medicare Beneficiaries. In *National Medical Expenditure Survey Findings 3*; National Center for Health Services Research: Rockville, MD, 1989. DHHS Publication No. (PHS) 89-3448.
17. Laska, E.M.; Meisner, M.; Siegel, C. Statistical inference for cost-effectiveness ratios. *Health Econ.* **1997**, *6*, 229–242.
18. Ubel, P.A.; Loewenstein, G.; Scanlon, D.; Kamlet, M. Individual utilities are inconsistent with rationing choices: A partial explanation of why Oregon's cost-effectiveness list failed. *Med. Decis. Mak.* **1996**, *16*, 108–116.

19. Calam, J. The somatostatin–gastrin link of *Helicobacter pylori* infection. *Ann. Med.* **1995**, *27*, 569–573.
20. Penston, J.G. Review article: Clinical aspects of *Helicobacter pylori* eradication therapy in peptic ulcer disease. *Aliment. Pharmacol. Ther.* **1996**, *10*, 469–486.
21. Tytgat, G.N. No *Helicobacter pylori*, no *Helicobacter pylori*-associated peptic ulcer disease. *Aliment. Pharmacol. Ther.* **1995**, *9* (Suppl 1), 39–42.
22. Fendrick, A.M. Outcomes research in *Helicobacter pylori* infection. *Aliment. Pharmacol. Ther.* **1997**, *11* (Suppl 1), 95–101.
23. Lim, A.G.; Walker, C.; Chambers, S.; Gould, S.R. *Helicobacter pylori* eradication using a 7-day regimen of low-dose clarithromycin, lansoprazole and amoxycillin. *Aliment. Pharmacol. Ther.* **1997**, *11*, 537–540.
24. Munnangi, S.; Sonnenberg, A. Time trends of physician visits and treatment patterns of peptic ulcer disease in the United States. *Arch. Intern. Med.* **1997**, *157*, 1489–1494.
25. Rune, S.J. Treatment strategies for symptom resolution, healing, and *Helicobacter pylori* eradication. *Scand. J. Gastroenterol.*, Suppl. **1994**, *205*, 45–47.
26. Riester, K.A.; Peduzzi, P., et al. Statistical evaluation of the role of *Helicobacter pylori* in stress gastritis: Applications of splines and bootstrapping to the logistic model. *J. Clin. Epidemiol.* **1997**, *50*, 1273–1279.
27. Malliwah, J.A.; Tabaqchali, M.; Watson, J.; Venables, C.W. Audit of the outcome of peptic ulcer disease diagnosed 10 to 20 years previously. *Gut* **1996**, *38*, 812–815.
28. Pilotto, A., et al. Cure of *Helicobacter pylori* infection in the elderly: Effects of eradication on gastritis and serological markers. *Aliment. Pharmacol. Ther.* **1996**, *10*, 1021–1027.
29. Jonsson, B.; Karlsson, G. Economic evaluation in gastrointestinal disease. *Scand. J. Gastroenterol.* **1996**, *220*, 44–51.
30. Greenberg, P.D.; Koch, J.; Cello, J.P. Clinical utility and cost effectiveness of *Helicobacter pylori* testing for patients with duodenal and gastric ulcers. *Am. J. Gastroenterol.* **1996**, *91*, 228–232.
31. O'Malley, P.G. The value of screening for psychiatric disorders prior to upper endoscopy. *J. Psychosom. Res.* **1998**, *44*, 279–287.
32. Benazzi, F. Dangerous interaction with nefazodone added to fluoxetine, desipramine, venlafaxine, valproate and clonazepam combination therapy. *J. Psychopharmacol.* **1997**, *11*, 1–190,191.
33. Benazzi, F. Severe anticholinergic side effects with venlafaxine–fluoxetine combination. *Can. J. Psychiatry* **1997**, *42*, 1–980,981.
34. Tiller, J.W.; Johnson, G.F.; Burrows, G.D. Moclobemide for depression: An Australian psychiatric practice study. *J. Clin. Psychopharmacol.* **1995**, *15* (4 Suppl 2), 31S–34S.
35. Goldstein, D.J.; Corbin, L.A.; Sundell, K.L. Effects of first-trimester fluoxetine exposure on the newborn. *Obstet. Gynecol.* **1997**, *89* (5 pt I), 713–718.
36. Chambers, C.D.; Birth outcomes in pregnant women taking fluoxetine. *N. Engl. J. Med.* **1996**, *335*, 1010–1015.
37. Bailer, J.C., 3rd. Preliminary reports: when is the evidence strong enough to change practice? *J. Clin. Epidemiol.* **1994**, *47* (3), 305–306.

Power

Kenneth B. Schechtman

Washington University School of Medicine, St. Louis, Missouri, U.S.A.

Abstract

The “power” of a statistical test is the probability of rejecting the null hypothesis when it is false. This brief entry summarizes statistical methods of analyzing this probability.

INTRODUCTION

The “power” of a statistical test is the probability of rejecting the null hypothesis when it is false. Closely related to statistical power is the false-negative rate, which is often referred to as the Type II error or beta error. The false-negative rate equals 1 minus the power. It is defined as the probability of not claiming statistical significance when the null hypothesis is false. A fundamental goal of every research protocol should be to define a study that has adequate statistical power and, thereby, a sufficiently small Type II error rate. Whereas there are no universally accepted guidelines, it is broadly accepted that a prospective study should not be initiated and is unlikely to be funded unless the statistical power that is associated with primary hypothesis tests is projected to be in the 0.8–0.9 range.

CORRELATES OF STATISTICAL POWER

Statistical power depends in obvious ways on several characteristics of the statistical test being performed. In addition to the sample size association, which is discussed next, the power of a statistical test depends also on the size of the between-group difference and on the amount of noise in the data. For example, the greater the difference between groups that are to be compared, the easier it is to establish statistical significance. Thus, statistical power is directly related to the magnitude of the failure of the null hypothesis. Similarly, the greater the variability in a quantitative measure, the greater is the noise in the data and the harder it is to establish significance. That is, an increase in the variance of a parameter implies a decrease in the statistical power of the associated hypothesis test.

STATISTICAL POWER AND SAMPLE SIZE

The most important correlate of statistical power is the sample size. An evaluation of the close relationship

between statistical power and sample size should be a key component of the design of every research protocol and of the interpretation of the resulting publications. Unfortunately, repeated reviews of the literature suggest that low power and inadequate sample size are extremely common, if not routine, in the medical literature.^[1–4] For example, Freiman et al.^[1] reviewed 71 randomized controlled trials whose primary conclusion was negative. They found that 67 of the 71 trials (94.4%) had greater than a 10% probability of missing a true 25% therapeutic improvement. That is, the power to detect a 25% treatment benefit was nearly always less than 0.9. The power to detect a very large treatment benefit of 50% was less than 0.9 in 70.4% of the reviewed papers. Although these authors do not present the precise results, it is clear from their tabulations that the power of many if not most of the reviewed hypothesis tests was such that clinically important differences in primary endpoints were more likely to be missed than to be reported as statistically significant. This suggestion that statistical power in the published literature is often extremely low is highlighted by Schechtman et al.,^[4] whose review of 37 papers on the surgical treatment of sleep apnea found that the median sample size would yield a power of under 0.2 to test key hypotheses of interest. When the power is this low, the claim that “there is no significant differences between groups” translates roughly into the more accurate assertion that “we have no idea if the two treatments are different in clinically important ways because the sample size was too small to address the question.”

In addition to confirming that many clinical trials are too small to detect large and clinically important differences, Moher et al.^[2] have noted that even in the most prestigious and carefully reviewed journals, there is a strong tendency to omit all discussion of sample size and statistical power. These authors reviewed the 383 randomized trials published in the *New England Journal of Medicine*, the *Journal of the American Medical Association*, and the *Lancet* in the years 1975, 1980, 1985, and 1990. They found that only 32% of the trials with negative results discussed statistical power and sample size calculations. This means

that unless the reader is energetic enough to do these computations independently, it may be difficult to determine whether negative conclusions mean that there are unlikely to be clinically important differences between groups or, alternatively, that the study was too small to address the question. This uncertainty is compounded by the fact that only 20 of the 102 negative studies that were reviewed in the Moher et al. paper made any statement as to the clinical significance of observed differences. The key bright spot in this report is that while the 0% rate at which sample size considerations were discussed in the reviewed 1975 publications mirrors the poor rates reported elsewhere at about that time,^[5,6] the more recent rate is a substantially improved 43% in 1990.

DETERMINING WHETHER A STUDY IS INADEQUATELY POWERED

Because they are discussed elsewhere in this volume, we do not focus on the many standard formulae that are commonly used to compute power and sample size. However, we demonstrate how an astute reader can quickly decide whether a study is too small using a standard unpaired *t* test as an illustration. Considerations similar to those described next can be applied in other settings.

The required per-group sample size *N* for an unpaired *t* test with the same number of patients in each group is computed using the equation

$$N = \frac{2(Z_{\alpha} + Z_{\beta})^2}{\Delta^2}$$

where Z_{α} is the standard normal deviate for a test with significance level α (e.g., $Z_{\alpha} = 1.96$ for a two-sided test at the 0.05 level of significance), Z_{β} is the normal deviate associated with a Type II error rate of β or a power of $1 - \beta$ (e.g., $Z_{\beta} = 1.28$ for a power of 0.9), and $\Delta = (\mu_2 - \mu_1)/\sigma$, where μ_1 and μ_2 are the means in the two groups and σ is the common standard deviation. The foregoing expression can be adjusted to compute the sample size in one of two groups when patient allocation is unequal by multiplying by $(r + 1)/r$, where r is the ratio of the total sample size to the sample size in the group of interest.

It follows from the foregoing equation that for a given significance level and a given statistical power, the required sample size depends totally on Δ , the ratio of the difference in means divided by the common standard deviation. Using the equation or the sample size tables that are presented in many statistics texts, it is easy to determine whether the sample size in a given study is reasonable. Thus, assuming a two-sided test at the 0.05 level of significance, a Δ of 1 means that 23 patients per group will yield a power of 0.9. If you are reading a paper where $\Delta = 1$ and the sample

size is 50 per group, it is overkill insofar as the relevant hypothesis test is concerned. If there are 10 patients in each group, the study is far too small unless Δ is in the 1.5 range. For smaller Δ s in the 0.5 range, 70 or so subjects must be studied in each group to yield adequate power. If a study with a Δ of 0.5 enrolls only 20 patients per group, a common state of affairs in much of the literature, you know immediately that it is underpowered and that nonsignificant results have very little meaning. An astute reader should apply this type of reasoning to every manuscript. A well-written paper should discuss power and sample size so the reader does not have to fill in the blanks.

CONCLUSION

The medical literature is filled with examples of studies that did not enroll a sufficient number of subjects to provide an adequate test of key hypotheses. When sample sizes are inadequate, the power to detect between group differences is limited, and it may be impossible to determine if nonsignificant results really do mean that between-group differences are clinically unimportant. It is incumbent on both authors and readers to critically consider the impact of sample size and power on the interpretation of results.

REFERENCES

1. Freiman, J.A.; Chalmers, T.C.; Smith, H., Jr.; Kuebler, R.R. The importance of bets, the type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 "negative" trials. *N. Engl. J. Med.* **1978**, *299*, 690–694.
2. Moher, D.; Dulberg, C.S.; Wells, G.A. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* **1994**, *272*, 122–124.
3. Williams, H.C.; Seed, P. Inadequate size of "negative" trials in dermatology. *Br. J. Dermatol.* **1993**, *128*, 317–326.
4. Schechtman, K.B.; Sher, A.E.; Piccirillo, J.F. Methodological and statistical problems in sleep apnea research: The literature on uvulopalatopharyngoplasty. *Sleep* **1995**, *18*, 659–666.
5. Chalmers, T.C.; Silverman, B.; Shareck, E.P.; Ambroz, A.; Schroeder, B.; Smith, H., Jr. Randomized Control Trials in Gastroenterology with Particular Attention to Duodenal Ulcer. In *Report to the Congress of the United States of the National Commission on Digestive Diseases*; Reports of the Workgroups: Part 2B: Subcommittee on Research-Targeted and Nondirected, National Institute of Arthritis, Metabolism, and Digestive Diseases: Bethesda, MD, 1978; Vol. IV, 223–255. Publ no. NIH 79-1885.
6. Mosteller, F.; Gilbert, J.P.; McPeck, B. Reporting standards and research strategies for controlled trials: Agenda for the editor. *Control. Clin. Trials* **1980**, *1*, 37–58.

Precision Medicine: Statistical Issues

Fuyu Song

Center for Food and Drug Inspection, China Food and Drug Administration, Beijing, China

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

In clinical trials, a typical approach for the evaluation of safety and efficacy of a test treatment under investigation is to first test for the null hypothesis of no treatment difference in efficacy based on clinical data collected under a valid trial design. If significant, the investigator would reject the null hypothesis of no treatment difference and then conclude the alternative hypothesis that there is a difference in favor of the test treatment. If there is a sufficient power for correctly detecting a clinically meaningful difference (improvement) when such a difference truly exists, we claim that the test treatment is efficacious. The test treatment will then be reviewed and approved by the regulatory agency such as the United States Food and Drug Administration (FDA) if the test treatment is well tolerated and there appears to be no safety concerns. We will refer to medicine developed based on this typical approach as *traditional medicine*.

Keywords: Individualized medicine; Personalized medicine; Positive predictive value; Traditional Chinese medicine; Traditional medicine.

INTRODUCTION

In clinical trials, a typical approach for the clinical investigation of safety and efficacy of a test treatment is to first test for the null hypothesis that there is no clinically meaningful difference between treatments based on clinical data collected under a valid study design. If significant, the investigator would reject the null hypothesis of no treatment difference and then conclude the alternative hypothesis that there is a difference in favor of the test treatment. If there is a sufficient power for correctly detecting a clinically meaningful difference (improvement) when such a difference truly exists, we claim that the test treatment is efficacious. The test treatment will then be reviewed and approved by the regulatory agency such as the United States Food and Drug Administration (FDA) if the test treatment is well tolerated and there appears to be no safety concerns. We will refer to medicine developed based on this typical approach as *traditional medicine*.

In his State of the Union address in early 2015, President Barack Obama announced that he was launching the *Precision Medicine Initiative*—a bold new research effort to revolutionize how we improve health and treat disease. As indicated by President Obama, precision medicine is an innovative approach that takes into account individual differences in people's genes, environments, and lifestyles. Unlike traditional approach (traditional medicine), most medical treatments have been designed for the average patient. For this *one-size-fits-all* approach, treatments

can be very successful for some patients but not for others. Precision medicine, on the other hand, gives medical professionals the resources they need to target the specific treatments of the illnesses we encounter.^[1] In response to President Obama's Precision Medicine Initiative, the National Institutes of Health kicked off cohort grants for precision medicine to develop treatments tailored to an individual based on their genetics and other personal characteristics subsequently.^[2] Seeking for precision medicine of cure has become the center of clinical research in pharmaceutical development since then.

PRECISION MEDICINE

Unlike traditional medicine, precision medicine is referred to as a medical model that proposes the customization of health care, with medical decisions, practices, and/or products being tailored to the individual patient.^[3] In this model, diagnostic testing is often employed for selecting appropriate and optimal therapies based on the context of a patient's genetic content or other molecular or cellular analysis. Tools employed in precision medicine could include molecular diagnostics, imaging, and analytics/software. This has led to biomarker development in genomics studies for target clinical trial. A validated biomarker (diagnostic tool) is then used to identify patients who are most likely to respond to the test treatment under investigation in the enrichment process of the target clinical trials.^[4–7]

As a result, precision medicine will benefit the subgroup of patients who are biomarker positive. In practice, however, there may exist no perfect diagnostic tool for determining whether a given patient is with or without molecular target in the enrichment process of the target clinical trials. Possible misclassification, which could cause a significant bias in the assessment of treatment effect in target clinical trials, is probably the most challenging issue in precision medicine (see, e.g., the work of Chow and Song^[8]).

As an example, let Y_{ij} be the responses of the j th subject in the i th group, where $j = 1, \dots, n_i$; $i = T, C$. Y_{ij} are assumed approximately normality distributed with homogeneous variances between the test and control treatments. Also, let μ_{T+} , μ_{C+} , and (μ_{T-}, μ_{C-}) be the means of test and control groups for the patients with (without) the molecular target. Table 1 summarizes the population means by treatment and diagnosis.

In target clinical trials, it is of interest to estimate the treatment effect for the patients truly having the molecular target, i.e., $\theta = \mu_{T+} - \mu_{C+}$. However, this effect may be contaminated due to misclassification, i.e., for those subjects who do not have the molecular target but got positive diagnosed results and those subjects who have the molecular target but got negative diagnosed results. The following hypothesis for detecting a clinically meaningful treatment difference in the patient population truly with the molecular target is of interest:

$$H_0 : \mu_{T+} - \mu_{C+} = 0 \quad \text{versus} \quad H_a : \mu_{T+} - \mu_{C+} \neq 0 \quad (1)$$

Let $\bar{y}_T$ and $\bar{y}_C$ be the sample means of test and control treatments, respectively. Since no diagnostic test is perfect for the diagnosis of the molecular target of interest without error, therefore, some patients with a positive diagnostic result may in fact not have the molecular target. It follows that

$$E(\bar{y}_T - \bar{y}_C) = \gamma(\mu_{T+} - \mu_{C+}) + (1 - \gamma)(\mu_{T-} - \mu_{C-}) \quad (2)$$

where γ is the positive predicted value, which is often unknown. Thus, an accurate and reliable estimate of γ is the key to success of target clinical trials^[7] and hence precision medicine.

PRECISION MEDICINE VERSUS PERSONALIZED MEDICINE

The term *precision medicine* is often mixed up with the term *personalized medicine*. To distinguish the difference

between precision medicine and personalized (or individualized) medicine, the National Research Council indicates that precision medicine refers to the tailoring of medical treatment to the individual characteristics of each patient. It does not literally mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease, in the biology and/or prognosis of those diseases they may develop, or in their response to a specific treatment. In summary, precision medicine is to benefit the subgroup of patients with the diseases under study, whereas personalized medicine is to benefit individual subjects with the diseases under investigation.

Statistically, the term *precision* is usually referred to the degree of closeness of the observed data to the truth. High degree of closeness is an indication of high precision. Thus, precision is related to the variability associated with the observed data. In practice, the variability associated with observed data include (1) intra-subject variability, (2) intersubject variability, and (3) variability due to subject-by-treatment interaction. As a result, precision medicine can be viewed as the identification of subgroup population with larger effect size (i.e., smaller variability), assuming that the difference in mean responses is fixed. Consequently, precision medicine focuses on minimizing intersubject variability, whereas personalized medicine focuses on minimizing intra-subject variability. Table 2 provides a comparison between precision medicine and personalized medicine.

FUTURE PERSPECTIVES

As indicated earlier, traditional medicine can only benefit average patients with the diseases under study, whereas precision medicine can further benefit a subgroup of patients who have certain characteristics (e.g., molecular target). On the other hand, the ultimate goal of personalized (or individualized) medicine is seeking cure for individual patients if it is not impossible. President Obama's Precision Medicine Initiative is an important step moving away from the traditional medicine and toward precision medicine, then personalized (or individualized) medicine so that individual patients could be beneficial. For this purpose, traditional Chinese medicine, which often consists of multiple components and focuses on global dynamic harmony (or balance) within individual patients, is expected to be the center of personalized medicine development moving

Table 1 Population means by treatment and diagnosis

Positive diagnosis	True target condition	Indicator of diagnostic	Test group	Control group	Difference
+	+	γ	μ_{T+}	μ_{C+}	$\mu_{T+} - \mu_{C+}$
	−	$1 - \gamma$	μ_{T-}	μ_{C-}	$\mu_{T-} - \mu_{C-}$

Note: γ is the positive predicted value

Table 2 Precision medicine versus personalized medicine

Characteristic	Traditional medicine	Precision medicine	Personalized medicine ^a
Active ingredient	Single	Single	Multiple
Target population	Population	Population	Individuals
Primary focus	Mean	Intersubject variability	Intra-subject variability
Dose/regimen	Fixed	Fixed	Flexible
Beneficial	Average patient	Subgroup of patients	Individual patients
Statistical method	Hypotheses testing Confidence interval	Hypotheses testing Confidence interval	Hypotheses testing Confidence interval
Use of biomarker	No	Yes	Yes
Blinding	Yes	Yes	May be difficult
Objective	Accuracy	Accuracy Precision	Accuracy Precision Reproducibility
Study design	Parallel/crossover	Parallel/crossover Adaptive design	Parallel/crossover Adaptive design
Probability of success	Low	Mild to moderate	High

^aPersonalized medicine = individualized medicine

toward the next century.^[9] For achieving this ultimate goal, regulatory requirement and quantitative/statistical methods for the assessment of treatment effect for drug products with multiple components are necessarily developed.

REFERENCES

1. News Release. *FACT SHEET: President Obama's Precision Medicine Initiative*; The White House, Office of the Press Secretary, January 30, 2015.
2. McCarthy, J. *NIH Kicks Off Cohort Grants for Precision Medicine*; www.govhealthit.com/news, November 23, 2015.
3. NRC. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*; Committee on a Framework for Developing a New Taxonomy of Disease, National Research Council; The National Academies Press, Washington, DC, 2011.
4. FDA. *The Draft Concept Paper on Drug-Diagnostic Co-Development*; The US Food and Drug Administration: Rockville, MD, 2005.
5. FDA. *Guidance on Pharmacogenetic Tests and Genetic Tests for Heritable Marker*; The US Food and Drug Administration: Rockville, MD, 2007.
6. FDA. *Draft Guidance on In Vitro Diagnostic Multivariate Index Assays*; The US Food and Drug Administration: Rockville, MD, 2007.
7. Liu, J.P.; Lin, J.R.; Chow, S.C. Inference on treatment effects for targeted clinical trials under enrichment design. *Pharm. Stat.* **2009**, *8*, 356–370.
8. Chow, S.C.; Song, F.Y. Some thoughts on precision medicine. *J. Biom. Biostat.* **2016**, *6*, 5. doi: 10.4172/2155-6180.1000269.
9. Chow, S.C. *Quantitative Methods for Traditional Chinese Medicine Development*; Chapman and Hall/CRC Press, Taylor & Francis, New York, NY, 2015.

Prediction Trees

Antonio Ciampi

McGill University, Montreal, Québec, Canada

Abstract

Prediction trees are powerful tools for extracting predictive information from data. This entry reviews different types of prediction trees while detailing the components of each type.

INTRODUCTION

Prediction trees are powerful tools for extracting predictive information from data. Their use is increasingly frequent in biomedical research, as in many other fields of application. The conceptual developments leading to the algorithms now in use concern many disciplines, statistics and machine learning among others. This article is written from a statistical perspective. In Statistics, one deals with variables measured on observational units. To formulate the prediction problem, one needs to distinguish between variables called predictors, and a (possibly multidimensional) variable of special interest called outcome. It is usually assumed that the expected value of the outcome μ is a function of the predictors' vector z : $\mu = \mu(z)$. A prediction rule associates to each possible value of the predictors, a characteristic of the probability distribution of the outcome, typically μ . The problem is to construct a prediction rule $\hat{\mu}(z)$ from a data matrix D , obtained from a series of observational units for which both predictors and outcome are known. The purpose of the construction is to be able to associate to new observational units, for which only the predictors are given, some reasonable guess about the outcome.

The analyst may decide to consider a particular functional form for $\mu(z)$, known except for a parameter of fixed dimensionality. The most popular examples of predictors defined this way are linear predictors: $\mu(z) = \beta z$ and generalized linear predictors:

$$h[\mu(z)] = \beta z$$

where h , the link function, is specified and β is a parameter. Then, the construction of the prediction rule from D reduces to the estimation of the parameter β . Alternatively, the analyst may choose to obtain a prediction rule through an algorithmic approach. In this case the function $\mu(z)$ is not specified through a parameter of fixed dimensionality, but through a series of algorithmic steps. The choice

between parametric and algorithmic approach is influenced by the current knowledge of the domain within which the prediction rule is needed. Ideally, a parametric approach should be based on a substantive theory justifying the chosen functional form, though, in actual applications, this is rarely the case. The algorithmic option is more appropriate in unstructured situations with limited theoretical knowledge. In these cases it would be highly desirable to propose algorithms that mimic cognitive strategies used by domain experts.

The construction of tree-structured predictors from data is algorithmic. The success of prediction trees may be partially explained by the fact that experts of many domains use some kind of recursive partitioning approach while making predictions. For instance, when making a prognosis, a physician may ask: "On the basis of my previous experience and knowledge, the most relevant prognostic characteristic is C_1 . Does this patient have C_1 ?" If the answer is yes, the physician may then make a prognosis. Or, if the information is not sufficient, the physician may evoke another prognostic characteristic, C_2 , say, as the most relevant prognostic information for patients having C_1 , and ask whether or not the patient has C_2 . Similarly, for patients not having C_1 , there may be another prognostic characteristic C_3 and a similar question: "Does this patient have C_3 ?" and so on. Clearly, this is an oversimplified model of the prognostic process, but it is a useful one for developing reasonable algorithms. Physicians are familiar with a graphic representation of this process, known as diagnostic key, or in the context of general medical decision, decision tree. It is tempting to say that the physician is actually applying a "virtual" prediction tree, which represents experience and knowledge. From the statistician's point of view, "experience and knowledge" have to be extracted from the data D . The challenge is to develop an algorithm that recursively searches for "most relevant predictors" and that also decides when to stop searching and actually predict. Thus, an algorithm for prediction tree construction must be based on a measure of "relevance" and a stopping rule.

BASIC NOTATIONS AND CONCEPTS

Fig. 1 represents a hypothetical prediction tree for predicting whether a cancer patient in remission will relapse within a year. It formalizes the process outlined above. The tree suggests asking a hierarchical set of questions about patients who belong to the population of interest. The presence of characteristic C_1 is ascertained for all patients. For patients with characteristic C_1 , one should ascertain the presence of C_3 , while for patients without C_1 , one should ascertain the presence of C_2 . The diagram implies that, on the basis of what is known at this stage (all measured variables), the prognostic process should be completed here. The labels at the nodes ($C_i?$) are called (binary) questions. By convention, circles represent points at which questions are asked. The top circle is termed root node or simply “root”; the other circles are termed internal nodes or, simply, nodes. Square boxes represent points at which predictions are made; they are termed terminal nodes or leaves. Circles and squares also represent subpopulations: For instance, the root node represents the general population and the left-most “leaf” represents the subpopulations of individuals with both C_1 and C_2 . Moreover, for a given data matrix D (a sample), nodes and leaves also represent sets of rows of D (subsamples). Graph theoretical language is also currently employed, but it hardly needs formal definitions. For instance, nodes issuing from the same “parent” node are known as its “children” and a subtree issuing from a node is a branch. Strictly speaking, the graph of Fig. 1 is a rooted binary “tree,” although tree is generally accepted.

Under the leaves, the probabilities of relapse within a year are reported. An alternative approach would be to indicate relapse or no relapse according to some decision rule, such as “if $p \geq 0.5$ then relapse,” “otherwise, no relapse.” Our preference is to express the leaf prediction as a probability, which is, indeed, the expected value of the underlying binary variable. In our case one can clearly see that the tree defines four prognostic groups: low, medium, medium-high, and high risk of relapse.

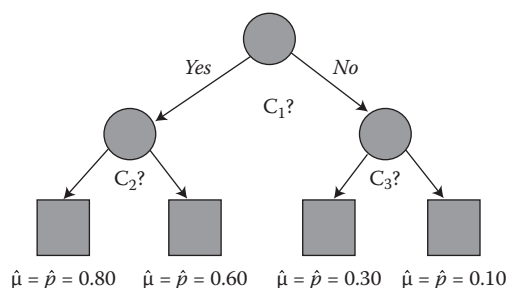


Fig. 1 Prediction tree for a binary outcome

PREDICTION TREES FOR A CLASS VARIABLE AND FOR A CONTINUOUS VARIABLE: CLASSIFICATION AND REGRESSION TREES

Tree-growing was introduced in the 1960s.^[1] The earliest algorithm was termed Automatic Interaction Detection. The goal was to predict a continuous univariate outcome, but extension to multivariate continuous outcome and categorical outcome was developed in the 1970s and early 1980s. The analog of the diagram of Fig. 1 would have, at each leaf, the conditional expectation of the outcome given the path joining the root to the leaf. The idea of the algorithm is extremely simple. At each node, the algorithm goes through all the predictors. For each value of a continuous predictor the data are split into subsets of cases above and below that value. For categorical predictors, the algorithm considers all binary groupings of its levels. The “goodness” of the split is assessed by a test statistic that compares the two subsets (t statistics, Hotelling’s T^2 , or chi-squared according to the type of outcome) and the “best split” is selected in consequence. The procedure is applied recursively to the subsets obtained by binary partitioning. The decision to stop branching is reached by assessing the “statistical significance” of the best split in the traditional sense. When no split is significant at some predetermined level the algorithm stops and the tree is obtained.

Until the mid-1980s, AID-like methods were used primarily in the social sciences and did not attract much attention within the statistical community. A change occurred, however, with the publication of the monograph *Classification and Regression Trees (CART)*.^[2] This seminal work formulated tree-growing within the framework of statistical learning theory, introducing ideas such as cross-validation and more general resampling schemes as reasonable remedies for the statistical problems implicit in intensive model searches. It developed a broad approach, known by its acronym CART, permitting the treatment of both categorical outcome (classification trees), and univariate continuous outcome (regression trees).

Recursive Partitioning in CART

The basic idea of the CART methodology is the separation of recursive partitioning from the choice of an optimal tree-structured predictor. We describe first the recursive partitioning phase. For a specified family of questions and a goodness of split measure, CART recursively partitions the root and children nodes by choosing at each node the question producing the best split. However, the stopping rule is not based on statistical significance, but rather on convenience considerations: In particular,^[2] propose to stop the recursive partitioning when the subsample of D at a node is too small, in a sense defined by the user, and/or, for classification trees, “pure,” i.e., consisting entirely of

elements of one class. Since statistical significance is not an issue, the choice of the goodness of split measure is now much broader. Moreover, the actual implementation of recursive partition may be governed entirely by computational efficiency requirements: For example, CART trees can be grown and stored by the efficient “depth first” strategy (add an extra split always at the leftmost node until no further split is possible). In contrast, earlier algorithms, and, indeed, some more recent ones, are “greedy” or best first algorithms: The current tree is updated by adding a split there where the best split is obtained.

For classification trees, one reasonable measure of goodness of split is based on the notion of node impurity, a quantity that assesses departure from a pure node. The impurity at a node t , $i(p|t)$, is indeed a function of the probability distribution of the outcome variable, $p = \{p(j|t); j = 1, 2, \dots, J\}$ which measures the “spread” of the distribution of the outcome: The more “peaked” the distribution, the smaller the impurity, with impurity zero associated to the case where one class is observed with probability 1 and the others are observed with probability 0 (pure node). The class probability distribution at node t , $p = \{p(j|t); j = 1, 2, \dots, J\}$, is generally unknown and has to be estimated from the data. Strictly speaking, we work with “estimates” of impurity $\hat{i} = i(\hat{p} | t)$. Several choices of impurity measures have been used, the most popular ones being Shannon’s entropy, and the Gini index. The definition is naturally extended to a split: The impurity of a split is the weighted sum of the impurities of the children nodes with weights given by the relative frequency of the children’s population sizes. Then, the impurity reduction due to the split (see CART for exact definitions) appears as a reasonable candidate for measuring goodness of split. One interesting property of impurity reduction is that it can be easily generalized to any partition, hence to the partition generated by a tree. In CART, the Gini index is recommended as an impurity measure; however, the authors also propose another measure of goodness of split, twoing, which does not have the properties of an impurity measure.

For regression trees, there is a natural measure of impurity at a node: the mean squared deviation from the mean of the continuous outcome, estimated on the subsample represented by the node. Impurity reduction due to a split (partition and tree) is defined accordingly. With this definition, regression tree-growing is seen as an adaptive least square regression algorithm—adaptive in the sense that binary variables are constructed from the original predictors as the algorithm unfolds. An alternative, equally natural definition of impurity at a node is the mean absolute deviation (MAD); with this choice tree-growing is an adaptive MAD regression algorithm.

Pruning in CART

The tree obtained in the recursive partitioning phase is termed large tree. It may be expected that the large tree

overfit the data. In other words, the predictive relationships between predictors and outcome suggested by the large tree reflect accidental properties of the data matrix D , which do not necessarily hold in the population from which D has been drawn. The optimal tree-structured predictor, or honest tree (honest because less prone to overfitting) is then selected among all the subtrees containing the root of the large tree. Following the tree metaphor, this second phase is termed “pruning”; in particular, the procedure proposed in CART is known as cost-complexity pruning. It is an algorithm that selects the honest tree by minimizing an appropriate cost-complexity function.

To define a cost-complexity function, one needs:

- A measure of cost associated to a predictor, which is smaller the more accurate the predictor. Cost must be a function of the tree structure T and of the data matrix D : $R(T, D)$; more specifically, cost depends on D through the estimates of $\mu(z)$ based on D . It would be tempting to identify $R(T, D)$ with the impurity reduction due to T . Indeed, CART adopts this definition for regression trees; however, for classification trees, the CART authors argue in favor of choosing for $R(T, D)$ the resubstitution–misclassification risk, which reduces to the resubstitution–misclassification error when all errors have the same cost. Resubstitution refers to the fact that these quantities are calculated on the same data set D used for the tree construction: It is a biased estimate of the true quantity.
- A measure of complexity associated to the tree structure, $g(T)$, and not depending on the data. Classification and regression trees define complexity as the number of terminal nodes of the tree.
- A complexity parameter α is a tuning parameter that has to be adjusted from the data through cross-validation, as we shall see later.

Then, the cost complexity of a tree T obtained from the data D is

$$CC_\alpha = R(T, D) + \alpha g(T)$$

Clearly, it is a linear combination of the cost and the complexity of a tree with the complexity parameter defining the tradeoff between the two. For $\alpha = 0$, only cost is important, and the large tree is the best tree. As α increases, complexity overrides cost and the best tree is eventually the trivial tree.

It can be shown that for a fixed α there is a unique subtree T_α of the large tree which minimizes cost complexity. In fact, one can construct a finite increasing sequence of α values associated to distinct subtrees:

$$T_{\text{large}} \supset T_{\alpha_1} \supset T_{\alpha_2} \supset \dots \supset T_{\alpha_{\max}} = \{\text{root}\}$$

This is known as the pruning sequence: It is uniquely associated to the data set D . Clearly, any tree of the pruning sequence is a honest tree for some values of α . Thus, to complete pruning, an algorithm to choose α is needed.

How should the complexity parameter be determined? We have already remarked that $R(T, D)$ is a biased estimate of a theoretical quantity—it is usually smaller than its true value, so that several authors call the bias optimism and describe the bias as optimistic. Now, if we have another set of independent data D^* from the same population as D , we could use the parameters estimated on D to calculate the cost function on D^* . The cost $R(T, D; D^*)$ estimated this way is less optimistic than $R(T, D)$. Ideally, it would be desirable to have a sequence of new independent data sets $D^{*(i)}, i = 1, 2, \dots, B$. Then, one would obtain an even better estimate of true cost by computing the average of the $R(T, D; D^{*(i)})$'s, $\bar{R}^B(T)$, with the associated standard error providing a rough measure of the variation of the estimate. Then a very reasonable choice of α would be the value α^* such that $\bar{R}^B(T_{\alpha^*})$ is minimum.

Since such a sequence of independent data sets is rarely available, the CART monograph proposes to determine α^* by a resampling scheme such as cross-validation or the bootstrap. They opt for a particular kind of cross-validation, termed V -fold cross-validation. The original data matrix D is divided into V approximately equal parts, $D^{(v)}, v = 1, \dots, V$. For each v , a tree is grown on $D/D^{(v)}$, a pruning sequence is obtained from the α 's defined on the large tree, and costs are evaluated on $D^{(v)}$ for each α . For each α , one obtains V costs, which are averaged to obtain $\bar{R}^V(T_{\alpha})$; then α^* is chosen as the value for which $\bar{R}^V(T_{\alpha^*})$ is minimum. We refer to the CART monograph for a thorough presentation and an alternative selection rule which relies also on the standard errors of the $\bar{R}^V(T_{\alpha})$'s.^[2]

Notice that the CART pruning algorithm grows $V + 1$ large trees and pruning sequences. Indeed, there is no reason to expect that distinct $D^{(v)}$ gives rise to the same large tree and pruning sequence. On the other hand, the honest tree is selected from the pruning sequence of the large tree obtained from data matrix D ; the other large trees remain “invisible” to the analyst.

Other Algorithms for Constructing CART

There is a large family of algorithms that attempts to improve on CART. Here, we will discuss work aimed at improving the accuracy of prediction. The elements of the CART algorithm that can be changed to improve predictive accuracy are mainly the search for the best split at a node and the approach to the selection of the honest tree.

One of the shortcomings of CART methodology affecting prediction accuracy is the bias inherent in split selection: Since a split is chosen by extensive searches, predictors with many levels are favored and tend to appear more often than they should in the large tree. To reduce this bias Loh and Vanichsetakul proposed the algorithm FACT

for classification trees.^[3] The authors separate the problem of variable selection from that of split selection. By performing a discriminant analysis at each node for each predictor, they select first one predictor, and then split on the selected predictor. The discriminant analysis approach, naturally leads to a multiway split. FACT also avoids the CART-style pruning approach, but uses a node-based stopping rule. A later development led to the algorithm QUEST, which yields binary trees and allows pruning as an option.^[4] FACT and QUEST reduce the bias in favor of predictors with many levels. Moreover, QUEST was shown to produce the best predictive accuracy among 33 “old and new classification algorithms.”^[5] The algorithm GUIDE extends the split selection approach to the case of regression trees.^[6]

Prediction accuracy can also be improved by merging nodes from different parents, if appropriate. This can be done by amalgamating leaves of the honest tree or during the process of tree construction.^[7–9] The resulting structure, however, is no longer a tree in the strict sense, but an “induction graph.”

GENERALIZED PREDICTION TREES

We have described so far CART and the elements of their construction. Several generalizations have been proposed with two distinct purposes:

1. Improving prediction accuracy.
2. Predicting aspects of the distribution of the outcome other than its expected value. In particular, this is useful when working with an outcome that is more complex than a categorical outcome or a one-dimensional continuous outcome.

One way to improve prediction further is to adopt predictive structures more general than a tree. On the other hand, one can build predictors for other aspects of the distribution while preserving the basic tree structure.

The approach proposed aims at both purposes.^[8,10] The key idea is to explicitly assume that the purpose of the prediction is to associate to a set of the predictor variables an aspect of the outcome distribution represented by a parameter θ that does not necessarily determine the distribution completely. To determine the distribution completely one needs (θ, ψ) , where ψ can be seen as a “nuisance” parameter to be treated in an appropriate way. The parameter of interest may also be affected linearly by other covariates, x , say, termed leaf variables (criterion variables in earlier versions). The nuisance parameter may also depend on other covariates, and the RECPAM algorithm distinguishes between root and virtual variables (global and local variables in an earlier version). Without entering in the details we will simply state here that this idea gives rise to an “enriched” tree structure.^[8,10] A generalized RECPAM

tree is indeed a tree, possibly with partially merged leaves; however, the prediction at the leaves occur through a regression equation containing the leaf variables, possibly supplemented by an additional root regression equation, the coefficients of which do not vary across leaves. Another important feature of RECPAM methodology is the replacement of the notion of impurity at a node with that of generalized information, basically a “badness of fit measure” that compares the estimated parameter θ with the data. In most cases developed so far, the badness of fit measure is the deviance or some quantity with properties similar to those of the likelihood—which justifies the term “information.” Then, one can easily define quantities, such as generalized information gain (corresponding to “impurity reduction” of CART) and information loss, which are at the heart of the RECPAM algorithm. In the standard cases of classification and regression trees and in many other examples, these quantities correspond to likelihood ratio statistics.

Tree-Structured Classification and Regression

The idea of “adding regression equations” to a tree structure, termed tree-structured regression or treed regression, has been independently pursued by other authors.^[11–13]

The RECPAM approach contains essentially the tree-structured models described above. Its main limitation is that the analyst has to specify a priori which variables are predictors, leaf, virtual, and root. This may be possible in highly structured problems with rich peripheral knowledge. For example, one may be interested in comparing the effect of two treatments as a function of specified variables (effect modifiers)—in other words, one wishes to assess the potential interactions between treatment and other covariates. Then, one simply specifies the binary treatment variable as leaf, the constant term as a nuisance parameter to be treated as a virtual variable, and all other variables as predictors. On the other hand, in unstructured problems such as those occurring in a data mining context, it is hard to decide which variables enter the linear predictors. When in doubt one can require that all variables are leaf variables, as some of the above-cited authors do. However, if only a few predictors are truly needed at each leaf, this option will be too noisy and will result in prediction inaccuracy. Therefore, it is highly desirable to reduce noise by automatic selection of the variables entering each of the regression equations at the leaves. This was achieved by Chaudhuri et al. and Kim and Loh, who proposed accurate and computationally efficient algorithms for tree-structured classification and regression.^[12,13]

Trees for More General Outcomes

Biomedical studies are increasingly based on more complex outcomes than a class variable or a single continuous variable. Count data occur, for instance, in the study of disease distribution, and they are treated through Poisson regression

or generalizations thereof (negative binomial regression or partially specified regression with parameter estimation based on quasi-likelihood). Survival outcomes with censoring have been common in clinical biostatistics for quite some time, and they have stimulated the development of a large family of regression models, among which the Cox model continues to occupy a central place. More recently, event history data have become increasingly common to describe the evolution of chronic diseases with recurrent acute phases. Repeated measurements of continuous, binary, or categorical variables and, more generally, clustered measurements such as those that occur in meta-analysis and in multilevel systems, are now active areas of methodological and applied statistical research. The interest often centers on assessing the effect of some covariates on outcome: linear regression of some kind (regression for correlated continuous or discrete outcome, event history regression, multilevel regression, etc.) is usually the first tool on which the analyst relies. When a full statistical model is available for the regression model, likelihood-based estimation of the parameters is most commonly used, as, for example, in mixed models for continuous outcomes. When only a portion of the statistical model can be specified, the likelihood is replaced by closely related quantities, such as a partial likelihood (Cox regression and its generalization for repeated events), quasi-likelihood (regression for categorical and count correlated data), etc. It is natural to develop tree-based methods as an alternative to or complement of nonstandard regression models and this has been done successfully for some situations.

The RECPAM approach, with its reliance on likelihood-based measures of information gain, offers a natural framework for converting a given linear regression model into a tree-growing algorithm, and indeed into a tree-structured regression model.^[8] The RECPAM approach has been developed for the generalized linear model, which includes, among others, trees for binary, Poisson (possibly with overdispersion), and censored exponential outcomes, and for the Cox model.^[7,14] The RECPAM algorithms based on the Cox model contain the unique feature of tree-structured subgroup analysis: a tree can be grown to predict differential treatment effects from predictors, while correcting for the possible direct effect of predictors on survival. RECPAM tree-growing algorithms, treed regression, and subgroup analysis have also been proposed for multivariate normal outcomes.^[15]

The limitation of the actually developed algorithms of the RECPAM family is that they are not always optimized for predictive accuracy and computational efficiency and further work is needed in most cases to improve on these features. On the other hand, there are very accurate and efficient algorithms that, while less general than RECPAM and based on different approaches, achieve similar purposes extremely well. See, for example, the tree-structured Cox regression algorithm proposed by Ciampi and Loh for survival outcome and the treed generalized regression algorithm of Chaudhuri et al.^[12,16,17]

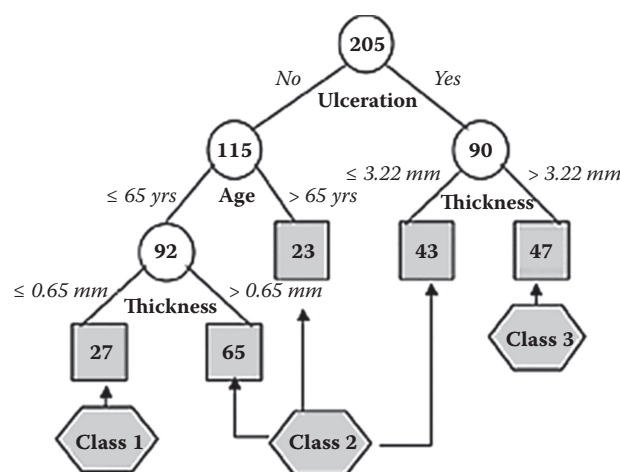


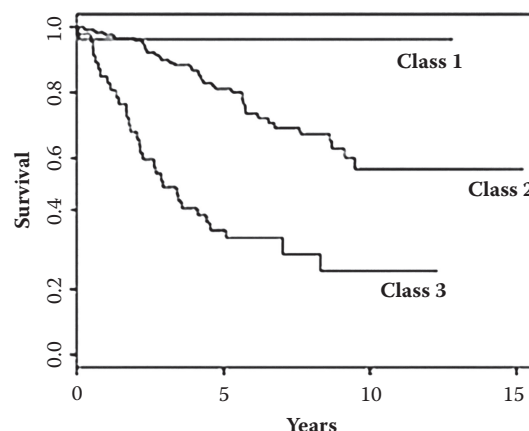
Fig. 2 Prediction tree for survival curves: melanoma data

Other approaches to tree-growing for survival outcome have been developed.^[18–23] Prediction trees for longitudinal outcomes have also been independently proposed by Segal.^[24] A tree-growing algorithm and treed regression for binary correlated outcomes has been developed by Ann and Chen.^[25] We refer to Zhang et al. for an excellent review of the use of trees in the health sciences and interesting original contributions.^[26]

To conclude this section, a prediction tree for survival curves, constructed from real data, is presented in Fig. 2. The data, included with S-plus 6.2, as data frame Melanoma, were collected on patients with malignant melanoma at Odense University Hospital, Denmark, by K.T. Drzewiecki. The outcome is censored survival time and the prognostic variables include sex, age, tumor thickness (in mm), and ulceration (a binary variable denoting the presence or absence of ulceration). The tree was constructed using the generalized linear model module of RECPAM, under the hypothesis of exponential survival times conditional on the covariates. This tree structure has five clearly identifiable leaves, which after the amalgamation step, yield three distinct prognostic groups (RECPAM classes). The survival curves of the prognostic group are also shown in the graph. It is easy to read off the graph the definition of each prognostic group, and to visualize the corresponding survival curve. Clearly, the “best prognosis” group comprises patients younger than 65, without ulceration and with a tumor thickness less than 0.65 mm; the “worst prognosis” group consists of patients with ulceration and tumor thickness greater than 3.22 mm; the remaining patients constitute an “intermediate prognosis” group.

SOME DIRECTION OF CURRENT RESEARCH

We have described a basic structure—a tree—possibly with equations at some leaves and some merger of nodes or



leaves and we have summarized a number of approaches to extract such a structure from data. Although minor improvements are still possible within the basic construction process, some of the most promising ongoing research is aimed at extending the basic model further. More general structures and constructive algorithms to extract them from data have been proposed. In general, the new structures lose some interpretability in exchange for better predictive accuracy.

Trees with Soft Nodes

One recent proposal is that of a tree with soft nodes.^[27] The idea is very simple and it applies to splits on continuous predictors. Consider, for example, a split on the variable age. It seems interesting to replace a question of the type: “Is the subject older than 60?” with “Is the subject old?” Since “being old” is not a sharply defined concept, one can interpret the question as “Should the subject be considered old and with what probability?” A “soft split” as in Fig. 3, therefore, replaces a “hard split,” where the sigmoid curve gives the probability of “going left.” A hard node then can be seen as a limiting case of a sigmoid curve that becomes a “step function.”

An algorithm to construct a soft tree to predict binary variables has been developed and empirical evidence shows that it produces generally better predictions than a hard tree. Work is in progress to extend the algorithm to continuous and categorical variables. The soft tree is strongly related to a much more general structure, a hierarchy of expert, proposed as a hybrid neural network in the literature of machine learning.^[28] In such a hierarchy, leaves are seen as low-level experts, and nodes are “superexperts” who make predictions by weighting the decision of their immediately subordinate lower-level experts (children, in tree language). In the general context, weighting is represented by a neural network or a linear combination of predictors, which results in a marked

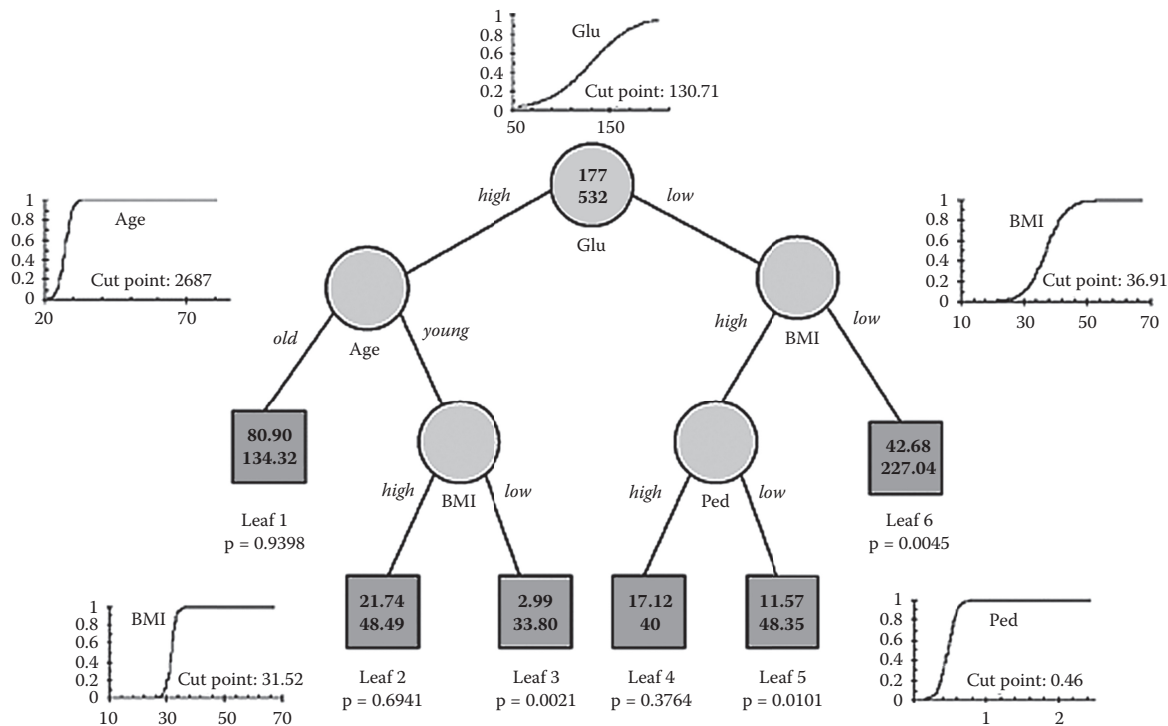


Fig. 3 Prediction tree with soft nodes for a binary outcome (From Ref. [27]; p. 1154.)

loss of interpretability. The tree with soft nodes reintroduces interpretability, by replacing neural nets with soft splits on a single variable. Moreover, while a hierarchy of experts is based on a prespecified number of experts disposed in a predefined hierarchical structure, the tree with soft nodes is entirely constructive in that it extracts from data both the hierarchical structure and the form of the soft splits.

Forests and Bayesian Trees

A completely different direction is being explored.^[29–31] We have already remarked at the end of the section Pruning in CART that the cross-validation algorithm involves the construction of several “large trees,” which are used in a very limited way and only as a means to choose a subtree of the main large tree. The main idea behind a variety of approaches is to use more than one tree as a predictor: This multiplicity of trees is sometimes called a forest. The problem of combing different trees has given rise to several algorithms such as “bagging,” “arcing,” and others.^[32] Gains in predictive accuracy are impressive, but the analyst has to give up the simple, easily interpretable single-tree predictor; instead, the concept of forest and averaging has to be explained to the nonstatistician expert, rarely an easy task.

There is another idea leading also to forests: Bayesian trees.^[33–35] In the Bayesian inferential paradigm, one needs

to specify priors on the space of all trees. Given the data, a family of models and some heuristics, one obtains a posterior probability on the space of all trees, which of course can be seen as a forest. In principle, a Bayesian prediction is obtained by averaging predictions from all possible trees using the posterior distribution. In practice, one reports the tree with the highest posterior probability and a few “meaningful” trees of posterior probability close to the maximum. In a way, Bayesian trees offer a principled approach to tree averaging within Bayesian inference. Bayesian trees, though very promising, need much further exploration. One of the main obstacles to more extensive use so far is scalability of the basic algorithms: It is not clear to what extent these algorithms work in reasonable time on massive data sets.

Predictive Clustering

While trees occupy an important position in supervised learning, they are beginning to appear also as a tool in unsupervised learning. The idea proposed is a new development of two earlier approaches.^[36–39] One can state the problem of clustering within a prediction framework as follows: given a vector of variables x , determine a partition of its components: $x = (y|z)$ such that z optimally predicts (in some sense) y . For example, one can require that the components of y be conditionally independent given z .

CONCLUSIONS

As we have seen, research on prediction trees is an extremely fertile area. Not only are very useful predictive tools now available, but also new tools are being developed constantly. By their nature, prediction trees continue to attract the attention of several communities of researchers: among others, the statistical community and the machine learning community. Indeed, one of the most promising trends in ongoing research is the convergence of independently developed traditions from different communities. Integration of computer science and statistical expertise, which has already proved extremely fruitful, will hopefully continue. Some advances will only be possible through a multidisciplinary combination of efforts: this is necessary to meet the challenges and opportunities arising from the need of tackling poorly structured prediction problems suggested by the availability of massive data sets.

REFERENCES

1. Morgan, J.N.; Sonquist, J.A. Problems in the analysis of survey data and a proposal. *J. Am. Statist. Assoc.* **1963**, *58*, 415–434.
2. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees*; Chapman and Hall: New York, 1984.
3. Loh, W.-Y.; Vanichsetakul, N. Tree-structured classification via generalized discriminant analysis. *J. Am. Statist. Assoc.* **1988**, *83*, 715–728.
4. Loh, W.-Y.; Shih, Y.-S. Split selection methods for classification trees. *Statist. Sinica* **1997**, *7*, 815–840.
5. Lim, T.-S.; Loh, W.-Y.; Shih, Y.-S. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mach. Learn. J.* **2000**, *40*, 203–228.
6. Loh, W.-Y. Regression trees with unbiased variable selection and interaction detection. *Statist. Sinica* **2002**, *12*, 361–386.
7. Ciampi, A.; Chang, C.-H.; Hogg, S.A.; McKinney, S.M. Recursive partition: a versatile method for exploratory data analysis in biostatistics. In *Biostatistics*; McNeil, I.B., Umphrey, G.J., Eds.; D. Reidel Publishing Company: Holland, 1987; *5*, 23–50.
8. Ciampi, A. Constructing prediction trees from data: the RECPAM approach. In *Computational Aspects of Model Choice*; Antoch, J., Ed.; Physica-Verlag: Heidelberg, Germany, 1992; 105–152.
9. Zighed, D.A.; Rakotomalala, R. *Graphes d'Induction—Apprentissage et Data Mining*; Hermes: France, 2000.
10. Ciampi, A. Generalized regression trees. *Comput. Statist. Data Anal.* **1991**, *12*, 57–78.
11. Chaudhuri, P.; Huang, M.-C.; Loh, W.-Y.; Yao, R. Piecewise-polynomial regression trees. *Statist. Sinica* **1994**, *4*, 143–167.
12. Chaudhuri, P.; Lo, W.-D.; Loh, W.-Y.; Yang, C.-C. Generalized regression trees. *Statist. Sinica* **1995**, *5*, 641–666.
13. Kim, H.; Loh, W.-Y. Classification trees with bivariate linear discriminant node models. *J. Comput. Graph. Statist.* **2003**, *12*, 512–530.
14. Ciampi, A.; Negassa, A.; Lou, Z. Tree-structured prediction for censored survival data and the Cox-model. *J. Clin. Epidemiol.* **1995**, *48* (5), 675–689.
15. Ciampi, A.; Hendricks, L.; Lou, Z. Discriminant analysis for mixed variables: integrating trees and regression models. In *Multivariate Analysis: Future Directions 2*; Cuadras, C.M., Rao, C.R., Eds.; Elsevier Science Publishers B.V.: New York, 1993; 3–22.
16. Ciampi Loh, W.-Y. Survival modeling through recursive stratification. *Comput. Statist. Data Anal.* **1991**, *12*, 295–313.
17. Ahn, H.; Loh, W.-Y. Tree-structured proportional hazards regression modeling. *Biometrics* **1994**, *50*, 471–485.
18. Ciampi, A.; Bush, R.S.; Gospodarowicz, M.; Till, J.E. An approach to classifying prognostic factors related to survival experience for non-Hodgkins's lymphoma patients. *Cancer* **1981**, *47*, 1–621.
19. Segal, M.R. Regression trees for censored data. *Biometrics* **1988**, *44*, 35–47.
20. Davis, R.B.; Anderson, J.R. Exponential survival trees. *Statist. Med.* **1989**, *8*, 947–961.
21. Ciampi, A.; Thiffault, J. Pruning regression trees for censored survival data: the RECPAM approach. *Commun. Statist. Theory Methods* **1989**, *18* (9), 3373–3388.
22. LeBlanc, M.; Crowley, J. Survival trees by goodness of split. *J. Am. Statist. Assoc.* **1993**, *88*, 457–467.
23. Negassa, A.; Ciampi, A.; Abrahamowicz, M.; Shapiro, S.; Boivin, J.-F. Tree-structured prognostic classification for censored survival data: validation of computationally inexpensive model selection criteria. *J. Am. Statist. Assoc.* **2000**, *67* (4), 289–318.
24. Segal, M.R. Tree-structured methods for longitudinal data. *J. Am. Statist. Assoc.* **1992**, *87*, 407–418.
25. Ahn, H.; Chen, J.J. Tree-structured logistic model for over-dispersed binomial data with application to modeling developmental effects. *Biometrics* **1997**, *53*, 435–455.
26. Zhang, H.; Singer, B. *Recursive Partitioning in the Health Sciences*; Springer: New York, 1999.
27. Ciampi, A.; Couturier, A.; Li, S. Prediction trees with soft nodes for binary outcomes. *Statist. Med.* **2002**, *21*, 1145–1165.
28. Jordan, M.I.; Jacobs, R.A. Hierarchical mixtures of experts and the EM algorithm. *Neural Comput.* **1994**, *6*, 181–214.
29. Breiman, L. Arcing classifiers (with discussion). *Ann. Statist.* **1998**, *26*, 801–849.
30. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140 (2).
31. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45* (1), 5–32.
32. Hastie, T.; Tibshirani, R.; Friedman, J.H. *The Elements of Statistical Learning; Data Mining, Inference, and Prediction*; Springer: New York, 2001.
33. Chipman, H.; George, E.I.; McCulloch, R.E. Bayesian CART model search (with discussion). *J. Am. Statist. Assoc.* **1998**, *93*, 937–960.
34. Chipman, H.; George, E.; McCulloch, R. Bayesian treed models. *Mach. Learn.* **2002**, *48*, 299–320.

35. Denison, D.G.T.; Mallick, B.K.; Smith, A.F.M. A Bayesian CART algorithm. *Biometrika* **1998**, *85*, 363–377.
36. Ciampi, A. Classification and discrimination: the RECPAM approach. In *COMPSTAT '94, Proceedings of the 11th International Symposium on Computational Statistics*, Vienna, Austria, Aug 20–26, 1994; Dutter, R., Grossmann, W., Eds.; Physica-Verlag: Heidelberg, Germany, 1994; 129–147.
37. Chavent, M. A monothetic clustering method. *Pattern Recognit. Lett.* **1998**, *19*, 989–996.
38. Lance, G.N.; Williams, W.T. Note on a new information—statistic classificatory program. *Comput. J.* **1968**, *11* (2), 195–195.
39. Gower, J.C. Maximal predictive classification. *Biometrics* **1974**, *30*, 643–654.

Premarket Applications: Registries

Nelson Lu, Yunling Xu, and Lilly Q. Yue

Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

National and international registries provide real-world evidence and can be used to potentially reduce the cost of clinical trials. These registries are playing a more and more important role in assessing the risk and benefit of medical devices. In this entry, related statistical and regulatory challenges and opportunities when utilizing registry data are discussed in detail from the following aspects: (1) adequacy of registry data quality; (2) bias introduced at every stage and aspect of study when incorporating registry data; (3) scientific validity of the study design; and (4) reliability and interpretability of the study results.

Keywords: Medical device; Registry; Data quality; Observational comparative studies; Study design; Propensity score.

INTRODUCTION

It is well perceived that randomized controlled trials (RCTs) provide the highest level of evidence in evaluating medical products, if they are designed and conducted properly. Sometimes for medical device clinical studies, conducting an RCT is not feasible due to ethical or practical issues, or a less burdensome approach is sufficient to provide the valid scientific evidence needed to support a marketing application.^[1] As a result, alternate study designs may be considered, including observational (non-randomized) clinical studies, to support medical device premarket applications.

Although RCTs have many advantages, they certainly have limitations. They require greater cost and time, and the generalizability of clinical findings to real-world medical practice may be limited, given the strict enrollment criteria that are typical in RCTs. Therefore, there has been increasing interest in expanding and integrating clinical research into more real-world settings and in utilizing real-world “big data” such as high-quality national or international registries.

A patient registry is an organized system using observational study methods to collect uniform clinical and other data to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and it serves one or more predetermined scientific, clinical, or policy purposes. A registry database is a file (or files) derived from the registry. Furthermore, registries are classified according to how their populations are defined. For example, product registries include patients exposed to biopharmaceutical products or medical devices; disease or condition registries are defined by patients having the same diagnosis.^[2]

The use of registry data has been explored for premarket regulatory decision-making for some medical devices. Here are some of such examples. The Interagency Registry for Mechanically Assisted Circulatory Support has been used to form a control group in the evaluation of a new ventricular assist system. Similarly, control groups have been formed from registries such as Extracorporeal Life Support Organization and the Vascular Quality Initiative. United Network for Organ Sharing (UNOS) registry data have been employed to supplement the clinical data collected in an RCT for the evaluation of an investigational device. The analysis of long-term survival data in UNOS provided great assistance in regulatory decision-making. To investigate an alternative access approach for transcatheter aortic valve replacement in inoperable patients with aortic stenosis, a study embedded in transcatheter valve therapy registry was designed.

In the following sections, related statistical and regulatory challenges and opportunities for utilizing registries for confirmatory medical device studies are discussed in detail.

POTENTIAL OPPORTUNITIES

For a well-established device technology and medical practice, a well-designed and executed national or international registry has the potential to: (1) complement the knowledge obtained from “traditional” clinical trials to answer scientific and clinical questions; (2) support applications, including therapeutic development, outcome research, patient care, quality improvement, medical product surveillance, and well-controlled effectiveness studies; and (3) allow researchers to answer health-care questions

efficiently, with savings in both time and cost, and for broader patient populations.^[3]

Some opportunities in the use of registry data exist for pivotal confirmatory medical device studies. The information obtained from a registry may generate a hypothesis to be tested in confirmatory clinical studies. Registry information may serve as supplementary data and provide additional information. For certain types of medical devices, the registry data may be used to derive a performance goal or objective performance criteria, which can be used in a single-arm clinical study.

Another usage of registry data is to form a control group (with patient-level data available) to serve as a compactor to the investigational device. Statistical methodology and considerations are extensively discussed in this entry. It is possible that a prospective clinical study can be embedded within a registry. In such a case, the statistical methodology used in traditional clinical trials is applicable.

For an existing approved medical device, the information from real-world use may be captured in some registries. Such information may be utilized to provide evidence for labeling updates or labeling extensions, such as expanded indications and expanded patient populations.

GENERAL CONSIDERATIONS

A registry needs to possess certain features in order to be utilized in the evaluation of confirmatory medical device studies.^[4] The discussion here focuses on the statistical and study design issues only.

Data Quality and Relevance

To obtain sound scientific evidence from a study, it is essential to have data with high quality. It is certainly true for evidence generated from registry data as well. The data quality of registries varies widely due to factors such as the purpose of the registry, carefulness in designing phase, various practical constraints, adequacy of site training, data handling, etc. To serve as the primary data source used to evaluate the risk and benefit of a medical device, it is important that the registry data have adequate quality to generate valid scientific evidence.

Important factors that need to be assessed include the relevance and reliability of the registry. They are extensively discussed in the U.S. Food and Drug Administration (FDA) Draft Guidance “Use of real-world evidence to support regulatory decision making for medical devices.”^[4] Please note that the Draft Guidance documents represent FDA’s thinking, but are not implemented until being finalized. Relevance indicates that the registry data can adequately address the applicable regulatory question or requirement. Ideally, the overall assessment of relevance would conclude that the existing observational data source is reliable, complete, consistent, accurate, and contains all

critical data elements necessary for evaluating the performance of a device in the applied regulatory context. Important factors for assessing the reliability of registry data include how the data were collected (data accrual), whether the data as collected are complete, accurate, and adequate for answering the question in hand (data adequacy), and whether the people and processes in place during data collection and analysis provide adequate assurance that bias is minimized and data quality and integrity are sufficient (data assurance).

Good practice to ensure high data quality in a well-conducted clinical study is important in registries as well. Some practices to enhance the relevance and reliability of a registry database are as follows: (1) accurately collect and record data; (2) adequately monitor the registry study to maintain the data integrity; (3) collect relevant information, including all key baseline covariates and clinical outcomes in the registry; (4) minimize missing data in baseline covariates and number of patients not completing the endpoint evaluation; and (5) utilize mechanisms such as central clinical event committees and core labs to reduce bias and variability of the study.^[5]

Registry versus Clinical Study

Device registries reflect the product use in the real-world. A wider patient population is generally expected in a registry, compared to a clinical study with defined inclusion/exclusion criteria. In addition, the skill levels and experience of the physicians are more varied in the real-world setting than in a clinical trial setting as well as the levels of ancillary care. Patients are treated under less-controlled environment in the real-world as there are usually no clinical protocols, and practices may vary among sites/physicians.

When it is intended to compare results of a clinical study with results of control groups formed from a registry database, careful considerations are needed. For a registry to serve as an acceptable data source, it needs to be relevant in terms of the ability to adequately address the applicable clinical questions under consideration. The clinical comparability between the registry database and the investigational study needs to be demonstrated with respect to patient population, treatment management, patient follow-up, definition of endpoints, the clinical event adjudication, etc.

It is important to assess whether clinical outcomes are affected due to different features between a registry and a well-conducted clinical study. In some cases, the clinical performance exhibited in a clinical trial tends to be better than that exhibited in a registry. Such situations may arise as, for example, patients in clinical studies are more likely to receive optimal medical treatment and/or guideline-recommended care. This could have important ramifications when the registry data are used as a comparative control, as the treatment effect of the investigational

device relative to the control is likely to be overestimated. In some situations, it is possible that the clinical events of subjects in a registry are less likely to be detected than subjects in a clinical trial. Note that oftentimes these features may be difficult to be captured in registries, and therefore such bias issues cannot be properly addressed via any statistical methodology.

CONTROL FORMED FROM REGISTRIES

One of the uses of registry data in the confirmatory medical device evaluation is to form a control group to serve as a comparator to the investigational device. For this purpose, the registry is not limited to a device registry and it could be a device, medication, surgery, etc. Appropriate registry selection, careful study design and planning, and pertinent statistical methodology are of paramount importance for reaching a valid comparative conclusion.

In selection of an appropriate registry, points discussed in section “General Considerations” need to be considered. The comparability between treatment groups is critical. Bias may be introduced from a lack of comparability. It is possible that treatment effect is confounded with time due to evolution of medical technology, learning effect of device use, changes in patient population, etc. Such a concern may be alleviated if there is sufficient overlap in the time frames between the investigational study and registry. If treatment effects are expected to be different among regions, the regions of a registry are better to be matched with where the investigational study is conducted.

In an RCT, the distributions of all observed and unobserved baseline covariates are expected to be balanced between two treatment groups. However, the same cannot be expected in an observational study as the treatment selection and study outcomes could be influenced by subject baseline characteristics. Without proper adjustment, great bias may be presented in the treatment effect estimates. A commonly used approach to reduce the bias is propensity score methodology, which was introduced by Rosenbaum and Rubin.^{[6][7]} The propensity score is the probability of treatment assignment conditional on observed baseline characteristics. Propensity scores are often estimated using a logistic regression model, with the treatment assignment as the dependent variable and baseline covariates as independent variables. With the estimated propensity scores, further design and analysis can be performed. One commonly adopted method in confirmatory medical device studies is to stratify subjects into groups (of roughly equal sizes) with similar propensity scores. Another method is to match each subject in the investigational device group with a unique control subject with a similar propensity score.

One critical feature of an RCT is that the study is prospectively designed. Naturally, an RCT is designed without access to any outcome data from either treatment groups.

This ensures convincing and interpretable treatment comparison results regarding outcomes. Such a feature is critical to be maintained in confirmatory, comparative observational studies under the regulatory setting. Without proper effort, the data can be analyzed with outcome data in sight. The objectivity and validity of the study design and the interpretability of the study results can greatly suffer. As pointed out by Rubin,^[8] “it is essentially impossible to be objective when a variety of analyses are being done, each producing an answer, favorable, neutral, or unfavorable to the investigator’s interests.” To address this issue, Yue, Lu, and Xu^[5] proposes a two-stage study design process.

The first design stage occurs before the initiation of the investigational study. In this stage, the preliminary sample size of the investigational device study is determined. An independent statistician is identified to perform the propensity score estimation and design. Masking mechanisms, such as a firewall, are planned to control access to any outcome data from this independent statistician during the entire design process. The patients’ baseline covariates to be collected and included in the propensity score modeling are specified.

The second design stage is conducted once the subject enrollment is completed and baseline covariates data are available. The previously identified independent statistician performs the task of estimating the propensity scores, grouping subjects based on the propensity scores, and assessing the covariate balance. An iterative process is usually taken until the covariate balance is reached. The statistical analysis plan is finalized at this stage. In the entire process, only the treatment assignment and baseline covariates data are needed. Any clinical outcome data and follow-up information are not needed nor are accessed.

The following lists some practical issues in implementing this practice.

Covariate Selection

In the first design stage, one task is to identify all covariates that are to be collected in the clinical study, included in the propensity score estimation model, and checked with balance between treatment and control groups. As reducing bias is an important task from the regulatory perspective, the list of covariates is usually prespecified to serve this objective. When a covariate that is related to both treatment assignment and outcomes is missing from the propensity score model, one of the key assumptions in propensity score methodology, ignorable treatment assignment assumption, may not be fulfilled.^[9,10] To reduce the possibility of such a scenario, a good practice is to include as many covariates in the list as possible (refer to the study by Yue, Lu, and Xu^[5] for more discussion).

Based on this principle, one of the key criteria to select a registry is that the information of all known important confounders is collected in the registry.

Missing Data in Covariates

In practice, it is common that there is a substantial amount of missing values in some of the covariates. If the analyses are performed based on data, excluding subjects with missing covariates, sample size is reduced and, more importantly, study conclusion may be questionable. Therefore, imputation of missing data is usually performed. For example, commonly used multiple imputation method or some adaptations thereof may be adopted (refer, for e.g., the study by Qu and Lipkovich^[11] for discussions). Simulation study results are presented in Mitra and Reiter^[12] to compare two approaches of estimating propensity scores after multiple imputation of missing covariates value.

Covariate Balance Diagnosis

In the propensity score adjusted analysis, inferences on treatment effect are valid only if covariates are balanced based on the selected propensity score design.^[6,13,14] Therefore, it is important to demonstrate the balance of baseline characteristics for the design based on the estimated propensity scores.

If the balance is not reached, alternative propensity score estimation modeling or a complete redesign of the control group may need to be performed. An iterative process between propensity score estimation/design and covariate balance assessment may be needed to finalize the propensity score estimation and design.

To demonstrate the balance of baseline characteristics, the common practice is to assess the marginal distribution of every observed covariate that is identified in the first stage design. Some appropriate methods include the graphical presentation and numerical quantities such as standardized (absolute) mean difference and variance ratios.^[14,15] The application of standardized (absolute) mean difference when it is utilized under propensity score stratification technique is discussed in the studies by Yue et al.^[3,5] and Li et al.^[16]

Selection of Subjects from Registry

It is important that the analysis set is specified in the second design stage. In an observational comparative study, oftentimes not all subjects from two treatment groups are comparable in terms of the distribution of propensity scores. In the practice, some subjects from one or both groups are removed to obtain better matches. While such a practice is common and well accepted in general research, it may be problematic in the confirmatory medical device studies. Discarding subjects in the investigational device can alter the intended population and make the resulting population difficult to be defined for labeling. Because of this regulatory constraint, it is important that the range of propensity scores for control group should cover the range of the subjects treated with the investigational device.

Sample Size and Power

The preliminary sample size is planned in the first design stage. The planning could be fairly challenging, and careful consideration may be demanded due to some uncertainties. The major challenge is due to the uncertainty regarding the degree of comparability in subjects' characteristics between both groups. Poorer comparability may result in the requirement of a larger sample size to achieve a certain power. Even if the number of subjects in the registry appears to be abundant, it is possible that the number of subjects that are comparable to the subjects in the investigational device group with respect to patient characteristics is much smaller.

When a control group is formed from a registry that is ongoing at the same time as the clinical study, the number of patients in the registry is unknown at the design stage. This may add even more complexity in the sample size planning.

To deal with these uncertainties, it may be better to take a more conservative approach. A larger sample size may be desirable in order to safeguard against the (unexpected) poor comparability and allow for greater flexibility in the second design stage.

CONCLUSION

National and international registries provide real-world evidence and can potentially be used to reduce the cost of clinical trials. To complement the major role of RCTs, these registries can play an important role in assessing the risk and benefit of medical devices in confirmatory studies. The key is to appropriately address statistical and regulatory challenges through the following actions: (1) ensuring the adequacy of data quality and (2) minimizing potential bias via scientifically valid study designs and proper statistical methods. By doing so, the study results could be reliable and interpretable, and the evidence may be generated to support regulatory decisions.

REFERENCES

1. 21CFR860.7(c)(2). Available at <http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfCFR/CFRSearch.cfm?FR=860.7> (accessed September 2016).
2. Gliklich, R.; Dreyer, N.; Leavy, M., Eds. Registries for Evaluating Patient Outcomes: A User's Guide. 3rd Ed.; *AHRQ Publication No. 13(14)-EHC111*; Agency for Healthcare Research and Quality: Rockville, MD, 2014.
3. Yue, Q.L.; Campbell, G.; Lu, N.; Xu, Y.; Zuckerman, B. Utilizing national and international registries to enhance pre-market medical device regulatory evaluation. *J. Biopharma. Stat.* **2016**, *26*, (6), 1136–1145.
4. US Food and Drug Administration. Use of real-world evidence to support regulatory decision making for medical

- devices. Draft guidance for industry and Food and Drug Administration staff. July 27, 2016. Available at <http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM513027.pdf> (accessed September 2016).
5. Yue, Q.L.; Lu, N.; Xu, Y. Designing premarket observational comparative studies using existing data as controls: Challenges and opportunities. *J. Biopharm. Stat.* **2014**, *24* (5), 994–1010.
6. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika*. **1983**, *70* (1), 41–55.
7. Rosenbaum, P.R.; Rubin, D.B. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* **1984**, *79* (387), 516–524.
8. Rubin, D.B. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Serv. Out. Res. Methodol.* **2001**, *2* (3), 169–188.
9. Rubin, D.B.; Thomas, N. Matching using estimated propensity scores: Relating theory to practice. *Biometrics*. **1996**, *52* (1), 249–264.
10. Hill, J.L.; Reiter, J.P.; Zanutto, E.L. A comparison of experimental and observational data analyses. In *Applied Bayesian Modeling and Causal Inference From an Incomplete-Data Perspective*; Andrew, G., Meng, X.L., Eds.; Wiley: Hoboken, NJ, 2004, 44–56.
11. Qu, Y.; Lipkovich, L. Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Stat. Med.* **2009**, *28* (9), 1402–1414.
12. Mitra, R.; Reiter, J.P. A comparison of two methods of estimating propensity scores after multiple imputation. *Stat. Methods Med. Res.* **2016**, *25* (1), 188–204.
13. Stuart, E.A. Matching methods for causal inference: A review and a look forward. *Stat. Sci.* **2010**, *25* (1), 1–21.
14. Austin, P. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* **2009**, *28* (25), 3083–3107.
15. Austin, P. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat. Med.* **2008**, *27* (12), 2037–2049.
16. Li, H.; Mukhi, V.; Lu, N.; Xu, Y.; Yue, Q.L. A note on good practice of objective propensity score design for premarket nonrandomized medical device studies with an example. *Stat. Biopharm. Res.* **2016**, *8* (3), 282–286.

Principal Component Analysis

Aiyi Liu and Enrique F. Schisterman

National Institute of Child Health and Human Development, Bethesda, Maryland, U.S.A.

Abstract

Principal components are obtained through a series of variance maximization of the linear functions of the original variables. This entry discusses the concepts, analysis, and some biomedical applications of principal components.

INTRODUCTION

This article discusses the concepts, analysis, and some biomedical applications of principal components. Originated by Pearson^[1] and later formalized by Hotelling,^[2] principal component analysis has become one of the best known techniques of exploratory multivariate data analysis and has found applications in many areas of scientific research. It is a topic that appears in almost every textbook on multivariate data analysis^[3–5] and is the main focus of several technical books.^[5–8] In the past several years, the widespread interest among medical researchers in microarray experiments and gene expression analyses, a rapidly growing technology in human genome studies,^[9,10] has become one of the few new forces that further vitalize this decades-old topic.^[11–15]

Although useful in practice, the aim of principal component analysis is somewhat simple: to reduce data dimensionality while retaining as much as possible the variation among the data. In doing so, the method linearly transforms a large number of intercorrelated variables, often referred to as the original variables, into the same or fewer number of uncorrelated variables, each called a principal component, in a way that the variances of these components are in descending order, so that the first several principal components explain most of the variation among the original variables. Selected according to a prespecified cut point of percentage of total variation explained, or some other criteria, these leading principal components contain all (or most of) the information inherent in the data and hence determine the dimensionality of the data; the remaining components are considered to be noise and contain little information about the data. Data analyses based on these leading principal components are then performed instead of analyses based on the original variables.

In addition to dimensionality reduction, principal component analysis has been shown to be a useful tool in other data analyses such as extracting information, seeking important regressors in regression analysis, and effectively visualizing and clustering subjects. In principle, any data

analyses can be conducted in a low-dimensional variable space based on the selected leading principal components, instead of a high-dimensional variable space with the original variables. However, for the analysis to be efficient and its results reliably interpreted, principal components being selected must retain, to a satisfactory extent, the information contained in the original variables.

PRINCIPAL COMPONENTS

The simplest way to present principal components is perhaps the one using matrix notations.^[4] Let $\mathbf{x} = (x_1, \dots, x_p)'$ be a p -variate random (column) vector $\mathbf{x}$ with mean vector $\mu = (\mu_1, \dots, \mu_p)'$ and positive semidefinite covariance matrix $\Sigma = (\sigma_{ij})$. From standard matrix theory,^[16] there exists a $p \times p$ orthogonal matrix $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_p)$ such that

$$\mathbf{W}'\Sigma\mathbf{W} = \text{diag}(\lambda_1, \dots, \lambda_p) \quad (1)$$

where

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0 \quad (2)$$

are the p eigenvalues of Σ in descending order. The k th column $\mathbf{w}_k$ of $\mathbf{W}$ is an eigenvector of Σ corresponding to the k th eigenvalue λ_k . Using the orthogonal matrix $\mathbf{W}$, we can then transform $\mathbf{x}$ into a new random vector $\mathbf{z} = (z_1, \dots, z_p)'$:

$$\mathbf{z} = \mathbf{W}'\mathbf{x} \quad (3)$$

The p variables z_k , ($k=1, \dots, p$), each being a linear combination of the p original variables x_k , are called the principal components of $\mathbf{x}$. Note that $\text{Cov}(\mathbf{z}) = \mathbf{W}'\text{Cov}(\mathbf{x})\mathbf{W}$, yielding $\text{Var}(z_k) = \lambda_k$. Thus, unlike the original variables x_k , which are correlated and may or may not have any order, the principal components are uncorrelated and ordered according to their variances: The first principal component z_1 has the largest variance, λ_1 , among all unit length linear combinations of $\mathbf{x}$, and the second principal

component z_2 has the largest variance, λ_2 , among all unit length linear combinations of $\mathbf{x}$ that are uncorrelated with z_1 , and so on, so that the k th principal component z_k has the largest variance, λ_k , among all unit length linear combinations of $\mathbf{x}$ that are uncorrelated with its $k-1$ preceding principal components. Another key property of principal components, as clearly seen from the construction process, is that the total variation in $\mathbf{x}$ as measured by

$$\text{tr}[\text{Cov}(\mathbf{x})] = \sum_{k=1}^p \text{Var}(x_k) = \sum_{k=1}^p \lambda_k \quad (4)$$

remains unchanged after transformation, i.e., $\text{tr}[\text{Cov}(\mathbf{z})] = \text{tr}[\text{Cov}(\mathbf{x})]$.

SAMPLE PRINCIPAL COMPONENTS

Often the principal components in Eq. 3 are called population principal components to reflect the dependence on the eigenvectors of the population covariance matrix Σ , which is in general unknown. In contrast, principal components derived based on the sample covariance $\mathbf{S}$, an estimator of Σ , are called sample principal components.

Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be n random observations of the p -variate random vector $\mathbf{x}$. The sample covariance matrix $\mathbf{S}$ is then given by

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \quad (5)$$

where $\bar{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i / n$ is the sample mean vector. Now suppose $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_p \geq 0$ are the p eigenvalues, and $\omega_1, \dots, \omega_p$ the corresponding eigenvectors, of the sample covariance $\mathbf{S}$. Then the k th sample principal component is given by

$$t_k = \omega_k' \mathbf{x}, \quad k = 1, \dots, p \quad (6)$$

Replacing $\mathbf{x}$ by $\mathbf{x}_i$ in Eq. 6 yields the k th principal component score of the i th observation. These component scores, especially the first few leading ones, play an essential role in principal component analysis; statistical analysis procedures are executed with leading principal component scores instead of the original observations.

RELATED INFERENCE ON PRINCIPAL COMPONENTS

Principal component analysis involves estimation of the eigenvalues λ_k and eigenvectors $\mathbf{w}_k$ of the covariance matrix Σ . Suppose that the random vector $\mathbf{x}$ follows a multivariate normal distribution $N(\mu, \Sigma)$. The maximum likelihood estimates of λ_k and $\mathbf{w}_k$ can be obtained accordingly from the sample covariance matrix $\mathbf{S}$, the maximum likelihood

estimate of Σ . (If some eigenvalues are identical, then certain averaging is needed; see Ref. [17].)

Exact joint distribution of the sample eigenvalues (of $\mathbf{S}$), although of little practical use, is derived in Ref. [18]. Asymptotic distributions of the sample eigenvalues are developed by, among others, James^[18] and Anderson.^[19] Chapter 9 of Ref. [4] has a nice inclusion of these developments.

These distributional theories allow us to test some important hypotheses regarding principal components. For example, one can test whether the first several leading principal components selected in the analysis actually account for the desired level of total variance (see Eq. 7). Another important example is to test whether some of the smallest eigenvalues are equal. For these and other testing problems, readers are referred to Chapter 9 of Ref. [4] and Chapter 3 of Ref. [8].

COMPUTATION OF PRINCIPAL COMPONENTS

Many standard statistical packages have the capacity to perform principal component analysis. In both S-PLUS and SAS, the built-in function or procedure share a common name, “princomp.” These two software packages are able to handle most aspects of principal component analysis. Another software, SPSS, also has similar features but allows more rotation and graphical presentation of the principal components via a pull-down menu.

Several methods for numerical calculation of principal components are summarized in Ref. [8] including Hotelling's^[2] power method to compute the largest eigenvalue and its corresponding eigenvectors of a matrix, the QL algorithm and singular value decomposition method to compute all principal components and the corresponding eigenvectors. See Ref. [8] and the references therein for more information.

An important issue that complicates computation of principal components arises when the number p of variables is large. In a typical microarray experiment or medical imaging analysis, for example, p , the number of genes or pixels, is often at least in several hundreds. To compute the eigenvalues and eigenvectors of the $p \times p$ sample covariance matrix $\mathbf{S}$ hence becomes a daunting task and sometimes “mission impossible.” In addition to the methods cited above, a new procedure, block principal component analysis, is developed in Ref. [13] to deal with high dimensionality. The method first groups the p variables into a number of blocks so that variables in the same block have higher correlation than in two different blocks. Principal components of variables in each block are derived and then combined to yield principal components of the original variables. This approach not only simplifies the computation when p (and possibly n) is large, but also has the advantage to get insight into how and what variables are related.

SELECTING PRINCIPAL COMPONENTS FOR ANALYSIS

When replacing the p original variables by the first q ($q < p$) leading principal components in executing a statistical analysis procedure, an important issue is to determine the number q . The most popular (and straightforward) approach is to first specify a cut point $0 < \alpha < 1$ and then choose q such that the first q components explain at least 100α percent of the total variance among the original variables, i.e.,

$$\frac{\sum_{k=1}^q \lambda_k}{\sum_{k=1}^p \lambda_k} \geq \alpha \quad (7)$$

If the eigenvalues λ_k are unknown, then their sample estimates $\hat{\lambda}_k$ are used in Eq. 7.

In general, some loss of information will occur unless $\alpha = 1$. For a reasonably large α (> 0.8), the loss of information is often minimal, and in some cases may lead to more efficient estimation of parameters. (See the section on “Principal Component Regression.”)

A more sophisticated and statistically guided cross-validation procedure in choosing the number q is developed by Krzanowski.^[20] For each q , a predicted value $\hat{x}_{ik}$ of the i th observation of the k th variable x_{ik} is obtained based on the first q principal components obtained using the subset of data that does not include x_{ik} , and the prediction sum of squares is computed. The number q to be chosen is the one that has no significant decrease in the prediction sum of squares when additional principal components are added to the prediction model.

DATA VISUALIZATION

Data visualization via use of principal components is an important application in exploratory multivariate data analysis. With reduced dimensionality by principal component analysis, it is possible to visualize the observations in a two- or three-dimensional space to detect interesting patterns among the observations. To do so, for each observation, we compute, using Eq. 6, the first two or three principal component scores, and then plot these scores in a two- or three-dimensional space. Such visualization process provides not only insight into relationship among observations, but useful preliminary information as well for more sophisticated clustering analysis of the observations.

It should be pointed out that since graphical presentation of the data uses only the first two or three principal components, the visualization results may not reveal the true data structure, unless these components account for a reasonably large amount of information contained in the data.

As an example, we examine the microarray gene expression data of five antiestrogen-responsive and

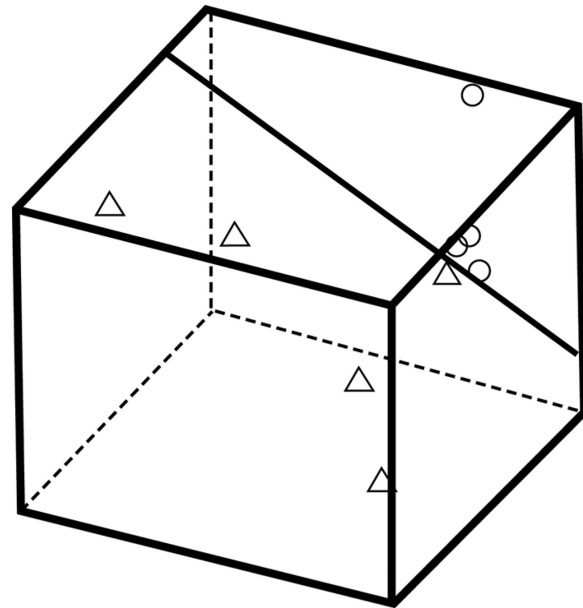


Fig. 1 Three-dimensional visualization of gene microarray data. Δ, responsive variants; ○, resistant variants (From Ref. [21])

four -resistant variants of a human breast cancer cell line. Expressions of a total of 597 genes are measured and the objective is to investigate whether these genes provide discrimination information between responsive and resistant variants. Before building a discriminant model, we can utilize principal components to get some insight into the gene expression data. Fig. 1 plots the first three principal component scores of these variants in a three-dimensional space, and shows clearly that responsive variants are mostly fairly close to each other and so are the resistant variants. Examination of the variance shows that the first three principal components account for 81.2% of the total variance, implying that patterns shown in the figure are a fairly reliable reflection of the true data structure.

PRINCIPAL COMPONENT REGRESSION

Another important application of principal components is their role as independent variables in linear regression analysis. When large numbers of regressor variables are present in regression analysis, multicollinearities—linear functional relationship among some regression variables—may occur. Multicollinearities result in unstable estimates of regression parameters, often with large variances. One approach^[22–24] to overcome multicollinearity in regression is to replace the regressor variables by their first several leading principal components and then obtain the least squares estimation for the regression parameters. This approach will in general yield biased estimators of the regression parameters. However, with properly chosen principal components as the new regressor variables, the resulting estimators often possess smaller mean squared error.

To be specific, consider the linear regression model

$$\mathbf{y} = \beta_0 \mathbf{1}_n + \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (8)$$

where $\mathbf{y}$ is the (column) vector of n observations of the dependent variable, $\mathbf{1}_n$ is the (column) vector of n 1s, and $\mathbf{X}$ is the design matrix of order $p \times n$, with the i th column often representing the i th vector of observations of the p regressor variables, the vector $\boldsymbol{\epsilon}$ consists of n i.i.d. random error terms with mean 0 and common variance σ^2 , and $\boldsymbol{\beta}$ is the vector of regression parameters. Without loss of generality, we can assume that the p rows of $\mathbf{X}$ are centered at 0, i.e., $\mathbf{X}\mathbf{1}_n = \mathbf{0}$, so that the matrix $\mathbf{X}\mathbf{X}'/n$ estimates the covariance matrix of the p regressor variables. With these assumptions, the least squares estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}\mathbf{X}')^{-1} \mathbf{X}\mathbf{y} \quad (9)$$

The mean squared error of $\hat{\boldsymbol{\beta}}$ is

$$\text{MSE}(\hat{\boldsymbol{\beta}}) = \sigma^2 \sum_{k=1}^p \lambda_k^{-1} \quad (10)$$

where $\lambda_1 \geq \dots \geq \lambda_p > 0$ are the p eigenvalues of the matrix $\mathbf{X}\mathbf{X}'$. Clearly, when multicollinearities occur, i.e., some of the λ s are close to 0, $\hat{\boldsymbol{\beta}}$ will have large variance, resulting in unstable estimation.

Now suppose that $\lambda_{q+1}, \dots, \lambda_p$ are the eigenvalues that are close to 0, causing multicollinearities. Let $\mathbf{W}$ be the matrix of order $p \times p$ whose k th column $\mathbf{w}_k$ is the eigenvector of $\mathbf{X}\mathbf{X}'$ corresponding to λ_k . Then the i th column of $\mathbf{Z} = \mathbf{W}'\mathbf{X}$ are the p principal component scores of the i th observation, $(x_{i1}, \dots, x_{ip})'$, on regressor variables. Put $\mathbf{W} = (\mathbf{W}_1, \mathbf{W}_2)$ with $\mathbf{W}_1 = (\mathbf{w}_1, \dots, \mathbf{w}_q)$ and $\mathbf{W}_2 = (\mathbf{w}_{q+1}, \dots, \mathbf{w}_p)$, and accordingly $\mathbf{Z} = (\mathbf{Z}_1', \mathbf{Z}_2')'$ with $\mathbf{Z}_1 = \mathbf{W}_1'\mathbf{X}$, and $\mathbf{Z}_2 = \mathbf{W}_2'\mathbf{X}$, corresponding to the first q and the remaining $p-q$ principal scores. Note that $\mathbf{X} = \mathbf{W}\mathbf{Z}$. The least squares estimate $\hat{\boldsymbol{\beta}}$ can be reexpressed as

$$\hat{\boldsymbol{\beta}} = (\mathbf{W}_1, \mathbf{W}_2) \text{diag}(\lambda_1^{-1}, \dots, \lambda_p^{-1}) (\mathbf{Z}_1', \mathbf{Z}_2')' \mathbf{y} \quad (11)$$

Retaining only the terms corresponding to the first q principal components leads to the principal component regression estimate of $\boldsymbol{\beta}$:

$$\tilde{\boldsymbol{\beta}} = \mathbf{W}_1 \text{diag}(\lambda_1^{-1}, \dots, \lambda_q^{-1}) \mathbf{Z}_1' \mathbf{y} \quad (12)$$

Applications of principal components in regression are not limited to linear models. For example, Barker and Brown^[25] investigated the application of principal components in logistic regression.

PRINCIPAL COMPONENTS BASED ON CORRELATION MATRIX

We have derived principal components based on the covariance matrix $\Sigma = (\sigma_{ij})$ or its estimator $\mathbf{S} = (s_{ij})$. When

some of the variables have relatively large variance as compared to others, often due to the use of different units of measurements of the variables, these variables may dominate the leading principal components, which is viewed as undesirable in many applications. An alternative approach is to use principal components of the standardized variables, obtained based on correlation matrix Ω with the (i, j) th elements $\omega_{ij} = \sigma_{ij}/(\sigma_{ii}\sigma_{jj})$, or its estimate $\hat{\Omega}$ with $\hat{\omega}_{ij} = s_{ij}/(s_{ii}s_{jj})$. Unlike principal components based on covariance matrix, the principal components derived this way are robust against large variances. It should be pointed out, however, that principal components derived from correlation matrix have no straightforward relation to those obtained from covariance matrix.

Recently, Boik^[26] developed distributional theory on correlation-based principal components. These theoretical developments allow us to test hypotheses regarding the correlation-based principal components.

INTERPRETATION OF PRINCIPAL COMPONENTS: A BIOMARKER EXAMPLE

In real studies, the original variables $\mathbf{x}$ are well designed and each has a clear practical meaning. However, the principal components, each being a linear combination of $\mathbf{x}$, may not be so, and interpretation of these components may be difficult and has to be carefully investigated. Here we present an example of biomarkers whose principal components turn out to be meaningfully interpreted.

Advanced laboratory procedures have provided a large collection of biomarkers of oxidative stress and antioxidant status. Individually, these markers contribute a broad spectrum of information on the oxidative stress status of an individual. However, because of their extensive number, limited knowledge is available on how to examine these biomarkers in combination, their relative significance on the oxidation/antioxidants process, and their interrelation. Considering measurements of prooxidant/antioxidant markers made in the same individuals, and by similar techniques, the likelihood of multicollinearity and multiple comparisons is of real concern when trying to clarify this intriguing process. These concerns need to be addressed when evaluating such biomarkers.

We utilize principal component analysis with 10 biomarkers measured in 925 men and women from a random sample of participants from Buffalo, New York, to obtain biologically meaningful linear combinations of these biomarkers related to oxidative stress and antioxidant status. We examine the first three principal components (based on sample covariance matrix), which explain 75% of the total variance of the 10 biomarkers. The three components are presented in Table 1, with coefficients that are 0 or close to 0 left blank.

Careful investigation reveals that these three components represent different aspects of the oxidative stress process.

Table 1 Principal Components of Oxidative Stress and Antioxidant Status Biomarkers

	Carotenoids	Oxidation	Antioxidants
β -Cryptoxanthin	.902		
β -Carotene	.676		
Lutein/Zeaxanthin	.649		
Lycopene	.462		
TBARS		.798	
Oxidized LDL		.701	-.206
Glucose		.666	
Triglycerides		.572	.309
Uric acid			.639
TEAC			.609

The first component consists of different types of vitamin A or carotenoids, including β -carotene, β -cryptoxanthin, lutein/zeaxanthin, and lycopene, giving the interpretation and name “carotenoids” to the component. The second factor reflects the effect of biomarkers related to oxidative stress, such as TBARS, oxidized LDL, glucose, vitamin E (part of the LDL molecule) and triglycerides, hence named “oxidation.” The third component, named “antioxidants” is characterized by antioxidant markers such as TEAC, uric acid, and triglycerides and negatively correlated with oxidized LDL.

In this application, 10 biomarkers are reduced to three leading principal components that retain most of the information. Each component has a sound scientific interpretation that requires deep understanding of subject matter.

CONCLUSION

Principal component analysis, as an exploratory multivariate data analysis technique, has been used in many other data analyses such as discriminant analysis, cluster analysis, and canonical correlation analysis, in addition to those described in the article.

Principal components are obtained through a series of variance maximization of the linear functions of the original variables. These optimal linear functions are, not surprisingly, different from the well-known Fisher's^[27] linear discriminant function because the objective functions are different; the latter is to minimize misclassification errors.

In principal component analysis, the first several leading principal components, which account for most of the variance, often play a far more important role than the others. There have been studies, however, that used the last few principal components to detect outliers and influential observations. (See Chapter 10 of Ref. [8] and the references therein.) When the first several principal components are used for analysis, it should be emphasized again that an

adequate number of components has to be selected to produce statistically reliable analysis results. If two or three components are used for data visualization (or pattern recognition), the information accounted for by these few components has to be evaluated.

REFERENCES

1. Pearson, K. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **1901**, 2, 559–572.
2. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, 24, 417–444.
3. Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*; Wiley: New York, 1958.
4. Muirhead, R.J. *Aspects of Multivariate Statistical Theory*; Wiley: New York, 1982.
5. Krzanowski, W.J. *Principles of Multivariate Analysis: A User's Perspective*, 2nd Ed.; Oxford University Press: Oxford, 2000.
6. Duntelman, G.H. *Principal Component Analysis*; Sage: Beverly Hills, CA, 1989.
7. Jackson, J.E. *A User's Guide to Principal Components*; Wiley: New York, 1991.
8. Jolliffe, I.T. *Principal Component Analysis*, 2nd Ed.; Springer: New York, 2002.
9. Cheung, V.G.; Morley, M.; Aguilar, F.; Massimi, A.; Kucherlapati, R.; Childs, G. Making and reading microarrays. *Nat. Genet. Suppl.* **1999**, 21, 15–19.
10. Brown, P.O.; Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nat. Genet. Suppl.* **1999**, 21, 33–37.
11. Yeung, K.Y.; Ruzzo, W.L. Principal component analysis for clustering gene expression data. *Bioinformatics* **2001**, 17, 763–774.
12. Mertens, J.A. Microarrays, pattern recognition and exploratory analysis. *Stat. Med.* **2003**, 22, 1879–1899.
13. Liu, A.; Zhang, Y.; Gehan, E.; Clarke, R. Block principal component analysis with application to gene microarray data classification. *Stat. Med.* **2002**, 21, 3465–3474.
14. Chipman, H.; Hastie, T.J.; Tibshirani, R. Clustering microarray data. In *Statistical Analysis of Gene Expression Microarray Data*; Speed, T., Ed.; CRC: New York, 2003, 159–200.
15. Parmigiani, G.; Garrett, E.S.; Irizarry, G.A.; Zeger, S.L. The analysis of gene expression data: An overview of methods and software. In *The Analysis of Gene Expression Data*; Parmigiani, G.; Garrett, E.S.; Irizarry, G.A., Zeger, S.L., Eds.; Springer: New York, 2003, 1–45.
16. Rao, C.R. *Linear Statistical Inference and its Applications*, 2nd Ed.; Wiley: New York, 1973.
17. Anderson, T.W. Asymptotic theory for principal component analysis. *Ann. Math. Stat.* **1963**, 34, 122–148.
18. James, A.T. The distribution of the latent roots of the covariance matrix. *Ann. Math. Stat.* **1960**, 31, 151–158.
19. Anderson, G.A. An asymptotic expansion for the distribution of the latent roots of the estimated covariance matrix. *Ann. Math. Stat.* **1965**, 36, 1153–1173.
20. Krzanowski, W.J. Cross-validation in principal component analysis. *Biometrics* **1987**, 43, 575–584.

21. Gu, Z.; Lee, R.Y.; Skaar, T.C.; Bouker, K.B.; Welch, J.N.; Lu, J.; Liu, A.; Zhu, Y.; Davis, N.; Leonessa, F.; Brunner, N.; Wang, Y.; Clarke, R. Association of interferon regulatory factor-1, nucleophosmin, nuclear factor-kB, and cyclic AMP response element binding with acquired resistance to faslodex (ICI 182,780). *Cancer Res.* **2002**, *62*, 3428–3437.
22. Massy, W.F. Principal components regression in exploratory statistical research. *J. Am. Stat. Assoc.* **1965**, *60*, 234–266.
23. Kendall, M.G. *A Course in Multivariate Analysis*; Griffin: London, 1957.
24. Hotelling, H. The relations of the newer multivariate statistical methods to factor analysis. *Br. J. Stat. Psychol.* **1957**, *10*, 69–79.
25. Barker, L.; Brown, C. Logistic regression when binary predictor variables are highly correlated. *Stat. Med.* **2001**, *20*, 1431–1442.
26. Boik, R.J. Principal components model for correlation matrices. *Biometrika* **2003**, *90*, 679–701.
27. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179–188.

Process Validation

James S. Bergum

Bristol-Myers Squibb Company, New Brunswick, New Jersey, U.S.A.

Merlin L. Utter

Pfizer Inc., Pearl River, New York, U.S.A.

Abstract

This entry focuses on the statistical techniques that aid in the validation of solid oral dosage forms, i.e., tablets and capsules. A general discussion of process validation is also provided.

INTRODUCTION

The Food and Drug Administration (FDA) defines validation in their 1987 publication *Guideline on General Principles of Process Validation*^[1] as “establishing documented evidence which provides a high degree of assurance that a specific process will consistently produce a product meeting its predetermined specifications and quality characteristics.” Because this definition contains the phrases “high degree of assurance” and “consistently produce,” which have statistical implications, statisticians have been involved in developing acceptance limits and sampling plans as well as analyzing validation data for process validation programs.

Although process validation is applied to all product forms, this entry will focus on the statistical techniques that aid in the validation of solid oral dosage forms, i.e., tablets and capsules. A general discussion of process validation will also be provided. This entry should be of most interest to statisticians who want an overview of the concepts of process validation and a discussion of those statistical techniques that are most relevant. By including a number of examples, this article is written so that those who conduct validation studies can obtain appropriate acceptance criteria for practical use.

HISTORY AND OVERVIEW

Validation started in the early 1970s with assay verification. During the mid-1970s, most of the attention given to validation dealt with sterile processes because of the critical nature of injectable products. In the early 1980s, the FDA turned its attention to such nonsterile processes as solid dosage, semisolid dosage, liquids, suspensions, and aerosols. The draft process validation guideline was made available for comment in March 1983 and was issued in May 1987. The guideline was written to help industry comply with the 1978 revision of the “Current Good

Manufacturing Practices (cGMP) Regulations for Finished Pharmaceuticals” and the “Good Manufacturing Practice Regulations for Medical Devices.” This was the first time that the word validation had appeared in the regulations. Because following the cGMP is a legal requirement, failure to comply would constitute a violation of the Food, Drug, and Cosmetic Act.

Many authors have written about process validation.^[2–16] This introduction is a compilation from these papers. Process validation is performed on new drug products (both prescription and over-the-counter), active pharmaceutical ingredients, products involved in technology transfer (e.g., to a new manufacturing site), marketed products that have not been validated, and previously validated products that incurred a major change (e.g., in a processing step, a new raw material source for the active, manufacturing equipment, or process parameters).

Elements of the validation concept should be incorporated at each stage of the product and process development continuum. Identification of the critical process variables and studies aimed at establishing optimal ranges for these variables should be completed during the prevalidation phases. Statistics can play a major role in the prevalidation phases of assay validation, in the setting of specifications, and in formula screening/optimization studies. Some of the critical variables that need to be studied are: particle size of the drug substance, bulk density of the active drug/excipients, blender load, binder concentration, amount and spray rate of granulating fluid, blender speed and blending times, moisture content of the dried granulation, particle size of the lubricated granulation, and lubricant blending times. In fact, a number of major steps are involved in the development of stable, bioavailable, manufacturable, and cost-effective solid oral dosage forms. These include the physicochemical characterization of the drug substance, formulation development of commercial dosage forms, development and validation of analytical test methods, specification setting, scale-up and process optimization, development of stability data on commercial formulations packaged in the container/closure

systems to be marketed, and process validation. During the validation phase, sampling plans as well as acceptance criteria are needed. After the data are collected, a statistical review may be needed to evaluate the data against the preset acceptance criteria.

TYPES OF VALIDATION

There are three different types of process validation: retrospective, concurrent, and prospective. Retrospective validation covers those situations where a product is already marketed without a documented process validation program. Retrospective validation was not intended to be a method of validating new processes but rather a safety net that could be used to validate products already on the market at the time the validation requirements were put into the regulations. Retrospective validation has been used when there have been no changes in formulation, procedure, or equipment, and when the efficiency of the process can be judged from in-process and finished-product results. In retrospective validation, it is not sufficient to examine only pass/fail records on previously manufactured lots; a statistical analysis of quantitative data of product attributes is necessary to predict process capability. Retrospective process validation involves using historical data to provide documented evidence that a system does what it purports to do. Because most companies have completed their application of retrospective validation to existing products, it will not be discussed in this entry. However, a discussion of some analysis techniques that might be used when performing a retrospective validation can be found in Kohberger.^[17]

Prospective process validation involves proving that the process does what it purports to do based on a preplanned protocol and before the process is put to commercial use. It involves the collection and evaluation of data on at least three consecutive successfully manufactured batches matching routine production in all aspects of production and control. The statistical aspects of prospective validation will be discussed in this entry. The protocol should state the process steps, identify the controls to be used, specify the variables to be monitored, state what samples are to be taken both for testing and as contingency, specify both the product performance characteristics/attributes to be evaluated and their acceptance criteria, and refer to the test methods to be used. Historical data from formulation development and scale-up studies may help to set these criteria.

Concurrent validation is similar to prospective validation except that batches are validated and released one batch at a time. Concurrent validation is only accepted in limited situations, e.g., for orphan drugs where it is deemed an economic hardship to produce three batches at one time when all cannot be sold. The validation procedures and acceptance criteria are similar to what would be carried out for a prospective validation.

SAMPLING

A manufacturing process may involve drying and granulation steps as well as intermediate and final mixing steps. Once a blend has been mixed, it may be transported to another location for screening or tableting (or encapsulation). Sampling can be performed at any of these steps in the manufacturing process. Samples can be taken when the blend is in a mixer, while being discharged from the mixer, when it is in a transport container (e.g., drum), throughout tablet compression or encapsulation, and after film coating (if appropriate). Sampling plans need to be developed at each of these stages.

There are two sampling plans that are generally used when testing blends or final product. In the first plan (Sampling Plan 1), a single test result is obtained from each location sampled. For example, in a blending step, a single test result would be obtained from each of a number of different locations within the blender. In a drum, a single test result might be obtained from the different locations within the drum or from each of a number of different drums. For final tablets, a single tablet may be tested from various time points throughout the tableting run. In the second plan (Sampling Plan 2), more than one test result is obtained from each of the sampled locations. For example, during the tableting operation, if a cup is placed under the tablet press at specific time points during the tableting run, several of the tablets from each cup sample would be tested for content uniformity. Sampling Plan 2 allows for estimation of between-location and within-location variability.

It is assumed for the remainder of this entry that the same number of units is tested from each of the sampled locations; that is, it is a balanced sampling plan. Regardless of what sampling plan is used to determine testing, multiple units are normally collected at each of the sample locations during validation to serve as contingency samples for possible later testing.

Powder Blend Sampling

Based on the interpretation of the Wolin court decision (*United States vs. Barr Laboratories*), the allowable size of sample taken from powder blends has been set at no more than three times the dosage unit weight. A perplexing problem facing manufacturers of oral solid dosage forms today is the difficulty in applying this unit dose sampling to blend uniformity validation because of the current limitations in sampling technologies. An excellent discussion of blend sampling is given in the PDA Technical Report on Blend Uniformity.^[18] Much of the following discussion is taken from that paper.

There is a great deal of frustration among oral solid dosage form manufacturers caused by unit dose sampling of blends. Companies have obtained very uniform results when testing the finished dosage form (i.e., tablets or capsules) while obtaining highly variable results when

attempting to comply with the current FDA position on blend uniformity sampling.

It is generally recognized that a thief is far from an ideal sampling device because of a propensity to provide non-representative samples, i.e., the sample has significantly different physical and chemical properties from the powder blend from which it was withdrawn. Although simple in concept, demonstrating blend uniformity is complicated by this potential for sampling error. The current technology does not yet provide a method for consistently obtaining small representative samples from large static powder beds. Using X-ray fluorescence and near-infrared spectroscopy methods to measure blend uniformity, it is hoped that these problems may soon be overcome.

As stated in the PDA Technical Report,^[18] sampling error can be influenced by: (i) the design of the thief; (ii) the sampling technique; and/or (iii) the physical and chemical properties of the formulation. The physical design of the thief can affect sampling error because the overall geometry of the thief can influence the sample that is collected. Surface material can fall down the side slit of a longitudinal thief as it is inserted into a powder bed. Sampling technique can also have an impact on sampling error. As the thief is inserted into a static powder, it will distort the bed by carrying material from the upper layers of the mixture downward toward the lower layers. The angle at which the thief is inserted into the powder bed can also influence sampling error. Another factor that can affect sampling error is the physical and chemical properties of the formulation. The force necessary to insert a long thief into a deep powder bed can be appreciable. This force, depending on the physical properties of the formulation, can lead to compaction, particle attrition, and further distortion of the bed. Ideally, the thief should be constructed from materials that do not preferentially attract the individual components of the formulation. In general, the potential for sampling error increases as the size of the sample and/or the concentration of drug in the formulation decreases. Samples obtained using thief probes can be subject to significant errors.

Finished Product Sampling

Two of the most common tests for finished product that have acceptance criteria are content uniformity and dissolution. The United States Pharmacopeia (USP 32) requirements for content uniformity for solid oral dosage forms (e.g., tablets and capsules) as well as for dissolution are summarized in [Tables 1,2](#).

When collecting samples to evaluate these tests, it is important to maintain the location identity of all samples taken and to maintain this identity throughout the testing regimen. Validation is the one time when the exact location in the batch is known for each of the individual dosage units tested. By showing that each of the sample locations tested provides acceptable results, a justification is developed for

Table 1 USP 32 Content Uniformity Test Requirements

Stage	Number tested	Pass stage if:
S ₁	10	$ M - \bar{X} + 2.4s \leq 15.0$, where M is defined below, $\bar{X}$ is the sample mean, and s is the sample standard deviation
S ₂	20	i. $ M - \bar{X} + 2.0s \leq 15.0$ using all 30 results (S ₁ + S ₂) ii. No dosage unit is outside the maximum allowed range of 0.75M to 1.25M.

M is defined as follows:

If T is less than or equal to 101.5%, and

- If $\bar{X}$ is less than 98.5%, then $M = 98.5\%$.
- If $\bar{X}$ is between 98.5 and 101.5%, then $M = \bar{X}$.
- If $\bar{X}$ is greater than 101.5%, then $M = 101.5\%$.

If T is greater than 101.5%, and

- If $\bar{X}$ is less than 98.5%, then $M = 98.5\%$.
- If $\bar{X}$ is between 98.5 and T, then $M = \bar{X}$.
- If $\bar{X}$ is greater than T, then $M = T$.

Where T is the Target content per dosage unit at the time of manufacture, expressed as a percentage of the label claim. Unless otherwise specified in the individual monograph, T is 100.0%.

All measurements of dosage units and criteria values are in percentage label claim.

Table 2 USP 32 Dissolution Test Requirements

Stage	Number tested	Pass stage if:
S ₁	6	Each unit is not less than $Q + 5\%$
S ₂	6	Average of 12 units (S ₁ + S ₂) is equal to or greater than Q and no unit is less than $Q - 15\%$
S ₃	12	Average of 24 units (S ₁ + S ₂ + S ₃) is equal to or greater than Q , not more than 2 units are less than $Q - 15\%$, and no unit is less than $Q - 25\%$

the later combining of tablets or capsules into a QC composite sample for the release testing of future batches. If a two-sided press is used for tableting, the identity of the side of the press from which the samples were taken should also be maintained.

It is recommended that individual dosage units be tested from as many different sample locations as possible. The number of units tested could even be tied to run length, with more units tested when the run length goes across multiple shifts. A discussion of the effect of sample size on one of the methods discussed, the Bergum approach, is provided in the "Effect of Sample Size" section. Because Sampling Plan 2 allows for the estimation of both between and within-location variability, this plan is generally recommended when testing individual dosage units for content uniformity. For dissolution, one might choose either Sampling Plan 1 or Sampling Plan 2, depending upon how many total units are tested.

STATISTICAL TECHNIQUES/APPROACHES

Since the start of validation in the late 1970s, there has been little published material on the statistical aspects of conducting a successful process validation. What follows are some of the statistical techniques that have been either suggested in the literature or used in practice when conducting validation studies. Their use in the development of validation criteria will be discussed in the “Comparison of Acceptance Criteria” section. The Bergum approach is discussed in two papers by Bergum.^[19,20] A discussion of the other techniques can be found in Hahn and Meeker.^[21] Other techniques that have been applied to validation data but are not discussed in detail in this article are analysis of variance (ANOVA) and process capability analysis. In the following subsections, let $\bar{X}$ and s denote the mean and standard deviation of a sample of size n and let t and F be the critical values for the t and F distributions with their associated degrees of freedom and confidence levels. Let MSB be the between-location mean square from the one-way ANOVA.

Tolerance Interval

A tolerance interval is an interval that contains at least a specified proportion P of the population with a specified degree of confidence, $100(1 - \alpha)\%$. This allows a manufacturer to specify that at a certain confidence level at least a fraction of size P of the total items manufactured will lie within a given interval. The form of the equation is: $\bar{X} \pm ks$ where k = the tabled tolerance factor and is a function of $1 - \alpha$, P , n and whether it is a one-sided or two-sided interval.

Prediction Interval

A number of prediction intervals can also be generated. A “two-sided prediction interval for a single future observation” may be of interest. This is an interval that will contain a future observation from a population with a specified degree of confidence, $100(1 - \alpha)\%$. The form of this equation is: $\bar{X} \pm ks$ where $k = t_{1-\alpha/2, n-1} \sqrt{1 + 1/n}$.

Another type of prediction interval that might be of interest is a one-sided upper prediction interval to contain the standard deviation of a future sample of m observations, again with a specified degree of confidence, $100(1 - \alpha)\%$. This is called the standard deviation prediction interval (SDPI). The form of this equation is: $s\sqrt{F_{1-\alpha, m-1, n-1}}$.

Confidence Interval

Confidence intervals can be generated for any population parameter. Specifically, a two-sided confidence interval about the mean is an interval that contains the true unknown mean with a specified degree of confidence,

$100(1 - \alpha)\%$. The form of this equation, which depends on the sampling plan, is as follows:

- Sampling Plan 1 $\bar{X} \pm ks$ where $k = t_{1-\alpha/2, n-1} / \sqrt{n}$
- Sampling Plan 2 $\bar{X} \pm k\sqrt{\text{MSB}}$

$$\text{where } k = \frac{t_{1-\alpha/2, \# \text{ locations}-1}}{\sqrt{(\# \text{ locations})(\# \text{ per locations})}}$$

Note that for any stated confidence level, the confidence interval about the mean is the narrowest interval, the prediction interval for a single future observation is wider, and the tolerance interval (to contain 95% of the population) is the widest.

Variance Components

Variance components analysis has been used in a number of applications within the pharmaceutical industry. The power of this statistical tool is the separation or partitioning of variability into nested components. The approach requires using Sampling Plan 2 so that the between-location and within-location variance components can be estimated. These estimates can be calculated using one-way ANOVA. The within-location variance is estimated by the mean square error, whereas the between-location variance is estimated by subtracting the mean square error from the mean square between locations and then dividing by the number of observations within each location. When applied to the blending operation, the method allows us to determine the between-location variance, which quantifies the distribution of active throughout the blend, and the within-location variance, which in turn is composed of sampling error, assay variance, and a component related to the degree of mixing on the “micro” scale. The total variance in the container or mixer is the sum of the two variance components. Similarly, one may also determine these components in the product. Here the within-location variance will again consist of the assay variance, the sampling error, and the “micro” mixing component, in addition to the weight variation. The between-location component is that variance associated with macro changes in the blend environment. It is this component that reflects the overall uniformity of the blend and is minimized when optimum blender operation is achieved.

Simulation

Monte Carlo simulation can be used to estimate the probability of passing multiple-stage tests such as content uniformity and dissolution. This technique is performed by generating computer-simulated data from a specific probability distribution (e.g., normal) and then using these generated sample data as if they were actual observations. The multiple-stage test is then applied to the data. This process can be repeated many times to evaluate various

test properties (e.g., determining the probability of passing the multiple stage test for specific values of the population mean and standard deviation of a normal distribution).

Bergum Approach

Bergum^[19,20] published a method for constructing acceptance limits that directly relates the acceptance criteria to multiple stage tests such as the USP 32 content uniformity and dissolution tests. These acceptance limits are defined to provide, with a stated confidence level $(1 - \alpha)100$, a stated probability (P) of passing the test. For example, one can make the statement that, with 95% confidence, there is at least a 95% probability of passing the USP 32 test. Both the USP 32 content uniformity and the USP 32 dissolution tests have been evaluated. In each case, the required limits are provided in “acceptance tables,” which are computer generated. These tables change with the confidence level $(1 - \alpha)$, the probability bound or coverage (P), and the sample size (n) for content uniformity as well as the Q value for dissolution. Confidence levels as well as values for P are typically 50%, 90%, or 95%. The PDA Technical Report^[18] suggests the use of a 90% confidence level to provide 95% coverage. The FDA prefers a 95% confidence level. A 50% confidence level can be considered a “best estimate” of the coverage. A SAS program has been written to construct these acceptance limit tables for the USP 32 content uniformity and dissolution tests for both Sampling Plans 1 and 2. The computer program can be obtained by contacting James Bergum at Bristol-Myers Squibb.

Statistical Basis

The Bergum approach uses the fact that these are multiple-stage tests. A multiple-stage test is a test with several stages, where each stage has requirements for passing the test. As can be seen in Tables 1–2, the USP 32 content uniformity and dissolution tests are multiple-stage

tests with multiple criteria at each stage. The lower bound, LBOUND (also called P), for the probability of passing the USP 32 content uniformity and dissolution tests uses the following relationship:

$$\text{Prob}(\text{passing USP 32 test}) \geq \max\{\text{Prob}(\text{passing } i\text{th stage})\}$$

Where $i = 1$ to S (S = number of stages in USP 32 test).

One requirement for this inequality to hold for a multiple-stage test is that failure of the overall test at any stage also results in failure of the overall test at any subsequent stage.

Assume that the test results follow a normal distribution with mean μ and standard deviation σ . Sigma (σ) is the standard deviation of a single observation. For a given value of μ and a given value of σ , LBOUND can be determined by calculating the probability of passing all of the requirements at each stage. Figs. 1 and 2 show the 95% contours for the calculated bound LBOUND. If μ and σ are on the 95% LBOUND contour, then at least 95% of the samples tested using the USP 32 test would pass the test. Tables 3 and 4 compare these lower bounds to the “actual” probability of passing these tests for selected values of μ and σ . The “actual” probabilities were determined by simulating the USP tests. Note that these probabilities are for passing the overall test. The probability of passing stage 1 can be much smaller, especially for dissolution (e.g., there is less than a 1.6% chance to pass stage 1 dissolution if the population mean is $Q + 5$ with a sigma of 7).

The LBOUND can be used to develop acceptance criteria by constructing a simultaneous confidence interval for μ and σ from the data. If a 90% confidence interval were constructed for μ and σ and the entire interval were below the 95% LBOUND, then, with 90% confidence, at least 95% of the samples tested would pass the USP 32 test. For Sampling Plan 1, the sample mean and sample standard deviation estimate the population parameters μ and σ . A simultaneous confidence interval for μ and σ is given in Lindgren.²² Because the variance of a single observation

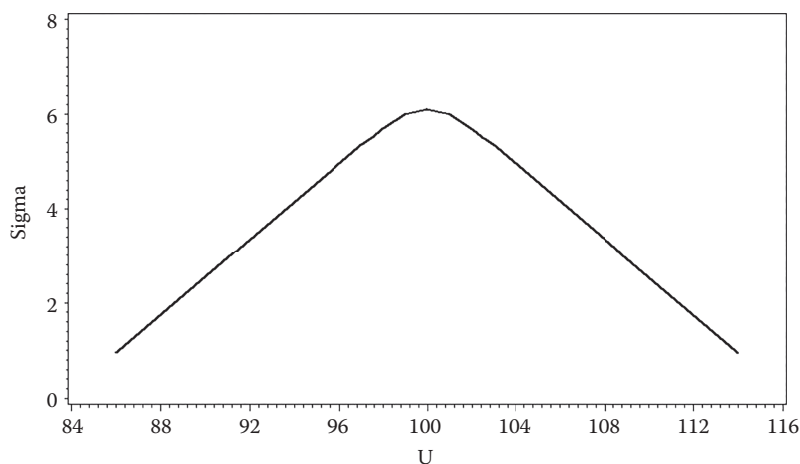


Fig. 1 Ninety-five percent contour plot for probability of passing USP 32 content uniformity test

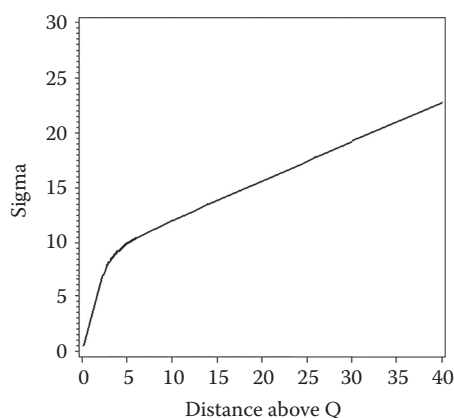


Fig. 2 Ninety-five percent contour plot for probability of passing USP 32 dissolution test

using Sampling Plan 2 is the sum of the between-location and within-location variances, σ (i.e., the standard deviation of a single observation) is estimated by calculating the square root of the sum of the between-location and within-location variance components. A confidence interval for σ is given by Graybill and Wang.^[23] The simultaneous confidence interval for μ and σ is constructed by using a Bonferroni adjustment on the two individual confidence intervals for μ and σ . Once the confidence interval is constructed, it must fall completely below the LBOUND specified. An acceptance limit table can be generated by finding the largest sample standard deviation for a fixed sample mean such that the resulting confidence interval remains below the prespecified LBOUND.

Effect of Sample Size

Through the use of operating characteristic (OC) curves, the effect of sample size on the ability to pass the Bergum approach for content uniformity (CU) can be evaluated. The OC curves provide estimates of the probability of passing the Bergum approach over a number of different population mean and standard deviation values. Figs. 3 and 4 provide OC curves for specific sample sizes using Sampling Plans 1 and 2, respectively. For these plots, a mean of 100% was assumed. A confidence level of 90% to obtain 95% coverage was also used. The estimated probability of passing the USP 32 content uniformity test was included for comparison.

Fig. 3 provides the OC curves using Sampling Plan 1 for sample sizes of 10, 30, 60, 100 and 2000. As expected, the probability of passing the acceptance limit table increases as the sample size increases. For example, if n is 30, the probability of passing the acceptance limit table for tablets when σ is 4.0% is approximately 75%. To increase the probability of passing the Bergum approach with this type of true quality, a larger sample size would be needed. The probability of passing the USP test is also provided to show how much more discriminating the Bergum Tables can be.

Table 3 Simulated (SIM) vs Lower Bound (LB) Probabilities of Passing CU Test

Population mean μ	Method	Population standard deviation σ		
		2	4	6
88	LB	0.800	0.024	0.004
	SIM	0.852	0.028	0.004
92	LB	1.000	0.674	0.072
	SIM	1.000	0.739	0.096
96	LB	1.000	1.000	0.610
	SIM	1.000	1.000	0.695
100	LB	1.000	1.000	0.964
	SIM	1.000	1.000	0.972
104	LB	1.000	1.000	0.610
	SIM	1.000	1.000	0.692
108	LB	1.000	0.674	0.072
	SIM	1.000	0.743	0.095
112	LB	0.800	0.024	0.004
	SIM	0.858	0.028	0.003

Table 4 Simulated (SIM) vs Lower Bound (LB) Probabilities of Passing Dissolution Test

μ Expressed as distance above Q	Method	Population standard deviation Σ		
		3	5	7
1	LB	0.949	0.836	0.753
	SIM	0.966	0.889	0.823
3	LB	1.000	0.998	0.981
	SIM	1.000	0.999	0.988
5	LB	1.000	1.000	0.999
	SIM	1.000	1.000	0.999

For example, when the probability of passing the USP test gets down to 95% or below, the probability of passing the Bergum tables for these OC curves is essentially zero.

Fig. 4 provides the OC curves using Sampling Plan 2 for a sample size of 60 but with different numbers of locations sampled; the results are compared to the use of Sampling Plan 1 without any replication. It is assumed for this plot that half of the total variation is due to between-location variance and half is due to within-location variance (i.e., factor = 0.5). Note that the number of locations has a significant effect on the probability of passing the acceptance limit table. This effect would have been even larger if the percent of variation due to locations was assumed to have been something greater than one half. It is recommended that when using Sampling Plan 2, the number of locations used be as large as possible. For example, if a total of 60 tablets are sampled across the batch, it is better to sample 3 from each of 20 locations than to sample 20 from each of 3 locations.

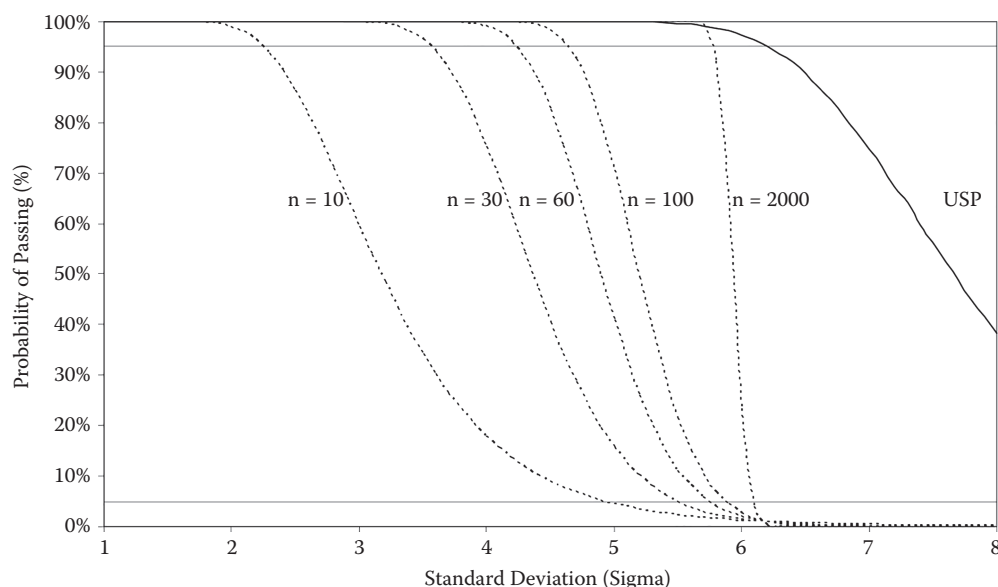


Fig. 3 Probability of meeting Bergum CU acceptance table for Sampling Plan 1, $\mu = 100.0\%$, 90% assurance/95% coverage, tablets

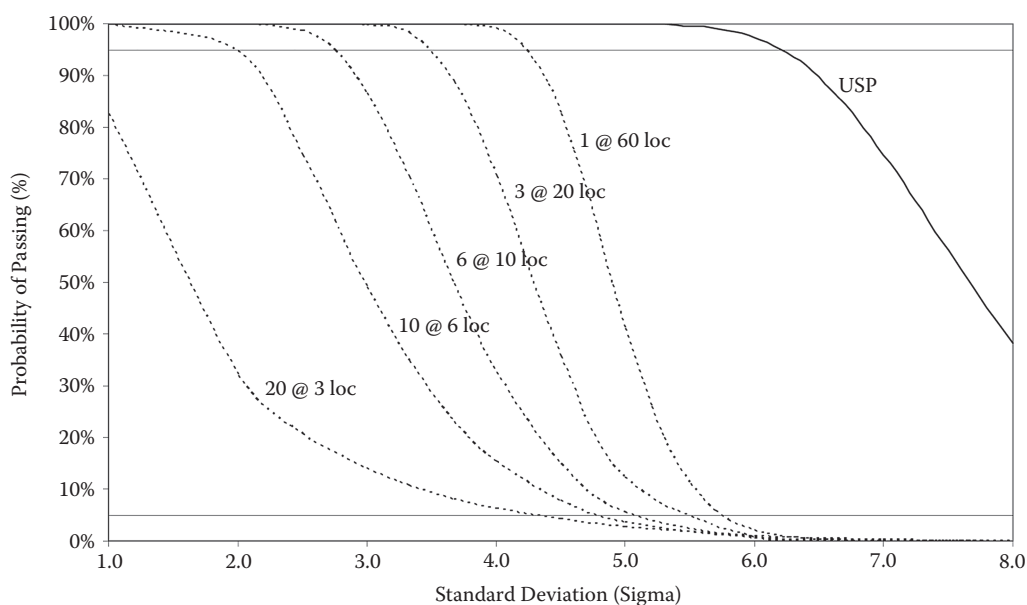


Fig. 4 Probability of meeting Bergum CU acceptance table for Sampling Plan 2, $\mu = 100.0\%$, 90% assurance/95% coverage, factor = 0.5, tablets

COMPARISON OF ACCEPTANCE CRITERIA

There are a number of tests that are performed during validation. In the blends, the primary interest is in showing that the blend is uniform in drug content. Uniformity can also be evaluated in the drums to ensure that segregation or demixing did not occur during transfer. The overall potency is generally not considered a critical variable in the blends because it is neither enhanced nor diminished by additional mixing. However, there may be a concern with potency loss during processing or storage between

processing steps, e.g., after emptying the blended powder into transports or as a result of tablet compression. In the final product, content uniformity and dissolution are of primary interest, and to a lesser extent, potency. Acceptance limits are generally developed for these tests. Other tests that are performed, such as particle size, bulk/tap density and flow, hardness, friability, and weight variation, may or may not have formal, statistically derived acceptance limits.

There are almost as many approaches to validation as there are companies performing validation. What follows

Table 5 Advantages and Disadvantages of Various Statistical Techniques for Powder Blends

Test	Advantages	Disadvantages
<i>Blend uniformity</i>		
1. FDA approach	Accepted by the FDA	Not statistically based Penalized for large n Adversely affected by constant loss of potency
2. SDPI	Rewarded for larger n Not affected by constant loss of potency Can be tied to part of Stage 1 USP 32 CU test	Difficult to apply with Sampling Plan 2 Not directly tied to full USP 32 CU test
3. Tolerance interval	Easy to calculate Rewarded for larger n	Not directly tied to full USP 32 CU test Difficult to apply for Sampling Plan 2 Factors can be hard to find for nonstandard coverage probabilities
4. Bergum approach	Rewarded for larger n Tied directly to USP 32 overall test Easy table look-up Provides high assurance of passing USP 32 test	Computer program required (but can be provided) Difficult to pass using Sampling Plan 2 with few locations
5. Holistic approach	Provides chance to recover from variable blender results	Substitution of variance components concept may be a hard sell Degrees of freedom can be significantly reduced

Table 6 Advantages and Disadvantages of Various Statistical Techniques for Finished Product (Tablets/Capsules)

Test	Advantages	Disadvantages
<i>CU/dissolution</i>		
1. Simulation	Directly tied to USP 32 Can be tied to either Stage 1 or full test Can handle non-symmetric potency limits	Not a function of n Does not provide high assurance level of passing USP 32 test (only point estimate)
2. Tolerance interval	Easy to calculate Rewarded for larger n	Not tied directly to USP 32 overall test Difficult to apply for Sampling Plan 2 Factors can be hard to find for nonstandard coverage probabilities
3. Bergum approach	Rewarded for larger n Tied directly to USP 32 overall test Easy table look-up Provides high assurance of passing USP 32 test	Does not directly address non-symmetrical potency limits Computer program required (but can be provided) Difficult to pass using Sampling Plan 2 with few locations
<i>Potency (composite assay)</i>		
1. Confidence interval	Provides strong statement that overall batch average potency is acceptable	Does not provide assurance that a given assay result will meet requirements
2. Tolerance/prediction interval	Provides strong statement that an individual assay result will meet assay requirements	Requires a large number of composite assays

is a discussion of some of the methods of statistical analysis along with their advantages and disadvantages. Two proposals, one from the FDA for blends and another by the Parenteral Drug Association (PDA), called a “Holistic” approach to validation, are also discussed.

The advantages and disadvantages of these methods are listed in [Tables 5](#) and [6](#) for powder blends and finished product (tablets/capsules), respectively. Note that what might be an advantage to one person can be a disadvantage to the next.

Powder Blends

Blend Uniformity

The Food and Drug Administration Approach The FDA has proposed the following acceptance criteria for blend uniformity testing.^[24]

- Each individual sample should meet compendial assay limits (e.g., 90.0–110.0%).
- The RSD should be no greater than 5%.
- A minimum of 10 samples should be tested.

The samples should include potential “dead spots.” The tighter RSD requirement is to allow for additional variation from possible demixing and weight or fill variation. It is stated^[24] that just meeting the USP 32 content uniformity criteria is not appropriate for blends. The blend criteria also should not be relaxed because of sampling difficulties. It is more appropriate to change the sampling procedure to ensure accurate results.

The advantages of these criteria are that they are easily understood and implemented and any firm that meets them would be highly confident of satisfactorily passing a GMP or preapproval inspection. However, these criteria have a number of disadvantages. A firm is penalized for taking more samples because the probability of finding an out-of-range sample increases accordingly. These criteria also assume that current sampling practices can always provide a consistent collection of unit dose samples representative of the powder blend.

Standard Deviation Prediction Interval Because uniformity is of primary interest in powder blend validation and because of a concern that a constant sampling error can occur, one approach is to base the criteria on variability only. The SDPI allows one to predict, from a sample of size n , and with a specified level of assurance, an upper bound on the standard deviation of a future sample of size m from the same population. This approach is recommended in the PDA paper on blend uniformity.^[18]

By setting the future sample size m to 10, which is the stage 1 sample size for the USP 32 content uniformity test, and by requiring that the upper bound on the standard deviation of a future sample of size 10 be less than 6.25% (i.e., 15/2.4), which is the USP 32 stage 1 upper limit on the standard deviation if $M - \bar{X} = 0$, the SDPI approach can be tied to the USP 32 content uniformity test. The SDPI equation in the “Statistical Techniques/Approaches” section can be rearranged to obtain the following equation: $S_{cr} = S_m / [F_{1-\alpha, m-1, n-1}]^{1/2}$ where n = size of current sample; s_{cr} = critical standard deviation; s_m = upper bound of a future sample of size m ; $1 - \alpha$ = confidence level (e.g., 0.90); F = critical F value.

S_{cr} becomes the maximum acceptable sample standard deviation to meet the acceptance criteria. If the sample

standard deviation, s_n , is less than s_{cr} , then we are guaranteed, with a minimum assurance of $100(1 - \alpha)\%$, that the upper prediction bound for a future sample of size 10 will not be greater than 6.25% of the target concentration.

Tolerance Interval Approach To use the tolerance interval as an acceptance criterion, the confidence level $100(1 - \alpha)\%$ and coverage level P need to be chosen. One approach is to assume that the blend samples are the same as the resulting final product from that blend. To tie the tolerance interval to the USP 32 content uniformity test, one choice might be to use a coverage of 96.7% because the USP 32 test was developed based on matching the operating characteristic operating curves to the previous USP content uniformity test for tablets which allowed 1 tablet out of 30 (96.7%) to be outside 85–115% of label claim. If the tolerance interval is completely contained within the 85–115% interval, these acceptance criteria would be met. Another choice of how to define the coverage P is discussed in the PDA technical report.^[18] Although it is difficult to find tolerance factors for nonstandard coverage levels in published tables, they can be generated using the interval statement in the SAS/QC[®] procedure CAPABILITY.^[25]

Tolerance intervals assume that sampling is performed using Sampling Plan 1. There is only one variance component used to estimate the variance of a single observation, i.e., the sample variance. The degrees of freedom used to determine the tolerance factor k is the degrees of freedom associated with the sample variance. However, if Sampling Plan 2 is used, there are two variance components used to estimate the variance of a single observation. Therefore the degrees of freedom must be approximated. This can be carried out using Satterthwaite' approximation.^[26]

Bergum Approach Because the USP 32 content uniformity test is applied only to the finished product, the application of the Bergum approach requires that the acceptance limits for the blend be tied to the limits for the dosage form. In addition, for any of the approaches, each result is generally expressed in percentage of label claim as a percentage of active in a theoretical tablet weight. An alternative to the SDPI approach, which is not dependent upon the mean, is to express each result as a percentage of the sample mean and to then apply the Bergum approach. This has the effect of removing the mean effect and just evaluating the variability. If this were carried out, then the acceptance limit would be the RSD associated with a sample mean of 100%.

Holistic Approach The PDA report^[18] proposes a “Holistic” approach to the validation, in which means and variances of the blend are compared to the means and variances of the final product. The validation is considered successful if all criteria are met for both the blend and the final dosage form. If the final product fails, the validation is unsuccessful regardless of the blend results. However,

there are situations where the final product is acceptable and the blend is unsuccessful but the true blend uniformity can be deemed acceptable. This is because the inconsistent results might be due to sampling error when sampling the blend. The blend may have good location-to-location variability, but, because of sampling errors, the within-location error causes the blend results to fail. One approach given by the PDA paper^[18] is to use “analysis by synthesis.” To employ this technique, Sampling Plan 2 must be used for both the blend and the final dosage form so that the between-location and within-location variance components can be estimated. These variance components, as well as the total variance, can be statistically tested using an *F* test to determine if there is a significant difference between the variances at the two stages. If the within-location variance component in the blend is significantly higher than that in the final product, then the within-location variance component for the final product is substituted for the within-location variance component of the blend, in an attempt to remove the effect of sampling error in the blend sample results. This reduced overall variance for the blend is compared to the acceptance criteria.

Average Potency

There may be a desire to assess possible potency loss between the different sample stages. There also may be an interest in assessing whether the average potency results are at the target potency. At the blender stage, the average potency can be determined either from taking the average of several potency assays or by using the average of the uniformity values if it is felt that there are no sampling issues associated with the smaller sample quantity and if the assay and content uniformity methods are the same. If it is not clear whether there will be sampling issues during validation, it is suggested, when possible, that a formulation study be conducted prior to the validation to determine whether the smaller sample quantity will provide consistent uniformity results and, if not, what sample quantity will produce consistent results. With this support in hand, the smallest sample quantity that will provide consistent results should be used for the validation. It is understood, however, that it may not always be possible to conduct such prevalidation studies.

If a comparison across stages is to be performed, it is recommended that all powder results be reported as percentage of label claim, and not as a percentage of theoretical, so as to provide a direct comparison of the average results to finished product. One must remember that the potency results obtained prior to any addition of lubricant must be adjusted down to account for the noninclusion of the lubricant at the time of sampling. To compare the average uniformity or potency results across stages, one can require that the averages at each stage be within some stated amount of each other or of target. Statistically based techniques such as ANOVA or confidence intervals using the variation of the data and a stated assurance level can also be used.

Finished Product

Content Uniformity/Dissolution

Bergum Approach This approach is recommended in the PDA paper^[18] for final product testing. To use the dissolution acceptance limit tables, the value of *Q* is required.

Simulation One approach is to assume that the sample mean and the standard deviation are the true population mean and standard deviation, to provide a “best estimate” of the true probability of passing. This has the advantage that it can provide estimates of the probability of passing at any stage and can handle the nonsymmetrical potency shelf life limits in the content uniformity test. The disadvantage is that it does not provide a bound on the probability with high assurance and is not a function of sample size. However, it can provide a good summary statistic of the content uniformity data.

Tolerance Interval See the “Tolerance Interval Approach” section for comments regarding the use of tolerance intervals as acceptance limits for content uniformity data. Tolerance intervals can also be used as acceptance limits for dissolution. Because the USP 32 dissolution test for stage 1 is that all six capsules be greater than $Q + 5$, the tolerance interval could be tied to the USP 32 test by requiring that the lower bound on the tolerance interval be greater than $Q + 5$. To obtain a 95% probability of passing at stage 1, the coverage *P* of the tolerance interval would need to be $(0.95)^{1/6}$, or 0.991. Using a tolerance interval based on stage 1 of the USP 32 test can be very restrictive.

Confidence Interval Confidence intervals are not recommended for evaluating content uniformity data. However, an approach that is less restrictive than tolerance intervals for evaluating dissolution data is to base the acceptance limits on meeting the second and third stage of the USP 32 dissolution test. Both the second and third stages require that the sample mean be less than *Q*. Therefore a lower one-sided confidence interval for the population mean could be used as an acceptance limit. The criterion is that the lower bound on the confidence interval must be greater than *Q*.

Potency

Potency can also be evaluated during validation. It is assumed that some number of composite assays is tested during validation. One criterion might be to generate a $100(1 - \alpha)\%$ confidence interval about the mean using all the potencies collected. This interval will contain the true batch potency with $100(1 - \alpha)\%$ confidence. This interval should be contained within the potency “in-house” or release limits. Enough potency results should be looked at to have sufficient power that this interval will be contained within the desired limits.

Meeting the foregoing criterion should not be interpreted to mean that an individual composite potency assay will meet the “in-house” limits with high assurance. If this is desired, a prediction interval for a single future observation or, better yet, a tolerance interval should be used. The validation specialist should be cautioned that additional composite assays might need to be tested to meet either one of these criteria with high confidence.

At a minimum, each of the composite assay results obtained should fall within the desired limits, either the potency shelf specifications or the potency “in-house” (or release) limits. The “in-house” limits are felt to be the more appropriate because these are the limits that ensure that the product will meet the shelf limits throughout expiry.

Other Validation Issues

Validation data should be plotted whenever possible. For example, content uniformity and dissolution can be plotted versus the sample locations. This allows for a visual check for trends. A criterion requiring “no trends of note” or that some specific trend rule be met (such as Nelson’s mean square successive difference trend test^[27]) might be included as part of the acceptance criteria. Some companies use more of a process-capability approach to determine consistency of test results across sample locations.

It is desirable that samples be sent to the laboratory for testing in a designed and ordered way, to be able to separate laboratory effects from process effects if it becomes necessary. For example, if four units were to be tested from each sample location, send one half of the units from each location to the laboratory on each of 2 days. In practice, the laboratory may resist doing this.

Weights of individual dosage units should be obtained for every unit tested for both content uniformity and dissolution at the time the units are tested. This may be useful information for later investigation if unacceptable test results are obtained.

For coated products, because this is the finished form, sampling and testing should also be conducted. However, the emphasis is usually on the cores, where the sample identity across the batch is known and can be evaluated. At the coated stage, the effect of the coating solution on dissolution is probably of most interest. Individual coating pans, either all of them or some portion of them, should be sampled and tested, with pan number identity maintained.

FUTURE DIRECTIONS

Recently, a new draft guidance for industry on process validation has been issued for comments.^[28] Though not yet approved, it reinforces the importance of statistics in the process validation process. This draft guidance incorporates the FDA’s initiative on “Pharmaceutical cGMPs for the 21 st century” and is aligned with ICH

guidance documents such as Q8, Q9, and Q10. It is based on the principle that process validation should take place over the lifecycle of the product and process, not just the first three batches produced, and it emphasizes the importance of building process knowledge and understanding, establishing a strategy for process control, and maintaining a state of control throughout the lifecycle. The basis of process understanding and process evaluation is the use of sound statistical methods and procedures such as are provided in this chapter.

EXAMPLES

The two examples given in this section demonstrate the application of some of the statistical techniques described in previous sections using both Sampling Plans 1 and 2. Example 1 uses Sampling Plan 1 and Example 2 uses Sampling Plan 2. In each example, samples are taken from the blend and from the final product. Samples from both the blend and final capsules are tested for content uniformity. The final capsules are also tested for dissolution. We assume that the USP 32 dissolution specification for this immediate release product has a Q of 85% at 30 minute. Suppose the blend samples are taken from a V blender. This type of blender looks like a V with a left and right side of the V. Samples are taken from the front and back of each side of the blender from the top, middle, and bottom of the granulation, for a total of 12 locations. Assume that the data are in percentage of label claim units. Although a 90% confidence level is used throughout the example, 95% is also a typical confidence level. For the Bergum approach, a 95% probability of passing is used throughout. All tolerance factors were calculated using the interval statement in the SAS/QC procedure CAPABILITY.^[25]

Example 1 (Sampling Plan 1)

Blend

Using Sampling Plan 1, a single content uniformity result is obtained from each of 12 locations in the V blender, with the following results:

Mean = 97.76; SD = 2.64; RSD (%) = 2.70.

Blend data display

Location		Side	
		Left	Right
Front	Top	100.76	92.53
	Middle	97.17	98.22
	Bottom	95.64	101.91
Back	Top	100.88	98.97
	Middle	97.93	96.30
	Bottom	95.63	97.13

For the tolerance interval approach, the 90% two-sided tolerance interval to capture 96.7% of the individual content uniformity results is $97.76 \pm 3.112(2.64) = (89.54, 105.98)$. Because the interval is completely contained within the 85–115% range, the criterion is met.

The s_{cr} based on the SDPI is $6.25/\sqrt{2.27} = 4.15\%$. Because the standard deviation for the example is 2.64%, which is less than s_{cr} , this sample meets the acceptance criterion.

To use the acceptance limits proposed by Bergum, an acceptance limit table is generated to give the upper bound on the sample RSD for various values of the sample mean. For this example, the table was constructed for content uniformity using a 90% confidence level with a lower bound (LBOUND) of 95%. A portion of the acceptance limit table is as follows:

^aTable entry of interest.

Mean (% claim)	RSD (%)
97.5	3.02
97.6	3.03
97.7	3.05 ^a
97.8	3.07

The sample mean for this example is 97.76%, so the upper limit for the sample RSD is 3.05% (marked with ^a). It is recommended that the means always be rounded to the more restrictive RSD limit so that the assurance level and lower bound specifications are still met. So in this case, 97.76% is rounded to 97.7%. Therefore because the sample RSD of 2.70% is less than the critical RSD of 3.05, the acceptance criterion is met. This means that, with 90% assurance, at least 95% of samples taken from the blender would pass the USP 32 content uniformity test.

Finished Product

Assume that during encapsulation, a sample was taken at each of 30 locations throughout the batch. One capsule from each location was tested for content uniformity and one for dissolution, with the following results:

Mean = 97.63; SD = 2.34;

RSD (%) = 2.40.

Data display

CU				
99.19	96.38	98.82	98.53	94.37
97.33	95.97	101.32	97.78	97.03
97.05	94.39	100.85	97.77	95.42
95.42	96.73	101.29	96.80	103.03
99.23	97.28	97.52	100.26	95.27
97.36	91.77	98.23	98.07	98.35

Mean = 93.84; SD = 3.47; RSD (%) = 3.69.

Data display

Dissolution					
93.78	94.65	87.83	96.81	92.57	87.68
92.17	88.01	96.59	101.46	93.75	99.44
95.27	92.47	98.46	96.34	93.52	90.73
92.75	94.53	88.72	89.58	97.37	96.41
90.93	96.11	93.41	96.60	94.45	92.82

A 90% tolerance interval to capture 96.7% of the individual content uniformity test results is $97.63 \pm 2.625 \times 2.34 = (91.49, 103.77)$. Because this interval is contained within the 85–115% interval, the criterion is met.

Using a criterion based on passing stage 1 of the USP 32 dissolution test, a lower one-sided 90% tolerance interval to capture 99.1% of the individual dissolution values is $93.84 - 2.930(3.47) = 83.67$. Using this criterion, dissolution would fail because the lower bound is less than $Q + 5$, which is 90.

Using a criterion based on stages 2 and 3 of the USP 32 dissolution test, a lower one-sided 90% confidence interval for the population mean is $93.84 - 1.311(3.47)/\sqrt{30} = 93.01$. Because the lower bound on the confidence interval for the mean is greater than Q , these results would pass the criterion.

The Bergum acceptance limit tables for content uniformity and dissolution are as follows:

^aTable entry of interest.

Content uniformity ($n = 30$)

Mean (% claim)	RSD (%)
97.5	3.91
97.6	3.93 ^a
97.7	3.96
97.8	3.98

^aTable entry of interest.

Dissolution ($n = 30$)

Mean (% claim)	RSD (%)
93.6	8.56
93.8	8.61 ^a
94.0	8.66
94.2	8.70

Because the sample RSD values of 2.40% for content uniformity and 3.69% for dissolution are less than the corresponding acceptance limits from the tables of 3.93% and 8.61%, both tests pass the acceptance criterion.

Example 2 (Sampling Plan 2)**Blend**

Two samples are taken from each location in the V blender, with the following results:

Location	Data (% label)		Summary statistics		
	1	2	Mean	Variance	SD
1	96.45	99.19	97.82	3.75	1.93
2	95.90	97.33	96.62	1.02	1.01
3	94.86	99.18	97.02	9.33	3.05
4	96.88	92.55	94.71	9.37	3.06
5	98.40	94.23	96.32	8.69	2.95
6	102.03	102.50	102.27	0.11	0.33
7	93.74	95.36	94.55	1.31	1.15
8	102.43	101.24	101.84	0.71	0.84
9	101.72	97.18	99.45	10.31	3.21
10	97.32	99.64	98.48	2.69	1.64
11	100.58	98.39	99.48	2.40	1.55
12	90.49	94.48	92.49	7.96	2.82

To apply the tolerance interval, SDPI, and Bergum approaches, it is necessary to compute the following variance components:

Variance components	
Source	Estimate (SD)
Between	2.483
Within	2.192
Total	3.312

The estimated standard deviation of a single observation is 3.312.

To use the tolerance interval approach, the Satterthwaite approximate degrees of freedom (df) is 16.82. The 90% tolerance interval to capture 96.7% of the individual capsule content uniformity results is $97.59 \pm 3.111 \times 3.312 = (87.29, 107.89)$. The tolerance factor was determined using linear interpolation. This would meet the criterion because the interval is completely contained within the interval 85–115%.

The s_{cr} using the SDPI is $6.25/\sqrt{2.055} = 4.36$ using the degrees of freedom from the Satterthwaite approximation. The sample standard deviation (3.312) passes this criterion.

The Bergum approach requires calculating the standard deviation of the location means, the within-location standard deviation, and the overall mean:

Mean	97.59
SE (within-location SD)	2.19
Standard deviation of location means	2.93

The standard deviation of location means is computed by taking the standard deviation of the location means. It is not the between-location variance component.

A portion of the acceptance limit table generated to meet the criterion is as follows:

**Table entry of interest.*

Standard deviation of location means

SE	2.8		2.9		3.0	
	LL	UL	LL	UL	LL	UL
2.1	97.6	102.4	98.0	102.0	98.5	101.5
2.2	97.7	102.3	98.1	101.9	98.6	101.4 ^a
2.3	97.8	102.2	98.2	101.8	98.7	101.3

The lower (LL) and upper (UL) acceptance limits for the sample mean are given for various values of the standard deviation of location means and the within-location standard deviation (SE). For our example, after rounding the standard deviation estimates up to the more restrictive values, the combination of 3.0 for the standard deviation of location means and SE of 2.2 is 98.6 to 101.4. Because the mean of 97.59 is not contained in this interval, the criterion fails.

Finished Product

Suppose that four capsules are tested at each of 15 locations throughout the batch for content uniformity and dissolution, with the following results

CU Location	Data (% label)				Summary statistics		
	1	2	3	4	Mean	Variance	SD
1	97.08	99.72	98.37	93.50	97.17	7.13	2.67
2	99.72	100.32	101.01	100.29	100.33	0.28	0.53
3	99.90	98.27	98.88	97.96	98.75	0.73	0.85
4	92.78	92.17	93.44	91.22	92.40	0.89	0.94
5	96.32	96.61	95.66	97.20	96.45	0.42	0.64
6	100.97	102.17	99.06	98.80	100.25	2.57	1.60
7	97.02	95.35	98.65	95.98	96.75	2.08	1.44
8	99.39	98.81	98.63	98.06	98.72	0.30	0.55
9	99.59	97.80	97.67	95.95	97.75	2.21	1.49
10	97.97	98.54	100.26	98.74	98.88	0.96	0.98
11	96.09	97.61	95.49	97.50	96.67	1.10	1.05
12	98.87	97.81	97.28	98.80	98.19	0.60	0.78
13	101.10	102.60	100.48	98.62	100.70	2.71	1.65
14	100.80	100.34	98.49	100.93	100.14	1.27	1.13
15	99.70	100.09	100.14	99.20	99.78	0.19	0.43

A 90% tolerance interval to capture 96.7% of the individual content uniformity results using the Satterthwaite approximation of 21.56 df is $98.20 \pm 2.746 (2.407) = (91.59,$

Variance components		
Source	Mean square	Estimate (SD)
Between	18.486	2.057
Within	1.563	1.250
Total		2.407

104.81). Because the tolerance interval is contained within the 85–115% interval, the tolerance interval indicates that the capsules have good content uniformity.

The descriptive statistics to use the Bergum approach are:

Mean	98.20
SE (within-location SD)	1.25
Standard deviation of location means	2.15

The portion of the table for this combination of results is:

^aTable entry of interest.

Standard deviation of location means						
SE	2.0		2.1		2.2	
	LL	UL	LL	UL	LL	UL
1.2	92.8	107.2	93.2	106.8	93.6	106.4
1.3	92.9	107.1	93.3	106.7	93.7	106.3 ^a
1.4	93.0	107.0	93.4	106.6	93.8	106.2

The lower and upper acceptance limits for the mean are 93.7 to 106.3. Because 98.2 falls within the interval, the capsules pass the acceptance criterion.

Dissolution							
Location	Data (% released)				Summary statistics		
	1	2	3	4	Mean	Variance	SD
1	101.4	99.5	92.9	94.9	97.16	15.55	3.94
2	106.6	101.4	98.0	100.0	101.51	13.53	3.68
3	103.9	100.6	95.3	100.5	100.07	12.64	3.56
4	96.6	93.5	92.6	94.5	94.28	2.89	1.70
5	89.4	93.1	84.6	92.4	89.89	14.97	3.87
6	90.9	90.7	93.2	91.9	91.67	1.39	1.18
7	93.8	92.6	94.8	99.8	95.27	10.08	3.17
8	99.8	98.6	98.1	92.4	97.23	11.03	3.32
9	92.4	96.0	98.4	88.8	93.90	17.86	4.22
10	100.8	99.5	90.6	99.0	97.50	21.50	4.64
11	95.9	98.2	95.9	95.9	96.47	1.39	1.18
12	103.8	103.4	100.8	104.0	102.99	2.28	1.51
13	95.2	92.2	96.1	94.2	94.43	2.88	1.70
14	96.4	98.7	95.4	101.7	98.03	7.69	2.77
15	95.7	96.7	96.2	95.9	96.13	0.17	0.41

Variance components		
Source	Mean square	Estimate (SD)
Between	48.253	3.130
Within	9.056	3.009
Total		4.342

A 90% one-sided tolerance interval to capture 99.1% of the individual dissolution values using the Satterthwaite approximation of 31.13 *df* is $96.44 - 2.907(4.342) = 83.82$. The tolerance interval indicates that the capsules are not assured of passing stage 1 of the USP 32 dissolution test. The confidence interval approach based on stages 2 and 3 of the USP 32 dissolution test has a lower bound for the population mean of $96.44 - 1.345\sqrt{48.25/\sqrt{60}} = 95.23$. Because the lower bound of 95.23 is greater than the *Q* of 85%, the criterion is met.

Using the Bergum approach, the descriptive statistics are

Mean	96.44
SE (within-location SD)	3.01
Standard deviation of location means	3.47

The portion of the acceptance limit table for this combination of results is:

^aTable entry of interest.

Standard deviation of location means			
SE	3.25	3.50	3.75
2.75	88.80	89.10	89.40
3.00	88.90	89.10	89.40
3.25	88.90	89.20 ^a	89.40

The lower acceptance limit for the mean is 89.20%. Because 96.44 is greater than 89.20, the capsules pass the acceptance criterion for dissolution.

Both the blend and final capsules passed content uniformity. Therefore, it is not necessary to perform the Analysis by Synthesis approach given in the PDA paper.^[18]

CONCLUSION

A number of statistical techniques are described for possible use in the analysis of prospective process validation data of tablets and capsules and some of their advantages and disadvantages are discussed. Detailed examples are provided to aid in the understanding of many of the techniques discussed. The authors hope that this entry will stimulate the use of the outlined statistical approaches for the analysis of validation data by industry and their acceptance by the FDA. For powder blends, industry feels compelled to use the FDA approach as it is most likely to be accepted by the FDA. However, a

number of other approaches, such as the SDPI and the Bergum approaches are also more constraining than the USP 32 test while providing a sound statistical basis for the development of acceptance criteria. For finished product testing of content uniformity and dissolution, the Bergum approach offers a number of advantages that the authors believe should be considered. It is hoped that this entry will not only encourage an increase in the use of statistical techniques for the analysis of validation data but also spur discussion of the relative merits of the various techniques.

REFERENCES

1. FDA. *Guideline on General Principles of Process Validation*; FDA: Rockville, MD, 1987.
2. Nash, R.A. Process validation for solid dosage forms. *Pharm. Technol.* **1979**, 3, 105–107.
3. Nally, J.D. Validation guidelines-industry' perspective. *Pharm. Eng.* **1984**, 4, 21–24, 26, 27, 30, 32.
4. Chapman, K.G. PAR approach to process validation. *Pharm. Technol.* **1984**, 8, 22, 24, 26, 28–29, 32–34, 36.
5. Fry, E.M. General principles of process validation. *Pharm. Eng.* **1984**, 4, 33–36.
6. Fry, E.M. Process validation policy. *Pharm. Ind. (Germany)* **1984**, 46, 601–605.
7. Fry, E.M. Process validation: The FDA' viewpoint. *Drug Cosmet. Ind.* **1985**, 137, 46–51.
8. PMA. Validation Advisory Committee Process validation concepts for drug products. *Pharm. Technol.* **1985**, 9–78, 80, 82.
9. Edwards, C.M. Validation of solid dosage forms: FDA view. *Drug Devel. Ind. Pharm.* **1989**, 15, 1119–1133.
10. Nash, R.A. Response to recent GMP-validation interpretation. *Clin. Res. Regul. Aff.* **1993**, 10, 253–264.
11. FDA. *Guide to Inspections of Validation of Cleaning Processes*; FDA: Rockville, MD, 1993.
12. *Pharmaceutical Process Validation*, 2nd Ed.; Marcel Dekker: New York, 1993.
13. O'Shea, E. Aspects of process validation in the manufacturing industry. *Irish Pharm. J.* **1995**, 73, 107–110, 112.
14. Berman, J.; Planchard, J.A. Blend uniformity and unit dose sampling. *Drug Dev. Ind. Pharm.* **1995**, 21, 1257–1283.
15. Murthy, K.S.; Bozzone, S.; Maximos, A.S. Process validation of solid oral dosage forms. *Pharm. Eng.* **1996**, 16, 42–44, 46, 50–52, 54–58.
16. Chapman, K.G. A suggested validation lexicon. *Pharm. Technol.* **1983**, 7, 51–57.
17. Kohberger, R.C. *Biopharmaceutical Statistics for Drug Development*; Marcel Dekker: New York, 1988, 605–629.
18. Blend Uniformity Analysis: Validation and In-Process Testing. *Technical Report No. 25*; PDA J. Pharm. Sci. Technol. **1997**, 51 (suppl).
19. Bergum, J.S. Constructing acceptance limits for multiple stage tests. *Drug Devel. Ind. Pharm.* **1990**, 16, 2153–2166.
20. Bergum, J.; Li, H. Acceptance Limits for the New ICH USP 29 Content Uniformity Test. *Pharm. Technol.* **2007**, 10, 90–110.
21. Hahn, G.J.; Meeker, W.Q. *Statistical Intervals: A Guide For Practitioners*; Wiley: New York, 1991.
22. Lindgren, B.W. *Statistical Theory*; Macmillan: New York, 1968.
23. Graybill, F.A.; Wang, C.M. Confidence intervals on non-negative linear combinations of variances. *J. Am. Stat. Assoc.* **1980**, 75, 869–873.
24. Dietrick, J. *Special Forum on Blend Analysis. Presentation sponsored by PDA*; Rockville, MD, 1996.
25. *SAS/QC Software®: Usage and Reference, Version 6*, 1st Ed.; SAS Institute: Cary, NC, 1995, 1, 175–186.
26. Satterthwaite, F.E. An approximate distribution of estimates of variance components. *Biometrics* **1946**, 6, 110–114.
27. Nelson, L.S. The mean successive difference test. *J. Qual. Technol.* **1980**, 12, 174–175.
28. FDA. Draft, *Guidance for Industry, Process Validation: General Principles and Practices*; FDA: Rockville, MD, 2008.

Profile Analysis

Bin Cheng

University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.

Shein-Chung Chow

Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

This entry describes the study design and bioequivalence criteria of particle size distribution for both nonprofile and profile analysis, and offers some statistical issues as well as simulated results in an example study for further illustration.

Keywords: Bioequivalence; Compositional data; In vitro test; Particle; Size distribution; Profile analysis.

INTRODUCTION

For some locally acting drug products, such as nasal aerosols and nasal sprays, bioavailability (BA) and bioequivalence (BE) assessments are complicated because delivery to the sites of action does not occur primarily after the systemic absorption. Although BA and BE studies for nasal aerosols and nasal sprays generally would consider both local delivery and systemic absorption, the study of the former is of primary importance. *In vitro* studies, which refer to the studies outside the living organism, are performed to evaluate the local delivery. According to a recent guidance for BA and BE studies for nasal aerosols and nasal sprays for local action issued by the U.S. Food and Drug Administration (FDA)^[1] in April 2003, *in vitro* BE studies are usually characterized by the following seven tests: (i) single actuation content through container life; (ii) droplet size distribution by laser diffraction; (iii) drug in small particles/droplets, or particle/droplet size distribution by cascade impactor (CI); (iv) drug particle size distribution by microscopy; (v) spray pattern; (vi) plume geometry; and (vii) priming and repriming. Detailed descriptions regarding purposes and requirements of the above tests can be found in^[1] and the entry of *In vitro* Bioequivalence Testing in this encyclopedia (Wang, et al.^[2]). The FDA^[1] recommends non-profile analysis be applied to study the content uniformity through container life, droplet size distribution, spray pattern, and priming and/or repriming, while the profile analysis be applied to the study of particle/droplet distribution by cascade impactor. The major difference between a nonprofile analysis and a profile analysis is that the data for nonprofile analysis are usually *independent* while the data for profile analysis are correlated. Although the CI profile data can be viewed as

growth curve data, due to special features described later in this entry, the usual parametric or non-parametric growth curve analysis methods can not be directly applied. There had been little literature available regarding the analysis of the CI profile data until the publication of a 1999 FDA guidance on *in vitro* bioequivalence testing for nasal aerosols and nasal sprays products (FDA^[3]). In this entry, we discuss the existing testing procedures for establishment of BE with CI profile data. For illustration and comparison purpose, we focus on droplet/particle size distribution test by cascade impactor (CI).

Study designs and BE criteria of particle size distribution (PSD) for both the nonprofile and profile analysis are described and compared in the section “Study Designs and BE criteria.” In the section “Testing Procedures for Profile Analysis,” statistical issues for assessment of BE based on PSD are discussed and two testing procedures are introduced. Simulation results and an example concerning *in vitro* test for PSD are provided to illustrate the two testing procedures in the section “Simulation Study and an Example.” A parametric method based on the additive logistic normal model is proposed in the section “Model-based Testing Methods” as an alternative approach, followed by a brief discussion in the last section.

STUDY DESIGNS AND BE CRITERIA

In pharmaceutical practice, PSD is an important factor for development of nasal aerosols and nasal sprays drug products since the sizes of particles/droplets by multistage cascade impactor (CI) are useful measurements of aerodynamic diameter based on inertial impaction, which is an important factor in the deposition of drug in the nasal

passages. The multi-stage CI used in the study is necessarily specified in order to minimize distortion and to ensure reproducibility. To ensure consistency of *in vitro* performance, the FDA^[1] recommends that at least three batches (or sublots) of each of the test (T) and reference (R) drug products be used. Although not stated in the recent FDA guidance, at least 10 containers (canisters or bottles) is usually considered. For each selected container, CI data are only required at the beginning lifestage and the number of actuations are kept to the minimum (generally no more than 10) for an accurate and reliable assessment. Depositions at different stages for each actuation are usually reported in mass unit. Let X_s be the deposition in mass unit at stage s , where $s = 1, \dots, S$. Then, data can be expressed as S -dimensional vector $X = (X_1, \dots, X_S)$. Depending on whether the study is for aerosol or spray, X can be used in different ways.

For nasal sprays drug products, small droplets/particles, defined as being smaller in diameter than the effective cut-off diameter of the top stage of the CI, are of main interest since they may be potentially delivered to the regions beyond the nose and thus cause a safety concern. For this reason, the total number of stage S may be chosen as $S = 2$. Let X_1 be the mass deposition at the top stage, and X_2 be the pooled mass of all the CI stages below the top stage, then the nonprofile analysis use only X_2 for each actuation. For simplicity, we denote X_2 as Y . For BE study, let Y_T, Y_R , and Y'_R be the pooled mass depositions at the below-the-top stages for one actuation of one test and two references canisters, then *in vitro* BE can be claimed if the following hypothesis is rejected.

$$H_0: \theta = \frac{E(Y_T - Y_R)^2 - E(Y_R - Y'_R)^2}{\max\{\sigma_0^2, E(Y_R - Y'_R)/2\}} \geq \theta_{BE}, \quad (1)$$

where σ_0^2 is a prespecified constant and θ_{BE} is the BE limit, which is based on the prior judgment and may be adjusted by the reference drug. Testing procedures based on confidence intervals have been developed. In its earlier draft guidance, FDA recommends a test proposed by Hyslop, Hsuan, and Holder^[4] (FDA^[3]). When replicates are available, Chow, Shao, and Wang^[5] provide an asymptotic test based on the linear mixed-effects model to address the within-canister and between-canister variations. The procedure proposed by Chow, Shao, and Wang^[5] is more consistent with the study design recommended by FDA.^[1] Both methods are based on normal assumption. For a general discussion of *in vitro* nonprofile BE tests, see Chow and Shao.^[6] For nasal sprays drug products, the profile analysis of BE assessment involves comparison of the profiles between the test and reference drug products. According to FDA,^[3] the profile of the PSD $X = (X_1, X_2)$ is defined as

$$p = (p_1, p_2) = \left(\frac{X_1}{X_1 + X_2}, \frac{X_2}{X_1 + X_2} \right).$$

The profile data p_1 and p_2 are correlated in a complicated way.

Unlike nasal sprays drug products in which the pooled mass of drug deposition below the top stage is of interest, the CI studies of nasal aerosols PSD focus completely on the profiles and nonprofile analysis is not required. For assessment of BE based on USP 25 Apparatus 1 (< 601 >), the FDA^[1] recommends determination of profiles based on depositions at 11 sites: (i) sum of valve stem plus actuator; (ii) induction port; (iii—x) eight individual stages; and (xi) filter. Again, deposition should be reported in mass unit. For nasal aerosols drug products, the total number of profile stages S is suggested to be at least 3.

To set up a BE criterion for CI profile data, a reasonable measure of profile distance should be determined first. The BE criterion under thus a measure should be able to calibrate the relative mean difference normalized/adjusted by the within-container and between-container variations at *each* profile stage (not the CI physical stage since the depositions at several CI stages could be combined). One possible choice of such a measure is the one suggested by the FDA.^[3] Let

$$p_T = (p_{T1}, p_{TS}), \quad \text{and} \quad p_R = (p_{R1}, p_{RS}),$$

be the profiles of a test drug and a reference drug, respectively, then their profile difference is defined as

$$d(T, R) = \sum_{s=1}^S \frac{(p_{Ts} - p_{Rs})^2}{(p_{Ts} + p_{Rs})/2},$$

and $d(R, R')$ can be defined similarly for any two reference profiles. The BE criterion given by FDA^[3] is based on a triplet (T, R, R') , formed by one test and two reference drugs. Denote $R + R'/2$ as the component-wise average of the two reference drug profiles, and define

$$rd = \frac{d\left(T, \frac{R + R'}{2}\right)}{d(R, R')}, \quad (2)$$

then the profile BE measure suggested by the FDA^[3] is $Rd = E(rd)$. The CI profile BE is established if the following hypothesis is rejected at the 5% level of significance

$$H_0: Rd \geq \theta_{BE}. \quad (3)$$

TESTING PROCEDURES FOR PROFILE ANALYSIS

As mentioned in the previous section, the study design for CI profile BE assessment usually consists of 3 lots from the test drug, 3 lots from the reference drug, and at least 10 canisters from each lot. Thus, a total of 60 canisters are

tested and profile vectors are observed. CI profile data are complicated in several aspects. First, the data are highly correlated. Second, the data are not normally distributed. In fact, it may be hard to characterize its distribution parametrically. Finally, due to limited data, which are highly correlated, asymptotic methods may not be applicable. For these reasons, model independent finite sample testing procedures are preferred. In what follows, we will introduce two finite sample approaches.

The FDA's Approach

To test (3), the FDA^[3] suggests a bootstrap type approach to estimate the 95% confidence upper limit of Rd. To obtain this upper limit, the six lots used in the study are matched into different combinations of lot-triplets, each of which contains one test lot and two different reference lots. Within each lot-triplet, the 30 canisters are matched into different combinations of canister-triplets, each of which contains one test canister and two reference canisters from different lots. A random sample of 500 canister-triplets are then selected from the population of all possible different canister-triplets. Let rd_i be the ratio in (2) computed based on the i th canister-triplet in the random sample, $i = 1, \dots, 500$, and $\overline{Rd}_{95}$ be the 95th percentile of $rd_1, \dots, rd_{500}$. Then $\overline{Rd}_{95}$ is used as a 95% upper bound of Rd and is recommended for testing (3). The BE limit θ_{BE} , which is not discussed in either FDA^[3] or FDA,^[1] has to be determined by simulations.

There are several major concerns for this approach. First, Rd may not be well defined, especially when the reference profiles are uniform or follow a compound multinomial distribution as speculated by the FDA.^[3] Second, even if Rd is well defined, rd_i is not an accurate estimator of Rd since it is based on one triplet. Consequently, $\overline{Rd}_{95}$ can not be expected to be a 95% upper bound for Rd. This point will be illustrated in the section "Simulation Study and an Example." Finally, although the FDA approach looks like either a bootstrap approach or a randomized test approach, neither approach is statistically justified since the different lot-triplets share at least one common reference lot and the rd values are neither independent nor identically distributed.

Some modification could be made to the FDA approach. First, instead of defining Rd as the mean of rd, we could define it as a certain percentile of rd. Second, instead of using one single triplet, we could use the mean of m independent triplets to compute rd_i . That is, let $\bar{T}$, $\bar{R}$, and $\bar{R}'$ be the component-wise sample means of m independent canister-triplet profiles for the test and two reference drug products, respectively, then

$$rd_i = \frac{d\left(\bar{T}, \frac{\bar{R} + \bar{R}'}{2}\right)}{d(\bar{R}, \bar{R}')} \quad (4)$$

Although after modification of the definition of Rd, the FDA BE criterion can still be valid, the validity of the FDA approach has not been statistically justified even the new rd_i in (4) is used.

Two-stage Random Sampling Approach

As noted in the previous section, the BE profile distance measure Rd suggested by the FDA^[3] may not be well defined. A modification of the definition of Rd may be made, as suggested by Cheng and Shao.^[7] Instead of defining Rd as the mean of rd, Cheng and Shao^[7] defined it as a certain percentile of rd. Along this line, a simple testing procedure is proposed. For illustration purpose, consider the case where Rd is the median of rd.

Cheng and Shao's method is based on a two-stage sampling procedure and therefore it is referred to as a two-stage random sampling (TRS) method. At the first stage, a lot-triplet is randomly selected with replacement from the nine possible lot-triplets given. At the second stage, for any sampled lot-triplet, one canister is randomly selected from each lot and rd is calculated. This process is repeated independently n times to produce independent and identically distributed $rd_1, \dots, rd_n$. Let F be the cumulative distribution function of rd_j and Rd be the median of F . Then, (3) is equivalent to

$$H_0 : 0.5 \geq F(\theta_{BE}). \quad (5)$$

Let Y be the number of rd_j 's $\leq \theta_{BE}$. Then Y has a binomial distribution with size n and probability $F(\theta_{BE})$. A level α test for (5) rejects H_0 if and only if $Y > b_{n,\alpha}$, where $b_{n,\alpha}$ is an integer satisfying

$$\sum_{k=0}^{b_{n,\alpha}} \binom{n}{k} \frac{1}{2^n} = 1 - \alpha, \quad (6)$$

and α is a positive number that is close to 5%. If $n = 18$, for example, we may take $\alpha = 0.0481$, which gives $b_{n,\alpha} = 12$. That is, the bioequivalence can be claimed with 95.19% statistical assurance when at least 12 of 18 canister-triplets have the rd values $\leq \theta_{BE}$. Note that this test procedure ensures that the level of the test is exactly α .

Since sampling of lot-triplets is random in this procedure, the actual number of sampled canisters from each lot (or lot-triplet) is random, which may not be practically appealing. This is also true for the FDA's procedure. Although 10 canisters are sampled from each lot in the FDA's bootstrap-type procedure, the number of canisters from each lot (or lot-triplet) is still random in the 500 randomly selected canister-triplets. To use a more balanced sampling design, the following modification is also considered by Cheng and Shao.^[7] Assume that in each test lot we sample $3k$ canisters and in each reference lot we sample $6k$ canisters. Then, we randomly group these canisters into

9k non-overlapped canister-triplets, with one test canister and two reference canisters from different lots in each canister-triplet. The rd value of each canister-triplet is calculated, which results in $rd_1, \dots, rd_{9k}$. Note that these rd values are identically distributed, but are slightly dependent. Let Z be the number of rd_j 's $\leq \theta_{BE}$. Then Z is *not* exactly binomial, but is close to binomial. If we still reject the null hypothesis H_0 in (5) when $Z > b_{n,\alpha}$, where $b_{n,\alpha}$ is defined by (6), then the test should be approximately of level α . Simulation shows that this second method (which can be called a random grouping method) has only a slightly inflated size. This method, however, is more balanced.

SIMULATION STUDY AND AN EXAMPLE

To evaluate the performances of the procedures described in this entry, two simulations studies were conducted by Cheng and Shao.^[7] The first simulation was to study the behavior of the FDA's procedure, while the second simulation was to study the properties of the TRS method. In the first simulation, it is assumed that the number of stage $S = 3$. Let $(M_{kj1}, M_{kj2}, M_{kj3})$ be a random vector distributed as multinomial $(100, P_{kj1}, P_{kj2}, P_{kj3})$, and ε_{kji} 's be independent random errors having gamma distributions with mean 1 and variances σ_k^2 , where $j (= 1, 2, 3)$ is the index for lots, $k (= T, R)$ is the index for test or reference, and ε_{kji} 's and M_{kji} 's are independent. The profile vectors are given by

$$(X_{kj1}, X_{kj2}, X_{kj3}) = (\varepsilon_{kj1} M_{kj1}, \varepsilon_{kj2} M_{kj2}, \varepsilon_{kj3} M_{kj3}).$$

The values of P_{kji} were obtained from estimates in a real data set:

$$\begin{aligned}(P_{T11}, P_{T12}, P_{T13}) &= (0.99735, 0.00235, 0.00030) \\(P_{T21}, P_{T22}, P_{T23}) &= (0.99780, 0.00150, 0.00070) \\(P_{T31}, P_{T32}, P_{T33}) &= (0.99670, 0.00285, 0.00045) \\(P_{R11}, P_{R12}, P_{R13}) &= (0.99805, 0.00170, 0.00025) \\(P_{R31}, P_{R32}, P_{R33}) &= (0.99675, 0.00240, 0.00085)\end{aligned}$$

Four different values of (σ_T, σ_R) were considered:

$$(0.05, 0.05), (0.05, 0.10), (0.10, 0.05), (0.10, 0.10).$$

For each fixed value of (σ_T, σ_R) , the random variable rd is defined according to (2) with the triplet T, R, R' randomly selected from the following nine possible of lot-triplets:

$$\begin{array}{lll}T1, R1, R2 & T1, R1, R3 & T1, R2, R3 \\T2, R1, R2 & T2, R1, R3 & T2, R2, R3 \\T3, R1, R2 & T3, R1, R3 & T3, R2, R3\end{array}$$

Table 1 lists the median rd_{50} and 95th percentile rd_{95} of rd for each value of (σ_T, σ_R) based on 10,000,000 simulations. Also included in Table 1 are simulation proportions of $\overline{Rd}_{95} < \text{the median}$, 2, or the 95th percentile based on 5,000 simulations.

When rd_{50} is considered as the parameter Rd, $P(\overline{Rd}_{95} < rd_{50})$ reported in Table 1 is a type I error probability of FDA's test procedure if $\theta_{BE} = rd_{50}$, whereas $P(\overline{Rd}_{95} < 2)$ in Table 1 is a value of the power of the FDA's test. Since all values of the median are substantially smaller than 2, the result in Table 1 indicates that the FDA's test is very conservative. On the other hand, when rd_{95} is considered as the Rd in Table 1 and $\theta_{BE} = rd_{95}$, $P(\overline{Rd}_{95} < rd_{95})$ in Table 1 is a type I error probability of FDA's test and the result in Table 1 shows that FDA's procedure is incorrect and too liberal. Even if $\overline{Rd}_{95}$ is viewed as an estimator of rd_{95} , it is an incorrect estimator since $P(\overline{Rd}_{95} < rd_{95})$ may be far away from 0.5, for example, in the cases where $(\sigma_T, \sigma_R) = (0.05, 0.10)$ and $(\sigma_T, \sigma_R) = (0.10, 0.05)$.

As indicated in Cheng and Shao,^[7] the second simulation was done to study the properties of the TRS method. In the second simulation, similar settings were considered. That is, $n = 18(b_{n,\alpha} = 12)$ two-stage random samples of rd are considered, since this leads to a total of 54 canisters, which is close to the total of 60 canisters in FDA's procedure. Results based on 5,000 simulations are given in Table 2.

From Table 2, when θ_{BE} is chosen to be the median of rd (given in Table 1), the probabilities in Table 2 are type I error probabilities. The nominal value is $\alpha = 0.0481(n = 18)$. Hence, the results in Table 2 are within the nominal value $\pm 0.006 = \pm 2$ estimated standard simulation error. When $\theta_{BE} = 1, 1.5$, or 2, the results in Table 2 are values of power except for the case where $\theta_{BE} = 1$ and $(\sigma_T, \sigma_R) = (0.10, 0.05)$. It can be seen that a reasonable power can be achieved when θ_{BE} is substantially larger than

Table 1 Simulation Results for FDA's Upper Bound $\overline{Rd}_{95}$

	(σ_T, σ_R)			
	(0.05, 0.05)	(0.05, 0.10)	(0.10, 0.05)	(0.10, 0.10)
rd_{50}	0.6753	0.5378	1.3305	0.7684
rd_{95}	117.3226	44.1099	230.5589	70.3797
$P(\overline{Rd}_{95} < rd_{50})$	0	0	0	0
$P(\overline{Rd}_{95} < 2)$	0	0	0	0
$P(\overline{Rd}_{95} < rd_{95})$	0.5628	0.0908	0.9626	0.5368

Table 2 Simulation Results on the Rejection Probability for the TRS Method

θ_{BE}	(σ_T, σ_R)			
	(0.05, 0.05)	(0.05, 0.10)	(0.10, 0.05)	(0.10, 0.10)
rd ₅₀	0.0494	0.0460	0.0538	0.0508
1.0	0.1748	0.3760	0.0132	0.1248
1.5	0.3596	0.6340	0.0752	0.3646
2.0	0.4842	0.7942	0.1560	0.5556

the true median of rd or the test variability σ_T^2 is smaller than the reference variability σ_R^2 .

Example

An *in vitro* BE profile study for nasal solutions was conducted to compare the test drug and the generic drug (reference). Three lots were selected for both the test solution and the reference solution. Within each lot, 18 canisters were drawn and their profiles at the three stages were recorded. A summary of the profile data is given in Table 3.

If we choose $\theta_{BE} = 2$, then for the TRS method, statistic Y defined in §3.2 is $Y = 13$. Therefore, the null hypothesis in (5), or (3) is rejected and the *in vitro* BE is claimed. For the FDA method, formula (4) with $m = 100$ is used to compute the rd's, the resulting 95% percentile is $\overline{\text{Rd}}_{95} = 270.94$, which is much larger than $\theta_{BE} = 2$. This example shows that the FDA method is both unstable and too conservative.

MODEL-BASED TESTING METHODS

For any profile $\mathbf{p} = (p_1, \dots, p_S)$, there is a sum-to-one constraint on its components, that is $p_1 + \dots + p_S = 1$ and $p_s > 0, s = 1, \dots, S$. This imposes a statistical challenge to the modeling and analysis of profile data. It is clear that neither one of the two methods described in the section “Testing Procedures For Profile Analysis” takes this constraint into consideration. In this section, we discuss some alternative testing methods based on modeling of this special type of data.

In view of the sum-to-one constraint for the profile data, one immediate choice for its distribution is a Dirichlet Model. However, the covariance matrix of a Dirichlet distribution has a special structure, which implies that it may not be too restrictive to model the profile data via Dirichlet distribution. On the other hand, it turns out that profile

data also arise in many other areas, particularly in geology, where such type of data is termed as “compositional data.” There is an extensive literature on statistical methods for compositional data. The most influential method, the logistic normal method proposed by Aitchison,^[8] will be described below. Other methods include generalized Liouville method by Rayens and Srinivasan,^[9] and the Bayesian method by Iyengar and Dey.^[10]

We adopt the logistic normal method for its simplicity, maturity, and wide applications. The method is based on the following additive log ratio transformation of the compositional data

$$\mathbf{y} = \left(\log \frac{p_1}{p_S}, \dots, \log \frac{p_{S^*}}{p_S} \right),$$

where $S^* = S - 1$. A profile vector $\mathbf{p}$ is said to have an additive logistic normal distribution, denoted as $\mathbf{p} \sim L_{S^*}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if $\mathbf{y} \sim N_{S^*}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, an S^* -dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. It can be proved that if a deposition vector $\mathbf{X} = (X_1, \dots, X_S)$ has a multivariate log-normal distribution, then $\mathbf{p}$ has an additive logistic normal distribution (Aitchison^[11]).

Let $\mathbf{p}_{Ti}$ be the profile of the i th subject in the treatment group and $\mathbf{p}_{Ri}$ be the i th subject in the treatment group, $i = 1, \dots, n$. We consider the following model

$$\begin{aligned} \mathbf{p}_{T1}, \dots, \mathbf{p}_{Tn} &\sim L_{S^*}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T), \\ \mathbf{p}_{R1}, \dots, \mathbf{p}_{Rn} &\sim L_{S^*}(\boldsymbol{\mu}_R, \boldsymbol{\Sigma}_R). \end{aligned} \quad (7)$$

Here we assume equal sample sizes for the two groups but allow $\boldsymbol{\Sigma}_T \neq \boldsymbol{\Sigma}_R$. A nature choice of BE measure is

$$\theta = (\boldsymbol{\mu}_T - \boldsymbol{\mu}_R)'(\boldsymbol{\Sigma}_T + \boldsymbol{\Sigma}_R)^{-1}(\boldsymbol{\mu}_T - \boldsymbol{\mu}_R),$$

based on which the profile BE hypothesis can be formulated as

$$H_0: \theta \leq \theta_{BE}, \quad (8)$$

where θ_{BE} is a prespecified BE limit. Here we briefly describe the testing method for (8). Further detail is provided in Cheng.^[12] Let $\bar{\mathbf{Y}}_T, \bar{\mathbf{S}}_T$ be the sample mean and sample covariance of $\mathbf{y}_{Ti}$'s, the additive log ratio transformation of $\mathbf{p}_{Ti}$'s, respectively. Define

$$T^2 = n(\bar{\mathbf{y}}_T - \bar{\mathbf{y}}_R)'(\mathbf{S}_T + \mathbf{S}_R)^{-1}(\bar{\mathbf{y}}_T - \bar{\mathbf{y}}_R).$$

Table 3 Summary of Profile Data for Test and Reference Nasal Solutions

	Overall mean profile	Within canister standard deviation	Between canister standard deviation
Test	(0.9983, 0.0010, 0.0007)	0.0011	0.0014
Reference	(0.9942, 0.0042, 0.0016)	0.0012	0.0016

It is easily shown that T^2 is asymptotically distributed as $\chi^2_{S^*}(n\theta)$, a non-central chi-square distribution with S^* degrees of freedom and noncentrality parameter $n\theta$. Thus, the test which rejects H_0 in (8) if $T^2 > \chi^2_{S^*,\alpha}(n\theta_{BE})$ is an asymptotically level α test, where $\chi^2_{S^*,\alpha}(n\theta_{BE})$ is the $(1-\alpha)$ th percentile of $\chi^2_{S^*,\alpha}(n\theta_{BE})$. Under further assumption that $\Sigma_T = \Sigma_R$, we have

$$\frac{2(n-S^*)}{(n-1)S^*}T^2 \sim F_{S^*, 2(n-S^*)}(n\theta),$$

a noncentral F-distribution with $(S^*, 2(n-S^*))$ degrees of freedom and non-centrality parameter $n\theta$. An exact test for (8) can be derived easily.

Some quick remarks are in order. First, this new transformation approach is consistent with methods used in other fields where compositional data arise. The new BE criterion is easier to interpret and incorporates the variance information. Second, the new test is based on large sample theory hence is robust. Third, other BE measures/criteria can also be utilized under model (7). For example, we could define BE measure as

$$\theta = (\mu_T - \mu_R)' \Sigma_R^{-1} (\mu_T - \mu_R) + \text{tr}(\Sigma_R^{-1/2} \Sigma_T \Sigma_R^{-1/2} - I)$$

and the corresponding test can be derived using the δ -method. This BE measure is analogous to the following nonprofile BE measure suggested by FDA for univariate responses

$$\frac{(\mu_T - \mu_R)^2}{\sigma_R^2} + \frac{(\sigma_R^2 - \sigma_T^2)}{\sigma_R^2}.$$

Finally, we may jointly model the total deposition $W = \sum_{s=1}^S X_s$ and the profile $\mathbf{p}$ as $(\log W, \mathbf{y})' \sim N_S(\tilde{\mu}, \tilde{\Sigma})$. This endeavor is relevant since the total deposition may also be important in establishing the *in vitro* BE.

DISCUSSION

Both of the approaches discussed in the section “Testing Procedures For Profile Analysis” are non-parametric approach. The major differences between the TRS method and the FDA method include (i) the TRS method produces independent, identically distributed rd values; while the rd values in the FDA method are dependent, and (ii) for a given total number of non-overlapping canister triplets, the FDA method is more computationally intensive.

It should be noted that a chi-square type measure is adopted by both methods. Other measures may also be

used to formulate a BE criterion. Other possible choices of profile distance measures discussed in Tsong, Sathe, and Shah^[13] may be considered.

Although nonparametric methods are generally preferred for the CI profile BE study, it is more convenient to model the PSD profile parametrically. In the section “Model-Based Testing Methods,” we provide one such method which is both easy to interpret and implement. Since the asymptotic test derived in section “Model-Based Testing Methods” is still valid even if the assumption in (7) fails to hold, the method should be considered as robust and nonparametric in nature. Finally, we point out that profile data can also be modeled semiparametrically using a quasi-likelihood approach (Cheng^[12]).

REFERENCES

1. FDA. *Guidance for Industry on Bioavailability and Bioequivalence Studies for Nasal Aerosols and Nasal Sprays for Local Action*; Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, MD, 2003.
2. Wang, H.; Zhang, Y.; Shao, J.; Chow, S.C. *In Vitro Bioequivalence Testing*. In *Encyclopedia of Biopharmaceutical Statistics*; Chow, S. C., Ed.; Marcel Dekker, Inc.: New York, 2003, 449–455.
3. FDA. *Guidance for Industry on Bioavailability and Bioequivalence Studies for Nasal Aerosols and Nasal Sprays for Local Action*; Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, MD, 1999.
4. Hyslop, T.; Hsuan, F.; Holder, D. J. A small sample confidence interval approach to assess individual bioequivalence. *Stat. Med.* **2000**, *19*, 2885–2897.
5. Chow, S. C.; Shao, J.; Wang, H. *In vitro* bioequivalence testing. *Stat. Med.* **2003**, *22*, 55–68.
6. Chow, S. C.; Shao, J. *Statistics in Drug Research—Methodologies and Recent Developments*; Marcel Dekker, Inc.: New York, 2002.
7. Cheng, B.; Shao, J. Profile analysis for assessing *in vitro* bioequivalence. *J. Biopharm. Stat.* **2002**, *3*, 323–332.
8. Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc. Ser. B* **1982**, *44*, 139–177.
9. Rayens, W. S.; Srinivasan, C. Dependence properties of generalized Liouville distributions on the simplex. *J. Am. Stat. Assoc.* **1994**, *89*, 1465–1470.
10. Iyengar, M.; Dey, D. Bayesian Analysis of Compositional Data. In *Generalized Linear Models: A Bayesian Perspective*; Dey, D. K.; Ghosh, S. K.; Mallick, B. K., Eds.; Marcel Dekker: New York, 2000, 349–363.
11. Aitchison, J. *The statistical Analysis of Compositional Data*; Chapman & Hall/CRC: New York, 1986.
12. Cheng, B. Assessing In Vitro Profile Bioequivalence: A Modeling Approach, submitted.
13. Tsong, Y.; Sathe, P. M.; Shah, V. P. *In Vitro* Dissolution Profile Comparison. *Encyclopedia of Biopharmaceutical Statistics*; Chow, S. C., Ed.; Marcel Dekker, Inc.: New York, 2003, 456–462.

Propensity Score Analysis and Its Application in Regulatory Settings

Lilly Q. Yue

Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

Propensity score analysis is a versatile statistical method for causal inference in observational studies. This entry provides the equation for determining propensity score and details the procedure of this type of analysis.

INTRODUCTION

First introduced by Rosenbaum and Rubin,^[1] the propensity score $e(X)$ for a subject with a vector X of observed covariates is the conditional probability of receiving treatment ($Z = 1$) rather than control ($Z = 0$) given X :

$$e(X) = \Pr(Z = 1 | X) \quad (1)$$

Propensity score analysis is a versatile statistical method for causal inference in observational studies that can improve treatment comparison by adjusting for up to a relatively large number of potentially confounding covariates.^[1–3] In recent years, there has been an increased interest in applying this methodology to non-randomized clinical studies,^[4–6] including those designed to support regulatory applications for marketing medical products.

The propensity score $e(X)$ is a balancing score in the sense that it is a function of the observed covariates X such that the conditional distribution of X given $e(X)$ is the same for subjects in the treatment group and subjects in the control group.^[1] If treatment assignment is strongly ignorable, that is, the treatment assignment Z and the potential outcome Y are conditionally independent given the observed covariates, X , i.e., $\Pr(Z | X, Y) = \Pr(Z | X)$, and if $0 < \Pr(e(X)) < 1$, for all X , then the average treatment effect at each value of the propensity score is an unbiased estimate of the true treatment effect at that propensity score.^[1] The above assumption requires that all covariates relevant to both treatment assignment and potential outcomes are measured (i.e., there is no hidden bias). In practice, the propensity score is estimated by modeling the probability of treatment group membership as a function of the observed baseline covariates via, for example, a multiple logistic regression or discriminant analysis. Propensity score matching, stratification, weighting, covariate adjustment, and some combination of those strategies can then

be used to produce unbiased estimates of the treatment effect (Refs. [1–3; 7–15] and references therein).

Rubin^[7] points out that as a balancing score, the propensity score is a one-dimensional summary of the observed covariates such that when the propensity scores are balanced across two treatment groups, the distribution of all the covariates are balanced in expectation across the two groups. Therefore, evaluating treatment-group overlap in terms of the distributions of propensity scores provides a straightforward assessment of whether the two treatment groups are sufficiently similar with respect to baseline covariates to allow a sensible treatment comparison. When sufficient overlap is present, the propensity score analysis can provide a straightforward treatment comparison that reflects adjustment for differences in all observed baseline covariates. Otherwise, as indicated by Rubin,^[7] either the propensity score prediction model would need to be reformulated or it would have to be concluded that the covariate distributions did not overlap sufficiently to allow propensity score analysis to adjust for these covariates. Therefore, in a regulatory environment, it is crucial to start with comparable patient populations, because propensity score analysis does not work well (in fact, no statistical method works well) when there is serious imbalance in baseline covariates between the two treatment groups. Unfortunately, it is sometimes impossible to predict in advance whether the patient population with a new treatment is comparable to that with a control, which is a risk that must be taken into consideration. In fact, an unsuccessful non-randomized study can be more burdensome than a randomized trial at the outset.^[16,17]

Like other covariate adjustment methods, propensity score methods can only adjust for observed covariates and not for unobserved ones. This is always a limitation of non-randomized studies compared with randomized trials where randomization tends to balance both observed and unobserved covariates. Therefore, some sensitivity analysis for hidden bias is often desirable in a medical product regulatory

submission based on a non-randomized study. It should also be noted that propensity score methods may not eliminate all the bias resulting from imbalance in the observed covariates, due to the limitation of propensity score modeling, which typically uses a linear combination of covariates.^[7]

Braitman and Rosenbaum^[10] state that the propensity score methods work better under three conditions: (i) when clinical outcome event is rare, (ii) when there are a large number of patients in each treatment group, and (iii) when there are many covariates measured. In a regulatory setting, it may be inappropriate to conduct propensity score analysis when sample size is small, which could result in some propensity score subclasses, such as quintiles, containing patients from only one treatment group, and the inappropriate exclusion of these patients from the data analysis.^[16] Furthermore, although it is expected that all observed covariates could be balanced between the two treatment groups by propensity scores, in a small non-randomized study, substantial imbalance of some covariates might be unavoidable despite the use of a sensible propensity score method. It is essential to collect a rich set of clinically relevant baseline covariates that are potentially related to both the treatment assignment and the clinical outcomes.^[15] These covariates should be well measured in the study and used in the data analysis.

Rubin^[9] emphasizes the importance of designing of observational studies prospectively and advocates that observational studies can and should be designed to approximate randomized experiment as closely as possible without any access to any outcome data, thereby assuring the objectivity of the design. Particularly in a regulatory environment, the prospective design of a non-randomized confirmatory study with propensity score analysis specified as primary analysis is essential, in order to provide valid and convincing evidence of safety and effectiveness of a medical product.^[16,17] As discussed above, it is crucial to select comparable patient populations, as no statistical method works well when there is severe imbalance in baseline covariates between the two treatment groups. In the study protocol, details of the propensity score data analysis need to be pre-specified, including the propensity score method, for example, stratification, missing-data handling strategy for baseline covariates, and sample size estimation in the context of propensity scores, as propensity score adjustment may be sensitive to these factors.^[12,18] It is also important to pre-specify all clinically relevant baseline covariates, collect them in the study, and use them in the data analysis. At the same time, sensitivity analysis should be planned to take into consideration unobserved covariates. Furthermore, the assessment of the success of the propensity score estimation should be performed by checking the resulting balance of the distributions of covariates after propensity score adjustment, using Rubin's^[8] balance criteria or Rosenbaum and Rubin's^[2] balance criteria. In addition, the evaluation of treatment group comparability with respect to the distribution of baseline covariates should be conducted by assessing treatment group balance in terms of the distribution of propensity scores.

REFERENCES

1. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, *70*, 41–55.
2. Rosenbaum, P.R.; Rubin, D.B. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* **1984**, *79*, 516–524.
3. Rosenbaum, P.R. Propensity score. In *Encyclopedia of Biostatistics*, 2nd Ed; Armitage, P., Colton, T.; Eds.; John Wiley & Sons: Chichester, 2005.
4. Blackstone, E.H. Comparing apples and oranges. *J. Thorac. Cardiovasc. Surg.* **2002**, *1*, 8–15.
5. Grunkemeier, G.L.; Payne, N.; Jin, R.; Handy, J.R., Jr. Propensity score analysis of stroke after off-pump coronary artery bypass grafting. *Ann. Thorac. Surg.* **2002**, *74*, 301–305.
6. Wolfgang, C.; Winkelmayr, W.C.; Glynn, R.; Mittleman, M.; Levin, R.; Pliskin, J.; Avorn, J. Comparing mortality of elder patients on hemodialysis versus peritoneal dialysis: A propensity score approach. *J. Am. Soc. Nephrol.* **2002**, *13*, 2353–2362.
7. Rubin, D.B. Estimating casual effects from large data sets using propensity scores. *Ann. Intern. Med.* **1997**, *127*, 757–763.
8. Rubin, D.B. Using propensity score to help design observational studies: Application to the tobacco litigation. *Health Serv. Outcomes Res. Methodol.* **2001**, *2*, 169–188.
9. Rubin, D.B. The design versus the analysis of observational studies for causal effects: Parallel with the design of randomized trials. *Stat. Med.* **2007**, *26*, 20–36.
10. Braitman, L.; Rosenbaum, P.R. Rare outcomes, common treatments: Analytic strategies using propensity scores. *Ann. Intern. Med.* **2002**, *137*, 693–696.
11. D'Agostino, R.B., Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat. Med.* **1998**, *17*, 2265–2281.
12. D'Agostino, R.B., Jr.; Rubin, D.B. Estimating and using propensity scores with partially missing data. *J. Am. Stat. Assoc.* **2000**, *95*, 749–759.
13. Lunceford, J.K.; Davidian, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat. Med.* **2004**, *23*, 2937–2960.
14. Hill, J.; Reiter, J. Interval estimation for treatment effects using propensity score matching. *Stat. Med.* **2006**, *25*, 2230–2256.
15. Brookhart, M.A.; Schneeweiss, S.; Rothman, K.J.; Glynn, R.J.; Avorn, J.; Sturmer, T. Variable selection for propensity score models. *Am. J. Epidemiol.* **2006**, *163*, 1149–1156.
16. Yue, L.Q. Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies. *J. Biopharm. Stat.* **2007**, *17*, 1–13.
17. Li, H.; Yue, L.Q. Statistical and regulatory issues in Non-randomized medical device clinical studies. *J. Biopharm. Stat.* **2008**, *18*, 20–30.
18. Qu, Y.; Lipkovich, I. Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Stat. Med.* **2009**, *28*, 1402–1414.

Propensity Score: Methodology and Application to Clinical Studies

Lilly Q. Yue and Heng Li

Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

Formulated by Rosenbaum and Rubin, the propensity score methodology is a versatile statistical technique used mainly in observational (non-randomized) studies for improving treatment comparison by adjusting for up to a relatively large number of potentially confounding covariates in estimating the effect of treatments on outcomes. Since its formulation, the methodology has been widely used in areas such as epidemiological and social science studies. There has been an increased interest in applying this methodology to non-randomized clinical studies, including those designed to support regulatory applications for marketing medical products. This entry gives a brief overview of propensity score methods and discusses important things to consider in the application of these methods in clinical studies, especially those submitted for regulatory review.

Keywords: Propensity score; Matching; Stratification; Weighting; Observational study.

INTRODUCTION

First introduced by Rosenbaum and Rubin,^[1] the propensity score $e(X)$ for a subject with a vector X of observed baseline covariates is the conditional probability of receiving treatment ($Z = 1$) rather than control ($Z = 0$), given X : $e(X) = \Pr(Z = 1 | X)$. The propensity score $e(X)$ is a balancing score in the sense that for subjects with the same propensity score, the distribution of observed covariates is the same between the treated group and the control group. If treatment assignment is strongly ignorable, that is, the treatment assignment Z and the potential outcome Y are conditionally independent, given the observed covariates, X , i.e., $\Pr(Z | X, Y) = \Pr(Z | X)$, and if $0 < e(X) < 1$, for all X , then the average treatment effect on outcome at each value of the propensity score is an unbiased estimate of the true treatment effect at that value of propensity score.^[1] The above assumption requires that all covariates relevant to both treatment assignment and outcome are measured and captured in X (i.e., there is no hidden bias). In practice, the propensity score is estimated by modeling the probability of treatment group membership as a function of the observed covariates, typically via logistic regression.

PROPENSITY SCORE METHODOLOGY

Propensity score methodology refers to a collection of versatile statistical tools for causal inference in observational

studies based on the concept of propensity score.^[1–12] Commonly used propensity score methods include propensity score matching, stratification (subclassification) on the propensity score, inverse probability of treatment weighting (IPTW) using the propensity score, covariate adjustment using the propensity score, and some combinations of these approaches.^{[1–8][10–17]} The propensity score methods could be used to design and analyze an observational study, mimicking some of the characteristics of a randomized controlled trial.^[13] Although the first three of the aforementioned methods enable the separation of the design and the analysis, an appealing feature that is critical in clinical studies submitted for regulatory review, covariate adjustment using the propensity score does not enable such separation.

Like other covariate adjustment methods, propensity score methods can only adjust for observed covariates and not for unobserved ones. The impossibility of adjusting for unobserved covariates is a fundamental limitation of non-randomized studies that contrasts them with randomized trials where randomization tends to balance both observed and unobserved covariates. Therefore, some sensitivity analysis for hidden bias is often desirable in a medical product regulatory submission based on a non-randomized study. It should also be noted that propensity score methods may not eliminate all bias resulting from imbalance in the observed covariates, due to the intrinsic limitation of propensity score modeling.

Braitman and Rosenbaum^[3] state that the propensity score methods work better when there are a large number of subjects in each treatment group and when there are many covariates measured. In a regulatory setting, it may be inappropriate to conduct propensity score analysis when sample size is small because small sample size could result in some propensity score subclasses, such as propensity score quintiles, containing subjects from only the treated group or only the control group.^[18] Also, although it is expected that all observed covariates could be balanced between the treated and control groups through propensity scores, in a small non-randomized study, substantial imbalance of some covariates might be unavoidable despite the use of a sensible propensity score model. On the other hand, it is essential to collect a rich set of clinically relevant baseline covariates that are potentially related to both treatment assignment and clinical outcomes, in order to minimize residual bias.^[17] These covariates should be well measured in the study and used in the study design and outcome analysis.

PROPENSITY SCORE APPLICATIONS

Regarding the application of the propensity score methodology, a notable distinction between the exploratory work of general research and the confirmatory study in the regulatory setting is the necessity of prespecification of study objectives and outcome-free study design in the latter.^[18–21] Rubin^[13–15] points out that the propensity score methodology can be utilized to design an observational study without any access to any outcome data, thereby ensuring the objectivity of the design. Here, the study design refers to the process of propensity score estimation and covariate balance assessment, as well as the development of statistical analysis plan (SAP). These should be carried out objectively, that is, in an outcome-free manner. The study design also encompasses identification of control group and baseline covariate to be collected as well as sample size and power considerations.

Objective Study Design

In the regulatory setting, the prospective design of a non-randomized confirmatory study or observational postmarket study based on propensity score is particularly essential for providing objective evidence of safety and effectiveness.^[19–22] The prospective design using propensity scores can be carried out in two stages as proposed by Yue, Lu, and Xu.^[22] The first design stage is completed before the initiation of the investigational study. One main task at the beginning of this stage is to identify an independent statistician. This statistician needs to be blinded to any outcome data throughout the entire design stage. Some masking mechanism, such as a firewall, should be planned in this stage to control access of outcome data.

The planning of the initial sample size should also be done during the first stage. The second design stage should be initiated as soon as subject enrollment is concluded and baseline covariates data are collected, cleaned, and locked.^[22,23] The independent statistician identified in the first design stage should perform the task of estimating the propensity scores, matching subjects based on the propensity scores, and assessing the covariate balance. This may involve an iterative process until the balance is reached in the covariate distributions. The SAP should be finalized at this stage. Note that in the study design phase, only the treatment assignment and baseline covariate data are needed. The outcome data are not required nor should they be accessed.

Covariate Balance Diagnostics

Covariate balance assessment is crucial in the iterative process of propensity score modeling. In mimicking randomized clinical trials, propensity score methods are used in observational studies to achieve the distributional balance of baseline covariates between the treated and control groups. However, it is by no means guaranteed that balance is achieved, and so balance should be checked. Some graphical methods (such as side-by-side boxplots and LOVE plots)^[24] and numerical diagnostics^[2–17] are available, none of which is sufficient on its own.

One graphical method compares the distributions of the estimated propensity scores between the treated and the control groups to see whether there is sufficient number of subjects in the two groups that are similar enough to make treatment comparison on outcomes (common support). Without sufficient overlap, one may have no choice but to conduct the study only on a subset of the patients. However, in so doing, the generalizability of the study may be limited. In the premarket setting in particular, exclusion of patients treated with the investigational medical product could lead to the dangerous consequence of possibly changing its intended use by changing the patient population and such post hoc change could be problematic. On the other hand, good overlap in the propensity score distributions does not guarantee covariate balance. Therefore, it is necessary to demonstrate that the distributions of the baseline covariates are similar between the treatment groups, after the propensity score adjustment. With propensity score matching, the covariate balance assessment should be made on the matched samples. When propensity score stratification is used, the covariates should be balanced within each stratum.

One of the most commonly used numerical balance diagnostics is the “standardized difference in means,” which is the absolute value of difference in means of each covariate, divided by the pooled standard deviation.^[16–25] It is pointed out that statistical significance testing should not be used as a measure of covariate balance,^[26–27] and the

c-statistic of the propensity score model does not provide an appropriate assessment of covariate balance.^[26–28]

Covariate Selection and Identification

There is no consensus in the literature as to which baseline covariates to include in the propensity model. Four possible sets of baseline covariates are considered:^[25] all observed baseline covariates, all baseline covariates that are associated with treatment assignment, all covariates that affect the outcome (i.e., the potential confounders), and all covariates that affect both treatment assignment and the outcome (i.e., the true confounders). To satisfy the assumption of ignorable treatment assignment in the propensity score methodology (i.e., no unobserved difference between the two treatment groups conditional on the observed covariates), it is important to include all covariates known to be related to both treatment assignment and outcomes.^[29,30] In practice, it may be difficult to make a judgment on whether a covariate is a confounding variable. Therefore, it is better to include as many variables as possible that are collected in a study, particularly in a premarket confirmatory study, as failure to include a key confounding variable in the propensity score model could lead to a substantially biased estimate of treatment effect on outcomes.^[23]

The covariates to be collected and included in the propensity score modeling should be prospectively identified. Identification of such covariates should be a thoughtful process based on scientific reasoning and up-to-date clinical research. This process needs to be outcome-free to avoid potential bias and to ensure the validity of study conclusion.^[13,22]

It should be emphasized that the propensity score model should only include covariates that are observed at baseline and not posttreatment covariates that may be influenced or modified by the treatment.^[25]

Control Group Selection

It is crucial to select the control group from a relatively comparable patient population to begin with, since propensity score methods may not work (in fact, no statistical method would work) when there is serious imbalance in baseline covariates between the two treatment groups.^[18,22] However, it is sometimes impossible to predict in advance whether the patient population with an investigational medical product is comparable to that with a control, which is a risk for all non-randomized studies.

Sample Size Estimation and Power Consideration

In general, the more comparable the baseline covariates between treatment and control groups are, the smaller the sample size that is required to achieve desired power. When

using propensity score methodology to design a study, the degree of non-comparability of two treatment groups and the propensity score adjustment should be taken into account in sample size estimation.^[22] However, it is sometimes difficult to predict in advance and hence hard to make assumptions on treatment group comparability.

Sample size determination may depend on the particular subject matching approach to be used. For example, stratification and 1:1 matching may lead to different sample sizes. It may also depend on analysis methods for outcomes. Therefore, it is important to plan the sample size under various design/analysis scenarios.^[22]

Missing Data in Covariates

It is common for a large number of baseline covariates to be measured and collected, as in cardiovascular device studies. This could result in a large proportion of patients with at least one covariate value missing. Without any action on the missing covariate values, these subjects could be automatically excluded from propensity score modeling and then from treatment comparison with respect to clinical outcome, leading to questionable study results.

Missing data in covariates can be dealt with by using the familiar method of multiple imputation or some adaptations thereof.^[9] After missing data have been imputed, propensity score models can be built, one for each of the m completed data sets. One approach to estimating the treatment effect is to obtain a single propensity score for each patient as the average of the m propensity scores from the m models and then use this average propensity score to generate a single estimate of the treatment effect. Another approach is to use the m propensity score models to generate m treatment effect estimates and then combine those m estimates. Mitra and Reiter^[31] compared those two approaches via simulation studies and suggested that the first approach has greater potential to produce substantial bias reductions.

Outcome Analysis

Once the propensity score model is finalized, matching or stratification is implemented, and covariate balance is found to be adequate, outcome data can be unblinded and outcome analysis can start. If propensity score matching is used, there is debate about whether outcome analysis needs to take into account the matched nature of the data.^[17,25] If propensity score stratification is used, then treatment effect is estimated as a weighted average across strata. One weighting scheme produces an estimate of the average treatment effect, which is the average effect of moving an entire population from control to treated. Another weighting scheme produces an estimate of the average treatment effect on the treated (ATT), which is the average effect of treatment on those subjects who ultimately received the treatment.^[25] When

choosing a weighting scheme, it should be kept in mind that the estimand needs to be clinically meaningful.

CONCLUSION

Propensity score methodology expands techniques that use just one confounding covariate, allows simultaneous adjustment for many observed covariates, and thus reduces bias in treatment comparisons. It could be used to approximate randomized clinical trials and transparently design and analyze observational studies. One unique feature of propensity score methods, be it propensity score matching, stratification, or IPTW, is the ability to separate study design and outcome analysis, which is an essential requirement in clinical studies submitted for regulatory review. The methodology has been used in premarket comparative observational (non-randomized) medical device studies, using a concurrent (but non-randomized) control, a historical control formed from patients with existing data collected from earlier studies of a previously approved device, or a control group extracted from a well-designed and executed national or international patient registry database. It has also been utilized in drug safety comparative observational studies, using electronic medical record databases or administrative health-care databases to estimate treatment effects of intervention.^[21] We believe that the application of propensity score methodology in clinical studies will continue to grow in the future.

REFERENCES

1. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, *70* (1), 41–55.
2. Austin, P. A tutorial and case study in propensity score analysis: An application to estimating the effect of in-hospital smoking cessation counseling on mortality. *Multivar. Behav. Res.* **2011**, *46* (1), 119–151.
3. Braitman, L.; Rosenbaum, P.R. Rare outcomes, common treatments: Analytic strategies using propensity scores. *Ann. Int. Med.* **2002**, *137* (8), 693–695.
4. Brookhart, M.A.; Schneeweiss, S.; Rothman, K.J.; Glynn, R.J.; Avorn, J.; Sturmer, T. Variable selection for propensity score models. *Am. J. Epidemiol.* **2006**, *163*, 1149–1156.
5. D'Agostino, R.B. Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat. Med.* **1998**, *17* (19), 2265–2281.
6. D'Agostino, R.B. Jr.; Rubin, D.B. Estimating and using propensity scores with partially missing data. *J. Am. Stat. Assoc.* **2000**, *95*, 749–759.
7. Hill, J.; Reiter, J. Interval estimation for treatment effects using propensity score matching. *Stat. Med.* **2006**, *25*, 2230–2256.
8. Lunceford, J.K.; Davidian, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Stat. Med.* **2004**, *23*, 2937–2960.
9. Qu, Y.; Lipkovich, L. Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Stat. Med.* **2009**, *28*, 1402–1414.
10. Rosenbaum, P.R. Propensity score. In *Encyclopedia of Biostatistics*. 2nd Ed.; Armitage, P.; Colton, T., Eds.; Wiley: New York, 2005, 4267–4272.
11. Rosenbaum, P.R.; Rubin, D.B. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* **1984**, *79*, 516–524.
12. Rubin, D.B. Estimating causal effects from large data sets using propensity scores. *Ann. Int. Med.* **1997**, *127*, 757–763.
13. Rubin, D.B. Using propensity score to help design observational studies: Application to the tobacco litigation. *Health Serv. Outcomes Res. Methodol.* **2001**, *2* (3), 169–188.
14. Rubin, D.B. The design versus the analysis of observational studies for causal effects: Parallel with the design of randomized trials. *Stat. Med.* **2007**, *26* (1), 20–36.
15. Rubin, D.B. For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2008**, *2* (3), 808–840.
16. Austin, P. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* **2009**, *28* (25), 3083–3107.
17. Stuart, E.A. Matching methods for causal inference: A review and a look forward. *Stat. Sci.* **2010**, *25* (1), 1–21.
18. Yue, L.Q. Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies. *J. Biopharm. Stat.* **2007**, *17* (1), 1–13.
19. Li, H.; Yue, L.Q. Statistical and regulatory issues in Non-randomized medical device clinical studies. *J. Biopharm. Stat.* **2008**, *18* (1), 20–30.
20. Yue, L.Q. Regulatory considerations in the design of comparative observational studies using propensity scores. *J. Biopharm. Stat.* **2012**, *22* (6), 1272–1279.
21. Levenson, M.S.; Yue, L.Q. Regulatory issues of propensity score methodology: Application to drug and device safety studies. *J. Biopharm. Stat.* **2013**, *23* (1), 110–121.
22. Yue, L.Q.; Lu, N.T.; Xu, Y. Designing pre-market observational comparative studies using existing data as controls: Challenges and opportunities. *J. Biopharm. Stat.* **2014**, *24* (5), 994–1010.
23. Li, H.; Mukhi, V.; Lu, N.T.; Xu, Y.; Yue, L.Q. A note on good practice of objective propensity score design for pre-market nonrandomized medical device studies with an example. *Stat. Biopharm. Res.* **2016**, *8* (3), 282–286.
24. Ahmed, A.; Husain, A.; Love, T.; Gambassi, G.; Dell'Italia, L.; Francis, G.; Gheorghiade, M.; Allman, R.; Meleth, S.; Bourge, R. Heart failure, chronic diuretic use, and increase in mortality and hospitalization: An observational study using propensity score methods. *Eur. Heart J.* **2006**, *27* (12), 1431–1439.
25. Austin, P.C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. Behav. Res.* **2011**, *46*, 399–424.
26. Austin, P.C.; Grootendorst, P.; Anderson, G.M. A comparison of the ability of different propensity score models to

- balance measured variables between treated and untreated subjects: A Monte Carlo study. *Stat. Med.* **2007**, 26 (4), 734–753.
27. Imai, K.; King, G.; Stuart, E.A. Misunderstandings between experimentalists and observationalists about causal inference. *J. Roy. Stat. Soc. Ser. A.* **2008**, 171 (2), 481–502.
 28. Austin, P.C. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Med. Decis. Making.* **2009**, 29 (6), 661–677.
 29. Rubin, D.B.; Thomas, N. Matching using estimated propensity scores: Relating theory to practice. *Biometrics* **1996**, 52 (1), 249–264.
 30. Hill, J.L.; Reiter, J.P.; Zanutto, E.L. A comparisons of experimental and observational data analyses. In *Applied Bayesian Modeling and Causal Inference From an Incomplete-Data Perspective*; Gelman, A.; Meng, X.-L., Eds.; Wiley: New York, 2004, 44–56.
 31. Mitra, R.; Reiter, J.P. A comparison of two methods of estimating propensity scores after multiple imputation. *Stat. Methods Med. Res.* **2016**, 25 (1), 188–204.

Proportion of Treatment Effect

Cong Chen

Merck & Co., Inc., West Point, Pennsylvania, U.S.A.

Hongwei Wang

Merck & Co., Inc., Rahway, New Jersey, U.S.A.

Abstract

A proportion of treatment effect is defined as a putative surrogate endpoint as a measure of ascertaining the magnitude of the causal pathway accounted for by the treatment effect on the surrogate biomarker. This entry focuses on the impact of the proportion treatment effect has on determining surrogate endpoint biomarkers.

INTRODUCTION

A surrogate endpoint of a clinical trial is a surrogate biomarker intended to supplant a clinical endpoint, characteristic, or variable that directly reflects how a patient feels or functions, or how long a patient survives.^[1] Examples of “true” clinical endpoints are death, stroke, myocardial infarction, loss of vision, progression to AIDS, and other events that cause a substantial reduction in quality of life. Since true clinical endpoints do not occur often, clinical trials that use them as primary endpoints are usually large and may take years to complete. During the development of a pharmaceutical product, it is often unfeasible or impractical to conduct such trials, and surrogate endpoints have to be considered to save time and cost. However, a surrogate biomarker has to be validated before it can be accepted as a surrogate endpoint.

Surrogate endpoints have been of clinical interest for decades, but it was not until Prentice published a seminal paper in 1989 that formal statistical validation started.^[2] Prentice defined a surrogate endpoint as “a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint.” This definition is stringent and impracticable. To overcome the problem, Freedman, Graubard, and Schatzkin^[3] proposed the so-called “proportion of treatment effect” (PTE), explained by a putative surrogate endpoint as a measure of ascertaining the magnitude of the causal pathway accounted for by the treatment effect on the surrogate biomarker. The proposal was made in the context of logistic regression models, and was subsequently extended to Cox regression models by Lin, Fleming, and De Gruttola,^[4] which is arguably more pertinent to surrogate validation and will be used as the primary platform for presentation.

The conventional procedure for a PTE estimation fits data into two working models separately to estimate the treatment effects before and after adjustment for the covariate corresponding to the surrogate biomarker of interest. The construction of confidence intervals for the PTE under the conventional procedure lacks support by standard statistical software such as SAS, and could be very computationally demanding even if the support is available in the future. To overcome this problem, Chen, Wang, and Snapinn^[5] proposed a procedure to simplify the computation. Under the new procedure, the treatment effects before and after adjustment for the covariate are simultaneously estimated from a single model. More important than saving computational effort, the idea of using a single model can also be effectively applied to multiple-covariate analysis for the decomposition of overall treatment effect and for the comparison of PTE among several surrogate biomarkers.

STATISTICAL CONCEPT AND METHODS

PTE Estimation Based on Two Separate Models

Consider survival data generated from a clinical trial with repeated measurements of a surrogate biomarker over time for each patient. Suppose there are two treatment groups (new treatment and placebo, say), and R is the treatment indicator, which takes value 1 for the new treatment and 0 for placebo. Denote by X the surrogate biomarker, and by $X(t)$ the measurement of X at time t . The PTE calculation starts with the following two Cox regression models:

$$\lambda_1(t | R) = \lambda_{10}(t) \exp\{\alpha_1 R\} \quad (1)$$

$$\lambda_2(t | R, X) = \lambda_{20}(t) \exp\{\alpha_2 R + \beta X(t)\} \quad (2)$$

where $\lambda_{10}(t)$ and $\lambda_{20}(t)$ are unspecified baseline hazard functions, and α_1 , α_2 , and β are unknown regression coefficients.

The instantaneous event rate is assumed to be proportional to $\exp\{\alpha_1 R\}$ under the first model, and is assumed to be proportional to $\exp\{\alpha_2 R + \beta X(t)\}$ under the second model. Note that α_1 estimates the overall treatment effect before any adjustment and α_2 estimates the treatment effect after the adjustment for the biomarker effect; the PTE explained by the biomarker is naturally defined to be $1 - \alpha_2/\alpha_1$. This definition is very intuitive and appealing to medical researchers.

After data are fit to the two models separately and the estimates of (α_1, α_2) are available, the estimation of PTE is straightforward. The PTE estimates the magnitude of the surrogate effect. As to what extent of surrogacy PTE suggests, it is arguably more important to construct the associated confidence intervals than fixating upon the point estimate. For example, a 95% confidence interval for PTE of (0.40, 1.05) suggests that the surrogate biomarker under consideration explains at least 40% of the treatment effect. To construct the confidence interval, we need the covariance matrix of the estimates for (α_1, α_2) . However, because the two models in general do not hold simultaneously, there is a lot of difficulty in the estimation of the covariance matrix. To address the issue, Lin, Fleming, and De Gruttola^[4] proposed a sandwich type covariance matrix to account for model misspecifications. But it is not supported by standard statistical software such as SAS, and, due to its jackknife nature, could be very computationally demanding even if the support is available in the future.

PTE Estimation Based on a Single Model

Chen, Wang, and Snapinn^[5] proposed to estimate PTE solely based on a single model so that the statistical property of the PTE estimate could be easily characterized. The idea is to separate information between the treatment indicator and the surrogate biomarker, and recover the unadjusted treatment effect directly from the second model, which is assumed to be the complete model. Rewrite the model as

$$\lambda_2(t | R, X) = \lambda_{20}(t) \exp\{(\alpha_2 + c\beta)R + \beta(X(t) - cR)\} \quad (3)$$

where c is a tuning parameter to retain the unadjusted treatment effect in $\alpha_2 + c\beta$, or in other words, to make the adjusted surrogate biomarker $X(t) - cR$ balanced between the two treatment groups.

Let $(\hat{\alpha}_2, \hat{\beta})$ be the maximum partial likelihood estimates of (α_2, β) from fitting data to the complete model. They asymptotically follow a multivariate normal distribution,^[6] with estimated mean of $(\hat{\alpha}_2, \hat{\beta})$ and estimated covariance

matrix $\begin{pmatrix} \hat{\sigma}_R^2 & \hat{\rho}\hat{\sigma}_R\hat{\sigma}_X \\ \hat{\rho}\hat{\sigma}_R\hat{\sigma}_X & \hat{\sigma}_X^2 \end{pmatrix}$. An estimate of c is obtained from

$\text{cov}(\hat{\alpha}_2 + c\hat{\beta}, \hat{\beta}) = 0$ and is given by $\hat{c} = -\hat{\rho}\hat{\sigma}_R/\hat{\sigma}_X$. This is equivalent to using the estimated information matrix as the basis to measure dependence as suggested by Zografos^[7] for information separation. Because $\hat{\alpha}_2 + \hat{c}\hat{\beta}$ and $\hat{\beta}$ are uncorrelated, the information in R with respect to clinical outcomes would be largely separated from that in $X(t) - \hat{c}R$ (the following section provides more details on information separation). Therefore, the unadjusted treatment effect is estimated to be $\hat{\alpha}_2 + \hat{c}\hat{\beta}$, and the corresponding estimate of the PTE is

$$\hat{P} = 1 - \frac{\hat{\alpha}_2}{\hat{\alpha}_2 + \hat{c}\hat{\beta}} = \frac{\hat{c}\hat{\beta}}{\hat{\alpha}_2 + \hat{c}\hat{\beta}} \quad (4)$$

Given $\hat{c}$, $\hat{P}$ is a ratio of two random variables with independent asymptotic normal distributions. From the estimated covariance matrix of $(\hat{\alpha}_2, \hat{\beta})$, the confidence interval of PTE can be easily constructed by using the delta-method, Fieller's method,^[8] or parametric bootstrap.^[9]

Proportion of treatment effect estimates based on the single-model approach and the two-model approach are not expected to be the same in general. This is because when $\beta \neq 0$, $\alpha_2 + c\beta$ is not equal to α_1 even when $X(t) - cR$ is independent of R .^[10] However, it is of interest to make a comparison. To illustrate, we draw an analogy to a problem in regular linear regressions. Let $Y = \alpha_2 R + \beta X + \varepsilon$ be the true model with ε independent of (R, X) . Suppose that R and X are standardized variables, both with mean 0 and variance 1, and $\text{corr}\{R, X\} = \theta$. Let $Y = \alpha_1 R + \omega$ be the reduced model with covariate X omitted. Then the least squares estimate of α_1 from the reduced model would converge to $\text{corr}\{R, Y\} = \alpha_2 + \theta\beta$, exactly the same as the unadjusted treatment effect derived from the true model using the information separation procedure as just discussed. Therefore, the PTE estimates converge to the same quantity under the two approaches.

The idea of using a single model for PTE estimation can be easily extended to multivariate analysis. Consider a Cox regression model with m time-varying covariates,

$$\lambda(t | R, X) = \lambda_0(t) \exp\left\{\alpha R + \sum_{k=1}^m \beta_k X_k(t)\right\} \quad (5)$$

where $X_k(t)$ is the measurement of the surrogate biomarker X_k at time t , $k = 1, \dots, m$. Similar to the single covariate case, the above model can be rewritten as

$$\lambda(t | R, X) = \lambda_0(t) \exp\left\{\left(\alpha + \sum_{k=1}^m c_k \beta_k\right)R + \sum_{k=1}^m \beta_k (X_k(t) - c_k R)\right\} \quad (6)$$

where $(c_1, \dots, c_m)$ are tuning parameters to retain the unadjusted treatment effect in $\alpha + \sum_{k=1}^m c_k \beta_k$, of which $c_k \beta_k$ is the

contribution from $X_k(t)$, and α is the contribution from sources other than $\{X_1(t), \dots, X_m(t)\}$. After data are fit to the Cox regression model^[5] and the maximum partial likelihood estimates of $(\alpha, \beta_1, \dots, \beta_m)$, denoted by $(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_m)$, together with an estimate of the covariance matrix are obtained, the tuning parameters are estimated from the following system of equations;

$$\text{Cov}\left(\hat{\alpha} + \sum_{k=1}^m c_k \hat{\beta}_k, \hat{\beta}_j\right) = 0 \quad \text{for } j = 1, \dots, m \quad (7)$$

Once the estimates of $(c_1, \dots, c_m)$, denoted by $(\hat{c}_1, \dots, \hat{c}_m)$, are available, the PTE by X_j is

$$\hat{c}_j \hat{\beta}_j / \left(\hat{\alpha} + \sum_{k=1}^m \hat{c}_k \hat{\beta}_k \right).$$

It is clear from the above discussion that PTE, by definition, can also take negative values or values >1 . This is not surprising, because overall treatment effect can be viewed as the net effect from all sources, which could take values of possibly opposite signs. Consequently a PTE estimate is just the estimate of a ratio that might take any value. When it is negative, it means the new treatment has a detrimental effect through the biomarker compared to the placebo; when it is >1 , it means that the new treatment has some other effects in the opposite direction to that reflected in the biomarker under evaluation.

Information Separation

The concept of information separation used in the single-model approach apparently has wide applications unconfined to Cox regression models. It can be directly adapted to any statistical models capable of providing point estimates of parameters of interest and corresponding estimate of the covariance matrix. We use linear regression models to provide a more detailed explanation behind the concept.

Let R be an indicator of treatment group that takes value 1 or -1 , each with probability $1/2$, and X be an m -dimensional column vector of standardized covariates, each with mean 0 and variance 1. Denote by F the covariance matrix of X . Assume that $E\{X|R\} = AR$. Suppose that the response variable Y is associated with (R, X) in the following linear regression model:

$$Y = \alpha R + \beta' X + \varepsilon \quad (8)$$

where ε is a standardized random error independent of (R, X) . Observe that, by definition, $\text{cov}(R, X) = A$. The adjusted treatment difference between $R = 1$ and $R = -1$ is naturally 2α as per the definition of R , and the unadjusted treatment difference is

$$E\{Y | R = 1\} - E\{Y | R = -1\} = 2(\alpha + \beta' A) \quad (9)$$

Denote by $(\hat{\alpha}, \hat{\beta}')$ the least squares estimator of (α, β') , and by M the covariance matrix of $(\hat{\alpha}, \hat{\beta}')$. Based on standard linear regression theory,

$$M = \begin{pmatrix} \text{var } \hat{\alpha} & \text{cov}(\hat{\alpha}, \hat{\beta}') \\ \text{cov } \hat{\beta} & \text{cov}(\hat{\beta}, \hat{\beta}') \end{pmatrix} = \frac{1}{n} \begin{pmatrix} 1 & A' \\ A & F \end{pmatrix}^{-1} \quad (10)$$

where n is the sample size. Under the information separation procedure, model (8) is rewritten as

$$Y = (\alpha + \beta' c) R + \beta' (X - cR) + \varepsilon \quad (11)$$

where c again is a vector of tuning parameters solved from $\text{cov}(\alpha + \beta' c, \hat{\beta}) = 0$. By applying standard matrix theory^[11] to (10),

$$\begin{aligned} c &= -\text{cov}(\hat{\beta}', \hat{\beta})^{-1} \text{cov}(\hat{\alpha}, \hat{\beta}) \\ &= -\text{cov}(\hat{\beta}', \hat{\beta})^{-1} \left(-\text{cov}(\hat{\beta}', \hat{\beta})^{-1} A \right) = A \end{aligned} \quad (12)$$

Therefore, the unadjusted treatment effect under the information separation procedure is the same as in (9), and so are the corresponding PTEs. However, the natural approach in (9) involves the calculation of $\text{cov}(R, X)$, which could be cumbersome when X is a time-varying covariate. The information separation procedure is apparently more computationally flexible, especially in complex or non-linear models. Besides, unlike the information separation approach, the natural approach does not take advantage of the optimality property inherent in an information matrix.

CONCLUSIONS

The road to the supreme status of surrogacy has always been bumpy. Many surrogate biomarkers have been investigated in the past, but only a few have passed scientific scrutiny and are now well accepted as surrogate endpoints. In cardiovascular diseases, systolic blood pressure and low-density lipoprotein (LDL or bad cholesterol) are among a handful of commonly accepted surrogate endpoints for clinically important cardiovascular events after a long search. Other examples include intraocular pressure for preservation of vision and HIV RNA reductions for AIDS progression-free survival. A new drug can now be approved solely based on reduction of systolic blood pressure or LDL. But, it does not necessarily mean that the drug is always safe to use or marketing claims regarding the ultimate endpoint are allowed. For example, since it caused fatal rhabdomyolysis, cerivastatin (a cholesterol-lowering drug) was withdrawn from the market in August 2001 after years in use. Examples of weakness and limitation of surrogate endpoints can also be found in Ref. [12].

As was also pointed out in Ref. [12] one reason why many surrogate endpoints need further confirmation is that there might be multiple pathways of intervention. While current statistical methods mainly (for exceptions, see e.g., Ref. [13]) aim at testing whether a surrogate biomarker qualifies to be a surrogate endpoint one at a time, the single-model approach has the advantage of simultaneously evaluating multiple surrogate biomarkers and helping identify multiple causal pathways. The framework of using a single model for PTE analysis or surrogacy validation makes relevant statistical problems very tractable, and a spectrum of research work is warranted. A similar idea of using the single-model approach can also be found in Ref. [14].

Just like Prentice's definition of surrogate endpoint, though intuitive and appealing the PTE idea is not immune to criticism. Proportion of treatment effect estimation could be very sensitive to the handling of errors in surrogate biomarker measurements^[15] and other modelling aspects. Researchers have also found that definite conclusions can seldom be drawn based on PTE estimation, as it usually has a wide confidence interval. But for the confidence interval to be narrow, clinical trials have to be large and the study duration has to be long, which is exactly what researchers try to avoid in the first place. Despite these caveats, interesting clinical conclusions such as the impact of serum uric acid on cardiovascular endpoint are routinely drawn based on PTE analysis.^[16] To partly overcome the dilemma with PTE, Buyse et al.^[17] and Molenberghs et al.^[18] proposed to validate surrogacy by combining data from similar trials, and to use two new quantities to replace PTE as measures of surrogacy. An alternative definition of PTE can also be found in Ref. [19].

Statistical methods such as PTE analysis provide powerful tools for surrogacy validation. However, they have a fundamental limitation. They can help identify associations or lack thereof, but statisticians need input from medical science to put forth mechanisms for possible explanation of whether or not the identified relationship is independent or causal. Statisticians do not work alone in surrogacy validation, just as in any applied statistical field. As new surrogate biomarkers are routinely discovered in bioscience and technology, statistical methods have to be improved to meet the growing need. This could continue to make the related statistical field one of the hottest in statistics for years to come.

REFERENCES

1. Biomarker Definition Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharm. Ther.* **2001**, 69, 89–95.
2. Prentice, R.L. Surrogate endpoint in clinical trials: definition and operational criteria. *Stat. Med.* **1989**, 8, 431–440.
3. Freedman, L.S.; Graubard, B.I.; Schatzkin, A. Statistical validation of intermediate endpoints for chronic diseases. *Stat. Med.* **1992**, 11, 167–178.
4. Lin, D.Y.; Fleming, T.R.; De Gruttola, V. Estimating the proportion of treatment effect explained by a surrogate marker. *Stat. Med.* **1997**, 16, 1515–1527.
5. Chen, C.; Wang, H.W.; Snapinn, S. Proportion of treatment effect (PTE) explained by a surrogate marker. *Stat. Med.* **2003**, 22, 3449–3459.
6. Anderson, P.K.; Gill, R.D. Cox's regression model for counting processes: a large sample study. *Ann. Stat.* **1982**, 4, 1100–1120.
7. Zografos, K. On a measure of dependence based on Fisher's information matrix. *Commun. Stat. Theory Meth.* **1998**, 27, 1715–1728.
8. Fieller, E.C. The biological standardization of insulin. *J. R. Stat. Soc.* **1940**, 7 (Suppl), 1–15.
9. Efron, B. *An Introduction to the Bootstrap*; Chapman & Hall: New York, 1993.
10. Struthers, C.A.; Kalbfleisch, J.D. Misspecified proportional hazard models. *Biometrika* **1986**, 73, 363–369.
11. Searle, S.R. *Linear Models*; John Wiley & Sons: New York, 1971.
12. Fleming, T.R.; DeMets, D.L. Surrogate end points in clinical trials: are we being misled. *Ann. Intern. Med.* **1996**, 125, 605–613.
13. Xu, J.; Zeger, S.L. The evaluation of multiple surrogate endpoints. *Biometrics* **2001**, 57, 81–87.
14. Li, Z.; Meredith, M.P.; Hoseyni, M.S. A method to assess the proportion of treatment effect explained by a surrogate endpoint. *Stat. Med.* **2001**, 20, 3175–3188.
15. Chen, C.; Snapinn, S. Separate modeling approaches for survival analysis of a covariate with inaccurate measurements. *J. Biopharm. Stat.* **2004**, 14 (2), 375–387.
16. Høiegggen, A.; Chen, C.; Dahlöf, B.; et al., The impact of serum uric acid on cardiovascular outcomes in the LIFE study. *Kid. Int.* **2004**, 65, 1041–1049.
17. Buyse, M.; Molenberghs, G.; Burzykowski, T.; Renard, D.; Geys, H. Statistical validation of surrogate endpoints: problems and proposals. *Drug Inf. J.* **2000**, 34, 447–454.
18. Molenberghs, G.; Buyse, M.; Geys, H.; Renard, D.; Burzykowski, T.; Alonso, A. Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clin. Trials.* **2002**, 23, 607–625.
19. Wang, Y.; Taylor, J.M.G. A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics* **2002**, 58, 803–812.

Proportional Hazards Regression Model

Bee Leng Lee

Department of Mathematics, San José State University, San José, California, U.S.A.

Jason Fine

Department of Statistics and Department of Biostatistics and Medical Informatics,
University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.

Abstract

The proportional hazards regression model has become, by a wide margin, the most used regression model for analyzing time-to-event data because of its flexibility and availability in software packages. This entry breaks down each part of this model in detailed terms.

Keywords: Proportional hazards; Regression model; Residuals.

INTRODUCTION

An objective of many medical studies is the prognostic assessment of time to an event, denoted T , such as the remission of symptoms, the progression of a disease, or the onset of death. Clinical studies can be designed to obtain information concerning possible prognostic factors, and a physician can usually decide as to which factors may be relevant. However, a meaningful summary of the data to reveal the underlying relationship between the time to the occurrence of the event and the various predictors can only be achieved through a comprehensive statistical analysis.

Time to an event is often called a *failure time*, *survival time*, or *event time* in the literature. It is the interval between a clearly defined time origin and the occurrence of an event of interest. The time origin need not be at the same calendar time for each individual. An important consideration in the choice of a time origin is that, subject to any known differences on individual characteristics, all individuals should be as comparable as possible at their time origin. Most *randomized clinical trials*, for example, have staggered entry; the event time of a subject is usually measured from the day the subject is randomized to receive one of the treatment protocols. It is customary to refer to the event of interest as *failure* in clinical trials because it is usually some negative individual experience such as death.

A mechanism that can lead to an incomplete observation of an event time is *censoring*. Broadly speaking, a censored observation arises when the exact event time is unknown, but can only be determined to lie within a certain interval. The most commonly encountered form of a censored observation is one in which the observation on a patient begins at the randomization, and ceases before the event of interest is realized, when the patient is lost to follow-up. Such observations are said to be *right-censored*

because the incomplete nature of the observation occurs in the right tail of the time axis. For example, in a study designed to evaluate survival after a confirmed diagnosis of HIV, some subjects may die from AIDS or AIDS-related causes before the end of the study, yielding exact survival times; others may withdraw before the end of the study; still others may be alive at the end of the study. The last two kinds of observations are right-censored observations because the survival status is only known until the last contact.

OVERVIEW

Other censoring mechanisms that can occur in practice include *left censoring*, when all that is known is that the event has already occurred before a certain time, and *interval censoring*, when the only information available is that the event occurs within some interval.

Whatever the cause of censoring, a key assumption is that the censoring is *noninformative* about the event of interest, that is, the cause of censoring is unrelated to the impending event. A systematic withdrawal of high-risk or low-risk patients from a study is an example of informative censoring, which may result in inaccurate statistical inference about the survival experience. Kalbfleisch and Prentice^[1] (Section 5.2) provide further discussion of noninformative censoring.

Censoring is an inherent feature in time-to-event data, which creates special problems in the analysis of such data. In particular, it renders the usual linear regression analysis inappropriate because the response, the event time, is incompletely determined for some observations. Regression models proposed for prognostic modeling of time-to-event data generally involve assumptions on the

hazard function, which describes the way in which the instantaneous risk of an individual experiencing an event changes over time. One such model is the *proportional hazards regression model* introduced by Cox.^[2] It has become, by a wide margin, the most used regression model for analyzing time-to-event data because of its flexibility and availability in software packages.

THE MODEL

Let $\mathbf{x} = (x_1, x_2, \dots, x_p)^T$ be p measured covariates (or predictor variables) on a given individual, where the symbol T denotes transpose. These can be patient characteristics, such as gender or race, or these can be factors of particular interest, such as the treatment received during a drug trial. The coding of categorical variables follows the usual rules in ordinary regression analysis. It is assumed, for now, that the values of the covariates for any given individual remain constant over time. Such covariates are called *fixed-time* or *time-independent* covariates.

The *Cox model*, in its simplest form, postulates that the hazard function at time t for an individual with covariate values $\mathbf{x}$ is:

$$\begin{aligned} \lambda(t | \mathbf{x}) &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} pr\{t \leq T \leq t + \Delta t | T \geq t, \mathbf{x}\} \\ &= \lambda_0(t) \exp(\beta^T \mathbf{x}) \end{aligned} \quad (1)$$

where $\lambda_0(t)$ is an arbitrary, non-negative function of time; $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$ is a vector of unknown regression coefficients; and $\beta^T \mathbf{x} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$. The function $\lambda_0(t)$ characterizes the dependence of the hazard function on time. The subscript 0 on $\lambda_0(t)$ reflects that it is the hazard function for an individual with all covariate values equal to zero; for this reason, $\lambda_0(t)$ is referred to as the *baseline hazard function*. The function $\exp(\beta^T \mathbf{x})$ characterizes how the hazard function changes as a function of covariate values. Notice that there is no intercept parameter (or constant term) in $\beta^T \mathbf{x}$, for it is subsumed in $\lambda_0(t)$. In prognostic modeling, attention is usually focused on inference about specific regression coefficients or on $\beta^T \mathbf{x}$ (which is sometimes known as the *risk score* or *prognostic index* for an individual with covariate values $\mathbf{x}$), while $\lambda_0(t)$ is often considered a nuisance parameter.

Besides the hazard function, a function of central interest in the analysis of time-to-event data is the *survival function*, which gives the probability that an individual has yet to experience the event of interest at a specified time. Under the Cox model, the survival function at time t for an individual with covariate values $\mathbf{x}$ has the form:

$$S(t | \mathbf{x}) = [S_0(t)]^{\exp(\beta^T \mathbf{x})} \quad (2)$$

where $S_0(t)$ is the *baseline survival function*, that is, the survival function for an individual with all covariate values

equal to zero. The baseline survival function is related to the baseline hazard function through the equation:

$$S_0(t) = \exp[-\Lambda_0(t)] \quad (3)$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u) du$ is called the *cumulative* (or *integrated*) *baseline hazard function*.

The Cox model can be derived as a special case of a class of linear transformation models (Ref. [3], Section 4.7) given by:

$$g(T) = -\beta^T \mathbf{x} + \varepsilon \quad (4)$$

where $g(\cdot)$ is an unknown monotone transformation and ε is a random variable with completely specified distribution. The Cox model results when $g(\cdot) = \log \Lambda_0(\cdot)$ and ε has a unit extreme value distribution.

The Cox model is a *semiparametric model* in the sense that the covariate effect is modeled parametrically, while no assumption is made regarding the functional form of the time effect. As we shall see later, that $\lambda_0(t)$ is left completely unspecified does not sacrifice the ability to obtain a reasonably good estimate of β . The log-linear form $\exp(\beta^T \mathbf{x})$ assumed for the covariate effect can be replaced by any known function of $\mathbf{x}$ and β that is nonnegative, such as a *cubic spline*.^[4] A variety of probability models can be used to specify $\lambda_0(t)$ to yield fully parametric proportional hazards regression models. For example, a *Weibull distribution* with scale parameter α and a shape parameter γ is flexible enough to accommodate increasing ($\gamma > 1$), decreasing ($\gamma < 1$), or constant ($\gamma = 1$) time effect: $\lambda_0(t) = \alpha \gamma t^{\gamma-1}$. The utility of the Cox model stems from the fact that in most applications, the form of $\lambda_0(t)$ is either unknown or complex, and if a parametric model is chosen incorrectly, it could lead to estimators of the wrong quantity. The Cox model is a *robust* model, in the sense that if the correct parametric model is, say, Weibull, then the use of the Cox model typically will give results comparable to those obtained using a Weibull model.

In the Cox model, as expressed in Eq. 1, the regression coefficient β_k (where $k = 1, 2, \dots$, or p) describes the change in the hazard function on a logarithmic scale for a change in the corresponding covariate x_k of one unit, while all other covariates are kept fixed. We note that the ratio of hazards for an individual with covariate values $\mathbf{x}$ compared with an individual with covariate values $\mathbf{z}$ is:

$$\frac{\lambda(t | \mathbf{x})}{\lambda(t | \mathbf{z})} = \frac{\lambda_0(t) \exp(\beta^T \mathbf{x})}{\lambda_0(t) \exp(\beta^T \mathbf{z})} = \exp \left\{ \sum_{k=1}^p \beta_k (x_k - z_k) \right\} \quad (5)$$

which is constant through time. Suppose x_k is a dichotomous covariate such as smoking status, which is often coded so that smokers correspond to a value of 1 and non-smokers correspond to a value of 0. If the value of the coefficient is $\beta_k = \log 2$, then the interpretation is that at any

given point in time, smokers are twice as likely to experience the event of interest (such as death) compared to nonsmokers, all else being equal. Part of the appeal of the Cox model lies in such interpretations of the hazard ratio as the *relative risk* of exposure to a potential hazard. The evaluation of relative risk allows, for example, an insurance company to determine if smokers should pay a higher premium than nonsmokers.

Although the primary application of the Cox model is the consideration of relative risk, the absolute (or overall) risk is of interest when the purpose is prediction. For example, the risk of death before a certain age is essential for calculating the actual premiums of smokers (or nonsmokers). The hazard function $\lambda(t|x)$ is useful for characterizing the absolute risk, for it allows us to determine the probability that the event of interest will occur in the next instant, given that the event has not occurred. A related quantity is the *cumulative hazard function*: $\Lambda(t|x) = \int_0^t \lambda(u|x) du$ which describes the accumulated risk up until time t .

ESTIMATION

A major contribution of the article by Cox^[2] is the enunciation of a clever method for estimating the regression coefficients separately from the arbitrary baseline hazard function. This method is appropriate for both uncensored and right-censored data. Based on a sample of n individuals, the observed data consist of the triplets (t_i, δ_i, x_i) , $i = 1, 2, \dots, n$, where the subscript i denotes individual i ; t_i is the “observation” time, which is either the event time or the censoring time, depending on whether the event of interest had occurred by time t_i ; δ_i is the event indicator, that is, $\delta_i = 1$ if the event of interest had occurred (so that t_i is the event time) and $\delta_i = 0$ otherwise (so that t_i is the censoring time); and $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ is the covariate values associated with the individual.

ESTIMATION OF REGRESSION COEFFICIENTS FOR DISTINCT EVENT TIME DATA

If there are no ties between the event times, the estimation of β is based on the *partial likelihood* introduced by Cox,^[2] which has the form:

$$L(\beta) = \prod_{i=1}^n \left[\frac{\exp(\beta^T x_i)}{\sum_{k \in R(t_i)} \exp(\beta^T x_k)} \right]^{\delta_i} \quad (6)$$

where $R(t_i) = \{k : t_k \geq t_i\}$, termed the *risk set* at time t_i , is the set of individuals at risk of experiencing the event of interest an instant before time t_i . Here the summation in the denominator is over all individuals whose observation

times are at least t_i . Note that the ordering of the observation times, rather than the actual observation times, is used in the calculation of $L(\beta)$. This renders the estimation procedure more robust to outliers in the observation times, as only the rankings of the times are used.

To understand the partial likelihood, it is useful to note that individuals whose event times were censored contribute factors of 1 (which contains no information about β) to the partial likelihood. The product in Eq. 6 is therefore effectively over all individuals who experienced the event of interest. For each of these individuals, the contribution to the partial likelihood represents the conditional probability of the individual experiencing the event given that there was an event at that time: among those individuals still at risk:

$$\begin{aligned} & \Pr\{\text{individual } i \text{ experiences the event at time } t_i | R(t_i) \\ & \quad \text{and an event at time } t_i\} \\ &= \frac{\Pr\{\text{individual } i \text{ experiences the event at time } t_i | R(t_i)\}}{\Pr\{\text{an event at time } t_i | R(t_i)\}} \quad (7) \\ &= \frac{\lambda_0(t_i) \exp(\beta^T x_i)}{\sum_{k \in R(t_i)} \lambda_0(t_i) \exp(\beta^T x_k)} \end{aligned}$$

Notice that $\lambda_0(t_i)$ cancels in Eq. 7, allowing inferences on β not involving the nuisance parameter.

The term “partial likelihood,” rather than (complete) likelihood, is used because the formula in Eq. 6 does not explicitly consider the probabilities for those individuals who were censored. Nevertheless, as discussed by Johansen,^[5] the Cox partial likelihood can be derived as a *profile likelihood* for β from the full likelihood. First, the full likelihood is set up as a function of both β and λ_0 :

$$\begin{aligned} L_F(\beta, \lambda_0) &= \prod_{i=1}^n [\lambda(t_i | x_i)]^{\delta_i} S(t_i | x_i) \\ &= \prod_{i=1}^n [\lambda_0(t_i) \times \exp(\beta^T x_i)]^{\delta_i} \exp[-\Lambda_0(t_i) \exp(\beta^T x_i)] \end{aligned}$$

We then maximize $L_F(\beta, \lambda_0)$ as a function of λ_0 only, using a discretization similar to the *Nelson-Aalen estimate* of the hazard function for data without covariates.^[6] Let $\hat{\lambda}_\beta$ denote the maximizer of $L_F(\beta, \lambda_0)$ for a given β . Note that $\hat{\lambda}_\beta(t) = 0$ except at observed event times. The profile likelihood is defined as the function: $pl(\beta) = \sup_{\lambda_0} L_F(\beta, \lambda_0) = L_F(\beta, \hat{\lambda}_\beta)$. The Cox partial likelihood is proportional to $pl(\beta)$.

The estimate for β is found by maximizing Eq. 6, or, equivalently, the logarithm of Eq. 6. This involves solving for $\beta_1, \beta_2, \dots, \beta_p$ in the set of p nonlinear equations:

$$\frac{\partial \log L(\beta)}{\partial \beta_j} = 0, \quad j = 1, 2, \dots, p \quad (8)$$

which can be done numerically by using the Newton-Raphson method of iteration (or some other iterative method). The solution to Eq. 8 is called the *maximum partial likelihood estimate* (MPLE), which we shall denote as $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)^T$. The estimator of the covariance matrix of $\hat{\beta}$ is obtained in the same manner as in most maximum likelihood estimation applications. Let $\mathbf{I}(\beta)$ denote the negative of the matrix of second-order partial derivatives of the partial likelihood function $L(\beta)$, that is, the entry in the j th row and k th column of $\mathbf{I}(\beta)$ is:

$$\frac{\partial^2 L(\beta)}{\partial \beta_k \partial \beta_j} \quad (9)$$

where $i = 1, 2, \dots, p$ and $j = 1, 2, \dots, p$. $\mathbf{I}(\beta)$ is known as the *observed information matrix*. The estimator of the covariance matrix of $\hat{\beta}$ is the inverse of the information matrix evaluated at $\hat{\beta}$, $\hat{Var}(\hat{\beta}) = [\mathbf{I}(\hat{\beta})]^{-1}$. For the j th estimated coefficient $\hat{\beta}_j$ (where $j = 1, 2, \dots, p$), the estimated standard error is given by the positive square root of the j th diagonal element in $\hat{Var}(\hat{\beta})$.

Algorithms for the estimation of β are available in many software packages. The procedure Phreg in SAS and the function coxpk in S-Plus, for example, provide the MPLE of β and its standard error.

ESTIMATION OF REGRESSION COEFFICIENTS IN THE PRESENCE OF TIES

The partial likelihood in Eq. 6 was developed under the assumption of continuous event times, which rules out the possibility of ties among the event times (ties among the censoring times are allowed). Modifications to the partial likelihood are therefore needed when the data contain tied event times. An estimation of β is otherwise carried out in the same manner as in the case of distinct event times.

Kalbfleisch and Prentice^[1] derived an exact expression for the partial likelihood for data containing tied event times. The basis for its construction is to assume that ties have been introduced into the data by the grouping of continuous times into intervals. This is a reasonable assumption because the recording of time is often rounded to the nearest unit of measurement. Let $t_{(1)} < t_{(2)} < \dots < t_{(m)}$ denote the distinct, ordered event times. Of course, $m \leq n$. Suppose that, of the n individuals in the sample, d_i individuals realized their events at $t_{(i)}$, where $i = 1, 2, \dots, k$ and $\sum_{i=1}^k d_i = n$; in other words, there are d_i ties at $t_{(i)}$. The tied values could have been observed in any one of the $d_i!$ orderings of their values. The exact partial likelihood function is obtained by modifying the denominator of Eq. 6 to include each of these possible orderings for each $t_{(i)}$. Both SAS and S-Plus include the option (discrete and exact, respectively) of using the exact partial likelihood. If the number of ties at any event time is large, the amount of computation time required for evaluating the exact partial likelihood can

be prohibitive, and approximations to the exact partial likelihood are often used instead.

Breslow^[7] considered each of the tied observations as distinct and derived an approximation to the exact partial likelihood function that is identical to Eq. 6. In other words, the partial likelihood function in Eq. 6 is applied without modification to data containing tied event times. If the number of ties at each event time is small relative to the size of the corresponding risk set, then Breslow's approximation works well; otherwise, it could produce "estimated β coefficients too close to 0 in absolute value"^[8] because of multiple counting of individuals who realized the event in the denominator of Eq. 6. An alternative approximation is because of Efron.^[9] Hertz-Picciotto and Rockhill^[10] presented a simulation study, which shows that the Efron approximation performed far better than the Breslow approximation, even with heavy ties and fairly small sample sizes.

To understand the difference between the two approximations, one may view the problem of ties in terms of population average effects in longitudinal data analysis. This has been studied by Zeger et al.^[11] who demonstrated that the population average model, when compared to subject-specific model, consistently underestimated the magnitude of the coefficients for the fixed effects in a linear logistic model. Hertz-Picciotto and Rockhill^[10] suggest that if tied data represent population averages, then the Breslow approximation could be considered an unadjusted population average model, in which the downward bias is attributed to the loss of information that results from the grouping of tied observations in the numerator of the likelihood. The Efron approximation adjusts the denominator downward to compensate for such loss of information.

The Breslow approximation is the default method in SAS. The Efron approximation is the default method in S-Plus and is implemented in SAS as the efron option.

ESTIMATION OF SURVIVAL FUNCTION

It may be of interest to estimate the survival function for an individual with a given set of covariate values, say x_0 . The expression for the survival function given in Eq. 2 suggests that once an estimate of the regression coefficients is available, all we need is an estimate of the baseline survival function. The Breslow^[7] estimate of $S_0(t)$ is derived through an estimate of $\Lambda_0(t)$, given by:

$$\hat{\Lambda}_0(t) = \sum_{t_i \leq t} \sum_{k \in R(t_i)} \frac{\delta_i}{\exp(\hat{\beta}^T x_k)} \quad (10)$$

Note that $\hat{\Lambda}_0(t)$ is a step function with jumps at the observed event times only; in other words, when graphed as a function of time, $\hat{\Lambda}_0(t)$ is flat between two consecutive event times. If there are no covariates, $\hat{\Lambda}_0(t)$ reduces

to the Nelson-Aalen estimate; in that case, all the exponential terms in the denominator of Eq. 10 are 1. The Breslow estimate of $S_0(t)$ is $\hat{S}_0^B(t) = \exp[-\hat{\Lambda}_0(t)]$, which is obtained using the expression for $S_0(t)$ shown in Eq. 3.

An alternative estimate of $S_0(t)$, because of Kalbfleisch and Prentice,^[12] is an extension of the *Kaplan-Meier estimate* to allow for covariates. In the particular case where the event times are distinct, the Kalbfleisch-Prentice estimate is given by:

$$\hat{S}_0^{KP}(t) = \prod_{i: t_i \leq t} \left[1 - \frac{\delta_i \exp(\hat{\beta}_i^T x_i)}{\sum_{k \in R(t_i)} \exp(\hat{\beta}_k^T x_k)} \right]^{\exp(\hat{\beta}_i^T x_i)} \quad (11)$$

Each term in the product can be regarded as an estimate of the probability that an individual survives through the interval defined by two consecutive observed event times.^[13] In the no-covariate case, $\hat{S}_0^{KP}(t)$ reduces to the Kaplan-Meier estimate. When there are tied event times, $\hat{S}_0^{KP}(t)$ is found numerically by solving a system of equations.

Making use of Eq. 2, the estimated survival function for an individual with covariate values x_0 is: $\hat{S}(t|x_0) = [\hat{S}_0(t)]^{\exp(\hat{\beta}^T x_0)}$ where $\hat{S}_0(t)$ is either $\hat{S}_0^B(t)$ or $\hat{S}_0^{KP}(t)$. The SAS procedure phreg uses $\hat{S}_0^{KP}(t)$ by default, and $\hat{S}_0^B(t)$ can be chosen with the method=ch option. S-Plus has both $\hat{S}_0^B(t)$ and $\hat{S}_0^{KP}(t)$ available in the function surv.fit, along with the variance estimates.

The variance of $\hat{S}(t|x_0)$ is complex (e.g., see Ref. [14]); in the no-covariate case, it equals that for the Kaplan-Meier estimate.

ASSESSMENT OF MODEL ADEQUACY

Central to the assessment of model adequacy in any setting is an appropriate definition of a residual. In the usual regression setup, a natural definition of the residual is the difference between the observed value of the response and that predicted by the fitted model. In the regression analysis of time-to-event data, however, the definition of the residual is not as clear-cut, due in part to the possibility of incompletely observed response, and in part to the hazard formulation. Several residuals have been proposed for examining the different aspects of the fit of the Cox model.

Schoenfeld Residuals

A key assumption of the Cox model is proportional hazards, that is, the hazard ratio of individuals with distinct covariate values is constant over time. The validity of this assumption is vital to the interpretation of the regression coefficients. *Schoenfeld residuals*^[15] are useful for assessing the proportional hazards assumption for each covariate as well as several covariates simultaneously. There is not a single value of the residual for each individual, but a set of

values—one for each covariate included in the fitted Cox model.

The Schoenfeld residual for the i th individual on the k th covariate is given by:

$$s_{ik} = \delta_i \{x_{ik} - \bar{x}_k(\hat{\beta}, t_i)\} \quad (12)$$

where $\hat{\beta}$ is the MPLE, and: $\bar{x}_k(\hat{\beta}, t_i) = \sum_{j \in R(t_i)} x_{jk} \exp(\hat{\beta}^T x_j) / \sum_{j \in R(t_i)} \exp(\hat{\beta}^T x_j)$ is a weighted average of the values of the k th covariate over those individuals still at risk at the event time t_i of the i th individual, with $\exp(\hat{\beta}^T x_j)$ as the weight assigned to the j th individual in the risk set. Note that if the event time of the i th individual is censored, then $s_{ik} = 0$ for all k .

Some authors call the Schoenfeld residual in Eq. 12 the *score residual*, because s_{ik} can be derived as the i th individual's contribution to the first derivative of the log partial likelihood function with respect to β_k :

$$\frac{\partial \log L(\beta)}{\partial \beta_k} = \sum_{i=1}^n \delta_i \{x_{ik} - \bar{x}_k(\beta, t_i)\} \quad (13)$$

That is, s_{ik} can be regarded as an estimate of the i th term in the above summation, obtained by substituting $\hat{\beta}$ for β . Because $\hat{\beta}$ is the solution to the equations obtained by setting Eq. 13 to zero, the Schoenfeld residuals must sum to zero. Furthermore, in large samples, the expected value of s_{ik} is zero and the residuals are uncorrelated with one another. For this reason, the Schoenfeld residuals, when plotted against the event times for each covariate in the fitted model, should scatter in a nonsystematic way about the horizontal axis.

Grambsch and Therneau^[16] suggest that scaling the Schoenfeld residuals by an estimator of its variance yields a residual with greater diagnostic power than the unscaled residuals. Most software packages compute an approximate scaled Schoenfeld residuals, obtained by multiplying $m\hat{Var}(\hat{\beta})$ to the Schoenfeld residuals, where m is the number of uncensored event times.

Numerous tests of the proportional hazards assumption have been proposed. Grambsch and Therneau^[16] showed that most of these are essentially tests for zero slopes in a generalized linear regression of scaled Schoenfeld residuals on chosen functions of time, such as the logarithm or the rank of observed event times, or the Kaplan-Meier estimate of the survival function. For example, the test proposed by Harrell,^[17] which uses the correlation between the unscaled residuals and the rank of the event times, is similar to a trend test applied to the plot of the scaled residuals vs. the rank of the event times. The function cox.zph in S-Plus implements the score test developed by Grambsch and Therneau^[16] using the scaled Schoenfeld residuals, which is based on an extension of the Cox model to include time-varying regression coefficients. An overview of this test, as well as illustrations, can be found in Hosmer and Lemeshow^[18] (Section 6.3).

Martingale Residuals

The construction of *martingale residuals*^[19,20] is based on a *counting process*^[6,21] representation of the Cox model. The method relies heavily on probability theory and stochastic processes, and will not be discussed here. The residual for the i th individual is defined as:

$$m_i = \delta_i - \exp(\hat{\beta}^T x_i) \hat{\Lambda}_0(t_i) \quad (14)$$

which can be interpreted as the difference between the observed number of events for the individual over the interval $(0, t_i)$, and the predicted number of events in that interval under the fitted Cox model. To see this, note that the observed number of events is one if the event time is uncensored, and zero if censored, which is equal to the value of δ_i . The second term in Eq. 14 is an estimate of the cumulative hazard function at time t_i , which can be viewed as the expected number of events over $(0, t_i)$.

Despite the similarity, in terms of interpretation, between the martingale residuals and the residuals encountered in linear regression analysis, the distribution of the martingale residuals does not aid in the assessment of the overall fit of the Cox model in the usual manner. For example, a common diagnostic tool in linear regression analysis is the plot of the residuals against the fitted values; if the model is adequate, the points should scatter in a nonsystematic way. This plot is useless for martingale residuals because they are negatively correlated with the fitted values.

Martingale residuals are useful for investigating the appropriate functional form for a given covariate using an assumed Cox model for the remaining covariates. Technical details are provided in Fleming and Harrington^[6] (Section 4.5). Therneau and Grambsch^[8] illustrate the use of martingale residuals with examples.

EXTENSIONS

Time-Dependent Covariates

At the introduction of the proportional hazards regression model, Cox noted that the covariates for any given individual may be *time-dependent* or *time-varying*, that is, the values of the covariates may change over time. The resulting model is often referred to as the *extended Cox model*.

Time-dependent covariates are usually classified as being either *internal* or *external*.^[11] An internal time-dependent covariate is one whose value is specific to the individual, such as blood pressure of a patient recorded at different times. It is measured only as long as the individual is still under observation, and thus its value carries information about the event time of the individual. In contrast, an external time-dependent covariate is one whose value at a particular time does not require individuals to

be under observation, such as air pollution index for a particular geographic area recorded at different times.

Let $x_i(t) = (x_{i1}(t), x_{i2}(t), \dots, x_{ip}(t))^T$ denote the covariate values at time t for the i th individual under study, where $i = 1, 2, \dots, n$. The notation is general in the sense that if a particular covariate, say the k th covariate, is time-independent then $x_{ik}(t) = x_{ik}(0) = x_{ik}$. The partial likelihood function for data involving time-dependent covariates is similar to the expression in Eq. 6, with the appropriate alterations of x_i to $x_i(t)$. The estimation of the regression coefficients also proceeds as described in “Estimation.” It is not necessary to differentiate between internal and external time-dependent covariates in the construction of the partial likelihood or the estimation of the regression coefficients. The distinction, however, becomes important when the estimation of the survival function is of interest. For internal time-dependent covariates, $S(t|x_0(s), s \leq t) = 1$ because observing $x_0(t)$ indicates that the event has not occurred [because the individual could not have realized the event if $x_0(t)$ is observed]. The hazard function no longer bears a relationship to the survival function when internal time-dependent covariates are involved. Thus even though it is mathematically easy to extend the Cox model to include time-dependent covariates, one should examine the nature of any time-dependent covariate when interpreting the fitted model, especially when the absolute risk is of interest.

Frailty Models

In medicine, it is a common observation that the natural course of a disease or a treatment effect varies from individual to individual. This condition is often referred to as *individual heterogeneity*. It is rarely possible in practice to include all the relevant covariates to account for individual heterogeneity, either because we are not aware that there exist covariates to be included, or the heterogeneity is manifest only indirectly. The unmeasured covariates, when present in the usual linear model settings, are called *random effects*. Such covariates are called *frailties* in survival analysis, based on the idea that individuals have different frailties, and those who are most frail will experience an adverse event earlier than the others.^[22]

The Cox model can be extended in a natural way to account for individual heterogeneity, which may lead to more accurate estimates of β and $\lambda_0(t)$, and hence improvements in the predictions of relative and absolute risks. Let x_{i0} denote the value of an unmeasured covariate (the frailty) for the i th individual. The hazard function, conditional on both x_{i0} and $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$, becomes:

$$\lambda(t|x_i, x_{i0}) = \lambda_0(t) \exp(\beta_0 x_{i0} + \beta^T x_i) = w_i \lambda_0(t) \exp(\beta^T x_i) \quad (15)$$

where $w_i = \exp(\beta_0 x_{i0})$. Because $w_1, w_2, \dots, w_n$ are unobserved, they are viewed as a sample from some family of distributions with possibly unknown parameters.

The extended model, as expressed in Eq. 15, is called a *proportional hazards frailty regression model*.

Much of the research in proportional hazards frailty regression has dealt with the choice of a statistical distribution for the frailty. It is popular to let the frailty have a gamma distribution with a mean equal to one and an unknown variance γ . The resulting model is equivalent to the linear transformation model given in Eq. 4, with $g(\cdot) = -\log \Lambda_0(\cdot)$ and with ε having a $\text{Pareto}(\gamma)$ distribution. Other distributions proposed in the literature include the inverse Gaussian distribution,^[23] the positive stable distribution,^[24] and the log-normal distribution.^[25] Note that if the frailty distribution has zero variance, then Eq. 15 reduces to the Cox model; otherwise, a model with non-proportional hazards results. This implies that an estimate of the variance of the frailty distribution, which we shall denote as γ , is useful for assessing the proportional hazards assumption in the Cox model.

An estimation of model^[15] with γ assumed known has been studied extensively (e.g., Refs. [26,27]). These methods are not applicable when γ is unknown. Recently, Kosorok et al.^[28] developed a unified approach to estimation in model^[15] that allows for general frailty distributions with unspecified γ . The proposed algorithm permits inference about γ and β separately from the arbitrary $\lambda_0(t)$, as in the Cox partial likelihood approach. When the model is correctly specified, the procedure is semi-parametric efficient; moreover, the bootstrap gives valid inferences even under misspecification.

REFERENCES

1. Kalbfleisch, J.D.; Prentice, R.L. *The Statistical Analysis of Failure Time Data*; John Wiley: New York, 1980.
2. Cox, D.R. Regression models and life-tables. *J. R. Stat. Soc. B* **1972**, *34*, 187–220.
3. Bickel, P.J.; Klaassen, C.A.J.; Ritov, Y.; Wellner, J.A. *Efficient and Adaptive Estimation for Semiparametric Models*; Springer-Verlag: New York, 1998.
4. Hastie, T.; Tibshirani, R. Generalized additive models (with discussion). *Stat. Sci* **1986**, *1*, 295–318.
5. Johansen, S. An extension of Cox's regression model. *Int. Stat. Rev* **1983**, *51*, 258–262.
6. Fleming, T.R.; Harrington, D.P. *Counting Processes and Survival Analysis*; Wiley: New York, 1991, 4.
7. Breslow, N.E. Covariance analysis of censored survival data. *Biometrics* **1974**, *30*, 1–89.
8. Therneau, T.M.; Grambsch, P.M. *Modeling Survival Data: Extending the Cox Model*; Springer-Verlag: New York, 2000, 1–49.
9. Efron, B. The efficiency of Cox's likelihood function for censored data. *J. Am. Stat. Assoc* **1977**, *72*, 557–565.
10. Hertz-Picciotto, I.; Rockhill, B. Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics* **1997**, *53*, 1151–1156.
11. Zeger, S.L.; Liang, K.-Y.; Albert, P.S. Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **1988**, *44*, 1049–1060.
12. Kalbfleisch, J.D.; Prentice, R.L. Marginal likelihoods based on Cox's regression and life model. *Biometrika* **1973**, *60*, 267–278.
13. Collett, D. *Modelling Survival Data in Medical Research*; Chapman and Hall: London, 1994, 96–97.
14. Klein, J.P.; Moeschberger, M.L. *Survival Analysis: Techniques for Censored and Truncated Data*; Springer-Verlag: New York, 1997, 260.
15. Schoenfeld, D. Partial residuals for the proportional hazards regression model. *Biometrika* **1982**, *69*, 239–241.
16. Grambsch, P.M.; Therneau, T.M. Proportional hazards tests in diagnostics based on weighted residuals. *Biometrika* **1994**, *81*, 515–526.
17. Harrell, F.E. The phglm Procedure. *SAS Supplemental Library User's Guide, Version 5*; SAS Institute: Cary, NC, 1986.
18. Hosmer, D.W.; Lemeshow, S. *Applied Survival Analysis: Regression Modeling of Time to Event Data*; Wiley: New York, 1998.
19. Barlow, W.E.; Prentice, R.L. Residuals for relative risk regression. *Biometrika* **1988**, *75*, 65–74.
20. Therneau, T.; Grambsch, P.M.; Fleming, T.R. Martingale-based residuals for survival models. *Biometrika* **1990**, *77*, 147–160.
21. Andersen, P.K.; Borgan, O.; Gill, R.D.; Keiding, N. *Statistical Models Based on Counting Processes*; Springer-Verlag: New York, 1993.
22. Aalen, O.O. Heterogeneity in survival analysis. *Stat. Med* **1988**, *7*, 1121–1137.
23. Hougaard, P. Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika* **1984**, *71*, 75–83.
24. Hougaard, P. Survival models for heterogeneous populations derived from stable distributions. *Biometrika* **1986**, *73*, 387–396.
25. McGilchrist, C.; Aisbett, C.W. Regression with frailty in survival analysis. *Biometrics* **1991**, *47*, 461–466.
26. Cheng, S.C.; Wei, L.J.; Ying, Z. Analysis of transformation models with censored data. *Biometrika* **1995**, *82*, 835–845.
27. Scharfstein, D.O.; Tsiatis, A.A.; Gilbert, P.B. Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data. *Lifetime Data Anal* **1998**, *4*, 355–391.
28. Kosorok, M.R.; Lee, B.L.; Fine, J.P. Robust inference for proportional hazards frailty regression models. *Ann. Stat* **2004**, *32*, 1448–1491.

FURTHER READING

- Bull, K.; Spiegelhalter, D.J. Tutorial in biostatistics: Survival analysis in observational studies. *Stat. Med* **1997**, *16*, 1041–1074.
- Harrell, F.E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*; Springer-Verlag: New York, 2001.
- Hougaard, P. *Analysis of Multivariate Survival Data*; Springer-Verlag: New York, 2000.
- Lee, E.T. *Statistical Methods for Survival Data Analysis*; John Wiley: New York, 1992.
- Oakes, D. *Biometrika centenary: Survival analysis*. *Biometrika* **2001**, *88*, 99–142.

Protocol Development

Robert D. Chew

Pfizer Global Research and Development, Groton, Connecticut, U.S.A.

Abstract

The goal of this entry is to abstract the guidance published by the European Medicines Agency and the U.S. Food and Drug Administration as a convenience for the statistical practitioner faced with the task of describing in the protocol the planned analysis of a clinical trial and also to offer some suggestions based on practical experience in the field.

Keywords: Keywords: Clinical trial; Protocol development; Statistics.

INTRODUCTION

The goal of this entry is to abstract the guidance published by the European Medicines Agency (EMA) and the U.S. Food and Drug Administration (FDA)^[1-7] as a convenience for the statistical practitioner faced with the task of describing in the protocol the planned analysis of a clinical trial and also to offer some suggestions based on practical experience in the field.

The point of view here is confirmatory. Protocols are of course also produced for research trials, proof-of-concept trials, and the like. It is reasonable that such protocols might be less inclusive in terms of elements or less specific in terms of given elements. However it is also reasonable to suppose that any protocol might be made better, and certainly made no worse, by explicit consideration of the features of confirmatory trials during its preparation.

NATURE AND PURPOSE OF THE PROTOCOL FOR A CLINICAL TRIAL

The protocol for a clinical trial is “[a] document that describes the objective(s), design, methodology, statistical considerations, and organization of a trial.”^[1] It describes the medical conjecture, the subject population, the treatment regimens, the data gathering, and the analysis and evaluation of the data.

The primary purpose of this entry is to describe the function of the protocol as the portrayal of the statistical design and data analysis for a medical experiment. However, a clinical trial is a complex enterprise. The clinical-trials statistician should be aware that the protocol also serves a number of administrative and organizational purposes. As a document, the protocol is a contract between the sponsor and the investigator.^[1] The protocol identifies the agencies and individuals involved in the

project and describes their responsibilities and obligations (e.g., for serious adverse event reporting, availability of records for inspection). It serves as the primary means for the Institutional Review Board to ensure the ethical purpose and conduct of the trial and should contain a statement that the trial will be conducted in compliance with good clinical practice and the applicable regulatory requirements. The protocol summarizes the nature of the known and potential risks and benefits to the subjects. It requires informed consent and adherence to the principles in the Declaration of Helsinki. Certain sections of a protocol will typically contain instructions concerning the handling of specimens for laboratory analysis or the administration of special (or even routine) medical procedures.

Most institutions, and nearly all pharmaceutical companies, have standard outlines and naming/numbering conventions for protocols. Sections generally include, but are by no means limited to, those in the following abbreviated list. There are a number of reasonable ways to organize these topics within the document, and different sponsors will choose different conventions. However, nearly everyone covers at least the following topics somewhere in their protocols.

- *Introduction:* Describes the test article, synthesizes what is known about its effects and previous research, and motivates the current project.
- *Objective:* Sharpens the medical conjecture.
- *Subject selection criteria:* Defines the subject eligibility criteria in such a way as to link to a target population of prospective users of the treatment and protect the safety of study subjects.
- *Study medication:* Thoroughly describes the test and control treatments, their packaging, and the mechanism for treatment masking.
- *Plan and procedures:* Lays out the schedule of visits, doses, tests, and evaluations to be performed on the subjects by the investigative staff, rules around discontinuing

subjects, reporting of serious adverse events, and administration of concomitant or rescue medication.

- *Efficacy and safety endpoints and criteria:* Describes the primary and secondary efficacy and safety endpoints and how those data are going to be collected and scored. Study monitoring and data-collection quality controls might be described in this section.
- *Data analysis and statistical considerations:* Defines the primary and secondary variables; describes the primary and secondary analyses and the statistical design considerations.
- *Committees:* Identifies various committees associated with the trial and states their mission and scope. Examples would include data monitoring committees, endpoint classification committees, steering committees, and executive committees.^[2,4]
- *Pharmacoeconomic investigations:* Contains many of the same elements as the main protocol; often inserted as a subprotocol.
- *Pharmacokinetic/pharmacodynamic investigations:* Increasingly, population pharmacokinetics, exposure-response, and the like are included as secondary analyses in confirmatory trials; often inserted as a subprotocol.

Sponsors should have established procedures for generating, approving, and amending protocols. These procedures should state who is to be involved in the process and what role they are to play. The ICH Section E6 on Good Clinical Practice^[1] asserts that

[t]he sponsor should utilize qualified individuals (e.g., biostatisticians, clinical pharmacologists, and physicians) as appropriate, throughout all stages of the trial process, from designing the protocol and CRFs and planning the analyses to analyzing and preparing interim and final clinical trial/study reports.

It is a sound practice to require that the statistical section be contributed by a qualified professional statistician.

CONTENT OF THE STATISTICAL SECTION OF A PROTOCOL FOR A CLINICAL TRIAL

The advice of the ICH is to prespecify the analysis: “The extent to which the procedures in the protocol are followed and the primary analysis is planned a priori will contribute to the degree of confidence in the final results and conclusions of the trial.”^[3] Hence, the a priori specification of the analysis is an important function of the statistical analysis section of a protocol.

Primary Variable(s) and Analyses

“[T]he protocol should identify the primary [efficacy variables] with an explanation of why they were

chosen.”^[4] “Only results from analyses envisaged in the protocol (including amendments) can be regarded as confirmatory.”^[3]

Because most clinical trials collect a number of different types of variables, there will typically be a number of different analyses conducted on the data from a given trial. “The protocol should make clear a distinction between aspects of a trial that will be used for confirmatory proof and the aspects that will provide data for exploratory analysis.”^[3] One means of drawing this distinction, and of addressing the multiplicity issue besides, is via the definition, in the statistical section of the protocol, of a primary variable and a primary analysis. The terms *primary*, *secondary*, *exploratory*, and the like are probably not susceptible to rigorous definition. However, one sense in which those in the clinical-trials field tend to understand them is in terms of this confirmatory/exploratory distinction. The *primary* variable (or endpoint) is the variable in terms of which the investigators intend either to confirm or to refute the conjecture upon which the trial is based. Because it is often possible to analyze a given variable in a number of different ways, it is also recommended to specify a primary (stochastic) model and a primary analysis strategy for the primary variable.

What, then, is entailed in the specification of the primary variable and its analysis in the statistical analysis section of a protocol? Pharmaceutical trial statisticians have generally found the following set of considerations to be useful.

Primary Variable Definition

This is the precise mathematical definition of the primary outcome variable. In many instances this variable is not raw data but is derived according to some algorithm from the raw data. Stating that algorithm avoids ambiguity; e.g., say, “the return-visit-3 measurement minus the baseline-visit measurement as a percentage of the baseline” as opposed to, “the percentage change from baseline.” Mathematical notation is appropriate and useful; but it should be supplemented by a verbal description. Note that, in this example, there is an implicit understanding of the definition of *baseline*. It is usually helpful to state explicitly the definition of *baseline* and other time-point dependencies. They may not be as obvious as one might suppose.

Primary Statistical Objective

This involves a translation of the primary objective of the investigation into statistical terms: a statistical model and an accompanying hypothesis testing or estimation objective. If it is a hypothesis test, as is frequently the case, the null and alternative hypotheses should be stated in terms of the primary statistical model.

“It is vital that the protocol of a trial designed to demonstrate equivalence or noninferiority contain a clear statement that this is its explicit intention.”^[3,5]

Primary Analyses

These are the analyses, or statistics, that will produce the answers to the primary investigational conjecture. The ICH E9 document recommends that “all the principal features of the proposed confirmatory analysis of the primary variables”^[3] be laid out in the protocol. It is not unreasonable that different, equally competent analysts might make differing judgments about what is an appropriate level of analytical detail to include in a protocol, i.e., what constitutes a “principal feature” of the analysis. Further detail can be given in an accompanying analysis plan document. Most practitioners would state in the protocol the main statistic upon which the primary analysis will be based, e.g., least-squares means, log-likelihood ratio, some ranks-based test. If not pre-specified, covariate adjustments, variable transformations, and methods to handle dropouts or missing data can lead to controversy later. It is appropriate and useful to employ mathematical notation.

DECISION RULES

A description of the statistical decision space and decision rules is essential. “The protocol should... designate the pattern of significant findings or other method of combining information that would be interpreted as supporting efficacy.”^[4]

Given that one has calculated a set of statistics, one must state a priori how this set of statistics will be used to evaluate the primary conjecture. Whether hypothesis tests are one- or two-sided, their significance (alpha) levels, and/or outcomes that would lead the experimenters to conclude that treatments are equivalent, would be included with this topic. If a number of statistical hypotheses are being evaluated, give the map from the subsets of possible outcomes to the overall substantive conclusions.

Here, the rubric of analysis emphasizes the arithmetic, whereas the rubric of decision rule emphasizes the inference to be made from that arithmetic. It is useful, at least, to do the thinking around the analytic strategy in these terms.

ERROR CONTROL

The method for controlling the error rates must be set up. Particularly when a number of hypotheses are being tested, the method for controlling the Type I error rates, and at what level they will be controlled, should be stated. The method of management of the joint coverage probabilities

for sets of interval estimates would be given. Type II error rates might also be addressed.^[6]

ANALYSIS SUBSETS

A description of the subset of subjects whose data will contribute to the primary analysis should be included. This is where such terms as “Intention-to-treat,” “per-protocol,” and the like come in. It is advisable not to assume that all readers have the same understanding of these terms, but rather to explicitly spell out the membership requirements for (at least) the primary analysis subset.

DESIGN AND NUMBER OF SUBJECTS

The statistical aspects of the experimental design should be set out. This section gives the reasoning that justifies the planned number of subjects. Working estimates of anticipated treatment effects and variability would be given, along with their source. Clinically meaningful treatment effects that the study is targeted to detect or, in the case of equivalence trials, exclude, would be stated. If the design is something besides one-period, double-blind, parallel-group (e.g., a crossover or dose escalation), or if the design allows treatment transitions, then the statistical implications of these characteristics should be accounted for.

Sometimes trials are designed to be stopped when a certain amount of information has been accumulated. This is a statistical design feature that may not be obvious and ideally should be fully explained.

In a multicenter study, it may be useful to state the number of subjects planned per site.

SECONDARY VARIABLE(S) AND ANALYSES

“Secondary variables are either supportive measurements related to the primary objective or measurements of effects related to the secondary objective.”^[3] The testing of the primary efficacy conjecture in some subpopulation is a common secondary objective. Secondary endpoints can be defined, and their analyses planned, in an exposition that parallels that for the primary variables: variable definition, statistical model, statistical objective, analytic strategy, and decision rules. The level of detail for secondary analyses is frequently reduced, as compared to the primary analyses, and they can be discussed as a group in order to make the exposition efficient and readable. However, just as for primary analyses, it is obvious that the statement of secondary conjectures and analyses a priori is an important factor in lending credibility to the conclusions.

FURTHER TOPICS

Interim Analyses

The timing, purpose, and scope of planned interim analyses must be stated. If decisions about stopping the trial, extending the trial, or otherwise modifying the design (e.g., by adding or dropping treatment arms) are to be based on interim analyses, then the decision criteria should be described in the protocol. Spelling out comprehensively and in detail the process by which accumulating data will be examined and acted upon (or not) will obviate concerns about bias and scientific validity. This material could be rather extensive. See the entry titled *Interim Analysis* for further advice on what to include.

Randomization and Blinding

A high-level description of the randomization scheme might be included. It is not recommended that the block size be revealed in the protocol. A small proportion of trials have somewhat complex randomization schemes, and in these cases, the description would be appropriately more detailed. If the trial is not of the usual double-blind variety, then it may be helpful to argue for the type of treatment masking that is planned. The point is to address concerns that critics of the trial may raise regarding bias.

Problem Data

If it is possible to anticipate conditions under which data will be invalid or would otherwise create analytical challenges, then it may be worth describing them in the protocol and prespecifying analysis methods to account for the problem data. Outliers, missing observations, premature terminations, losses to follow-up, and protocol violations are some of the general categories of difficulties that can be anticipated. Another common issue is created by low-enrolling centers in a multicenter study where the analysis is blocked by center. Some sponsors may prefer to base their analytical approach to problem data on a treatment-masked review of the data conducted prior to any actual analysis. If methods for handling the problem data result from such a blinded pre-analysis review, and those methods are documented, then concerns around potential bias do not arise.

Because they can be rather intricate, details regarding the handling of problem data in the analyses are better left for the analysis plan rather than the protocol. However, the attention that the missing data problem

has received in recent years has created the expectation that the protocol for any confirmatory trial will include at least a high-level description of a strategy for dealing with missing data.^[7]

The ICH E9 document advises that “[t]he protocol should also specify procedures aimed at minimizing any anticipated irregularities in study conduct that might impair a satisfactory analysis.”^[3] This material may fit more naturally in the plan and procedures section of the protocol than in the statistical analysis section, however.

Safety Data

The planned analyses of safety data should be set out. According to conventions used by many sponsors, this description is terse, emphasizing summarization and display. However, if there are known safety issues, it is clearly in a sponsor's interest to prespecify in some detail an inferential strategy to address them. General analytical issues around safety data include the definition of incidence (per subject, per dose, per subject-year, etc.) and how to treat data from subjects who have no or minimal exposure to the study treatments.

REFERENCES

1. European Medicines Agency. *ICH E6 (R1) Guideline for Good Clinical Practice*; July 1996, <http://www.emea.europa.eu/htms/human/humanguidelines/efficacy.htm>.
2. European Medicines Agency. *Data Monitoring Committee*; July 2005, <http://www.emea.europa.eu/htms/human/humanguidelines/efficacy.htm>.
3. European Medicines Agency. *ICH E9 Statistical Principles for Clinical Trials*; March 1998, <http://www.emea.europa.eu/htms/human/humanguidelines/efficacy.htm>.
4. European Medicines Agency. *ICH E3 Structure and Content of Clinical Study Reports*; December 1995, <http://www.emea.europa.eu/htms/human/humanguidelines/efficacy.htm>.
5. European Medicines Agency. *Choice of a Non-Inferiority Margin*; July 2005, <http://www.emea.europa.eu/htms/human/humanguidelines/efficacy.htm>.
6. European Medicines Agency. *Points to Consider on Multiplicity Issues in Clinical Trials*; September 2002, <http://www.emea.europa.eu/htms/human/humanguidelines/efficacy.htm>.
7. European Medicines Agency. *Missing data in confirmatory clinical trials; release for consultation*; April 2009, <http://www.emea.europa.eu/htms/human/humanguidelines/efficacy.htm>.

P-Values

Stephen Senn

Department of Statistics, University of Glasgow, Glasgow, U.K.

Abstract

A *p*-value is the probability of observing a result as extreme or more extreme than that observed given that the null hypothesis is true. This entry reviews the history of *p*-values in research and discusses some of the issues and criticisms some researchers have in regards to *p*-values.

Keywords: Distribution; Jeffreys-Good-Lindley paradox; *P*-value.

INTRODUCTION

A *p*-value is the probability of observing a result as extreme or more extreme than that observed given that the null hypothesis (H_0) is true. To be able to calculate a *p*-value, one thus requires at least three things: a null hypothesis, a probability model, and an agreed ordering of the possible outcomes, so that those that are more extreme than those actually observed may be identified. This ordering is generally conveniently arranged in terms of a “test statistic” T so that higher or lower values of T , as the case may be, or higher absolute values (as perhaps in two-sided *p*-values) identify more extreme cases. In fact, the requirement to order the outcomes also generally entails a fourth requirement—that all possible relevant outcomes have been defined or restricted. These relevant outcomes form the so-called *sample space*. Usually, because *p*-values are a frequentist concept, the sample space is defined in terms of the possible outcomes that could occur from an infinite repetition of the experiment (e.g., a clinical trial).

In practice, it often turns out to be far from simple to agree what exactly such an infinite repetition would look like. For example, the trial protocol may not have fixed exactly in advance the number of patients on each arm, even if the trial is not a sequential trial; yet except where the trial is a sequential trial, generally no account is taken of this. In the analysis of binary data, a 2×2 table is often formed, and it is common (but not uncontroversial) to condition on both margins, even though one of them will not have been fixed at all in advance and even though the other may only have been fixed approximately.

These and other difficulties, which will be discussed in due course, mean that *p*-values have been a critical failure. However, if frequency of use is any guide, they have also been a huge popular success. It seems that *p*-values, despite having been banned by at least one journal, are here to stay for the foreseeable future.

HISTORY

David,^[1] in an article attempting to trace the first use of statistical terms, could find no use of the term “*p*-value” earlier than that of Brownlee^[2] in 1960. However, similar phrases were certainly used by Fisher^[3] as early as 1922 (e.g., “values of P ”) and the first use of the *concept* is often dated to be at least two centuries earlier than that, with the famous significance test of Arbuthnot.^[4] According to Arbuthnot, “Among innumerable Footsteps of Divine Providence to be found in the Works of Nature, there is a very remarkable one to be observed in the exact Balance that is maintained, between the numbers of Men and Women.” Arbuthnot had data on the christening of male and female children from London for the years 1629–1710. In each of these years, there was an excess of male christening, which he took to reflect an excess of births, which excess he regarded as necessary to compensate for higher male mortality. He then proceeds to calculate the probability of such a result occurring by chance. Observing that there is a finite but small probability of an exact equality of the sexes, Arbuthnot makes a conservative concession to what we would now call the “null hypothesis” by assuming that the probability of an excess of male births in a given year is 0.5. He then argues that the probability of such an excess 82 years in row is $(1/2)^{82}$, “which will be found easily by the Table of Logarithms to be 1/48360 0000 0000 0000 0000 0000,” going on to conclude, “from whence it follows that [it] is Art, not Chance, that governs.”

Arbuthnot’s probability can clearly be interpreted as a *p*-value but it is an atypical representative of the species because Arbuthnot has the most extreme case—all years show an excess of male births.^[5] In fact, Arbuthnot’s probability can be given two further alternative interpretations. Because it only reflects the observed result (and does not include more extreme cases), it is also a likelihood. In fact, as a function of the value of the probability

of the parameter in question—the probability of an excess of males births in a given year, say θ —and assuming a Bernoulli process, we have the likelihood as $L(\theta) = \theta^{82}$, which Arbuthnot merely calculated as $L(1/2)$. Yet another interpretation can be given. If Divine Providence is intervening to maintain the excess of male births, then the likelihood of observing such an excess is 1. Thus the ratio of likelihoods for the observed data given the hypothesis of “chance” (under Arbuthnot’s assumption) to that given the hypothesis of Divine Providence is also given by the probability calculated by Arbuthnot.

An early example of a p -value that incorporates the more usual difficulty of having to be calculated when the most extreme chance has not occurred is given in the prize-winning essay of Daniel Bernoulli^[6] in 1734. This considers, among other matters, whether the then known (six) planets had orbits that were coplanar to a degree that cannot be explained by chance. Bernoulli looks at this question in a number of different ways. One of these compares the planetary orbits to the sun’s equator with the following results:

Mercury	2°56′
Venus	4°10′
Earth	7°30′
Mars	5°49′
Jupiter	6°21′
Saturn	5°58′

The most extreme of these is that of the Earth at 7°30′. Because the maximum possible is 90°, this is $7°30′/90° = 1/12$. Bernoulli then argues that the probability that all six planets have inclinations to the sun’s equator of less than 7°30′ is $(1/12)^6 = 1/2985984$, thus concluding that this is no coincidence.

Bernoulli’s example has more difficulties than Arbuthnot’s example. He does not have the most extreme case and is thus on the horns of a dilemma. Either he would have to conclude that the pattern is uninteresting (the theory of exact coplanar orbits being clearly falsified), or he has to include the more extreme cases, otherwise he would get the same answer whatever the arrangement. For instance, if the precision to which the planetary orbits can be measured is 1 min, or $1′/90° = 1/5400$, then the probability of any observed arrangement is simply $(1/5400)^6 \simeq 1/(2.5 \times 10^{22})$ so that if a small probability of the observed result were a reason for rejecting the null hypothesis, it would be rejected whatever the result. This particular difficulty is one that critics of the p -value have repeatedly stressed.

If we turn to the more recent history of p -values, then Fisher is often credited (or blamed) for their popularity. This appears to be a result of his espousal of significance tests, with which p -values are often regarded as being closely associated. There are several problems with this view, however. The first is that tail area probabilities (for that is what p -values are) were calculated long before Fisher. Indeed,

they were used by both Student^[7] and Pearson,^[8] both of whom were Bayesians, and the latter gives them the modern interpretation of the probability of a result as extreme or more extreme than that observed. The second reason is that it was actually Fisher, together with Yates in their joint tables, who introduced the habit of inverse tabulation using fixed percentage points such as 5%, 1%, etc.^[9] This accords more easily with the practice of noting whether a result is or is not significant at a given significance level than with calculating the tail area probability. The third reason is that there is no unique association between p -values and the Fisherian significance test to the exclusion of the Neyman-Pearson hypothesis testing framework. As Lehmann,^[10] writing within the latter framework, puts it:

“In applications there is usually available a nested family of rejection regions, corresponding to different significance levels. It is then good practice to determine not only whether the hypothesis is accepted or rejected at the given significance level, but also to determine the smallest significance level $\tilde{\alpha} = \tilde{\alpha}(x)$, the significance probability or p -value, at which the hypothesis would be rejected for the given observation.”

Whatever the reason, whoever is or are historically responsible for promoting p -values as an inferential device, their widespread use cannot now be denied. For example, although there have been considerable efforts in recent years by medical statisticians to promote the use of confidence intervals when reporting the results of clinical trials^[11] and although others have suggested how Bayesian approaches might be employed,^[12] it is still the case that the overwhelming proportion of papers published in the medical literature will not only use p -values but will also give them pride of place in both abstract and report. This is true despite the fact that research has shown that physicians do not even understand how p -values are defined, let alone their properties.^[13,14] Some of these properties are examined below.

DISTRIBUTION OF THE P-VALUE

Under the null hypothesis, when based upon a continuous statistic, the p -value has a uniform distribution $U(0,1)$ (see below), so that every value between 0 and 1 is equally likely. As Fisher^[15] noted in 1932, in the fourth edition of his famous book, *Statistical Methods for Research Workers*, it thus follows that minus twice the logarithm of the p -value has a chi-square distribution under the null hypothesis. This can be used to combine the results of independent tests to perform a sort of meta-analysis (to use a more modern term), because if k in such log-transformed p -values is added, the distribution is χ^2_{2k} .

Where a problem is well structured in the sense that a model is specified up to an unknown parameter θ , then likelihood is a far superior device to the p -value for examining the support afforded by the data to different possible

values of θ . Even in the less well-structured case where the model has one or more nuisance parameters, the profile likelihood will form a better way of investigating support for given values of θ . Nevertheless, if for no other reason than that their use is so common, it is useful to get some general feelings for the distribution of the p -value under the family of alternative hypotheses, as it increases understanding of the way p -values behave and thus serves as a warning against rash interpretation. A number of authors have looked at this (e.g., Refs. 16–19), but it is implicit in many older Bayesian criticisms of p -values.^[20,21]

Take a simple but unrealistic case, which can nevertheless act as an approximation to many more realistic cases, where T is distributed normally with known standard error σ_T and unknown mean $E(T) = \theta$. We assume that under H_0 , $\theta = 0$. Suppose, most simply, that we are interested in a one-sided p -value and suppose without further loss of generality that we regard lower values of T as more “extreme.” (In the hypothesis testing framework, this would apply if we had as alternative hypothesis $H_1: \theta < 0$.) As a function of an observed value t of T , the p -value is then defined as $P(t) = P(T < t | \theta = 0)$. In this case, we have:

$$P(T < t | \theta = 0) = P(T/\sigma_T < t/\sigma_T | \theta = 0) = \Phi(t/\sigma_T) = \Phi(z)$$

where $\Phi(\cdot)$ is the distribution function of the standard normal and z is the ratio of statistic to standard error, which will have the standard normal distribution under H_0 . If Z is a random standard normal deviate, then its probability density function (pdf) is $\phi(z)$, where $\phi(\cdot)$ is the pdf of the standard normal. On the other hand, under the alternative hypothesis, we require:

$$\begin{aligned} P(T < t | \theta = \Delta) \\ &= P\{(T - \Delta)/\sigma_T < (t - \Delta)/\sigma_T | \theta = \Delta\} \\ &= \Phi(t/\sigma_T - \Delta/\sigma_T) = \Phi(z - \Delta/\sigma_T). \end{aligned}$$

so that the pdf of Z is:

$$\phi(z - \Delta/\sigma_T) \quad (1)$$

However, if $P = \Phi(z)$ is the p -value, then we have $dz = dP/\phi(z)$, $z = \Phi^{-1}(P)$ so that making the necessary substitutions in Eq. 1, we have the pdf of P as:

$$\frac{\phi\{\Phi^{-1}(P) - \Delta/\sigma_T\}}{\phi\{\Phi^{-1}(P)\}} \quad (2)$$

which is clearly equal to 1 when the null hypothesis is true and hence $\Delta = 0$.

It may be instructive to reparameterize Eq. 2 in terms of the power $\Psi(\Delta, \sigma_T)$ of a significance test associated with a given significance level α . We then have:

$$\Psi(\Delta, \sigma_T) = P\{T/\sigma_T < \Phi^{-1}(\alpha) | \theta = \Delta\} = \Phi\{\Phi^{-1}(\alpha) - \Delta/\sigma_T\}$$

which we may solve for Δ/σ_T to obtain:

$$\Delta/\sigma_T = \Phi^{-1}(\alpha) + \Phi^{-1}(\beta) \quad (3)$$

where $\beta = 1 - \Psi$ is the “Type II error rate.” Substituting Eq. 3 in Eq. 2, we obtain:

$$\frac{\phi\{\Phi^{-1}(P) - [\Phi^{-1}(\alpha) + \Phi^{-1}(\beta)]\}}{\phi\{\Phi^{-1}(P)\}} \quad (4)$$

Fig. 1 illustrates the probability density of the p -value for different values of the type I error rate (“size”) and

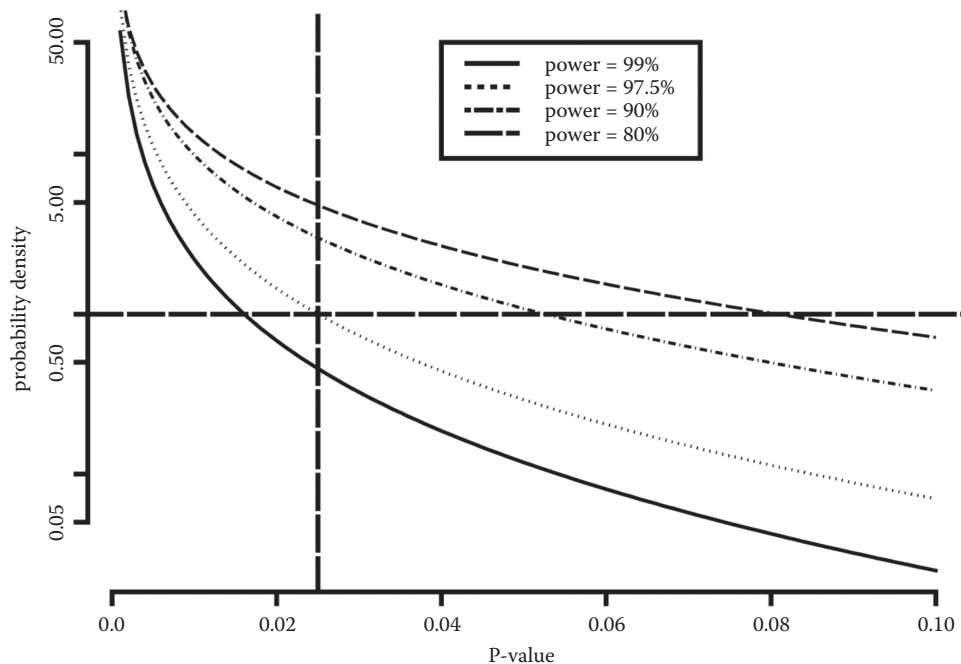


Fig. 1 Probability density (logarithmic axis) of the P -value (for a limited range 0–0.1) for trials with differing power for a size of 2.5%

power. It should be noted that as the power increases, for any p -value, however small, eventually a point will be reached where the result is more probable under the null hypothesis than under the alternative hypothesis. In the family of curves represented in Fig. 1, the break-even point is reached when the power is 97.5% for a test with 2.5% size and the p -value is observed to be exactly 0.025. (For the higher power of 99%, the pdf is actually lower under the alternative hypothesis rather than the null hypothesis.) This is because where the Type I and Type II error rates are identical to the p -value, the point estimate must be halfway between the clinically relevant difference and zero and thus gives equal “support” to these two.

An alternative way of looking at this is to consider the likelihood for a given p -value and predetermined size as a function of the power. We may do this by taking the pdf given by Eq. 4 and considering this as a function of $1 - \beta$ for a given P rather than as a function of P for a given β . (The size and power together merely determine the noncentrality parameter so that this is simply another way of looking at the likelihood as a function of the noncentrality parameter. However, although it is revealing to look at the likelihood of the p -value to gain theoretical understanding, in practice, such analysis should never be performed on the p -value scale.)

A family of curves is given in Fig. 2. The curves all correspond to p -values that would conventionally be considered significant, the least significant being $P = 0.025$. Initially, as the power increases, the likelihood increases. However, eventually a point is reached where the likelihood declines so that a given p -value, however significant, offers less plausible support for the clinically relevant difference

than for the null hypothesis if obtained from a very large trial. This point is at the heart of an important Bayesian criticism of p -values that will be considered in “Some Criticisms.” However, before it is considered, one red herring must be disposed of.

It is sometimes claimed that a significant result is more convincing if it comes from a very large trial than from a small one.^[22] This would appear to contradict the message of Fig. 2, which suggests that a given p -value, especially if marginally significant, might be less convincing from a large trial than a small one. In fact, there is no contradiction. Two different things are being discussed.^[19,23] The one is a statement of the form $P \leq 0.05$; the other is a statement of the form $P = 0.049$ (say). The situation is that in the former case, if this is all that is known (so that, for example, the statement $P \leq 0.05$ does not imply $0.01 \leq P \leq 0.05$ because if $P \leq 0.01$, this would have been stated),^[24] the result is more convincing with increasing sample size. This is because, for a given alternative, the likelihood of the observed result is, effectively, the power function. It is, of course, trivially true that power increases with increasing power. However, for the latter case, the likelihood is not the same as the power and hence the situation illustrated in Fig. 2 is obtained.

SOME CRITICISMS

In this section, some criticisms (mainly Bayesian) of the p -value are considered. The first relates closely to the discussion above.

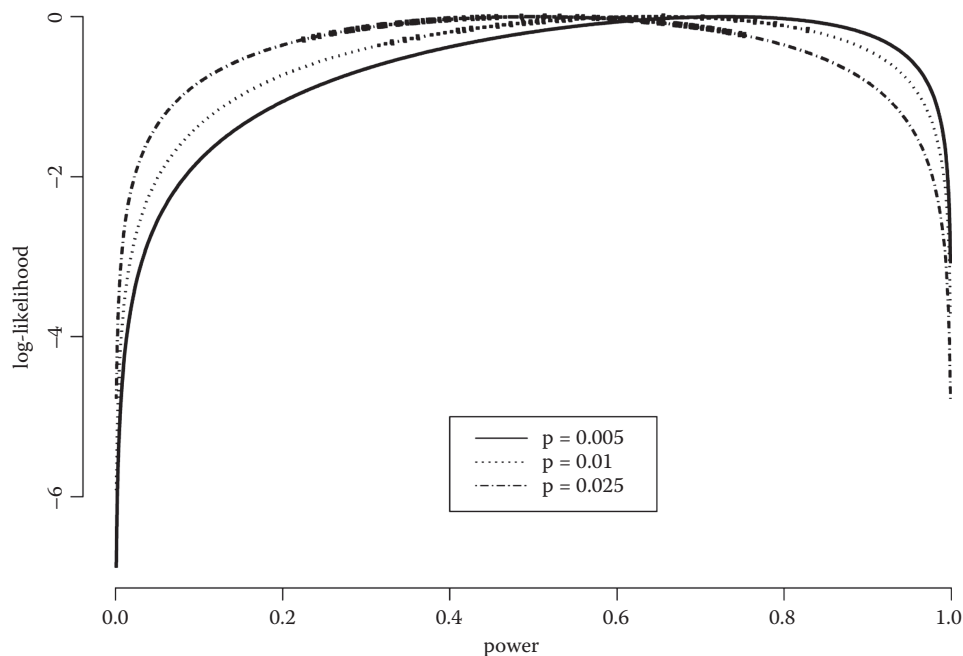


Fig. 2 Log likelihood (scaled to be zero at maximum) as a function of power for a size of 0.025 (one-sided) for three different observed P -values

Jeffreys-Good-Lindley Paradox

As illustrated above, a consideration of the likelihood of a given p -value shows that, eventually, as the sample size increases, more support is given to the null hypothesis than for any given value under the alternative hypothesis. However, because the alternative hypothesis will typically comprise an (often open) interval of values, it does not at all follow that the parameter value under the null hypothesis will be better supported than for every parameter value under the alternative. On the other hand, the null may be better supported than most values under the alternative. This is the basis of the Bayesian criticism that eventually as the sample size increases, a given p -value is less indicative of the falsity of the null and may eventually support it.^[20,21,25] Below, a simplified version of this paradox is presented.

Consider, for example, Eq. 2 not as a function of P but as a function of Δ , so that it forms a likelihood for Δ . Under H_0 , $\Delta = 0$, and Eq. 2 = 1 so that the ratio of the likelihood for some given value under the alternative hypothesis is also equal to Eq. 2. However, $\phi(\cdot)$ is the pdf of a standard normal, and this reaches its maximum when the argument is zero so that because the denominator is constant, Eq. 2 is maximized when:

$$\begin{aligned}\Phi^{-1}(P) - \Delta/\sigma_t &= 0 \\ \Delta &= \sigma_t \Phi^{-1}(P)\end{aligned}\quad (5)$$

at which point it equals:

$$\frac{\phi(0)}{\phi\{\Phi^{-1}(P)\}}\quad (6)$$

This expression depends on P alone and so is constant for a given P , whatever the power of the test, and would appear to give support to what has been referred to as the “alpha postulate”: equal p -values from different trials given equal evidence against the respective null hypotheses. For example, if $P = 0.025$ (one-sided), the value of Eq. 6 is ≈ 6.8 .

However, if one were to adopt a Bayesian perspective, then one would have to consider the prior probability of a given alternative hypothesis. These prior probabilities qua prior probabilities do not change as data are obtained (it is not permitted to change one’s prior beliefs) so that the fact that there is always one possible value under the alternative hypothesis that remains better supported than the null hypothesis is not directly relevant. If this value is itself a priori unlikely—but others which receive little support from the data are more likely given that the alternative hypothesis is true—then the alternative hypothesis as a whole may not be well supported.

If one wishes to choose between what Cox has referred to as a *precise* hypothesis,^[26] which is to say a null hypothesis that specifies some precise value for the unknown

parameter (say $\Delta = 0$), and an *alternative* hypothesis that merely specifies a range for the parameter, say $\Delta > 0$, then a simple ratio of the form given by Eq. 4 is not appropriate. It is necessary to integrate the likelihood over the region of the hypothesis first.

Consider, for the moment, a common two-sided testing problem and suppose now that we believe with probability θ that H_0 : $\Delta = 0$ and hence believe with probability $(1 - \theta)$ that H_1 : $\Delta \neq 0$. Conditional on H_1 , we have a prior distribution for Δ which is $n(0, \gamma^2)$. Suppose that we run a trial that will yield a standard error for our test statistic T of σ_t and that we observe t to within a precision of $\pm \epsilon$. Our predictive distribution for $T = t$ given H_0 is $n(0, \sigma_t^2)$, so that the conditional probability of the observed result is approximately:

$$2\phi\left(\frac{t}{\sigma_t}\right)\frac{\epsilon}{\sigma_t}\quad (7)$$

On the other hand, conditional on H_1 , the distribution is $n(0, \sigma_t^2 + \gamma^2)$ so that we have a conditional probability of:

$$2\phi\left(\frac{t}{\sqrt{\sigma_t^2 + \gamma^2}}\right)\left(\frac{\epsilon}{\sqrt{\sigma_t^2 + \gamma^2}}\right)\quad (8)$$

The ratio of Eqs. 7–8 is the factor:

$$\frac{\phi(t/\sigma_t)}{\phi(t/\sqrt{\sigma_t^2 + \gamma^2})} \frac{\sqrt{\sigma_t^2 + \gamma^2}}{\sigma_t}\quad (9)$$

by which the prior odds in favor of H_0 , $\theta/(1 - \theta)$, must be multiplied in order to obtain the posterior odds. This is the product of two ratios. For a given P value, which implies a given value of t/σ_t , the first term approaches a limit, which is the ratio of the probability density at the observed standardized value to the value at 0. However, the same is not true of the second ratio. As the size of the trial increases, σ_t reduces, but γ^2 is unaffected and so the second term of Eq. 9 can be made arbitrarily large as the trial increases, and hence the factor by which the posterior odds will be increased, arbitrarily large.

VIOLATION OF THE LIKELIHOOD PRINCIPLE

The likelihood principle, originally due to Barnard,^[27] states (roughly) that evidence about an unknown parameter in a model depends on the data only through the likelihood. This means that the stopping rule is irrelevant. However, in applying hypothesis tests to sequential trials, it is usual to adjust the conclusion according to the stopping rule. Thus such sequential analyses (obviously) violate the likelihood principle as do hypothesis tests (at least to the extent that they are given evidential interpretations) and hence p -values.

This is a simple example from Pawitan.^[28] Compare a binomial and a negative binomial experiment to examine an unknown probability θ of success X , each with eight successes out of 10 trials. For the binomial experiment, $n = 10$ was fixed but for the negative binomial experiment, it was determined in advance so that the trial would stop as soon as there were two failures. The likelihood in both cases is: $L(\theta) = k\theta^8(1-\theta)^2$ where k is some constant that does not depend on θ . Under the likelihood principle, inferences in the two cases would be the same given identical prior distributions.

However, if we test $H_0: \theta = 0.5$, then although the p -value is given by $P(X \geq 8 | \theta = 0.5)$, the calculation is not the same in the two cases. For the binomial, we have:

$$P_{\text{bin}} = \sum_{x=8}^{10} \binom{10}{x} \left(\frac{1}{2}\right)^{10} = 0.055$$

However, for the negative binomial, we have:

$$P_{\text{neg}} \sum_{x=8}^{\infty} (x+1) \left(\frac{1}{2}\right)^{x+2} = 0.02$$

Thus because we have the same likelihoods but different p -values, there is a clear violation of the likelihood principle. The implication of this would seem to be (at the very least) that where we have well-structured models, likelihood is preferable as a means of forming inferences about the unknown parameter. This would only leave two justifications, if at all, for using p -values. The first is where the problems are so poorly structured that it is not possible to form a likelihood. (Daniel Bernoulli's would be one, for example, where it appears very difficult.) The second case would be as a means of helping to calibrate the likelihood in multiparameter models, likelihood being difficult to interpret under such cases. It may be noted that, in fact, a common practice in sequential trials is not to adjust the p -value for repeated significance tests but to adjust the reference level to which the p -value is referred. In fact, defining p -values in a way that reflects adjustment directly can be extremely difficult as the ordering of the sample space can be quite complex.

REPEATABILITY OF P-VALUES

A practical question of some interest in the context of drug development is what is the probability of repetition of the p -value. For example, given the observation $P = 0.05$ in trial 1, what is the probability that $P \leq 0.05$ in trial 2? Goodman^[29] has shown that this is plausibly about 50%. A very simple intuitive understanding why this is so may be gained by considering that if the median for the predictive probability distribution for the point estimate from the second trial is the point estimate from the first trial, then there is 50% probability that second point estimate will be

closer to the null than was the first. However, if the trial is the same size, the standard error will be similar so that there is half a chance that the standardized value will be less than it was before. But because it was just significant from the first trial, there is a 50% probability that it will not be significant from the second.

This is a practical point that should be better understood, but it is not, contrary to what has been claimed, an inferential weakness of p -values for a number of reasons.^[30] The first is that there is nothing special from the inferential point of view of looking at the predictive probability for a trial to follow of exactly the same size. If the trial is much larger, then the probability of significance is also larger. The second reason is that if it were the case that a significant result carried with it a high probability that a further result would also be significant, then anticipated evidence would have the same value as actual evidence, a paradoxical and unsatisfactory state of affairs. In fact, in this respect, the behavior of p -values closely resembles Bayesian probability statements.^[30]

ISSUES AFFECTING THE USE OF P-VALUES IN PRACTICE

In this section, some practical issues that affect the use of p -values are discussed. The first is particularly relevant to their use in drug development.

Two-Trials Rule

It is unusual in the context of drug regulation that a pharmaceutical sponsor will achieve the registration of a pharmaceutical by presenting the results of a Phase III program, which only contains one significant trial. A convention, which is usually but not universally observed, has arisen over the years that two Phase III trials should be significant. In the simplest case, where there are only two trials in the Phase III program, because the conventional level of significance employed is $0.025 = 1/40$ one-sided and if the results from both trials are regarded as trustworthy, the overall Type I error rate for the program becomes $(1/40)^2 = 1/16000 = 0.00625$. If this were the only purpose of the two-trials rule, the regulator could be provided with equivalent protection by running a single trial and requiring $P < 0.00625$, one-sided.^[19,31]

To get some feel for this alternative requirement, we may consider the unrealistic but instructive case where data from the two trials we would run are homogenous, the variance is known a priori, and the trials are of identical design and size. If we use the common two-trials rule, we require that the z -scores z_1, z_2 , which is to say the standardized differences from trials 1 and 2, should be less than -1.96 so that we have $z_1 \leq -1.96$. However, the sum $z_1 + z_2$ has a variance of 2 and a standard error of $\sqrt{2}$ and because the quantile of the standard normal corresponding

to 0.00625 is -3.227 , this alternative requirement would have $z_1 + z_2 \leq -3.227\sqrt{2} = -4.564$.

Note that if this alternative requirement is adopted, then individual p -values of 0.025 one-sided will no longer produce significance. The minimal p -value requirement that would *guarantee* significance if both trials satisfied it would now be one corresponding to $\Phi(-4.564/2) = \Phi(-2.282) = 0.011$. On the other hand, there is no requirement for both p -values to attain this limit because significance attained in one is traded off against the other.

However, this alternative approach (which is almost identical, under these circumstance, to an analysis of the pooled data fitting trial as a block) would be more efficient. The only advantage of the common current approach would be if one feared that every now and then circumstances would arise to make a trial result unusable. If that is feared, however, the very common current practice of running two similar trials with almost identical protocols recruiting similar patients in similar centers makes no sense.^[19]

Other Approaches to Combining P-Values

Where a number of trials has been carried out, it may sometimes (rather rarely in practice) be desirable to combine

the results using p -values alone. Fisher's approach has already been discussed above. An early alternative proposal is that of Tippet, which simply uses the minimum p -value to judge the significance of the result. If k trials are being combined, then the probability under H_0 that the minimum p -value is at least as low as the observed minimum value is:

$$P_{\text{Tippet}} = 1 - (1 - P_{\min})^k \quad (10)$$

This is a slightly sharper adjustment than the common Bonferroni adjustment of $kP_{\min}$ for target significance for multiplicity because it exploits the independence of the p -values under the null for a series of trials.^[32] Yet another approach is to base a test on the sum (or equivalently the average) p -value.^[33]

Fig. 3 illustrates the contours of “global” p -values for four of these schemes for the case where there are two trials. However, in practice, meta-analytic approaches using original data are superior and because the sponsor usually has access to all data, there is no reason to summarize the results of a drug development using a direct combination of p -values.^[34]

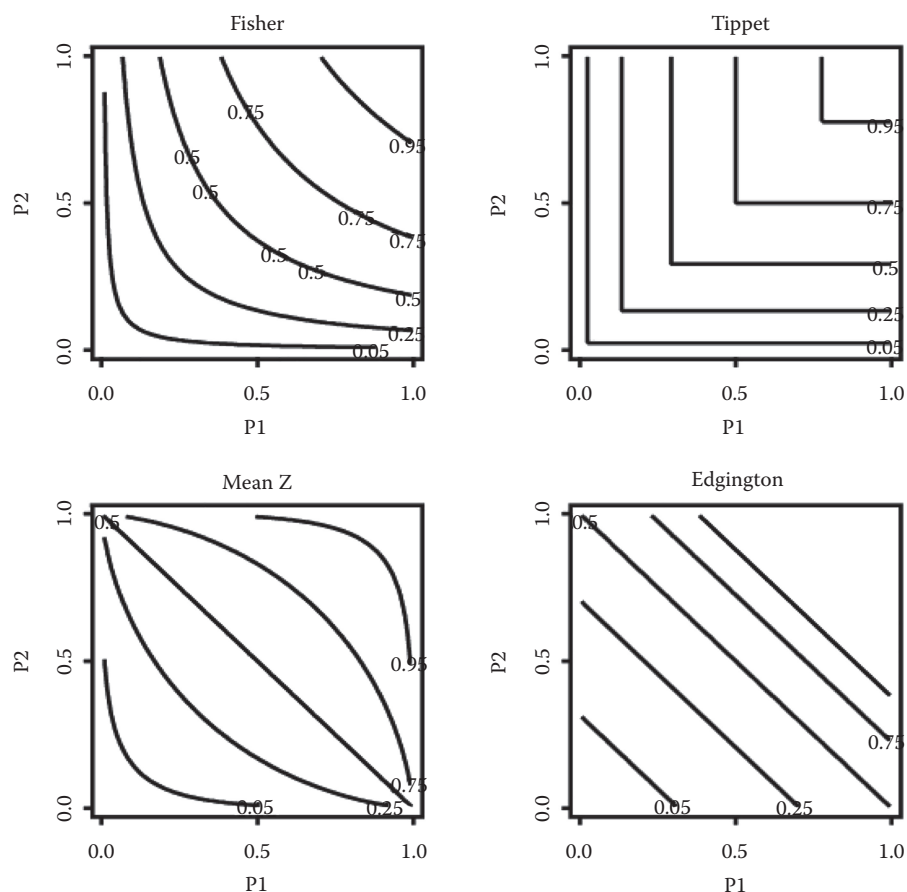


Fig. 3 Contour plots of global P -values calculated from individual P -values from two trials according to four different methods

One-Sided or Two-Sided *P*-Values

The usual convention in drug regulation is that a result is accepted as significant when the two-sided *p*-value is less than 0.05. In practice, however, the regulator will not register a new drug if it is significantly worse than the comparator, whether this is a placebo or a standard treatment. Consequently, one might suppose that if θ represents the difference between the experimental drug and the comparator (defined in such a way that low values of θ are “good”), then in a hypothesis testing framework, the practical problem is to decide between pairs of hypotheses: $H_{0A} : \theta = 0, H_1 : \theta < 0$ or, more realistically, between: $H_{0B} : \theta \geq 0, H_1 : \theta < 0$. This corresponds to a one-sided testing situation and hence one might suppose that one-sided *p*-values would be more appropriate.

To the extent that two-sided *p*-values are calculated by doubling the one-sided *p*-value, this makes no difference because the values commonly adopted for significance are merely conventional. Faced with a sponsor who wished to use one-sided *p*-values, the regulator could accede to the request but then simply insist on using 2.5% as the threshold for significance. The practical effect would be the same as using a 5% one-sided test.

A technical difficulty and possible inferential controversy arises when using test statistics whose distribution is not symmetric.^[28] For example, suppose we want to test whether the mean λ of a random variable X with a Poisson distribution is equal to 13, against the hypothesis that it is not. We take a single observation and observe $X = 6$. The exact one-sided *p*-value is $P(X \leq 6 | \lambda = 13) = 0.0259$. If we double this, we get $P = 0.052$. However, we could instead consider what would constitute a value as extreme or more extreme in the other direction. There are various ways we could define extreme. One might be on the basis of the ratio of the likelihood given the maximum likelihood estimate to the likelihood under the null. This, for given $X = x$, is $LR = e^{x-\lambda}(\lambda/x)^x$ and for the observed value of six this yields 10.6. However, for $x \geq 22$, $LR \geq 13.1$, whereas for $x \leq 21$, $LR \leq 7.9$ so that values of $x \geq 22$ should lead to rejection of the hypothesis. However, $P(X \geq 22) = 0.014$ so that using an approach of adding together tail areas, one would have $P = 0.026 + 0.014 = 0.04$.

This difficulty must be counted as a further inferential drawback of *p*-values. In practice, a rule that has many advantages is simply to double the one-sided value when a two-sided value is wanted.^[35]

A Bayesian examination of *p*-values is given by Freedman^[36] who concludes that there is really no justification from this perspective for two-sided *p*-values.

Problems with Discrete Data

Part of the problem with the Poisson case above concerns the asymmetry of the distribution. However, a further part of it has to do with the discrete nature of the data.

This causes difficulties for hypothesis testing also in that a particular target Type I error rate (say 5%) may not be attainable unless, as part of the decision rule, an auxiliary randomization device were used.^[37] For instance, returning to the Poisson distribution of the previous example, suppose we are interested in testing the alternative $\lambda < 13$. If we reject if $X \leq 7$, the Type I error rate is 0.054, whereas if we reject if $X \leq 6$, we have an error rate of 0.026 (as discussed above). However, suppose we have an auxiliary random variable $R, U(0,1)$, a device first suggested by Tocher,^[37] note that $(0.05 - 0.026)/(0.054 - 0.026) = 0.86$. The following rule will give a Type I error rate of 5%: reject if $(X \leq 6)$ or $(X = 7 \cap R \leq 0.86)$. Of course, in practice, no regulator will accept such an arbitrary external randomizing device but it should be noted that methods, such as hypothesis testing, which themselves make use of the sample space to make decisions, use a partially arbitrary internal randomization device (see Ivanova and Berger^[38] for a particularly marked example). Be that as it may, it is at least a theoretical possibility in the hypothesis testing framework and one which has at least one possible application, namely that of comparing different tests in terms of power in a way that avoids distortion because of particular significance levels that one test or the other test could attain.^[39]

If it seems repugnant to use an auxiliary device for testing hypotheses, it will seem even more so for calculating *p*-values. However, the analogous strategy would appear to be to calculate the probability of observing a result just more extreme than that observed and add some random fraction of the probability of the result observed. Clearly, in practice, no one will do this. However, alternatively, if one imagines that a “remote scientist,” who may have a different habitual Type I error rate to the trialist, would like to use one’s own auxiliary device, one will need to provide two *p*-values: $P(T \leq t)$ and $P(T < t)$, where t is the observed value of the test statistic T . These two *p*-values are sometimes used to calculate a third value, as will now be explained.

The “lumpiness” of testing that accompanies discrete data carries over into *p*-values. In particular, where combining *p*-values is important, this can be unfortunate. However, the distributional properties of *p*-values can be improved by calculating a so called mid P ^[40,41] as $\{P(T \leq t) + P(T < t)\}/2$. This has theoretical advantages over the classical *p*-value for this purpose but in practice is little used.

Multiplicity

Where many significance tests are being performed at a common level of α , the probability of at least one being significant is greater than α . If it is desired to control this family-wise Type I error rate, then it will be necessary to adopt some special scheme, perhaps a predefined order of tests, some adjustment downward of target levels of significance or some adjustment upward of *p*-values.

This is a vast and controversial topic.^[42,43] There is no general agreement that such adjustments are sensible. However, of all the approaches, the upward adjustment of p -values seems the least satisfactory in that it is the one that makes it most difficult for any “remote scientist” to come to an independent inference.

Adaptive Designs

In recent years, there has been a great deal of interest in adaptive designs starting with the influential paper in 1994 by Bauer and Kohne^[44] which used approaches to combining p -values as the basis for controlling the type I error rate. Because p -values, rather than more conventional test statistics are used, such approaches can be extremely flexible since one can even change the outcome measure between different stages of a design. The approach is not without its critics since it can (in theory) do considerable violence to the likelihood principle. A critical evaluation is given by Burman and Sonneson.^[45]

CONCLUSION

A possible historical interpretation is that p -values represent a way of discussing compatibility between null hypotheses and data, which, although not without logical difficulties, is so intuitively appealing that scientists in all areas have tended to use them. Just how natural this way of thinking is may be illustrated using an example from a paper criticizing p -values. In his interesting article, Matthews not only urges scientists to abandon p -values but also considers some aspects of scientific honesty and reporting. In discussing Millikan’s famous determinations of the charge of the electron he writes, “the discrepancy is so large that the probability of generating it by chance alone is less than 1 in 10^3 .”^[46] However, he might just as well have written $P < 0.001$ because this is exactly what is reported here.

It seems that p -values are one of the ways we choose to look at data and models: the most popular and the least respectable. They are perhaps the “pulp fiction” of statistical inference—shallow but not completely without merit. However, there is no denying that they can be dangerous, that where we have well-structured problems, we can do far better, and that the current degree of reliance on them in medical research and drug development is to be deplored.

FURTHER READING

There is a very extensive literature critical of p -values. The papers by Berger and Berry,^[47] Freeman,^[48] Goodman,^[49] and Matthews^[46] are particularly readable introductions. See also papers by Johnstone^[50] and particularly Berger and Sellke^[51] for more detailed criticisms.

Royall^[23] specifically considers the effect of sample size on the meaning of p -values, a point that was famously dealt with by Lindley^[21] (but see Bartlett^[52] for a necessary correction). A classic paper proving the incompatibility of p -values and the likelihood principle is that of Birnbaum.^[53] An excellent and thorough overview of likelihood methods is the book by Pawitan.^[28] For a general discussion of the Bayesian and likelihood approaches to inference and the possible role (if any) of p -values, see the book by Good.^[25] An important paper discussing the use of significance tests in general is that of Cox^[26] and an examination from a Bayesian perspective with particular attention to one-sided and two-sided tests is provided by Freedman.^[36]

There are rather fewer papers defending p -values. Barnard^[40] (despite being the author of the likelihood principle) finds a limited use for them. Senn^[54,55] gives a lukewarm defense. Whitmore and Xekalaki^[56] consider an extension of the p -value concept to consider the predictive validity of models. One of the best discussions of approaches to combining p -values from a number of trials is the article in German by Sonneman.^[32] Another important paper on this topic is that of Birnbaum.^[57]

REFERENCES

1. David, H.A. First(questionable) occurrence of common terms in mathematical statistics. *Am. Stat.* **1995**, *49*, 121–133.
2. Brownlee, K.A. *Statistical Theory and Methodology in Science and Engineering*; Wiley: New York, 1960.
3. Fisher, R.A. The goodness of fit of regression formulae and the distribution of regression coefficients. *J. R. Stat. Soc.* **1992**, *85*, 597–612.
4. Arbuthnot, J. An argument for Divine Providence taken from the constant regularity observed in the births of both sexes. *Philos. Trans. R. Soc. Lond.* **1710**, *27*, 186–190.
5. Shoesmith, E. *Leading Personalities in Statistical Science*; Wiley: New York, 1997, 7–10.
6. Bernoulli, D. *Die Werke von Daniel Bernoulli*; Birkhäuser Verlag: Basle, 1987, 303–326.
7. Student The probable error of a mean. *Biometrika* **1908**, *6*, 1–25.
8. Pearson, K. On the criterion that a given system of deviation from the probable in a correlated system of variables is such that it can reasonably supposed to have arisen from random sampling. *Philos. Mag* **1900**, *50*, 157–175.
9. Fisher, R.A.; Yates, F. *Statistical Tables for Biological Agricultural and Medical Research*; Longman: Harlow, 1974.
10. Lehmann, E.L. *Testing Statistical Hypotheses*, 2nd Ed.; Chapman and Hall: New York, 1994, 70.
11. Gardner, M.J.; Altman, D.G. Confidence intervals rather than P-values: Estimation rather than hypothesis testing. *Br. Med. J. (Clin. Res. Ed.)* **1986**, *292*, 746–750.
12. Spiegelhalter, D.J.; Freedman, L.S.; Parmar, M.K.B. Bayesian Approaches to randomized trials. *J. R. Stat. Soc. Ser. A Stat. Soc.* **1994**, *157*, 357–387.

13. Friedman, S.B.; Phillips, S. What's the difference? Pediatric residents and their inaccurate concepts regarding statistics. *Pediatrics* **1981**, *68*, 644–646.
14. Altman, D.G.; Bland, J.M. Improving doctors' understanding of statistics (with discussion). *J. R. Stat. Soc. Ser. A Stat. Soc.* **1991**, *154*, 223–268.
15. Fisher, R.A. *Statistical Methods, Experimental Design and Scientific Inference*; Oxford University Press: Oxford, 1925.
16. Miettinen, O. *Theoretical Epidemiology: Principles of Occurrence Research in Medicine*; Delmar Publishing, 1985.
17. Senn, S.J. Suspended judgment of n-of-1-trials. *Control. Clin. Trials* **1993**, *14*, 1–5.
18. Hung, H.M. The behavior of the P-value when the alternative hypothesis is true. *Biometrics* **1997**, *53*, 11–22.
19. Senn, S.J. Statistical Issues in Drug Development. *Statistics in Practice*; Wiley: Chichester, 1997.
20. Jeffreys, H. *Theory of Probability*, 3rd Ed.; Clarendon Press: Oxford, 1961.
21. Lindley, D.V. A statistical paradox. *Biometrika* **1957**, *44*, 187–192.
22. Peto, R. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br. J. Cancer* **1976**, *34*, 585–612.
23. Royall, R.M. The effect of sample size on the meaning of significance tests. *Am. Stat.* **1986**, *40*, 313–315.
24. Johnstone, D.J.; Lindley, D.V. Bayesian-inference given data significant-at-alpha-test of point hypotheses. *Theory Decis.* **1995**, *38*, 51–60.
25. Good, I.J. *Good Thinking: The Foundations of Probability and Its Applications*; University of Minnesota Press: Minneapolis, 1983.
26. Cox, D.R. The role of significance tests. *Scand. J. Stat.* **1977**, *4*, 49–70.
27. Barnard, G.A. Statistical inference (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1949**, *11*, 115–149.
28. Pawatan, G.A. *All Likelihood: Statistical Modelling and Inference Using Likelihood*; Clarendon Press: Oxford, 2001.
29. Goodman, S.N. A comment on replication, p-values and evidence. *Stat. Med.* **1992**, *11*, 875–879.
30. Senn, S.J. A note on p-values and replication probabilities. *Stat. Med.* **2002**, *21*, 2437–2444.
31. Fisher, L.D. One large, well-designed, multicenter study as an alternative to the usual FDA paradigm. *Drug Inf. J.* **1999**, *33*, 265–271.
32. Sonneman, E.S. *Kombination unabhängiger test in Biometrie in der Chemisch—Pharmazeutischen Industrie 4*; Fischer Verlag: Stuttgart, 1991.
33. Edgington, E.S. An additive method for combining probability values from independent experiments. *J. Psychol.* **1972**, *80*, 351–363.
34. Senn, S.J. The many modes of meta. *Drug Inf. J.* **2000**, *34*, 535–549.
35. Yates, F. Tests of significance for 2×2 contingency tables. *J. R. Stat. Soc. A* **1984**, *147*, 426–463.
36. Freedman, L.S. An analysis of the controversy over classical one-sided tests. *Clin. Trials* **2008**, *5*, 635–640.
37. Tocher, K.D. Extension of the Neyman-Pearson theory of tests to discontinuous variates. *Biometrika* **1950**, *37*, 130–144.
38. Ivanova, A.; Berger, V.W. Drawbacks to integer scoring for ordered categorical data. *Biometrics* **2001**, *57*, 567–570.
39. Barnard, G.A. On alleged gains in power from lower P-values. *Stat. Med.* **1989**, *8*, 1469–1477.
40. Lancaster, H.O. Significance tests in discrete distributions. *J. Am. Stat. Assoc.* **1961**, *56*, 223–234.
41. Barnard, G. Must clinical trials be large? The interpretation of P-values and the combination of test results. *Stat. Med.* **1990**, *9*, 601–614.
42. Cook, R.J.; Farewell, V.T. Multiplicity considerations in the design and analysis of clinical trials. *J. R. Stat. Soc. Ser. A Stat. Soc.* **1996**, *159*, 93–110.
43. Hsu, J.C. *Multiple Comparisons Theory and Methods*; Chapman & Hall/CRC: Boca Raton, FL, 1996.
44. Bauer, P.; Kohne, K. Evaluation of experiments with adaptive interim analyses. *Biometrics* **1994**, *50*, 1029–1041.
45. Burman, C.-F.; Sonneson, C. Are Flexible designs sound? *Biometrics* **2006**, *62*, 664–669.
46. Matthews, R. Fact Versus Factions: The Uses and Abuse of Subjectivity in Scientific Research. *European Science and Environment: Forum*; 1998.
47. Berger, J.O.; Berry, D.A. Statistical-Analysis and the illusion of objectivity. *Am. Sci.* **1998**, *76*, 159–165.
48. Freeman, P.R. The role of p-values in analysing trial results. *Stat. Med.* **1993**, *12*, 1443–1452. discussion 1453–1458.
49. Goodman, S.N. P-values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *Am. J. Epidemiol.* **1993**, *137*, 485–496. discussion 497–501.
50. Johnstone, D.J. Tests of significance in theory and practice. *Statistician* **1986**, *35*, 491–504.
51. Berger, J.O.; Sellke, T. Testing a point null hypothesis—the irreconcilability of p-values and evidence. *J. Am. Stat. Assoc.* **1987**, *82*, 112–122.
52. Bartlett, M.S. A comment on D.V. Lindley's statistical paradox. *Biometrika* **1957**, *44*, 533–534.
53. Birnbaum, A. On the foundations of statistical inference (with discussion). *J. Am. Stat. Assoc.* **1962**, *57*, 269–326.
54. Senn, S.J. Discussion of paper by Professor Peter Freeman. *Stat. Med.* **1993**, *12*, 1453–1458. (Further comment in same issue pp. 1403, 1437, 1524).
55. Senn, S.J. Two cheers for P-values. *J. Epidemiol. Biostat.* **2001**, *6*, 193–204.
56. Whitmore, G.A.; Xekalaki, E. P-values as measures of predictive-validity. *Biom. J.* **1990**, *32*, 977–983.
57. Birnbaum, A. Combining independent tests of significance. *J. Am. Stat. Assoc.* **1954**, *49*, 559–575.

QT Analysis

Lang Li

Department of Medicine, Indiana University, Indianapolis, Indiana, U.S.A.

Stephen Hall

Division of Clinical Pharmacology, Department of Medicine, Indiana University, Indianapolis, Indiana, U.S.A.

Mehul Desai

Division of Cardiovascular and Renal Products, Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

This entry summarizes four prescribed QT analysis methods and compares them in terms of their biases and efficiencies in QT and QTc prolongation estimations are compared theoretically, and also discusses clinical trial design issues.

INTRODUCTION

Cardiac repolarization is a normal, physiologic process that occurs in man by the outward movement of potassium ions through specific channels in myocardial cell membranes. An undesirable feature of some pharmaceutical agents is their ability to delay cardiac repolarization, an effect that can be indirectly measured as prolongation of the QT/QTc interval on the surface electrocardiogram (ECG). The QT/QTc interval (Fig. 1) represents the time duration of ventricular depolarization and subsequent repolarization, beginning at the start of the QRS complex and ending where the T wave returns to isoelectric baseline. A delay in cardiac repolarization (as manifested by QT prolongation) can predispose to the development of life-threatening cardiac arrhythmias [e.g., torsade de pointes (TdP)]. TdP may degenerate into ventricular fibrillation leading to sudden death.

Most drugs that have caused TdP (fatal and nonfatal) associated with the use of a drug have resulted in regulatory actions, including withdrawal from the market (e.g., terfenadine and cisapride), restriction of use (e.g., thioridazine and droperidol), or denial of marketing authorization. To regulate QT/QTc interval prolongation assessment in non-clinical and clinical trials, Food and Drug Administration (FDA) is announcing the availability of a guidance entitled “E14 Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs.”^[1] The guidance was prepared under the auspices of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). The guidance provides recommendations to sponsors concerning clinical studies to assess the potential of a new drug to

cause cardiac arrhythmias, focusing on the assessment of changes in the QT/QTc interval on the ECG as a predictor of risk. The guidance is intended to encourage the assessment of drug effects on the QT/QTc interval as a standard part of drug development.

This E14 ICH guideline^[1] provides recommendations to sponsors concerning the design, conduct, analysis, and interpretation of clinical studies to assess the potential of a drug to prolong the QT/QTc intervals. More detailed statistical reviews on design and analysis of QT/QTc data are described in the report from Pharmaceutical Manufacturers and Research Association (PhRMA) Statistics Expert Team.^[2] We briefly summarize some important aspects. In designing a QT trial, both positive and placebo controls are important for showing effect or no effect; replicated QT ECGs at a specific time help to reduce the within-subject variation, hence increase the power; the collections of baseline ECGs from multiple resources are more reliable than using a single baseline ECG data set; a definitive QT/QTc study should generally provide dose-response information, including the effect at dose higher than those anticipated for therapeutic purpose; and the choice of end points (QTc prolongation or above/below a QTc threshold) should be dictated by the primary objective of the study, as well as different analyses performed to achieve that objective.

In analyzing QT/QTc data, because of its inverse relationship to heart rate, the QT interval is routinely transformed (normalized) by means of various formulae into a heart rate-independent “corrected” value known as the QTc interval. Two common heart rate correction methods used in clinical practice are Bazett’s^[3] correction and Fridericia’s^[4] correction. By convention, these two correction methods correct the QT interval to a standard heart rate of 60 bpm. However, it has been

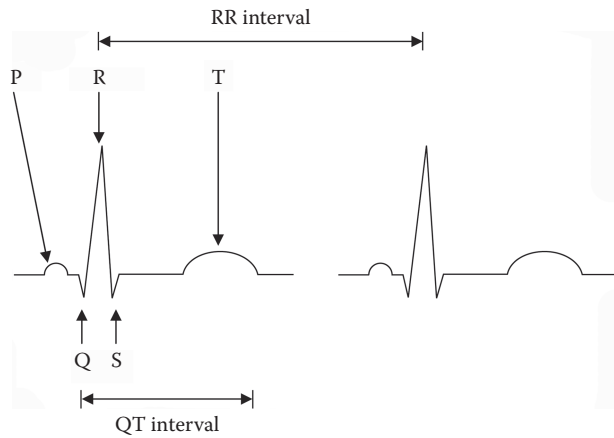


Fig. 1 The P wave represents depolarization occurring within the atria of the heart. The QRS complex represents depolarization occurring within the ventricles of the heart. The T wave represents repolarization occurring within the ventricles of the heart. The QT interval is the period of time it takes to complete both depolarization and repolarization within the ventricles and is measured from the beginning of the Q wave to the end of the T wave

recognized that the relationship between heart rate and QT interval is individual specific. Thus the application of a population level correction such as Bazett's or Fridericia's may not be optimal to correct heart rate in a specific individual and may increase variability when analyzing data. Malik^[5] proposed to characterize an individual's QT/RR relationship obtained from off-drug or placebo data, and then apply subject-specific correction factors derived from the off-drug data to the on-drug dataset. However, because this method assumes an unchanged QT/RR relationship after the drug treatment, it could lead to a biased QTc prolongation estimate. This problem was addressed by Li et al.^[6] in a joint QT correction for both off- and on-drug data.

In this article, for the first time, four prescribed QT analysis methods are summarized together. Their biases and efficiencies in QTc and QTc prolongation estimations are compared theoretically. Their performances are illustrated through a haloperidol QT analysis. In addition, clinical trial design issues are discussed, specifically crossover vs. parallel group studies.

ANALYSIS OF ECG DATA FROM CLINICAL TRIALS

QT Correction Method Overview

Historically, QT analysis is usually composed of two steps.^[3,4] Malik^[5] gives us a comprehensive review of this topic. The first step is to correct the QT interval for the effect of heart rate (inverse of RR interval). Heart rate correction of the QT interval can be based

on a uniform correction for all subjects or based on a subject-specific correction derived from individual regression analysis of that subject's placebo (off-drug) QT vs. RR data. The second step in analyzing whether a drug-induced QT interval change has occurred is to perform statistical analyses based on the heart rate corrected QT (QTc) between the treatment (on-drug) and placebo (off-drug) groups.

From a statistical perspective, it is clear that a one-step joint statistical model for both on- and off-drug QT analysis justified by RR intervals would be the method of choice. As linear mixed models^[7] and nonlinear mixed models^[8] have become popular tools for many biomedical researches, this QT analysis is a simply problem. However, because clinical pharmacologists would be more comfortable with a two-step analysis, it is important to determine to what extent that the two-step analysis would lose information or be biased relative to the one-step joint analysis. Although this problem is discussed in detail before,^[6] this entry provides a comprehensive summary of various QT analysis.

There are many proposed heart rate QT correction methods cited in the medical literature. What is considered the best method of correction is debatable. Bazett's formula^[3] is by far the most common correction used by clinicians, because it is practical and convenient. Based on his empirical experience, Bazett assumed the following relationship $QT = \alpha \times RR^{1/2}$, and $QTc = QT \times 1000^{1/2}/RR^{1/2}$.

QTc has the interpretation of QT when RR = 1000 ms. Fridericia's^[4] formula is another commonly used QT correction method and has been suggested by some to be a better alternative to Bazett's method. He proposed a different empirical model, $QT = \alpha \times RR^{1/3}$, and $QTc = QT \times 1000^{1/3}/RR^{1/3}$. Neither method acknowledged the variation of the QT vs. RR relationship from population to population or subject to subject. Hence, corrected QT inevitably has large variations. The limitations of these uniform corrections were evaluated by Malik^[5] and Desai et al.^[9,10]

Malik^[5] recognized the limitation of assuming a constant relationship between QT vs. RR among different individuals and proposed a subject-specific model to correct the QT interval at the first step through Eqs. (1) and (2). The subject-specific $\hat{\beta}_i$ is a least square estimate of β_i from Eq. (1) that utilized the off-drug (placebo) QT and RR data from subject i only

$$QT_{ij} = \alpha_i \times RR_{ij}^{\beta_i} + \varepsilon_{ij} \quad (1)$$

$$QTc_{ij} = QT_{ij} \times 1000^{\hat{\beta}_i} / RR_{ij}^{\hat{\beta}_i} \quad (2)$$

Although Malik's method has greatly improved the subject-wise QT correction, it has three potential limitations. Firstly, the validity of the equal variance assumption for the measure error of a QT interval is questionable. Based

on our experience, the variance of a QT interval is usually proportional to its magnitude. A better alternative model is a linear model for the log-transformed QT and RR, which is described in Eq. (3). More detailed discussion of the variance model and transformation can be found in Carroll and Ruppert.^[9]

$$\log \text{QT}_{ij} = \gamma_i + \beta_i \times \log \text{RR}_{ij} + \varepsilon_{ij} \quad (3)$$

Secondly, subject-specific estimate of $\tilde{\beta}_i$ in Eq. (2) may not be practical, because it can often be difficult to acquire such large numbers of ECGs (e.g., 41) in a single subject^[5] during a pharmacokinetic/pharmacodynamic study. Although Malik's method can provide asymptotically unbiased subject-specific estimate of β_i , it has larger variance than the best linear unbiased prediction of β_i based on the log-likelihood function for Eq. (1) or (3). This property was carefully addressed in both linear mixed models^[7] and non-linear mixed models.^[8]

Thirdly, at step one, Malik's QT correction term $\tilde{\beta}_i$ in Eq. (2) is estimated from the placebo (off-drug) group only. It is equivalent to assume that β_i is unchanged on and off the drug, which may or may not be true. Hence, the bias of the QTc prolongation estimates caused by this assumption violation needs to be justified. The first two limitations can be addressed by well-developed linear and non-linear mixed model theories, while the third assumption was addressed by Li et al.^[6]

In this section, theoretical comparisons among three methods are performed: fixed coefficient QT corrections proposed by Bazett and Fridericia (fixed QTc), off-drug only QT correction proposed by Malik (off-drug QTc), and off-on-drug joint QT correction proposed by Li et al.^[6] (off-on QTc). The biases and efficiencies of these QTc methods on QTc and QTc prolongation estimates will be derived.

Model Specification

We chose $\log(\text{QT})$ as the outcome variable because QT is often modeled as a power function of RR in the literature.^[5] Log transformation can render a linear relationship. In addition, as the measurement error of the QT interval usually has a constant coefficient of variation (CV), a log transformation can make the variance homogeneous. Davidian and Carroll^[11] provided detailed discussion of the model and transformation of the non-constant variance. It has been shown that women differ from men with regard to the QT vs. RR relationship at heart rates below 100 (or $\text{RR} < 600$ ms) with women having greater QTc prolongation for any given heart rate,^[12] and it is well known that gender can influence drug-induced QTc changes.^[13] Therefore, $\log(\text{QT})$ and $\log(\text{RR})$ should have different linear relationships for the males and the females.

The statistical literatures on linear mixed models can be traced back at least as far as Harville^[14] and Laird and Ware.^[15] Those models have been broadly applied to the biomedical research. Eq. (4) is a linear mixed model representation, where $y_{p,gij}$ denotes the $\log(\text{QT})$ measurement during the placebo period for subject i of gender g at the time point t_j , $g = 1$ (male), 2 (female), $i = 1, \dots, m$ and $j = 1, \dots, n$. Similarly, $y_{t,gij}$ is the $\log(\text{QT})$ measurement during the treatment period for subject i at the time point t_j . The terms, $x_{p,2gij}$ and $x_{t,2gij}$, denote the $\log(\text{RR}/1000)$ measurements within the placebo and treatment periods, respectively. The $\log(\text{RR})$ is linearly transformed to $\log(\text{RR}/1000)$ because the intercept has the heart rate corrected $\log(\text{QT})$ or $\log(\text{QTc})$ interpretation when $\text{RR} = 1000$ ms in model (4). In addition, $x_{p,1gij}$ and $x_{t,1gij}$ are the indicator variables for the gender. In model (4), $\beta_{t,1g}$ and $\beta_{p,1g}$ are $\log(\text{QTc})$ s (intercepts) for the treatment and placebo periods of gender g , respectively, $\beta_{t,2g}$ and $\beta_{p,2g}$ are two corresponding slopes between the $\log(\text{QT})$ and $\log(\text{RR})$. The errors, $\varepsilon_{t,gij}$ and $\varepsilon_{p,gij}$, are $i.i.d.N(0, \sigma_0^2)$, random effect follows $b_{1gi} \sim N(0, \sigma_1^2)$, b_{2gi} follows $N(0, \sigma_2^2)$, and $\text{corr}(b_{1gi}, b_{2gi}) = \rho$. Model (5) is a matrix expression of (4).

$$y_{t,gij} = x_{t,1gij}\beta_{t,1g} + b_{1gi} + (\beta_{t,2g} + b_{2gi})x_{t,2gij} + \varepsilon_{t,gij} \quad (4)$$

$$y_{p,gij} = x_{p,1gij}\beta_{p,1g} + b_{1gi} + (\beta_{p,2g} + b_{2gi})x_{p,2gij} + \varepsilon_{p,gij}$$

$$\begin{aligned} y_t &= x_{t,1}\beta_{t,1} + x_{t,2}\beta_{t,2} + z_{t,1}b_1 + z_{t,2}b_2 + \varepsilon_t \\ y_p &= x_{p,1}\beta_{p,1} + x_{p,2}\beta_{p,2} + z_{p,1}b_1 + z_{p,2}b_2 + \varepsilon_p \end{aligned} \quad (5)$$

$$\begin{aligned} \text{cov}(b_1) &= I\sigma_1^2, \text{cov}(b_2) = I\sigma_2^2, \quad \text{cov}(b_1, b_2) = I\rho\sigma_1\sigma_2 \\ \text{cov}(\varepsilon_t) &= \text{cov}(\varepsilon_p) = I\sigma_0^2 \end{aligned}$$

The QTc is defined as the QT when $\text{RR} = 1000$ ms. In terms of our model (4), the true QTc in the treatment and placebo periods for the males and females are:

$$\begin{aligned} \text{true_QTc}_{t,\text{male}} &= e^{\beta_{t,11}}, \quad \text{true_QTc}_{t,\text{female}} = e^{\beta_{t,12}} \\ \text{true_QTc}_{p,\text{male}} &= e^{\beta_{p,11}}, \quad \text{true_QTc}_{p,\text{female}} = e^{\beta_{p,12}} \end{aligned} \quad (6)$$

The null hypotheses are $\text{true_QTc}_{t,\text{male}} = \text{true_QTc}_{p,\text{male}}$ and $\text{true_QTc}_{t,\text{female}} = \text{true_QTc}_{p,\text{female}}$. Particularly, $\text{true_QTc}_{t,\text{male}} - \text{true_QTc}_{p,\text{male}}$ and $\text{true_QTc}_{t,\text{female}} - \text{true_QTc}_{p,\text{female}}$ represent the QTc prolongations owing to the treatment for the males and females, respectively.

Fixed QT Correction Analysis

In Bazett and Fridericia's correction formula,^[3,4] the QT correction coefficients owing to RR are 0.33 and 0.50, respectively. Here we would like to investigate their QTc and QTc prolongation's biases and variances. Denote β^F as

a vector of fixed QT correction coefficients. The corrected log(QT) for both treatment and placebo groups is:

$$\begin{aligned} y_t^F &= x_{t,1}\beta_{t,1} + x_{t,2}(\beta_{t,2} - \beta^F) + z_{t,1}b_1 + z_{t,2}b_2 + \varepsilon_t \\ y_p^F &= x_{p,1}\beta_{p,1} + x_{p,2}(\beta_{p,2} - \beta^F) + z_{p,1}b_1 + z_{p,2}b_2 + \varepsilon_p \end{aligned} \quad (7)$$

In the analysis step, by assuming that the slopes of log(QT) vs. log(RR) to be β^F on and off the drug, and there is no subject to subject variation, some terms in (7), $x_{t,2}(\beta_{t,2} - \beta^F)$, $z_{t,2}b_2$, $x_{p,2}(\beta_{p,2} - \beta^F)$, and $z_{p,2}b_2$ are zeros. Thus the model can be expressed in Eq. (8), where $\text{cov}(a_i) = I\tau_i^2$, and $\text{cov}(e_i) = \text{cov}(e_p) = I\tau_0^2$. In particular, $\eta_{t,1} - \eta_{p,1}$ is the log(QTc) prolongation based on the mis-specified model (8).

$$\begin{aligned} y_t^* &= x_{t,1}\eta_{t,1} + z_{t,1}a_1 + e_t, \\ y_p^* &= x_{p,1}\eta_{p,1} + z_{p,1}a_1 + e_p \end{aligned} \quad (8)$$

The restricted maximum likelihood (REML) of $(\eta_{t,1}^T, \eta_{p,1}^T)^T$ from (8) is $(\tilde{\eta}_{t,1}^T, \tilde{\eta}_{p,1}^T)^T = (x_1^T w x_1)^{-1} x_1^T w y^F$, where $w = (\tau_0^2 I + \tau_1^2 z_1 z_1^T)^{-1}$, $x_1 = \text{diag}(x_{t,1}, x_{p,1})$, and $z_1 = (z_{t,1}^T, z_{p,1}^T)^T$. On the other hand, $(\tilde{\eta}_{t,1}^T, \tilde{\eta}_{p,1}^T)^T$ can be expressed in terms of (9) and (10) from its true model (7)

$$\begin{pmatrix} \tilde{\eta}_{t,1} \\ \tilde{\eta}_{p,1} \end{pmatrix} = \begin{pmatrix} c_0^{-1}c_{11} & 0 \\ 0 & c_0^{-1}c_{22} \end{pmatrix} \times \begin{pmatrix} x_{t,1}\beta_{t,1} + x_{t,2}(\beta_{t,2} - \beta^F) + z_{t,1}b_1 + z_{t,2}b_2 + \varepsilon_t \\ x_{p,1}\beta_{p,1} + x_{p,2}(\beta_{p,2} - \beta^F) + z_{p,1}b_1 + z_{p,2}b_2 + \varepsilon_p \end{pmatrix} \quad (9)$$

$$\begin{aligned} \tilde{\eta}_{t,1} - \tilde{\eta}_{p,1} &= c_0^{-1} \left(c_{11}x_{t,1}\beta_{t,1} - c_{22}x_{p,1}\beta_{p,1} \right) \\ &\quad + c_0^{-1} \left[c_{11}x_{t,2}(\beta_{t,2} - \beta^F) \right. \\ &\quad \left. - c_{22}x_{p,2}(\beta_{p,2} - \beta^F) \right] \\ &\quad + c_0^{-1} (c_{11}z_{t,1} - c_{22}z_{p,1})b_1 \\ &\quad + c_0^{-1} (c_{11}z_{t,2} - c_{22}z_{p,2})b_2 \\ &\quad + c_0^{-1} (c_{11}\varepsilon_t - c_{22}\varepsilon_p) \end{aligned} \quad (10)$$

Taking expectation on both sides of (9) and (10), Eqs. (11) and (12) display the quantities that fixed QT correction method estimates. Apparently, the estimate of log(QTc) is biased according to Eq. (11). The bias, the second terms of Eq. (11), comes from the products of $\bar{x}_{t,2}$ and $(\beta_{t,2} - \beta^F)$ in treatment group, and $\bar{x}_{p,2}$ and $(\beta_{p,2} - \beta^F)$ in placebo group, where $(\bar{x}_{t,2}, \bar{x}_{p,2})$ represents the closeness of average RR measurements to 1000 ms. The covariance of $\tilde{\eta}_{t,1} - \tilde{\eta}_{p,1}$ is given in Eq. (13), its asymptotic approximation.

$$\begin{pmatrix} E(\tilde{\eta}_{t,1}) \\ E(\tilde{\eta}_{p,1}) \end{pmatrix} = \begin{pmatrix} \beta_{t,1} + \bar{x}_{t,2}(\beta_{t,2} - \beta^F) \\ \beta_{p,1} + \bar{x}_{p,2}(\beta_{p,2} - \beta^F) \end{pmatrix} \quad (11)$$

$$E(\tilde{\eta}_{t,1} - \tilde{\eta}_{p,1}) = (\beta_{t,1} - \beta_{p,1}) + (\bar{x}_{t,2}\beta_{t,2} - \bar{x}_{p,2}\beta_{p,2}) - (\bar{x}_{t,2} - \bar{x}_{p,2})\beta^F \quad (12)$$

$$\text{cov}(\tilde{\eta}_{t,1} - \tilde{\eta}_{p,1}) = O\{\Delta\bar{z}_2^2\} + O\left\{\frac{1}{mn}\right\} \quad (13)$$

$$\begin{aligned} \bar{x}_{t,2} &= \begin{pmatrix} 1/nm \sum_{ij} \log(\text{RR}_{t,21ij}/1000), & 0 \\ 0, & 1/nm \sum_{ij} \log(\text{RR}_{t,22ij}/1000) \end{pmatrix}, \\ \bar{x}_{p,2} &= \begin{pmatrix} 1/nm \sum_{ij} \log(\text{RR}_{p,21ij}/1000), & 0 \\ 0, & 1/nm \sum_{ij} \log(\text{RR}_{p,22ij}/1000) \end{pmatrix} \end{aligned}$$

$$\begin{aligned} \Delta\bar{z}_2^2 &= \left\{ \sum_i \left[\sum_j \log(\text{RR}_{t,21ij}/\text{RR}_{p,21ij})/n \right]^2 / m, \right. \\ &\quad \left. \sum_i \left[\sum_j \log(\text{RR}_{t,22ij}/\text{RR}_{p,22ij})/n \right]^2 / m \right\}^T \end{aligned}$$

Off-Drug QT Correction Analysis

The detailed off-drug QT correction analysis can be found in Ref. [6]. We present a summary here. In Malik's QT correction step, based on the placebo data, $y_p = x_{p,1}\beta_{p,1} + x_{p,2}\beta_{p,2} + z_{p,1}b_1 + z_{p,2}b_2 + \varepsilon_p$, the REML^[12] of $\beta_{p,2}$ and b_2 are $\tilde{\beta}_{p,2}$ and $\tilde{b}_2$, respectively. The corrected log(QT) for both treatment and placebo groups are:

$$\begin{aligned} y_t^* &= x_{t,1}\beta_{t,1} + x_{t,2}(\beta_{t,2} - \tilde{\beta}_{p,2}) + z_{t,1}b_1 \\ &\quad + z_{t,2}(b_2 - \tilde{b}_2) + \varepsilon_t \\ y_p^* &= x_{p,1}\beta_{p,1} + x_{p,2}(\beta_{p,2} - \tilde{\beta}_{p,2}) + z_{p,1}b_1 \\ &\quad + z_{p,2}(b_2 - \tilde{b}_2) + \varepsilon_p \end{aligned} \quad (14)$$

In the analysis step, by assuming unchanged slopes of log(QT) vs. log(RR) on and off the drug, some terms in Eq. (14), $x_{t,2}(\beta_{t,2} - \tilde{\beta}_{p,2})$, $z_{t,2}(b_2 - \tilde{b}_2)$, $x_{p,2}(\beta_{p,2} - \tilde{\beta}_{p,2})$, and $z_{p,2}(b_2 - \tilde{b}_2)$ are canceled out. Thus the model becomes Eq. (15), in which $\text{cov}(a_i) = I\tau_i^2$, and $\text{cov}(e_i) = \text{cov}(e_p) = I\tau_0^2$. In particular, $\eta_{t,1} - \eta_{p,1}$ is the log(QTc) prolongation based on the misspecified model (15) under the unchanged slope assumption.

$$\begin{aligned} y_t^* &= x_{t,1}\eta_{t,1} + z_{t,1}a_1 + e_t \\ y_p^* &= x_{p,1}\eta_{p,1} + z_{p,1}a_1 + e_p \end{aligned} \quad (15)$$

The REML of $(\eta_{t,1}^T, \eta_{p,1}^T)^T$ from Eq. (15) is $(\tilde{\eta}_{t,1}^T, \tilde{\eta}_{p,1}^T)^T = (x_1^T w x_1)^{-1} x_1^T w y^*$. Based on Li et al.^[6], their expectations are displayed in Eqs. (16) and (17). Apparently, treated group's log(QTc) estimate is biased, and the estimate of log(QTc) prolongation is biased. The bias, the second term of Eq. (17), comes from the product of $\bar{x}_{t,2}$ and $(\beta_{t,2} - \beta_{p,2})$. Eq. (18) provides log(QTc) prolongation estimate variance's asymptotic approximation.

$$E(\tilde{\eta}_{p,1}) = \beta_{p,1}, \quad E(\tilde{\eta}_{t,1}) = \beta_{t,1} + \bar{x}_{t,2}(\beta_{t,2} - \beta_{p,2}) \quad (16)$$

$$E(\tilde{\eta}_{t,1} - \tilde{\eta}_{p,1}) = (\beta_{t,1} - \beta_{p,1}) + \bar{x}_{t,2}(\beta_{t,2} - \beta_{p,2}) \quad (17)$$

$$\begin{aligned} \text{cov}(\hat{\beta}_{t,1} - \hat{\beta}_{p,1}) &= O\left\{\frac{1}{mn}\right\} + O\left\{\frac{\Delta \bar{x}_2^2}{mn}\right\} + O\left\{\frac{\Delta \bar{z}_2^2}{mn}\right\} \\ \Delta \bar{x}_2^2 &= \left\{ \left(\frac{1}{mn} \sum_{ij} \log(RR_{t,21ij} / RR_{p,21ij}) \right)^2 \right. \\ &\quad \left. \left(\frac{1}{mn} \sum_{ij} \log(RR_{t,22ij} / RR_{p,22ij}) \right)^2 \right\}^T \end{aligned} \quad (18)$$

Off-On-Drug QT Correction Analysis

Estimating the QTc from the joint models in (5) is called the off-on-drug QT correction analysis. Obviously all the β s in Eq.(5) can be estimated jointly by maximizing REML. Following the theories of linear models,^[7,16] the QTc prolongation estimates are unbiased and the most efficient. However, we would like to describe the REML from the joint models (5) as if they are solved through the following two-step procedure. This procedure provides us a way to compare the fixed QT correction analysis and off-drug QT correction analysis to the off-on-drug QT correction analysis. Given that $\hat{\beta}_{t,2}$, $\hat{\beta}_{p,2}$, and $\hat{b}_2$ are REML estimates of $\beta_{t,2}$, $\beta_{p,2}$, and b_2 from Eq. (5), respectively, the REML of $\beta_{t,1}$, $\beta_{p,1}$, and b_1 can be solved through a reduced model (19) by treating $\hat{\beta}_{t,2} = \hat{\beta}_{t,2}$, $\hat{b}_2 = \hat{b}_2$, and $\hat{\beta}_{p,2} = \hat{\beta}_{p,2}$.

$$\begin{aligned} y_t^{**} &= x_{t,1} \beta_{t,1} + x_{t,2} (\beta_{t,2} - \hat{\beta}_{t,2}) + z_{t,1} b_1 \\ &\quad + z_{t,2} (b_2 - \hat{b}_2) + \varepsilon_t \\ y_p^{**} &= x_{p,1} \beta_{p,1} + x_{p,2} (\beta_{p,2} - \hat{\beta}_{p,2}) + z_{p,1} b_1 \\ &\quad + z_{p,2} (b_2 - \hat{b}_2) + \varepsilon_p \end{aligned} \quad (19)$$

In particular, $(\hat{\beta}_{t,1}^T, \hat{\beta}_{p,1}^T)^T = (x_1^T w x_1)^{-1} x_1^T w y^{**}$, where $w = (\sigma_0^2 I + \sigma_1^2 z_1 z_1^T)^{-1}$, $x_1 = (x_{t,1}^T, x_{p,1}^T)^T$, $z_1 = (z_{t,1}^T, z_{p,1}^T)^T$, and $y^{**} = (y_t^{**T}, y_p^{**T})^T$. With the same derivations as in Eqs. (16) and (17), $E(\hat{\beta}_{t,1}) = \beta_{t,1}$, $E(\hat{\beta}_{p,2}) = \beta_{p,2}$, and $E(\hat{\beta}_{t,1} - \hat{\beta}_{p,2}) = \beta_{t,1} - \beta_{p,1}$.

They reinforce that REML estimate $\hat{\beta}_{t,1} - \hat{\beta}_{p,2}$ is unbiased. The asymptotic covariance of $\hat{\beta}_{t,1} - \hat{\beta}_{p,2}$ follows Eq. (20),

$$\text{cov}(\hat{\beta}_{t,1} - \hat{\beta}_{p,1}) = O\left\{\frac{1}{mn}\right\} + O\left\{\frac{\Delta \bar{x}_2^2}{mn}\right\} + O\left\{\frac{\Delta \bar{z}_2^2}{mn}\right\} \quad (20)$$

SUMMARY

- The fixed QT correction analysis leads to both biased log(QTc) and its prolongation estimates. The degree of bias for the log(QTc) estimates depends on $\bar{x}_{t,2}(\beta_{t,2} - \beta^F)$ in treatment group, and $\bar{x}_{p,2}(\beta_{p,2} - \beta^F)$ in placebo group; the degree of log(QTc) prolongation estimates depends on $(\bar{x}_{t,2}\beta_{t,2} - \bar{x}_{p,2}\beta_{p,2}) - (\bar{x}_{t,2} - \bar{x}_{p,2})\beta^F$. The variance of log(QTc) prolongation estimate is as the same order as $O\{\Delta \bar{x}_2^2\} + O\{1/mn\}$.
- Off-drug QT correction has unbiased log(QTc) estimate in placebo group, but bias estimate in treatment group. Thus, it leads to a biased log(QTc) prolongation estimate, and the bias is $\bar{x}_{t,2}(\beta_{t,2} - \beta_{p,2})$. The variance of log(QTc) prolongation estimate is as the same order as $O\{1/mn\} + O\{\Delta \bar{x}_2^2/mn\} + O\{\Delta \bar{z}_2^2/mn\}$.
- Off-on-drug method has no bias on log(QTc) and log(QTc) prolongation estimates. The variance of log(QTc) prolongation estimate is as the same order as $O\{1/mn\} + O\{\Delta \bar{x}_2^2/mn\} + O\{\Delta \bar{z}_2^2/mn\}$.

HALOPERIDOL DATA ANALYSIS

We studied 22 nonsmoking healthy volunteers aged between 21 and 41 years. The study was conducted in Georgetown University's General Clinical Research Center (GCRC). All volunteers gave written informed consent prior to participation in the study. The study protocol was approved by the Institutional Review Board of Georgetown University Medical Center. This protocol was a randomized, double blind, placebo controlled crossover clinical trial. There were two study periods with each period lasting four days. There was a three-day washout between the two study periods. Subjects were randomly assigned to receive either a single placebo capsule or a single haloperidol tablet (Geneva Pharmaceuticals Lot # 114748, Broomfield, CO) at 8 AM during the first study period and the alternative treatment at the identical time during the second study period. ECGs were collected before and 2, 4, 6, 8, 10, 24, 36, 48, 60, 72, 84, and 96 hr after the administration of either the placebo or haloperidol. A total of 13 recordings were performed in the placebo period and 13 recordings performed in the treatment period. ECGs were obtained with the subject resting supine for approximately 15 min prior to acquisition using a MAC 5000 ECG machine (GE Medical Systems, Milwaukee, WI). We acquired 12 lead ECGs

at a paper speed of 50 mm/sec and an amplitude of 20 mm/mV. The ECGs were printed in a format that displayed rhythm strips from all 12 leads simultaneously to facilitate identification of the earliest Q wave and latest T wave observable.

Fig. 2 shows the patterns of $\log(\text{QT})$ vs. $\log(\text{RR})$ relationships between the males and females in the treatment and placebo periods separately. The $\log(\text{QTc})$ is $\log(\text{QT})$ when $\log(\text{RR}) = \log(1000) \cong 6.91$. Fig. 3 shows the fitting of

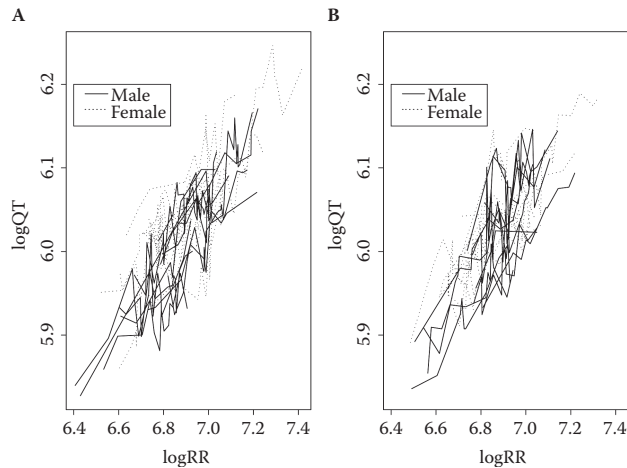


Fig. 2 Plots of $\log(\text{QT})$ vs. $\log(\text{RR})$ (raw data). The solid lines are for males, and the dotted lines are for females. (A) is the placebo group and (B) is the treatment group

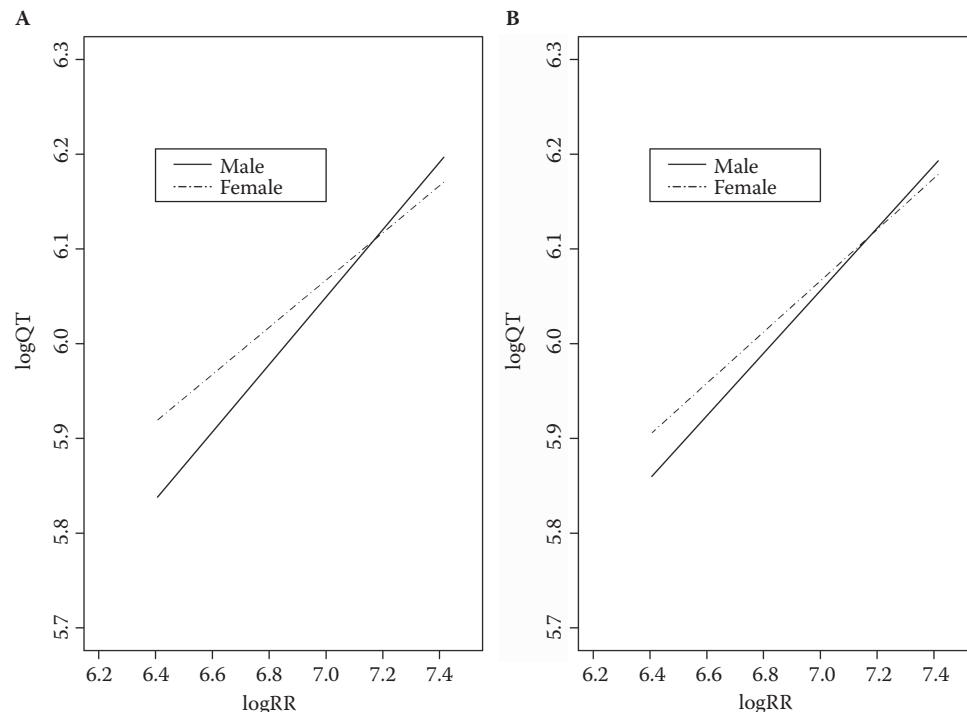


Fig. 3 Plots of $\log(\text{QT})$ vs. $\log(\text{RR})$ based the fitting of linear mixed models (3) and (4). The solid lines are for males, and the broken lines are for females. (A) is the placebo group and (B) is the treatment group

model (4) based on our method. Males and females have different QT vs. RR relationships in both the placebo and treatment periods. In addition, after treatment, male and female's QTc tend to be closer than they are in the placebo period. Table 1 summarizes the QT and QTc prolongation estimates among all four methods, Fridericia, Bazett, off-drug, and off-on-drug QT analysis.

- **QTc estimates:** Treatment does not alter the $\log(\text{QT})$ and $\log(\text{RR})$ relationships in both female and male groups according to our off-on-drug analysis, i.e., the slopes are 0.35 and 0.32 for treatment and placebo groups in male group, and they are 0.26 and 0.26 in the female group. As we expected, female and male have different QT/RR relationship. The off-on-drug correction term for male is more close to Fridericia coefficient (0.33) than Bazett's coefficient (0.50). Therefore, Fridericia's QTc estimates should be more close to off-on-drug QTc than Bazett's QTc estimates, and they are confirmed in Table 1. For example, in the treated female group, off-on-drug QTc is 419.17ms, Fridericia's QTc is 421.70 ms, and Bazett's QTc is 424.31ms. On the other hand, as the drug does not change the slope of $\log(\text{QT})$ and $\log(\text{RR})$ much, off-on-drug and off-drug QTc estimates are close. For example, in the treated female group, off-on-drug QTc is 419.17 ms and off-drug QTc is 420.35 ms.
- **QTc prolongation estimates:** Following our theoretical discussion, off-on-drug QTc prolongation

Table 1 Haloperidol Data Analysis

Parameter	Period	Sex	OFF-ON	OFF	Fridericia	Bazett
$e^{\beta_{p,11}} = QTc_{p,11}$	P	M	409.00(4.21)	409.55(4.47)	409.67(4.65)	410.50(4.90)
$e^{\beta_{p,12}} = QTc_{p,12}$	P	F	420.73(4.21)	422.01(4.47)	422.57(4.65)	423.24(4.90)
$e^{\beta_{t,11}} = QTc_{t,11}$	T	M	413.72(4.21)	414.06(4.47)	414.23(4.65)	415.95(4.91)
$e^{\beta_{t,12}} = QTc_{t,12}$	T	F	419.17(4.21)	420.35(4.47)	421.70(4.65)	424.31(4.91)
$\beta_{t,11} - \beta_{p,11}$			0.011(0.003)	0.012(0.003)	0.011(0.003)	0.013(0.004)
$\beta_{t,12} - \beta_{p,12}$			-0.004(0.003)	-0.004(0.003)	-0.002(0.003)	0.003(0.004)
$QTc_{t,11} - QTc_{p,11}$			4.51(1.39)	4.72(1.38)	4.56 (1.47)	5.45(1.88)
$QTc_{t,12} - QTc_{p,12}$			-1.56(1.39)	-1.66(1.39)	-0.87(1.46)	1.07(1.89)
$\beta_{p,21}$	P	M	0.35(0.03)	0.35(0.02)		
$\beta_{p,22}$	P	F	0.26(0.02)	0.26(0.02)		
$\beta_{t,21}$	T	M	0.32(0.02)			
$\beta_{t,22}$	T	F	0.26(0.02)			
$\beta_{p,21} - \beta_{p,22}$			0.09(0.02)	0.09(0.03)		
$\beta_{t,21} - \beta_{t,22}$			0.06(0.03)			
$\beta_{t,21} - \beta_{p,21}$			-0.03(0.02)			
$\beta_{t,22} - \beta_{p,22}$			0.004(0.02)			
σ_0			0.026	0.027	0.029	0.036
σ_1			0.032	0.033	0.031	0.032
σ_2			0.057	0.066		
ρ			0.788	0.864		

OFF represents the off-drug QT correction analysis, OFF-ON represents the off-on drug QT correction analysis, and Fridericia and Bazett are fixed QT correction analysis.

estimate is unbiased, but the other three are biased. The bias largely depends on the difference between the estimated or fixed QTc correction slope and the true QTc correction slope. In off-drug analysis, as the difference of slopes between treated and placebo groups is small in both male and female groups, the off-drug QTc prolongation estimates have small biases. For example, in the female group, the off-drug QTc prolongation is estimated as -1.66 ms, and the off-on-drug estimate is -1.56 ms. On the other hand, as Fridericia's slope, 0.33, is closer to the ones from off-on method than Bazett's 0.50, Bazett's QTc prolongation estimate, 1.07 ms, is more biased than the Fridericia's -0.87 ms.

- **Variance of QTc prolongation estimates:** Both two fixed coefficient methods have larger standard errors (SEs) than the regression based methods, i.e., 1.47 ms (Fridericia) and 1.88 ms (Bazett) vs. 1.38 ms in both off-on-drug and off-drug. These results confirmed our asymptotic approximations in Eqs. (13), (18), and (20).

CLINICAL TRIAL DESIGN AND SAMPLE SIZE CONSIDERATION

Clinical Trial Design

There are many clinical trial design elements to consider when designing a study to assess the ability of a pharmaceutical agent to prolong the QT interval. Some of these elements include selecting the most appropriate dose or doses to study, appropriate timing of ECG measurements, appropriate blinding, the use of a positive control agent to demonstrate assay sensitivity, and appropriate powering of the study.

With respect to study power and adequately sizing a study, it is important to identify a clinically irrelevant (or clinically relevant) magnitude of QT/QTc interval prolongation. It is not completely clear what magnitude of QT/QTc interval prolongation is of no clinical consequence. It has been suggested by some that drugs that prolong the mean (baseline and placebo-subtracted) QTc interval by 5 ms or less do not appear to cause TdP.^[15]

Table 2 Power and Sample Size Analysis

Total sample 2 <i>mn</i>	Number of subject <i>m</i>	QT#/subject <i>n</i>	QTc prolongation			
			5 <i>ms</i>	10 <i>ms</i>	20 <i>ms</i>	40 <i>ms</i>
200	100	1	0.047	0.057	0.185	0.890
400	200	1	0.051	0.074	0.455	0.992
600	300	1	0.059	0.115	0.690	1.000
800	400	1	0.061	0.201	0.840	1.000
1000	500	1	0.066	0.262	0.911	1.000
1200	600	1	0.069	0.340	0.975	1.000
200	20	5	0.057	0.167	0.487	0.950
300	30	5	0.097	0.227	0.628	1.000
400	20	10	0.087	0.272	0.765	1.000
400	40	5	0.107	0.288	0.897	1.000
600	30	10	0.141	0.358	0.910	1.000
800	40	10	0.163	0.498	0.979	1.000

The number *m* denotes subject number in a clinical trial, and *n* denotes the QT samples per subject. All the power calculations are based on 10,000 simulation data sets.

One design topic that is discussed in more detail in this review deals with the use of a crossover or parallel group design. We will see later in section 5 that parallel group design usually needs more samples (QT measurements) than crossover design, because the crossover design uses within-subject variation more efficiently.

Crossover or parallel group study designs can be suitable for trials assessing the potential of a drug to cause QT/QTc interval prolongation. Crossover studies at least have two potential advantages:

- They usually call for smaller numbers of subjects than parallel group studies, as the subjects serve as their own controls and hence reduce variability of differences related to intersubject variability.
- They might facilitate heart rate correction approaches based on individual subject data.

Parallel group studies might be preferred under certain circumstances:

- For drugs with long elimination half-lives for which lengthy time intervals would be required to achieve steady state or complete washout.
- If carryover effects are prominent for other reasons, such as irreversible receptor binding or long-lived active metabolites.
- If multiple doses or treatment groups are to be compared.

Power and Sample Size Analysis

Following our discussion in the clinical trial design, it is worthwhile to provide more details in crossover and parallel designs. According to the variation information we obtained from the data analysis, it would be very helpful

to investigate the sample size and power in testing QTc prolongation hypotheses.

In the simulation, the within-subject CV for QT is $\sigma_0 = 0.026$, the between-subject CV at the average level (random intercept) for QT is $\sigma_1 = 0.32$, and the between-subject standard deviation for log(QT) and log(RR) slope is set as $\sigma_2 = 0.057$. The slope is set as $\beta_{21} = 0.33$ for both placebo and treated groups. For the sake of simplicity, we treat male and female subjects equally. Practically, they do show different QTc and log(QT) and log(RR) relationship, but they only take two degrees of freedom to estimate them, and won't affect the sample size much. The simulated data for both crossover and parallel designs are generated through model (21) and (22), in which, dQTc is the QTc prolongation that we expect to detect from the data. In our simulation, dQTc is specified as 5, 10, 20, and 40 ms. Different sample size combinations for subject number, $m = (20, 30, 40)$, and samples per subject, $n = (1, 5, 10, \text{ and } 15)$ are simulated. Please note that, *m* denotes the number of subjects in each treatment group. The total QT sample size is 2 *mn*.

$$\begin{aligned} \log(\text{QT})_{t,ij} &= \beta_{t,1} + b_{1i} + (\beta_{t,2} + b_{2i}) \\ &\quad \log(\text{RR}_{t,2ij}/1000) + \varepsilon_{t,ij}, \\ \log(\text{QT})_{p,ij} &= \beta_{p,1} + b_{1i} + (\beta_{p,2} + b_{2i}) \\ &\quad \log(\text{RR}_{p,2ij}/1000) + \varepsilon_{p,ij} \end{aligned} \quad (21)$$

$$\begin{aligned} \text{cov}(b_{1i}) &= 0.032^2, \quad \text{cov}(b_{2i}) = 0.057^2, \\ \text{corr}(b_{1i}, b_{2i}) &= 0.85, \\ \text{cov}(\varepsilon_t) &= \text{cov}(\varepsilon_p) = 0.026^2, \quad \beta_{t,2} = \beta_{p,2} = 0.33, \\ \beta_{t,1} &= \log(420 \text{ ms} + \text{dQTC}), \quad \beta_{p,1} = \log(420 \text{ ms}), \\ (\text{RR}_{t,ij}, \text{RR}_{p,ij}) &\sim N[\log(1000 \text{ ms}), 160^2] \end{aligned} \quad (22)$$

When there are more than one QT sample measured from each subject in a clinical trial (i.e., crossover trial), off-on QTc analysis will be performed. For parallel group trial, in which only one QT sample is measured for a subject, a linear model (23) is employed to test the QTc prolongation hypothesis. In this linear model (23), $e_{t,i}$ contains both within-subject variation, $\varepsilon_{t,i}$, and between-subject variations (b_{1i} , b_{2i}).

$$\begin{aligned}\log(\text{QT})_{t,i} &= \beta_{t,1} + \beta_{t,2} \log(\text{RR}_{t,2i}/1000) + e_{t,i}, \\ \log(\text{QT})_{p,i} &= \beta_{p,1} + \beta_{p,2} \log(\text{RR}_{p,2i}/1000) + e_{p,i}\end{aligned}\quad (23)$$

Each power in Table 2 is calculated from 10000 simulation data sets.

In summary:

- Parallel design requires larger samples (QT measurements) than crossover design to achieve the similar power. For example, for $d\text{QT}_c = 10$ ms, parallel design needs 1200 samples to achieve 34% power, while crossover design needs only 600 samples to achieve 35.8% power. For $d\text{QT}_c = 20$ ms, parallel design needs 1000 samples to achieve 91% power, while crossover design needs only 600 samples to achieve 91% power.
- In the crossover design, when the total sample size is fixed, less samples per subject tend to provide more power than more samples per subject. For example, for $d\text{QT}_c = 10$ ms, 20 subjects and 10-sample/subject provides 27.2% power; and 40 subjects and 5-sample/subject provides 28.8% power. For $d\text{QT}_c = 20$ ms, these two combinations give 76.5% and 89.7% power, respectively.

CONCLUSIONS

Fixed QT correction analysis, such as Bazett and Fridericia, not only have biased QTc prolongation estimates, but also lead to larger variances. Off-drug QT correction analysis proposed by Malik has small bias in QTc prolongation estimate. Its variance is almost the same as those based on off-on-drug correction analysis. Our proposed off-on-drug correction analysis has no bias in QTc and QTc prolongation estimates. Although the off analysis can produce reliable QTc prolongation estimates, off-on analysis is a better choice by all means. As the standard statistical softwares for linear mixed models, i.e., SAS PROC MIXED (SAS Institute) and Splunx nlme (Mathsoft Inc.), become more popular, we highly encourage scientists to implement the off-on analysis.

Generally crossover design needs smaller sample size than parallel design. In the crossover design, when the

total sample size is fixed, less samples per subject tend to provide more power than more samples per subject.

Although off-on analyses assumed a constant drug effect during the study, this assumption need not be true. If this assumption is not true, this analysis will yield a diluted drug effect on the QTc interval because drug concentrations at the very beginning and tail end of the study are much lower than during the middle of the study. This may be one reason that we failed to see a drug effect in females administered haloperidol in our study. Thus, modeling the QTc time profile and evaluating its association with the drug concentration are important clinical questions. Adding time effects and drug concentrations into the $\log(\text{QT})$ vs. $\log(\text{RR})$ model is not trivial work, because allowing both the intercept and slope of model (4) to be functions of time will greatly increase the colinearity. This can become a good research topic.

ACKNOWLEDGMENTS

Dr. Lang Li is supported by NIH grant, R01 GM74217 (LL).

REFERENCES

1. http://www.ich.org/MediaServer.jserv?@_ID=1476&@_MODE=GLB (accessed 12th October, 2005).
2. http://www.findarticals.com/p/articals/mi_qa3899/is_200501/ai_n14905765 (accessed 12th October, 2005).
3. Bazett, J.C. An analysis of time relations of electrocardiograms. *Heart*. **1920**, 7, 353–367.
4. Fridericia, L.S. Die Systolendauer im Elektrokardiogramm bei normalen Menschen und bei Herzkranken. *Acta. Med. Scand.* **1920**, 53, 469–486.
5. Malik, M. Problems of heart rate correction in assessment of drug-induced QT interval prolongation. *J. Cardiovasc. Electrophysiol.* **2001**, 12, 411–420.
6. Li, L.; Desai, M.; Desta, Z.; Flockhart, D. QT analysis: a complex answer to a 'simple' problem. *Stat. Med.* **2004**, 23, 2625–2643.
7. Searle, S.R.; Casella, G.; McCulloch, C.E. *Variance Components*; Wiley: New York, 1992.
8. Davidian, M.; Giltinan, D.G. *Nonlinear Models for Repeated Measurement Data*; Chapman & Hall: London, 1995.
9. Desai, M.; Tanus-Santos, J.; Li, L.; Gorski, J.C.; Arefayene, M.; Liu, Y.; Desta, Z.; Flockhart, D. Pharmacokinetics and QT interval pharmacodynamics of oral haloperidol in poor and extensive metabolizers of CYP2D6. *Pharmacogenomic. J.* **2003**, 3 (2), 105–113.
10. Desai, M.; Li, L.; Desta, Z.; Malik, M.; Flockhart, D. Variability of heart rate correction methods for the QT interval: potential for QTc overestimation with a low dose of haloperidol. *Br. J. Clin. Pharmacol.* **2003**, 55 (6), 511–517.
11. Davidian, M.; Carroll, R.J. Variance function estimation. *J. Am. Stat. Assoc.* **1987**, 82, 1079–1091.

12. Mayuga, K.A.; Parker, M.; Sukthanker, N.D.; Perlowski, A.; Schwartz, J.B.; Kadish, A.H. Effects of age and gender on the QT response to exercise. *Am. J. Cardiol.* **2001**, *87*, 163–167.
13. Benton, R.E.; Sale, M.; Flockhart, D.A.; Woosley, R.L. Greater quinidine-induced QTc interval prolongation in women. *Clin. Pharmacol. Ther.* **2000**, *67*, 413–418.
14. Harville, D. Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* **1977**, *72*, 320–340.
15. Laird, N.M.; Ware, J.H. Random effects models for longitudinal data. *Biometrics.* **1982**, *38*, 963–974.
16. Lehmann, E.L. *Theory of Point Estimation*; Chapman & Hall: London: 1986.

Quality by Design in Biopharmaceutical Development

Harry Yang

MedImmune, LLC, Gaithersburg, Maryland, U.S.A.

Abstract

In recent years Quality by design (QbD) has been in the forefront of biopharmaceutical development. A cornerstone of QbD is to build quality into a product based on a thorough understanding of the product and manufacturing process. To that end, QbD involves defining desired clinical performance, identifying critical quality attributes that have an impact on product safety and efficacy, determining necessary raw material attributes and manufacturing process parameters to achieve the desired product quality attributes, and designing manufacturing controls to ensure consistent manufacturing. QbD enhances development capability, increases development speed, and enables manufacturing robustness. It also helps proactively prevent manufacturing failures. In certain instances QbD enables a manufacturer to make post-approval changes or scale-up operations without prior approval from regulatory authorities. However, implementation of QbD is often challenging. This article discusses key elements of QbD and statistical methods that can be utilized to enable the successful implementation of QbD.

Keywords: Analytical method; Bayesian analysis; Critical quality attributes; Design space; Process validation; Quality-by-Design; Risk assessment.

INTRODUCTION

Quality by design (QbD) has been in the forefront of biopharmaceutical development ever since the publication in [year] by the US Food and Drug Administration (FDA) of “*Pharmaceutical Current Good Manufacturing Practices (cGMPs) for the 21st Century*.” This initiative was intended to address a number of issues concerning pharmaceutical industry, which include low efficiencies in bringing innovative medicines to the market, waste of materials, and rework of finished products. Coupled with other subsequent regulatory publications including ICH Q8 (R2) Pharmaceutical Development, ICH Q9 Quality Risk Management, ICH Q10 Pharmaceutical Quality Systems, and Q11 Concept Paper, the initiative paved the way for the industry to adopt a QbD approach to process and product development, utilizing new technological advances and quality risk management (QRM) techniques. To deliver on the promise of this initiative, an increased emphasis has been placed on product and process understanding. This includes understanding the mechanism of action of the product, its safety profile, variability of raw materials, bulk and finished products, analytical method variations, relationships between process parameters and the product’s critical quality attributes (CQAs), and the impact of specific CQAs on clinical responses and overall product quality. This enhanced product and process knowledge not only enables product quality to be built by

design, but also provides a basis for the development of robust control strategies.

However, despite the plethora of opportunities to reduce manufacturing costs while ensuring product quality, implementation of QbD is often challenging due to the large number of variables to consider, interactions among the variables, uncertainties in measurements, and missing data involved in product and process development. Such intrinsic complexities argue for advanced statistical methods in both study design and analysis. This article is intended to discuss key elements of QbD, statistical methods and other analytical tools that can be utilized to help with implementation of QbD.

QUALITY BY DESIGN

In the traditional development paradigm, product and process design are often based on studies that explore one variable at a time, resulting in little understanding of the interdependence of variables that may have a joint effect on an output. Process validation is traditionally a documentation exercise, and primarily focused on the initial full-scale batches from a process. The overall product quality demonstration relied on extensive testing, including the inspection of raw materials, in-process material testing, and final drug substance and product release testing. The lack of product and process understanding and limited

knowledge of variability of analytical methods and manufacture process often resulted in stringent specifications that prohibited the release of products that otherwise may have had acceptable clinical performance.^[1] As changes to the process and/or test methods required post approval submissions, there were no incentives for the manufacturers to improve their manufacturing processes.

The QbD approach distinguishes itself from the traditional product and process development paradigm in many aspects. Key to this is that quality should be designed into a product, rather than demonstrated by testing. Under QbD, product and process development begins with pre-defined quality objectives, followed by the identification of critical quality attributes (CQAs) that have a significant impact on the drug product. Experiments are carried out to understand process variability and establish the relationships between process performance and input materials and process parameters. This knowledge serves as the basis for developing effective process controls focused on the areas of high risk. It also lends the manufacturer the ability to develop a manufacturing process that is flexible and consistent in delivering high quality product.

As outlined in ICH Q8 (R2),^[2] key steps to the implementation of QbD include: (1) defining quality target product profile (QTPP); (2) determining CQAs through risk assessment; (3) linking raw material attributes (MAs) and process parameters (PPs) to CQAs; (4) developing design space for key MAs and PPs; (5) developing and implementing effective control strategies; and (6) managing the product life cycle through continued process improvement. The overall QbD approach is illustrated in Fig. 1. The successful implementation of QbD requires significant effort in each of the above steps and relies on novel applications of statistical methods.^[3]

Quality Target Product Profile

The quality target product profile (QTPP) defines a new drug product in terms of its desired quality characteristics, and serves as the focal point for the development of the product and for all stages of the product lifecycle. Specifically, the QTPP is defined by ICH as “[a] prospective summary of the quality characteristics of a drug product that ideally will be achieved to ensure that the desired quality, taking into account safety and efficacy of the drug product.”^[2] ICH Q8 (R2) also lists some considerations for the QTPP, including: “Intended use in clinical setting, route of administration, dosage form, delivery systems; Dosage strength(s); Container closure system; Therapeutic moiety release or delivery and attributes affecting pharmacokinetic characteristics (e.g., dissolution, aerodynamic performance) appropriate to the drug product dosage form being developed; and Drug product quality criteria (e.g., sterility, purity, stability and drug release) appropriate for the intended marketed product.”^[2] While all of these aspects need to be taken into account, more attention may

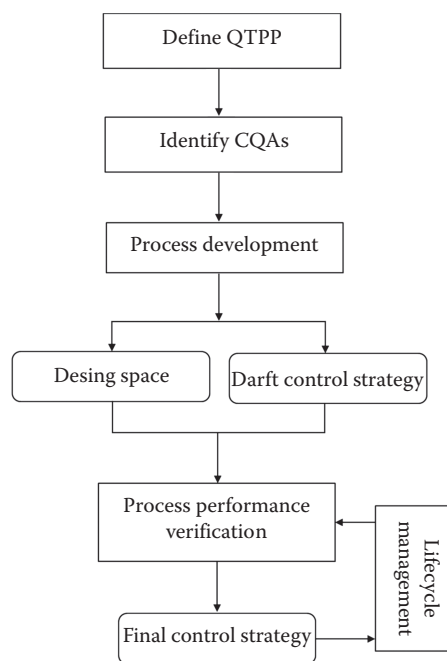


Fig. 1 QbD process development paradigm (Adapted from CMC Biotech Working Group^[4])

be given to some aspects than others for a specific product. Tools for developing a robust QTPP are discussed in literature. Of note are two white papers concerning the use of QbD for development of monoclonal antibodies and vaccines.^[4,5]

Product and Process Design

Critical Quality Attributes

The identification of CQAs follows immediately after the QTPP is developed. This allows for the evaluation and control of product characteristics that have impact on quality, safety and efficacy. A CQA is defined as “a physical, chemical, biological, or microbiological property or characteristic that should be within an appropriate limit, range, or distribution to ensure the desired product quality.”^[2] CQAs collectively indicate whether the manufacturing process delivers product that meets its QTPP. The identification of CQAs is usually accomplished through a criticality assessment that evaluates the risk associated with each attribute, including the severity (consequences to the patient), probability (likelihood of occurrence) and detectability of non-conformance of the CQA. Both prior product knowledge and process capability are useful sources of information for such an evaluation. The former includes cumulative laboratory, non-clinical, and clinical experience related to the quality attribute under evaluation, as well as data from molecules in the same class and published literature. The latter consists of data from process development studies and relevant manufacturing

experience from the same or related products. Both product knowledge and process capability serve as the basis for establishing the acceptable range for a CQA.

Since changes in CQAs might have a significant impact on product safety and efficacy, it is critical to establish the relationships between CQAs and clinical performance, and to use this knowledge to define acceptable ranges for CQAs. This knowledge provides a foundation for developing robust control strategies for the manufacturing process. However, it can be difficult to link a particular quality attribute to clinical performance (safety and efficacy).

In the literature, various models have been proposed to understand the relationships between CQAs and product safety and efficacy. For example, for process-related impurities, an impurity safety factor (ISF) is used to link to the toxicity of an impurity to the maximum acceptable level in the product through the following function:^[6]

$$\text{ISF} = \text{LD}_{50}/r$$

where LD_{50} is the amount of an impurity that results in lethality in 50% of animals tested, and r is the maximum amount of an impurity in the final product dose. The higher the ISF, the lower the safety risk caused by the impurity is. If a lower limit on the ISF is available (for example, from regulatory guidelines), it can be translated into a bound on the maximum amount of impurity in a final product dose.

Another example by Yang^[7] describes a method that provides a functional relationship between three correlated CQAs of a monoclonal antibody and a surrogate efficacy marker, namely, area under the drug concentration curve (AUC). Such a linkage allows for the assessment of the impact of changes in these CQAs on the product efficacy, and also allows for the establishment of clinically relevant joint acceptance ranges for the CQAs.

Process Development

After the CQAs are defined, the focus is shifted to the development of a manufacturing process that produces product meeting its QTPP. At this stage of development, studies are carried out, using multivariate design of experiment (DOE) strategies, to evaluate the effects of process variables and their interactions on process performance. Knowledge gleaned from the results of these studies, perhaps combined with platform knowledge and data from other molecules in the same class, will be used to define critical process parameters (CPPs) through a risk assessment. The next step is to optimize the process based on the CPPs identified from the risk assessment. This experimentation also gives rise to a “design space”, which will be defined later. The above sequence of activities is repeated for each unit operation, establishing a design space for the CPPs in each.

Process Control

The implementation of control strategies enables a manufacturing process to consistently deliver quality product. QbD principles call for using a risk-based approach to develop control strategies, aimed at keeping the manufacturing risks at or below acceptable levels. Knowledge of CQAs and the impact of CPPs and other input materials on these CQAs enables manufacturers to avoid unnecessary testing as a means to demonstrate product quality. In essence, the level of control of each quality attribute should reflect the criticality of the attribute. Effective control strategies integrate a number of elements including input material controls, procedural controls, process parameter controls, in-process testing, specification testing, characterization/comparability testing and process monitoring, as appropriate.

DESIGN SPACE

Design space is a key concept in the implementation of QbD. According to ICH Q8 (R2) (2006), a design space is “[t]he multidimensional combination and interaction of input variables (e.g., material attributes) and process parameters that have been demonstrated to provide assurance of quality.” For example, a design space for a cell culture system may have ranges for temperature, pH, feed volume, and culture duration that ensure quality product. Design space is intimately related to quality risk management (QRM) principles. By linking manufacturing variations with the variability of CQAs, it sets the stage for developing effective manufacturing controls. A design space also has regulatory implications for post approval changes as “[w]orking within the design space is not considered as a change. Movement out of the design space is considered to be a change and would normally initiate a regulatory post-approval change process”.^[2] Thus, a well-developed design space may enable a manufacturer to continuously improve the manufacturing process by adapting to novel technologies without incurring additional risk and creating more regulatory hurdles. Most recently, the concept of design space has also been used in analytical method development, which will be discussed in a later section.

STATISTICAL METHODS FOR DESIGN SPACE

Typically, statistical design of experiments (DOE) concepts are used to gain experimental efficiency in obtaining the necessary information regarding the effects of process parameters and input material attributes on the CQAs in order to establish a design space. These experiments are more effective and efficient than “one-factor-at-a-time” (OFAT) studies. Note that not all the process parameters

and material attributes have a significant impact on CQAs. Prior to the characterization studies, a risk analysis is performed to select experimental factors that are deemed to have high or medium impact on CQAs. Knowledge gained from these studies, coupled with other historical data and manufacturing experience, serves as the basis for establishing a design space, product specifications, and manufacturing control strategies.

Several statistical methods have been proposed to determine design space through multivariate regression analysis. Of note are two traditional approaches: overlapping mean response surfaces and the composite desirability function. Using these approaches, a design space is determined to be either a multivariate region in which the mean response values of CQAs are within specifications, or a region in which the composite desirability function (a measure of how well a combination of factor settings optimize the response) exceeds a pre-specified threshold. As pointed out by Peterson,^[8] these traditional methods do not take into account correlations among CQAs, nor do they account for variability in the model prediction. An appropriately constructed design space needs to account for measurement uncertainties, uncertainty about the parameters of the statistical models, and correlations among the measurements, in order to provide high assurance of product quality. In recent years, significant advances have been made in developing design spaces using Bayesian multivariate analysis techniques^[8–17] and further statistical opportunities for design space development still exist. For example, so far the concept of design space is by and large applied to input materials and process parameters. In addition, a design space is usually developed based on experiments and data from small-scale processes. How to update such design spaces, based on data from pilot and full-scale data, remains to be further researched. It will also be of great benefit to bring these statistical methods into common practice through software development.^[18]

Overlapping Mean Response Surfaces

Let $\mathbf{Y} = (Y_1, \dots, Y_p)'$ be a $p \times 1$ vector of measures of p CQAs, A the set of joint acceptable ranges of the CQAs, and $\mathbf{x}$ a $k \times 1$ vector of process parameters and other controllable inputs, such as material attributes. Suppose that the relationship between $\mathbf{Y}$ and $\mathbf{x}$ can be characterized through the model:

$$\mathbf{Y} = \mathbf{f}(\mathbf{x}; \boldsymbol{\theta}) + \boldsymbol{\varepsilon}$$

where $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta})$ is a mathematical function, $\boldsymbol{\theta}$ model parameters, and $\boldsymbol{\varepsilon}$ measurement errors.

Let $\hat{\boldsymbol{\theta}}$ be the vector of estimates of the model parameters. Therefore, the estimate of expected mean values at $\mathbf{x}$ are given by $\hat{E}[\mathbf{Y}|\mathbf{x}] = \mathbf{f}(\mathbf{x}; \hat{\boldsymbol{\theta}})$. The design space based on the overlapping mean approach is defined as

$$\{\mathbf{x} : \hat{E}[\mathbf{Y}|\mathbf{x}] \in A\}.$$

Consider an example regarding the performance of a liquid chromatography method.^[9,10] Data were collected from a Box-Behnken experiment, which has three factors, $\mathbf{x} = (x_1, x_2, x_3) = (\% \text{ isopropyl, temperature, pH})$, and four responses, $\mathbf{Y} = (y_1, y_2, y_3, y_4) = (\text{resolution, run time, signal to noise ratio, tailing})$. The aim of the experiment was to determine the design space of the assay parameters $\mathbf{x}$ that allow for high probabilities for the responses to meet specifications. For this assay, the acceptance region A for the responses is given as

$$A = \{\mathbf{y} : y_1 \geq 1.8, y_2 \leq 15, y_3 \geq 300, 0.75 \leq y_4 \leq 0.85\}.$$

LeBlond^[19] fit the data using a simple regression model:

$$\mathbf{Y} = \mathbf{z}(\mathbf{x})\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

where

$$\mathbf{z}(\mathbf{x}) = (1, x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2),$$

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_0^{(1)} & \dots & \theta_0^{(4)} \\ \vdots & \ddots & \vdots \\ \theta_7^{(1)} & \dots & \theta_7^{(4)} \end{pmatrix},$$

and $\boldsymbol{\varepsilon}$ is normally distributed with mean 0 and covariance matrix Σ . As depicted in Fig. 2, the design space for temperature and % isopropyl alcohol with pH being set at the middle level 0.175 is the area between the two contours, $y_3 = 300$ and $y_4 = 0.85$, where the condition $\hat{E}[\mathbf{Y}|\mathbf{x}] \in A$ is met.

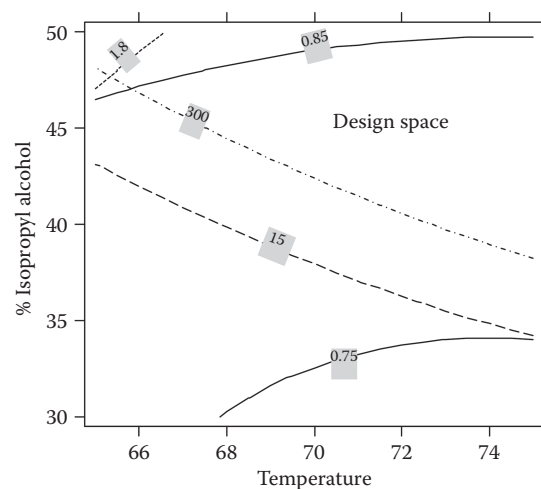


Fig. 2 Design space determined as the “sweet spot” of settings for two CPPs in which the mean responses for two CQAs meet their respective acceptance criteria (adapted from LeBlond^[19])

As noted by Peterson,^[9] there are several serious drawbacks of this approach. First, the method does not account for uncertainties in the model parameter estimates. Second, it ignores potential correlations among the responses. Third, it does not quantify the level of assurance for meeting specifications when the parameters are inside the design space.

Bayesian Approach

Definition of Bayesian Design Space

As an alternative approach, and to address the deficiencies of the overlapping mean contours and desirability function methods, Peterson^[8] suggested a Bayesian approach to establishing a design space, and an application of the method was demonstrated by Stockdale and Cheng^[14] through two case studies. A Bayesian design space and various extensions were also explored by Lebrun (2012). In general terms, a Bayesian design space is defined as

$$\{\mathbf{x} : \Pr[\mathbf{Y} \in A | \mathbf{x}, \text{data}] \geq R\} \quad (1)$$

where $\mathbf{Y}$, A , and $\mathbf{x}$ are defined as before, “Pr” stands for posterior predictive probability, *data* is the data set from a controlled experiment, which includes measured values of the CQA(s) and various settings of the process parameters, and R is a pre-selected level of reliability that must be met.

Because the posterior predictive probability takes into account the uncertainty in the model parameters and correlation among the response variables, it overcomes the drawbacks of the overlapping mean response surface method. In addition, the Bayesian method can be easily extended to accommodate many different types of experiments such as split-plot^[13] as well as experiments involving mixed effects.^[16] The Bayesian design space can be determined by estimating the probability $\Pr[\mathbf{Y} \in A | \mathbf{x}, \text{data}]$ over a grid of $\mathbf{x}$ values. The posterior predictive probability can be estimated either through a closed-form solution or by Markov Chain Monte Carlo (MCMC) simulations. The concept of a Bayesian design space is illustrated in Fig. 3 for the liquid chromatography method example described earlier.

Regression Model

Several regression models can be used to describe the measured responses of the CQAs ($\mathbf{Y}$). For example, the following model was used by Peterson^[8] and Stockdale and Cheng:^[14]

$$\mathbf{Y} = \mathbf{B}\mathbf{z}(\mathbf{x}) + \mathbf{e} \quad (2)$$

Where $\mathbf{B}$ is a $p \times q$ matrix of regression coefficients, $\mathbf{z}(\mathbf{x})$ is a $q \times 1$ vector function of $\mathbf{x}$, and $\mathbf{e}$ is a $p \times 1$

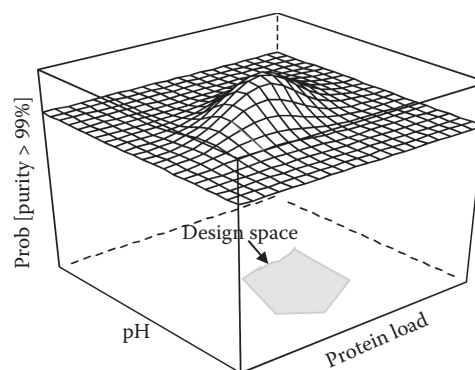


Fig. 3 Design space (DS) based on Bayesian predictive modeling that links input process parameters x and material attribute z with CQAs Y . The DS is determined such that there is a high probability that the CQAs jointly meet their acceptance criteria when the process parameters and material attributes are within the DS

vector of measurement errors having a multivariate normal distribution with mean $\mathbf{0}$ and covariance-variance matrix Σ .

It is assumed that $\mathbf{z}(\mathbf{x})$ is the same for each CQA though the method described in this section can be extended to the seemingly unrelated regressions (SUR) model (Peterson, 2006), in which each response Y_i ($i = 1, \dots, p$) has a different function $\mathbf{z}_i(\mathbf{x})$. Under such circumstances, the SUR model provides greater flexibility and accuracy in modeling the CQAs (Peterson, 2007). A SUR model takes the form:

$$Y_j = \mathbf{z}_j(\mathbf{x})'\beta_j + e_j, \quad j = 1, \dots, p.$$

The SUR model includes the standard multivariate regression model as a special case where $\mathbf{z}_j(\mathbf{x}) = \mathbf{z}(\mathbf{x})$. A design space based on the SUR model and posterior predictive probability can be similarly obtained.^[20]

Prior Information

Use of prior information (regarding process/method parameters, CQAs, etc.) is a critical step in Bayesian inference. Peterson^[8] discussed various ways in which prior information can be elicited for a design space model. For example, an informative prior distribution may be established from experiments done at the pilot scale, with the intention of being used for developing a design space for commercial scale manufacturing. However, oftentimes data from pilot scale experiments alone may not be sufficient for constructing informative priors. Peterson (2008) discussed three additional sources of information that can be potentially used for specifying informative prior distributions. Please refer to Peterson^[8] for detail.

Posterior Predictive Probability and Design Space

Let $\mathbf{Y}_{obs} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)'$ be an $n \times p$ matrix consisting of n observations of the $1 \times p$ response vector $\mathbf{Y}$.

Let $\mathbf{Z}_{\text{exp}} = (\mathbf{Z}_1, \dots, \mathbf{Z}_n)'$ be an $n \times q$ matrix with $\mathbf{Z}_i = \mathbf{z}(\mathbf{x}_i)$, $i = 1, \dots, n$, and let $\mathbf{x}_i$ be the i^{th} condition of the controllable input variables and process parameters. Assuming that a non-informative prior is used to describe the parameters $(\mathbf{B}, \Sigma)$:

$$p(\mathbf{B}, \Sigma) \propto |\Sigma|^{-(p+1)/2}. \quad (3)$$

It is well known^[21] that the posterior predictive distribution of a future observation $\tilde{\mathbf{Y}}$ in this case is a multivariate-t with degrees of freedom $\nu = n - p - q + 1$:

$$\tilde{\mathbf{Y}} | \mathbf{x}, \text{data} \sim t_{\nu}(\hat{\mathbf{B}}\mathbf{z}(\mathbf{x}), \mathbf{H}) \quad (4)$$

where

$$\mathbf{H} = [1 + \mathbf{z}(\mathbf{x})' \mathbf{D}^{-1} \mathbf{z}(\mathbf{x})] \hat{\Sigma},$$

$$\mathbf{D} = \sum_{i=1}^n \mathbf{Z}_i \mathbf{Z}_i',$$

and $\hat{\mathbf{B}}$ and $\hat{\Sigma}$ are the least squares estimates of $\mathbf{B}$ and Σ given by

$$\hat{\mathbf{B}} = (\mathbf{Z}_{\text{exp}}' \mathbf{Z}_{\text{exp}})^{-1} \mathbf{Z}_{\text{exp}}' \mathbf{Y}_{\text{obs}}$$

$$\hat{\Sigma} = [\mathbf{Y}_{\text{obs}} - (\hat{\mathbf{B}}\mathbf{Z})'] [\mathbf{Y}_{\text{obs}} - (\hat{\mathbf{B}}\mathbf{Z})'] / \nu \text{ with } \nu = n - p - q + 1.$$

Because the posterior predictive distribution is a multivariate t -distribution, it can be simulated as follows:^[20]

1. Draw $\mathbf{W}$ from the multivariate normal distribution $N(\mathbf{0}, \mathbf{H})$;
2. Draw $\mathbf{U}$ from the chi-square distribution χ_{ν}^2 ;
3. Calculate $Y_j = \sqrt{\nu} W_j / \sqrt{\mathbf{U}} + \hat{\mu}_j$, for $j = 1, \dots, p$, where Y_j , W_j , and $\hat{\mu}_j$ are the j^{th} elements of $\mathbf{Y}$, $\mathbf{W}$, and $\hat{\mathbf{B}}\mathbf{z}(\mathbf{x})$, respectively.

As suggested by Peterson (2009), the posterior predictive probability can be approximated using Monte Carlo simulation:

$$p(\mathbf{x}) = \Pr[\mathbf{Y} \in A | \mathbf{x}, \text{data}]$$

$$\approx \frac{1}{N} \sum_{s=1}^N I(\mathbf{Y}^{(s)} \in A)$$

where $\mathbf{Y}^{(s)}$, $s = 1, \dots, N$, are independent random multivariate-t variables simulated from the above-mentioned procedure, and $I(\cdot)$ is an indicator function taking values of either 0 or 1. The design space defined in (1) can be obtained by estimating the posterior predictive probability $p(\mathbf{x})$ over a grid of $\mathbf{x}$ values and comparing it to the reliability threshold R .

Alternatively, as discussed previously, informative prior distributions may be used in the derivation of the posterior predictive probability. For example, one may use conjugate prior distributions for $p(\mathbf{B} | \Sigma)$ and $p(\Sigma)$:^[16]

$$\mathbf{B} | \Sigma \sim N_{p \times q}(\mathbf{B}_0, \Sigma, \Sigma_0)$$

$$\Sigma \sim W_1^{-1}(\Omega, \nu_0)$$

Where $\mathbf{B}_0$ is the mean vector and Σ and Σ_0 are the covariance matrices of the columns and rows of $\mathbf{B}$, respectively; Σ follows an inverse-Wishart distribution with Ω being an *a priori* response scale matrix, and ν_0 the degrees of freedom.

It was shown by Lebrun (2012) that the posterior predictive probability in this case is also a multivariate t -distribution. Thus, using the Monte Carlo simulation procedure previously described, the Bayesian design space can be constructed.

Example

For the liquid chromatography method example previously discussed, LeBlond (2014) presented a design space calculation based on method (2). The design space is $\{\mathbf{x} : \Pr[\mathbf{Y} \in A | \mathbf{x}, \text{data}] \geq 0.99\}$. To calculate the posterior predictive probability, minimally informative priors b_0 , b_3 , and $b_6 \sim N(0, 0.0001)$, b_1 and $b_2 \sim N(0, 0.001)$, and b_4 , b_5 , and $b_7 \sim N(0, 0.01)$ are used for the eight regression coefficients in model (2), based on examining the results of a frequentist analysis of the data, and a scalar inverse Wishart prior is used for the variance-covariance matrix.^[18] The prior for the variance-covariance matrix is assigned based on a method suggested by Gelman and Hill.^[22] To begin, the variance-covariance matrix is first expressed as

$$\Sigma = \text{Diag}(\xi) \mathbf{Q} \text{Diag}(\xi),$$

where $\text{Diag}(\xi)$ is a diagonal matrix with diagonal elements $\xi = (\xi_1, \xi_2, \xi_3, \xi_4)$ and $\mathbf{Q}$ is a correlation matrix given by

$$\mathbf{Q} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{pmatrix}.$$

The parameters $\xi = (\xi_1, \xi_2, \xi_3, \xi_4)$ are chosen as $U(0, S_i)$ ($i = 1, \dots, 4$) with S_i being the frequentist estimate of the 99% upper confidence limit, of the square root of the i^{th} diagonal element of Σ , and $\mathbf{Q}$ is distributed according to $IW(\mathbf{I}, 5)$, where $\mathbf{I}$ is the 4×4 identity matrix. This representation of the priors has the advantage of using separate priors for the variances and the correlation coefficients. Using MCMC methods, the posterior predictive probability for a future CQA response is estimated and the resulting design space is depicted, along with the overlapping mean contours for comparison, in Fig. 4.

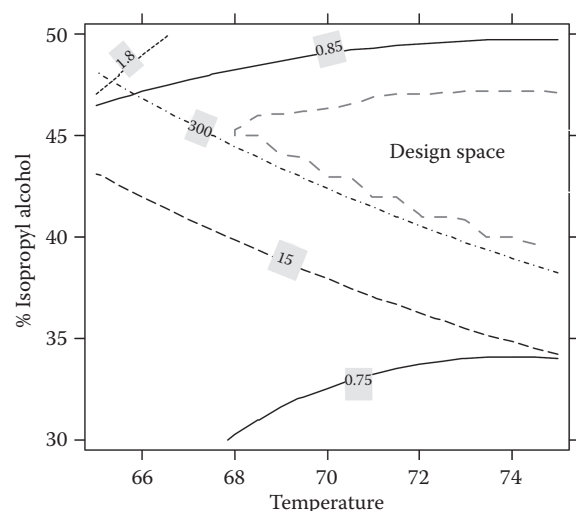


Fig. 4 Bayesian design space in which the predictive posterior probability for the responses to meet the acceptance criteria exceeds 99%. Also shown are overlapping mean contours from a regression model (adapted from LeBlond^[18])

Such a design space provides over 99% confidence that movement within the design space has no impact on the assay's ability to meet its specification.

CONTINUED PROCESS VERIFICATION

Continued process verification (CPV) is the next step in the lifecycle of the product/process, following successful process performance qualification. The goal of CPV is to demonstrate that the process continues to consistently produce product that meets its quality standard. In addition, CPV provides opportunities to either confirm that the control strategies are appropriate or identify other sources of variations that might require modification of specifications, in-process controls, or process improvements. Central to the fulfilment of these objectives is a system or systems for detecting unplanned departures from the process as designed, through monitoring and review of data collected from post approval batches. It is also critical that the data collection and analysis is performed in accordance with cGMP requirements, and that proper corrective and preventive action(s) be taken when deviations are detected. The data should include relevant product quality attributes and process parameters, incoming materials or components measurements, in-process tests. The use of statistical tools such as statistical process control (SPC) and process analytical technologies may potentially provide real-time process monitoring.

Traditionally, univariate SPC charts such as Shewhart, CUSUM, and EWMA charts have been used to monitor process performance over time.^[23] However, a major drawback of these univariate methods is that they do not account

for interdependence among the variables being monitored. As large and complex data sets including process parameters, quality attributes, properties of raw materials, and environmental factors of manufacturing areas are likely to include many and complex correlations, using independent univariate control charts may fail to identify out of control observations. The issue was illustrated by Kourti and MacGregor,^[24] as reproduced in Fig. 5. The process data of two quality attributes (y_1, y_2) are plotted, with the confidence region for (y_1, y_2) indicated by the ellipse and the respective univariate control limits by the dotted lines. It is clear that the out of control batch indicated by "*" would have been misidentified if only the univariate controls charts were used.

Multivariate analysis techniques such as principal components analysis (PCA) and partial least squares (PLS) are effective statistical tools dealing with interdependence among variables. Through reducing the dimensionality of multivariate data, these methods make monitoring of complex data manageable. They have been often used by many companies for monitoring the performance of manufacturing processes and analytical methods, qualifying batches of raw materials, demonstrating comparability pre- and post-changes.

In the following, we use an example from the a-Mab CMC case study^[4] to illustrate how to verify a scale-down model (2L) that was used to predict large scale (15, 000) production bioreactor performance, based on commercial scale batch data. This can be used as a CPV effort. Both scales use the same cell host, expression system and a similar cell culture process, and are characterized through thirteen quality attributes. A PCA was performed, and the results indicated that 99.4% of the variability in the data set was explained by the first five principal components. To demonstrate the comparability between the two scales, a 95% confidence ellipsoid for the five principal components was constructed based on the production scale data set and is shown in Fig. 6. The corresponding principal

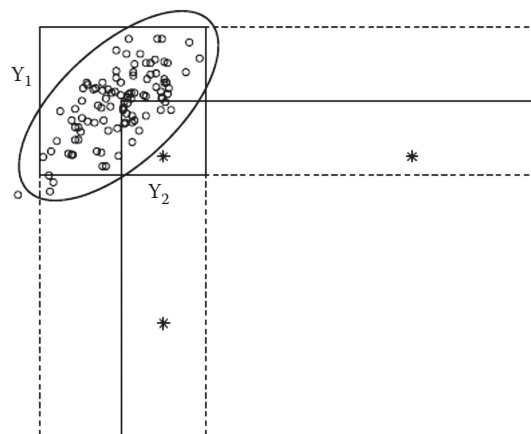


Fig. 5 Performance of multivariate control chart versus univariate control charts (adapted from Kourti and MacGregor^[24])

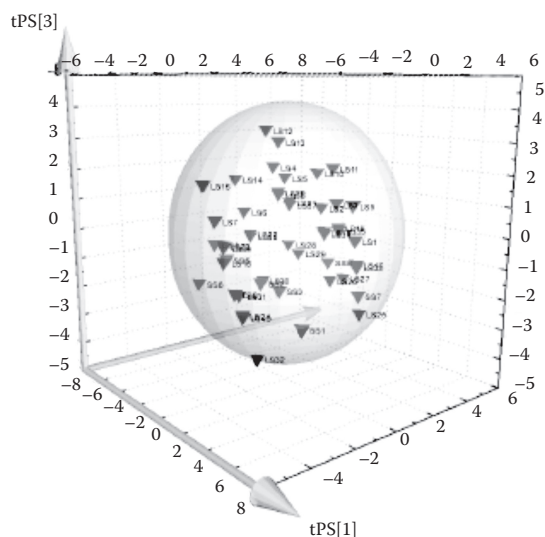


Fig. 6 Scale comparison based on PCA with black and blue triangles corresponding to small and production scales, respectively (Adopted from CMC Biotech Working Group^[4])

components of the small-scale batches are plotted and shown to be within the confidence ellipsoid, indicating comparable performance between the two scales.

QBD APPROACH TO ANALYTICAL METHOD DEVELOPMENT

Critical enablers to ensuring manufacturing consistency, analytical methods play a central role in biopharmaceutical research and development. Traditionally analytical method was developed and validated in accordance with regulatory requirements such as ICH Q2, Validation of Analytical Procedures: Methodology. For example, method validation consists of a series of experiments intended to demonstrate that an analytical method meets pre-specified acceptance criteria for accuracy, precision, linearity and etc. In recent years, there has been a shift within the industry from a compliance-driven analytical method development paradigm towards a lifecycle approach rooted in QbD. For example, Borman et al.^[25] suggested using a QbD approach for analytical methods that includes risk assessment, robustness and ruggedness testing, and assessment of method variability compared with the specification limits to determine whether the method is fit for its intended purpose. Viewing an analytical procedure as a process and data as its product, Nethercote et al.^[26] argued that QbD concepts for manufacturing process development can be applied to analytical methods, including the concepts of lifecycle validation. Analytical QbD, aligning understanding of analytical method variability with its intended use, is further explored and described in several publications.^[25,27,28] In recent years, analytical QbD also has attracted much attention from USP and FDA.

The culmination of these efforts is the proposal that the three stages of process validation be applied to analytical method validation (Nethercote and Ermer, 2012; Martin et al. 2014), as depicted in Fig. 7.

The starting point of this approach is to define an analytical target profile (ATP), which describes the quality of the data that the method is required to produce. The concept of ATP was first proposed by the EFPIA/PhRMA working group^[29] and adopted by the USP Validation and Verification Expert Panel.^[28] The ATP is constructed based on understanding the measurement uncertainty of the method and measurand, and the required level of assurance regarding the reportable results. Several considerations need to be taken into account when defining the ATP.^[30] They include (1) the range over which the analytical method is expected to reliably quantify the amount or potency of analyte; (2) the total acceptable uncertainty or error, which measures the closeness between the reportable value and the true value, and which is expressed as the sum of systematic and random errors; and (3) a description of the analyte to be tested, including the sample or matrix in which it will be tested. The ATP is used to guide method development, performance qualification, and continued verification throughout the lifecycle of the method. It is critical to link the traditional assay characteristics such as accuracy and precision to the ATP. Such a link enables the definition of acceptance criteria for these performance parameters in the method performance qualification study.

After the ATP is defined, the analyst proceeds to Stage 1 Method Design, in which the analyst selects an appropriate analytical technique and develops a draft method that may meet the requirements defined in the ATP. As understanding sources of variability is one of the primary focuses at this stage, risk assessment tools such as Ishikawa or fishbone diagrams, FEMA, etc., can be used to gain understanding of variables that may have an impact on the method performance, and help prioritize experimentation.^[28] DOEs are carried out to determine ranges of key method parameters in which the ATP may be met. Stage 2 Method Qualification is intended to confirm that the method is fit for its intended purpose through formal studies. This stage

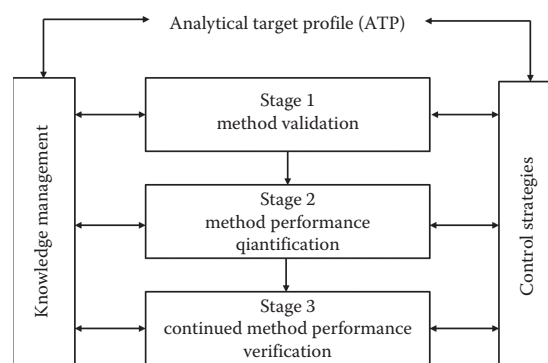


Fig. 7 Life cycle approach to analytical procedure validation

involves the collection of data from performing the analytical procedure under routine operating conditions, and demonstration that the data meet acceptance criteria in the ATP. In addition, knowledge of the analytical procedure gleaned from the method qualification studies can be used to develop control strategies. The objective of Stage 3 Continued Procedure Performance Verification is to provide assurance that the method remains in a state of control during routine use. This involves routine monitoring to confirm the analytical procedure's performance, and also performance verification after any changes made to the method. Statistical process control charts can be used to capture any special-cause variation.

CONCLUSIONS

Since the launch of the FDA initiative "Pharmaceutical cGMPs for the 21st Century" in 2002, there has been a significant progress in developing QbD approaches for biopharmaceutical products. Biopharmaceutical product and process development, due to the high level of complexity, would benefit from greater use of available data and tools for risk assessment and development of design space and control strategies. How effectively to synthesize information from various sources, quantify manufacturing variability and risk, identify CQAs, determine design space, and devise effective control strategies is the key to the implementing a QbD approach. All of these rely on statistical methodologies for experimental design and data analysis. Statistical advances, particularly in the area of Bayesian modeling and computation have brought about new opportunities for applying QbD principles in biopharmaceutical development.

ACKNOWLEDGEMENT

The author would like to thank Ms. Katherine Giacoletti and Dr. Lorin Roskos for their helpful review of the article.

REFERENCES

1. Yu, L.X., Amidon, G., Khan, M.A., Hoag, S.W., Polli, J., Raju, J.K., and Woodcock, J. Understanding pharmaceutical quality by design. *The AAPS Journal* **2015**, *16* (4), 771–783.
2. ICH Q8(R2) Pharmaceutical Development, 2006.
3. Yang, H. Emerging Non-Clinical Biostatistics in Biopharmaceutical Development and Manufacturing. Chapman & Hall/CRC: Boca Raton, FL, 2016.
4. CMC Biotech Working Group. A-Mab: A Case Study in Bioprocess Development, 2009. www.casss.org/associations/9165/.../A-Mab_Case_Study_Version_2-1.pdf.
5. CMC-Vaccine Working Group. A-VAX: Applying Quality by Design to Vaccines, 2012. www.ispe.org
6. Schenerman, M.A.; Axley, M.J.; Oliver, C.; Ram, K.; Wasserman, G.F. Using a risk assessment process to determine criticality of product quality attributes. In *Quality by Design for Biopharmaceutical: Principles and Case Studies*; Rathore, A.S.; Mhatre, R., Eds.; Wiley: Hoboken, NJ, 2009.
7. Yang, H. Setting specifications of correlated quality attributes. *PDA J. of Pharm. Science and Technology* **2013**, *67*, 533–543.
8. Peterson, J. J. A Bayesian approach to the ICH Q8 definition of design space. *Journal of Biopharmaceutical Statistics* **2008**, *18*, 959–975.
9. Peterson, J. J. A Posterior predictive approach to multiple response surface optimization. *Journal of Quality Technology* **2004**, *36*, 139–153.
10. Peterson, J. J. What your ICH Q8 design space needs: a multivariate predictive distribution. *Pharmaceutical Manufacturing* **2009**, *8* (10), 23–28.
11. Peterson, J. J.; Miró-Quesada, G.; del Castillo, E. A Bayesian Reliability Approach to Multiple Response Optimization with Seemingly Unrelated Regression Models. *Quality Technology and Quality Management* **2009**, *6* (4), 353–369.
12. Peterson, J. J.; Yahyah, M. A Bayesian design space approach to robustness and system suitability for pharmaceutical assays and other processes. *Statistics in Biopharmaceutical Research* **2009**, *1* (4), 441–449.
13. Peterson, J.J.; Snee, R.D.; McAllister, P.R.; Schoefield, T.L.; Carella, A.J. Statistics in pharmaceutical development and manufacturing (with discussion). *Journal of Quality Technology* **2009**, *41*, 111–147.
14. Stockdale, G.; Cheng, A. Finding Design Space and a Reliable Operating Region using a Multivariate Bayesian Approach with Experimental Design. *Quality Technology and Quantitative Management* **2009**, *6* (4), 391–408.
15. Peterson, J. J.; Lief, K. The ICH Q8 Definition of Design Space: A Comparison of the Overlapping Means and the Bayesian Predictive Approaches. *Statistics in Biopharmaceutical Research* **2010**, *2*, 249–259.
16. LeBrun, P. Bayesian design space applied to pharmaceutical. Unpublished dissertation. University de Liege, 2012. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=3&ved=0ahUKewijkrXoq8HKAhX-Ilx4KHxjoDEYQFggoMAI&url=https%3A%2F%2Forbi.ulg.ac.be%2Fbitstream%2F2268%2F126503%2F1%2Fthesis.pdf&usg=AFQjCNG-IxFztbhvUfL_3qMwMqKX-H4LvZw&bvm=bv.112454388,d.dmo.
17. Lebrun, P.; Giacoletti, K.; Scherder, T.; Rozet, E.; Boulanger, B. A quality by design approach for longitudinal attributes. *Journal of Biopharmaceutical Statistics* **2015**, *25*, 247–259.
18. Hofer, D. Discussion. *Journal of Quality Technology* **2009**, *41* (2), 137–139.
19. LeBlond, D. Applied non-clinical Bayesian Statistics. Presented at 2014 Nonclinical Biostatistics Conference, Villanova, PA, October 13, 2015.
20. Peterson, J. J. A Review of Bayesian Reliability Approaches to Multiple Response Surface Optimization. In *Bayesian Process Monitoring, Control & Optimization*, Chapman & Hall/CRC: Boca Raton, FL, 2007.
21. Press, S. J. *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*. R. E. Krieger Publication Co.: Malabar, FL, 1982.

22. Gelman, A.; Hill, J. *Data Analysis Using Regression and Multilevel Hierarchical Models*. Cambridge University Press: New York, 2006.
23. Montgomery, D.C. *Introduction to Statistical Quality Control*. 6th Ed.; John Wiley & Sons: New York, 2008.
24. Kourti, T.; MacGregor, J.F. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems* **1995**, *28*, 3–21.
25. Borman, P.; Nethercote, P.; Chaffield, M.; Thompson, D.; Truman, K. The Application of quality by design to analytical methods. *Pharm. Technol.* **2007**, *31* (10), 142–152.
26. Nethercote, P.; Borman, P.; Bennett, T.; Martin, G.; McGregor, P. QbD for better method validation and transfer. *Pharmaceutical Manufacturing*, 2010, <http://www.pharmamanufacturing.com/articles/2010/060/> (accessed on April 17, 2016).
27. Nethercote, P.; Ermer, J. Quality by design for analytical methods: implications for method validation and transfer. *Pharm. Technol.* **2012**, *36* (10), 74–79.
28. Martin, G.P.; Barnett, K.L.; Burgess, C.; Curry, P.D.; Ermer, J.; Gratzl, G.S.; Hammond, J.P.; Herrmann, J.; Kovacs, E.; LeBlond, D.J.; LoBrutto, R.; McCasland-Keller, A.K.; McGregor, P.L.; Nethercote, P.; Templeton, A.C.; Thomas, D.P.; Weitzel, J. Stimuli to the revision process: lifecycle management of analytical procedures: method development, procedure performance quantification, and procedure performance verification. *Pharmaceutical Forum* **2013**, *39* (5). <http://www.usp.org/uspnf/stimuli-article-life-cycle-management-analytical-procedures-posted-comment>.
29. EFPIA/PhRMA. Pharmaceutical Research and Manufacturers of America Analytical Technical Group and European Federation of Pharmaceutical Industries and Associations (EFPIA) Analytical Design Space Topic Team. Implications and opportunities of applying QbD principles to analytical measurements. *Pharm. Technol.* **2010**, *34* (2), 52–59.
30. Colgan, C.; Hanna-Brown, M.; Pellett, J.; Wrisley, L.; Sluggett, G.; Morgado, J.; Harrington, B.; Szucs, R.; Barnett, K.; Graul, T.; Steeno, G. Using Quality by Design to develop robust chromatographic methods. *Pharmaceutical Technology* **2014**, *38* (8), <http://www.pharmtech.com/using-quality-design-develop-robust-chromatographic-methods> (accessed on April 14, 2016).

Randomization

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Jun Shao

University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.

Abstract

Randomization plays an important role in the conduct of clinical trials. This entry discusses several models and methods of clinical trial randomization and reviews the statistical effects randomization has on clinical trials as well as examples of when randomization is not feasible.

INTRODUCTION

Randomization plays an important role in the conduct of clinical trials. Randomization not only generates comparable groups of patients who constitute representative samples from the intended patient population but also enable valid statistical tests for clinical evaluation of the study drug. Randomization in clinical trials involves random recruitment of the patients from the targeted population and random assignment of patients to the treatments. For a valid statistical assessment of the efficacy and safety of a study drug, it is important that a representative sample of qualified patients be randomly selected from the targeted patient population. A qualified patient is referred to as a patient who meets the inclusion and exclusion criteria and has signed the informed consent form. Randomization avoids subjective selection bias for the integrity and the scientific and/or statistical validity of the intended clinical trials. Patients participating in the clinical trials are randomly assigned to one of the treatments under study which avoids subjective assignment of treatments. When there are heterogeneity in demographics and/or patient characteristics, randomization with blocking and/or stratification is helpful in removing the potential bias that might occur due to the differences in demographics and/or patient characteristics. Under randomization, statistical inference can be drawn under some probability distribution assumption of the intended patient population. The probability distribution assumption depends on the method of randomization under a randomization (population) model. A study without randomization will result in the violation of the probability distribution assumption; consequently, no accurate and reliable statistical inference on the study drug can be drawn.

RANDOMIZATION MODELS

In practice, the selection of patients may be performed under different principles of randomization. Lachin^[1] provided a comprehensive summarization of randomization basis for statistical tests under various models which are summarized in this section.

Population Model

The concept for selecting a representative sample from a patient population by some random procedures for obtaining statistical inference by which clinicians can draw conclusions regarding the patient population is called the *population model*.^[1,2] Under the assumption of a homogeneous population, the clinical responses of all patients in the trial are independent and have the same distribution. Lachin^[1] pointed out that the significance level and power will not be affected by random assignment of patients to the treatments as long as the patients in the trial represents a random sample from the homogeneous population. If we split a random sample into two subsamples, statistical inferential procedures are still valid even if the first half patients are assigned to the study drug and the other half are assigned to a placebo.

Invoked Population Model

When planning a clinical trial, we usually select investigators first and then recruit/enroll qualified patients at each selected investigator's site sequentially. As a result, neither the selection of investigators (or study centers) nor the recruitment of patients is random. However, patients who enter a trial are assigned to treatment groups at random. In practice, the collected clinical data

are usually analyzed as if they were obtained under the assumption that the sample is randomly selected from a homogeneous patient population. Lachin^[1] referred to this process as the *invoked population* model because the population model is invoked as the basis for statistical analysis as if a formal sampling procedure were actually performed. In current practice, the invoked population model is commonly employed in data analysis for most clinical trials. However, it should be noted that the invoked population model is based on the assumption that is inherently instable.

Randomization Model

For current practice, although the process for study site selection and for patient recruitment are not random, the assignment of treatments to patients is performed based on some random mechanism. Thus, treatment comparisons can be made based on the so-called randomization or permutation tests. We will refer this process as the *randomization model*. Because the statistical procedures based on the concept of permutation require the enumeration of all possible permutations, it is feasible only for small samples. When the sample size increases, the probability distribution derived under permutation will approach to a normal distribution according to the Central Limit Theorem. Lachin^[3] indicated that the probability distributions for the family of linear rank statistics for large samples are equivalent to those of the tests obtained under the assumption of the population model. As a result, if patients are randomly assigned to the treatments, statistical tests for evaluation of treatments should be based on permutation tests because the exact *p*-value can be easily calculated for small samples. For large samples, the data can be analyzed by the methods derived under the population model as if the patients were randomly selected from a homogeneous population.

RANDOMIZATION METHODS

In general, randomization methods for treatment assignment can be classified into three types according to the restriction of the randomization and the change in probability for randomization with respect to the previous treatment assignments. These randomization methods include the complete randomization, the permuted-block randomization, and the adaptive randomization. In what follows, we will describe these randomization methods and compare their relative merits and limitations whenever possible.

Complete Randomization

Simple randomization or complete randomization is the procedure in which no restrictions are enforced on the

nature of randomization sequence except for the number of patients required for achieving the desired statistical power and the ratio of patient allocation between treatments. For a clinical trial comparing a study drug with a placebo, the method of simple randomization is referred to as a completely binomial design^[4] or a simply complete randomization^[3] if it has the following properties: (i) the chance that a patient receives either the test drug or the placebo is 50%, and (ii) randomization of assignments are performed independently for each patient. The randomization codes based on the method of complete randomization can be generated either by random numbers^[5] or by some statistical computing software such as SAS. However, it should be realized that a computer cannot generate *true* random numbers, but pseudo random numbers, because only a fixed number of different long series of almost unpredictable permuted numbers is generated. Therefore, Lehmann^[2] recommended that a run test be performed to verify the randomness of the generated randomization codes.

Permuted-Block Randomization

One of the major disadvantages of the simple randomization is that treatment imbalance may occur periodically. One resolution to this major disadvantage of simple randomization is to force balance of the number of patients assignment to each treatment periodically. In other words, we first divide the whole series of patients who are to enroll into the trial into blocks with appropriate (typically equal) blocking sizes. We then randomize the patients within each block to ensure that an equal number of patients will be assigned to each of the two treatment groups. For example, a blocking size of two guarantees that each treatment group will be assigned one patient for the first two patients enrolled into the study. A blocking size of four indicates that two patients of the first four patients who enter the study will be randomly assigned to each of the two treatment groups. This method of randomization is known as the *permuted-block randomization*. Permuted-block randomization is probably the most frequently employed method for the assignment of patients to treatments in clinical trials.

Permuted-block randomization is in fact a stratified randomization. Therefore, a stratified analysis should be performed with blocks as strata to properly control the overall type I error rate and, hence, to provide the optimal power for detection of a possible treatment effect. In practice, the block effect is usually ignored when performing data analysis. Matts and Lachin^[6] showed that the test statistic ignoring the block is equal to 1 minus the intrablock correlation coefficient times the test statistic which takes the block into account. Because the patients within the same block are usually more homogeneous than those between blocks, the intrablock correlation coefficient is often positive. As a result, the tests which ignore block

will produce more conservative results. However, because most clinical trials are multicenter studies with a moderate block size, it is not clear from their discussion whether a saturated model with factors treatment, center, and block and their corresponding two-factor and three-factor interactions should be included in the analysis of variance. It is not clear whether the Mantel–Haenszel test and linear rank statistics based on permutation model are adequate for a complete combination of all levels of all strata with a very small number of patients in each stratum. In addition, when block size is large, there is a high possibility that a moderate number of patients fall in the last block at which the randomization codes are not entirely used up. If the trial is also stratified according to some covariates, the number of patients in such incomplete block will become quite sizable. Therefore, a stratified analysis will be quite complicated because of these incomplete blocks from all strata and the possible presence of intrablock correlation coefficient.

A practical issue related to permuted-block randomization is the choice of block size. A small blocking size, such as two, may prevent treatment imbalance; however, it may result in a high intrablock correlation coefficient and, thus, a large loss of power. Note that a smaller block size leads to a higher probability of guessing the treatment code right, which may decrease the reliability of blinding and, consequently, the integrity of the study.

Adaptive Randomization

In addition to the permuted-block randomization, it is also of interest to adjust the probability of assignment of patients to treatments during the study. This type of randomization is called the *adaptive randomization* because the probability for the assignment of current patient is adjusted based on the assignment of previous patients. Unlike other randomization methods described above, the randomization codes based on the method of adaptive randomization cannot be prepared before the conduct of the study because the randomization process is actually performed at the time when a patient is enrolled into the study and that the adaptive randomization requires the information of previous randomized patients. In clinical trials, the method of adaptive randomization is often applied with respect to treatment, covariate, or clinical response. Therefore, the adaptive randomization is also known as the treatment adaptive randomization, the covariate adaptive randomization, or the response adaptive randomization. Treatment adaptive randomization has been discussed in literature by many researchers (see, e.g., Efron,^[7] Wei,^[8,9] and Wei and Lachin^[10]). The response adaptive randomization has also received much attention since introduced by Zelen^[11] (see, e.g., Wei et al.^[12] and Rosenburger and Lachin^[13]). More details regarding the covariate adaptive randomization can be found in Taves,^[14] Pocock,^[5] and Spilker.^[15]

EFFECT OF MIXED-UP RANDOMIZATION CODES

A problem that commonly encountered during the conduct of a clinical trial is that a proportion of treatment codes are mixed up in randomization schedules. Mixing up treatment codes can distort the statistical analysis based on the population or randomization model. In what follows, we introduce a method proposed by Chow and Shao^[16] to quantitatively study the effect of mixed-up treatment randomization codes.

Consider a two-group parallel design for comparing a test drug and a control (placebo), where n_1 patients are randomly assigned to the treatment group and n_2 patients are randomly assigned to the control group. When randomization is properly applied, the population model holds and responses from patients are normally distributed. Consider the simple case where two patient populations (treatment and control) have the same variance σ^2 . Let μ_1 and μ_2 be the population means for the treatment and the control, respectively. The null hypothesis that $\mu_1 = \mu_2$ (i.e., there is no treatment effect) is rejected at the 5% level of significance if

$$T = \frac{|\bar{x}_1 - \bar{x}_2|}{\hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{0.975; n_1 + n_2 - 2} \quad (1)$$

where

$$\hat{\sigma}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$\bar{x}_1$ and s_1^2 are the sample mean and variance of responses from patients in the treatment group, respectively; $\bar{x}_2$ and s_2^2 are the sample mean and variance of responses from patients in the control group, respectively; and $t_{0.975; n_1 + n_2 - 2}$ is the 97.5th percentile of the t distribution with $n_1 + n_2 - 2$ degrees of freedom. The power of the test defined by Eq. 1, i.e., the probability of correctly detecting a treatment difference when $\mu_1 \neq \mu_2$, is

$$\begin{aligned} p(\theta) &= P(T > t_{0.975; n_1 + n_2 - 2}) \\ &= 1 - \mathcal{T}_{n_1 + n_2 - 2}(t_{0.975; n_1 + n_2 - 2} | \theta) \\ &\quad + \mathcal{T}_{n_1 + n_2 - 2}(-t_{0.975; n_1 + n_2 - 2} | \theta) \end{aligned}$$

where

$$\theta = \frac{\mu_1 - \mu_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

and $\mathcal{T}_{n_1 + n_2 - 2}(\cdot | \theta)$ is the noncentral t distribution function with $n_1 + n_2 - 2$ degrees of freedom and the noncentrality

parameter θ . This follows from the fact that under the population model, $\bar{x}_1 - \bar{x}_2$ has the normal distribution with mean $\mu_1 - \mu_2$ and variance $\sigma^2[(1/n_1) + (1/n_2)]$; $(n_1 + n_2 - 2)\hat{\sigma}^2$ has the chi-square distribution with $n_1 + n_2 - 2$ degrees of freedom; and $\bar{x}_1 - \bar{x}_2$ and $\hat{\sigma}^2$ are independent. When $n_1 + n_2$ is large,

$$p(\theta) \approx \Phi(\theta - 1.96) + \Phi(-\theta - 1.96)$$

where Φ is the standard normal distribution function.

Suppose that in each treatment group, patients independently have probability p to mix up their treatment codes. A straightforward calculation shows that the mean of $\bar{x}_1 - \bar{x}_2$ is

$$E(\bar{x}_1 - \bar{x}_2) = (1 - 2p)(\mu_1 - \mu_2)$$

and the variance of $\bar{x}_1 - \bar{x}_2$ is

$$\text{Var}(\bar{x}_1 - \bar{x}_2) = \left[\sigma^2 + p(1-p)(\mu_1 - \mu_2)^2 \right] \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Under the null hypothesis that $\mu_1 = \mu_2$, the distributions of $\bar{x}_1 - \bar{x}_2$ and $\hat{\sigma}^2$ are the same as those in the case of no mixed-up treatment codes; hence, mixing up codes does not have any effect on the significance level of the test defined by Eq. 1. When $\mu_1 \neq \mu_2$ and $p > 0$, $\bar{x}_1 - \bar{x}_2$ is no longer normally distributed but is approximately normal when both n_1 and n_2 are large. Also, $\hat{\sigma}^2$ is a consistent estimator of $\sigma^2 + p(1-p)(\mu_1 - \mu_2)^2$ when $n_i \rightarrow \infty$ for $i = 1, 2$. Thus, the power of the test defined by Eq. 1 is approximately

$$p(\theta_m) = \Phi(\theta_m - 1.96) + \Phi(-\theta_m - 1.96) \quad (2)$$

where

$$\begin{aligned} \theta_m &= \frac{(1-2p)(\mu_1 - \mu_2)}{\sqrt{\left[\sigma^2 + p(1-p)(\mu_1 - \mu_2)^2 \right] \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \\ &= \frac{(1-2p)\theta}{\sqrt{1 + \frac{p(1-p)(\mu_1 - \mu_2)^2}{\sigma^2}}} \\ &= \frac{(1-2p)\theta}{\sqrt{1 + p(1-p)\theta^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \end{aligned}$$

(assuming that $0 < p < 0.5$). Note that $\theta_m = \theta$ if $p = 0$; that is, there is no mixed-up treatment codes.

The effect of mixed-up treatment codes can be measured by comparing $p(\theta)$ with $p(\theta_m)$. Suppose

that $n_1 = n_2 = 100$, and when there is no mixed-up treatment codes, $p(\theta) = 80\%$, which gives that $|\mu_1 - \mu_2|/\sigma = 0.397$ and $|\theta| = 2.81$. When 5% of treatment codes are mixed up, i.e., $p = 5\%$, $p(\theta_m) = 71.23\%$. When $p = 10\%$, $p(\theta_m) = 60.68\%$. Hence, a small proportion of mixed-up treatment codes may seriously affect the probability of detecting treatment effect when such an effect exists. The effect of mixed-up treatment codes is higher when n_1 and n_2 are not large. Table 1 lists simulation approximations to the values of power of the test defined by Eq. 1 when $n_1 = n_2 = n_0$ is 10, 15, 20, or 30 and $p = 1\%$, 3%, 5%, or 10%. It can be seen that when $n_0 \leq 30$, the actual power is substantially lower than that given by formula 2 with $n_0 = 100$.

When $n_1 = n_2 = n_0$ is large, we may plan ahead to ensure a desired power when the maximum proportion of mixed-up treatment codes is known. Assume that the maximum proportion of mixed-up treatment codes is p and the desired power is 80%, i.e., $|\theta| = 2.81$. Let n_w be the sample size for which the power is equal to 80% when there are 5% of mixed-up treatment codes. Then

$$\frac{(1-2p) \left| \frac{\mu_1 - \mu_2}{\sigma} \right| \sqrt{\frac{n_w}{2}}}{\sqrt{1 + p(1-p) \left| \frac{\mu_1 - \mu_2}{\sigma} \right|^2}} = \left| \frac{\mu_1 - \mu_2}{\sigma} \right| \sqrt{\frac{n_0}{2}}$$

which is the same as

$$(1-2p)^2 n_w = n_0 \left[1 + p(1-p) \left| \frac{\mu_1 - \mu_2}{\sigma} \right|^2 \right] = n_0 + 2p(1-p)\theta^2$$

Therefore,

$$n_w = \frac{n_0}{(1-2p)^2} + \frac{2p(1-p)\theta^2}{(1-2p)^2}$$

will maintain the desired power when the proportion of mixed-up treatment codes is no larger than p . For example, if $p = 5\%$ and 80% power is desired, then $\theta^2 = 2.81^2$ and $n_w = 1.235n_0 + 1.026$. Hence, $n_w = 1.24n_0$; that is, a 24% increase of the sample size will offset a 5% mix-up in randomization schedules.

When n_1 and n_2 are small, it is difficult to estimate a new sample size to ensure a desired power even when the maximum proportion of mixed-up treatment codes is known. A simulation study may be helpful. For example, suppose that $n_1 = n_2 = n_0 = 10$ and $(\mu_1 - \mu_2)/\sigma = 1.26$, which results in a power of 75.84% when $p = 0$ (Table 1). If $p = 5\%$, then the power is 63.59%. According to the results in Table 1, the power is 74.1% when $p = 5\%$ and $\theta = 3.24$. Hence, we need to increase the sample size so that $\theta = 3.24$ in order to ensure a power comparable to the original power 75.84%; that is, the new sample size should be $n_w = 2(3.24/1.26)^2 = 13.22$. Thus, a recommended new sample size is $n_1 = n_2 = 14$.

Table 1 Power of the Two-Sample t Test Approximated by Monte Carlo of Size 10,000

n_0	$(\mu_1 - \mu_2)/\sigma$	θ	p (%)				
			0	1	3	5	10
10	1.26	2.81	0.7584	0.7372	0.6844	0.6359	0.5156
	1.45	3.24	0.8671	0.8490	0.8015	0.7410	0.6120
15	1.03	2.81	0.7727	0.7518	0.7052	0.6650	0.5449
	1.19	3.24	0.8751	0.8647	0.8261	0.7818	0.6602
20	0.89	2.81	0.7828	0.7649	0.7170	0.6796	0.5619
	1.03	3.24	0.8903	0.8549	0.8378	0.7918	0.6762
30	0.73	2.81	0.7892	0.7747	0.7429	0.7032	0.5834
	0.84	3.24	0.8945	0.8791	0.8423	0.8101	0.6963

CONCLUSION

In some situations, deliberate unequal allocation of patients between treatment groups may be desirable. For example, it may be of interest to allocate patients to the treatment and a control in a ratio of 2:1. Such situations include the following: (i) the patient population is small; (ii) previous experience with the study drug is limited; (iii) the response profile of the competitor is well known; and (iv) there are missing values, and the rates of missing depend on the treatment groups.

Randomization is one of the key elements for the success of clinical trials intended to address scientific and/or medical questions. However, it should be noted that in many situations, randomization may not be feasible in clinical research. For example, nonrandomized observational or case-controlled studies are often conducted to study the relationship between smoking and cancer. However, if the randomization is not used for some medical considerations, the Food and Drug Administration (FDA) requires that statistical justification should be provided with respect to how systematic selection bias can be avoided.

REFERENCES

- Lachin, J.M. Statistical properties of randomization in clinical trials. *Control. Clin. Trials* **1988**, *9*, 289–311.
- Lehmann, E.L. *Nonparametrics: Statistical Methods Based on Ranks*; Holden Day: San Francisco, CA, 1975.
- Lachin, J.M. Properties of simple randomization in clinical trials. *Control. Clin. Trials* **1988**, *9*, 312–326.
- Blackwell, D.; Hodges, J.L. Jr., Design for the control of selection bias. *Ann. Math. Stat.* **1957**, *28*, 449–460.
- Pocock, S.J. *Clinical Trials: A Practical Approach*; Wiley: New York, 1983.
- Matts, J.P.; Lachin, J.M. Properties of permuted-block randomization in clinical trials. *Control. Clin. Trials* **1988**, *9*, 327–344.
- Efron, B. Forcing a sequential experiment to be balanced. *Biometrika* **1971**, *58*, 403–417.
- Wei, L.J. A class of designs for sequential clinical trials. *J. Am. Stat. Assoc.* **1977**, *72*, 382–386.
- Wei, L.J. The adaptive biased-coin design for sequential experiment. *Ann. Stat.* **1978**, *9*, 92–100.
- Wei, L.J.; Lachin, J.M. Properties of the urn randomization in clinical trials. *Control. Clin. Trials* **1988**, *9*, 345–364.
- Zelen, M. Play the winner rule and the controlled clinical trial. *J. Am. Stat. Assoc.* **1969**, *64*, 131–146.
- Wei, L.J.; Smythe, R.T.; Lin, D.Y.; Park, T.S. Statistical inference with data-dependent treatment allocation rules. *J. Am. Stat. Assoc.* **1990**, *73*, 840–843.
- Rosenburger, W.F.; Lachin, J.M. The use of response-adaptive designs in clinical trials. *Control. Clin. Trials* **1993**, *14*, 471–484.
- Taves, D.R. Minimization: a new method of assessing patients to treatment and control groups. *Clin. Pharmacol. Ther.* **1974**, *15*, 443–453.
- Spilker, B. *Guide to Clinical Trials*; Raven Press: New York, 1991.
- Chow, S.C.; Shao, J. *Statistics in Drug Research—Methodology and Recent Development*; Marcel Dekker, Inc.: New York, 2002.

Randomized Trials with Clusters: Design and Analysis

Sin-Ho Jung, Ph.D

Department of Biostatistics and Bioinformatics Duke University, Durham,
North Carolina, U.S.A.

Abstract

In clustered data, each cluster consists of multiple subjects, called subunits, and observations are collected from the subjects of each cluster. Since the subjects in each cluster share some common frailties, their outcome data are usually positively correlated. In a randomized study, we can consider one of two randomization strategies. The first strategy, called cluster randomization, is to randomize clusters among treatment arms. In this case, all subjects within each cluster receive the same treatment, so that dependency of the outcome data exists within each treatment arm only and the outcome data among different arms are independent. The second strategy, called subunit randomization, is to randomize the subjects within each cluster between different treatment arms. In this case, subjects within each cluster receive different treatments as randomized, so that dependency of the outcome data exists both within each treatment arm and across different treatment arms. In comparing the mean outcome level among different arms, we have to appropriately account for the dependency in both the sample size calculation and the subsequent statistical analysis. In this chapter, we investigate some estimators of binary response rate and test statistics for comparing the response rates between two arms using these estimators. We consider both cluster and subunit randomization strategies and we assume that the number of subunits is possibly random among clusters.

Keywords: Chi-square statistic; Cluster randomization; Intraclass correlation coefficient; Optimal weight; Subunit randomization; Varying cluster size.

1 INTRODUCTION

Clustered data occur frequently in many fields of application. Examples include the development of tumors in one or more animals of a litter, the presence of retinitis in either or both eyes of an AIDS patient, and community trials for vaccine development. In these examples, litters, AIDS patients, and communities are clusters, and the observations are made from animals, eyes, and community members which are called subunits. For the practical purpose, we often randomize clusters between different treatment arms. In this case, while the clusters are independent, the subunits within each cluster are not. Consequently, the standard test statistics to compare between treatment arms using independent observations are not valid.

In the presence of cluster effect, the application of Pearson chi-square statistic is not valid to compare the response rates among treatment arms. Donner^[1] and Donner and Banting^[2] proposed an adjusted chi-square statistic by assigning an equal weight to each subunit for testing the homogeneity of response rates from such clustered data. Lee and Dubin^[3] and Lee^[4] proposed a weighted estimator which assigns the same weight to each cluster regardless of the number of subunits in each cluster, called cluster size. These two weighting methods are identical if the cluster

size is constant among clusters. Jung and Ahn^[5] introduced an optimal weight estimator that minimizes the variance of the response probability.

Sometimes, we randomize the subunits of each cluster among different treatment arms, which is called subunit randomization. An example arises from a clinical trial conducted to evaluate influence of HL-A (human lymphocyte antigen) on survival of allogeneic grafts (defined as time to rejection of graft) in burned patients^[6]. The donor and recipient were first matched for ABO blood types and then either closely or poorly matched for HL-A antigens. Selected skin grafts (subunits) from at least two unrelated types of donors were applied to each patient (cluster), so that each patient has skin grafts of both close and poor HL-A match. The objective of the study was to discover whether avoiding severe HL-A incompatibility will extend allograft survival time.

In this chapter, we assume that we observe a binary response from each subunit, and propose modified chi-square test statistics accounting for the cluster effect under both cluster randomization and subunit randomization. If the cluster sizes are random, we can consider three different weighting systems to estimate the response probability for each treatment group, (i) equal weights to subunits, (ii) equal weights to clusters, and (iii) optimal weights.

Readers may refer to Donner and Klar^[7] for an excellent review of history, design issues, statistical methods of cluster randomization trials.

2 MODIFIED χ^2 TESTS

We restrict our discussion in this chapter to the comparison of response rates between two arms, but extension to more than two arms is straightforward. Under each of cluster or subunit randomization, we investigate modified χ^2 test accounting for cluster effect by using three different weighting methods. These three test statistics are identical if the cluster size is identical for all clusters.

2.1 Under Cluster Randomization

Suppose that n_k clusters are randomized to arm $k (= 1, 2)$. Let m_{ki} denote the number of subunits for cluster $i (= 1, \dots, n_k)$ in arm k . Under cluster randomization, all subunits on each cluster belong to the same arm. Let y_{kij} denote a binary variable indicating a response ($y_{kij} = 1$) or no response ($y_{kij} = 0$) for subunit $j (= 1, \dots, m_{ki})$ with $P(y_{kij} = 1) = p_k$. We assume that subunits in each cluster are exchangeable with intraclass correlation coefficient $\rho_k = \text{corr}(y_{kij}, y_{kij'})$ for $j \neq j'$. However, units from different clusters are assumed to be independent. Let $y_{ki} = \sum_{j=1}^{m_{ki}} y_{kij}$, $y_k = \sum_{i=1}^{n_k} y_{ki} = \sum_{i=1}^{n_k} \sum_{j=1}^{m_{ki}} y_{kij}$, $M_k = \sum_{i=1}^{n_k} m_{ki}$, $M = M_1 + M_2$, and $n = n_1 + n_2$.

To test the hypothesis $H_0: p_1 = p_2$ against $H_1: p_1 \neq p_2$, Donner^[1] and Donner and Banting^[2] proposed the adjusted χ^2 statistic which is an adjustment of usual Pearson chi-square statistic based on an empirical estimate of the intraclass correlation coefficient,

$$X_U^2 = X_1^2/C_1 + X_2^2/C_2,$$

where

$$X_k^2 = (y_k - M_k \hat{p})^2 / M_k \hat{p} + (M_k - y_k - M_k \hat{q})^2 / M_k \hat{q}$$

$$C_k = 1 + \left(\sum_{i=1}^{n_k} m_{ki}^2 / M_k - 1 \right) \hat{\rho},$$

$\hat{p} = (y_1 + y_2)/M$ and $\hat{q} = 1 - \hat{p}$. Here, $\hat{p}$ is the analysis of variance (ANOVA) estimator for the intraclass correlation coefficient (ICC), ρ , given by

$$\hat{\rho} = \frac{\text{msc} - \text{mse}}{\text{msc} + (\tilde{m} - 1)\text{mse}},$$

where

$$\text{msc} = \sum_{k=1}^2 \sum_{i=1}^{n_k} m_{ki} (\hat{p}_{ki} - \hat{p}_k)^2 / (n - 2),$$

$$\text{mse} = \sum_{k=1}^2 \sum_{i=1}^{n_k} y_{ki} (1 - \hat{p}_{ki}) / (M - n),$$

$$\tilde{m} = (M - \sum_{k=1}^2 M_k^{-1} \sum_{i=1}^{n_k} m_{ki}^2) / (n - 2), \hat{p}_{ki} = y_{ki} / m_{ki} \quad \text{and} \quad \hat{p}_k = y_k / M_k.$$

The adjusted χ^2 statistic, χ_U^2 , assigns an equal weight to each subunit. Equal weights to subunits are interpreted as weights to clusters proportional to the cluster size.

As an alternative method, Lee and Dubin^[3] and Lee^[4] proposed to assign equal weights to clusters regardless of the cluster size. This weighting system avoids the dominance of a few clusters with large number of observations in the sample. In this case, p_k is estimated by the simple average of a response rate of each cluster in group k ,

$$\tilde{p}_k = \frac{\sum_{i=1}^{n_k} \tilde{p}_{ki}}{n_k}.$$

A variance estimator for $\tilde{p}_k$ is given as

$$\tilde{u}_k = \frac{\sum_{i=1}^{n_k} (\tilde{p}_{ki} - \tilde{p}_k)^2}{n_k (n_k - 1)}.$$

We can show that $\tilde{p}_k$ is consistent and approximately normal with mean p_k and estimated variance $\hat{v}_k$. Therefore, the weighted χ^2 statistic, X_C^2 which assigns an equal weight to each cluster, can be written as

$$X_C^2 = \frac{(\tilde{p}_1 - \tilde{p}_2)^2}{\tilde{v}_1 + \tilde{v}_2},$$

Note that these two weighting methods are identical if all clusters have equal cluster size. Jung and Ahn^[5] introduced an optimal weighting method reflecting the amount of intraclass correlation when the cluster size is varying among clusters. For group k , let $(w_{k1}, w_{k2}, \dots, w_{kn_k})$ be a set of weights such that $w_{ki} \geq 0$ and $\sum_{i=1}^{n_k} w_{ki} = 1$. The optimal weight estimator is given as

$$\bar{p}_k = \sum_{i=1}^{n_k} w_{ki} \hat{p}_{ki}.$$

where

$$w_{ki} = \bar{\sigma}_{ki}^{-2} / \sum_{i=1}^{n_k} \bar{\sigma}_{ki'}^{-2},$$

$$\bar{\sigma}_{ki}^2 = \frac{\bar{s}^2 + (m_{ki} - 1)\bar{c}_k}{m_{ki}},$$

$$\bar{s}^2 = \bar{p}(1 - \bar{p}),$$

and

$$\bar{c}_k = \sum_{i=1}^{n_k} \frac{y_{ki}(y_{ki} - 1)}{m_{ki}(m_{ki} - 1)} - \bar{p}^2$$

The optimal estimator $\bar{p}_k = \sum_{i=1}^{n_k} w_{ki} \hat{p}_{ki}$ is consistent and approximately normal with mean p_k and variance $\bar{v}_k = (\sum_{i=1}^{n_k} \bar{\sigma}_{ki}^{-2})^{-1}$, so that

$$Z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{v}_1 + \bar{v}_2}}$$

follows a normal distribution with mean 0 and variance 1 under $H_0: p_1 = p_2$. Based on these results, Ahn, Jung, Kang^[8] propose a modified χ^2 statistic using the optimal weights

$$X_o^2 = \frac{(\bar{p}_1 - \bar{p}_2)^2}{\bar{v}_1 + \bar{v}_2},$$

If the cluster size is constant for each arm (i.e. $m_{ki} = m_k$), then all three test statistics are identical.

Example 1

As a real study example, we take a teratologic animal study (Weil^[9]) that has been analyzed by several authors including Williams^[10], and Ochi and Prentice^[11]. In this study, 32 pregnant rats were randomly assigned to either receive a control diet ($n_1 = 16$) or a chemically treated diet ($n_2 = 16$) during pregnancy and lactation, see Table 1. This is an animal study, but its statistical design is very similar to that of a randomized clinical trial. The 21-day survival rates among the pups in each litter are shown in Table 4 for the control and treatment groups. The estimated 21-day survival rates (p_1, p_2) for two groups are (0.899, 0.772) using equal weights to the subunits, (0.893, 0.746) using equal weights to cluster, and (0.898, 0.750) using the optimal weights. The purpose of the study is to examine if the rats in the treatment group have significantly lower survival rate than those in the control group.

For the data set, ICC is 0.03 for the control and 0.37 for the treatment group. Williams^[10] noted that the ICC of the treatment group is substantially higher than that of the control group, possibly due to treatment effect. Jung, Ahn and Donner^[12] proposed the conditions for checking the validity of X_U^2 . Even though the ICC is substantially different between the two groups, it can be shown that X_U^2 satisfies the validity conditions for the analysis of this toxicological data. So, the use of a common ICC, $\hat{p} = 0.25$, remains appropriate purpose of testing the null hypothesis.

The estimated imbalance parameter, $\hat{k}$, is 0.95 in this data set. The weighted chi-square statistics (and p-values) for

testing equal survival probability between two groups are $X_C^2 = 2.582(p = 0.108)$, $X_U^2 = 2.761(p = 0.097)$, and $X_O^2 = 3.554(p = 0.059)$. These results suggest that there is no significant difference in 21-day survival rate between the two groups at two-sided $\alpha = 0.05$ level.

2.2 Under Subunit Randomization

Under subunit randomization, we use slightly different notations or the same notations with different definitions from those under cluster randomization. Suppose that there are n clusters, and cluster $i (= 1, \dots, n)$ has m_i subunits of which m_{ki} are randomized to arm $k (= 1, 2)$, $m_i = m_{1i} + m_{2i}$. Under subunit randomization, subunits of each cluster are allocated to different arms, so that there exist dependent observations both in each arm and among different arms. Let y_{kij} denote a binary variable indicating a response ($y_{kij} = 1$) or no response ($y_{kij} = 0$) for subunit $j (= 1, \dots, m_{ki})$ randomized to arm $k (= 1, 2)$ from cluster $i (= 1, \dots, n)$ with $P(y_{kij} = 1) = p_k$. Let $y_{ki} = \sum_{j=1}^{m_{ki}} y_{kij}$, $y_k = \sum_{i=1}^n y_{ki} = \sum_{i=1}^n \sum_{j=1}^{m_{ki}} y_{kij}$, $M_k = \sum_{i=1}^n m_{ki}$, and $M = M_1 + M_2$.

At first, we consider estimating p_k by assigning equal weight to subunits using

$$\hat{p}_k = \frac{y_k}{M_k} = \frac{\sum_{i=1}^n y_{ki}}{M_k},$$

and testing the hypothesis $H_0: p_1 = p_2$ against $H_1: p_1 \neq p_2$ using $\hat{p}_1 - \hat{p}_2$. Under H_0 ,

$$\hat{p}_1 - \hat{p}_2 = \sum_{i=1}^n \left(\frac{y_{1i} - m_{1i}p_1}{M_1} - \frac{y_{2i} - m_{2i}p_2}{M_2} \right) = \sum_{i=1}^n \epsilon_i,$$

where $\epsilon_i = M_1^{-1}(y_{1i} - m_{1i}p_1) - M_2^{-1}(y_{2i} - m_{2i}p_2)$ are 0 mean independent random variables. So, under H_0 by the central limit theorem, $\hat{p}_1 - \hat{p}_2$ is approximately normal with mean 0 with variance that can be estimated by

$$\hat{v} = \sum_{i=1}^n \hat{\epsilon}_i^2,$$

where $\hat{\epsilon}_i = M_1^{-1}(y_{1i} - m_{1i}\hat{p}_1) - M_2^{-1}(y_{2i} - m_{2i}\hat{p}_2)$. Hence, a modified χ^2 test can be obtained as

$$X_U^2 = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\hat{v}}.$$

Table 1 Number, y , of pups that survived the 21-day lactation period and the number, m , of pups alive at 4 days for each litter

Group	(y, m)							
Control	(13, 13)	(12, 12)	(9, 9)	(9, 9)	(8, 8)	(8, 8)	(12, 13)	(11, 12)
	(9, 10)	(9, 10)	(8, 9)	(11, 13)	(4, 5)	(5, 7)	(7, 10)	(7, 10)
Treated	(12, 12)	(11, 11)	(10, 10)	(9, 9)	(10, 11)	(9, 10)	(9, 10)	(8, 9)
	(8, 9)	(4, 5)	(7, 9)	(4, 7)	(5, 10)	(3, 6)	(3, 10)	(0, 7)

An estimator of p_k assigning equal weights to clusters is given as

$$\tilde{p}_k = \frac{\sum_{i=1}^n \hat{p}_{ki}}{n}.$$

Under H_0 ,

$$\tilde{p}_1 - \tilde{p}_2 = \frac{1}{n} \sum_{i=1}^n (\hat{p}_{1i} - \hat{p}_{2i}) = \sum_{i=1}^n e_i,$$

where $e_i = n^{-1}(\hat{p}_{1i} - \hat{p}_{2i})$ are 0 mean independent random variables. So, by the central limit theorem, $\tilde{p}_1 - \tilde{p}_2$ is approximately normal with mean 0 with variance that can be estimated by

$$\hat{v} = \sum_{i=1}^n e_i^2$$

under H_0 . Hence, a modified χ^2 test can be obtained as

$$X_C^2 = \frac{(\tilde{p}_1 - \tilde{p}_2)^2}{\hat{v}}.$$

Now we derive optimal weights for a modified χ^2 test. Under subunit randomization, we consider

$$W = \sum_{i=1}^n w_i (\hat{p}_{1i} - \hat{p}_{2i})$$

for $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$. Since, for $i = 1, \dots, n$, $\hat{p}_{1i} - \hat{p}_{2i}$ are independent 0 mean random variables under H_0 , by the central limit theorem, W is approximately normal with mean 0 and variance that can be approximated by

$$\bar{v} = \sum_{i=1}^n w_i^2 (\hat{p}_{1i} - \hat{p}_{2i})^2.$$

In order to minimize $\bar{v}$, we use the inverse of the variance of $\hat{p}_{1i} - \hat{p}_{2i}$ as the weight w_i . Let $\sigma_k^2 = p_k(1 - p_k)$, and $\rho_0 = \text{corr}(y_{1ij}, y_{2ij})$, the correlation coefficient between two subunits that are assigned to different treatment arms, and $\rho_k = \text{corr}(y_{kij}, y_{kij})$, the correlation coefficient between two subunits that are assigned to the same treatment arms from each cluster. Then,

$$\hat{p}_{1i} - \hat{p}_{2i} = \frac{1}{m_{1i}} \sum_{j=1}^{m_{1i}} y_{1ij} - \frac{1}{m_{2i}} \sum_{j=1}^{m_{2i}} y_{2ij}$$

has variance

$$s_i^2 = \frac{\sigma_1^2}{m_{1i}} \{1 + (m_{1i} - 1)\rho_1\} + \frac{\sigma_2^2}{m_{2i}} \{1 + (m_{2i} - 1)\rho_2\} - 2\rho_0\sigma_1\sigma_2.$$

Note that σ_k^2 , ρ_0 and ρ_1 can be consistently estimated by $\hat{\sigma}_k^2 = \hat{p}_k(1 - \hat{p}_k)$,

$$\bar{\rho}_0 = \frac{\bar{c}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2}$$

and

$$\bar{\rho}_k = \frac{\bar{c}_k}{\hat{\sigma}_k^2}$$

respectively, where

$$\bar{c}_{12} = \frac{\sum_{i=1}^n \sum_{j=1}^{m_{1i}} \sum_{j'=1}^{m_{2i}} (y_{1ij} - \hat{p}_1)(y_{2ij'} - \hat{p}_2)}{\sum_{i=1}^n m_{1i} m_{2i}}$$

is a consistent estimator of the covariance between two subunits assigned to different treatment arms from the same cluster and

$$\bar{c}_{12} = \frac{\sum_{i=1}^n \sum_{j=1}^{m_{ki}} \sum_{j'=1}^{m_{ki}} (y_{kij} - \hat{p}_k)(y_{kij'} - \hat{p}_k)}{\sum_{i=1}^n m_{ki}(m_{ki} - 1)}$$

is a consistent estimator of the covariance between two subunits assigned to treatment arm k from the same cluster. We obtain $\bar{s}_i^2$ from s_i^2 by replacing σ_k^2 , ρ_0 and ρ_k with their consistent estimators $\hat{\sigma}_k^2$, $\bar{\rho}_0$ and $\bar{\rho}_k$, respectively. Finally, we obtain an optimal weights

$$w_i = \frac{\bar{s}_i^{-2}}{\sum_{i'=1}^n \bar{s}_{i'}^{-2}}$$

and an optimal χ^2 test

$$X_O^2 = \frac{W^2}{\bar{v}} = \frac{\left\{ \sum_{i=1}^n w_i (\hat{p}_{1i} - \hat{p}_{2i}) \right\}^2}{\sum_{i=1}^n w_i^2 (\hat{p}_{1i} - \hat{p}_{2i})^2}.$$

If the cluster size is constant (i.e. $m_i = m$) and the allocation proportion is fixed (i.e. m_{1i}/m_{2i} is identical for all clusters), then the all three test statistics are identical.

Example 2

The data, shown in Table 2, consist of survival times (days) of closely ($k = 1$) and poorly ($k = 2$) matched skin grafts on each burned patient. We want to test if the probability that the survival time is longer than or equal to 30 days or not between closely and poorly matched skin grafts. We have $\hat{p}_1 = 0.32$, $\hat{p}_2 = 0.14$, $X_U^2 = 3.126$ (p-value = 0.077) using equal weights to subunits, $\hat{p}_1 = 0.42$, $\hat{p}_2 = 0.18$, $X_C^2 = 2.286$ (p-value = 0.131) using equal weights to clusters, and $X_O^2 = 2.650$ (p-value = 0.104). Since too

Table 2 Days of skin graft survival on each burned patient

Patient	1	2	3	4	5	6	7	8	9	10	11
Close match	37	19	57+, 57+	93	16	22	20	18	77, 63, 29	29	60+
Poor match	29	13	15	26	11	17	26	21	43	15, 18	40

+ indicates censored observations.

few patients have multiple grafts within each graft group, the estimator $\bar{\rho}_k$ may not be reliable enough, so that the optimal χ^2 does not have more significance than that using equal weights for subunits.

3 SIMULATION STUDIES

At first, we conduct some simulations. We set the number of clusters at $n_1 = n_2 = 10, 20, 30$. For cluster i ($i = 1, \dots, n_k$) in arm k ($k = 1, 2$), the cluster size m_{ki} is generated from a negative binomial distribution truncated below 1. We used the methodology proposed by Donner and Koval^[13] where the cluster size m_{ki} is generated from the negative binomial distribution truncated below 1 with parameters s and P . This distribution is

$$P_{(m)} = \frac{(s+m-1)!Q^{-s}(P/Q)^m}{(s-1)!m!(1-Q^{-s})},$$

where $Q = 1 + P$, $m = 1, 2, \dots$.

From the parameters of this distribution, we can derive new parameters that describe the variability of the cluster size. By Johnson and Kotz^[14], the mean and variance of this distribution are $\mu = sP/(1 - P_0)$ and $\sigma^2 = \mu[1 + P - sP P_0/(1 - P_0)]$, where $P_0 = (1 + P)^{-s}$. The measure of imbalance is given by $\kappa = 1/(1 + \nu^2)$, where ν is the coefficient of variation, that is, $\nu = \sigma/\mu$. It is clear that κ equals 1 when all the cluster sizes are identical. We have a relationship between the variance and the imbalance parameter, $\sigma^2 = \mu^2(1 - \kappa)/\kappa$. The variance of cluster size decreases in κ . Table 3 provides the value of κ under various design settings with $n = 10$. We use numerical methods to obtain the values of s and P corresponding to specified values of μ and κ .

Once the cluster size, m_{ki} , is obtained, a random variable y_{ki} is generated from a beta-binomial distribution with true success probability p_k for clusters in group k and an ICC $\rho_1 = \rho_2 = \rho = 0.05, 0.1, 0.2, 0.3$. In order to study the effect of varying the degree of imbalance (κ) on the performance of the estimators, we vary κ over a specified range for a mean cluster size μ . We allow the cluster size to vary with a mean cluster size of $\mu = 10$, and the imbalance parameter of $\kappa = 0.6, 0.7, 0.8$, and 0.9 .

We compute the empirical type I error and power X_U^2, X_C^2 , and X_O^2 , by running 5,000 simulations for each combination of parameters. Table 4 shows the empirical type I error rates for testing $H_0: p_1 = p_2 (= p)$ under $p = 0.1, 0.3$, or 0.5 . The weighted chi-square statistics, X_U^2 and X_O^2 , are slightly anti-conservative when $n_1 = n_2 = 10$. When $n_1 = n_2 = n = 20$ or 30 , all the weighted chi-square statistics have empirical type I errors close to 5% significance level and the differences in empirical type I errors are negligible among weighted chi-square statistics considered here. Table 5 shows the empirical powers for testing $H_0: p_1 = p_2$ under $p_1 = 0.1, 0.3$ or 0.5 , and $p_2 = p_1 + 0.2$. The optimal weight chi-square statistic, X_O^2 , yields higher empirical power levels than X_U^2 and X_C^2 . When $n_1 = n_2 = 10$, X_C^2 performs better in empirical type I errors than X_U^2 and X_O^2 . However, X_C^2 has lowest powers among weighted chi-square statistics

Table 3 Illustrative values of $\hat{\kappa}$ for $n = 10$

Design	$(m_{k1}, m_{k2}, \dots, m_{k10})$	$\hat{\kappa}$
1	6, 7, 8, 8, 9, 10, 10, 12, 14, 17	0.9
2	4, 4, 7, 8, 10, 11, 14, 16, 18, 19	0.8
3	2, 3, 5, 7, 10, 12, 15, 17, 21, 24	0.7
4	1, 2, 3, 5, 7, 13, 17, 19, 22, 29	0.6

Table 4 Empirical type I error probability for testing $H_0: p_1 = p_2 = p$ with $n_1 = n_2 = n$, a common intraclass correlation coefficient ρ and an imbalance parameter κ

κ	ρ	p	$n = 10$			$n = 20$			$n = 30$		
			X_U^2	X_C^2	X_O^2	X_U^2	X_C^2	X_O^2	X_U^2	X_C^2	X_O^2
.9	.05	.1	.0522	.0466	.0532	.0558	.0518	.0558	.0518	.0474	.0508
		.3	.0596	.0584	.0596	.0532	.0502	.0554	.0566	.0516	.0562
		.5	.0608	.0566	.0616	.0582	.0532	.0596	.0540	.0546	.0542
	.10	.1	.0578	.0470	.0578	.0582	.0530	.0592	.0548	.0518	.0560
		.3	.0626	.0548	.0642	.0582	.0540	.0578	.0520	.0470	.0502
		.5	.0614	.0520	.0616	.0584	.0530	.0556	.0582	.0516	.0554

(Continued)

Table 4 (Continued)

κ	ρ	p	$n = 10$			$n = 20$			$n = 30$		
			X_D^2	X_C^2	X_O^2	X_D^2	X_C^2	X_O^2	X_D^2	X_C^2	X_O^2
.8	.20	.1	.0546	.0452	.0566	.0504	.0500	.0528	.0488	.0502	.0510
		.3	.0556	.0468	.0552	.0542	.0486	.0550	.0512	.0506	.0528
		.5	.0664	.0554	.0638	.0546	.0514	.0528	.0576	.0546	.0562
	.30	.1	.0486	.0418	.0525	.0500	.0476	.0536	.0572	.0520	.0540
		.3	.0622	.0560	.0632	.0516	.0520	.0548	.0518	.0490	.0524
		.5	.0648	.0556	.0640	.0508	.0498	.0532	.0518	.0540	.0546
	.50	.1	.0584	.0476	.0588	.0566	.0524	.0575	.0570	.0468	.0554
		.3	.0664	.0566	.0634	.0524	.0452	.0516	.0572	.0560	.0586
		.5	.0544	.0488	.0544	.0566	.0530	.0576	.0526	.0500	.0536
	.10	.1	.0556	.0504	.0600	.0579	.0540	.0580	.0556	.0454	.0516
		.3	.0634	.0544	.0620	.0588	.0526	.0572	.0566	.0540	.0562
		.5	.0642	.0568	.0618	.0586	.0584	.0590	.0576	.0540	.0592
.7	.20	.1	.0520	.0402	.0510	.0528	.0458	.0534	.0486	.0498	.0530
		.3	.0624	.0538	.0628	.0606	.0522	.0570	.0532	.0500	.0538
		.5	.0626	.0534	.0626	.0554	.0512	.0560	.0564	.0486	.0520
	.30	.1	.0488	.0330	.0480	.0490	.0490	.0548	.0472	.0490	.0544
		.3	.0626	.0560	.0606	.0518	.0450	.0496	.0594	.0512	.0546
		.5	.0658	.0582	.0654	.0610	.0584	.0612	.0540	.0578	.0565
	.05	.1	.0582	.043	.0592	.0612	.05	.06	.052	.0506	.0534
		.3	.0606	.053	.0584	.057	.0538	.0584	.0582	.0506	.0566
		.5	.0648	.0536	.064	.0568	.0476	.0542	.056	.0468	.0544
	.10	.1	.062	.0482	.0625	.0554	.0488	.056	.0564	.052	.057
		.3	.0614	.0504	.0562	.0582	.0528	.0596	.0518	.0514	.0512
		.5	.0636	.0538	.061	.0552	.048	.0542	.0554	.051	.0544
.6	.20	.1	.0604	.0468	.0638	.0546	.045	.0554	.0516	.045	.0518
		.3	.0614	.0482	.0618	.0596	.0568	.058	.0562	.056	.0594
		.5	.069	.0582	.0652	.0572	.0538	.0576	.0556	.0546	.0528
	.30	.1	.0478	.037	.0546	.0456	.0486	.0554	.0534	.0512	.0536
		.3	.0612	.0468	.0564	.0572	.0546	.0608	.0534	.0506	.0545
		.5	.0684	.0568	.0676	.0584	.0524	.0586	.0528	.051	.0506
	.05	.1	.0586	.0458	.0608	.0586	.0448	.0594	.059	.0448	.06
		.3	.0624	.049	.0604	.0592	.0578	.0606	.058	.0518	.057
		.5	.0616	.0548	.062	.06	.0524	.0636	.054	.0464	.0539
	.10	.1	.0652	.0426	.0652	.0584	.0512	.0586	.0568	.0498	.0568
		.3	.0674	.0514	.061	.0588	.0554	.058	.054	.0484	.0544
		.5	.0616	.054	.061	.06	.0472	.0556	.0534	.0576	.0546
	.20	.1	.064	.0432	.0648	.053	.049	.0578	.0538	.0472	.056
		.3	.0664	.0562	.0668	.0576	.0472	.0562	.0568	.0554	.0594
		.5	.0692	.0564	.0682	.055	.0516	.0566	.0596	.0522	.0598
	.30	.1	.052	.0342	.0574	.0482	.0506	.057	.0492	.051	.0556
		.3	.0646	.05	.059	.0504	.0466	.0514	.0572	.0512	.0544
		.5	.0674	.0504	.0592	.0594	.0578	.0592	.0562	.0478	.0516

Table 5 Empirical powers for testing $H_0: p_1 = p_2$ with $n_1 = n_2 = n$, a common intraclass correlation coefficient ρ and an imbalance parameter κ

κ	ρ	p_1	p_2	$n = 10$			$n = 20$			$n = 30$		
				X_D^2	X_C^2	X_O^2	X_D^2	X_C^2	X_O^2	X_D^2	X_C^2	X_O^2
.9	.05	.1	.3	.8302	.79	.835	.9859	.9791	.9857	.9987	.9987	.9989
		.3	.5	.6526	.6064	.6584	.9194	.899	.923	.9835	.9761	.9847
		.5	.7	.6506	.609	.6588	.9136	.888	.9152	.9833	.9781	.9851
	.10	.1	.3	.7208	.7	.7362	.9445	.9407	.9491	.9941	.9919	.9945
		.3	.5	.5384	.5016	.5458	.8252	.8162	.8356	.9423	.9359	.9451
		.5	.7	.5408	.5068	.5484	.829	.8096	.8364	.9403	.933	.9433
	.20	.1	.3	.569	.5498	.5842	.8436	.845	.8572	.9463	.9525	.9579
		.3	.5	.4128	.3868	.4262	.6474	.652	.6702	.8246	.8298	.8422
		.5	.7	.4008	.38	.4098	.6534	.6494	.6726	.8286	.8332	.8434
	.30	.1	.3	.4626	.4468	.478	.7348	.7512	.7632	.8838	.896	.9036
		.3	.5	.326	.3132	.3376	.5356	.5406	.5542	.713	.7342	.743
		.5	.7	.3246	.3162	.3408	.5454	.5512	.5638	.7008	.7162	.7256
.8	.05	.1	.3	.8058	.743	.8206	.9833	.9577	.9867	.9979	.9933	.9987
		.3	.5	.6272	.545	.6378	.9012	.8446	.9098	.9789	.9545	.9825
		.5	.7	.6454	.5638	.655	.9068	.8546	.9088	.9763	.9559	.9779
	.10	.1	.3	.714	.6678	.7378	.9328	.9168	.9463	.9905	.9869	.9947
		.3	.5	.5322	.4854	.5482	.8064	.7672	.8232	.9216	.9052	.9351
		.5	.7	.5182	.4692	.5382	.787	.7622	.8096	.9258	.911	.9407
	.20	.1	.3	.5362	.5162	.5652	.8104	.8228	.8484	.9308	.9411	.9547
		.3	.5	.3888	.3668	.4054	.6222	.6266	.662	.7838	.7962	.8242
		.5	.7	.3802	.3672	.408	.6292	.6372	.6768	.7892	.8054	.8266
	.30	.1	.3	.4316	.433	.4694	.7002	.7248	.7458	.8488	.879	.895
		.3	.5	.3064	.305	.3284	.4942	.5286	.5466	.6716	.7034	.7254
		.5	.7	.3068	.3048	.3296	.5044	.5278	.5528	.67	.7034	.7242
.7	.05	.1	.3	.7902	.6796	.8066	.9749	.9246	.9817	.9981	.9881	.9985
		.3	.5	.6176	.499	.6312	.8874	.7774	.8988	.9717	.909	.9763
		.5	.7	.6118	.497	.6274	.8808	.784	.8972	.9711	.9232	.9755
	.10	.1	.3	.6848	.6206	.7152	.9204	.8834	.9419	.9811	.9697	.9895
		.3	.5	.5054	.4342	.5262	.7666	.708	.8008	.9106	.8742	.9341
		.5	.7	.5054	.432	.5244	.7802	.7244	.8058	.899	.8808	.9302
	.20	.1	.3	.5126	.499	.5644	.7804	.7894	.8392	.9142	.9288	.9527
		.3	.5	.3742	.3534	.409	.5816	.6036	.6554	.7424	.769	.8154
		.5	.7	.3772	.3518	.415	.5904	.5886	.6494	.7544	.7656	.8092
	.30	.1	.3	.4084	.4132	.4602	.6576	.703	.7334	.8178	.8568	.8808
		.3	.5	.2996	.2904	.3322	.4664	.506	.5428	.6262	.6884	.7152
		.5	.7	.2852	.2814	.3184	.472	.5112	.544	.621	.6766	.7094
.6	.05	.1	.3	.7588	.6232	.7866	.9619	.8854	.9755	.9943	.9705	.9963
		.3	.5	.5992	.4542	.6186	.8626	.7184	.8796	.9645	.8718	.9735
		.5	.7	.5932	.4306	.612	.8628	.7086	.8884	.9603	.8684	.9713
	.10	.1	.3	.639	.5594	.683	.8908	.8354	.926	.9747	.9555	.9893
		.3	.5	.4814	.3864	.5062	.7232	.6498	.7754	.8774	.8308	.917
		.5	.7	.4754	.3864	.5056	.731	.666	.7854	.8772	.8216	.9194
	.20	.1	.3	.4874	.4736	.5506	.7306	.7424	.8112	.8782	.8936	.9337

(Continued)

Table 5 (Continued)

κ	ρ	p_1	p_2	$n = 10$			$n = 20$			$n = 30$		
				X_D^2	X_C^2	X_O^2	X_D^2	X_C^2	X_O^2	X_D^2	X_C^2	X_O^2
.30		.3	.5	.3568	.3288	.3906	.557	.564	.6392	.6968	.728	.7938
		.5	.7	.3442	.3032	.385	.5482	.5404	.616	.7034	.7198	.793
	.30	.1	.3	.3772	.3938	.4488	.606	.6784	.7238	.768	.8386	.8716
		.3	.5	.2796	.2764	.325	.4308	.485	.5306	.579	.6548	.702
		.5	.7	.278	.2754	.3234	.4356	.4806	.5324	.5948	.6662	.7088

considered here. As the imbalance parameter (κ) increases, the variability in cluster size decreases and the empirical power increases. The difference in empirical power between X_O^2 and the other two weighted chi-square statistics increases as the variability in cluster size increases. Table 5 shows that the empirical power increases as the number of clusters n_k increases. As the ICC ρ decreases, the empirical power increases. Based on the results from the simulation study, we recommend using X_O^2 because of its higher power compared to the other weighted chi-square tests.

REFERENCES

- Donner, A. Statistical methods in ophthalmology: an adjusted chi-square approach. *Biometrics* **1989**, *45* (2), 605–611.
- Donner, A.; Banting, D. Analysis of site-specific data in dental studies. *Journal of Dental Research* **1988**, *67* (11), 1392–1395.
- Lee, E.; Dubin, N. Estimation and sample size considerations for clustered binary responses. *Statistics in Medicine* **1994**, *13* (12), 1241–1252.
- Lee, E. Two sample comparison for large groups of correlated binary responses. *Statistics in Medicine* **1996**, *15* (11), 1187–1197.
- Jung, S.; Ahn, C. Estimation of response probability in correlated binary data: A new approach. *Drug Information Journal* **2000**, *34*, 599–604.
- Batchelor, J.R.; Hackett, M. HL-A matching in treatment of burned patients with skin allografts. *Lancet* **1970**, *2*, 581–583.
- Donner, A.; Klar, N. *Design and analysis of cluster randomization trials in health research*. Arnold: London, 2000.
- Ahn, C.; Jung, S.H.; Kang, S.H. An evaluation of weighted chi-square statistics for site-specific binary data. *Drug Information Journal* **2003**, *37* (1), 91–99.
- Weil, C.S. Selection of the valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. *Food Cosmet. Toxicol.* **1970**, *8*, 177–182.
- Williams, D.A. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **1975**, *31* (4), 949–952.
- Ochi, Y.; Prentice, R.L. Likelihood inference in a correlated probit regression model. *Biometrika* **1984**, *71* (3), 531–543.
- Jung, S.; Ahn, C.; Donner, D. Evaluation of an adjusted chi-square statistic as applied to observational studies involving clustered binary data. *Statistics in Medicine* **2001**, *20* (14), 2149–2161.
- Donner, A.; Koval, J. A procedure for generating group sizes from a one-way classification with a specified degree of imbalance. *Biometrical Journal* **1987**, *29* (2), 181–187.
- Johnson, N.L.; Kotz, S. *Distributions in Statistics. Discrete Distributions*. Wiley: New York, 1969.

Rank Regression: Stability Analysis

Annpey Pong

Merck Research Laboratories, Summit, New Jersey, U.S.A.

Ying Qing Chen

Fred Hutchinson Cancer Research Center, Seattle, Washington, U.S.A.

Xiudong Lei

Division of Biostatistics, School of Public Health, University of California, Berkeley, California, U.S.A.

Abstract

This entry presents the semiparametric model of stability analysis and discusses the rank-based inference procedures to determine the shelf-life of drug products.

Keywords: Expiration dating period; Rank regression; Semiparametric; Model; Shelf-life; Stability analysis.

INTRODUCTION

In the pharmaceutical industry, stability study is performed to determine the shelf-life of drug product. It provides the assurance that the capacity of a drug product maintains its identity, strength, quality, and purity before the drug expiration date. The stability data are often collected on the product samples stored in controlled conditions to measure the drug characteristics changing over time.

As indicated in a draft consensus guideline of *Evaluation of Stability Data* issued by the International Conference on Harmonization (ICH) Steering Committee (2002),^[1] “regression analysis is considered an appropriate approach to evaluating the stability data for a quantitative attribute and establishing a retest period or shelf life” (p. 6). However, most of these regression models rely heavily on their underlying parametric assumptions, such as normality of the continuous characteristics or their transformations. In this article, the alternative approach, semi-parametric regression models, is presented. The rank-based inference procedures to determine shelf-life are discussed. Numerical studies including Monte Carlo simulations and practical examples are demonstrated as well.

STABILITY ANALYSIS BASED ON RANK REGRESSION

Semiparametric Regression Models

In the FDA’s 1987 guideline,^[2] the linear regression models shown in Eq. 1 were used in the stability analysis. Denote

Y_{ij} the continuous outcome of the characteristic for the j th measurement of the i th batch, $i = 1, 2, \dots, n$, and $j = 1, 2, \dots, m_i$.

$$Y_{ij} = \alpha_i + \beta_i X_{ij} + e_{ij} \quad (1)$$

where X_{ij} are the associated covariates, $(\alpha_i, \beta_i)^T$ are the parameters, and e_{ij} are the independent random errors. The random errors are assumed to be identically zero-mean normal deviates with constant standard variance σ_e^2 . The superscript T denotes the transpose of vector or matrix.

However, it is well known that the fully parametric approaches in Eq. 1 have to depend heavily on the validity of their parametric assumptions. Without strong belief in these assumptions, the following alternative semiparametric model can be considered:

$$h(Y_{ij}) = \alpha_i + \beta_i X_{ij} + e_{ij} \quad (2)$$

where e_{ij} are independent zero-mean deviates with an unknown density function of $f(\cdot)$, and $h(\cdot)$ is some known monotonic transformation function.

Note that the model in Eq. 2 is quite flexible without assuming the underlying distribution of normality or constant variance. Moreover, there is no assumption on the parametric distributions on $(\alpha_i, \beta_i)^T$ as random effects but leaving them as fixed effects, when appropriate rank-based inference procedures discussed later are utilized. Additionally, the parametric component in Eq. 2 is still the linear combination of parameter β_i and covariates X_{ij} . So the sign and quantity of β_i characterize the direction and magnitude of the linear trend of Y_{ij} with respect to X_{ij} , respectively.

For example, when X_{ij} are the times at which Y_{ij} are measured, β_i may stand for “degradation” of measurements over time for its negative sign.

RANK-BASED SHELF-LIFE DETERMINATION

Shelf-Life for a Single Batch

Considering one single batch of stability data for stability analysis, $n = 1$. Without loss of generality, assume $h(\cdot)$ is an identity link. Eq. 2 becomes $Y_j = \alpha + \beta X_j + e_j$. Under the null hypothesis of $\beta = 0$, a standard rank test statistic is $U = \sum_{j=1}^m (X_j - \bar{X})R(Y_j)$ where $\bar{X}$ is the mean of X , and $R(Y_1), R(Y_2), \dots, R(Y_m)$ are the ranks of $Y_1, Y_2, \dots, Y_m$, respectively. When β is not necessarily zero, the residuals of $e_j = Y_j - (\alpha + \beta X_j)$ are considered $U(\beta) = \sum_{j=1}^m (X_j - \bar{X})R(e_j)$. Apparently, $E[U(\beta); \beta] = 0$ for any $\beta \in \mathbf{B} \subset \mathbf{R}$, where $\mathbf{B}$ is the parameter space and $\mathbf{R}$ is the real line. Therefore β is estimated by solving the unbiased estimating equation $U(\beta) = 0$. It is noticeable that the parameter of α is not estimable from $U(\beta)$, because it serves as a baseline effect for every observation and therefore the ranks are invariant to the constant intercept. One remedy to determine α is using the median of $(Y_j - \beta X_j), j = 1, 2, \dots, m$. When the distribution of error terms is further symmetric, α can be estimated by the median of the Walsh average of residuals of $(e_i + e_j)/2, 1 \leq i < j \leq m$.^[3,4]

As $U(\beta)$ is a monotonically decreasing step function in β , there may not exist values of β to allow $U(\beta) = 0$ exactly. One approach is to define $\hat{\beta}$ that satisfies $U(\hat{\beta}+)U(\hat{\beta}-) \leq 0$. The other alternative is to obtain the rank estimate $\hat{\beta}$ by minimizing the $D(\beta) = \sum_{j=1}^m w_j \{R(e_j)\} e_j$ where $w_j(\cdot)$ are the nonconstant sequence of scores such that $w_j + w_{m-j+1} = 0$.^[5] Then $D(\beta)$ is a nonnegative, continuous, and convex function of β . Specifically, consider the Wilcoxon scores of $w(j) = \phi\{j/(m+1)\}$, where $\phi(u) = \sqrt{12}(u - 1/2)$. Let β_0 be the true value of β . According to the regularity conditions and Theorem 5.2.3 in Ref. [3], $m^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0, \tau^2 - 1)$ where $\tau^2 = (\sqrt{12} \int_0^1 f^2(u) du)^{-1}$ and $\Sigma = \lim_{m \rightarrow \infty} m^{-1} X_c' X_c$ is to be positive-definite. Here X_c is the centered X by $X_c = X - 1(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)$, where $\bar{x}_i$ is the mean of the i th column in X . Furthermore, when $\hat{\alpha}$ is the median of $\hat{e}_j = Y_j - \hat{\beta} X_j$, it can be shown that $m^{1/2}[(\hat{\alpha}, \hat{\beta})^T - (\alpha_0, \beta_0)^T]$ would have an asymptotic distribution of multivariate normal with mean zero and variance-covariance of $V = \tau^2 \begin{bmatrix} 2f(0)\tau^{-2} + \mu^T - 1 & \mu^T - 1 \\ -\mu^T - 1 & -1 \end{bmatrix}$ where μ is the mean vector of Y 's, and $f(\cdot)$ is the density function of e_j 's. When $\hat{\alpha}$ is the median of the Walsh averages, $m^{1/2}[(\hat{\alpha}, \hat{\beta})^T - (\alpha_0, \beta_0)^T]$ follows an asymptotic distribution of multivariate normal with mean zero and variance-covariance of $V = \tau^2 \Lambda^{-1}$ where $\Lambda^{-1} = \lim_{m \rightarrow \infty} [1, X]^T [1, X]$. Here $\mathbf{1}$ is a unit vector and $\mathbf{X}$ the design matrix.

As a result, an approximately 95% pointwise confidence bound for EY is then (LB, UB), where

$$LB = \hat{\alpha} + \hat{\beta} X + Z_{0.025} \sqrt{(1, X) \hat{V}(1, X)^T} \quad \text{and} \quad UB = \hat{\alpha} + \hat{\beta} X + Z_{0.975} \sqrt{(1, X) \hat{V}(1, X)^T}$$

Therefore an estimate of shelf-life is the point of time, T_{shelf} , at which LB or UB intersects the predefined most allowable specification, η , say, depending on whether β is negative or positive, respectively.

Shelf-Life for Multiple Batches

According to the FDA's 1987 guideline, at least three batches are preferable to allow the estimate of batch-to-batch variation in stability analysis. When the batch-to-batch variation is suspected, there are three occasions from batch-to-batch:

- Different intercepts, but same slope.
- Different slopes, but same intercept.
- Both intercepts and slopes are different.

The estimation equations along with the estimates of α and β of three occasions, and shelf-life determination are discussed as below.

When the batch-to-batch variation is present as in occasion (1): $Y_{ij} = \alpha_i + \beta X_{ij} + e_{ij}$ where β is the common slope for every batch. The estimation equation is $U(\beta) = \sum_{i=1}^n U_i$ where $U_i = \sum_{j=1}^{m_i} i(X_{ij} - \bar{X}_i)R(e_{ij})$. Here $\bar{X}_i = m_i^{-1} \sum_j X_{ij}$ and $e_{ij} = Y_{ij} - (\alpha_i + \beta X_{ij})$. Similarly, the medians of $Y_{ij} - \beta X_{ij}, j = 1, 2, \dots, m_i$, can be used to estimate $\alpha_i, i = 1, \dots, n$, individually.

When the batch-to-batch variation is present as in occasion (2): $Y_{ij} = \alpha + \beta_i X_{ij} + e_{ij}$ where α is the common intercept for every batch. To estimate individual β_i , consider: $U_i(\hat{\beta}_i) = 0$ for every individual i to solve for $\hat{\beta}_i$. Then the medians of $Y_{ij} - \hat{\beta}_i X_{ij}, j = 1, 2, \dots, m_i, i = 1, \dots, n$, are used to estimate α .

When the batch-to-batch variation is present as in occasion (3): $Y_{ij} = \alpha_i + \beta X_{ij} + e_{ij}$ where the intercepts and slopes are both different from batch-to-batch. To estimate individual β_i , consider: $U_i(\hat{\beta}_i) = 0$. And the medians of $Y_{ij} - \hat{\beta}_i X_{ij}, j = 1, 2, \dots, m_i, i = 1, \dots, n$, respectively.

If the batch-to-batch variation exists, the straightforward approach is to use the minimum approach required by the FDA's guidelines. That is, first compute $T_{\text{shelf}, i}$ for each individual batch, and the final shelf-life is then $\min(T_{\text{shelf}, i}, i = 1, 2, \dots, n)$. Or, when the number of batches is large enough and, potentially, the so-called “future” batches show variability, then the shelf-life can be computed based on the prediction bounds of the average values of α_i and β_i , following the approach by Shao and Chen.^[6] If in fact, α_i 's are random effects, the randomness of α_i still has an impact on the asymptotic distribution of the slope estimates. In this case, the parameter τ and the joint distribution of the intercepts and slopes will need to be modified.^[7]

PRELIMINARY TEST FOR BATCH SIMILARITY

The batch similarity can be evaluated by the following null hypotheses of testing the equality of slopes and intercepts among batches. $H_{\alpha:\beta} : \beta_i = \beta_{i'} \text{ for all } i \neq i'$. $H_{\alpha:\alpha} : \alpha_i = \alpha_{i'} \text{ for all } i \neq i'$. To test the batch-to-batch variation, the method in Ref. [8], which is similar to the Woolf's test for stratified samples in homogeneity, can be extended to the obtained rank-based estimates. Let $(\hat{\alpha}_i, \hat{\beta}_i)$ be the estimates of (α_i, β_i) for individual batch i based on the methods discussed before. Let $(\hat{\alpha}, \hat{\beta})$ be one estimate by pooling all the batches together, i.e., without batch-to-batch variation. Then the statistics $TS = \sum_{i=1}^n (\hat{\alpha}_i - \hat{\alpha}, \hat{\beta}_i - \hat{\beta})^T \hat{V}_i^{-1} (\hat{\alpha}_i - \hat{\alpha}, \hat{\beta}_i - \hat{\beta})$ where $\hat{V}_i$ are the estimates of the variance of $(\hat{\alpha}_i, \hat{\beta}_i)$, respectively, under the null hypothesis that there is no batch-to-batch variations. It follows a χ^2 -distribution with degrees of freedom of $n - 1$ approximately. The null hypothesis is rejected if the associated p -value is larger than 0.25 as in the current guideline from FDA.

NUMERICAL STUDIES

The computation of the proposed rank-based approaches is tedious especially when the covariates are of relatively higher dimensions. In this case, the "recursive bisection" programming^[9] can be used to solve the estimating function. To compute $\hat{V}$ of the confidence bands of (LB, UB), $\hat{\tau}$ can be estimated by $\int f^2(u) du$,^[10] which is by obtaining a kernel type of estimate of $f(u)$: $f(u) = \frac{1}{mh} \sum_{j=1}^m k\left(\frac{u - e_j}{h}\right)$ where $k(\cdot)$ is a uniform kernel function $k(u) = \frac{1}{2} \quad u \in [-1/2, 1/2]$ and 0 otherwise and h is kernel bandwidth. Then $\delta = \int f^2(u) du = \int f(u) dF(u)$ can be estimated by $\hat{\delta} = \frac{1}{m^2 h} \sum_{i=1}^m \sum_{j \neq i}^m I\left(\pi |e_i - e_j| \frac{h}{2}\right)$ Hettmansperger^[3] proposed a modified version of $\hat{\delta}_c$ to ease the computation, such that $\hat{\delta}_c = \frac{1}{mc} + \frac{1}{m(m-1)h} \sum_{i=1}^m \sum_{j \neq i}^m k\left(\frac{e_i - e_j}{h}\right)$ where c is any fixed constant. The modified $\hat{\delta}_c$ is also consistent with δ when $h = cm^{-1/2}$ (Theorem 5.2.4 in Ref. [3]), The actual implement of $\hat{\delta}_c$ can be found in Ref. [7].

In the simulations, the performance of rank-based estimation procedure under different error structures based on a common slope model shown below is considered: $Y_{ij} = \alpha_i + \beta X_{ij} + e_{ij}$ for $i = 1, \dots, n$ and $j = 1, \dots, m$ where Y_{ij} is the characteristic of interest of the j th measurement for the i th batch, X_{ij} is the time of the j th measurement for the i th batch, e_{ij} is the random error of the j th measurement for the i th batch, α_i is the batch-specific intercept, and β is the common slope.

Considering the FDA's (1987) recommendation of at least three batches to obtain the shelf-life for an NDA submission, both $n = 1$ and 3 are performed in the simulation

study. The slope is defined as $\beta = -0.5$, i.e., 0.5 unit decreasing change per month. In addition, $\alpha = 100$ for $n = 1$ and $\alpha = \{100, 101, 102\}$ for $n = 3$. The sampling time points in month are based on a full FDA sampling plan, i.e., $x = (0, 3, 6, 9, 12, 18)$. There are three error distributions considered in the simulation:

- Normal [$e_{ij} \approx N(0, 1)$]
- Uniform [$e_{ij} \approx U(-2, 2)$]
- Log-normal [$\log e_{ij} \approx N(0, 1)$]

One thousand sets of data for each error distribution and for number of batch $n = 1$ and $n = 3$, respectively. The mean and standard deviation of the estimated intercepts and slopes are summarized in Table 1. It is noticeable that the parameter estimates of rank regression models are comparable in terms of efficiency with those of normal linear regression models with the ordinary least squares (OLS) for data with normal errors, or other errors that do not deviate much from normal errors. The OLS estimates for normal linear models tend to be a bit more efficient than the rank regression estimates with the normal assumption. This can be seen from their smaller standard deviations. However, the rank regression is more robust and efficient for data with errors that deviate much from normal errors.

ALTERNATIVE APPROACHES

In addition to the approach outlined in the previous sections, there are also some other rank-based methods that can be used in stability analysis as well. For a specific example, consider the model in the special single-batch occasion, $Y_j = \alpha + \beta X_j + e_j$. As assumed, the random errors are identically distributed; therefore any pair of (e_i, e_j) are comparable, in the sense that $E\{I(e_i e_j)\} = 1/2$. Therefore the following estimating equations can be used to reasonably estimate the slope parameter of β :

$$U(\hat{\beta}) = \sum_{i=1}^n \sum_{j=i+1}^n I[y_i - y_j - \beta(x_i - x_j) \leq 0] - (n^2 - n)/4 = 0$$

It is not difficult then to show that $E[U(\beta)] = 0$ and $\text{var}[U(\hat{\beta})] = \text{var}\left[\sum_{i=1}^n \sum_{j=i+1}^n I(e_i \leq e_j)\right] = (n^2 - n)/8$. To estimate the intercept α , the similar estimator of the medians of $Y_i - \beta X_i$ can be used as $\hat{\alpha}$.

For the three different situations in the presence of the batch-to-batch variations, the above estimating equations can be modified accordingly to estimate the slopes and intercepts, as was done in the previous sections similarly.

In addition to the exact method to compute the confidence intervals, a bootstrap method can be also used to obtain the confidence intervals. Specifically,

Table 1 Mean and Standard Deviation of Estimated Model Parameters

Parameter		$e_{ij} \approx N(0,1)$	$e_{ij} \approx U(-2,2)$	$e_{ij} \approx N(0,1)$
[$n = 1$]				
RRM	α	100.0048 (0.6057)	99.9829 (0.8052)	101.1343 (0.7491)
RRM	B	-0.4995 (0.0328)	-0.5010 (0.0384)	-0.4941 (0.0462)
NLM	α	100.0077 (0.5489)	99.9992 (0.6230)	101.6468 (1.1458)
NLM	B	-0.4994 (0.0316)	-0.5012 (0.0359)	-0.4997 (0.0623)
[$n = 3$]				
RRM	α_1	99.9860 (0.5051)	100.0252 (0.6642)	101.0751 (0.4958)
RRM	α_2	100.9893 (0.5012)	101.0363 (0.6622)	102.1150 (0.5241)
RRM	α_3	101.9936 (0.4877)	102.0038 (0.6615)	103.1259 (0.5476)
RRM	B	-0.4996 (0.0193)	-0.4998 (0.0227)	-0.4981 (0.0182)
NLM	α_1	99.9915 (0.4411)	100.0097 (0.4955)	101.6286 (0.9150)
NLM	α_2	100.9961 (0.4302)	101.0180 (0.4847)	102.6333 (0.9056)
NLM	α_3	101.9827 (0.4275)	102.0043 (0.4960)	103.6305 (0.8960)
NLM	B	-0.4995 (0.0176)	-0.4997 (0.0205)	-0.4993 (0.0382)

RRM: Rank regression model. NLM: Normal linear model (fitted by OLS). Standard deviations are enclosed in parenthesis. $n = 1$ or 3 represents number of batches. True parameters are $\beta = -0.5$, $\alpha = 100$ for $n = 1$; and $\alpha_1 = 100$, $\alpha_2 = 101$, and $\alpha_3 = 102$ for $n = 3$.

1. For a batch data i , m_i pairs of (x_a, x_b) are picked randomly with replacement.
2. Use the proposed method to get $\hat{\beta}$.
3. Repeat the above two steps for B times, thus we have $\hat{\beta}_b$ for $b = 1, 2, \dots, B$.
4. Take the 2.5 percentile of $\hat{\beta}_b$ as the lower confidence bound.
5. $\hat{\alpha}$ is estimated as the median of $Y_i - \hat{\beta}_{LB} X_i$.
6. For a continuous X^* including the range of X , the estimate of Y^* is computed as: $Y^* = \hat{\alpha} + \hat{\beta}_{LB} X^*$.
7. The time at which Y^* intersects the prespecified limit is the shelf-life.

Simulations have been conducted to verify the validity of the alternative approaches. Similar results were obtained with the data example in Ref. [11], which are consistent with the ones in Ref. [7].

CONCLUSION

The semiparametric regression models and their associated rank-based inference procedures in obtaining the shelf-life provide an alternative approach when the parametric assumptions are not true in the fully parametric models. Although they may carry disadvantages in efficiency when the parametric models may fit well, their flexibility to allow nonstandard continuous outcomes in stability data is appealing in practice when the regulatory agencies or drug manufacturers tend to be conservative in determining the shelf-life of actual products. In practical computing, the subroutines of PROC IML in SAS, such as NLPNMS (Nelder-Mead algorithm) and NLPQN (Quasi-Newton algorithm), can be used

to minimize $D(\beta)$ without requiring the second derivatives, and an updated estimator of τ can be found in Ref. [12].

The feature of the nonparametric components in this model also has profound practical interpretation. They do not need parametric assumptions. The proposed approaches do not explore the statistical justification whether or not the determined shelf-life estimates its true counterpart, but simply adopts the minimum approach of the FDA guidelines because of their simplicity and uniformity as an algorithm in choosing an intuitive shelf-life. Thus the results are also comparable to the available ones that conform to the FDA guidelines.

REFERENCES

1. International Conference on Harmonization. *Guideline for Evaluation of Stability Data*; ICH, 2003.
2. Food and Drug Administration. *Guideline for Submitting Documents for the Stability of Human Drugs and Biologics*; FDA: Rockville, MD, 1987.
3. Hettmansperger, T.P. *Statistical Inferences Based on Ranks*; John Wiley: New York, 1984.
4. Hettmansperger, T.P.; McKean, J.W. *Robust Nonparametric Statistical Methods*; Arnold: New York, 1998.
5. Jaeckel, L.A. Estimating regression coefficients by minimizing the dispersion of residuals. *Ann. Math. Stat.* **1972**, *43*, 1449–1548.
6. Shao, J.; Chen, L. Prediction bounds for random shelf-lives. *Stat. Med.* **1997**, *16*, 1167–1173.
7. Chen, Y.Q.; Pong, A.; Xing, B. Rank regression in stability analysis. *J. Biopharm. Stat.* **2003**, *13* (3), 463–479.
8. Chow, S.-C.; Shao, J. *Statistics in Drug Research: Methodologies and Recent Developments*; Marcel Dekker: New York, 2002.

9. Chen, Y.Q.; Jewell, N.P. On a general class of semiparametric hazards regression models. *Biometrika* **2001**, *88*, 687–702.
10. Schuster, E. On the rate of convergence of an estimate of a functional of a probability density. *Scand. Actuar. J.* **1974**, *1*, 103–107.
11. Shao, J.; Chow, S.-C. Statistical inference in stability analysis. *Biometrics* **1994**, *50*, 753–763.
12. Koul, H.L.; Sievers, G.L.; McKean, J.W. An estimator of the scale parameter for the rank analysis of linear models under general score functions. *Scand. J. Stat.* **1987**, *14*, 131–141.

Regulatory Statistics

Estelle Russek-Cohen, Tsai-Lien Lin, and Shiohjen Lee

U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

Regulatory statistics is the scientific discipline consisting of the development and application of statistical methods, tools, approaches and other relevant processes derived from statistical science (and its sister disciplines) used to assess the safety, efficacy, quality and performance of regulated pharmaceutical products. In this article we highlight why such research is essential and touch on a number of topics in this arena. Finally we discuss topics in this area likely to gain prominence over the next few years.

Keywords: FDA; Innovative Trial Designs; PK and Bioequivalence; Postmarket Surveillance; Bayesian statistical methods; Biomarkers and Bioassays.

INTRODUCTION

Regulatory Science is “the science of developing new tools, standards and approaches to assess the safety, efficacy, quality and performance of FDA-regulated products”. Obviously such methods could be applicable to products regulated by FDA counterparts in other regions of the world.^[1]

Regulatory statistics is analogous. “Regulatory statistics is the scientific discipline consisting of the development and application of statistical methods, tools, approaches and other relevant processes derived from statistical science (and its sister disciplines) used to assess the safety, efficacy, quality and performance of FDA regulated products.” The definition is broad, its boundaries are vague and methods keep evolving.

In this article, we highlight areas of statistical research motivated by a regulatory need. In particular, we highlight areas that tend to have greater emphasis in a pharmaceutical regulatory environment though the US Food and Drug Administration (FDA) regulates more than just drugs and biologics. Statistical methods impact all aspects of product development ranging from pre-clinical, review of clinical trial protocols, marketing applications to post-market surveillance. Some of the assessments relate to manufacturing, evaluation of key biomarkers, and moving from pilot lots in pre-market studies to full production lots in post-market. However, many in the statistical community associate regulatory review at the FDA to be focused on clinical trials from trial design to trial outcomes.

The views expressed here are those of the authors and not necessarily those of the U.S. Food and Drug Administration.

Statisticians in a Regulatory Agency

Statisticians perform a key role in a regulatory agency such as the FDA. While they review protocols and results for many kinds of studies, they are in a position to identify gaps in statistical methodology; methodology and its consistency with policy and/or regulations; and make significant research contributions both in refereed journals and in regulatory guidance.

Guidances

There are many requirements for the approval of medical products at the FDA and key ones include those imbedded in the Code of Federal Regulations (e.g. 21CFR 314.126^[2] calls for adequate and well controlled studies). However in the need for transparency and to assist companies and FDA staff, current thinking is often documented in guidance and we will cite them as appropriate below. Each regulatory agency issues them (e.g. the FDA and its European counterpart the European Medicines Agency (EMA)). In addition, the International Council on Harmonization (ICH) develops guidelines that are recognized by multiple regulatory agencies. Among these, E9 Statistical Principles for Clinical Trials,^[3] is among the most cited.

FDA also participates in guideline development that is the result of public private partnerships, e.g. the US Pharmacopeia. The latter guidelines are mentioned in the discussions on statistical aspects of assessing non-clinical data below.

It should be noted that recommendations in guidance do not equate to regulation and companies can suggest alternatives in many situations.

STUDIES WITH ANIMAL MODELS

Pharmacology and toxicology studies in animals are required before the first-in-human trial. Depending on the product type and the intended population, other long term toxicology studies may need to be conducted at a later stage and submitted for marketing application, for example, carcinogenicity, genotoxicity and reproductive toxicity studies.^[4]

For vaccines or therapeutic products that cannot be evaluated for efficacy in humans (e.g. products to evaluate bioterrorism agents) there is the potential to be licensed or approved via the animal rule pathway, efficacy is demonstrated in proper animal models.^[5] Biomarkers that bridge data from animals to responses in humans often generate interesting statistical challenges. Note that safety for these products will need assessments in humans though that may not pose unique statistical challenges.

EARLY STAGE CLINICAL/DATA IN HUMANS

Prior to the approval of a drug or biologic product, there are often several development stages, in part because it would be unethical to expose a large number of subjects to an investigational product without a preliminary assessment of safety in humans. These sequential stages usually start from preliminary assessment of safety, to exploration of preliminary efficacy and additional safety data, and to the confirmation of safety and efficacy of the product. They are mostly referred to as Phase 1, Phase 2 and Phase 3; or early Phase and late Phase; or pilot Phase and pivotal Phase. Each stage has its role in product development. Depending upon how much we know about the investigational product, not all stages are needed prior to a product's approval. In this section we identify several major classes of studies along with some methodological concerns, some are randomized clinical trials but others are not.

Early Phase Exploratory Studies

Early phase studies are conducted to investigate an initial assessment of the safety and tolerability of the product, types of adverse events observed, and preliminary activity in treating disease. Often smaller studies are performed. The size of early phase studies may vary by therapeutic area but they are always smaller than confirmatory studies within the same area. Subjects in these studies may not be representative of the desired indication for use population, for example in early phase cancer studies, subjects may not have the same type of cancer or may be in very late stage of disease with no treatment options remaining. Often multiple doses are considered in early phase studies and Bayesian,^[6] frequentist and sometimes purely descriptive approaches are seen. In oncology^[7] therapeutic area, one goal of an early phase study is to identify the

maximum tolerated dose (MTD) but in studies for less serious conditions, the endpoints in these early phases are likely to differ.

Pharmacokinetic (PK) Studies

For new drugs and therapeutic proteins such as monoclonal antibodies, part of the requirement for submitting to FDA is a PK study that looks at the single-dose and/or steady-state concentration-time profile of the drug and/or key metabolites or other biomarkers thought to be related to clinical performance following administration of a product. The PK parameters estimated from the data, such as area under the curve (AUC), the maximum concentration observed for each subject or C_{max} , the timing of the maximum concentration or T_{max} , clearance rate, and half-life, etc., can inform both efficacy and safety and support dose and dosing schedule considerations. For generic drugs, the approval is often based on a PK study only, to demonstrate bioequivalence between a generic product and a licensed reference product.^[8]

In most PK studies, a simpler non-compartmental model approach is often used to derive the PK parameters for each individual subject, which generally yields comparable results as a compartmental modelling approach. However, in some situations a modeling approach is desired or needed. For example, when blood samples are collected from patients and thus are sparse or at different times across different patients or when the goal is to predict PK parameters or concentration-time profiles, a population PK model is often used to analyze the data from all subjects simultaneously. A population kinetic modeling analysis may involve the use of non-linear mixed effects models which include subject specific terms as random effects. Covariates such as body weight and age can be included in the model as well. Both Bayesian and frequentist approaches are used though frequentist methods are much more common. Bayesian approaches to population PK modeling may allow for more personalized dose regimes while still limiting the number of blood samples that have to be collected on any subject but could provide a path to developing pediatric dose recommendation.^[9]

LATE PHASE/CONFIRMATORY CLINICAL TRIALS

The role of phase 2 or proof of concept studies may vary with therapeutic area. These studies can include somewhat complex dose finding algorithms^[10] that aim to identify a dose that is safe and efficacious in treating the intended disease. With respect to study size, phase 2 studies are generally larger than in phase 1 in order to collect more safety data or/and obtain a reliable treatment effect size for future planning. Sometimes the investigational product is compared with an already approved treatment, or with

a placebo. Sometimes the phase 2 study is a randomized controlled trial.^[11] If phase 2 study results show that the investigational product may be as good as the existing treatment, or better than a placebo, it then moves into confirmatory phase that aims to confirm the safety and efficacy of the product for the intended disease for registration purposes. Phase 2 studies ought to be designed to gather data for planning an appropriate future phase 3 study with a reasonable likelihood of success. Sometimes more than one phase 2 study is needed because questions are generated after an initial phase 2 study.

Confirmatory studies are sometimes called phase 3 studies or registration studies as the results are often captured in a product label. In these phase 3 studies it is considered important to pre-specify how the results will be evaluated and how type 1 error will be controlled. Evidence that the study has adequate power to meet its objectives is also an important consideration. Statistical methods in the literature have most often been focused on the later phases of development but in recent years have provided means of combining more than one phase in a single study.

In planning a late phase study it is important to anticipate what can go wrong or plan for how some data is handled. For example, patients may want to go onto rescue medication if they fail to respond to a therapy and even if they stay in the study, how that information is included in the assessment of treatment benefit ought to be discussed in advance.^[12] In general, the interpretation of study results could be problematic if there is a high rate of missing data or a high rate of protocol deviations in the study since study results calculated using different approaches to handling missing data or dealing with protocol deviations may lead to inconsistent interpretations.^[13,14]

Randomization

The most common feature in phase 3 clinical trials is randomization. How one does randomizations and cautions over getting it done appropriately has been a common topic in the biostatistics literature and of concern in a regulatory submission. The objective of randomization in clinical trials is to balance subjects' demographic and baseline characteristics before receiving study therapies, so a sensible treatment comparison can be made at study completion. The most important advantage of randomization when properly performed is that it minimizes selection and allocation bias by balancing both known and unknown factors in the random assignment of treatments. As a result, randomization serves the basis for a statistical test of treatment effect that requires very few assumptions.

There are different types of randomization schemes.^[15] The most basic one is complete randomization (or simple randomization). Complete randomization on average achieves balance between treatment groups on both known and unknown factors. Imbalances do occur sometimes, especially in small trials and when there are many

covariates. One remedy is stratification with the use of permuted block randomization within subgroups defined by pre-specified prognostic factors. However, stratified randomization can become counterproductive when there are many prognostic factors to balance possibly resulting in a small number of patients within each stratum. Dynamic allocation has been proposed as an alternative when there are a large number of covariates to be balanced. Minimization is a type of dynamic allocation which has been used commonly, especially in oncology clinical trials.^[16] FDA would generally expect an analysis that respects the kind of randomization used in the study design so some cautions exist over the use of dynamic allocation methods.

Some of the advantages of randomization can be lost if protocol deviations or missing data disproportionately impact one treatment group and not another. If the reasons for missing data vary by treatment arm, e.g. placebo recipients drop out because of lack of efficacy and treatment recipients drop out because of side effects, there can still be some concerns in interpretability of trial outcomes.

Adaptive Randomized Designs

Adaptive randomized trial designs allow modification of trial procedures and/or statistical methods of ongoing clinical trials using accumulated data learned from interim analyses. A potential beneficial feature of adaptive design is that it may speed the development process of drug/biological products and therefore reduce costs, streamline the product development process and thereby leads to earlier availability of products.

There are numerous adaptive design features that can be applied to clinical trials including adaptive randomization, adaptive dose finding, sample size re-estimation^[17] and can include Bayesian approaches for dose escalation and dose finding. There have been many proposed adaptive design methods^[18,19] in the literature. When adaptive design protocols are submitted to a regulatory agency, how the design would generate valid and interpretable outcomes is essential. Some of the key considerations of adaptive design clinical trials are captured in 2010 FDA draft Guidance,^[20] *Adaptive Design Clinical Trials for Drugs and Biologics*. The emphasis of the FDA draft Guidance document is on adequate and well-controlled studies since that is consistent with the regulations. Though adaptive designs allow modification of design features during the course of trials, all adaptations should be planned ahead and pre-specified at the design stage. Study designs and proposed analyses will be evaluated for adequate control of type 1 error and the potential for operational biases.

Ambitious study designs that incorporate a number of adaptive features may be hard to interpret. Therefore, in therapeutic areas where high dropouts are common, or where there is heterogeneity in the patient population being studied, careful attention needs to be made to the ultimate interpretation of study results. Clinical and statistical

judgment may be factored into the evaluation of any study but with adaptive designs, simulations can be particularly useful. Simulations can be used for not only just demonstrating power and/or type 1 error control but also to generate a discussion between clinician and statistician at the planning stages. When simulations are conducted for others to examine, clear documentation and justification for what the results demonstrate are expected.

Bias can be an issue in any study, but the bias that comes because of one or more interim looks of data to modify study design is of special concern with adaptive designs. Other articles in this encyclopedia expand on the bias issues.

Companies that elect to use an adaptive design may often engage a data monitoring committee (DMC) to both monitor for safety and evaluate interim analyses.^[21] The role of DMC in adaptive design clinical trials could become more complex.

Before someone engages in planning an adaptive design, common sense ought to be that they examine whether a simpler design will suffice to demonstrate the safety and efficacy of the proposed product and/or answer scientific questions. It is in the best interest of public health to get new products to market for patients in need. However a poorly planned adaptive design is both time consuming and costly and may not bring new products to market sooner.

Bayesian Approach

Bayesian approaches can be applied to just about any phase of product development and sometimes have adaptive features though at FDA we have seen them more commonly in early phase studies. This is likely to change in the future and one would anticipate more research around Bayesian designs suitable for a regulatory submission. As noted earlier, in oncology a variant of a continual reassessment method is common in phase 1 but one may see a different type of Bayesian dose finding algorithm in phase 2. It is less common to see a Bayesian approach in late phase or confirmatory trials in drug development, though recently more have been agreed to. Type 1 error rates are not part of a classical Bayesian formulation but because a regulatory agency would be uncomfortable approving a product based on a study that came with a high potential for a type 1 error, companies still need to demonstrate, often via simulation, that type 1 error is sufficiently small and controlled. Current research in this area includes best practices for demonstrating type 1 error control for the purposes of a regulatory submission.

One of the strengths of Bayesian methods is the natural way one can combine data from multiple sources. Bayesian methods have been proposed for “extrapolation” to pediatric populations where adult data for the same indication can form an informative prior or be used in some form of Bayesian hierarchical model. Although FDA has issued a guidance for pediatric extrapolation for medical devices

including some suggestions that a Bayesian approach to combining adult and pediatric data would be appropriate,^[22] many of the principles are not unique to devices and may find their way into drug or biologic approvals.

Among the most innovative use, Bayesian methods have been used in the development of platform designs where multiple products from multiple companies are tested in the same trial and treatment arms may be added or dropped during the trial. The ISPY2 Phase 2 trial^[23] for the treatment of neoadjuvant breast cancer is one such example but more have been proposed. Each one is likely to be unique and geared to what is known about a given disease and can consider stratification of patients within the disease (e.g. using biomarkers to define patient subgroups).

Noninferiority Trials

An active-control trial is one of four kinds of concurrently controlled trials that provide evidence of effectiveness described in FDA’s regulation on adequate and well-controlled studies (21 CFR 314.126). Active-control trials often rely on non-inferiority testing. The objective of a non-inferiority trial, per FDA’s Guidance on Non-Inferiority Clinical Trials,^[24] is to show that the difference between the investigational product and active control treatment is small, small enough to allow the known effectiveness of the active control to support the conclusion that the investigational product is also effective.^[25]

Multiple Endpoints

In evaluations of an investigational product, clinical endpoints are measures in a clinical trial to reflect the effects of the product. Because of complexity of diseases, there is often interest in more than one endpoint for a study indication. For example, with a pneumococcal vaccine, there may be an immunogenicity endpoint for each of several types of pneumococcal subtypes and the goal would be the success on all such endpoints. No adjustment in type 1 error control would be made if all co-primary endpoints need to be successful. When there are many endpoints of interest in a clinical trial and success of any endpoint could be considered evidence of effectiveness, failing to adjust the assumed type 1 error for each of several possible success scenarios would result in an increasing rate of falsely concluding a therapy is efficacious when in fact it is not. Such a consideration is particularly important for regulatory decision making in adequate and well-controlled studies that serve as the primary evidence in pre-market applications. For additional discussions, see Bretz et al^[26,27] and Huque et al.^[28]

Historically Controlled Trials

Historically controlled trials are trials intended to support a comparative conclusion regarding an investigational therapy, where the ‘comparator’ is not a concurrent group

of subjects. Rather, data of the ‘comparator’ are from a prior study, studies, or possibly a registry. Therefore, one would need to have legal access to the historical data on the patient level and all the appropriate baseline covariates need to have been measured and collected. This may be a challenge. Such trials are not commonly seen in drug or biologic product development for regulatory purposes. However, they have been considered for rare diseases for which few therapeutic options exist. This kind of study could be more plausible in those rare disease settings where the natural history of the disease and its progression are well documented.

Obviously historically controlled trials don’t have the interpretative advantage of a randomized clinical trial and so one would expect them to be less common in most pre-market therapeutic settings. Propensity score methods^[29] originally developed for observational settings have been used largely in confirmatory medical device trials in an effort to account for some differences between subjects that are in a current trial versus those that come from an older study. The methods have advantages over not using propensity scores especially when one knows what covariates belong in the score but unlike randomization cannot balance for covariates that are not part of the score.

Sometimes one sees a study that is a single arm study with a pre-specified success criteria based on literature and clinical judgment. This type of study tends to be seen in those therapeutic areas in which the outcome of a control arm is thought to be well understood; e.g. when the natural course of the disease is well documented. In addition, ethical factors can weigh in. So with some rare diseases or diseases that are often fatal such as late stage cancers, it may be considered unethical or unnecessary to use a control arm. Although we have not seen this often in a drug and biological setting, perhaps a Bayesian approach that borrows some historical control information is preferable to a study that relies solely on external data to characterize the control group.

Biomarkers In Late Stage Trials

Biomarkers are used in various stages of drug development. Some are defined in terms of patients characteristics at baseline and can include prognostic markers related to disease severity or anticipated outcomes in the absence of treatment. It can also include predictive markers that identify patients who may benefit from a specific therapy.^[30]

In the academic literature, there is a definition of a surrogate endpoint that can be used to predict treatment benefit using an endpoint of primary concern.^[31] For example, improvements in HIV viral load are used in place of advancing from being HIV positive to development of AIDS related symptoms.^[32] Within a regulatory setting and defined in regulations, a surrogate endpoint is an endpoint reasonably likely to predict clinical benefit and may form the basis of accelerated approval. The two definitions are not equivalent.

Subgroups In Clinical Trials

Heterogeneity in treatment response within a single trial is a recognized outcome and regulatory submissions have to address treatment response by demographic subgroups, and possibly treatment response by region (e.g. by U.S. versus non-U.S.). The complexity of addressing this is not trivial because of possible inflation of type 1 error rate and also inadequate power to identify treatment differences.^[33,34]

Quality By Design In Clinical Trials

There is a recognition that modern Clinical Trials have become quite complex and many can fail to meet key objectives if not planned correctly.^[35] Having an eye on the study objectives throughout the planning process is helpful and statisticians can play a role in developing study design options that lead to more interpretable outcomes. For example, by selecting subjects more likely to comply with directions (“enriching for compliers”), we may have fewer missing data points. Getting investigators to work together, and standardizing more subjective endpoints can lead to a cleaner result. Statisticians can also help with algorithms to monitor trials in order to identify problems (e.g. an unforeseen rise in a particular safety concern).^[36] Statisticians need to be engaged in all aspects leading up to a clinical trial including planning, conduct and analysis.

Bioequivalence And Biosimilarity

In academic training the emphasis is on clinical trials designed to assess treatment differences or superiority of one treatment over another. In a regulatory environment while there is always a need for superiority trials, there is also a need for demonstrating two products are sufficiently similar. For generic drugs the term is bioequivalence and for a biologic, the term is biosimilarity.^[37,38] There is a well-established literature on bioequivalence but a more limited literature on biosimilars.^[39]

Vaccines For Preventing Disease

The Center for Biologics Evaluation and Research (CBER) at FDA regulates vaccine license applications and works with other agencies to monitor vaccine efficacy and safety in post-licensure. Even when a new vaccine is compared to a placebo control, it is often not adequate to demonstrate that the vaccine is superior to a placebo, but a more stringent criteria needs to be set. Vaccine recipients are often healthy and in the context of balancing risk and benefit, one may ask that the vaccine demonstrate that it can prevent at least a pre-specified percentage of cases of the disease. This would lead to a super-superiority hypothesis of primary concern.^[40,41] In addition to clinical efficacy studies, vaccine lot to lot consistency usually

needs to be demonstrated in a clinical study using immunogenicity endpoints to assess equivalence among three consecutive lots.

A particular topic of interest in the development of a novel vaccine is a biomarker that can be used to determine who is or is not protected by the vaccine, a correlate of protection. Study designs to establish the existence of a correlate of protection at individual or population level are of great interest in a regulatory setting.^[42–44] Immunobridging methods to assess vaccine efficacy in a population different from the population in which the correlate of protection is established may also present challenges. A few such studies have been submitted to FDA.

Large Simple Safety Trials

The typical premarket assessment of a novel product involves a detailed assessment of any observed adverse events. Some caution is needed in separating events associated with the disease being treated and events that may be attributable to the drug investigated in the trial. If the events are rare but it is unclear if a signal, though rare, may exist, companies may be asked to conduct a large simple safety trial. The concerns over evaluating some of these have been documented particularly in the context of cardiovascular outcome trials^[45] and new methods have been proposed.^[46] Because safety is of paramount concern in vaccines, we have also seen an efficacy study embedded in a larger safety study (e.g. REST trial).^[47] They are often called ‘simple’ largely because they focus on a limited amount of safety data collected on each subject in the trial.

Benefit-Risk

Regulators and companies are increasingly being asked to consider the balance of risks and benefits of medical products. This has led to a formal literature including a recent book by Qi Jiang et al.^[48] The more quantitative approaches including Bayesian methods are likely to provide opportunities for formal statistical approaches that capture uncertainty in the various pieces of evidence contributing to the assessment.

POSTMARKET SURVEILLANCE

Once a product is approved or licensed, FDA and companies often engage in collecting additional post-market data. Most premarket clinical trials only provide a preliminary evaluation of product safety and may fail to capture rare but serious adverse events. There are two very common types of surveillance.

FDA provides for voluntary reporting of suspected adverse reactions to a drug or biologic or possibly a reaction when two or more products are taken together. Passive surveillance refers to the idea that only complaints are

registered and we recognize that while we see complaints we may not know how many units of the product were dispensed. This has led to a number of statistical algorithms focused on data mining in passive surveillance systems. Empirical Bayesian methods^[49] and more recent Likelihood methods^[50] are designed to find unanticipated safety signals that are then further investigated using other data sources.

Active surveillance refers to the idea that FDA and others may elect to pursue adverse events using a system that provides numerators and denominators and is studied in a prospective manner. The Vaccine Safety Datalink is among the oldest such systems (since 1991) and methods such as Rapid Cycle Analysis^[51] have been used to successfully study a pre-specified potential safety concern using data from electronic health records. FDA engages in joint activities with other federal agencies and some interesting statistical advances have come of that. The biggest FDA-sponsored post-market active surveillance system to date is the Sentinel system and the reader is encouraged to visit the Sentinel system website (<http://www.sentinel-system.org>). Even though Sentinel is considered an active surveillance system, it is often used to take a look a retrospective look at insurance claim data and/or electronic health records with millions of individuals in the system. Claims data are often updated over a period of months because of claims being settled and then recorded in the dataset. One may need to account for a lag in reporting of relevant data in any model that analyzes data as rapidly as possible. This can be a challenge with evaluating safety of new products and/or seasonal influenza vaccines that are reformulated each year.^[52] Claims data can also have limitations since one may track what procedures have been billed for but sometimes a final diagnosis may not be recorded. When combining claims data from multiple providers, one needs methods that work in a distributed system since the detailed patient level data needs to stay behind a firewall.^[53]

ESSENTIALS OUTSIDE OF CLINICAL DATA

Regulatory agencies regulate the entire product and look at more than just data from one or more clinical trials. This can include evaluating lab assays that are critical to product evaluation and manufacture. For example, the release and stability assays and their specifications for all quality attributes of a drug or biological product must be established prior to licensure. Many changes related to product manufacture and release test may occur at different stages of the product life cycle. For example, a biological product may be produced in a pilot scale during the product’s development but need to be scaled up to produce millions of doses annually before licensure. Potency assays may change post licensure as technology advances, which may or may not require change of product specifications. Manufacturing

site, with or without scale-up, or testing site may change. Critical manufacturing materials, critical test reagents, or instrument/equipment may need to be replaced. FDA is regulating the product for the purpose of protecting public health and modifications to manufacture and acceptance criteria may evolve from early testing to later phases and in post-market. Comparability/bridging before and after each change needs to be evaluated to ensure the product remains efficacious and safe as the clinical lots used in the clinical trials to demonstrate efficacy and safety of the product. Statistical contributions in this area are substantive (see FDA Guidance on process validation).^[54]

Laboratory assays are essential in the support of clinical development of a product too. For generic drugs, approval is often based on bioavailability/bioequivalence studies which evaluate the kinetic profile of drug concentrations measured by assays. For new drugs or biologics, when a clinical efficacy trial is unfeasible, unethical (e.g., anthrax, pandemic flu, Ebola), or takes too long to conduct, biomarker data will be the key evidence supporting efficacy under accelerated approval or animal rule pathway. For therapeutic biologics, an immunogenicity assay is also an important tool for safety evaluation (see FDA draft Guidance on immunogenicity assay).^[55] For vaccines, serology assays used to measure vaccine induced immune response are key for developing a vaccine and interpreting modifications of manufacturing. Selection of vaccine dose, schedule, or formulation and lot consistency studies are usually based on immunogenicity. When there is an established correlate of protection, efficacy of a vaccine may be demonstrated by immunogenicity studies. Therefore, assays play a critical role in the preclinical and clinical development programs of many therapeutic and preventive products. The validation of bioanalytical assays used for quantifying analytes in complex biological matrix involves issues different from assays used for testing product quality, for example, the need of using incurred samples (see FDA draft Guidance on bioanalytical method validation),^[56] and thus presents different statistical challenges in the design and analysis of validation studies.

FDA engages many stakeholders in this setting. In addition to FDA guidance, FDA staff participate in many external organizations to further best practices. For example, U.S. Pharmacopeia is an organization that develops many guidelines^[57–60] associated with bioassays for drug and biologic development and assessment. Physicians may elect to recommend a drug for a patient based on the results of a laboratory test and the Clinical Laboratory Standards Institute (CLSI) publishes guidelines that are used in the in vitro diagnostics industry and are the result of an agency and industry working group; multiple CLSI guidelines have been recognized by FDA.^[61] Guidelines may address the development of an assay analysis method by discussing issues in model selection, transformation, weighting, and calculation of potency, etc. and the performance of an assay by recommending designs and statistical analysis

methods to assess precision, accuracy, linearity, specificity, robustness, stability and so on. ICH Q2(R1)^[62] provides definitions to assay validation parameters recognized by drug and biologic regulators. Statistical methods for evaluating assay performance have been evolving over the recent years, and have been or are being updated in some guidances. For example, equivalence approach is now well recognized as the appropriate way for testing parallelism, a fundamental assumption of dilution assays, and for testing assay comparability. How to set equivalence margins, however, remains a difficult issue for many assays. Differences in parameter definitions and statistical approaches may still exist across guidances by different organizations or from different countries, for different types of assays, and for different uses of assays. There are plenty opportunities for new statistical ideas in this area.

TOPICS LIKELY TO GAIN MORE PROMINENCE

Precision Medicine

Precision medicine is and will continue to challenge statisticians. We are still far from tailoring a treatment to an individual but we are getting closer to recommendations for groups of individuals. Many groups in a study usually mean fewer subjects within any given group. Topics in this vein would include adaptive enrichment designs,^[63] the evaluation of companion and complementary diagnostics (diagnostics that provide useful information in the selection of therapies for patients).^[64]

Real World Evidence

Everyone is aware of the limitations of clinical trials relative to what happens once a drug is approved. Other kinds of studies could help inform the use of a drug. This can include alternative designs, e.g. a pragmatic trial^[65] that evaluates real world use of a clinical product for a broad group of subjects. This can sometimes include a group or cluster randomized trial in which e.g. a nursing home or a medical practice is assigned a treatment group rather than individual subjects (see Hayes et al^[66] and special issue in Clinical Trials October 2015). Registries and/or large databases may also be used to inform efficacy as well as safety. The methodology in this arena is continuing to be developed as is the context for the use of such studies in a regulatory setting such as FDA.

Science Of Patient Input

Regulatory agencies are listening to patient groups and would like to have formal assessments of patient reported outcomes (PRO)^[67] and patient preference information (PPI).^[68,69] Methodology normally seen in psychometrics and in the survey world (e.g. questionnaire

development) will likely influence how one develops a PRO and the methodology for eliciting patient preferences may be influenced by marketing research methodology.

CONCLUSIONS

As the discussion presented indicates, statisticians and statistical methods have a role to play in just about every phase of pharmaceutical development and assessments, whether by a company or a regulatory agency such as FDA. The role of statisticians in regulation in particular is often downplayed and too often trivialized as “guardians of a p-value”. Approval of medical products has become very complex and often information from multiple sources needs to be brought together to form the totality of evidence. Statisticians, with their quantitative training and their understanding of evidence synthesis and quantitative decision making will have an increasing role in a regulatory setting. We have highlighted more traditional activities but also tried to note some new methods likely to have increased visibility in a regulatory environment. Because the theme of this article is regulatory statistics we have not mentioned the substantial literature related to drug development from a pharmaceutical company perspective.

REFERENCES

1. US FDA. Advancing Regulatory Science for Public Health, 2010, www.fda.gov/downloads/scienceresearch/special-topics/regulatory-science/ucm228444.pdf.
2. Code of Federal Regulations. CFR 21, 314.126.
3. International Council on Harmonisation ICH E9: Statistical Principles for Clinical Trials, 1998, www.ich.org.
4. Lin, K.K. Carcinogenicity Studies of Pharmaceuticals. *Encyclopedia of Biopharmaceutical Statistics*, 2nd Ed., 2003, 160–174.
5. US FDA. Product development under the Animal Rule, 2015, <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM399217.pdf>.
6. O’Quigley, J.; Shen, L. Continual reassessment method: a likelihood approach. *Biometrics* **1996**, *52*, 673–684.
7. Le Tourneau, C.; Lee, J.; Siu, L. Dose Escalation Methods in Phase I Cancer Clinical Trials. *J. Natl Cancer Inst.* **2009**, May 20; *101* (10), 708–720.
8. Shen, M.; Russek-Cohen, E.; Slud, E.V. Exact Calculation of Power and Sample Size in Bioequivalence Studies Using Two One-Sided Tests. *Pharmaceutical statistics* **2015**, *14*, 95–101.
9. Bjorkman, S.; Oh, M.S.; Spotts, G.; Schroth, P.; Fritsch, S.; Ewenstein, B.M.; Casey, K.; Fischer, K.; Blanchette, V.S.; Collins, P.W. Population pharmacokinetics of recombinant factor VIII: the relationships of pharmacokinetics to age and body weight. *Blood* **2012**, *119*, 612–619.
10. Pinheiro, J.; Bornkamp, B.; Glimm, E.; Bretz, F. Model-based dose finding under model uncertainty using general parametric models. *Statist. Med.* **2014**, *33*, 1646–1661.
11. Rubinstein, L.; Crowley, J.; Ivy, P.; LeBlanc, M.; Sargent, D. Randomized Phase II Designs. *Cancer Clinical Research* **2009**, *15*, 1883–1890.
12. Mehrotra, D.; Hemmings, R.; Russek-Cohen, E. on behalf of the ICH E9/R1 Expert Working Group. Seeking harmony: estimands and sensitivity analyses for confirmatory clinical trials. *Clinical Trials* **2016**, *13*, 456–458.
13. National Research Council. The prevention and treatment of missing data in clinical trials. Panel on handling missing data in clinical trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. The National Academies Press: Washington, DC, 2010.
14. LaVange, L.; Permutt, T. A regulatory perspective on missing data in the aftermath of the NRC report. *Statistics in Medicine* **2016**, *35* (17), 2853–2864.
15. Rosenberger, W.F.; Lachin, J.M. *Randomization in clinical trials: Theory and Practice*; 2nd Ed.; Wiley: New York, 2015.
16. Xu, Z.; Proschan, M.; Lee, S. Validity and Power Considerations on Hypothesis Testing under Minimization. *Statistics in Medicine* **2016**, *35*, 2315–2327.
17. Lin, M.; Lee, S.; Zhen, B.; Scott, J.; Horne, A.; Solomon, G.; Russek-Cohen, E. CBER’s Experience with Adaptive Design Clinical Trials. *Therapeutic Innovation and Regulatory Science* **2016**, *50*, 195–203.
18. Chow, S.-C.; Chang, M. *Adaptive Design Methods in Clinical Trials*. Chapman and Hall: London, 2007.
19. Berry, S.M.; Carlin, B.P.; Lee, J.J.; Muller, P. *Bayesian Adaptive Methods for Clinical Trials*. Chapman & Hall/CRC Biostatistics Series: Boca Raton, FL, 2010.
20. US FDA. Adaptive design clinical trials for drug and biologics draft guidance, 2010. www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm201790.pdf.
21. US FDA. Guidance for Clinical Trial Sponsors on the Establishment and Operation of Clinical Trial Data Monitoring Committees; FDA; Rockville, MD, 2006. www.fda.gov/regulatoryinformation/guidances/ucm127069.htm.
22. US FDA. Guidance for Industry and Food and Drug Administration Staff: Leveraging Existing Clinical Data for Extrapolation to Pediatric Uses of Medical Devices, 2016.
23. Barker, A.D.; Sigman, C.C.; Kelloff, G.J.; Hylton, N.M.; Berry, D.A.; Esserman, L.J. I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy. *Clinical Pharmacology and Therapeutics* **2009**, *86*, 97–100.
24. US FDA. Guidance for Industry: Non-Inferiority Clinical Trials to Establish Effectiveness, 2016. <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm202140.pdf>.
25. Rothmann, M.D.; Wiens, B.L.; Chan, I.S.F. *Design and Analysis of Non-Inferiority Trials*. Chapman and Hall/CRC Biostatistics Series: Boca Raton, FL, 2011.
26. Bretz, F.; Maurer, W.; Brannath, W.; Posch, M. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* **2009**, *28*, 586–604.
27. Bretz, F.; Maurer, W.; Hommel, G. Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. *Statistics in Medicine* **2011**, *30*, 1489–1501.

28. Huque M.F.; Dmitrienko, A.; D'Agostino, R.B. Multiplicity issues in clinical trials with multiple objectives. *Statistics in Biopharmaceutical Research* **2013**, *5* (4), 321–337.
29. Yue, L.Q. Statistical and Regulatory Issues with the application of propensity score analyses to nonrandomized medical device clinical studies. *J. Biopharm. Statistics* **2007**, *17*, 1–13.
30. BEST (Biomarkers, EndpointS, and other Tools) Resource, 2016. <http://www.ncbi.nlm.nih.gov/books/NBK338448/>.
31. Prentice, R.L. Surrogate endpoints in Clinical Trials: Definition and Operational Criteria. *Statistics in Medicine* **1989**, *8*, 431–440.
32. Buyse, M.; Molenberghs, G.; Burzykowski, T.; Renard, D.; Geys, H. Statistical Validation of Surrogate Endpoints: Problems and Proposals, *Therapeutic Innovation & Regulatory Science* **2000**, *34* (2), 447–454.
33. Alosch, M.; Fritsch, K.; Huque, M.; Mahjoob, K.; Pennello, G.; Rothmann, M.; Russek-Cohen, E.; Smith, F.; Wilson, S.; Yue, L. Statistical Considerations on Subgroup Analysis in Clinical Trials. *Statistics in Biopharmaceutical Research* **2015**, *7*, 286–303.
34. International Council on Harmonisation ICH E17. General Principles for Planning and Design of Multi-Regional Clinical Trials, 2016. <http://www.ich.org>.
35. Sprenger, K.; Nickerson, D.; Meeker-O'Connell, A.; Morrison, B. Quality by Design in Clinical Trials: A Collaborative Pilot with FDA, 2012, *Therapeutic Innovation & Regulatory Science* **2012**, *47*, 161–166.
36. Zink, R., Risk-Based Monitoring and Fraud Detection in Clinical Trials Using JMP® and SAS®, 2014, SAS Institute.
37. US FDA. Guidance for Industry: Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Considerations, 2003.
38. US FDA. Guidance for Industry: Scientific Considerations in Demonstrating Biosimilarity to a Reference Product, 2015.
39. Chow, S.-C. *Biosimilars: Design and Analysis of Follow-on Biologics*. Chapman and Hall/CRC: Boca Raton, FL, 2013.
40. Mehrotra, D.V. Vaccine Clinical Trials – A statistical primer. *Journal of Biopharmaceutical Statistics* **2006**, *16*, 403–414.
41. Nauta, J. *Statistics in Clinical Vaccine Trials*, Springer: London, 2010.
42. Chan, I.S.F.; Li, S.; Matthews, H.; Chan, C.; Vessey, R.; Sadoff, J.; Heyse, J. Use of statistical models for evaluating antibody response as a correlate of protection against varicella. *Statistics in Medicine* **2002**, *21*, 3411–3430.
43. Gilbert, P.B.; Qin, L.; Self, S.G. Evaluating a surrogate endpoint at three levels, with application to vaccine development. *Statistics in Medicine* **2008**, *27* (23), 4758–4778.
44. Siber, G.R.; Chang, I.; Baker, S.; Fernsten, P.; O'Brien, K.L.; Santosham, M.; Klugman, K.P.; Madhi, S.A.; Paradiso, P.; Kohberger, R. Estimating the protective concentration of anti-pneumococcal capsular polysaccharide antibodies. *Vaccine* **2007**, *25*, 3816–3826.
45. Marchenko, O.; Jiang, Q.; Chuang-Stein, C.; Chakravarty, A.; Ke, C.; Ma, H.; Maca, J.; Russek-Cohen, E.; Sanchez-Kam, M.; Zink, R.C. Evaluation and Review of strategies to assess cardiovascular risk in clinical trials in patients with type 2 diabetes mellitus. *Statistics in Biopharmaceutical Research* **2015**, *7*, 253–266.
46. Little, R.; Wang J.; Sun X.; Tian H.; Suh E.Y.; Lee M.; Sarich T.; Oppenheimer L.; Plotnikov A.; Wittes J.; Cook-Bruns N.; Burton P.; Gibson C.M.; Mohanty S. The treatment of missing data in a large cardiovascular clinical outcomes study. *Clin Trials* **2016**, *13* (3), 344–351.
47. Vesikari, T.; Matson, D.O.; Dennehy, P.; Van Damme, P.; Santosham, M.; Rodriguez, Z.; Dallas, M.; Heyse, J.; Goveia, M.; Black, S.; Shinefield, H.; Christie, C.; Ylitalo, S.; Itzler, R.; Coia, M.; Onorato, M.; Adeyi, B.; Marshall, G.; Gothefors, L.; Campens, D.; Karvonen, A.; Watt, J.; O'Brien, K.; DiNubile, M.; Clark, H.; Boslego, J.; Offit, P.; Heaton, P. for the Rotavirus Efficacy and Safety Trial (REST) Study Team. Safety and Efficacy of a Pentavalent Human–Bovine (WC3) Reassortant Rotavirus Vaccine. *New England J. Medicine* **2006**, *354*, 23–33.
48. Jiang, Q.; He, W. *Benefit-Risk Assessment Methods in Medical Product Development: Bridging Qualitative and Quantitative Assessments*. Chapman and Hall/CRC Press: Boca Raton, FL, 2016.
49. Dumouchel, W. Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System. *The American Statistician* **1999**, *53*, 177–190.
50. Huang, L.; Zalkikar, J.; Tiwari, R.C. A Likelihood Ratio Test Based Method for Signal Detection With Application to FDA's Drug Safety Data. *Journal of the American Statistical Association* **2011**, *106*, 1230–1241.
51. Kulldorff, M.; Davis, R.L.; Kolczak, M.; Lewis, E.; Lieu, T.; Platt, P. A Maximized Sequential Probability Ratio Test for Drug and Vaccine Safety Surveillance. *Sequential Analysis* **2011**, *30*, 58–78.
52. Franks, R.; Sandhu, S.; Avagyan, A.; Lu, Y.; Hong, H.; Garcia, B.; Worrall, C.; Kelman, J.; Ball, R.; MacCurdy, T. Robustness properties of a sequential test for vaccine safety in the presence of misspecification. *Statistical Analysis and Data Mining* 19 Aug., **2014**, doi: 10.1002/sam.11234.
53. Cook, A.J.; Tiwari, R.C.; Wellman, R.; Heckbert, S.; Li, L.; Heagerty, P.; Marsh, T.; Nelson, J. Statistical approaches to group sequential monitoring of postmarket safety surveillance data: current state of the art for use in the Mini-Sentinel pilot, *Pharmacoepidemiology and Drug Safety* **2012**, *21*, 72–81.
54. US FDA. Guidance for Industry: Process Validation: General Principles and Practices, 2011.
55. US FDA. Draft Guidance: Assay Development and Validation for Immunogenicity Testing of Therapeutic Protein Products, 2016.
56. US FDA. Draft Guidance: Bioanalytical Method Validation, 2013.
57. US Pharmacopeia. General chapter <1225> Validation of Compendial Procedures. USP39-NF34 S1, 2016.
58. US Pharmacopeia, General chapter <1032> Design and Development of Biological Assays. USP39-NF34 S1, 2016.
59. US Pharmacopeia. General chapter <1033> Biological Assay Validation. USP39-NF34 S1, 2016.
60. US Pharmacopeia. General chapter <1034> Analysis of Biological Assays. USP39-NF34 S1, 2016.
61. US FDA. CDRH recognized standards, 2016. <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfStandards/search.cfm>.

62. International Council on Harmonisation ICH Q2(R1): Validation Analytical Procedures: Text and Methodology, 2005. www.ich.org.
63. Simon, N.; Simon, R. Adaptive Enrichment Designs for Clinical Trials. *Biostatistics* **2013**, *14*, 613–625.
64. US FDA. Draft Guidance for Industry and Food and Drug Administration Staff: Principles for Codevelopment of an In Vitro Companion Diagnostic Device with a Therapeutic Product, 2016.
65. Ford, I.; Norrie, J. Pragmatic Trials. *New England Journal of Medicine* **2016**, *375*, 454–463.
66. Hayes, R.J.; Moulton, L.H. *Cluster Randomised Trials*. Chapman and Hall/CRC Press: Boca Raton, FL, 2009.
67. Cappelleri J.C.; Zou K.H.; Bushmakina, A.G.; Alvir, J.M.; Alemanyeh, D.; Symonds, T. *Patient Reported Outcomes: Measurement, Implementation and Interpretation*. Chapman & Hall/CRC Biostatistics Series: New York, 2014.
68. Ho, M.P.; Gonzalez, J.M.; Lerner, H.P.; Neuland, C.Y.; Whang, J.M.; McMurry-Heath, M.; Hauber, A.B.; Irony, T. Incorporating patient-preference evidence into regulatory decision making, *Surg Endosc* **2015**, *29*, 2984–2993, 2987.
69. Johnson, F.R.; Lancsar, E.; Marshall, D.; Kilambi, V.; Mühlbacher, A.; Regier, D.A.; Bresnahan, B.W.; Kanninen, B.; Bridges, J.F.P. Constructing Experimental Designs for Discrete-Choice Experiments: Report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. *Value in Health* **2013**, *16*, 3–13.

Release Targets

Greg C. G. Wei

Pfizer Global Research and Development, Groton, Connecticut, U.S.A.

Abstract

This entry discusses how manufacturers of pharmaceutical drugs need to set the in-house specification on the initial potency high enough such that a batch that satisfies this specification at time zero will meet the LRL with high certainty, a specification standard known as release target.

INTRODUCTION

Manufacturers of drug products must demonstrate that their products meet with United States Pharmacopeia specifications before the products can be released for sale. One of the important specifications is the low registered limit (LRL) on drug's potency. After a drug product is manufactured, the natural degradation process gradually reduces the drug's potency. A drug's potency must remain above the LRL at the end of its registered shelf life. An important factor in the determination of the potency at the end of shelf life is the drug's potency at time zero, or the initial potency. Manufacturers need to set the in-house specification on the initial potency high enough such that a batch that satisfies this specification at time zero will meet the LRL with high certainty. This in-house specification is called the release limits (RLs) or the release targets. A typical scenario is illustrated in Fig. 1.

Some statistical approaches have been developed to determine the RLs of the initial drug potency. Allen et al.^[1] proposed a simple method for calculating the RLs. Wei^[2] suggested a similar but more statistically justifiable approach. Shao and Chow^[3] adopted a Bayesian decision theory approach to construct the RLs. This article will only describe the methods by Allen et al. and Wei. The Bayesian method by Shao and Chow will not be discussed in this article due to its technical complexity.

ALLEN ET AL. METHOD

The idea of their approach is very intuitive. To ensure potency of the shelf life above the registration limit, the RL was made equal to the registration limit plus a "buffer." The buffer consists of the potency decrease over the shelf life, the variability associated with the rate of decrease, and the variability of the assay. These quantities are estimated from stability studies using linear regression analysis.

Case I: Product with No Degradation

The expression of the RL is shown in the following equation:

$$RL = LRL + t \times \frac{S}{\sqrt{n}}$$

where RL is the release limit; LRL is the lower registration limit; S is the assay standard deviation; DF is the degrees of freedom; t is the 95% confidence (one-sided) t value with DF; and n is the number of replicate assays used for batch release.

Case II: Product with Degradation

For products whose potency decrease over the shelf life due to degradation process, RL can be calculated as follows:

$$RL = LRL + bT + t \sqrt{S_T^2 + \frac{S^2}{n}}$$

where b is the average degradation rate; T is the shelf life; and S_T is the standard error of bT .

The degrees of freedom used to determine t can be calculated by the Satterthwaite approximation.

Case III: Product Requiring Reconstitution Prior to Use

Some products must be reconstituted before use. The potency decrease due to reconstitution and its variability must be taken into account in RL calculation. RL can be calculated as follows:

$$RL = LRL + bT + rT + t \sqrt{S_T^2 + S_R^2 + \frac{S^2}{n}}$$

where r is the average degradation rate of the reconstituted solution; and S_R is the standard error of the average degradation rate of the reconstituted solution.

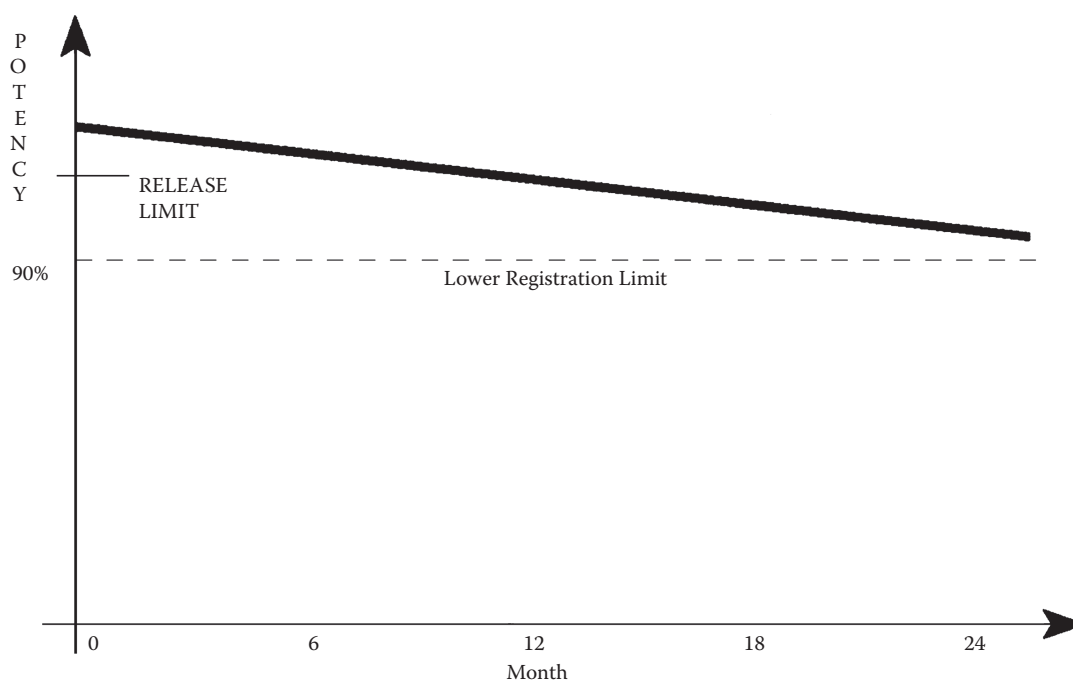


Fig. 1 Release limit of drug potency

The degrees of freedom used to determine t can be calculated by the Satterthwaite approximation.

CONDITIONAL AND UNCONDITIONAL METHODS

Allen et al.'s method for RL determination is very appealing due to simplicity and intuitiveness. However, it lacks a rigorous statistical justification. Wei^[2] proposed a more justifiable statistical derivation. His method is briefly described here.

Decision Rules

First, let us define two types of RLs in the following decision rules, expressing each one in a probability statement. The basic idea is that RL should be set at a value such that the probability of failing LRL at the end of shelf life is controlled.

Conditional Rule

Let y_{i0} be the observed sample average of time-zero potency. Find an RL such that if $y_{i0} \geq \text{RL}$, then

$$P(\text{Lower 95\% confidence interval (CI) of true potency at time } T < \text{LRL} | y_{i0}) \leq \alpha$$

This rule defines RL as a value such that if the potency measurement at time zero is above this value, then the probability of failure for this batch at time T will not be more than α .

Unconditional Rule

Find RL such that

$$P(\text{Lower 95\% CI of true potency at time } T < \text{LRL} | \text{RL} \leq y_{i0} \leq M) \leq \alpha$$

This rule defines RL as a value such that as long as the potency measurement at time zero is above this value and below the upper limit M , the probability of failure will not be more than α . The RLs defined in the conditional and the unconditional rules both control the chance of failure at a given α . However, the two rules have different meanings. The RLs based on the conditional rule, denoted by RL(C), keeps the chance of failure for *each* released batch under α , whereas the RLs based on the unconditional rule, denoted by RL(UC), keeps the chance of failure for *all* future released batches under α .

Release Limits Based on the Conditional Rule

Let the observed potency follows the following model with parameters b , σ_a^2 , σ_β^2 , and σ_e^2 , then for k th replicate from i th batch at time t ,

$$y_{ik} = \mu + \alpha_i + (b + \beta_i)t + e_{ik}, \\ t = 0, t_1, t_2, \dots, T, \quad k = 1, 2, \dots, n_i$$

where y_{ik} is the assayed potency of k th replicate, i th batch, at time t ; μ is the mean batch potency at time zero; α_i is the random batch effect on potency at time zero, $\alpha_i \sim N(0, \sigma_a^2)$; b is the average potency change rate per time unit; β_i is the random batch effect on slope, $\beta_i \sim N(0, \sigma_\beta^2)$; t is the

sampling time; e_{itk} is the total random assay error, $e_{itk} \sim N(0, \sigma_e^2)$; α_i and β_i are independent; and (α_i, β_i) and e_{itk} are independent.

The same model can also be expressed as

$$y_{itk} = \mu_i + b_i t + e_{itk}, \quad t = 0, t_1, t_2, \dots, T, \quad k = 1, 2, \dots, n_i$$

where $\mu_i = \mu + \alpha_i$; and $b_i = b + \beta_i$.

Having observed the sample average of n_0 assay values at time zero, y_{i0} , the future observed assay value at time t follows a new model below,

$$y_{itk}^* = y_{i0} + \eta_i + (b + \beta)_i t + e_{itk}, \quad t = t_1, t_2, \dots, T, \quad k = 1, \dots, n_i \quad (1)$$

where $\eta_i \sim N(0, \sigma_e^2/n_0)$.

Let

$$\begin{aligned} Y_{it} &= \begin{pmatrix} y_{it1} \\ y_{it2} \\ \vdots \\ y_{itn_i} \end{pmatrix}, & X_{1t} &= \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \\ X_{2t} &= \begin{pmatrix} t \\ t \\ \vdots \\ t \end{pmatrix}, & E_{it} &= \begin{pmatrix} e_{it1} \\ e_{it2} \\ \vdots \\ e_{itn_i} \end{pmatrix}, \\ Y_i &= \begin{pmatrix} Y_{i0} \\ Y_{i1} \\ \vdots \\ Y_{iT} \end{pmatrix}, & X_1 &= \begin{pmatrix} X_{10} \\ X_{11} \\ \vdots \\ X_{1T} \end{pmatrix}, \\ X_2 &= \begin{pmatrix} X_{20} \\ X_{21} \\ \vdots \\ X_{2T} \end{pmatrix}, & E_i &= \begin{pmatrix} E_{i0} \\ E_{i1} \\ \vdots \\ E_{iT} \end{pmatrix} \end{aligned}$$

also

$$\begin{aligned} Y_i^* &= \begin{pmatrix} Y_{i1}^* \\ Y_{i2}^* \\ \vdots \\ Y_{iT}^* \end{pmatrix}, & X_1^* &= \begin{pmatrix} X_{11}^* \\ X_{12}^* \\ \vdots \\ X_{1T}^* \end{pmatrix}, \\ X_2^* &= \begin{pmatrix} X_{21}^* \\ X_{22}^* \\ \vdots \\ X_{2T}^* \end{pmatrix}, & E_i^* &= \begin{pmatrix} E_{i1}^* \\ E_{i2}^* \\ \vdots \\ E_{iT}^* \end{pmatrix} \end{aligned}$$

The model 1 in vector form becomes

$$Y_i^* = (X_1^* X_2^*) \begin{pmatrix} y_{i0} \\ b \end{pmatrix} + X_1^* \eta_i + X_2^* \beta_i + E_i^*$$

Now the point estimate of potency and its variance at time T under the new model can be calculated as

$$\begin{aligned} \hat{y}_{iT}^* &= y_{i0} + \hat{b}_i T \\ \text{Var}(\hat{y}_{iT}^*) &= T^2 \text{Var}(\hat{b}_i) = T^2 (X_2^{*T} X_2^*)^{-2} X_2^{*T} \Sigma X_2^* \\ &= T^2 \left[\sigma_\beta^2 + (X_2^{*T} X_2^*)^{-1} \sigma_e^2 \left(1 + \frac{N}{n_0} \right) \right] \end{aligned}$$

where $N = n_{i1} + n_{i2} + \dots + n_{iT}$.

To control the probability of failure for this batch, it is sufficient to control the probability that the 95% lower CI for the mean potency at time t will be below LRL under certain level (say α). It is clear that this probability is a decreasing function in y_{i0} .

$$P(\hat{y}_{iT} - z_{0.05} \sqrt{\text{Var}(\hat{y}_{iT})} < \text{LRL} | y_{i0})$$

RL is the minimum value of y_{i0} such as

$$P(\hat{y}_{iT} - z_{0.05} \sqrt{\text{Var}(\hat{y}_{iT})} < \text{LRL} | y_{i0}) \leq \alpha$$

In fact, the solution for RL can be expressed as

$$\begin{aligned} \text{RL}(C) &= \text{LRL} + bT + (z_\alpha + z_{0.05})T \\ &\quad \times \sqrt{\sigma_\beta^2 + (X_2^{*T} X_2^*)^{-1} \sigma_e^2 \left(1 + \frac{N}{n_0} \right)} \end{aligned}$$

when samples are only drawn at time zero and the end of the study.

The above expression is reduced to

$$\text{RL}(C) = \text{LRL} + bT + (z_\alpha + z_{0.05}) \times \sqrt{\sigma_\beta^2 T^2 + \sigma_e^2 \left(\frac{1}{nT} + \frac{1}{n_0} \right)}$$

Notice the similarity between the above expression and the expression in Case II of Allen et al. method. The $\text{RL}(C)$ is batch-specific. For each batch, y_{i0} will be obtained and compared to $\text{RL}(C)$ to decide whether the batch can be released. For every batch released in such a way, the probability of failure at time T is controlled. This method does not utilize any prior information about distribution of time-zero potency among batches.

Release Limits Based on the Unconditional Rule

Now suppose the distribution of time-zero potency among batches is known. Recall model

$$y_{ik} = \mu_i + b_i t + e_{ik}, \quad t = 0, t_1, t_2, \dots, T, \quad k = 1, 2, \dots, n_i$$

Use the same vector notation as above and let $X = (X_1 X_2)$. Then the lower 95% CI for the true potency at time T is Lower C for true potency at time

$$T = \hat{y}_{iT} - z_{0.05} \sqrt{\text{var}(\hat{y}_{iT})}$$

Because the lower 95% CI for the true potency at time T follows a normal distribution as can be seen from the above expression, the probability of failure at time T shown below can therefore be calculated.

$$P(95\% \text{ lower CI of true potency at time } T < \text{LRL})$$

Because the estimated time-zero potency is negatively correlated with the predicted potency at time T , the probability of failure at time T is negatively associated with the RLs at time zero. The basic idea is that for those batches which pass the RLs, the probability of failure at time T will be under an allowable value. The RL(UC) can then be solved from the following inequality, and $\bar{y}_{i0}$ is the sample mean assay value,

$$P(95\% \text{ lower CI of true potency at time } T < \text{LRL} | \text{RL(UC)} \leq y_{i0} \leq M) \leq \alpha \quad (2)$$

or

$$P\left(\hat{y}_{iT} - z_{0.05} \sqrt{\text{var}(\hat{y}_{iT})} < \text{LRL} | \text{RL(UC)} \leq y_{i0} \leq M\right) = \frac{P\left(\hat{y}_{iT} - z_{0.05} \sqrt{\text{var}(\hat{y}_{iT})} < \text{LRL} \cap \text{RL(UC)} \leq y_{i0} \leq M\right)}{P(\text{RL(UC)} \leq y_{i0} \leq M)} \leq \alpha$$

Under normality, RL(UC) can then be solved using the joint distribution of y_{i0} and the predicted y_{iT} as given below.

$$\begin{pmatrix} y_{i0} \\ \hat{y}_{iT} - z_{0.05} \sqrt{\text{var}(\hat{y}_{iT})} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu + bT - z_{0.05} \sqrt{\text{var}(\hat{y}_{iT})} \end{pmatrix}, \text{cov}\begin{pmatrix} y_{i0} \\ \hat{y}_{iT} \end{pmatrix}\right)$$

and

$$\text{var}(y_{i0}) = \sigma_\alpha^2 + \frac{\sigma_e^2}{n_0}$$

$$\text{cov}(y_{i0}, \hat{y}_{iT}) = \frac{1}{n_0} (1'_{n_0}, 0'_{n_1}, \dots, 0'_{n_T}) \sum X(X'X)^{-1} \begin{pmatrix} 1 \\ T \end{pmatrix}$$

Then the minimum value for RL(UC) which satisfies the inequality 2 can be easily found using a numerical search routine.

Expected Loss Function Approach

RL(UC) keeps probability of failure for all future batches under α . A smaller α means a fewer number of batches fail LRL at time T , but a smaller α will also result in a larger number of batches fail to be released. A failed batch at time zero or time T will both be a financial loss to the manufacturer. Therefore, an alternative way to determine RL(UC) is to find one which minimizes the total expected loss provided that the financial loss due to failing a batch at time zero and time T are known.

Suppose

Financial loss due to failing a batch at time 0 = L_0 .

Financial loss due to failing a batch at time T = L_T .

Then

Expected Loss = $L_0 P(\text{a batch failed at time 0}) + L_T P(\text{a batch failed at time } T)$

$$\text{Expected Loss} = L_0 P(y_{i0} < \text{RL(UC)}) + L_T P(\text{RL(UC)} \leq y_{i0} \leq M \cap \hat{y}_{iT} < \text{LRL})$$

RL(UC) that minimizes the above function can then be found using a numerical search routine.

All three methods described in this section require prior knowledge about certain parameters. For the conditional method, they are b , σ_β^2 , and σ_e^2 , and μ , b , σ_α^2 , σ_β^2 , and σ_e^2 for the unconditional method. In practice, these knowledge can be obtained from different sources such as assay validation studies, previous stability studies, and published results.

CONCLUSION

The Allen et al. method is very appealing to practitioners for its simplicity and intuitiveness. However, their method was not established on a justifiable statistical argument. Having this drawback, the property of their RL is not clear. Nonetheless, the performance of their method may still be satisfactory in practice.

The three methods proposed by Wei were derived from a probability argument of failure rate control. The conditional method is the easiest one to calculate and requires the least information. It should be used in the situation where no prior knowledge about the distribution of batch potency at time zero. The unconditional method is an improvement over the conditional method at the expense of availability of prior knowledge about the distribution of batch potency at time 0. However, these two methods both focus on controlling failure rate of batch at time T , they do not account for cost due to batch rejection at time zero. The loss function method minimizes the expected loss due to batch failure at time 0 and T when cost ratio is available. In practice, which method to use will depend on the availability

of the required information. The estimated RL should be treated as an important component in the whole process of RL determination along with other deciding factors such as production capability and other practical issues.

REFERENCES

1. Allen, P.V.; Dukes, G.R.; Gerger, M.E. Determination of release limits: A general methodology. *Pharm. Res.* **1991**, *8* (0).
2. Wei, G.C.G. Simple methods for determination of the release limits for drug products. *J. Biopharm. Stat.* **1998**, *8* (1), 103–114.
3. Shao, J.; Chow, S.-C. Constructing release targets for drug products: A Bayesian decision theory approach. *Appl. Stat.* **1991**, *40* (3).

Reliability

Valentin Rousson and Theo Gasser

University of Zurich, Zurich, Switzerland

Abstract

This entry reviews some issues of reliability for continuous measurements with a focus on three kinds of reliability: intrarater, interrater, and test-retest reliability.

INTRODUCTION

Classical statistical analysis usually assumes that data are measured without error. Yet even in exact sciences such as physics, data are recorded with errors, and these are even more important when data are the result of ratings. One usually distinguishes between two kinds of measurement error: systematic error (or bias) and random error (or variability). The first one arises when the “measuring instrument,” such as a rater, gives values that are systematically too large, or systematically too low, compared to the “truth.” These errors are especially annoying when data are obtained from different sources (such as different raters) with a varying amount of systematic error. Random errors are attributable to the fact that an instrument does not exactly give twice the same value when twice measuring the same thing. Reasons for this may be slight changes in the external conditions, for example. Random errors lead to underestimating correlations between variables, and more generally make it more difficult to obtain significant results. Thus, before introducing new techniques or rating scales, it is advisable to perform an analysis of measurement error, and such an analysis is called a *reliability analysis*.

Reliability is sometimes confounded with validity. Whereas reliability is about measurement error, validity refers to the appropriateness, meaningfulness, and usefulness of what is being measured. For example, an IQ test is said to be reliable if it twice provides similar results when applied twice to the same subject. On the other hand, an IQ test is said to be valid if it measures what is meant by intelligence. Validity is especially important in social sciences where the variables of interest, such as intelligence or satisfaction, may not be directly measured (they are referred to as “latent” variables). In natural sciences, one usually measures more “concrete” variables, and the problem of validity is less important. Interestingly enough, both reliability and validity can be improved by the building of “components,” i.e., by combining several variables together (as performed in an IQ test). But while validity may be improved by considering relevant variables that are

not necessarily correlated with each other, the reliability of a component will be generally higher when the combined variables are correlated, as shown below.

Another concept related to reliability is agreement. For categorical data, the agreement between two measuring instruments is usually defined as the percentage of times that the same value is achieved. For continuous data, one is interested in differences between measurements provided by two instruments (in the actual scale of the data), and hence agreement measures “accuracy.” By way of contrast, reliability informs us about the ability of the instruments to differentiate among the entities measured. Thus, reliability will depend on the population of entities considered, much like a correlation coefficient, such that the more heterogeneous the population, the higher the reliability. Agreement and reliability might therefore provide a very different picture of the situation.

In this entry, we review some issues of reliability for continuous measurements. We shall focus on three kinds of reliability: intrarater, interrater, and test-retest reliability. The standard method for assessing reliability in this context is intraclass correlation. But we shall see that different kinds of intraclass correlation coefficients should be used depending on the situation. The key difference will be the treatment of the systematic error that might or might not be a deficiency of the measuring procedure.

SOME CONSEQUENCES OF MEASUREMENT ERROR

Before discussing in detail how to assess reliability, let us illustrate some effects of measurement error on scientific research by first considering the classical two-sample problem. Let X and Y be normally distributed variables with expectations μ_X and μ_Y and with common variance σ^2 . The test statistic is given by $t = (n_X n_Y / (n_X + n_Y))^{1/2} (\bar{X} - \bar{Y}) / s$, where $\bar{X}$ and $\bar{Y}$ are the sample means of data in both groups, s^2 is an estimate of σ^2 , and n_X and n_Y denote sample sizes. If data are recorded with errors, they are actually sampled from $\tilde{X} = X + \epsilon_X$ and $\tilde{Y} = Y + \epsilon_Y$, where ϵ_X and

ϵ_r denote measurement errors. In what follows, we shall assume ϵ_X and ϵ_r to be normally distributed with expectation v_X and v_r , and with common variance τ^2 , and to be independent from each other and from X and Y . Thus variables $\tilde{X}$ and $\tilde{Y}$ are normally distributed with expectations $\mu_X + v_X$ and $\mu_Y + v_r$ and with common variance $\sigma^2 + \tau^2$.

Quantities v_X and v_r represent the amount of systematic error. For example, if v_X is larger than 0, data in the first group will be on average too large. The numerator of the test statistic is hence an estimate of $\mu_X - \mu_Y + v_X - v_r$, instead of $\mu_X - \mu_Y$. Thus, systematic errors are annoying only if v_X differs from v_r (i.e., if we have a “differential bias”). In this case, the numerator of t will be too large (if $v_X > v_r$), or too low (if $v_X < v_r$), and this will result in an invalid test.

On the other hand, the variance τ^2 of measurement errors represents the amount of random error in the data (which is taken equal in both groups in this example). Thus, the denominator of the test statistic is an estimate of $\sigma^2 + \tau^2$, instead of σ^2 , such that the larger τ^2 , the smaller the test statistic. Data recorded with random errors lead therefore to fewer significant results.

In regression, random errors do not only lead to a loss of power for the statistical tests, but also to a systematic underestimation of the strength of the association between the variables. To see this, let us consider data sampled from a pair of variables (X, Y) with marginal variances τ_X^2 and τ_Y^2 and with covariance σ_{XY} . The sample correlation $r_{XY} = s_{XY}/(s_X s_Y)$ is an estimate of the true correlation $\sigma_{XY}/(\sigma_X \sigma_Y)$, where s_X^2 and s_Y^2 denote sample variances and s_{XY} sample covariance. Again, in case of measurement errors, data are actually sampled from $(\tilde{X}, \tilde{Y}) = (X + \epsilon_X, Y + \epsilon_r)$ instead of (X, Y) . Let us assume that measurement errors ϵ_X and ϵ_r have expectation 0 and variances τ_X^2 and τ_r^2 , respectively. Then the covariance of $\tilde{X}$ and $\tilde{Y}$ is given by

$$\begin{aligned} \text{Cov}(\tilde{X}, \tilde{Y}) &= \text{Cov}(X, Y) + \text{Cov}(X, \epsilon_r) + \text{Cov}(\epsilon_X, Y) \\ &\quad + \text{Cov}(\epsilon_X, \epsilon_r) \end{aligned}$$

If measurement errors are independent from each other, as well as from X and Y (which is a reasonable assumption), the covariance of $\tilde{X}$ and $\tilde{Y}$ is equal to the covariance of X and Y , such that the numerator of r_{XY} is still an unbiased estimate of σ_{XY} . On the other hand, sample variances estimate $\tau_X^2 + \tau_X^2$ and $\sigma_Y^2 + \tau_r^2$, and have a positive bias as estimates of τ_X^2 and τ_r^2 . As a consequence, the denominator of r_{XY} will be too large, and the sample correlation will underestimate the true correlation, even for large sample size. Such underestimated correlations are referred to as “attenuated correlations.” This phenomenon is partly responsible for the low correlations sometimes found in the social sciences, when larger ones are expected.

For similar reasons, estimates of the slope of a regression line obtained from data where the explanatory variable is recorded with error will be systematically too small. This phenomenon is sometimes referred to as the “regression dilution bias.” For an interesting example, the reader

is referred to MacMahon et al.^[1] In this study, the authors consider the association between diastolic blood pressure (DBP) with the incidence of stroke and of coronary heart disease. They note that previous analyses were carried out with DBP measured at “baseline,” which is subject to a lot of random fluctuations, leading to underestimating the true association above. They consider instead a variable called “usual DBP,” defined as the average blood pressure of an individual over several years, and they obtain slopes about 60% steeper.

DIFFERENT ASPECTS OF RELIABILITY

A reliability analysis uses repeated measurements of a variable Y . In the most common and simplest case, a random sample of n entities with each d repeated measurements is considered (this is the case of “balanced data”). In what follows, we shall denote by Y_{ij} the j th measurement made on the i th entity (for $i = 1, \dots, n$ and $j = 1, \dots, d$). When measurements are the result of ratings made on subjects performing a given task, three aspects of reliability are usually of interest: intrarater, interrater, and test–retest reliability. For intrarater data, the d measurements refer to d ratings of a same rater, while for interrater data, they are ratings of d different raters. Thus, in the former case, only one rater is needed and Y_{ij} denotes the j th measurement made by this rater on the i th subject. In the latter case, d raters are needed and Y_{ij} denotes the measurement made by the j th rater on the i th subject. Finally, in a test–retest situation, each subject should perform the task d times and Y_{ij} denotes the result of the j th trial made by the i th subject. The same three concepts may be found in other kinds of reliability studies, even if the terminology may sometimes be different. Let us, as an example, consider measurements of laboratory parameters from blood sample, performed via some instruments. The equivalent of intrarater reliability will then be the analysis of repeated measurements made on a same blood sample by a given instrument and the equivalent of interrater reliability will be the analysis of measurements by different instruments on this sample. Finally, test–retest reliability will be the analysis of measurements made on different samples from a same subject (where the subject could be measured once a week, for example). Note that when the variable to be measured varies fast in time, intrarater and interrater reliability are sometimes confounded with test–retest reliability (e.g., this may arise for blood pressure).

We shall illustrate our point with a real example. Largo et al.^[2] introduced a test battery to study the development of neuromotor functions in children and adolescents. The test battery consists of various finemotor tasks for which the time needed to complete them is recorded with a watch (or “rated” because time has, in this case, a clear subjective component). Evidently, a good reliability of such a test battery is a prerequisite for its clinical and scientific

usefulness, and hence data were collected to assess intrarater, interrater, and test–retest reliability. In what follows, intra- and interrater reliability are derived from a sample of $n = 30$ children. Each child was rated once by a first rater (in what follows, Rater 1), and twice by a second rater (Rater 2) from videotape. The two ratings of Rater 2 were performed within 8 weeks. Test–retest reliability is derived from a sample of $n = 22$ children, who performed the tasks twice 1 week apart. For all three kinds of reliability, hence we have $d = 2$.

An analysis of reliability starts with scatter plots (and other exploratory statistics) to identify peculiar patterns. As stated in the “Introduction” section, we distinguish between two kinds of discrepancy among repeated measurements. Systematic error refers to those differences that are systematically in the same direction. Random errors refer to differences that do not lie predominantly in the same direction. Fig. 1 shows intrarater, interrater, and test–retest data for one of the tasks of the test battery. While intrarater data exhibit practically no systematic error and low random error (all but one point are pretty well aligned on the identity line), there is a clear systematic error for both interrater and test–retest data. For interrater data, Rater 2 rated systematically lower than Rater 1. This is undoubtedly a failure in reliability because it indicates a lack of training or standardization. For test–retest data, the second attempt of a child was systematically better than

the first one. However, this systematic error is attributed to a learning effect, which is a natural phenomenon and not a defect of the test battery. This example shows that the systematic error should be treated differently according to the situation. This will be the subject of the section “Assessing Intrarater, Interrater, and Test–Retest Reliability.”

CONCORDANCE CORRELATION

From a bivariate sample of continuous repeated measurements as described in the previous section, reliability is often evaluated by using either Pearson (or Spearman) correlation, a paired t -test, or tests on the intercept and the slope of the least-squares regression line, to check whether they are equal to 0 and 1, respectively. Figs. 1–3 in Lin^[3] accurately illustrate why these three approaches are not appropriate to assess reliability. A perfect reliability is achieved if reliability data, as those in Fig. 1, lie all on the identity line. Thus, a high correlation does not always imply a high reliability. For example, a Pearson correlation of 1 just means that the data are perfectly aligned, but not necessarily on the identity line. In particular, it is not possible to detect a systematic error using correlation. On the other hand, the null hypothesis of a paired t -test and of tests on the coefficients of the regression line do not involve random errors at all.

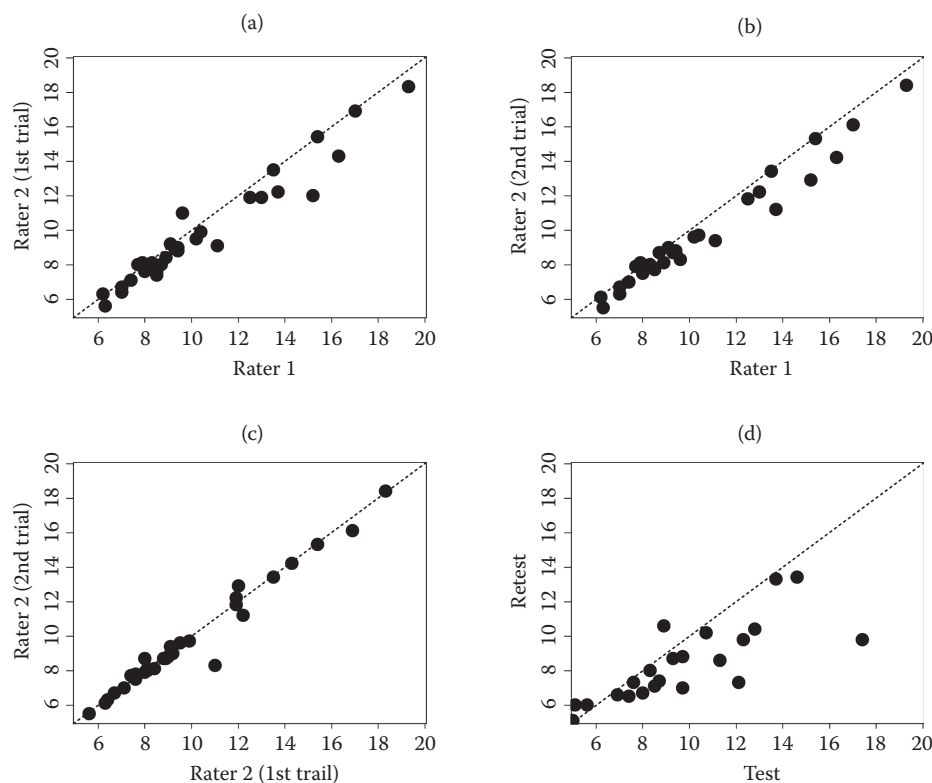


Fig. 1 Scatter plots of reliability data for the timed performance (in seconds) of some neuromotor task: (a and b) interrater data, (c) intrarater data, (d) test–retest data. Identity line is plotted as reference

Let Y_1 and Y_2 be two repeated measurements of a continuous variable Y . As a perfect reliability is achieved if and only if $E[(Y_1 - Y_2)^2] = 0$, Lin^[3] proposes to use this quantity for assessing reliability. If μ_1 and μ_2 denote means, and σ_1^2 and σ_2^2 denote variances of Y_1 and Y_2 , respectively, and if ρ denotes the correlation between Y_1 and Y_2 , we have

$$E[(Y_1 - Y_2)^2] = (\mu_1 - \mu_2)^2 + (\sigma_1^2 - \sigma_2^2) + 2(1 - \rho)\sigma_1\sigma_2$$

Thus a perfect reliability is attained if and only if $\mu_1 = \mu_2$, $\sigma_1 = \sigma_2$, and $\rho = 1$. To obtain a coefficient with values between -1 and 1 , Lin³ defines the *concordant correlation coefficient* as

$$\rho_{con} = 1 - \frac{E[(Y_1 - Y_2)^2]}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}$$

Equivalently, we have $\rho_{con} = \rho C_b$, where $C_b = 2/(v + 1/v + u^2)$, $v = \sigma_1/\sigma_2$, and $u = (\mu_1 - \mu_2)/(\sigma_1\sigma_2)^{1/2}$. The factor C_b takes its values between 0 and 1 , and is said to be a bias correction for ρ (when the issue is evaluating reliability) because $|\rho_{con}| \leq |\rho|$. Concordance correlation can be easily estimated by plugging in sample versions of μ_1 , μ_2 , σ_1^2 , σ_2^2 , and ρ .^[3]

Concordance correlation for assessing reliability is an attractive concept that can be generalized to more than two repeated measurements. However, it does not take into account the “random source” of the d measurements. For example, in interrater reliability, concordance correlation will inform us about the reliability of the d raters in the sample. But the result cannot be generalized to other raters. If the raters in the sample are thought to be randomly selected from a larger population of raters, one has to use the concept of intraclass correlation. This is the topic of the next two sections.

INTRACLAS CORRELATION

Let us consider a sample of repeated measurements Y_{ij} , for $i = 1, \dots, n$ and $j = 1, \dots, d$, as in the section “Different Aspects of Reliability.” We may distinguish two situations. If j indexes a particular criterion, such as time or rater, the ordering of the d measurements $Y_{i1}, \dots, Y_{id}$ made on an entity i is not arbitrary and one cannot exchange them. For example, the j th measurement may refer to the result of the j th trial (for test–retest reliability), or to the rating of the j th rater (for interrater reliability). On the other hand, if the ordering of j within i is arbitrary, it is not meaningful to refer to the j th measurement of an entity because the d measurements are interchangeable.

A historical explanation for $d = 2$ can be found in the work of Fisher,^[4] where a (fictive) example with measurements made on pairs of brothers is considered. The first situation above arises when a pair of brothers can be systematically classified into two categories (or two classes),

such as the younger and the elder brother. If one is interested to calculate the correlation between the younger brother and the elder brother, one may use the standard product–moment correlation coefficient, which in this case may be referred to as *interclass correlation*. But if no such classifying criterion is available, for example if we consider twins, the ordering of the two brothers is arbitrary and this is the second situation above. In this case, one cannot define something like a correlation between “this” brother and “that” brother. One can only refer to “the correlation between two brothers,” and this is an example of what is called *intraclass correlation*.

When analyzing repeated measurements of the first kind, i.e., when one may distinguish among the d measurements, a useful model is given by

$$Y_{ij} = \mu + s_i + r_j + e_{ij} \quad (1)$$

where μ is a fixed parameter, and where s_i , r_j , and e_{ij} are mutually independent random effects which are normally distributed with mean 0 and variances σ_s^2 , σ_r^2 , and σ_e^2 , respectively. Note that model 1 may be overly simplistic in some cases. In particular, it assumes that variances σ_r^2 and σ_e^2 do not depend on the size of measurements. In some cases, the assumptions of the model are better fulfilled for transformed data than for raw data.

Let us first consider the (interclass) correlation between two *specific* measurements j_1 and j_2 . Under model 1, this correlation does not depend on the particular values of j_1 and j_2 and is given by

$$\rho_{ic1} = \frac{\text{Cov}(Y_{ij1}, Y_{ij2})}{\sqrt{\text{Var}(Y_{ij1})\text{Var}(Y_{ij2})}} = \frac{\sigma_s^2}{\sqrt{(\sigma_s^2 + \sigma_e^2)(\sigma_s^2 + \sigma_e^2)}} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}$$

Note that the denominator of ρ_{ic1} is equal to the variance of measurements Y_{ij} conditionally on j (and hence does not involve σ_r^2), such that ρ_{ic1} may also be referred to as the correlation between Y_{ij1} and Y_{ij2} conditionally on j_1 and j_2 (with $j_1 \neq j_2$).

If one wishes to define the (intraclass) correlation between two *arbitrary* measurements in this context, one should consider the correlation between Y_{ij1} and Y_{ij2} unconditionally on j_1 and j_2 (with $j_1 \neq j_2$). One will obtain

$$\rho_{ic2} = \frac{\text{Cov}(Y_{ij1}, Y_{ij2})}{\text{Var}(Y_{ij})} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_r^2 + \sigma_e^2}$$

Here the denominator is equal to the variance of measurements Y_{ij} unconditionally on j such that σ_r^2 is involved and $\rho_{ic2} \leq \rho_{ic1}$.

When the ordering of the d measurements is arbitrary, the term r_j in model 1 does not make sense. Thus, one has to assume a simpler model, for example

$$Y_{ij} = \mu + s_i' + e_{ij}' \quad (2)$$

where μ is a fixed parameter, and where s'_i and e'_{ij} are mutually independent random effects which are normally distributed with mean 0 and variances σ_s^2 and σ_e^2 , respectively. In this context, it is not possible to define the correlation between two specific measurements, while the correlation between two arbitrary measurements is given by

$$\rho_{ic3} = \frac{\text{Cov}(Y_{ij1}, Y_{ij2})}{\text{Var}(Y_{ij})} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}$$

At first sight, one may think that models 1 and 2 are equivalent by letting $s'_i = s_i$ and $e'_{ij} = r_j + e_{ij}$. But one would have then $\text{Cov}(e'_{i1j1}, e'_{i2j2}) = \sigma_r^2/d$ for $i_1 \neq i_2$, such that the independency assumptions of model 2 would no longer hold as soon as $\sigma_r^2 > 0$.

All three coefficients ρ_{ic1} , ρ_{ic2} , and ρ_{ic3} are sometimes referred to as intraclass correlation in the literature, but ρ_{ic1} should better be referred to as interclass correlation, as explained above.

ASSESSING INTRATER, INTERRATER, AND TEST-RETEST RELIABILITY

The ordering of repeated measurements for interrater data and test-retest data is clearly not arbitrary and one should use model 1. The term s_i represents the subject effect, whereas r_j is the systematic error (a “rater effect” for interrater data, or a “trial effect” for test-retest data) and e_{ij} is the random error. Thus, the amount of systematic error and random error may be quantified by σ_r^2 and σ_e^2 , respectively. From the definition of coefficients ρ_{ic1} and ρ_{ic2} given in the previous section, we see that ρ_{ic2} takes into account the systematic error, whereas ρ_{ic1} does not.

Coefficient ρ_{ic1} is the “correlation between two specific raters” for interrater data and the “correlation between two specific (for example two successive) trials” for test-retest data. In contrast, ρ_{ic2} should simply be referred to as the “correlation between two raters” or as the “correlation between two trials” (where raters and trials might change from one entity to another). Following Rousson et al.,^[5] we recommend the use of ρ_{ic2} to assess interrater reliability, because a systematic error clearly indicates a failure in reliability and should be penalized. For test-retest data, on the other hand, a systematic error is often attributable to a learning effect (improvement in the retest situation) or to a fatigue effect (worsening in the retest situation), which are not directly related to reliability. Thus, we advocate the use of ρ_{ic1} to assess test-retest reliability.

For intrarater data, it is not always clear whether or not the ordering of repeated measurements is arbitrary. If one suspects a learning effect for the rater, one may consider the ordering in which the ratings were carried out and use model 1. Because such a learning effect should not occur for well-trained raters, and hence indicates a lack of reliability,

one may then use ρ_{ic2} to assess intrarater reliability. On the other hand, if no learning effect is expected, one may use model 2 and ρ_{ic3} . But as the systematic error is, in general, negligible for intrarater data, the choice between models 1 and 2, and coefficients ρ_{ic1} , ρ_{ic2} , and ρ_{ic3} is not crucial.

In some instances, models different from Eq. 1 or Eq. 2 are needed. For interrater data, model 1 implicitly assumes that the d raters are a random sample from a potentially large population of raters. If the d raters are the only raters of interest, one may prefer to consider a fixed rater effect r_j , rather than a random one (such that $r_1, r_2, \dots, r_d$ are the whole population of r_j). In this case, model 1 should be parametrized such that $\sum_j^d = 1$ $r_j = 0$, while the amount of systematic error may be quantified by $\tilde{\sigma}_r^2 = \sum_{j=1}^d r_j^2 / (d-1)$. Then, interrater reliability can be assessed by $\tilde{\rho}_{ic2} = \sigma_s^2 / (\sigma_s^2 + \tilde{\sigma}_r^2 + \sigma_e^2)$. For test-retest data, the distinction between a model with fixed or random r_j is not important because we do not take into account the effect r_j in the definition of reliability.

Finally, one may simultaneously assess several aspects of reliability, such as intrarater and interrater reliability, by using a three-dimensional array of data, where Y_{ijk} denotes the k th rating of the j th rater made on the i th subject (for $i = 1, \dots, n$, $j = 1, \dots, d$ and $k = 1, \dots, m$). In this case, one could consider a model such as

$$Y_{ijk} = \mu + i_i + j_j + k_k + ij_{ij} + ik_{ik} + jk_{jk} + ijk_{ijk} \quad (3)$$

where μ is a fixed parameter, and where i_i , j_j , k_k , ij_{ij} , ik_{ik} , jk_{jk} , and ijk_{ijk} are mutually independent random effects which are normally distributed with mean 0 and variances σ_i^2 , σ_j^2 , σ_k^2 , σ_{ij}^2 , σ_{ik}^2 , σ_{jk}^2 , and σ_{ijk}^2 , respectively. The “total variance” of measurements Y_{ijk} is given by the sum of all these variances. Then, one could assess interrater reliability by the percentage of this total variance which is not attributable to the raters, i.e., by

$$\frac{\sigma_i^2 + \sigma_k^2 + \sigma_{ik}^2}{\sigma_i^2 + \sigma_j^2 + \sigma_k^2 + \sigma_{ij}^2 + \sigma_{ik}^2 + \sigma_{jk}^2 + \sigma_{ijk}^2}$$

Similarly, intrarater reliability can be assessed by

$$\frac{\sigma_i^2 + \sigma_j^2 + \sigma_{ij}^2}{\sigma_i^2 + \sigma_j^2 + \sigma_k^2 + \sigma_{ij}^2 + \sigma_{ik}^2 + \sigma_{jk}^2 + \sigma_{ijk}^2}$$

whereas a “global” coefficient of reliability can be obtained as

$$\frac{\sigma_i^2}{\sigma_i^2 + \sigma_j^2 + \sigma_k^2 + \sigma_{ij}^2 + \sigma_{ik}^2 + \sigma_{jk}^2 + \sigma_{ijk}^2}$$

All these reliability coefficients are defined as ratios of variances and assume values between 0 and 1. Thus, they have the particularity to be always positive (unlike their estimates introduced in the section “Estimation,” which may happen to be negative) and may be interpreted

as “percentage of variance explained” (without being squared). In general, they represent the percentage of variance in the measurements that is due to the subject effect, i.e., which is not attributable to measurement error, such that the higher this percentage, the better the reliability. Simplicity of interpretation makes these coefficients particularly attractive to assess reliability, compared to other approaches.

ESTIMATION

While sample correlation will generally underestimate ρ_{ic2} and ρ_{ic3} (as already mentioned in the section “Concordance Correlation”), it is a valid estimate of ρ_{ic1} (in the case $d = 2$). On the other hand, sample correlation calculated from all $nd(d-1)$ pairs of observation (X_{ij1}, X_{ij2}) (with $i = 1, \dots, n, j_1, j_2 = 1, \dots, d$ and $j_1 \neq j_2$), where each observation X_{ij} appears $2(d-1)$ times, is a valid estimate of ρ_{ic3} . More efficient estimates using the assumptions of the model are obtained by performing an analysis of variance for repeated measurements.^[6] From data Y_{ij} , we calculate the following sums of squares

$$\begin{aligned} SS_s &= d \sum_{i=1}^n (\bar{Y}_i - \bar{Y}_{..})^2 \\ SS_r &= n \sum_{j=1}^d (\bar{Y}_j - \bar{Y}_{..})^2 \\ SS_e &= \sum_{i=1}^n \sum_{j=1}^d (Y_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y}_{..})^2 \end{aligned}$$

where $\bar{Y}_i = \sum_{j=1}^d Y_{ij} / d$, $\bar{Y}_j = \sum_{i=1}^n Y_{ij} / n$ and $\bar{Y}_{..} = \sum_{i=1}^n \sum_{j=1}^d Y_{ij} / (nd)$. Under model 1, quantities $MS_s = SS_s / (n-1)$, $MS_r = SS_r / (d-1)$, and $MS_e = SS_e / (n-1)(d-1)$ have expectations $d\sigma_s^2 + \sigma_e^2$, $n\sigma_r^2$ and σ_e^2 , respectively. Thus, unbiased estimates of σ_s^2 , σ_r^2 and σ_e^2 are obtained as

$$\begin{aligned} \hat{\sigma}_s^2 &= \frac{MS_s - MS_e}{d} \\ \hat{\sigma}_r^2 &= \frac{MS_r - MS_e}{n} \\ \hat{\sigma}_e^2 &= MS_e \end{aligned}$$

Then, an estimate of ρ_{ic1} is given by

$$R_{ic1} = \frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \hat{\sigma}_e^2} = \frac{MS_s - MS_e}{MS_s + (d-1)MS_e}$$

while an estimate of ρ_{ic2} is given by

$$R_{ic2} = \frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \hat{\sigma}_r^2 + \hat{\sigma}_e^2} = \frac{MS_s - MS_e}{MS_s + (d-1)MS_e + (d/n)(MS_r - MS_e)}$$

Under model 1 with fixed effect r_j , R_{ic2} is also an estimate of $\tilde{\rho}_{ic2}$.

For model 2, R_{ic1} and R_{ic2} are similar and both are valid estimates of ρ_{ic3} . A more efficient estimate of ρ_{ic3} was proposed in 1899 by Pearson et al.^[7] and is given by

$$R_{ic3} = \frac{MS_s - MS_w}{MS_s + (d-1)MS_w}$$

where $MS_w = SS_w / (n(d-1))$ and

$$SS_w = \sum_{i=1}^n \sum_{j=1}^d (Y_{ij} - \bar{Y}_i)^2 = SS_r + SS_e$$

Contrary to R_{ic1} and R_{ic2} , estimate R_{ic3} provides exactly the same value when the ordering of repeated measurements is altered for some entities.

Coefficient R_{ic3} is the most widely used estimate of intraclass correlation in the literature. Note, however, that $MS_w = \hat{\sigma}_r^2 + \hat{\sigma}_e^2$ and $(MS_s - MS_w)/d = \hat{\sigma}_s^2 - \hat{\sigma}_r^2/d$. Thus, under model 1, R_{ic3} is an estimate of

$$\frac{\sigma_s^2 - \sigma_r^2/d}{\sigma_s^2 + \sigma_r^2 + \sigma_e^2 - \sigma_r^2/d}$$

and will underestimate both ρ_{ic1} and ρ_{ic2} . If an ordering of repeated measurements can be defined, it is important to sort the data (such that the j th measurement has the same meaning for each entity) and to use R_{ic1} (if one does not wish to penalize σ_r^2) or R_{ic2} (if one wishes to penalize σ_r^2) rather than R_{ic3} , otherwise one will underestimate intraclass correlation.

If one wishes to estimate coefficients for more complicated models, such as model 3, one will face a problem of variance components estimation. For a good overview of the topic, see, e.g., Searle et al.^[8] There the problem of estimation for “unbalanced data” is also considered, for which the method of restricted maximum likelihood can be used.

To approximately satisfy the normality and homoscedasticity assumptions, reliability data of Fig. 1 were log-transformed. Coefficients R_{ic1} , R_{ic2} , and R_{ic3} calculated on such transformed data are given in Table 1. Following the arguments given in the previous section, reliability is estimated to be 0.98 for intrarater data, 0.96 for interrater data and 0.81 for test–retest data.

Table 1 Estimates of Intraclass Correlation for the Data of Fig. 1 (log-Transformed)

Panel	Type of reliability	R_{ic1}	R_{ic2}	R_{ic3}
(a)	Interrater data	0.97	0.96	0.96
(b)	Interrater data	0.98	0.96	0.96
(c)	Intrarater data	0.98	0.98	0.98
(d)	Test–retest data	0.81	0.74	0.73

CONFIDENCE INTERVALS

It is not meaningful to test for a “statistically significant” reliability (because this would just prove that the reliability is larger than 0, which is trivial). To assess how well reliability is estimated, a confidence interval is needed. In practice, the minimum acceptable value for a reliability coefficient is often set to 0.75 or 0.80.^[9] Thus a necessary condition to prove that reliability is acceptable is to obtain at least 0.75 as lower boundary of a 95% confidence interval.^[10]

Exact confidence intervals for ρ_{ic1} and ρ_{ic3} can be derived. Under model 1, $SS_s/(d\sigma_s^2 + \sigma_e^2)$ follows a chi-square distribution with $n-1$ degrees of freedom, while SS_e/σ_e^2 follows a chi-square distribution with $(n-1)(d-1)$ degrees of freedom. Moreover, these two quantities are independent. As a consequence, a confidence interval for ρ_{ic1} at the level $1-\alpha$ may be obtained from

$$1-\alpha = \Pr \left\{ \frac{\frac{MS_s}{MS_e} - F_{1-\alpha/2}}{\frac{MS_s}{MS_e} + (d-1)F_{1-\alpha/2}} \leq \rho_{ic1} \leq \frac{\frac{MS_s}{MS_e} - F_{\alpha/2}}{\frac{MS_s}{MS_e} + (d-1)F_{\alpha/2}} \right\}$$

where F_α denotes the α -quantile of an F -distribution with $n-1$ and $(n-1)(d-1)$ degrees of freedom. Similarly, under model 2, a confidence interval for ρ_{ic3} at the level $1-\alpha$ may be obtained from

$$1-\alpha = \Pr \left\{ \frac{\frac{MS_s}{MS_w} - F_{1-\alpha/2}^*}{\frac{MS_s}{MS_w} + (d-1)F_{1-\alpha/2}^*} \leq \rho_{ic3} \leq \frac{\frac{MS_s}{MS_w} - F_{\alpha/2}^*}{\frac{MS_s}{MS_w} + (d-1)F_{\alpha/2}^*} \right\}$$

where F_α^* denotes the α -quantile of an F -distribution with $n-1$ and $n(d-1)$ degrees of freedom.

Confidence intervals for ρ_{ic2} are much more problematic. The approximate confidence interval for ρ_{ic2} proposed by Fleiss and Shrout¹¹ is often used in practice. But while asymptotically exact when both n and d tend to infinity, this method produces confidence intervals that are too liberal when d is small, and this becomes even worse when n increases. The proposal of Zou and McDermott¹² does not solve this problem and often produces confidence intervals that are not informative, i.e., which have a very low lower bound (close to 0, or even a negative value), and this holds also when the true value of ρ_{ic2} is high and n is large. This phenomenon persists when using an asymptotically exact solution for fixed d and n tending to infinity,^[13,14] and is unavoidable when d is small and the ratio σ_r^2/σ_e^2 is

not negligible.^[15] For interrater reliability, this means that even when the true reliability ρ_{ic2} is large, one will not be able to prove it from a sample with few raters if the ratio “systematic error to random error” is not negligible.

To obtain informative and conservative confidence intervals for interrater reliability, Rousson et al.¹⁵ proposed an approach related to a sensitivity analysis. This approach assumes that the raters are well-trained such that σ_r^2/σ_e^2 is not too large, for example not larger than 1. In that paper, an informative and conservative confidence interval for coefficient $\tilde{\rho}_{ic2}$ is also provided, which uses quantiles of a noncentral F -distribution.

HOW TO IMPROVE RELIABILITY

A low intrarater or interrater reliability may indicate that the raters are not well trained or that the variable to be measured is not well defined or standardized. Thus, a first step to improve reliability is a better training of the raters and the introduction of better standards of measurement. In particular, the rater effect r_j in model 1 should become negligible for well-trained raters. Test-retest reliability is related to fluctuations in the state of a subject and is hence more difficult to improve.

If reliability remains low despite a good standardization, one may use an average of m measurements instead of a single measurement. This will reduce the amount of random error by a factor m . For example, blood pressure is often routinely assessed by an average of two measurements made some minutes apart. For interrater reliability, the use of an average of m ratings made by m different raters will also reduce the amount of systematic error by a factor m . From a sample of d repeated measurements, reliability of an average of m measurements can thus be estimated by

$$R_{ic1,m} = \frac{MS_s - MS_e}{MS_s + (d-m)MS_e/m}$$

$$R_{ic2,m} = \frac{MS_s - MS_e}{MS_s + (d-m)MS_e/m + (d/n)(MS_r - MS_e)/m}$$

$$R_{ic3,m} = \frac{MS_s - MS_w}{MS_s + (d-m)MS_w/m}$$

for test-retest, interrater, and intrarater reliability, respectively. Note that when $m = d$, the first coefficient above is equal to the well-known Cronbach's alpha^[16] for assessing the internal consistency of the d measurements.

In social sciences, it is common to use “component of variables” rather than single variables to obtain fewer and better interpretable information. As a by-product, reliability is also improved. For simplicity, let us consider m variables that are equally reliable and equally correlated with each other. Let us consider the case of zero systematic error, such that the reliability of a single variable is given

by $\sigma_s^2/(\sigma_s^2 + \sigma_e^2)$, and let $\rho \geq 0$ be the correlation between two of these variables. Then, the reliability of the average (or the sum) S_m of the m variables is given by

$$\frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_e^2}{1 + \rho(m-1)}}$$

Thus, reliability of component S_m will be higher than reliability of a single variable whenever $\rho > 0$ and $m > 1$. Moreover, reliability of S_m is an increasing function of ρ and m . In particular, it tends to 1 when m tends to infinity (if $\rho > 0$). For neuromotor data, for example, test–retest reliability of an average of 12 different timed performances (appropriately standardized) is higher than test–retest reliability of a single timed performance.^[5]

MEASURING AGREEMENT

As stated in the “Introduction” section, reliability and agreement are related but nevertheless two different concepts. Bland and Altman^[17] point out the drawbacks of intraclass correlation for evaluating the agreement between two measuring instruments. They note that it depends on the range of measurements, such that a low value of

intraclass correlation may just be the consequence of a low variability between the entities in the sample. Moreover, it is not related to the size of measurement error, which might be clinically tolerable.

To assess the agreement between two measuring instruments (such as a new technique and an established one to see whether the former is accurate enough to replace the latter), Bland and Altman^[17,18] propose to calculate differences $D_i = Y_{i2} - Y_{i1}$ of measurements provided by the two instruments (on $i = 1, \dots, n$ entities). When D_i is approximately normally distributed, about 95% of them will lie in the interval $[\bar{D} - 2S_D, \bar{D} + 2S_D]$, which they call *limits of agreement*, where $\bar{D} = \sum_{i=1}^n D_i/n$ and $S_D^2 = \sum_{i=1}^n (D_i - \bar{D})^2/(n-1)$. This method does not depend on the range of measurements, and limits of agreement indicate whether discrepancies are acceptable in practice based on clinical arguments. Moreover, a plot of these differences D_i vs. average measurements $M_i = (Y_{i1} + Y_{i2})/2$ allows the detection of possible trends in systematic error and/or in random error. If some transformation of the data is needed to satisfy the normality assumption, however, limits of agreement are no longer directly related to the scale of the data and are more difficult to interpret.

Let us see how limits of agreement are related to the concepts introduced in the section “Intraclass Correlation.” Under model 1, quantities $\bar{D}$ and S_D^2 are estimates of $r_2 - r_1$

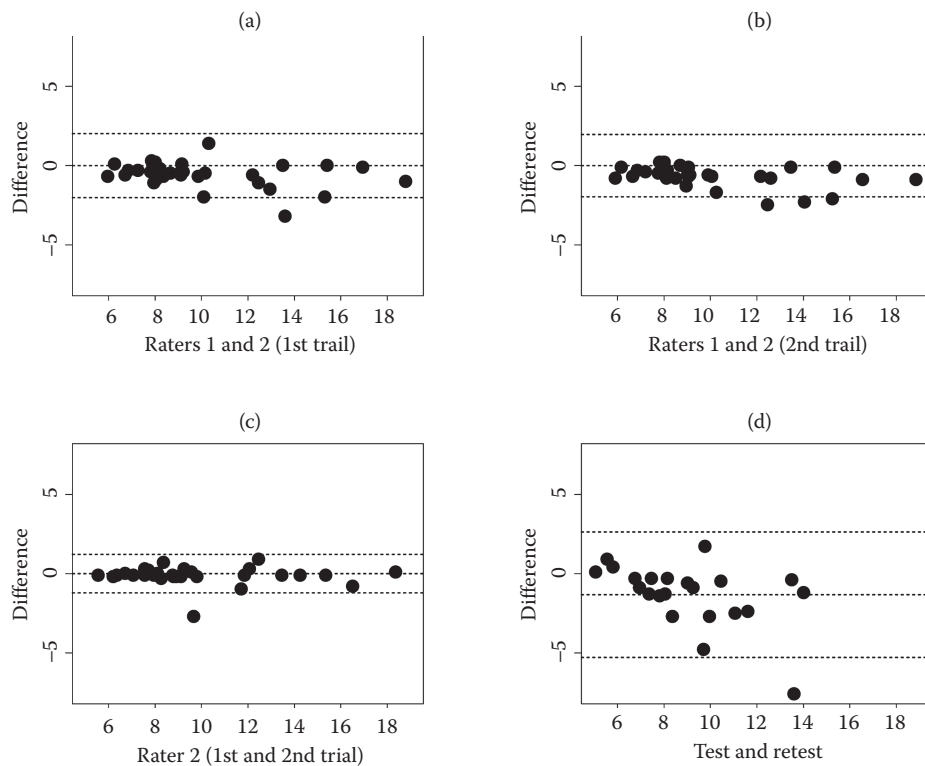


Fig. 2 Limits of agreement for the timed performance (in seconds) of some neuromotor task: (a and b) interrater data, (c) intrarater data, (d) test–retest data. Estimate of interval I_2 was used for (a)–(c), and of interval I_1 for (d). A line indicating the center of the intervals is also plotted. The x-axis refers to the average of two measurements and the y-axis to the differences between two measurements

and $2\sigma_e^2$, respectively (we actually have $S_D^2 = 2\hat{\sigma}_e^2$). Therefore limits of agreement of Bland and Altman estimate the interval

$$I_1 = [r_2 - r_1 - 2\sqrt{2\sigma_e^2}; r_2 - r_1 + 2\sqrt{2\sigma_e^2}]$$

which refers to differences between two specific measurements. On the other hand, if one wishes limits of agreement for differences between two arbitrary measurements, one should consider the interval

$$I_2 = [-2\sqrt{2(\sigma_r^2 + \sigma_e^2)}; 2\sqrt{2(\sigma_r^2 + \sigma_e^2)}]$$

which can be estimated by plugging in estimates $\hat{\sigma}_r^2$ and $\hat{\sigma}_e^2$ introduced in the section “Estimation.” Contrary to I_1 , interval I_2 is centered at 0 and is also defined for $d > 2$ repeated measurements.^[5]

Following the arguments in the section “Assessing Intrarater, Interrater, and Test–Retest Reliability,” limits of agreement I_1 could be used to assess the accuracy of measurements in a test–retest situation, while I_2 could be used for assessing accuracy of interrater or intrarater data. Fig. 2 shows limits of agreement for the data plotted in Fig. 1. The estimate of I_2 is $[-1.2, 1.2]$ for intrarater data and $[-2.0, 2.0]$ for interrater data. Thus, 95% of the times, two ratings of a same rater will not differ more than 1.2 sec. On the other hand, two ratings of two different raters (randomly selected from a larger population of raters) will not differ more than 2 sec. For test–retest data, the estimate of I_1 is $[-5.3, 2.6]$ and hence the second trial can be up to 5.3 sec faster than the first one. From the plot of test–retest data, one may also see some possible negative trend in systematic error (meaning that the learning effect is more important for those children with a bad performance) and some positive trend in random error.

While useful for assessing the accuracy of measurements for a normally distributed variable, limits of agreement are inappropriate for assessing reliability, as soon as several variables with different scales are involved. For example, one cannot compare the reliability of different tools for assessing neuromotor development, or of different psychological tests, using limits of agreement. For this, intraclass correlation is the method of choice because it does not depend on the scale of the data.

CONCLUSION

Reliability can be appropriately assessed using intraclass correlation. This informs us about the percentage of variability in measurements which is not a result of measurement error, and is hence easily interpretable and useful for comparing the reliability of different variables. As intraclass correlation depends on the range of measurements, it is important to use a sample which is representative of the

population of interest. Different definitions and estimates of intraclass correlation should be used depending on the situation. In this entry, we have focused on three estimates denoted by R_{ic1} , R_{ic2} , and R_{ic3} . We have argued that one should use the first one if there is a systematic error which is not related to reliability (as for test–retest data), the second one if there is a systematic error which is related to reliability (as for interrater data), and the third one if no systematic error is expected (as it may happen for intrarater data).

ACKNOWLEDGMENTS

The authors wish to thank Prof. Remo Largo for the permission to use the neuromotor data and the Swiss National Science Foundation for financial support (Project no. 3200-064047.00/1).

REFERENCES

- MacMahon, S.; Peto, R.; Cutler, J.; Collins, R.; Sorlie, P.; Neaton, J.; Abbott, R.; Godwin, J.; Dyer, A.; Stamler, J. Blood pressure, stroke, and coronary heart disease. Part 1. Prolonged differences in blood pressure: Prospective observational studies corrected for the regression dilution bias. *Lancet* **1990**, 335 (8692), 765–774.
- Largo, R.; Caflisch, J.; Hug, F.; Muggli, K.; Sheehy, A.; Gasser, Th.; Molinari, L. Neuromotor development from 5 to 18 Years. Part I: Timed performance. *Dev. Med. Child Neurol.* **2001**, 43 (7), 436–443.
- Lin, L.I.K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **1989**, 45 (1), 255–268.
- Fisher, R.A. *Statistical Methods for Research Workers*; Oliver and Boyd: Edinburgh, 1925.
- Rousson, V.; Gasser, Th.; Seifert, B. Assessing intrater, interrater and test–retest reliability of continuous measurements. *Stat. Med.* **2002**, 21 (22), 3431–3446.
- Bartko, J.J. The intraclass correlation as a measure of reliability. *Psychol. Rep.* **1966**, 19 (1), 3–11.
- Pearson, K.; Lee, A.; Bramley-Moore, L. Mathematical contributions to the theory of evolution—VI. Genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred racehorses. *Philos. Trans. R. Soc. Lond., A.* **1899**, 192, 257–330.
- Searle, S.R.; Casella, G.; McCulloch, C.E. *Variance Components*; Wiley: New York, 1992.
- Burdock, E.L.; Fleiss, J.L.; Hardesty, A.S. A new view of interobserver agreement. *Pers. Psychol.* **1963**, 16 (4), 373–384.
- Lee, J.; Koh, D.; Ong, C.N. Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Comput. Biol. Med.* **1989**, 19 (1), 61–70.
- Fleiss, J.L.; Shrout, P.E. Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika* **1978**, 43 (2), 259–262.
- Zou, K.H.; McDermott, M.P. Higher-moment approaches to approximate interval estimation for a certain intraclass correlation coefficient. *Stat. Med.* **1999**, 18 (15), 2051–2061.

13. Arteaga, C.; Jeyaratnam, S.; Graybill, F.A. Confidence intervals for proportions of total variance in the two-way cross component of variance model. *Commun. Stat., Theory Methods* **1982**, *11* (15), 1643–1982.
14. Gui, R.; Graybill, F.A.; Burdick, R.K.; Ting, N. Confidence intervals on ratios of linear combinations for non-disjoint sets of expected mean squares. *J. Stat. Plan. Inference* **1995**, *48* (2), 215–227.
15. Rousson, V.; Gasser, Seifert, B. Confidence intervals for intraclass correlation in inter-rater reliability. *Scand. J. Stat.* **2003**, *30* (3), 617–624.
16. Cronbach, L.J. Coefficient alpha and the internal structure of tests. *Psychometrika* **1951**, *16* (3), 234–297.
17. Bland, J.M.; Altman, D.G. A note on the use of the intra-class correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput. Biol. Med.* **1990**, *20* (5), 337–340.
18. Bland, J.M.; Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1986**, *1* (8476), 307–310.

Repeated Measurement Designs: Missing Values

Keumhee Carriere Chough

University of Alberta, Edmonton, Alberta, Canada

Yuanyuan Liang

Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, San Antonio, Texas, U.S.A.

Rong Huang

Statistics Canada, Ottawa, Ontario, Canada

Xiaoming Sheng

University of Utah, Salt Lake City, Utah, U.S.A.

Abstract

This entry reviews and summarizes various strategies for dealing with missing data problems in repeated measurement designs that have been suggested to date.

INTRODUCTION

Repeated measurement designs are popular in clinical trials for studies comparing the efficacy of several non-curative treatments because they can effectively remove within-subject variances, by taking repeated measurements from each subject with different treatments. However, by the very nature of taking repeated measurements, they also suffer from the common problem of missing values on one or more occasions. Complete data sets are typically unavailable for various reasons.

In repeated measurement designs, the missing data problem can be viewed as a general problem of missing observations in multivariate data. Analysis issues related to the bivariate or multivariate normal distribution with missing observations has been quite extensively discussed in the literature.

In this entry, we review and summarize various strategies for dealing with missing data problems in repeated measurement designs that have been suggested to date. Our discussion is limited to data with an approximately multivariate normal distribution where the missing data occur on the outcome variable. We first discuss various forms of missing data mechanisms and analytic strategies for dealing with missing data problems. Then, we focus on strategies for use with incomplete repeated measures data in small samples. We also review another approach that utilizes proxy information in lieu of dealing with missing values in the traditional manner.

MISSING DATA MECHANISM

When taking repeated observations,^[1–4] there are various reasons for having missing values. In most studies, unobserved data are treated as “missing,” in the sense that there are true underlying values that could have been observed but were not. However, depending on how the data have become missing, treating unobserved data simply as missing can lead to a serious estimation bias.^[5–7] This section reviews several possible types of missing data, limiting to the case where the missing data occur on the outcome variable.

The most appropriate way to handle missing or incomplete data depends on how the data points became missing. Little and Rubin^[8] define three unique types of missing data mechanisms, distinguishing among the different situations according to how the data became missing and incomplete. They are divided into three different situations. The data can be missing completely at random (MCAR), missing at random (MAR), or there can be nonignorable missing data (NMD).

The MCAR situation assumes that the reasons why the data are missing are completely unrelated to the data (both response and covariates). That is, cases with complete data are indistinguishable from cases with incomplete data. Observed units are random samples of the sampled units. Heitjan^[9] provides an example of MCAR data where a research associate shuffles raw data sheets and arbitrarily discards some of the sheets. The discarded

sheets have no particular reason to be discarded and completely random and independent of what could have been observed and the associated covariates. Graham et al.^[10] illustrate the use of planned missing data patterns of this type. In such situations, data analyses based on complete subset data are unbiased, although it may lose statistical power.

The MAR situation occurs when the observed units are not necessarily a random subsample of the sampled units, but they are a random sample of the sampled values within subclasses defined by values of covariates. For example, if participants with low self-esteem in a study are less likely to return for follow-up sessions, then the observed data are not a random subsample in the study, but they are a random sample of the sampled values within subclasses defined by the values of self-esteem. In these two MCAR and MAR situations, the missing-data mechanisms are ignorable for likelihood-based inferences.^[8]

In NMD situation, the probability that the data are observed depends on the value of the outcome and varies across the levels of the covariates. In the above example of MAR situation, if the observed data are not a random subsample even among the subclasses defined by self-esteem or any other covariates, we have the NMD situation. Therefore the missing data are nonignorable, and analyses on the reduced sample that does not allow for the missing data mechanism are subject to bias.

In practice, it is often difficult to make the MCAR assumption, because there is often some degree of relationship between missing values and other observed data. Therefore the pattern of MAR is an assumption that is most often, but not always, tenable.

GENERAL APPROACHES TO MISSING DATA

Analysis techniques that deal with incomplete or missing data are useful and necessary, especially for experiments involving human subjects, because in these cases, the data are usually expensive, scarce, or significant. For any missing data methods to be effective, they should incorporate a specific missing data mechanism into their development. For missing data problems, three general approaches have been used, most of which were built based on the MAR assumption.

The first is the traditional approach of using only complete data subsets. However, this approach has been widely criticized, because it produces inconsistent estimators when missing data are not missing at random.^[11] This shortcoming has led many investigators to develop a second approach: efficient incomplete data analysis methods, using all available data (e.g., Refs. [5,6,12–14]). A third approach considers imputing missing data.^[15] An advantage of this latter approach is that one can use standard statistical software to analyze the resulting completely filled data.

Some missing data approaches tend to be computer intensive. Bayesian and/or simulation approaches have also been used. With the advancement of computer technology in recent years, more computer-intensive techniques such as the Gibbs sampler have gained popularity. See Ref. [16] for reviews on these various missing data-analysis methods.

The distribution theory for most of the resulting estimators has been based on an asymptotic normal distribution theory assuming large samples. The testing procedures involving missing data with small samples need to be developed further, even for bivariate sets of data. It has been shown that using almost maximum likelihood estimators is generally most powerful when the number of complete pairs of observations is moderately large (larger than 20) and the correlation is in the range of 0.3–0.9.^[17,18]

INCOMPLETE REPEATED MEASURES DATA ANALYSES

Investigators of repeated measurement designs (RMD) typically deal with small sample data. For this reason, an analysis method that does not rely too heavily on computational aspects may be desirable and feasible where the approximate distributions can be obtained. This entry reviews such methods for small sample repeated measures data analyses with missing values.

We denote a repeated measurement design with t treatments, p periods, and s sequences (groups) as RMD (t, p, s). For example, the simple two-period, two-treatment (A and B) crossover design, where subjects are assigned to the treatment sequences AB and BA, occurs frequently in practice. Data are obtained over two periods on two treatments, with two observations per subject. Each subject receives one treatment during the first period, and the other treatment during the second period. This idea has a strong appeal among physicians and pharmaceutical researchers especially. See Refs.[1–4] for general design issues with repeated measures data.

Generally, crossover designs have been the design of choice, when recruiting study subjects is expensive or the response is expected to show high variability among study subjects. However, if some subjects have missing data in one or more occasions in a general RMD (t, p, s), we are concerned with how to best and effectively deal with the situation.

Consider the model popularly employed:

$$y_{ijk} = \mu + \pi_i + \tau_t + \gamma_t + \lambda_k + \varepsilon_{ijk} \quad (3.1)$$

where y_{ijk} is the response from subject j in period i and sequence k , for $i = 1, \dots, p$, $j = 1, \dots, n_k$, and $k = 1, \dots, s$, μ is the overall mean, π_i is the effect of period i , τ_t is the effect of treatment t , γ_t is the residual effect of treatment t , λ_k is the effect of sequence k , and ε_{ijk} is the measurement error. Another popular form of model 3.1 can include random

subject effects instead of the fixed sequence effects λ_k to accommodate dependencies among the repeated measures data.^[2-5]

Rewrite model 3.1 in matrix form as

$$\mathbf{y} = \mathbf{Z}\mathbf{X}\boldsymbol{\beta} + \mathbf{e} = \mathbf{Z}\boldsymbol{\mu} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of responses, $\mathbf{Z}$ is an incidence matrix of ones and zeros, which equals to an identity matrix when there are no missing values, and $\mathbf{X}$ is a design matrix that relates the cell mean parameters to the repeated measurement design model parameters $\boldsymbol{\beta}$. The $\mathbf{e}$ is a vector of errors that has a mean zero with $\text{Cov}(\mathbf{e}) = \mathbf{V}$. When the variance-covariance matrix $\mathbf{V}$ is known, the maximum likelihood estimator for $\boldsymbol{\beta}$ is obtained as $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}^{-1}\hat{\boldsymbol{\mu}})$ with $\text{Cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}$ where $\mathbf{W} = (\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-1}$ and $\hat{\boldsymbol{\mu}} = (\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{y})$. For $\mathbf{V}$ unknown, Carrière^[5,6] used $\hat{\mathbf{V}}$ for $\mathbf{V}$. Then, the estimated treatment effect τ_t and its variance are $\tau_t = \mathbf{c}'_{1t}\boldsymbol{\beta}$ and $\text{var}(\hat{\tau}_t) = \mathbf{c}'_{1t}(\mathbf{X}'\hat{\mathbf{W}}^{-1}\mathbf{X})^{-1}\mathbf{c}_{1t}$ and those for residual effect γ_t are $\hat{\gamma}_t = \mathbf{c}'_{2t}\boldsymbol{\beta}$ and $\text{var}(\hat{\gamma}_t) = \mathbf{c}'_{2t}(\mathbf{X}'\hat{\mathbf{W}}^{-1}\mathbf{X})^{-1}\mathbf{c}_{2t}$ for treatment and residual effect vectors $\mathbf{c}_{1t}$ and $\mathbf{c}_{2t}$ for treatment t . Carrière^[5,6] gave detailed estimation results for specific designs.

Consider a two-treatment design, where missing data occur at random. The hypotheses we are interested in testing are that there are no direct treatment effect contrast

$$H_{01} : \tau = 0$$

for $\tau = (\tau_1 + \tau_2)/2$, and no residual treatment effect contrast

$$H_{02} : \gamma = 0$$

for $\gamma = (\gamma_1 + \gamma_2)/2$. The test statistics investigators consider are of the form

$$t_\gamma = \frac{\hat{\gamma}}{\sqrt{\text{var}(\hat{\gamma})}}, \text{ and } t_\tau = \frac{\hat{\tau}}{\sqrt{\text{var}(\hat{\tau})}} \quad (3.2)$$

where var is the estimated variance for estimator $\hat{\tau}$ or $\hat{\gamma}$, obtained as above.^[5,6]

The distribution of the statistics in Eq. 3.2 is then approximated suitably.^[5,6] Specifically, assuming that missing values occur in a nested pattern, which is quite plausible in general longitudinal data, a small sample testing procedure with missing data has been developed.^[5-7,18] Such an approach makes maximal use of all available information and is more powerful than an analysis that considers complete subset data only.

The derivation of an approximate distribution of the statistics in Eq. 3.2 first recognized that the analysis based on complete subset data uses the exact t distributions for the statistics in Eq. 3.2 with degrees of freedom of $N^{(L)} - s$ and $(p - 1)(N^{(L)} - s)$, respectively, for the model under a general covariance assumption and that under a compound symmetry model.

With incomplete data, Carrière^[5,6] and Patel^[7,18] discussed the rationale for deciding the approximate degrees of freedom for testing H_{01} and H_{02} . For compound symmetry covariance structures, the degrees of freedom are based on those used in the estimated variance component. For the case where the missing data occur in two stages ($L = 2$), $(p - 1)(N^{(L)} - s)$ was suggested for testing both τ and γ , where $N^{(L)}$ is the total number of subjects who completed the trial. When no pattern can be specified for the covariance matrix, the degrees of freedom were proposed once again based on the estimated variances and covariances involved in estimating the variances for τ and γ . They were proposed as $N^{(L)} - s$ for τ , but with the average degrees of freedom for γ , which is $(N^{(L)} + N - 2s - p_1 p_2)/2$, where p_1 is the number of periods with complete set of data up through the first p_1 periods and p_2 is the number of remaining periods with missing data. Their strategies were demonstrated to work well in repeated measurement data in small samples. See Refs. [5,6] for details.

A simulation study of the relative powers and levels of significance for small samples was conducted based on 1000 samples, and the procedure was demonstrated to be valid and preferable to the practice of discarding the incomplete sets. The following is the result of applying the procedure of Carrière^[5,6] to some useful and popular RMDs.

The incomplete data procedure was found to be most helpful for RMD (2, 2, 2) in testing no residual effects, because it increases the power of testing for no residual treatment effects. If one follows the two-step analysis strategy suggested by Grizzle,^[19] it is important to make an effort to increase the power of deciding accurately that there are no residual effects. Hence the incomplete data procedure is strongly recommended for RMD (2, 2, 2).

RMD (2, 3, 2) benefits greatly from the incomplete data procedure. The power increases rather significantly in multiperiod designs, and even more so when the missing data methods are incorporated. The power increases with the increasing level of ρ for testing both no treatment and no residual effects.

For this multiperiod design with $p = 3$, the missing data problem can be considered from another perspective that the design is a result of a planned two-period experiment, with observations of one additional period on a fraction of the subjects. It was shown that even with a dropout rate as high as 80% (or observations available on one additional period from 20% of the subjects), a significant improvement in efficiency can be achieved over two-period designs. Carrière^[2] reports that when one cannot observe a fraction of subjects for one more period, it seems clear that the resulting RMD (2, 2, 2) is of questionable value.

Carrière^[6] also discussed the conventional assumption that any approach would be reasonable with small sample data, because various covariance patterns are indistinguishable because of the sample size being too small. However, she demonstrated that, even for small sample

data, recognizing the correct covariance structure of the data is extremely important in terms of both empirical size and power of testing hypotheses of parameters of interest. The incomplete data procedure is not robust against the assumption about the covariance structure.

Generally, an incomplete data procedure that uses no specific covariance assumptions was clearly robust in comparison with procedures based on the compound symmetry assumption. Furthermore, an incomplete data procedure that assumes a special covariance structure did not maintain the empirical size at the nominal level for data when the assumptions were violated. The impact was greater when the correlation is high and the sample size is large.^[6]

USE OF PROXY INFORMATION

In this section, we consider incorporating proxy information by replacing missing values on some study subjects with approximate information about their possible responses. Huang^[20] considered a strategy for dealing with incomplete data by using proxy information, when such information is available in an attempt to avoid dealing with missing data situations. For example, in cancer clinical trials, some patients may be too sick to respond, but their caregivers are usually on hand to provide pertinent information instead. As another example, unavailable information about a subject's socioeconomic status can be filled in with proxy data about neighborhood affluence. As these two examples illustrate, such information is available in diverse forms and varies a great deal in quality.

The issue then is how to treat proxy information to maximize its usefulness. In principle, this strategy combines all of the three approaches mentioned earlier. It uses a complete data analysis method, fills in for missing values with proxy information, and utilizes all available data. However, this approach differs from the traditional imputation methods in that ours are based on better information than is provided by such impersonal methods as mean imputation or regression imputation using computer software.

Many investigators have shown empirically that there is a high degree of agreement among the responses of study subjects and their proxies, i.e., their care providers.^[21,22] However, we must account for possible systematic bias and heterogeneity of proxy data from the actual responses in statistical analyses of the data. The biggest advantage of incorporating proxy information may be that it avoids the missing data problem so that standard statistical analysis methods can be applied. See Refs. [23–28] for various situations, where the use of proxy was discussed in terms of some statistical criteria.

Assuming that the missing data occur at random, Huang^[20] presented the merits of using proxy data in place of missing values in a multivariate repeated measures data setting. A multiple linear model was considered that

accommodates another set of model parameters and also a different form of covariance matrix due to the use of proxies such that:

$$y_{ijk} = \mu + \pi_i + \tau_j + \gamma_i + \lambda_k + \varepsilon_{ijk} \quad (4.1)$$

$$p_{ijk} = \mu^a + \pi_i^a + \tau_j^a + \gamma_i^a + \lambda_k^a + \varepsilon_{1ijk} + \varepsilon_{2ijk} \quad (4.2)$$

where all terms in Eq. 4.1 are the same as in Eq. 3.1 from the actual observed data except that ε_{ijk} denotes the measurement error. Eq. 4.2 is to model the proxy data, where p_{ijk} is the missing data filled in with its proxy data for subject j who failed to give an actual response y_{ijk} in period i and sequence k , other model parameters with the superscript “a” are the same as in model 3.1, which are to model possible bias of using the proxy data in place of actual responses, and ε_{1ijk} and ε_{2ijk} are the measurement error terms from the actual data and proxy data, respectively. In particular, the proxy data have no bias if all terms in Eq. 4.2 with the superscript “a” are the same as those in Eq. 4.1. Further, the proxy data have the same variability as the actual data, if $\varepsilon_{2ijk} = 0$ for all i, j, k .

Utilizing proxy information and under the assumption of normality, the statistics for testing for the direct treatment and residual treatment effects similar to Eq. 3.2 are also shown to have approximate t -distributions with certain degrees of freedom,^[20] using much the same argument as Carrière^[5,6] and Patel,^[7,18] which was outlined in the previous section. The approximate degrees of freedom depend on the estimated variance and covariances used to build the t -ratios in Eq. 3.2 with proxy data. See also Refs. [5–7,18]. We summarize the findings for the case of monotone missing data in two stages with $p_2 = 1$ (i.e., only the last period has data missing).

If the actual data and the proxy data share a common covariance matrix, the degrees of freedom were set at $d_1 = (p - 1)(N - 2s)$ under the compound symmetry covariance assumption, because it is the degrees of freedom associated with the sum of squared error used. When the covariance matrix is common, but no pattern can be specified, the degrees of freedom were set at $d_2 = N - (3s + p_1)/2$, where d_2 is the average of the degrees of freedom associated with the two associated submatrices of the estimated covariance matrix.^[20]

When the actual data and proxy data do not share a common covariance matrix but are compound symmetry, the same degrees of freedom as above were used, namely $(p - 1)(N^{(L)} - s)$, as in the incomplete data method^[6] because in this case, it was found that proxy information did not affect the estimation of the mean and the variances of τ .^[20] When the different covariance matrix forms are unspecified, the average of the degrees of freedom associated with the three submatrices of the estimated covariance matrix was used and set at $d_3 = (2N - 3s - 2p_1)/3$.^[20] These approximations were demonstrated to work well in the simulation study. In particular, this approach was compared to the

incomplete analytical method summarized in the previous section.^[5,6]

To summarize the impact of using proxy information in lieu of missing values in RMD for selected designs, we consider RMD (2,2,2) and RMD (2,3,2) in small samples of size five subjects in each sequence. As anticipated, the results are highly dependent mainly on whether we can assume that the proxy and actual observed data share a common covariance matrix.^[20]

For RMD (2, 2, 2), when the proxy and actual data share a common covariance matrix, the testing procedure for the residual treatment effect using proxies provided results similar to those for the traditional incomplete data analysis, although the use of proxies appeared to be less efficient. However, the use of proxies appears to be significantly more powerful in testing the direct treatment effects. When the proxy data and actual data do not share a common covariance matrix, there was very little difference between the two approaches. Overall, they performed similarly in this case, although the incomplete method was slightly more powerful for testing. However, use of proxy does not help to alleviate the estimation problem of the traditional design under the unequal residual effect model, because use of proxy information to fill in the missing second period data does not result in any gain or loss in efficiency.

In RMD (2, 3, 2), however, there were substantial differences between the two approaches. When they share a common covariance matrix, for testing τ , the procedure that uses proxies was both more powerful and more prominent than in the RMD (2, 2, 2). For testing γ , the procedure using proxies was also better. In general, the method using proxies appears to be more powerful than is the traditional incomplete data-analysis strategy, especially when the sample size is small. However, the two approaches performed similarly in large samples. When they do not share a common covariance matrix, the incomplete analysis method was substantially more powerful for testing both τ and γ . Huang^[20] considered the impact of proxy data on design issues. For example, it was indicated that even with about 50% of the data filled in with proxies in the third period, the RMD (2, 3, 2) can attain 80% efficiency, as compared to the ideal situation of complete data with actual observations. In general, she discussed that one could design the experiment so that 80% efficiency can be attained, possible because of available proxies.

Therefore it appears that using proxy data to replace missing values is an excellent alternative to conventional strategies for addressing missing data problems. However, this is true only if the proxy data are similar in variability as the actual observed data. Also, the benefit occurs mainly in small samples. Nonetheless, the results are substantial and significant in that, even when over 50% of the data were replaced by proxy information, the efficiency, as compared to the ideal situation of complete actual data, could be over 90%.^[20] Analysts can use the proxy method

while enjoying all the conveniences of standard analysis methods.

CONCLUSION

In this entry, we have discussed the general approaches for dealing with missing data situations when using repeated measurement designs on small samples. Various investigators have shown that using complete subset data while discarding incomplete pairs is inefficient and proposed alternate methodologies. However, to be effective, any missing data procedure must properly recognize the reasons for having missing values. Also, any testing procedure should be suitable for use in practice as well as theory. The methods we reviewed in this entry are not based on iterative techniques or simulated imputations. As with any approximate procedures, appropriate adjustment to the distribution of the test statistics is important.

We focused on incomplete analytical methods for small sample repeated measures data, because most investigators of clinical trials work with such small samples. We applied the incomplete data procedure to selected RMDs and found them to be most helpful for testing for no residual treatment effects in the traditional two-period, two-treatment, two-sequence design because they increase the power of the test. The effect was even more significant in three-period designs. The power increases significantly in multiperiod designs, and even more so when the missing data methods are incorporated.

We also assessed the role of proxy information with a main goal to establish whether collecting proxy information in an effort to optimize the data collection process is useful in missing data analyses. Utilizing proxy information can be quite competitive and more efficient than other incomplete analytical methods without proxy, but only when the proxy and the observed data share a common covariance matrix. Proxy information loses its value when it is highly variable as compared to the actual observed data. We suggest that detailed exploratory data analyses are necessary before deciding to incorporate proxy information in the analysis. If the proxy and actual data are similar in their variability, incorporating them in the analysis is efficient while controlling its bias. It is also beneficial from a design point of view. However, when the proxy data are highly variable compared to the actual observed data, we recommend that investigators discard the proxy data and use the traditional strategies of incomplete data analysis methods.

ACKNOWLEDGMENTS

This work was funded by grants from the Natural Sciences and Engineering Research Council of Canada and the Alberta Heritage Foundation for Medical Research.

REFERENCES

1. Carrière, K.C.; Reinsel, G.C. Investigation of dual-balanced crossover designs for two treatments. *Biometrics* **1992**, *48*, 1157–1164.
2. Carrière, K.C. Crossover designs for clinical trials. *Stat. Med.* **1994b**, *13*, 1063–1069.
3. Carrière, K.C.; Huang, R. Traditional Paradigms in Pharmacoeconomics: Consideration for Cost-Effective Designs. In *Introduction to Applied Pharmacoeconomics*; McGraw-Hill Medical Publishing Division: New York, 2001, 19–39.
4. Carrière, K.C.; Huang, R. Crossover designs for two-treatment clinical trials. *J. Stat. Plan. Inference* **2000**, *87*, 125–134.
5. Carrière, K.C. Incomplete repeated measures data analysis in the presence of treatment effects. *J. Am. Stat. Assoc.* **1994a**, *89*, 680–686.
6. Carrière, K.C. Methods for repeated measures data analysis with missing values. *J. Stat. Plan. Inference* **1999**, *77*, 221–236.
7. Patel, H.I. Analysis of incomplete data from a clinical trial with repeated measurements. *Biometrika* **1991**, *78*, 609–919.
8. Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*; John Wiley: New York, 1987.
9. Heitjan, D.F. Annotation: What can be done about missing data? Approaches to imputation. *Am. J. Public Health* **1987**, *87*, 548–550.
10. Graham, J.W.; Hofer, S.M.; MacKinnon, D.P. Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivar. Behav. Res.* **1996**, *31*, 197–218.
11. Schafer, J.L. *Analysis of Incomplete Multivariate Data*; Chapman & Hall: London, 1997.
12. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc., B* **1977**, *56*, 463–474.
13. Efron, B. Missing data, imputation, and the bootstrap. *J. Am. Stat. Assoc.* **1994**, *89*, 463–479.
14. Efron, B.; Tibshirani, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* **1986**, *1*, 54–77.
15. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; John Wiley: New York, 1987.
16. Wothke, W. Longitudinal and Multi-Group Modeling with Missing Data. In *Modeling Longitudinal and Multiple Group Data: Practical Issues, Applied Approaches and Specific Examples*; Little, T.D., Schnabel, K.U., Baumert, J., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, 1998.
17. Lin, P.; Stivers, L.E. Testing for equality of means with incomplete data on one variable: A Monte Carlo study. *J. Am. Stat. Assoc.* **1975**, *70*, 190–193.
18. Patel, H.I. Analysis of incomplete data in a two-period crossover design with reference to clinical trials. *Biometrika* **1985**, *72*, 411–418.
19. Grizzle, J.E. The two-period changeover design and its use in clinical trials. *Biometrics* **1965**, *21*, 467–480.
20. Huang, R. Issues with Repeated Measures Data. Unpublished Ph.D. Thesis; University of Alberta, 2001.
21. Grootendorst, P.; Feeny, D.; Furlong, W. Does it matter whom and how you ask? An investigation into inter- and intra-rater agreement in the 1996 Ontario Health Survey. *J. Clin. Epidemiol.* **1997**, *50*, 127–135.
22. Jalukar, V.; Funk, G.F.; Christensen, A.J.; Karnell, L.H.; Moran, P.J. Health states following head and neck cancer treatment: Patient, health-care professional, and public perspectives. *Health Neck* **1998**, *20*, 600–608.
23. McCallum, B.T. Relative asymptotic bias from errors of omission and measurement. *Econometrica* **1972**, *40*, 757–758.
24. Wickens, M.R. A note on the use of proxy variables. *Econometrica* **1972**, *40*, 759–761.
25. Barnow, B.S. The use of proxy variables when one or two independent variables are measured with error. *Am. Stat.* **1976**, *30*, 119–121.
26. Frost, P.A. Proxy variables and specification bias. *Rev. Econ. Stat.* **1979**, *61*, 323–325.
27. Ohtani, K. On the use of a proxy variable in prediction: An MSE comparison. *Rev. Econ. Stat.* **1981**, *63*, 627–629.
28. Trenkler, G.; Stahlecker, P. Dropping variables versus use of proxy variables in linear regression. *J. Stat. Plan. Inference* **1996**, *50*, 65–75.

Repeated Measures Data with Missing Values Analysis: Overview of Methods

Keumhee Carriere Chough

University of Alberta, Edmonton, Alberta, Canada

Yuanyuan Liang

Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, San Antonio, Texas, U.S.A.

Taesung Park

Seoul National University, Seoul, Korea

Abstract

In this entry we discuss the merits and drawbacks of available small sample approaches with approximately normally distributed repeated measures of data with missing values. This discussion is limited to cases where the missing data occurs on the outcome variable.

INTRODUCTION

The missing data problem, which persists in much of empirical scientific investigations, is particularly common in repeated measures data. One reason for this is that the same subjects are used repeatedly over time. One of the main goals in dealing with missing data is to try to remove biases and reduce the estimation variances in an attempt to improve the overall estimation efficiency, thereby increasing study power. With current advancements in computer technology, many computational analysis methods have been developed, but most have a major drawback: they rely heavily on large sample theory.^[1-3]

In their overview of missing data methods for approximately normally distributed repeated measures data in small samples, Carrière et al.^[4] discuss non-iterative procedures for data with compound symmetry and unstructured covariance matrices, as well as the use of proxy information. In general, the approach to incomplete data involves identifying appropriate missing data mechanisms.

In this work, which supplements the previous work done by Carrière et al.,^[4] we discuss the merits and drawbacks of available small sample approaches with approximately normally distributed repeated measures data with missing values. This discussion is limited to cases where the missing data occurs on the outcome variable. In particular, this entry expands the discussion to include multiple imputation approaches. We use a numerical example to demonstrate the practical implications.

MISSING DATA MECHANISMS AND IMPLICATIONS

Most missing data methods are based on an assumption that missingness indicators contain true values that are meaningful for analysis.^[5] Therefore, we must make an effort to reveal the true values. Obviously, procedures based only on the complete subset data can create serious biases and result in inefficient analyses.^[1, 5]

We consider repeated measures data y_{ij} for subject $j = 1, \dots, N$ in period $i = 1, \dots, p$ in the presence of treatment effects, along with a covariate or design vector x_{ij} with $m_{ij} = 0$ if y_{ij} is missing and 1 if y_{ij} is observed. The goal is to efficiently estimate the contrast of treatment effects in the presence of missing values.

Little and Rubin^[5] define three unique of types missing data mechanisms that occur in different situations: missing completely at random (MCAR), missing at random (MAR), and nonignorable missing (NIM). In the MCAR situation, cases with complete data are indistinguishable from cases with incomplete data, so that $E(y_{ij} | m_{ij} = 1) = E(y_{ij} | m_{ij} = 0)$. Also, we have $p(m_{ij} = 1 | x_{ij}, y_{ij}) = p(m_{ij} = 1)$. This implies that the investigator can get consistent results by examining only the complete subset data. There is no danger of biased estimation, whether or not incomplete pairs receive appropriate attention. However, data can be missing for uncontrolled events in the course of data collection, and missing data are often associated with study variables—both the outcome and the covariates. The MCAR is too strong an assumption in reality.

In the MAR situation, cases with incomplete data differ from cases with complete data. The probability of observing a missing value depends on the observed values, but not on the unobserved values (both the covariates and outcome). Here, we have $E(y_{ij} | x_{ij}, m_{ij} = 1) = E(y_{ij} | x_{ij}, m_{ij} = 0)$ and within some subclasses of the data, they are still a random sample of cases.^[5] Therefore, missing values are traceable or predictable from other variables available in the database already observed. Investigators have approached this situation in a variety of ways, including imputation,^[6–11] weighting,^[5,12,13] resampling methods,^[7,14] data augmentation,^[15] and the Gibbs sampler.^[16] The success of these techniques depends on correctly specifying the missingness mechanism and/or the imputation model. Rubin and Little^[5] show that, in the case of MAR data, the likelihood inference produces consistent and efficient results even if the missing data mechanism is ignored. In the context of a generalized estimating equations approach, Liang and Zeger^[17] have found that the normal-model GEE is consistent when the missing data depend on any number of previous observations (i.e., MAR), if the mean function is correctly specified. It appears that both the bias and the standard error of estimates can be reduced as long as all available cases are used for both MCAR and MAR data, especially when using likelihood-based methods.^[5] Little^[10] and Park and Lee^[18,19] advocated simple pattern mixture models to test the dependence of responses on the missing data mechanisms, assuming that the missing data are ignorable. If the test is rejected, the missing data are determined not to be of type MCAR, but MAR.

In the NIM situation, the missing data mechanism is not completely random (non-MCAR), nor is it predictable from other variables in the database (non-MAR). Here, the reason for the missing data is explainable, but unmeasurable, because the very variable causing the data to be missing is unobservable.^[5] For valid analysis results, it is important to handle nonignorable missing data appropriately. Imputation, or a sampling-importance resampling approach, has been used to achieve the desired efficiency and unbiased analysis.^[12] Sheng and Carrière^[14] conclude that, in item response data analysis, the bootstrap under imputation approach can accommodate all missing data mechanisms, including the NIM situation, efficiently and without bias. However, this procedure is based on large sample theory.

In summary, consistent and efficient analysis depends on the investigator's ability to choose an appropriate missing data mechanism. It is especially important to devise a model for the nonignorable missing values, which accommodates a particular situation. Since any model for nonignorable data is specific to a given study and cannot be discussed in general terms, we have limited this discussion to MAR and MCAR situations.

AVAILABLE DATA ANALYSIS

Many investigators developed methods to utilize all available data rather than discarding incomplete pairs. Little and Rubin^[5] noted that available case analysis methods generally lack practical appeal due to the imbalance problem of varying sample bases. When the sample bases vary from one variable to another variable within a study, obtaining estimates of the parameters, or even asymptotic standard errors, can become very complicated. However, this is not an issue when using the ML method under an MCAR or MAR missing data mechanism with large samples, as there are valid standard errors and tests available based on large sample theory.^[3] For small sample cases, though, further work is still needed. Several studies, although limited in scope, have indicated that the available case analysis method is superior to complete subset analysis.^[1,2,20]

Approximate inference procedures have also been proposed for small samples to utilize all available data, if not all available cases. Carrière et al.^[4] present an overview of missing data strategies, with a focus on available data analysis methods for approximately normal repeated measures data. Their concern is with small sample data, and they discuss approximate distributions of estimators for making inferences for parameters of interest.

In particular, Carrière^[1,2] has developed small sample testing procedures based on the maximum likelihood method for two particular situations of the within-subject covariance matrix—when it has a compound symmetry pattern and when it is unstructured. Approximate solutions for small sample inference procedure were found upon obtaining the explicit forms of the standard errors of the parameter estimators of interest. Carrière^[1,2] also suggested an approximate degrees of freedom approach based on the Satterthwaite^[21] approximation method and on the assumption that the variability of higher-order terms is negligible. Although rather complex, these methods have been demonstrated to work well in computer simulation studies.

Other small sample available data analysis methods can also be applicable when using common software (for example, PROC MIXED in SAS^[22]), with approximation techniques similar to those suggested by Carrière,^[1,2] but based on different assumptions. Comparing her procedure with the SAS procedure, Carrière^[1] noted that the inference based on the available data method depends on the design structure. PROC MIXED in SAS approximates the distribution of the estimators uniformly to make inferences for all model parameters. It appears that caution is advised if the design is not orthogonal, as the analysis results can be more liberal or more conservative than expected.^[1]

Although the Carrière procedure^[1,2] can generally apply to any repeated measures data with missing values occurring monotonically and at random, the estimation methods are applicable to only two, very specific cases.

The procedure can be extended to other forms of covariance structures in an attempt to obtain approximate or even asymptotic distributions of the estimators. But because of the complexity involved, it may not be practical or even possible to use the procedure without making many unrealistic assumptions. For this reason, we echo Little and Rubin^[5] who concluded that the current state of available data/case analysis is not generally satisfactory for small sample repeated measures data. Further development is needed.

IMPUTATIONS

Imputations (both single and multiple) require creating a predictive distribution based on the observed data, either implicitly or explicitly.^[9] Examples include mean, regression, stochastic regression, hot or cold deck imputations, and substitutions. Composite methods combining some of these imputations and other techniques can also be used. Typically, imputation builds on the conditional mean of the missing outcome given observed data. Practically, imputation approaches are attractive because of their easy implementation and computational convenience, because the data are completely filled in even if the values are artificial, and the pseudocomplete data can be analyzed using standard software.

Single Imputations

Single imputation approaches impute a single data point for each missing value. Single imputations were used in empirical research even before the theory had been formally developed. Using imputation approaches for each draw of the data, a complete data set $\mathbf{y} = (\mathbf{y}_{\text{obs}}, \mathbf{y}'_{\text{mis}})$ is obtained by combining the observed data set $\mathbf{y}_{\text{obs}}$ and the imputed data set for the missing values $\mathbf{y}'_{\text{mis}}$. The most popular of these single imputations is probably the mean substitution based on a specific covariate pattern. See Rubin^[9] for details.

From a slightly different perspective, Huang et al.^[23] discussed a single imputation strategy where the imputation uses proxy information from caregivers or family members of a respondent who can provide approximate information on behalf of the respondents. Then, this proxy information can be used for analysis instead of dealing with missing responses. The principal idea of this approach is implicit, in that it assumes an underlying model where the respondents and their proxies share a common mean and variance. Possible differences in the mean and variance can be accommodated. Huang et al.^[23] also provided an approximate degrees of freedom solution to the testing hypothesis problem for missing repeated measures data with proxy.

Huang et al.^[23] reported that a single imputation with proxy information can play a significant role in missing data analyses. They also discussed design implications in

order to provide general guidelines for utilizing available proxy information to improve data collection strategy for missing data situations. However, when the variability between the actual and proxy data sets is different, it is better to use other available case analyses so as to maintain the type I error rate and to increase the power of testing parameters of interest.

Multiple Imputations

In the multiple imputation strategy, which was first proposed by Rubin,^[8] the imputed values are supposed to represent repeated random draws under a given model for each missing value. Overall inferences can then be drawn by combining results from complete data sets.

Multiple imputation involves drawing multiple values for each missing data point and there are various methods available for drawing the values to impute.^[9] For each draw of the data under a chosen imputation method, a complete data set $\mathbf{y}^m = (\mathbf{y}_{\text{obs}}, \mathbf{y}^m_{\text{mis}})$ is created by combining $\mathbf{y}_{\text{obs}}$ and $\mathbf{y}^m_{\text{mis}}$, $m = 1, \dots, M$, as defined earlier. Substantial empirical work^[24–26] has shown that multiple imputation with $M = 3$ or 5 works well with typical fractions (<30%) of missing data in surveys. Rubin^[9] suggests using standard inference methods to analyze each of the complete data sets $\mathbf{y}^m$ and then combining them for overall analysis results. In this section, we discuss the imputation strategy for repeated measures data. We assume that missing data occur monotonically, so that the data can be clustered into L groups according to the number of periods the subjects completed. Carrière^[1] has noted that some intermediate missing values, which are frequently few in number, can be imputed or removed before proceeding with the analysis, without much loss of information.

Let $\mathbf{y}_j = (y_{1j}, \dots, y_{pj})^T$ be the vector of responses from subject j where y_{ij} is from subject j in period i , $i = 1, \dots, p$ and $j = 1, \dots, N$. We can express the data from this repeated measurement design by the model

$$\mathbf{y}_j = \boldsymbol{\mu}_j + \boldsymbol{\varepsilon}_j = \mathbf{x}_j \boldsymbol{\beta} + \boldsymbol{\varepsilon}_j \quad (1)$$

for a $q \times 1$ vector of parameters $\boldsymbol{\beta}$ linking responses to some design or covariate matrix $\mathbf{x}$.

Richardson and Flack^[3] propose a multiple imputation strategy for analyzing the data based on regression models. Called the residual draw (RD) method, this approach is based on Rubin^[9] and is specific to three-period crossover trial data. The RD method imputes the conditional predictive mean of the incomplete case with additional noise. Basically, the investigator draws noise from the empirical distribution, and is therefore less sensitive to violations of the normal assumption.

The RD method is implemented separately for each group/sequence. It involves estimating the least squares estimators to fit a regression of the observed $(p + 1)$ th

period data on the observed first p periods data to obtain $\hat{\beta}$ for each regression model. Then, the error variance is set as $\sigma^{*2} = \text{SSE}/W$ where SSE is the residual sum of squares from the least squares fit and W is a draw from a chi-square distribution with the degrees of freedom of SSE. Then, a β^* is drawn from the $N(\hat{\beta}, \sigma^{*2}(x^T x)^{-1})$. Finally, to impute for an incomplete case i , the method imputes a value $Y_i^* = x_i^T \beta^* + r_0^* = Y_0 + (x_i^T - x_0^T) \beta^*$ where those with a subscript 0 are data values observed for a complete case drawn at random from the subjects in the given sequence. The r_0^* is the residual of that case when $\beta = \beta^*$. Through simulations, Richardson and Flack^[3] show that their strategy works well. However, their simulation uses relatively large sample data.

Huang and Carrière^[27] suggest that a less parametric approach than the one proposed by Richardson and Flack would be more robust. They impute the missing values using conditional distribution of the missing data, given the observed data in previous periods. The p_1 -vector $y_{(p_1)j}^T = (y_{1j}, \dots, y_{p_1j})$ for the complete subset data up to the first p_1 periods is assumed to be distributed as multivariate normal with mean $\mu_{(p_1)}^T = (\mu_1, \dots, \mu_{p_1})$ and covariance matrix Σ_{11} , the $p_1 \times p_1$ subcovariance matrix for the first p_1 periods. Then, the conditional distribution of $y_{p_1+1,j}$ given $y_{(p_1)j}$ is normal with mean $\mu_{p_1+1} + \sigma_{21} \Sigma_{11}^{-1} (y_{(p_1)j} - \mu_{(p_1)})$, and variance $\sigma_{22} - \sigma_{21} \Sigma_{11}^{-1} \sigma_{12}$, where Σ_{11} is the submatrix of Σ for the first p_1 rows and p_1 columns, σ_{21} is that for the $(p_1 + 1)$ th row and the first p_1 columns, with $\sigma_{12} = \sigma_{21}^T$, and σ_{22} is the $(p_1 + 1)$ th diagonal element of Σ . For subsequent periods, replace p_1 with $p_1 + 1$, and repeat the process. See also Carrière.^[2] Since the covariance matrix is usually unknown, the investigator uses the respective sample estimates, obtained based on complete subsets of the data and denoted by s_{ij} , s_{ij} , and S_{ij} . Then, Huang and Carrière^[27] extended the imputation strategy that Rubin^[9] suggested for a univariate normal model to a multivariate normal model.

This strategy involves estimating the conditional mean $\hat{\mu}_{p_1+1} = \bar{y}_{p_1+1} + s_{21} S_{11}^{-1} (y_{(p_1)j} - \bar{y}_{(p_1)})$, and the variance $\hat{\sigma}^2 = s_{22} - s_{21} S_{11}^{-1} s_{12}$. then, the conditional estimators are updated by drawing a chi-square random variable g with degrees of freedom $N^{(l)} - s$ and a random variable from a standard normal distribution z as $\sigma^* = \hat{\sigma} (N^{(l)} - s)/g$ and $\mu_{p_1+1}^* = \hat{\mu}_{p_1+1} + \sigma^* z / \sqrt{N^{(l)}}$, where $N^{(l)}$ is the total number of observations at the missing data stage l and s is the number of sequences. Then, a random variable z is drawn from a standard normal distribution to impute for the missing values in the period $p_1 + 1$, as $y_{p_1+1,j} = \mu_{p_1+1}^* + \sigma^* z$. This is repeated for all missing components in the period $p_1 + 1$. Treating the imputed values as if they were actual values, the steps are repeated for the next periods, with p_1 replaced by $p_1 + 1$. This whole process is repeated M times to create M multiply imputed data sets.

The multiply imputed data sets are analyzed to obtain an overall result. First, the usual data analysis is performed for

each of the M imputed data sets. The M analysis results are then combined to give a repeated-imputation inference as follows. Let $\hat{\beta}_m$ and W_m be the estimators and their associated variance-covariance matrix for β from the complete data set m , $m = 1, \dots, M$. The overall estimator of β from M data sets is obtained from $\bar{\beta}_M = \sum_{m=1}^M \hat{\beta}_m / M$ and its variance as $V(\bar{\beta}_M) = T_M = \bar{W}_M + (1 + M^{-1}) B_M$. The M within-imputation variances W_m are averaged to obtain $\bar{W}_M = \sum_{m=1}^M W_m / M$, and $B_M = \sum_{m=1}^M (\hat{\beta}_m - \bar{\beta}_M)(\hat{\beta}_m - \bar{\beta}_M)^T / (M - 1)$ is the between-imputation variance.

To test a hypothesis of a linear contrast $\theta = l^T \beta$, Rubin^[9] considers an approximate distribution for θ given by

$$(\theta - \bar{\theta}_M) [l^T T_M l]^{-1/2} \sim t_v \quad (2)$$

where $\bar{\theta}_M = l^T \bar{\beta}_M$, and the degree of freedom v is

$$\tilde{v} = v_0 \{ [f(v_0)(1 - r_M)]^{-1} + v_0 / v \}^{-1} \quad (3)$$

where $f(v_0) = (v_0 + 1) / (v_0 + 3)$, $r_M = (1 + M^{-1}) \text{tr}(B_M T_M^{-1}) / q$, and v_0 is the degree of freedom based on the complete subset data. The r_M estimates the fraction of information on β that is missing due to nonresponses. This fraction can be no larger than that of all missing data.^[9]

IMPUTATION OR AVAILABLE CASE ANALYSIS?

The available data analysis method is not always satisfactory for the reasons stated in the section Available Data Analysis above. Then, the imputation approach might be one alternate method to use. There are others, but most of them build on simulation methods (for example, data augmentation^[15] and Gibbs sampler^[16]). This entry compares available data analysis methods and imputations, among other possible approaches.

Many investigators favor multiple imputations over single imputations because they can generate random variations in imputed artificial data. However, many studies have also found that this method does not perform as effectively as expected. For example, in the context of generalized linear models, Xie and Paik^[13] considered four multiple imputation strategies and sample average imputations. They conclude that as long as one uses all available data, all approaches are consistent and efficient. For continuous responses, Huang and Carrière^[27] found that there is no real advantage in creating multiple complete data sets, as the analysis method that uses only available data performs as well as or better than the multiple imputation method.

Specifically, Huang and Carrière^[27] note that the corresponding asymptotic distributions appear to fit reasonably well for the two approaches they considered (ML and MI approaches) as long as all available data are used. In that study, the multiple imputation method performed as well as

the ML method^[1,2] for all situations considered. However, the ML method was found generally superior to the multiple imputation method, especially when the correlation is large and the covariance structure is unspecified.

In light of the technical limitations noted earlier regarding the available data method, Huang and Carrière^[27] suggest adopting the multiple imputation method when types of covariance structures other than compound symmetry or unstructured covariance matrices provide a good fit to the data in small samples. Although their implementation of the multiple imputation method for small samples was not entirely satisfactory in terms of keeping the type I error and testing power, they recommended that the violation may be far less serious than that by the ML approach with large sample theory assumptions. Also, it could be considerably more complicated to try to calculate the asymptotic standard error of the ML estimators than using multiple imputations. See also Richardson and Flack.^[3]

As noted in the previous section, the real advantage of using multiple imputations is not in obtaining efficient estimators, but in obtaining unbiased estimators for parameters by attempting to reveal the true values for missing data. However, the one advantage of the multiple imputation (and of any other imputation techniques) is its capacity to use standard analysis statistical software. As long as all available data are used, both ML and MI approaches seem to be valid for univariate data as well as for repeated measures data analyses. Further work needs to be done on deriving the asymptotic small sample procedures for available data analysis. Also, there is a need to investigate the sample sizes required for the asymptotic ML procedures that can be validly used with standard software.

This study has not dealt with the NIM case. We suggest that the alternative imputation strategy using proxy information, as in Huang et al.,^[23] could be effective. This approach would likely work if proxy providers carefully

assess reasonable responses from the respondents with a full understanding of the reasons for the missing data in the given situation.

ANALYSIS OF BRONCHIAL ASTHMA DATA

To contrast the methods discussed in this paper, we analyzed the bronchial asthma data in Patel,^[28] which utilizes the traditional two-period, two-treatment, two-sequence design with eight and nine patients in each sequence, respectively, and $N = 17$. We conducted an exploratory data analysis to determine the covariance structure, and we observed the compound symmetry covariance structure to provide an adequate fit to the data.

We analyzed the original complete data using PROC MIXED in SAS, and compared the (exact) test for the treatment and carryover effects to the t distribution with the degrees of freedom computed from $(N - 2) = 15$ from the complete actual data. Grizzle^[29] suggested a two-stage analysis for assessing the significance of the residual effects, using a P -value of 0.15 to determine insignificance of a residual effect. Based on the two-stage analysis, we did not remove the residual effect, and thus concluded that, in the presence of a nonnegligible residual effect, the test of a direct treatment effect was significant at 0.05 (first row of Table 1).

Next, we induced missing values by deleting the measurements in the second period from the four subjects in the BA treatment sequence to produce MAR data. We then compared the analysis based on the complete subset data omitting both measurements from the four subjects with missing observations in the second period to t distribution, with the exact degrees of freedom computed from the complete subset data of $17 - 4 = 13$ patients (second row of Table 1). The results are similar to the original data analysis, with a slightly lower estimate for the treatment effect

Table 1 Analysis Results of Bronchial Asthma Data

Method	$(\hat{\tau})$	se ($\hat{\tau}$)	df	P -value	$\hat{\gamma}$	se ($\hat{\gamma}$)	df	P -value
Full original data	-0.384	0.169	15	0.038	-0.512	0.315	15	0.125
Complete subset data	-0.404	0.178	11	0.044	-0.503	0.323	11	0.148
Incomplete method ^a	-0.384	0.152	11	0.028	-0.471	0.285	11	0.127
Incomplete method ^b	-0.384	0.163	10	0.040	-0.470	0.309	10	0.159
Proxy ^c	-0.384	0.177	13	0.049	-0.468	0.340	13	0.193
Proxy ^c	-0.384	0.180	13	0.053	-0.468	0.348	13	0.206
Proxy ^d	-0.384	0.166	13	0.038	-0.468	0.319	13	0.166
Proxy ^d	-0.384	0.169	13	0.041	-0.468	0.326	13	0.174
MI ^e	-0.384	0.164	9.429	0.043	-0.554	0.333	6.554	0.143

^aMethod by Carrière.^[1]

^bMethod by PROC MIXED of SAS.

^cMethod by Huang et al.^[23]

^dMethod by PROC MIXED of SAS with df adjustment of Huang et al.^[23]

^eMultiple imputation method by Huang and Carrière.^[27]

and a higher estimate for the residual effect, but higher standard errors for both.

Applying the available case analysis method of Carrière,^[1] the tests are compared against t_{11} , the approximate degrees of freedom of $(N^{(2)} - 2)$, where $N^{(2)}$ is the number of subjects with complete observations. We obtained results similar to the original data analysis (third row of Table 1). We see improved power over the complete subset analysis. A slightly different approximation procedure by SAS PROC MIXED produced results similar to complete subset analysis (fourth row of Table 1). In particular, the residual effect is now insignificant according to Grizzle.^[29] Some comparisons of these two available case methods can be found in Carrière.^[1]

Applying the single imputation approach of Huang et al.,^[23] we considered: (1) choosing values close to their first period values, taking into account that they tend to be smaller in period 2 than in period 1 (*proxy I* = 2.5, 1.5, 3.0, 1.0 and (2) slightly overestimating the missing data (*proxy II* = 3.0, 2.0, 3.5, 1.25). We first examined the bias and possible variance heterogeneity of the proxy data from the actual responses. The variances of the proxy data sets are a little larger than those of the actual data, but the test of a common variance is not rejected. Hence, the tests are compared against t_{13} , with approximate degrees of freedom of $(N - 4)$, assuming a common variance between proxy and the real data. Rows 5 and 6 in Table 1 show the results of the analysis described in Huang et al.,^[23] and rows 7 and 8 show those using SAS PROC MIXED. When using PROC MIXED, we adjusted the degrees of freedom as suggested by Huang et al.^[23] These proxy approaches reflected the mechanism used in choosing the proxy values with slightly less sensitive results in the proxy II than in proxy I, and an indication that the residual effects are non-existent. Huang et al.^[23] note that tests of treatment effects are affected by the presence of proxy data, resulting in a less sensitive outcome than for nonproxy approaches in this case. However, if we remove the insignificant residual effects from the model, we reject the null hypothesis of equal treatment effects in all cases, and conclude that the treatment effects are significantly different, as in the analysis of full original data.

Finally, we applied the non-regression-based multiple imputation approach of Huang and Carrière.^[27] The results are reported in row 9 of Table 1. The advantage of this approach is the convenience of using any standard software of choice for the analysis, but the penalty is quite high, as reflected in the adjusted degrees of freedom. The qualitative results are similar to those from the original data analysis and the available case analysis (see rows 1 and 3).

CONCLUSIONS

The purpose of this entry is twofold: the first is to supplement the Carrière et al.^[4] by including the imputation

procedure for small sample repeated measures data and the second is to compare implications of various incomplete data methods. Available data analyses are said to be generally unsatisfactory, because the calculation of the asymptotic standard errors of estimators is quite complex even for MCAR. However, when available, they are found to be more powerful than multiple imputations. Imputation-based procedures fill in for the missing values to analyze them by standard methods. When desired, the strategy of the multiple imputation method discussed by Huang and Carrière^[27] may be used for small sample repeated measures data; its overall performance was not substantially worse than that of the alternative. The discussion was limited to the MCAR and MAR cases and future work dealing with the NIM case will be forthcoming.

ACKNOWLEDGMENTS

This work was funded by grants from the Natural Sciences and Engineering Research Council of Canada, the Alberta Heritage Foundation for Medical Research, and the Korea Federation of Science and Technology Societies (Brain Pool) to K.C. Carrière and from the Korean Research Foundation (KRF-2004-015-C0086) to T.S. Park.

REFERENCES

1. Carrière, K.C. Incomplete repeated measures data analysis in the presence of treatment effects. *J. Am. Statist. Assoc.* **1994**, *89*, 680–686.
2. Carrière, K.C. Methods for repeated measures data analysis with missing values. *J. Statist. Plann. Inference.* **1999**, *77*, 221–236.
3. Richardson, B.A.; Flack, V.F. The analysis of incomplete data in the three-period two-treatment cross-over design for clinical trials. *Stat. Med.* **1996**, *15* (2), 127–143.
4. Carrière, K.C.; Huang, R.; Sheng, X.; Liang, Y. Missing values in repeated measures designs. *Encyclopedia of Biopharmaceutical Statistics*, 2nd Ed.; Marcel Dekker Inc.: New York, 2003.
5. Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*; John Wiley: New York, 1987.
6. Barnard, J.O.; Rubin, D.B. Small-sample degrees of freedom with multiple imputation. *Biometrika.* **1999**, *86* (4), 948–955.
7. Efron, B. Missing data, imputation, and the bootstrap. *J. Am. Statist. Assoc.* **1994**, *89*, 463–478.
8. Gelfand, A.E.; Smith, A.F.M. Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* **1990**, *85*, 398–409.
9. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; John Wiley: New York, 1987.
10. Little, R. Pattern-mixture models for multivariate incomplete data. *J. Am. Statist. Assoc.* **1993**, *88*, 125–134.
11. Rubin, D.B.; Schenker, N. Interval estimation from multiple imputed data: a case study using agriculture industry codes. *J. Official Statist.* **1987**, *3*, 375–387.

12. Paik, M.C. The generalized estimating equation approach when data are not missing completely at random. *J. Am. Statist. Assoc.* **1997**, *92*, 1320–1329.
13. Xie, F.; Paik, M.C. Generalized estimating equation model for binary outcomes with missing covariates. *Biometrics* **1997**, *53*, 1458–1466.
14. Sheng, X.; Carrière, K.C. Strategies for analyzing missing item response data with an application. *Biom. J.* **2005**, *47* (5), 605–615.
15. Tanner, M.A.; Wong, W.H. The calculation of posterior distributions by data augmentation (C/R: p541–550). *J. Am. Statist. Assoc.* **1987**, *82*, 528–540.
16. Gelman, A.; Rubin, D.B. Inference from iterative simulation using multiple sequences (Disc: p. 483–501, 503–511). *Statist. Sci.* **1992**, *7*, 457–472.
17. Liang, K.-Y.; Zeger, S. Longitudinal data analysis using generalized linear models. *Biometrika* **1986**, *73*, 13–33.
18. Park, T.; Lee, S.Y. A test of missing completely at random for longitudinal data with missing observations. *Statist. Med.* **1997**, *16*, 1859–1871.
19. Park, T.; Lee, S.Y. Simple pattern-mixture models for longitudinal data with missing observations. *Statist. Med.* **1999**, *18*, 2933–2941.
20. Kim, J.O.; Curry, J. The treatment of missing data in multivariate analysis. *Sociol. Methods Res.* **1977**, *6*, 215–240.
21. Satterthwaite, F.E. An approximate distribution of estimates of variance components. *Biometrics Bull.* **1946**, *2*, 110–114.
22. SAS Institute. SAS Technical Report P-229 (6.07); SAS Institute, Inc.: Cary, NC, 2002.
23. Huang, R.; Liang, Y.Y.; Carrière, K.C. The role of proxy information in missing data analysis. *Statist. Methods Med. Res.* **2005**, *14* (5), 457–471.
24. Li, K.H.; Raghunathan, T.E.; Rubin, D.B. Large-sample significance levels from multiply imputed data using moment-based statistics and an *F* reference distribution. *J. Am. Statist. Assoc.* **1991**, *86*, 1065–1073.
25. Li, K.H.; Meng, X.L.; Raghunathan, T.E.; Rubin, D.B. Significance levels from repeated *p*-values with multiply-imputed data. *Statist. Sinica* **1991**, *1*, 65–92.
26. Meng, X.L.; Rubin, D.B. Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **1992**, *79*, 103–111.
27. Huang, R.; Carrière, K.C. Comparison of methods for incomplete repeated measures data analysis in small samples. *J. Statist. Plann. Inference* **2006**, *136* (1), 235–247.
28. Patel, H.I. Analysis of incomplete data from a clinical trial with repeated measurements. *Biometrika* **1991**, *78*, 609–919.
29. Grizzle, J.E. The two-period change-over design and its use in clinical trials. *Biometrics* **1965**, *21*, 467–480.

Reproductive and Developmental Studies

James J. Chen

U.S. Food and Drug Administration, Jefferson, Arkansas, U.S.A.

Abstract

The ultimate goal of reproductive and developmental studies is to assess reproductive risk to mature adults and risk to the developing individual at all stages from exposure to drugs and environmental compounds. This entry discusses statistical design methods and analysis and touches briefly on regulatory requirements in reproductive and developmental studies.

INTRODUCTION

Soon after the thalidomide tragedy in the late 1950s and early 1960s, which resulted in over 8000 malformed babies, there has been a great deal of interest in predictive tests of reproductive and developmental effects. The ultimate goal of the studies is to assess reproductive risk to mature adults and risk to the developing individual at all stages—from conception to sexual maturity, and from exposure to drugs and environmental compounds. Adverse reproductive and developmental effects include effects on male and female fecundity, spontaneous abortion, infant and child death, congenital malformations, growth retardation, and mental retardation. Some compounds may cause rather specific adverse effects, while others may have a broad spectrum of disturbance to the embryonic and fetal development resulting in all types of adverse effects.

Human data provide the most appropriate information for determining potential adverse effects of a compound. Human studies include epidemiologic studies and case reports. There are several types of epidemiologic studies; each has certain advantages and limitations. Cohort studies compare the results of clinical endpoints between exposed groups and unexposed groups to establish an association between a specific exposure and the outcome. Case-control studies compare the individuals with a particular disease to similar but nondiseased individuals on their exposure history, retrospectively. Because of ethical and practical considerations, dosing in humans to study adverse effects is excluded. Useful information for assessing human risks may be derived from human studies with adequate design and statistical power. However, the information for potential effects on humans is generally obtained using laboratory animal experimental models. Substantial evidence indicates that agents which have been associated with human effects can also be associated with adverse effects in animals.

The views presented in this paper are those of the author and do not necessarily represent those of the U.S. Food and Drug Administration.

Reproductive Studies

Reproductive studies evaluate the potential adverse effects of test compounds on reproductive systems of both adult males and females as well as the on the postnatal maturation and reproductive capacity of the offspring, and the cumulative effects on generations. The studies are designed to evaluate: (1) the effects on male and female fertility; (2) the effects during gestation on the mother and on the fetus; (3) the effects appearing after parturition on mother's lactation and on offspring growth, development, and sexual maturation; and (4) the mutagenic effects through several generations.

Developmental Studies

Developmental studies evaluate the effects of test compounds on a developing organism that result from the exposure of parent(s), during perinatal development, or from postnatal development to the time of sexual maturation. The studies are designed to evaluate: (1) the death of the developing organism, (2) structural malformations and variations, (3) altered growth, and (4) functional impairment.

REGULATORY REQUIREMENTS

The U.S. Food and Drug Administration (FDA) and other countries require that new drugs and certain medical devices must be approved for safety and effectiveness for their intended use before being marketed. Food additives, color additives, compounds for use in food-producing animals, and pesticide residues in foods must be shown to be safe for the proposed use. In 1966, the U.S. FDA^[1] published guidelines for the performance of reproductive and developmental tests of drugs in animals. Three segments of study are required in preclinical animal testing for each new drug depending on how women might be exposed to the

drug. These are referred to as Segment I (fertility and general reproductive performance), Segment II (developmental effects), and Segment III (perinatal and postnatal evaluations). These studies may be run more or less concurrently, and may be run in conjunction with chronic tests.

Three-Segment Design

The Segment I study is aimed at providing an overall evaluation of the effects of drugs on fertility in both sexes, the course of gestation, early and late stages of the development of the embryo and fetus, and postnatal development. The studies may be conducted by treating animals of only one sex and by mating with untreated animals of the opposite sex, or by treating both male and female animals. Segment II is primarily aimed at detecting teratogenic effects. The drug is given to the pregnant females during the period of organogenesis (e.g., days 6–15 for rats and mice, and days 6–18 for rabbits). The offspring are removed 1 or 2 days before term, and the corpora lutea, resorption sites, and live and dead fetuses are examined. Fetuses are weighed and examined for anomalies. Segment III is aimed at the evaluation the effects of drugs on the late stages of gestation and on parturition and lactation. The drug is given to pregnant females in the final one-third of gestation and continued throughout lactation to weaning (e.g., gestation day 15 to postnatal day 21 for rats or mice). The effects on duration of gestation are determined. Pup birth and developmental data including litter size, weight, and postnatal growth and mortality, along with impaired maternal behavior, are recorded and measured.

Most countries adopted the general principle of requiring the three segments of investigations in the preclinical testing of new drugs. Recently, the International Conference on Harmonization (ICH), representing the United States, the European Community, and Japan, published guidelines to address uniformity in conducting Segment I, II, and III studies.^[2–4] The ICH guidelines allow a standardization of protocols and a deletion of duplication of procedures with emphasis on flexibility in testing strategy. Testing for food additives, color additives, and animal drugs requires a Segment II developmental study with the incorporation of multigeneration reproductive studies.^[5–8] Testing requirements for environmental agents such as pesticides and industrial chemicals, regulated by the U.S. Environmental Protection Agency,^[9,10] are similar to those for food additives; that is, a standard Segment II developmental study and two-generation reproductive studies are required. In general, testing protocols for pharmaceutical and nonpharmaceutical chemicals are similar; however, the data for nonpharmaceutical chemicals would be used to establish an acceptable level of exposure using quantitative risk assessment.

When a drug has been approved for marketing by the FDA, the information for the safe and effective use of the drug must be on the label. The FDA^[11] published a

uniform labeling format for human prescription drugs. The pregnancy labeling provides information with regard to use during pregnancy for all drugs. The FDA has devised five categories to describe levels of potential risk to human pregnancy based on findings from animal and human studies.^[11,12] Category A is to be used for drugs where adequate, well-controlled studies in pregnant women have failed to demonstrate risk to fetus. Category B is used for those drugs for which no controlled studies have been conducted in human pregnancy, and animal studies have not indicated the potential for human developmental risk. Alternatively, animal studies may indicate some risk but controlled human studies do not. Category C is used for drugs where there are no controlled studies in human pregnancy and animal studies demonstrate an adverse effect, or have not been conducted. Potential benefits, however, may justify the potential risks. Category D is for drugs where there is a positive evidence of human developmental risks. Investigational and postmarketing surveillance data may show human developmental risks, but potential benefits may outweigh the human risks. Category X is used for agents where there is a documented evidence of human development toxicity from reports in the literature or postmarketing surveillance. The risk to human conceptus is considered to outweigh any possible benefit.

STATISTICAL DESIGN/METHOD/ANALYSIS

Experimental Design

Mice, rats, and rabbits are the most commonly used species for reproductive and developmental studies. The various guidelines generally require testing in a rodent and nonrodent species, usually rats and rabbits. An experiment typically consists of an untreated control and three dose groups. The highest dose is chosen to produce minimal maternal toxicity, ranging from marginal body weight reduction to not more than 10% mortality. The low dose should be a no-observed-adverse-effect level for both maternal and developmental toxicity. The midlevel dose usually is halfway between the low dose and the high dose. A fourth dose group may be added to avoid excessive dosage intervals and to assess dose–response relationships. Sometimes doses are based on known or expected human therapeutic or toxic levels. Typically, dosage is measured in milligrams per kilogram body weight per day. The U.S. regulatory guidelines generally recommend about 20 pregnant rodents and 15 nonrodent animals per dosage group. The ICH guideline recommends the use of 16–20 pregnant animals. Typical litter sizes (number of viable offspring) for the control animals range from 8 for rabbit to 12 and 14 for mice and rats, respectively.

Depending on the design, a test compound is administered to either parent prior to conception, and to the female

during prenatal development, and postnatally. Animals can be assigned to dose groups totally randomly or be assigned by stratified body weight (but randomly within body weight classes). Each animal is monitored throughout the experiment to detect toxic and pharmacologic effects of the test compounds. In a Segment II study, the pregnant dams normally are sacrificed just prior to term, and fetuses are then examined to assess developmental effects. In a Segment III study, females are allowed to deliver and rear their offspring. Postnatal functional evaluations are performed. All adult animals are necropsied at terminal sacrifice.^[3,13]

During the reproductive process, ovarian follicles release eggs, which are fertilized by sperms to form embryos. These embryos subsequently implant into the wall of the uterus. The discharged ovarian follicles differentiate into corpora lutea, which secrete hormones necessary to sustain pregnancy. The embryos undergo organogenesis to form fetuses. Each step in this process is a potential target for toxic effects of a test compound. Depending on when treatment begins, a test compound may affect the number of ovulating follicles, the number of corpora lutea, the number of embryos implanting into the wall of the uterus, or the number of live fetuses surviving to term. For surviving fetuses, growth retardation may occur, or fetuses may exhibit one or more types of structural malformations. Fig. 1 is a sketch of the outcomes from a dam in the developmental and reproductive toxicity experiment.

Regulatory requirements specify that a wide range of endpoints must be measured, recorded, and analyzed. The endpoints can be divided into two categories: parental and embryonic/fetal endpoints. Parental endpoints to assess the reproductive effect include: body weight

and weight gain, mating index, fertility index, changes in gestation length, sperm count, numbers of corpora lutea, implantation sites, dead or resorbed implants, viable fetuses, and target organ weights, as well as food and water consumption. Embryonic/fetal endpoints include: individual malformations (external or gross, visceral, and skeletal) and variations in viable fetuses, number of normal fetuses, fetal weight, fetal length, sex ratio, preweaning and postweaning survival and body weight, physical development, and functional tests.^[2,3,9,10,13]

Experimental Unit

In the analysis of animal developmental endpoints, the experimental unit is the entire litter rather than an individual fetus. In the experiment, the test compound is administered to the adult animal; the primary variables of interest are fetal responses. The effect of the test compound occurs in the female that receives the compound, or that is mated to a male that receives the compound. The treatment affects the fetuses indirectly via the dam. Thus the development of individual fetuses in a dam is not independent. The fetal responses from the same dam are expected to be more alike than responses from different dams. This phenomenon is referred to as the “litter effect.” The proper experimental unit in the analysis should be the litter with the fetal responses representing multiple observations from a single experimental unit. The failure to account for the intralitter correlations by using each fetus as the experimental unit will inflate the Type I error and will reduce the validity of the test.

Sample Size

Sample size determination is an important issue in detecting a dose effect. The probability of detecting a statistically significant effect increases with the increased magnitude of effect and the increased sample size. It is customary to specify a range of the magnitudes of the effect of interest; an optimal sample size then can be calculated to provide a reasonable power of detecting the effect. Sample size calculations and references for various tests are provided in Desu and Raghavarao.^[14] It should be emphasized that sample size determination is preferable to arbitrary requirements of testing 10 or 20 animals, as indicated in various guidelines for assessing reproductive and developmental effects.

Statistical Analysis

Statistical analyses of various endpoints are composed of two aspects: qualitative testing for adverse effects and quantitative estimation for risk assessment. The qualitative testing is used to determine if there is a statistically significant difference among the groups. Three statistical tests are conducted: (1) an overall test for significant differences

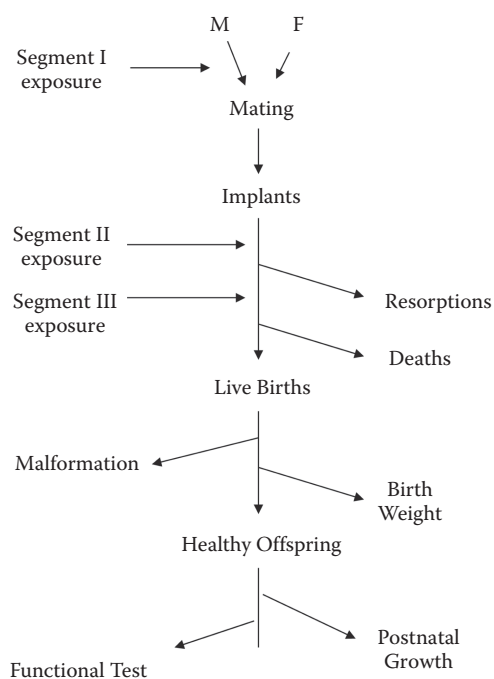


Fig. 1 Schematic representation of Segment I, II, III Studies

among groups, (2) pairwise comparisons between two specific groups (usually control vs. a dose group), and (3) a trend test to identify a dose-related increase or decrease. The trend test is more sensitive (powerful) to detect dose effect than the pairwise comparisons because it uses information from all dose groups. The quantitative estimation is used to determine a safe human level of exposure from an assessment of the dose–response relationship. The quantitative risk estimation typically does not apply to pharmaceutical drugs.

Methods

Endpoints are typically measured according to one of the three scales: continuous, count, and quantal (binary). Histological data such as severity or levels of achievement scores in behavioral testing may be recorded on the ordinal scale. Statistical methods appropriate for these different types of data are discussed separately.

Continuous Data

Continuous data such as body weights, organ weights, or behavioral measurements conducted on offspring following birth are measured on a continuous scale. The continuous endpoints are measured either at the litter level in an adult animal (e.g., maternal body weights) or at the individual fetus level (fetal body weights). Analysis of variance (ANOVA) is the most commonly used procedure for analysis of continuous data.^[15,16] The ANOVA procedure is used to assess the overall significant difference among groups. The ANOVA method assumes that data are independently and normally distributed with homogeneous variance. The logarithmic, square root, and arc–sine transformations are often applied to satisfy the normality assumption and to stabilize the variance. A simple one-way ANOVA analysis is the comparison of maternal endpoints among groups. Developmental endpoints traditionally are analyzed similarly but in terms of the average within each litter—a litter-based analysis.^[17] Nonparametric methods are used when the assumption of normality fails. The nonparametric analysis is initiated by ranking all observations of the combined groups. The nonparametric method is then applied to ranked data. Further details may be found in general texts addressing nonparametric methods.^[18]

There are a wide variety of post hoc tests available for pairwise comparisons or trend analysis after finding a significant result in an ANOVA. Dunnett's test^[19] and Williams' test^[20] are the two mostly used tests for comparisons of each of the dose groups to the control. Williams' test assumes an increasing (decreasing) dose–response trend. When this assumption is valid, Williams' test is more powerful in detecting differences between control and the dose groups. The Duncan multiple range test^[21] and the Tukey studentized range test^[22] are two frequently used procedures for all possible pairwise comparisons. Linear

regression is the simplest method to assess dose–response trends. But contrast tests, with a proper choice of contrast coefficients, in conjunction with ANOVA, are more commonly used for the dose–response trend test. Other trend tests are Tukey's test,^[23] which has been proposed to identify the highest dose at which there is no statistically significant trend, and Bartholomew's test of ordered alternatives.^[24]

ANOVA is also used in more complex experiments involving crossed (e.g., dose levels, replicates) and nested (e.g., litter effects) factors. The repeated-measures ANOVA is often used for the analysis of postnatal behavioral data.^[25] Multivariate analysis of variance (MANOVA) is used for a simultaneous analysis of several endpoints including repeated-measures data.^[25,26] A general approach to modeling developmental data can be carried out in terms of a mixed-effects model. Dempster et al.^[27] proposed a normal mixed-effects model with two levels of variance, in which litter effect is modeled by a nested random factor and dose is modeled by a fixed factor, for analyzing fetal weights.

Count Data

A number of primary reproductive endpoints are measured in counts. In a dominant lethal assay, male mice are treated with a suspect mutagen, and then are mated with females. The numbers of corpora lutea, implantations, lives, and dead conceptuses are counted to assess reproductive effects of the test compound. Count data are often normalized by the square root transformation; the transformed data are then analyzed as continuous data using the parametric ANOVA methods. Count data can also be analyzed by nonparametric methods.

Count data are generally modeled by a Poisson distribution. The mean of a Poisson is often expressed as a log linear function of dose and other covariates in the Poisson regression analysis. The Wald test, score test, or likelihood ratio test is used for group comparisons and trend test.^[28] A common complication in the analysis of count data is that the observed variation exceeds or falls below the variation that is predicted from a Poisson model. The negative binomial (gamma Poisson) distribution is a generalization of the Poisson model to account for the extra variation. Parametric and quasi-likelihood analyses of reproductive data with extra Poisson variation are given by Chen.^[29]

Binary Data

The binary endpoints, as continuous endpoints, can be measured either at the parent level, such as success or failure of pregnancy, or at the individual fetal level, such as presence or absence of a particular malformation type. Statistical methods for the analysis of the prenatal and fetal responses are different.

Two common approaches for the analysis of prenatal binary endpoints are the asymptotic tests and exact tests

(permutation tests). Asymptotic tests include the Pearson chi-square test and likelihood ratio chi-square test for the comparisons of the incidence rates among several groups.^[30] A continuity correction is often applied if only two groups are compared. The Cochran–Armitage test is used to test for trend.^[31,32] The Fisher exact test is the best known permutation test for comparing two groups. The permutation test computes the values of a “proper test statistic” from the observed data for all permutations (sampling without replacement) under the null hypothesis. The observed p value can be obtained by comparing the value of the observed test statistic with the empirical distribution. Ciminera^[33] proposed a randomization test for dose–response trend. General computational algorithms and software to perform all possible permutations are given by Mehta et al.^[34]

Because the litter rather than the fetus is the basic experiment unit, the analysis of fetal endpoints should account for the litter effect. The general approach to the analysis of developmental binary endpoints is to consider the proportion per litter such as the proportion of live fetuses with a certain type of malformation. Two approaches are commonly used for the analysis of proportions: litter-based analysis and extra binomial modeling. Both approaches assume that the denominator (“litter size”) is fixed.

Litter-based analysis has been widely used in the analysis of developmental data. Typically, the proportions are transformed by an arc-sine transformation, and then the parametric ANOVA methods are used. The litter-based approach does not use the data effectively because it does not account for the litter size. For example, one of two is treated the same as 5 of 10. Extra binomial modeling has been an important development in the methodology for the analysis of developmental data. The beta-binomial model and quasi-likelihood method have been well studied for the analysis of developmental data.^[28,35,36] The mean of the proportion is often expressed in terms of a logistic function. The Wald test, score test, or likelihood ratio test is used for group comparisons and trend test in the analysis with the beta-binomial model. The parametric and quasi-likelihood analyses of Segment II developmental data have been given by Chen.^[29]

Analysis

The commonly used litter-based method is presented in the data analysis. A typical experiment consists of g groups, a control, and $g - 1$ dose groups. Assume that the i th group contains m_i female animals. Let y_{ijk} be the response from a fetus out of n_{ij} examined or tested for a particular developmental outcome, $1 \leq i \leq g$, $1 \leq j \leq m_i$, and $1 \leq k \leq n_{ij}$. Note that y_{ijk} may be an indicator variable representing the presence or absence of a particular malformation type or a continuous variable representing a fetal weight or postnatal performance measurement. Depending on the endpoint of interest, n_{ij} may represent the number of viable

fetuses, number of implants, or number of measurements. The litter-based analysis is based on the per-litter response $y_{ij} = \sum_k y_{ijk} / n_{ij}$. For a continuous response, y_{ij} will represent the mean fetus response; for a discrete variable, it will represent the sum of the fetal responses. The y_{ij} can be viewed and analyzed as a maternal endpoint.

A statistical analysis of a study of effects of maternal exposure to diethylhexyl phthalate (DEHP) in rats is presented here.^[37] Diethylhexyl phthalate was administered in the diet on gestational days 0 through 20 at 0.00%, 0.5%, 1.0%, 1.5%, and 2.0%. The study was conducted with two replicates. Approximately equal numbers of dams were assigned to each group, and gestational day 0 body weights were matched across dose within each replicate. At gestational day 20, animals were sacrificed and all fetuses were examined.

A two-way (dose $\times$ replicate) ANOVA model was used for the analysis of continuous endpoints or per-litter proportion data. Prior to analysis, the arc-sine square root transformation was performed to all maternal or per-litter proportion data. A test of linear dose–response trend was performed using contrast tests. When a significant ($p < 0.05$) dose effect occurred, Duncan’s multiple range test was used for pairwise comparisons between control and each dose group. A one-sided test was used for all pairwise comparisons except for the maternal and fetal body weights and percentage of males per litter. Nominal (incidence) data were analyzed by a chi-square test for differences among groups, and by a Cochran–Armitage linear trend test on proportions. A one-sided Fisher’s exact test was used for pairwise comparisons when the chi-square test was significant. Table 1 contains a summary of the analysis for selected endpoints. For a detailed analysis, the reader is referred to Tyl et al.^[38]

SCIENTIFIC ISSUES

The main objectives in the analysis of reproductive and developmental experimental data are group comparisons and dose–response modeling. The statistical analysis of developmental endpoints presents a number of important issues. Developmental endpoints consist of both continuous and discrete outcomes. The analysis needs to account for the correlations induced by the clustering effects within litters. Tests for group differences for continuous endpoints are commonly conducted using the ANOVA with litter nested, MANOVA, or mixed-effects models. In contrast to the multivariate normal distributions for multiple continuous outcomes, distributions for the analysis of multiple binary outcomes are modeled in terms of sums of correlated binary variables under the exchangeability assumption. The resulting distribution is a generalized binomial. The distribution for the common developmental endpoints, such as the number of malformations per litter, often has an overdispersed binomial, but the distribution of sex combinations can be an underdispersed binomial.^[39]

Table 1 Reproductive parameters after the exposure of pregnant Fisher 344 rats to DEHP in the feed on gestational days 0–20

Endpoint	DEHP percent in feed				
	0.0	0.5	1.0	1.5	2.0
Number of pregnant dams	24	23	22	24	25
Maternal weight, gestational day 0	173.60 (3.25)	175.37 (3.37)	172.42 (3.12)	171.77 (2.80)	173.33 (2.95)
Maternal weight, gestational day 20	248.30 (3.64) ^a	246.00 (2.90)	232.73 (3.25) ^b	217.76 (3.25) ^b	184.15 (4.28) ^b
Maternal weight gain	74.69 (1.71) ^a	70.63 (2.48)	60.31 (1.44) ^b	45.99 (1.78) ^b	10.82 (3.00) ^b
Gravid uterine weight	49.79 (1.39) ^a	44.83 (2.49)	46.81 (1.31)	43.86 (0.96)	19.08 (3.71) ^b
Maternal liver weight	9.75 (0.12) ^a	11.72 (0.17) ^b	12.06 (0.17) ^b	12.21 (0.19) ^b	11.11 (0.15) ^b
Food consumption	312.04 (4.47) ^a	282.15 (6.75) ^b	252.71 (5.68) ^b	211.00 (5.04) ^b	179.48 (4.77) ^b
Water consumption	465.18 (7.47) ^a	495.47 (9.70)	515.66 (13.0) ^b	462.64 (9.35) ^b	376.76 (8.51) ^b
Corpora lutea	10.91 (0.33)	10.96 (0.22)	11.18 (0.30)	11.17 (0.18)	10.52 (0.61)
Implantation sites	10.92 (0.31)	9.83 (0.54)	10.59 (0.24)	10.58 (0.22)	10.40 (0.41)
Percent preimplantation loss	3.57 (1.14)	11.59 (4.20)	5.35 (1.28)	5.37 (1.48)	11.30 (2.84)
Percent resorption	4.14 (1.22) ^a	4.17 (1.20)	4.92 (1.67)	4.70 (1.25)	54.48 (8.10) ^b
Percent nonlive	4.14 (1.22) ^a	4.50 (1.29)	4.92 (1.67)	4.70 (1.25)	56.78 (8.04) ^b
Percent viable	10.46 (0.32) ^a	9.39 (0.53)	10.05 (0.26)	10.08 (0.25)	8.00 (0.70) ^b
Percent male	53.56 (3.65)	45.77 (4.51)	46.39 (3.31)	54.44 (3.00)	50.66 (6.10)
Percent malformation	1.27 (0.71) ^a	0.00 (0.00)	1.92 (1.11)	3.13 (1.06)	2.87 (1.64)
Fetal weight	3.022 (0.029) ^a	3.143 (0.035) ^b	2.852 (0.053) ^b	2.557 (0.034) ^b	2.266 (0.041) ^b

^aSignificance in linear trend.^bSignificantly different from control.

The beta-binomial distribution includes a non-negative dispersion parameter to model intralitter variations. The beta-binomial model uses a full likelihood-based analysis. This model has disadvantages in that it does not allow for negative intralitter correlations, and its maximum likelihood estimates can be biased or numerically unstable.^[40,41] The quasi-likelihood method offers an alternative approach; the quasi-likelihood estimates are generally more stable than the beta-binomial maximum likelihood estimates. The quasi-likelihood approach does not provide inference on the dispersion parameter. Liang and Zeger^[42] proposed a general method for the analysis of correlated data. This method is often referred to as generalized estimating equation (GEE). The analysis of continuous as well as discrete developmental endpoints with the GEE approach has been described by Dempster et al.^[27]

In a typical reproductive and developmental study, various endpoints are collected in order to obtain the maximum information from the study. Statistical analysis often involves a large number of tests, both in terms of multiple comparisons among groups and multiple tests on many endpoints. Because of the conduct of a large number of statistical tests, the chance of false positive findings could increase considerably. Commonly, each endpoint is analyzed separately. A significant result for an endpoint tested may be statistical, biological, or both. The issues of false positive and false negative error rates in the testing

of multiple reproductive and developmental endpoints has not been addressed. The main consideration should be to maximize the power of identifying every significant endpoint while controlling the overall Type I or family-wise error rate.

CONCLUSION

Recently, several authors have proposed multivariate models for simultaneous analysis of multiple endpoints. Joint modeling of multiple developmental outcomes can have some advantages: It can increase the power of detecting effects if the multiple outcomes are manifestations of some common biological effect, and it allows investigations of associations among the multiple outcomes if they are the result of different biological mechanisms.

In the analysis of developmental data, two types of correlation are inherent: the (intralitter) correlation between fetuses of the same litter and the (intrafetus) correlation between the endpoints in the same fetus. A joint modeling of multiple binary outcomes of different malformation types has been proposed by Lefkopoulou et al.^[43] and Chen and Ahn.^[44] A joint modeling of two overdispersed binomial outcomes for the numbers of deaths/resorptions and malformations has also been proposed.^[45–48] Recently, several authors have proposed conditional approaches to modeling malformation and fetal weight.^[49–51] However, a joint

modeling of continuous and discrete outcomes within a cluster is not straightforward, and requires further research.

Likelihood-based and GEE approaches are developed for dose–response modeling of univariate or multivariate reproductive endpoints. The likelihood-based approach requires strong assumptions on the distribution and the variance–covariance structures. The maximum likelihood estimates may be computationally unstable. The GEE^[42] and GEE2^[52] methods can be used to estimate the mean and variance/covariance parameters. Examples of dose–response analysis of reproductive toxicity data using the GEE approach are given in Lefkopoulou et al.^[43] and Chen and Ahn.^[44] The software for the GEE method is commercially available.^[53,54]

REFERENCES

1. U.S. Food and Drug Administration. *Guidelines for Reproductive Studies for Safety Evaluation of Drugs for Human Use*; Bureau of Drugs, Food and Drug Administration: Rockville, MD, 1966.
2. U.S. Food and Drug Administration. International Conference on Harmonisation. *Guideline on Detection of Reproduction for Medicinal Products*; Food and Drug Administration: Rockville, MD, 1966.
3. Christian, M.S.; Hoberman, A.M. *Handbook of Developmental Toxicology*; Hood, R.D., Ed.; CRC Press: New York, 1996, 551–595.
4. Manson, J.M. *Developmental Toxicology*, 2nd Ed.; Kimmel, C.A.; Buelke-Sam, J., Eds.; Raven Press: New York, 1996, 379–402.
5. Fitzhugh, O.G. *Modern Trend in Toxicology*; Boyland, E.; Goulding, R., Eds.; Butterworth: London, 1968, Vol. 1, 75–85.
6. Collins, T.F.X. *Handbook of Teratology*; Wilson, J.G.; Fraser, F.C., Eds.; Plenum: New York, 1978, Vol. 4, 191–214.
7. U.S. Food and Drug Administration. Toxicological Principles and Procedures for Priority Based Assessment of Food Additives (Red Book). *Guidelines for Reproduction Testing with a Teratology Phase*; Bureau of Foods, Food and Drug Administration: Washington, DC, 1982.
8. U.S. Food and Drug Administration. Toxicological Principles for Safety Assessment of Direct Food Additives and Color Additives Used in Food (Red Book II). *Guidelines for Reproduction and Developmental Toxicity Studies*; Center for Food Safety and Applied Nutrition: Washington, DC, 1993.
9. U.S. Environmental Protection Agency. Guidelines for the health assessment of suspect developmental toxicants. Fed. Regist. **1986**, *51*, 34028–34040.
10. U.S. Environmental Protection Agency. Guidelines for developmental toxicity risk assessment. Fed. Regist. **1991**, *56*, 63798–63826.
11. U.S. Food and Drug Administration. Prescription drug advertising; Content and format for labeling human prescription drugs. Fed. Regist. **1979**, *44*, 37434–37467.
12. *Physicians' Desk Reference*; Medical Economics Company: Oradell, NJ, 1992.
13. Tyl, R.W.; Marr, M.C. *Handbook of Developmental Toxicology*; Hood, R.D., Ed.; CRC Press: New York, 1996; 175–225.
14. Desu, M.M.; Raghavarao, D. *Sample Size Methodology*; Academic Press: New York, 1990.
15. Snedecor, G.W.; Cochran, W.G. *Statistical Methods*, 7th Ed.; Iowa State University Press: Ames, IA, 1980.
16. Winer, B.J.; Brown, D.R.; Michels, K.M. *Statistical Principles in Experimental Design*, 3rd Ed.; McGraw-Hill: New York, 1991.
17. Healy, M.J.R. Animal litters as experimental units. Appl. Stat. **1972**, *21*, 155–159.
18. Lehman, E.L. *Nonparametrics: Statistical Methods Based on Ranks*; Holden-Day: San Francisco, CA, 1975.
19. Dunnett, C.W. J. Am. Stat. Assoc. **1955**, *50*, 1096–1121.
20. Williams, D.A. Biometrics **1972**, *28*, 519–531.
21. Duncan, D.B. Biometrics **1955**, *11*, 1–42.
22. Tukey, J.W. 1953; unpublished manuscript.
23. Tukey, J.W.; Ciminera, J.L.; Heyse, J.F. Biometrics **1985**, *41*, 295–301.
24. Barlow, R.E.; Bartholomew, D.J.; Bremner, J.M.; Brunk, H.D. *Statistical Inference Under Order Restrictions*; John Wiley and Sons: New York, 1972.
25. Karpinski, K.F. *Statistics in Toxicology*; Krewski, D.; Frankin, C., Eds.; Gordon and Breach Science: New York, 1991, 393–434.
26. Rencher, A.C. *Methods of Multivariate Analysis*; John Wiley and Sons: New York, 1995.
27. Dempster, A.P.; Selwyn, M.R.; Patel, C.M.; Roth, A.J. Appl. Stat. **1984**, *33*, 203–214.
28. McCullagh, P.; Nelder, J.A. *Generalized Linear Model*, 2nd Ed.; Chapman and Hall: London, 1989.
29. Chen, J.J. *Design and Analysis of Animal Studies in Pharmaceutical Development*; Chow, S.C.; Liu, J.P., Eds.; Marcel Dekker: New York, 1998.
30. Agresti, A. *Categorical Data Analysis*; John Wiley and Sons: New York, 1990.
31. Cochran, W.G. Biometrics **1954**, *12*, 417–451.
32. Armitage, P. Biometrics **1955**, *11*, 375–386.
33. Ciminera, J.L. *Proceedings of the Symposium on Long-Term Animal Carcinogenicity Studies: A Statistical Perspective*; American Statistician Association: Washington DC, 1985, 26–35.
34. Mehta, C.R.; Patel, N.R.; Senchaudhuri, P. J. Comput. Graph. Stat. **1992**, *1*, 21–40.
35. Williams, D.A. Biometrics **1975**, *31*, 949–952.
36. Williams, D.A. Appl. Stat. **1982**, *31*, 144–148.
37. Tyl, R.W.; Price, C.J.; Marr, M.C.; Kimmel, C.A. Fundam. Appl. Toxicol. **1988**, *10*, 395–412.
38. Tyl, R.W.; Price, C.J.; Marr, M.C.; Kimmel, C.A. *Teratologic Evaluation of Diethylhexyl Phthalate in Fisher 344 Rats*; Research Triangle Institute: Research Triangle Park, NC, 1983.
39. Brook, R.J.; James, W.H.; Gray, E. Biometrics **1991**, *47*, 403–417.
40. Kupper, L.L.; Portier, C.; Hogan, M.D.; Yamamoto, E. Biometrics **1986**, *42*, 85–89.
41. Williams, D.A. Biometrics **1988**, *44*, 305–308.

42. Liang, K.Y.; Zeger, S.L. *Biometrika* **1986**, *73*, 13–22.
43. Lefkopoulou, M.; Moore, D.; Ryan, L.M. *J. Am. Stat. Assoc.* **1989**, *84*, 810–815.
44. Chen, J.J.; Ahn, H. Marginal models for analysis of binomial data with overdispersion. *J. Agric. Biol. Environ. Stat.* **3**.
45. Chen, J.J.; Kodell, R.L.; Howe, R.B.; Gaylor, D.W. *Biometrics* **1991**, *47*, 1049–1058.
46. Chen, J.J.; Li, L.A. *Stat. Sin.* **1994**, *4*, 265–274.
47. Ryan, L.M. *Biometrics* **1992**, *48*, 163–174.
48. Zhu, Y.; Krewski, D.; Ross, W.H. *Appl. Stat.* **1994**, *48*, 583–598.
49. Catalano, P.J.; Ryan, L.M. *J. Am. Stat. Assoc.* **1992**, *87*, 651–658.
50. Chen, J.J. *Risk Anal.* **1993**, *13*, 559–564.
51. Fitzmaurice, G.M.; Laird, N.M. *J. Am. Stat. Assoc.* **1995**, *90*, 845–852.
52. Prentice, R.L. *Biometrics* **1988**, *44*, 1033–1048.
53. SAS Institute. *SAS/STAT Version 8*; SAS Institute Inc: Cary, NC, 1999.
54. Cytel Software Corporation. *ToxTools*; 1999, Cambridge, MA.

Response Surface Methodology

William R. Myers

The Procter & Gamble Company, Mason, Ohio, U.S.A.

Abstract

Response surface methodology (RSM) is defined as a collection of mathematical and statistical methods that are used to develop, to improve, or to optimize a product or process. This entry works to provide a detailed definition of this type of methodology by presenting several design examples.

Keywords: Mathematical methods; Response surface methodology; Statistical methods.

INTRODUCTION

Response surface methodology (RSM) is defined as a collection of mathematical and statistical methods that are used to develop, to improve, or to optimize a product or process. It comprises of statistical experimental designs, regression modeling techniques, and optimization methods. Most applications of RSM involve experimental situations in which several independent variables (or factors) potentially influence one (or more) response variable. The independent variables are controlled by the experimenter, in a designed experiment, whereas the response variable is an observed output of the experiment. Fig. 1 illustrates the estimated relationship between a response variable and the two independent variables x_1 and x_2 .

In many applications of RSM, a sequential process is performed. At the start, a researcher may have numerous independent variables that are being studied. In order to determine which of the factors have an effect on the response variable, a *screening design* is often performed. This will potentially reduce the number of factors that need to be investigated in further experimentation. A researcher hopes to eliminate unimportant factors so that ensuing experiments (e.g., second-order response surface design) will be less costly and more efficient. In addition, an existing screening design could also be augmented with additional design points in order to estimate a second-order (or response surface) model. Another RSM tool is the *method of steepest ascent*. This is an optimization technique that will allow one to move iteratively toward an optimum set of experimental conditions. This is often implemented when the researcher is initially experimenting in a suboptimal region. It would not be very prudent to invest in a large—and perhaps costly—experiment unless one feels that the *region of optimum* is either inside the design region or very close to the periphery. After a *region of interest* has been defined, a response surface design is used to estimate a second-order model. This provides an approximation of

the true response surface over the region of interest, thus providing a better overall understanding of the estimated response surface and to determine the optimum operating conditions. RSM has been used extensively in many areas of biopharmaceutical research such as drug formulation, process development, and drug discovery. Many applications of RSM in biopharmaceutical research will be presented throughout the discussion.

MODEL BUILDING

A statistical model is used in order to estimate the relationship between the response variable and the independent variables. The estimated statistical model is called a *response surface model*. Regression techniques are used to estimate the response surface model. A first-order or second-order model is typically used to characterize this relationship. A first-order model assumes that each independent variable has a linear effect on the response variable. A first-order model is defined as: $Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \epsilon$ where Y is the response variable, β_0 is the model parameter that represents the intercept, β_i is the model parameter associated with the i th independent variable, x_i represents the level of the i th independent variable, and ϵ represents the random error. The parameter β_i represents the change in the response variable Y per unit change in x_i , while holding all the other variables constant. In many instances, an interaction between two independent variables is present. In those situations, the first-order model will include interaction terms as follows: $Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{ij} \beta_{ij} x_i x_j + \epsilon$ where β_{ij} is the model parameter that represents the interaction between the i th and the j th independent variable. When β_{ij} is present, the effect of the i th independent variable on the response variable Y depends on the level of the j th independent variable. The first-order model is typically used in cases where no curvature is present in the response surface, when the region of interest is narrow, or when the researcher

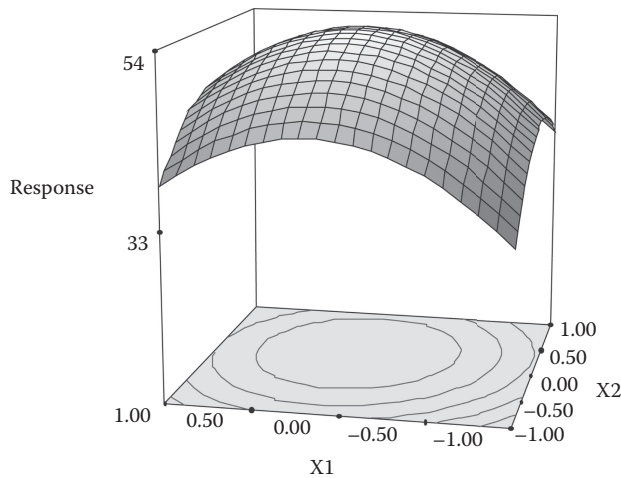


Fig. 1 An example of a response surface

attempts to reduce the number of factors for further experimentation through variable screening. In those cases where the aforementioned does not apply, a second-order model is commonly used. A second-order model assumes that the independent variables have a quadratic (or curvature) effect on the response variable. The second-order model is defined as $Y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i=1}^k \beta_{ii} x_i^2 + \sum_{ij} \beta_{ij} X_i X_j + \varepsilon$ where β_{ii} is the model parameter that represents the quadratic effect for the i th independent variable. The second-order model is extensively used in RSM. Myers, Montgomery, and Anderson-Cook^[1] point out that the second-order model is very flexible, in that it can approximate a variety of functional forms. Practical experience indicates that the second-order model works well in solving many real response surface problems.

The parameters for the previously discussed statistical models are estimated by the method of *ordinary least squares*. The parameter estimates are often called *regression coefficients*. Using matrix notation, the least squares estimator of the model parameter β is: $\hat{\beta} = (X'X)^{-1} X'y$ where X is an $(n \times p)$ model matrix and y is an $(n \times 1)$ vector of observed responses. All the standard techniques that are routinely used for model building in multiple linear regression are applied. This includes testing for significance, confidence intervals of the regression coefficients, and checking model adequacy. It is customary to use coded variables in RSM. The coded variable is computed in the following manner: $x_{ij}^c = x_{ij} - \bar{x}_j / x_{hi(j)} - \bar{x}_j$ where x_{ij} is the value in the natural units, $\bar{x}_j$ is the mean value for the j th variable, and $x_{hi(j)}$ is the maximum value in the natural units for the j th variable. This results in the values of x_{ij}^c ranging from -1 to $+1$. One reason for using coded variables is that the magnitude of each regression coefficient represents the change in the response variable for one unit change in the independent variable. Therefore, the impact of each independent variable can be evaluated in an equivalent manner. To read more about linear regression and model building, see Myers^[2] or Draper and Smith.^[3]

2^k FACTORIAL DESIGNS

The most commonly used experimental design when several factors are being studied is a factorial design. A more specific group of factorial designs is the 2^k factorial design, where each of the k factors has only two levels. The two levels, for each factor, represent a low ($-$) and a high ($+$) level and thus define the region of interest. An example of a 2^2 factorial design is shown in Table 1, with a graphical description in Fig. 2. A 2^k factorial design can be used when there is knowledge that no curvature is present in the experimental region. This will often be applicable when studying a narrow *region of interest*. A 2^k factorial design can naturally be extended to create a second-order response surface design which will be discussed later. A 2^k factorial design is very applicable in the case of *screening designs*, where a researcher has numerous factors to study and needs to eliminate those factors that do not have an effect on the response variable. Those factors, which are found to be important, would be studied further in a second-order response surface design.

FRACTIONAL FACTORIAL DESIGNS

If the number of factors in a 2^k factorial design is relatively large, then the size of the experiment may be impractical. For example, a 2^7 factorial design would consist of 128 experimental runs. One way to combat this problem is to perform only a specific fraction of the factorial experiment. These types of designs are appropriately called *fractional factorial designs*. Fractional factorial designs are very practical in the case of screening experiments. A key assumption when implementing a *fractional factorial*

Table 1 A 2^2 Factorial Design

Experimental run	X_1	X_2
1	$-$	$-$
2	$-$	$+$
3	$+$	$-$
4	$+$	$+$

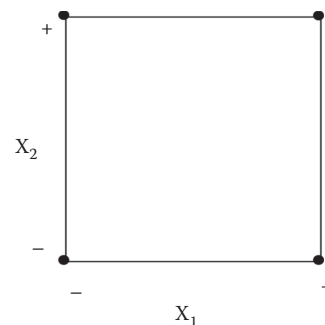


Fig. 2 A 2^2 factorial design

design is that higher-order interactions are negligible. Let us consider a simple example where a researcher has three factors of interest and only four experimental runs can be performed. In addition, there is the belief that two-way interactions are not present. In this case, a one-half fraction of a 2^3 factorial design (i.e., 2^{3-1} design) would be a reasonable choice. Fig. 3 provides a graphical description of this design. The following algorithm can be used in order to determine which experimental runs are selected to create an appropriate one-half fraction of a 2^k factorial design:

- Select a 2^{k-1} full factorial design.
- Then the k th factor (or additional column) of the design is generated by multiplying the (+) and (−) of each of the $k - 1$ columns.

In most cases a users will not need to rely on the aforementioned algorithm, as many statistical software packages can generate fractional factorial designs.

Table 2 provides the one-half fraction of a 2^3 factorial design. Because only a fraction of the complete factorial design is performed, the ability to differentiate between some of the factor effects will be sacrificed. When a factorial design is fractionated, some of the factor effects have the same estimate. Two or more effects that have identical estimates are said to be *aliased* with each other. In the one-half fraction of a 2^k factorial design, each effect has one alias. By observing Table 3, it should be apparent that the column for X_1 is identical with X_2X_3 (also $X_2 = X_1X_3$ and $X_3 = X_1X_2$). Therefore one cannot differentiate between the effect of X_1 and X_2X_3 . In most practical fractional factorial designs, higher-order interactions are aliased with main effects or lower-order interactions. This is why the assumption of negligible high-order interactions is vital when implementing a fractional factorial design.

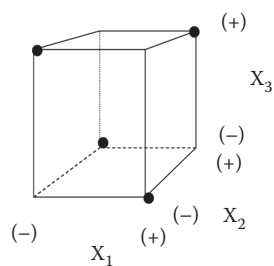


Fig. 3 One-half fraction of a 2^3 factorial design

Table 2 One-half Fraction of the 2^3 Factorial Design

Experimental run	X_1	X_2	$X_3 = X_1X_2$
1	−	−	+
2	−	+	−
3	+	+	+
4	+	−	−

Table 3 Confounding Factors in One-half of a 2^3 Factorial Design

Experimental run	X_1	X_2	X_1X_2	X_1X_3	X_1X_3	X_2X_3
1	−	−	+	+	−	−
2	−	+	−	−	+	−
3	+	+	+	+	+	+
4	+	−	−	−	−	+

Iskandarani et al.^[4] used a 2^{4-1} design with two center runs in order to optimize the capsule and tablet formulation for an experimental drug. A fractional factorial design was very appropriate in their situation, with the belief that higher-order interactions (e.g., three-way interactions) and several two-way interactions were expected to have little or no effect on the response variables.

There are some attractive flexible features of fractional factorial designs. First, they can be projected into a complete factorial design for any subset of $k - 1$ factors. Suppose a researcher used a one-half fraction of a 2^5 factorial design in order to screen out unimportant factors. If no more than four of the five variables are found to have a significant effect on the response variable, then a full factorial of the significant variables can be constructed. Thus, the main effects and each two-way interaction for each of the significant variables can be estimated. The utility of fractional factorial designs relies on the *sparsity-of-effects principle*, which states that a system is usually dominated by a few of the main effects and low-order interactions. If the number of variables under study becomes large, then it may be necessary to consider a smaller fraction of the 2^k design (e.g., one-quarter fraction of a 2^k design), which will result in more aliasing among effects. Myers, Montgomery, and Anderson-Cook^[1] Box and Draper,^[5] and Montgomery^[6] provide additional discussion on fractional factorial designs.

OTHER SCREENING DESIGNS (PLACKETT-BURMAN)

Another very popular class of screening designs is the Plackett-Burman design, which was developed by Plackett and Burman.^[7] These designs are two-level designs, in which the number of experimental runs is in multiples of four. Plackett-Burman designs fill the sample size void that exists with the fractional factorial designs. An example of a Plackett-Burman design with $N = 12$ and $k = 7$ is presented in Table 4. The alias structure (i.e., confounding of factors) for the Plackett-Burman is very complex; thus one must take great care when the estimation of interaction effects is of interest. Lin and Draper^[8] discuss projective properties of the Plackett-Burman design, which allows the experimenter to follow up an initial design with runs that allow an efficient separation of main effects and

Table 4 Plackett-Burman Design With $N = 12$ and $k = 7$

Experimental run	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	+	−	+	−	−	−	+
2	+	+	−	+	−	−	−
3	−	+	+	−	+	−	−
4	+	−	+	+	−	+	−
5	+	+	−	+	+	−	+
6	+	+	+	−	+	+	−
7	−	+	+	+	−	+	+
8	−	−	+	+	+	−	+
9	−	−	−	+	+	+	−
10	+	−	−	−	+	+	+
11	−	+	−	−	−	+	+
12	−	−	−	−	−	−	−

interaction effects. Sastry and Khan^[9] executed a Plackett-Burman design in order to evaluate the effect of formulation variables on drug release measurements. They were able to reduce the number of factors from seven to three, which significantly decreased the number of experimental runs required for the CCD. To read more on screening designs, see Myers, Montgomery, and Anderson-Cook,^[1] Box and Draper,^[5] or Montgomery.^[6]

CENTER RUNS IN RESPONSE SURFACE METHODOLOGY

A center run (or center point) is an experimental condition that is performed at the center of the design region $[(x_1, x_2, \dots, x_k) = (0, 0, \dots, 0)]$. Including (or augmenting) center runs to a 2^k factorial design has a very important role in RSM. The addition of center points provides the ability to determine if *curvature* (or quadratic effects) is present in the experimental region. This can formally be evaluated by a single-degree-of-freedom test. The mean square for *curvature* is the following: $n_f n_c (\bar{y}_f - \bar{y}_c)^2 / n_f + n_c$ where n_f and n_c represent the sample size for the factorial portion and center runs, respectively; and $\bar{y}_f$ and $\bar{y}_c$ represent the mean response for the factorial portion and center runs, respectively. If the mean response for the factorial design points is drastically different from the mean response for the center runs, then the mean square for *curvature* would be large and thus there would be evidence that curvature is present in the design region. Including center runs in a design will determine if curvature is present but cannot formally estimate and test the curvature effects for each factor. The test for curvature tests the following hypothesis:

$$H_0 : \sum_{j=1}^k \beta_{jj} = 0 \quad H_1 : \sum_{j=1}^k \beta_{jj} \neq 0$$

Sendrek et al.^[10] included center runs to a 2^4 factorial design in order to determine if curvature was present in their experimental region. By using the center runs, they were able to determine that a second-order design would

better approximate the response surface (a center composite design with six center runs was implemented). Their application involved the optimization of the turbidity and cloud point of a nonionic surfactant solution. Another advantage of implementing center runs in a 2^k factorial design is to get an economically appealing estimate of “pure error” that is independent of model selection. In other words, the experimental error is purely a function of replication, rather than potentially insignificant factors. When pure replication is not present in an experimental design, the estimate for experimental error is a function of insignificant factors.

STEEPEST ASCENT/DESCENT

A primary objective of RSM is to optimize a product or process. There are many instances when the researcher is not operating in the region of the optimum. After a screening experiment has been completed, a typical next step is to determine if one is close to the *region of the optimum*. The *method of steepest ascent* (or *descent*) can be used so that the experimenter can move along the path of maximum (or minimum) increase in the response variable. If one is trying to maximize the response, then the user will be looking for the steepest ascent, whereas if the user is trying to minimize the response, then the user will be looking for the steepest descent. Fig. 4 provides a contour plot, which demonstrates how the method of steepest ascent works. In this hypothetical example, the researcher would move along the path whereby Temperature is decreased and Speed is increased. The following are general steps for the method of steepest ascent:

1. Fit a first-order model (screening design).
2. Compute the path of steepest ascent (or descent), so one can expect the maximum increase (or decrease) in the response variable.

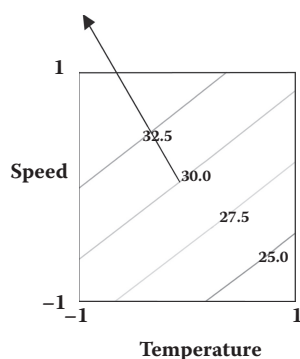


Fig. 4 Method of steepest ascent

3. Conduct experimental runs along the path and observe the response. At some point along the path, the improvement will decline and eventually disappear. The reason for this is because the first-order model, which was used to create the path, will no longer represent the true experimental system.
4. Select a base for a second experiment, where the maximum (or minimum) response is observed along the path. The second experiment could be either a factorial design or a response surface design. Testing for lack of fit (interaction terms) and using center runs for evaluating curvature are often useful at this stage.

The improvement found with steepest ascent will often not be as great at later iterations. In addition, one may not have the resources or time to perform numerous iterations of the method. However, after one has completed the method of steepest ascent, one has now developed a base for conducting a subsequent experiment (e.g., second-order response surface design) in order to optimize. Myers, Montgomery, and Anderson-Cook^[1] indicate that no more than two steepest ascent procedures are expected to be helpful.

The path of steepest ascent is a function of the regression coefficients in the first-order model. The magnitude of the regression coefficients determines the actual path, whereas the sign of the coefficients establishes the direction. The regression equation, which produced the contour in Fig. 4, was: $\hat{y} = 29.5 - 3.0 \times \text{Temperature} + 4.0 \times \text{Speed}$ where the design range for Temperature was 200–300°F and the design range for Speed was 300–500 rpm. Select one variable (e.g., Temperature) as the reference variable. Typically, one would move one coded unit along the path for the reference variable. Therefore one would move one coded unit in the negative direction of the variable Temperature. The ratio of the regression coefficients is used to calculate the points along the path for all other variables. In this example, the incremental steps for the Speed variable, in coded units, would be $(4/3) = 1.33$ in the positive direction. Table 5 provides the points on the path for both the coded and natural units. The method of steepest ascent has a very intuitive appeal in that one would expect an increase in the

Table 5 Points on the Path of Steepest Ascent

	Natural units		Coded units	
	Speed	Temperature	Speed	Temperature
Base	400	250	0	0
Delta	133	–50	1.33	–1
Base + delta	533	200	1.33	–1
Base + 2delta	666	150	2.66	–2

response variable by decreasing Temperature and increasing Speed, with Speed having a slightly greater impact.

EXPERIMENTAL DESIGNS FOR FITTING A SECOND-ORDER RESPONSE SURFACE

After a researcher has identified important variables via a screening experiment and has an understanding of the location of the region of the optimum (based on either steepest ascent/descent or scientific knowledge), it is time to consider implementing a second-order response surface design in order to be able to fit a second-order model and as a result better characterize the true response surface and find optimum experimental conditions.

One of the more popular designs for fitting a second-order response surface is the central composite design (CCD). These designs were introduced by Box and Wilson.^[11] The CCD consists of a 2^k factorial design, $2 \times k$ axial points, and center points. The CCD for three variables is displayed in Fig. 5. The flexibility of the CCD evolves around the selection of the axial points and the number of center points. The axial points can reside on the faces of the cuboidal region, which provides three levels per variable, or extend outside the faces of the cube, which provides five levels of each factor (as shown in Fig. 5). The choice of axial points is often dependent on the *region of interest*. For example, if the region of interest is defined as cuboidal, then it would be reasonable to place the axial points at the faces of the cube. Both Sastry and Khan^[9] (discussed earlier) and Lipps and Sakr^[12] used a CCD, with axial points located at the face of the cuboidal region. Lipps and Sakr used RSM in order to understand the effects of key process parameters in the top spray fluidized bed on measured granule and tablet properties.

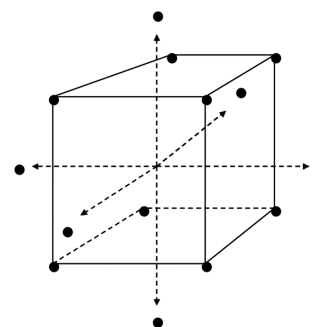


Fig. 5 A CCD for three variables

If the *region of interest* can be extended outside the cuboidal region (e.g., spherical), then the axial points can be extended outside the faces of the cube. Axial points typically range from 1.0 to $\sqrt{k}$, where $\sqrt{k}$ would represent a spherical design. Ogawa et al.^[13] used a spherical CCD in order to study the scale-up problem in the manufacturing of granules. The location of the axial points impacts the design property of rotatability. A design is rotatable if the prediction variance is constant for all points that are a common distance from the center of the design region. A CCD is rotatable if the axial point equal $\sqrt[4]{F}$, where F is the number of factorial design points. Having a stable prediction variance throughout the design region is certainly a desirable feature, given that a researcher may typically not know where one will want to predict. Branchu et al.^[14] implemented a four-factor rotatable CCD, with axial points that were two coded units from the center of the design, in order to model and to optimize the thermal stability of a model protein. The number of center runs also plays a key role in achieving the stability of the prediction variance throughout the design region. Myers, Montgomery, and Anderson-Cook^[1] indicate that for a spherical design, three to five center runs should be used, whereas for a cuboidal design, one to two center runs should be applied.

The CCD is very appealing because it allows sequential experimentation. For example, assume that a researcher is studying k variables via a screening design (e.g., 2^k factorial design or 2^{k-1} fractional factorial design) and based on the analysis only r variables ($r < k$) are found to be significant. Then $2r$ axial points and center points could be augmented to the original screening design in order to develop a CCD for r variables. Wehrle et al.^[15] demonstrated the sequential ability of RSM in order to determine which process factors affected the characteristics of the final produced granules and tablets. They began with a 2^2 factorial design and then used center runs to determine whether curvature was present in their experimental region. After it was determined that significant curvature was present, the original design was augmented with axial points at the face of the cube, which resulted in a CCD. Another biopharmaceutical application, which demonstrated the sequential nature of RSM, was presented by Hutchings et al.^[16] They used a 2^2 factorial design with center runs to determine which variables affected the response and if a second-order design would be necessary to approximate the response surface. Based on the analysis of their initial design, it was determined that a second-order design would more appropriately approximate the response surface. Axial points at $\alpha = \sqrt{2}$ and additional center runs augmented the factorial design in order to form a spherical CCD. This application involved understanding how cure time and curing temperature affected a commercially available ethylcellulose latex coating dispersion.

Another popular class of designs for fitting a second-order response surface are Box-Behnken designs (BBD), which were introduced by Box and Behnken.^[17] These designs

consist of three levels. Table 6 provides the design matrix for a Box-Behnken design with three variables and three center runs. A unique characteristic of the BBD is that it is spherical. Fig. 6 displays the BBD in the case of three variables. By comparing the geometric shapes of the BBD and the CCD, it should become apparent how to select between the two types of designs. Because the BBD is spherical in nature, it does not predict well at the extremes (or corners) of the design region. Therefore, the BBD should be limited to cases when predicting at the corners of the cube is not important. It is recommended that three to five center runs be used for a Box-Behnken design.

COMPUTER-GENERATED DESIGNS

There are instances when the classical response surface designs discussed in the previous section cannot be implemented in practice. Some possible reasons are as follows:

1. The number of experimental runs for the classical response surface designs are impractical

Table 6 Box-Behnken Design for Three Variables and Three Center Runs

X_1	X_2	X_3
+	+	0
−	+	0
+	−	0
−	−	0
−	0	−
+	0	+
+	0	−
−	0	+
0	−	−
0	+	+
0	−	+
0	+	−
0	0	0
0	0	0
0	0	0

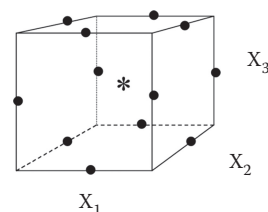


Fig. 6 A Box-Behnken design for three variables with center points

2. The design region under study is constrained, therefore classical response surface designs are not feasible
3. There are numerous qualitative variables being studied (qualitative variables are those factors that are not continuous in nature, e.g., type of solvent)

Computer-generated designs are based on various design-optimality criteria. *D-optimality*, which creates designs that do a good job of estimating model parameters, is the most popular design-optimality criterion. Another important design-optimality criterion is *I-optimality* (or *IV*). This criterion is used when prediction is of utmost importance. The user will need to provide several inputs (e.g., the assumed statistical model, sample size, range of the design variables, set of candidate points, design region constraints, and optimality criterion) in order to create a computer-generated design. The specific inputs for the user will depend on the statistical computer package. Another opportunity for using computer-generated designs is to augment (or improve) an existing experimental design with additional experimental runs. Statistical software packages like JMP®, SAS®, or Design-Expert® can create computer-generated designs. See Myers, Montgomery, and Anderson-Cook^[1] for a detailed discussion and many examples of computer-generated designs.

STATISTICAL ANALYSIS OF A RESPONSE SURFACE

After a researcher has completed the experimental work, often the goal is to locate the optimum operating condition (e.g., maximum, minimum, or target response). However, even more important is to truly understand the nature of the estimated response surface. A graphical representation of the estimate response surface is a vital part in understanding the response surface. Fig. 7 shows a 3-D plot of an estimate response surface, where a maximum estimated response is located in the experimental region. The maximum response is called a *stationary point*, which is where the derivatives of estimated response with respect to the design variables are all zero. There are situations where the stationary point is neither a maximum nor a minimum. In this particular case, the stationary point is classified as a *saddle point*. Fig. 8 illustrates an estimated response surface, where a saddle point exists. Another graphical method used to interpret a response surface is a contour plot. Figs. 9 and 10 illustrate contour plots in the case of a maximum response and a saddle point, respectively. Most of the applications of RSM mentioned previously used 3-D plots and/or contour plots to better understand how the factors affect the response and to locate possible optimum experimental conditions.

It can be more difficult to make clear interpretations of how the various factors are affecting the response

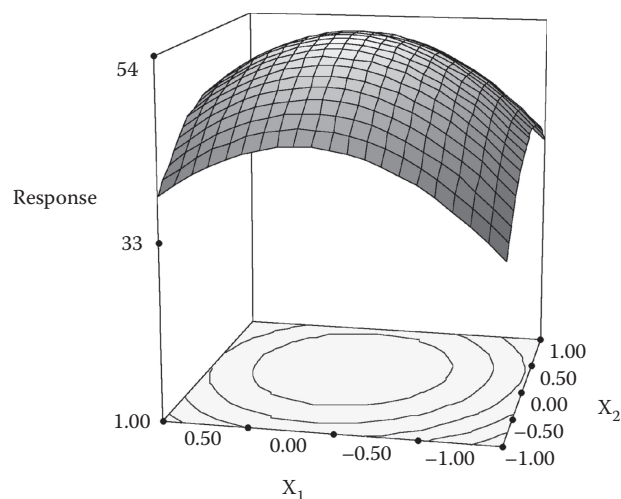


Fig. 7 A 3-D plot of a response surface

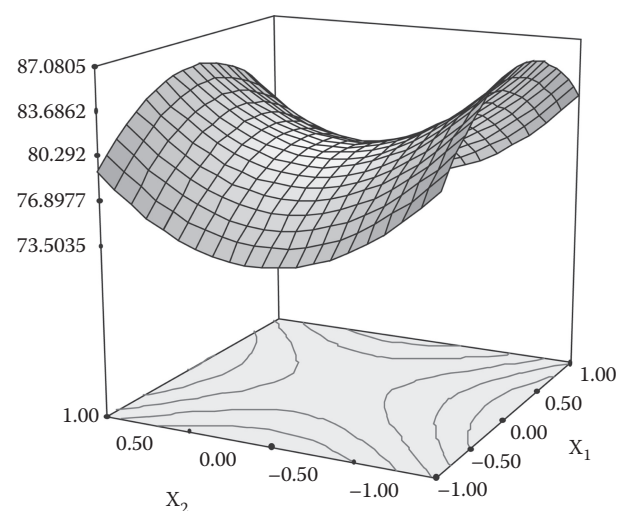


Fig. 8 An estimated response surface of a saddle point

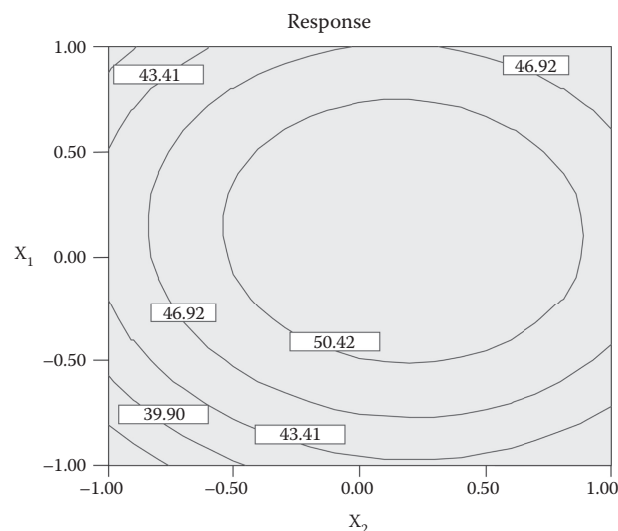


Fig. 9 A contour plot of a maximum response

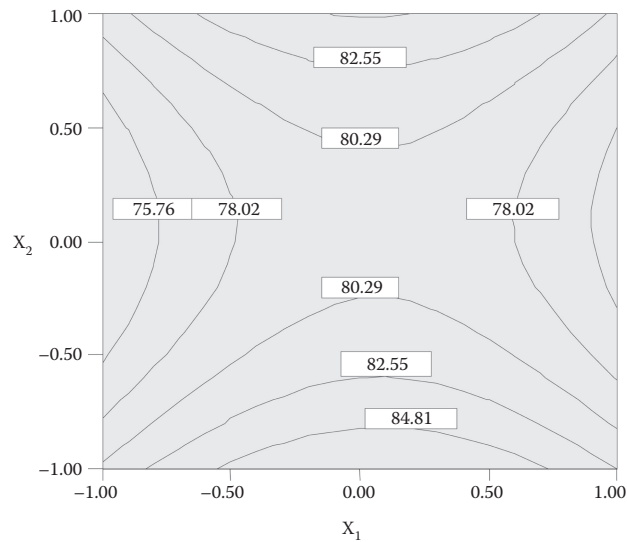


Fig. 10 A contour plot of a saddle point

variable via the 3-D or contour plot when the number of independent variables becomes large. This is especially true if the interaction between independent variables is present.

Canonical analysis is routinely used to determine whether the stationary point is a maximum, a minimum, or a saddle point. Canonical analysis starts with a rotation of the response function to a new set of axes, which correspond to the principal axes of the actual contour system. Looking at Fig. 10, the contour system with the new set of axes is provided. The second-order response surface model, in matrix notation, is $\hat{y} = \hat{\beta}_0 + \mathbf{x}'\hat{\beta} + \mathbf{x}'\hat{B}\mathbf{x}$ where $\hat{\beta}_0$ represents the estimated intercept, $\hat{\beta}$ represents linear coefficients, and $\hat{B}$ represents second-order coefficients. The nature of the stationary point is determined by looking at the signs and the magnitude of the eigenvalues of $\hat{B}$. If all the eigenvalues of are positive, then the stationary point is a minimum response. Therefore a move in any direction from the stationary point will result in an increase in the estimated response. If all the eigenvalues of are negative, then the stationary point is a maximum response. In this case, a move in any direction from the stationary point will result in a decrease in the estimated response. However, if the eigenvalues differ in sign, the stationary point is a *saddle point* and the change in the estimated response from the stationary point will depend on the direction one moves from the center of the design region.

The magnitude of the eigenvalue determines how sensitive the estimated response is as one moves away from the center of the experimental region in a given direction. For example, assume that both eigenvalues, in a two-variable case, are negative. Assume that one eigenvalue representing v_1 (similar to the x_1 direction) has a much larger absolute value than the eigenvalue representing

v_2 (this scenario is illustrated in Fig. 9). In this example, as one moves along the v_1 (or x_1) axis, there is a greater change in the estimated response, than when moving along the v_2 axis. Sendrek et al.^[10] used canonical analysis to determine the nature of the stationary point.

Situations can occur where there is a region, rather than a point, where the estimated optimum response exists. This is often termed a *stationary ridge*. Observing a stationary ridge can be very useful to the researcher because it indicates flexibility in operating conditions. Fig. 11 illustrates a contour plot, in a two-variable case, where a stationary ridge exists. In this example, the optimum region is a line. A rising (or falling) ridge can be present when the optimum region is found to be outside the experimental region. In this case, the experimenter may consider further experimentation along the rising ridge. This would be similar to steepest ascent. Fig. 12 provides an example of a rising ridge.

Even after canonical analysis has been performed, there are situations in which a researcher needs additional help in determining the best operating conditions. For example, assume that the stationary point is outside the experimental design (e.g., rising ridge), or that the stationary point is a saddle point. In these cases, *ridge analysis*, which determines the optimum response at a given distance from the center of the design region, can aid the statistical analysis. For more on analyzing response surfaces, see Myers, Montgomery, and Anderson-Cook.^[1] and Box and Draper.^[5]

MIXTURE EXPERIMENTS

All the discussions up to this point have involved experiments where the levels chosen for any control variable are independent of the levels chosen for other control

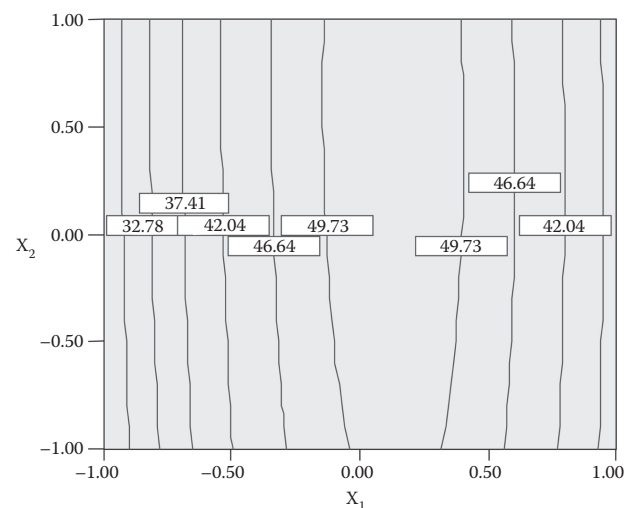


Fig. 11 A stationary ridge

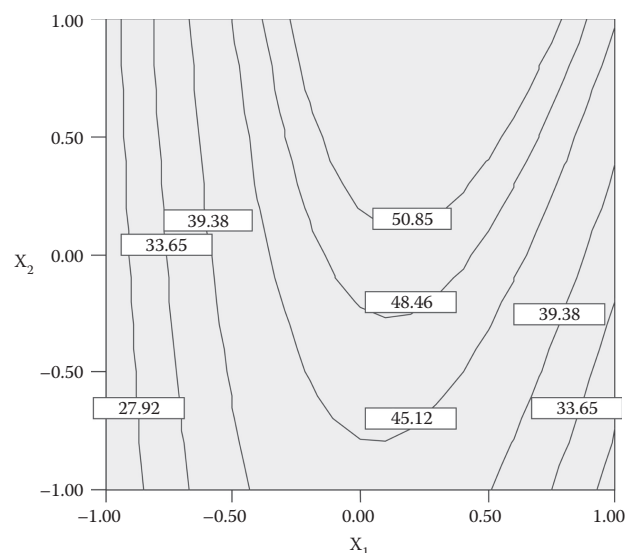


Fig. 12 A rising ridge

variables. However, experimental situations do exist where the response variable is assumed to depend only on the relative proportions of the components (or ingredients) and not on the amount. These types of situations are classified as *mixture experiments*. Drug formulation is one area within biopharmaceutical research in which mixture experiments are commonly used. Because the response variable is dependent on the proportion of the components, the following constraint is imposed: $\sum_{i=1}^k x_i = 1$.

Because of this constraint, the experimental region is defined as a simplex, which is a regularly sided figure with k vertices in $k - 1$ dimensions. Fig. 13 provides the simplex coordinate system for three components. Each of the vertices of the simplex represents *pure blends*, which

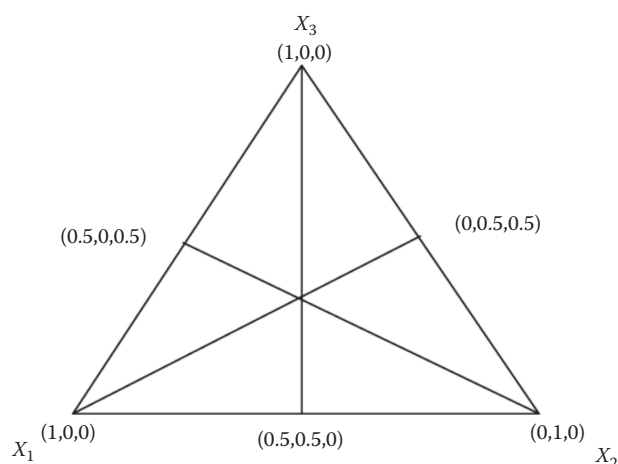


Fig. 13 A simplex coordinate system for three components

are mixtures that make up 100% of a single ingredient. The edges of the simplex, halfway between each of the vertices, represent *binary blends*, which are mixtures that represent equal amounts of two components. The location of the simplex, where the three lines intersect represent a *ternary blend*, is a mixture that represents an equal amount of all three components. Mixture experiments require different types of designs than those previously discussed. Because the mixture design region is a simplex, all design points must be at the vertices, on the edge, or in the interior of the simplex.

One of the most commonly used designs in the case of a mixture experiment is the *simplex lattice design*. A $\{k,m\}$ simplex lattice design for k components contains design points that are uniformly placed on the simplex. The actual design points are determined whereby the proportions assumed by each component take the $m + 1$ equally spaced values from 0 to 1: $x_i = 0, 1/m, 2/m, \dots, 1$.

For example, a $\{3,2\}$ simplex lattice design would consist of six design points (three pure blends and three binary blends). Another commonly used design in a mixture experiment is the *simplex centroid design*. In a k -component simplex centroid design, $2^k - 1$ distinct design points are used. These design points are the k permutations of primary blends, the $\binom{k}{2}$ permutations of binary blends, and so on, and the overall centroid $(1/k, 1/k, \dots, 1/k)$. Chu et al.^[18] and Shah et al.^[19] provide two examples of mixture experiments in biopharmaceutical applications. Chu et al.^[18] used a 10-point simplex centroid lattice augmented with three interior points to study the viscosity characteristics of a polymeric mucoadhesive formulation in multi-component solvent vehicles, whereas Shah et al.^[19] used a non-traditional simplex lattice to study the effects of varying quantitative compositions of the three key components (micronized drug, corn starch, and anhydrous lactose) on the compactibility and dissolution rates of tablets.

In many mixture experiments, there are constraints on the proportions of the components. In these cases standard mixture experiments may not be practical and thus computer-generated designs based on a given optimality criterion (e.g., D-optimality) would be a viable option. In addition, experimenters will encounter situations where both mixture and non-mixture variables are studied in one experiment. For example, an experimenter may want to study several mixture components and several process variables that could affect the blending properties of the ingredients. One must take extreme care with the statistical analysis of mixture experiments given that they are not performed identically as those with factorial experiments. Software packages like JMP® and Design Expert® can design and analyze mixture experiments. Cornell^[20] and Myers, Montgomery, and Anderson-Cook^[1] provide additional discussions of mixture experiments.

ROBUST PARAMETER DESIGNS

In the 1980s, Dr. Genichi Taguchi introduced a concept called robust parameter design (RPD), which increased the awareness of reducing variation in products and processes. RPD is a concept that stresses the proper choice of the *control factors* in order to reduce variation caused by a second set of factors called *noise variables*. These noise variables often contribute a large portion of the total variability. Some examples of noise factors are environmental conditions, raw materials, and operators. The primary objective of robust parameter design is to select the conditions of the control factors that are robust to changes in the noise variable. Therefore the conditions of the control factors that produce the best combination of bias error and variance error should be selected.

There are times in the development stage where the noise variables can be controlled in an experiment. This provides opportunities where both the control factors and noise variables can be combined into a single experiment. Taguchi uses a signal-to-noise ratio (SNR) performance criterion in order to determine the conditions of the control variable that are most robust to noise variables. Since the introduction of RPD by Taguchi numerous alternative modeling, analysis, and experimental design strategies using RSM have been proposed. Myers, Montgomery, and Anderson-Cook^[1] discuss, in great detail, how RSM can be applied to the robust parameter design problem.

CONCLUSION

RSM has been shown to be a very valuable tool in biopharmaceutical applications. The ability to effectively develop an optimum product or process has been demonstrated through many real-life examples. An important component of RSM involves statistical experimental design, which involves using efficient screening designs in order to eliminate unimportant variables from further experimentation. Another experimental design aspect involves using efficient second-order response surface designs in order to accurately estimate the response surface within the design region. The sequential nature of RSM provides great value and flexibility to the researcher. If a researcher is not in close proximity to the region of optimum experimental conditions, then steepest ascent (or descent) allows one to more efficiently locate the region of the optimum. A second sequential feature of RSM is the ability to augment existing screening designs in order to produce designs that allow one to estimate a second-order response surface. A second key component of RSM involves statistical analysis. This involves canonical analysis and ridge analysis, which are valuable tools to better understand the estimated response surface. They allow a researcher to better locate potential optimum experimental conditions. In addition, graphical tools like 3-D

plots and contour plots are also very useful tools to better understand how the response variable changes within the design region. Even though RSM has been used extensively in biopharmaceutical applications since the 1980s, there are certainly more opportunities to use RSM in biopharmaceutical research in the future.

REFERENCES

1. Myers, R.H.; Montgomery, D.C.; Anderson-Cook, C.M. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*; John Wiley & Sons: New York, 2009.
2. Myers, R.H. *Classical and Modern Regression with Application*; PWS-KENT Publishing: Boston, MA, 1990.
3. Draper, N.R.; Smith, H. *Applied Regression Analysis*, 2nd Ed.; John Wiley & Sons: New York, 1981.
4. Iskandarani, B.; Clair, J.H.; Patel, P.; Shiromani, P.K.; Dempski, R.,E. Simultaneous optimization of capsule and tablet formulation using response surface methodology. *Drug Dev. Ind. Pharm.* **1993**, *19* (16), 2089–2101.
5. Box, G.E.P.; Draper, N.R. *Empirical Model-Building and Response Surfaces*; John Wiley & Sons: New York, 1987.
6. Montgomery, D.C. *Design and Analysis of Experiments*, 3rd Ed.; John Wiley & Sons: New York, 1991.
7. Plackett, R.L.; Burman, J.P. The design of optimum multi-factorial experiments. *Biometrika* **1946**, *33*, 305–325.
8. Lin, D.K.J.; Draper, N.R. Projection properties of Plackett and Burman designs. *Technometrics* **1992**, *34*, 423–428.
9. Sastry, S.V.; Khan, M.A. Aqueous-based polymeric dispersion: Face-centered cubic design for the development of atenolol gastrointestinal therapeutic system. *Pharm. Dev. Technol.* **1998**, *3* (4), 423–432.
10. Sendrek, E.; Bonsignore, H.; Mungan, D. Response surface methodology as an approach to optimization of an oral solution. *Drug Dev. Ind. Pharm.* **1993**, *19* (4), 405–424.
11. Box, G.E. P.; Wilson, K.B. On the experimental attainment of optimum conditions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1951**, *13*, 1–45.
12. Lipps, D.M.; Sakr, A.M. Characterization of wet granulation process parameters using response surface methodology. *J. Pharm. Sci.* **1994**, *83* (7), 937–947.
13. Ogawa, S.; Kamijima, T.; Miyamoto, Y.; Miyajima, M.; Sato, H.; Takayama, K.; Nagai, T. A new attempt to solve the scale-up problem for granulation using response surface methodology. *J. Pharm. Sci.* **1994**, *83* (3), 439–443.
14. Branchu, S.; Forbes, R.T.; York, P.; Nyqvist, H. A Central Composite Design to investigate the thermal stabilization of lysozyme. *Pharm. Res.* **1999**, *16* (5), 702–708.
15. Wehrle, P.; Nobelis, P.; Cuine, A.; Stamm, A. Response surface methodology: An interesting statistical tool for process optimization and validation: An example of wet granulation in a high-shear mixer. *Drug Dev. Ind. Pharm.* **1993**, *19* (13), 1637–1653.
16. Hutchings, D.; Kuzmak, B.; Sakr, A. Processing considerations for an EC latex coating system: Influence of curing time and temperature. *Pharm. Res.* **1994**, *11* (10), 1474–1478.

17. Box, G.E. P.; Behnken, D.W. Some new three-level designs for the study of quantitative variables. *Technometrics* **1960**, 2, 455–475.
18. Chu, J.S.; Amidon, G.L.; Weiner, N.D.; Goldberg, A.H. Mixture experimental design in the development of a mucoadhesive gel formulation. *Pharm. Res.* **1991**, 8 (11), 1401–1407.
19. Shah, N.H.; Phuapradit, W.; Pinnamaraju, P.; Williams, L.; Infeld, M.H.; Malick, A.W. Strategies in the development of a tablet formulation of a poorly soluble drug for double-blind clinical studies. *Pharm. Technol.* **1994**, September, 152–160.
20. Cornell, J.A. *Experiments with Mixtures: Designs, Models, and Analysis of Mixture Data*, 2nd Ed.; John Wiley & Sons: New York, 1990.

Response-Adaptive Designs

Feifang Hu

University of Virginia, Charlottesville, Virginia, U.S.A.

Anastasia Ivanova

University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.

Abstract

This entry presents several classes of response-adaptive designs, and discusses the advantages and disadvantages of these designs, and the analysis of clinical trials based on adaptive designs.

INTRODUCTION

Response-adaptive designs, or response-adaptive randomization procedures, are designs that change allocation away from 50/50 based on responses observed so far in the trial. The desired allocation proportion is usually motivated by an ethical consideration of assigning more patients to better treatments. This review presents several classes of response-adaptive designs, and discusses the advantages and disadvantages of these designs, and the analysis of clinical trials based on adaptive designs. We also discuss goals for the allocation proportion and the power of treatment comparison in the trial where response-adaptive design is used.

RESPONSE-ADAPTIVE DESIGNS

The two main goals of response-adaptive designs^[1] are: (1) to maximize the individual patient's personal experience in a trial under certain restrictions, and (2) to reduce the overall number of patients (sample size) of a randomized clinical trial. The early ideas of adaptive designs can be traced back to Thompson^[2] and Robbins.^[3] Since then, two main families of response-adaptive designs have been proposed: (1) target-based, which is based on an optimal allocation target, where a specific criterion is optimized based on a population model; and (2) design-driven, where designs are driven by intuition and are not optimal in a formal sense.^[1,4]

Urn Models

One large class of response-adaptive designs is based on urn models (design-driven). Consider the simplest situation where two independent treatments "A" and "B" with binary outcomes are compared in the course of a trial. Let p_A be the probability of a success on treatment A and let p_B be the success probability of treatment B, with $q_A = 1 - p_A$ and

$q_B = 1 - p_B$. In addition, let n be the total number of patients in a trial, let n_A be the number of patients assigned to treatment A, and let n_B be the number of patients assigned to B, so that $n_A + n_B = n$. It is assumed that the outcome can be observed relatively quickly. Following the ethical imperative, one wants to change the allocation in the course of a trial so that more patients receive the treatment performing better thus far in the trial.

Play-the-Winner Rule

One of the first response-adaptive designs, the play-the-winner rule, was introduced by Zelen.^[5] The first patient is equally likely to receive one of the two treatments. The subsequent patient is assigned to the same treatment following a successful outcome, and to the opposite treatment following a failure. On average, the proportion of patients (n_A/n) assigned to treatment A is $q_B/(q_A + q_B)$.^[1] Therefore this design assigned more patients to the better treatment.

Randomized Play-the-Winner Rule

Wei and Durham^[6] and Wei^[7] suggested the randomized play-the-winner rule, a response-adaptive design that randomizes patients. An urn contains one ball of each type corresponding to the two treatments. When a patient arrives, a ball is drawn from the urn and then replaced. The patient receives corresponding treatment. If the outcome is a success, one ball of that type is added to the urn; otherwise, one ball of the opposite type is added to the urn. This design has the same limiting allocation proportion as the play-the-winner rule.

Drop-the-Loser Rule

According to the drop-the-loser rule,^[8] the urn starts with one ball of each treatment type and one ball of the so-called immigration type. When a patient arrives to be

randomized to a treatment, a ball is taken from the urn. If it is an immigration ball, the ball is replaced and two additional balls (one of each treatment type) are added to the urn. If it is an actual treatment ball, a patient is assigned to corresponding treatment and response is observed. If response is a failure, the ball is not returned to the urn. If response is a success, the ball is returned to the urn; hence the urn composition remains unchanged. On average, $nq_B/(q_A + q_B)$ patients are assigned to treatment A.

For clinical trials with K treatment ($K \geq 2$), an important family of urn models is the generalized Friedman's urn^[9] (also called generalized Polya's urn in literature), which includes the play-the-winner rule and the randomized play-the-winner rule as special cases. The properties of generalized Friedman's urn have been studied extensively.^[9–13] Another family of urn models is based on birth-and-death urn with immigration,^[14] which includes the drop-the-loser rule as a special case. Some special cases of birth-and-death urn have been proposed and studied recently.^[15–17] Many other urn models^[18–20] have been suggested for response-adaptive designs. These urn models and their properties are reviewed in Rosenberger.^[21]

There are several advantages of designs based on urn models: (1) urn models are intuitive and easy to implement in clinical trials; (2) they shift the allocation probability to better treatments; and (3) the properties of urn models have been studied extensively. For comparing two treatments ($K = 2$), the play-the-winner rule is a deterministic allocation procedure; therefore there is a possibility of selection bias.^[1] The randomized play-the-winner rule is a randomized design. Simulation studies^[22,23] have shown that the power of the treatment comparison might get lost when an adaptive urn model is used for allocation. The drop-the-loser rule is found to preserve power better than other adaptive urn designs.^[4,24,25]

There are several drawbacks of adaptive design based on urn models: (1) it can only target a specific allocation proportion; this allocation is usually not optimal in a formal sense; and (2) it can only be applied to the clinical trials with binary or multinomial responses. Modifications are necessary to apply urn models to clinical trials with continuous responses or survival responses.

Optimal Allocations

Response-adaptive designs were introduced following the ethical goal to maximize an individual patient's personal experience in the trial. Another important goal is to maximize the power of treatment comparison. These two goals usually compete with each other. Optimal allocation proportions are usually determined through some multiple-objective optimality criteria.^[26–29] Jennison and Turnbull^[29] described a general procedure for determining optimal allocation.

Consider the situation when two binomials are compared and the measure of interest is $p_A - p_B$. Let $\hat{p}_A$ and $\hat{p}_B$ be the corresponding maximum likelihood estimators of

p_A and p_B . The allocation that minimizes the variance of $\hat{p}_A - \hat{p}_B$ is the *Neyman allocation* with:

$$\frac{n_A}{n} = \frac{\sqrt{p_A q_A}}{\sqrt{p_A q_A} + \sqrt{p_B q_B}} \quad (1)$$

Note that if, for example, both p_A and p_B are less than 0.5, Neyman allocation will assign more patients to the inferior treatment. Another goal, a compromise between the ethical goal and efficiency, is to minimize the expected number of failures for a fixed variance of $\hat{p}_A - \hat{p}_B$. The *optimal allocation*^[23] for this goal is:

$$\frac{n_A}{n} = \frac{\sqrt{p_A}}{\sqrt{p_A} + \sqrt{p_B}} \quad (2)$$

For treatments with continuous outcomes, Dunnett^[30] considered an allocation rule for a set of multiple comparisons of $K - 1$ treatments vs. a single control. Dunnett proposed the following unbalance rule with an allocation ratio 1:1:1:...:1:. This allocation proportion is optimal when the variances of all $K - 1$ treatments and control are the same. It is still unclear how to find optimal allocations for general K treatment clinical trials.

Target-Based Response-Adaptive Designs

Because optimal allocation usually depends on some unknown parameters, we cannot implement it in practice, unless we estimate the unknown parameters sequentially. In this section, we introduce two allocation procedures that use sequential estimation of unknown parameters. We consider the case of comparing two treatments with binary responses. We also suppose that the optimal allocation is the target proportion.

Sequential Maximum Likelihood Procedure

1) To start, allocate $n_0 \geq 2$ patients to both treatments A and B, and 2) assume that we have assigned j ($j \geq 2n_0$) patients in the trial and their responses have been observed. Let $\hat{p}_A$ and $\hat{p}_B$ be the maximum likelihood estimates based on the j responses. Then we allocate the $(j + 1)$ st patient to treatment A with probability:

$$\frac{\sqrt{\hat{p}_A}}{\sqrt{\hat{p}_A} + \sqrt{\hat{p}_B}} \quad (3)$$

This design has been studied by Melfi and Page.^[31] The limiting allocation of this design is the optimal allocation^[32], $\left(\sqrt{p_A}/(\sqrt{p_A} + \sqrt{p_B})\right)$ and its properties have also been studied in Hu and Zhang.^[33]

Doubly Adaptive Biased Coin Design

This procedure is a generalization of the sequential maximum likelihood procedure.^[31] At first, allocate $n_0 \geq 2$

patients to both treatments A and B. Suppose we have assigned j ($j \geq 2n_0$) patients to treatment and are about to assign patients $j + 1$. Let $p = \sqrt{p_A}/(\sqrt{p_A} + \sqrt{p_B})$ be the desired allocation and let $\hat{p}$ be the desired allocation, replaced by the current estimates $\hat{p}_A$ and $\hat{p}_B$. Let j_A/j and j_B/j be the current proportions of patients that have already been allocated to A and B, respectively. Then the probability of assigning patient $j + 1$ to treatment A is given by:

$$g(j_A/j, \hat{p}) \quad (4)$$

where g is an allocation function. This design was first proposed by Eisele.^[34] The properties of the design have been studied recently.^[33,35] The following family of functions g has been proposed:^[33]

$$g(x, y) = \frac{y(x/y)^\gamma}{y(x/y)^\gamma + (1-y)((1-y)/(1-x))^\gamma} \quad (5)$$

where $\gamma \geq 0$.

Here we use a simple example with $\gamma = 2$ to illustrate the allocation procedure. Suppose we have already assigned $j = 12$ patients, $j_A = 6$ to A and $j_B = 6$ to B. We have observed a success rate of $\hat{p}_A = 2/3$ on A and $\hat{p}_B = 1/3$ on B. If we are interested in optimal allocation, then we can compute:

$$\hat{p} = \frac{\sqrt{\hat{p}_A}}{\sqrt{\hat{p}_A} + \sqrt{\hat{p}_B}} = \frac{\sqrt{2/3}}{\sqrt{2/3} + \sqrt{1/3}} = 0.586 \quad (6)$$

Then the probability of assigning the 13th patient to treatment A is computed as:

$$g(0.5, 0.586) = \frac{0.586(0.586/0.5)^2}{0.586(0.586/0.5)^2 + 0.414(0.414/0.5)^2} = 0.739 \quad (7)$$

Both sequential maximum likelihood procedures and doubly adaptive biased coin designs can be extended to K treatments.^[33]

Treatment Effective Mapping

Another family of response-adaptive designs is called treatment effect mapping.^[36] Rosenberger^[36] examined treatment effect mappings for continuous outcomes by using nonparametric linear rank statistic. This idea has been applied to several statistical models.^[37–39]

IMPORTANT ISSUES IN ADAPTIVE DESIGN

Variability and Power of Using Adaptive Designs

When using response-adaptive designs in clinical trials, the number of patients assigned to each treatment is a random variable. Therefore the power (for most testing hypotheses

and alternatives) is also a random variable.^[40] The variability of allocation has a strong effect on the power. This has been demonstrated by some simulations studies.^[22,23] Recently, Hu and Rosenberger^[24] show that the average power of a randomization procedure is a decreasing function of the variability of the procedure. Furthermore, a lower bound of the asymptotic variability of the allocation proportions for general-purpose adaptive designs is derived.^[25] Based on these theoretical results, we can compare different response-adaptive designs.

The drop-the-loser rule has minimum variability (lower bound) in the class of all response-adaptive procedures targeting urn allocation $q_B/(q_A + q_B)$. The doubly adaptive biased coin design can directly target any given allocation (may depend on unknown parameters). The parameter γ determines the variability of the procedure. At the extreme, when $\gamma = 0$, we have the sequential maximum likelihood procedure, which has the highest variability. As γ tends to ∞ , we have a procedure that attains the lower bound of any randomization procedure targeting the same allocation. But this procedure ($\gamma = \infty$) is entirely deterministic (except when $j_A/j = \hat{p}$). As γ becomes smaller, we have more randomization, but also more variability.

Overall, the doubly adaptive biased coin procedure is the most favorable design:^[24,25,33] 1) it is flexible in terms of targeting any desired allocation; 2) it can be applied to different types of responses; and 3) one can tune the parameter γ to reflect the tradeoff between the degree of randomization and variability, and the tradeoff between power and expected treatment successes. In a recently study of binary response,^[4] it has been shown that the doubly adaptive biased coin design ($\gamma = 2$) is always as powerful or more powerful than complete randomization (50/50) with fewer treatment failures.

Sample Size of Adaptive Designs

It is usually difficult to determine the required sample size for response-adaptive designs. This is because the allocation probabilities keep changing during a clinical trial when using response-adaptive design. For a fixed sample size, the number of patients assigned to each treatment is a random variable.^[40] Therefore the power is also a random variable for a fixed sample size.

In the literature, sample sizes of response-adaptive designs are calculated by ignoring the randomness of the allocations.^[1,35] As shown in one example of Hu,^[40] to achieve a target power of 0.8, the sample size is 62 from the formula^[1] (where the randomness of the design is ignored). If the sample size 62 is used, then, with 20% chance, the real power is less than 0.76.

Hu^[40] studied the properties of the power function of randomized designs and derived an approximate distribution of the power function for comparing two treatments when sample size is fixed. Based on the approximate distribution, a formula of the required sample size is then

obtained. In the above example, the required sample size should be 72, instead of 62. It also has been shown^[40] that a well-selected response-adaptive design requires a much smaller sample size compared with the equal allocation in certain cases.

Applying Adaptive Designs and Analyzing Data from Adaptive Designs

The randomized play-the-winner rule was used in the highly controversial extracorporeal membrane oxygenation (ECMO) study.^[41] A total of 12 patients were assigned, with only one patient assigned to conventional therapy and 11 assigned to ECMO. The only failure observed was in conventional therapy arm. It is reasonable to ask about the validity of such a trial.^[42,43] What went wrong? The randomized play-the-winner rule has a high variability of the allocation proportion (especially for small sample size $n = 12$). That is why only one patient was randomized to conventional therapy arm.

Adaptive designs have been implemented in several clinical trials, including a trial of crystalloid preload in hypotension for cesarean patients,^[44] a trial of fluoxetine vs. placebo in depression patients,^[45] and a four-arm trial in depression.^[46] Adaptive designs have also been used extensively in adaptive learning algorithms, game theory, as well as dynamical systems.^[47,48] Applications of adaptive designs in many different disciplines can be found in Flournoy and Rosenberger.^[49] Some general issues of implementing response-adaptive designs are discussed in Chapter 12 of Rosenberger and Lachin.^[1]

Inference for response-adaptive design is complicated because the treatment assignments depend on the responses. Likelihood-based inference was studied for different response-adaptive designs.^[50–53] In these papers, it has been shown that the maximum likelihood estimators are consistent and asymptotic normal under certain regularity conditions. For a clinical trial (binary responses) using the randomized play-the-winner rule,^[13,54] exact confidence intervals of the parameters p_A and p_B were derived. Confidence intervals can also be constructed following response-adaptive designs for censored survival data and binary responses.^[55] Rosenberger and Hu^[56] studied bootstrap confidence intervals for response-adaptive designs of K treatments.

CONCLUSION

In the above discussion, we assume that individual patient outcome will be immediately available. For clinical trials with delayed responses, we can update the urn when the response becomes available.^[13] The properties of urn model (the generalized Friedman's urn) with delayed responses are studied in some papers.^[57,58] For doubly adaptive biased coin designs, the unknown parameters are

estimated sequentially and the delayed responses effect these estimators. The properties of doubly adaptive biased coin design with delayed responses are unknown, and this is a topic of future research. It is also unclear how delayed responses will affect the birth-and-death urn and the treatment effective mapping.

We have focused on reviewing response-adaptive designs. In the literature, the term *adaptive designs* has been also used for designs that balance treatment assignments. The biased coin design of Efron^[59] has been proposed to force a sequential experiment to be balanced. After that, adaptive biased coin designs^[60,61] and generalized adaptive biased coin designs^[62] were proposed. Details and comparisons of balancing properties can be found in Rosenberger and Lachin.^[1]

Covariate information is often available and usually very important in clinical trials. In the literature, researchers have focused on how to balance treatment allocation on known covariates.^[63–66] Very few response-adaptive designs^[37,38] attempt to incorporate covariate information. How to use covariate information in response-adaptive design is a critical and conceptually difficult problem. This remains a very important topic of future research.

We have presented here our view on the use of response-adaptive designs in clinical trials. We apologize for not covering some important topics in adaptive designs. We hope that the discussion will point the readers in the right direction for a more comprehensive discussion of the different topics introduced.

REFERENCES

1. Rosenberger, W.F.; Lachin, J.M. *Randomization in Clinical Trials: Theory and Practice*; Wiley: New York, 2002.
2. Thompson, W.R. On the likelihood that one unknown probability exceeds another in the review of the evidence of the two samples. *Biometrika* **1933**, *25*, 275–294.
3. Robbins, H. Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.* **1952**, *58*, 527–535.
4. Rosenberger, W.F.; Hu, F. Maximizing Power and Minimizing Treatment Failures in Clinical Trials, **2004**, *1* (2), 141–147.
5. Zelen, M. Play the winner rule and the controlled clinical trial. *J. Am. Stat. Assoc.* **1969**, *64*, 131–146.
6. Wei, L.J.; Durham, S. The randomized play-the-winner rule in medical trials. *J. Am. Stat. Assoc.* **1978**, *73*, 840–843.
7. Wei, L.J. The generalized Polya's urn design for sequential medical trials. *Ann. Stat.* **1979**, *7*, 291–296.
8. Ivanova, A. A play-the-winner-type urn design with reduced variability. *Metrika* **2003**, *58*, 1–13.
9. Athreya, K.B.; Karlin, S. Embedding of urn schemes into continuous time branching processes and related limit theorems. *Ann. Math. Stat.* **1968**, *39*, 1801–1817.
10. Bai, Z.D.; Hu, F. Asymptotic theorem for urn models with nonhomogeneous generating matrices. *Stoch. Process. Their Appl.* **1999**, *80*, 87–101.

11. Bai, Z.D.; Hu, F. *Asymptotics in Randomized Urn Models*; Advances In Statistics **2008**, 334–360.
12. Smythe, R.T. Central limit theorems for urn models. *Stoch. Process. Their Appl.* **1996**, *65*, 115–137.
13. Wei, L.J. Exact two sample permutation tests based on the randomized play-the-winner rule. *Biometrika* **1988**, *75*, 603–606.
14. Ivanova, A.; Rosenberger, W.F.; Durham, S.D.; Flournoy, N. A birth and death urn for randomized clinical trials: Asymptotic methods. *Sankhya B* **2000**, *62*, 104–118.
15. Ivanova, A.; Flournoy, N. *A Birth and Death Urn for Ternary Outcomes: Stochastic Processes Applied to Urn Models*; Charalambides, C.A.; Koutras, M.V.; Balakrishnan, N., Eds.; Probability and Statistical Models with Applications; Chapman and Hall/CRC: Boca Raton, FL, 2001, 583–600.
16. Ivanova, A.; Rosenberger, W.F. Adaptive designs for clinical trials with highly successful treatments. *Drug Inf. J.* **2001**, *35*, 1087–1093.
17. Durham, S.D.; Flournoy, N.; Li, W. Sequential designs for maximizing the probability of a favorable response. *Can. J. Stat.* **1998**, *26*, 479–495.
18. Bai, Z.D.; Hu, F.; Shen, L. An adaptive design for multi-arm clinical trials. *J. Multivar. Anal.* **2002**, *81*, 1–18.
19. Andersen, J.; Faries, D.; Tamura, R.N. Randomized play-the-winner design for multi-arm clinical trials. *Commun. Stat., Theory Methods* **1994**, *23*, 309–323.
20. Bandyopadhyay, U.; Biswas, A. A class of adaptive designs. *Seq. Anal.* **2000**, *19*, 45–62.
21. Rosenberger, W.F. Randomized urn models and sequential design. *Seq. Anal.* **2002**, *21*, 1–28.
22. Melfi, V.; Page, C. Variability in Adaptive Designs for Estimation of Success Probabilities. In *New Developments and Application in Experimental Design*; Flournoy, N.; Rosenberger, W.F.; Wong, W.K., Eds.; Institute of Mathematical Statistics: Hayward, CA, 1998, 106–114.
23. Rosenberger, W.F.; Stallard, N.; Ivanova, A.; Harper, C.; Ricks, M. Optimal adaptive designs for binary response trials. *Biometrics* **2001**, *57*, 833–837.
24. Hu, F.; Rosenberger, W.F. Optimality, variability, power: Evaluating response-adaptive randomization procedures for treatment comparisons. *J. Am. Stat. Assoc.* **2003**, *in press*.
25. Hu, F.; Rosenberg, W.F.; Zhang, L.-X. *Asymptotically Best Response—Adaptive Randomization Procedures*. *Journal of Statistical Planning and Inference* **2006**, *136* (6), 1911–1922.
26. Hayre, L.S. Two-population sequential tests with three hypotheses. *Biometrika* **1979**, *66*, 465–474.
27. Hayre, L.S.; Turnbull, B.W. Estimation of the odds ratio in the two-armed bandit problem. *Biometrika* **1981**, *68*, 661–668.
28. Hardwick, J.; Stout, Q.F. Exact Computational Analysis for Adaptive Designs. In *Adaptive Designs*; Flournoy, N.; Rosenberger, W.F., Eds.; Institute of Mathematical Statistics: Hayward, CA, 1995.
29. Jennison, C.; Turnbull, B.W. *Group Sequential Methods with Application to Clinical Trials*; Chapman and Hall/CRC: Boca Raton, FL, 2000.
30. Dunnett, C.W. A multiple comparison procedure for comparing several treatments with a control. *J. Am. Stat. Assoc.* **1955**, *50*, 1096–1121.
31. Melfi, V.; Page, C. Estimation after adaptive allocation. *J. Stat. Plan. Inference* **2000**, *87*, 353–363.
32. Melfi, V.F.; Page, C.; Gerald, M. An adaptive randomized design with application to estimation. *Can. J. Stat.* **2001**, *29*, 107–116.
33. Hu, F.; Zhang, L.X. Asymptotic properties of doubly adaptive biased coin designs for multi-treatment clinical trials. *Ann. Stat.* **2004**, *32*, 268–301.
34. Eisele, J.R. The doubly adaptive biased coin design for sequential clinical trials. *J. Stat. Plan. Inference* **1994**, *38*, 249–261.
35. Eisele, J.; Woodroffe, M. Central limit theorems for doubly adaptive biased coin designs. *Ann. Stat.* **1995**, *23*, 234–254.
36. Rosenberger, W.F. Asymptotic inference with response adaptive treatment allocation designs. *Ann. Stat.* **1993**, *21*, 2098–2107.
37. Bandyopadhyay, U.; Biswas, A. Adaptive designs for normal responses with prognostic factors. *Biometrika* **2001**, *88*, 409–419.
38. Rosenberger, W.F.; Vidyashankar, A.N.; Agarwal, D.K. Covariate-adjusted response-adaptive designs for binary response. *J. Biopharm. Stat.* **2001**, *11*, 227–236.
39. Wei, L.J.; Yao, Q. Play the winner for phase II/III clinical trial. *Stat. Med.* **1996**, *15*, 2413–2423.
40. Hu, F. *Sample Size and Power of Randomized Designs*; 2003, Submitted for publication.
41. Bartlett, R.H.; Roloff, D.W.; Cornell, R.G.; Andrews, A.F.; Dillon, P.W.; Zwischenberger, J.B. Extracorporeal circulation in neonatal respiratory failure: A prospective randomized trial. *Pediatrics* **1985**, *76*, 479–487.
42. Ware, J.H. Investigating therapies of potentially great benefit: ECMO. *Stat. Sci.* **1989**, *4*, 298–306. C/R: pp. 306–340.
43. Faries, D.E.; Tamura, R.N.; Andersen, J.S. Adaptive designs in clinical trials. *Biopharm. Rep.* **1995**, *3*, 1–11.
44. Rout, C.C.; Rocke, D.A.; Levin, L.; Gouws, E.; Reddy, D. A reevaluation of the role of crystalloid preload in the prevention of hypotension associated with spinal anesthesia for elective cesarean section. *Anesthesiology* **1993**, *79*, 262–269.
45. Tamura, R.N.; Faries, D.E.; Andersen, J.S.; Heiligenstein, J.H. A case study of an adaptive clinical trial in the treatment of out-patients with depressive disorder. *J. Am. Stat. Assoc.* **1994**, *89*, 768–776.
46. Andersen, J. Clinical trials designs—Made to order. *J. Biopharm. Stat.* **1996**, *6*, 515–522.
47. Artur, V.B.; Ermol'ev, Yu.M.; Kaniiovskii, Yu.M. Adaptive growth processes modelled by urn schemes. *Cybernetics* **1987**, *23*, 49–57.
48. Posch, M. Cycling in a stochastic learning algorithm for normal form games. *Evol. Econ.* **1997**, *7*, 193–207.
49. *Adaptive Designs*; Flournoy, N.; Rosenberger, W.F., Eds.; Institute of Mathematical Statistics: Hayward, CA, 1995.
50. Rosenberger, W.F.; Flournoy, N.; Durham, S.D. Asymptotic normality of maximum likelihood estimators from multiparameter response-driven designs. *J. Stat. Plan. Inference* **1997**, *60*, 69–76.
51. Rosenberger, W.F.; Sriram, T.N. Estimation for an adaptive allocation design. *J. Stat. Plan. Inference* **1997**, *59*, 309–319.
52. Hu, F.; Rosenberger, W.F.; Zidek, J.V. Weighted likelihood for dependent data. *Metrika* **2000**, *51*, 223–243.

53. Hu, F.; Rosenberger, W.F. Analysis of time trends in adaptive designs with application to a neurophysiology experiment. *Stat. Med.* **2000**, *19*, 2067–2075.
54. Wei, L.J.; Smythe, R.T.; Lin, D.Y.; Park, T.S. Statistical inference with data-dependent treatment allocation rules. *J. Am. Stat. Assoc.* **1990**, *85*, 156–162.
55. Coad, D.S.; Woodroffe, M.B. Approximate confidence intervals after a sequential clinical trial comparing two exponential survival curves with censoring. *J. Stat. Plan. Inference* **1997**, *63*, 79–96.
56. Rosenberger, W.F.; Hu, F. Bootstrap methods for adaptive designs. *Stat. Med.* **1999**, *18*, 1757–1767.
57. Bai, Z.D.; Hu, F.; Rosenberger, W.F. Asymptotic properties of adaptive designs for clinical trials with delayed response. *Ann. Stat.* **2002**, *30*, 122–139.
58. Hu, F.; Zhang, L.X. *Asymptotic Normality of Adaptive Designs with Delayed Response*; Bernoulli **2004**, *10* (3), 447–463.
59. Efron, B. Forcing a sequential experiment to be balanced. *Biometrika* **1971**, *62*, 347–352.
60. Wei, L.J. A class of designs for sequential clinical trials. *J. Am. Stat. Assoc.* **1977**, *72*, 382–386.
61. Wei, L.J. The adaptive biased coin design for sequential experiments. *Ann. Stat.* **1978**, *6*, 92–100.
62. Smith, R.L. Sequential treatment allocation using biased coin designs. *J. R. Stat. Soc., B* **1984**, *46*, 519–543.
63. Atkinson, A.C. Optimal biased coin designs for sequential clinical trials with prognostic factors. *Biometrika* **1982**, *69*, 61–67.
64. Atkinson, A.C. Optimal biased-coin designs for sequential treatment allocation with covariate information. *Stat. Med.* **1999**, *18*, 1741–1752.
65. Pocock, S.J.; Simon, R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* **1975**, *31*, 103–115.
66. Zelen, M. The randomization and stratification of patients to clinical trials. *J. Chronic Dis.* **1974**, *28*, 365–375.

Risk Ratio Analysis

Kung-Jong Lui

Department of Mathematics and Statistics, San Diego State University, San Diego, California, U.S.A.

Abstract

The goal of this entry is to provide readers with an overview of characteristics and statistical analysis of the risk ratio analysis.

INTRODUCTION

As the risk ratio (RR), defined as the ratio of probabilities of developing a disease between the population with exposure and the population without exposure to a risk factor, was introduced,^[1] the RR has been the most important index to measure the strength of association between a risk factor and a disease in etiological studies.^[2–4] When the underlying disease under investigation is rare and chronic, directly estimating RR will require us to follow up a large cohort for a long time period. This can often cause the practical difficulty. For a rare disease, however, the RR can be approximated well by the odds ratio (OR).^[1,3,4] Furthermore, as the OR is invariant and estimatable under various study designs, we can estimate the RR indirectly by assessing its approximation—the OR from a retrospective case–control study.

For a randomized clinical trial (RCT), we may define the RR as the ratio of probabilities of having an adverse event between an experimental treatment and a placebo group. As the probability of having the adverse event in an RCT may not be small, the RR and OR are not necessarily approximately equal to each other. Thus, we cannot estimate the RR through the estimation of OR in an RCT as done in many epidemiological studies. The goal of this article is to provide readers with an overview of characteristics and statistical analysis of the RR. First, we note the relations between RR and other quantitative indices used in RCTs. We further note some asymptotic properties of commonly used estimators for the RR. We discuss estimation of the RR under stratified binomial sampling in large stratum size and sparse data, test of homogeneity of the RR across strata, and estimation of the rate ratio (which is relevant to the RR) under stratified Poisson sampling. Finally, we include a brief discussion on the estimation of the RR under other study designs, including the cluster randomization trial and simple compliance study.

DEFINITION AND CHARACTERISTICS OF RR

Let p_1 and p_0 denote the probability of suffering an adverse event in the experimental treatment and placebo group, respectively. The RR is simply defined as $\gamma = p_1/p_0$. The range for the RR is $0 < \gamma < \infty$. When $RR = 1$, the risks of having the adverse event for two comparison groups are equal. When $RR < 1$, the risk of having the adverse event in the experimental treatment is less than that in the placebo. Similarly, when $RR > 1$, the risk of having the adverse event in the experimental treatment is larger than that in the placebo. Note that the ratio p_1/p_0 is generally different from $(1-p_1)/(1-p_0)$, the ratio of probabilities of not having the adverse event. However, all the following statistical methods for the RR can be slightly modified to estimate the other ratio if the latter is our interest.

RELATIONS BETWEEN RR AND OTHER INDICES

Suppose that based on our subjective knowledge, we can assume that the experimental treatment tends to be beneficial, so that $RR (=p_1/p_0) \leq 1$. For example, consider the example of poliomyelitis vaccine trial.^[5,6] Here, p_1 and p_0 denote the probability of being a case with poliomyelitis in the vaccinated and placebo groups. As the vaccine is expected to protect a fraction δ of those who would have developed poliomyelitis if they had not been vaccinated, we may assume that $RR \leq 1$. The relative difference δ ,^[5,6] representing the protection effect owing to vaccinating those subjects who would otherwise have poliomyelitis if they were in the placebo group is, by definition, equal to $1 - RR (= (p_0 - p_1)/p_0)$. Thus, the relative difference^[4–6] may also be called the relative risk reduction, which is one of the most frequently used indices to measure the efficacy of a treatment in RCTs.

If the outcome (or response) of interest is rare, as noted previously, the RR can be approximated by the OR, $p_1(1-p_0)/[p_0(1-p_1)]$. Furthermore, because the OR can be estimated from the retrospective data and be expressed by parameters of the logistic regression or log-linear model, which can easily account for the confounding effects, the OR is likely to be the most useful index to measure the association between a risk factor and a disease in epidemiology. As the probability of outcome may often not be small in RCTs, however, the good property of the OR that can approximate the RR well may not hold. Because it is difficult for clinicians to comprehend the treatment effect in terms of the ratio of two odds, the OR is less frequently used in RCTs than other indices, such as the RR, relative difference, and risk difference.^[3]

The “number needed to treat” (NNT), defined as “the average number of patients needed to treat in order to prevent one patient with the adverse event in the placebo group,” has been recently proposed to quantify the treatment effect in RCTs and evidence-based medicine.^[7] Because this index can be easily understood by clinicians, the NNT has increasingly become popular to summarize the treatment effect.^[4,8–10] When $p_0 > p_1$, by definition, we can rewrite the NNT as $1/[(1-RR)p_0]$. This suggests that for a fixed baseline adverse event rate p_0 , the closer the RR to 0, the smaller is the NNT.

When one can assume that the time $T_i(>0)$ of an adverse event for a randomly selected patient in treatment $i(i=1,0)$ follows the negatively exponential distribution: $\lambda_i \exp(-\lambda_i t)$, we can easily see that the RR of probabilities for obtaining a patient with the adverse event between treatments $i=1$ and 0 in a given fixed follow-up time T^* is simply equal to $(1-\exp(-\lambda_1 T^*)) / (1-\exp(-\lambda_0 T^*))$. This ratio can be easily shown to approximately equal λ_1/λ_0 when both $\lambda_i T^*$ are small. In other words, the RR and the hazards ratio would be approximately equal if the follow-up time is relatively short to the mean length of time for the adverse event under investigation.

ESTIMATION AND ASYMPTOTIC PROPERTIES

Suppose that we randomly assign n_1 patients to receive the experimental treatment and n_0 patients to receive the placebo. Suppose further that we obtain X_i ($i=1,0$) number of patients with adverse events. The number X_i then follows the binomial distribution with parameters n_i and p_i . The likelihood of the observed data is given by

$$L(p_1, p_0 | x_1, x_0) = \binom{n_1}{x_1} p_1^{x_1} (1-p_1)^{n_1-x_1} \times \binom{n_0}{x_0} p_0^{x_0} (1-p_0)^{n_0-x_0}$$

On the basis of the above likelihood, the maximum likelihood estimator (MLE) for p_i is

$$\hat{p}_i = X_i/n_i$$

which is the sample proportion of patients with adverse events for treatment i . By the functional invariance property of the MLE,^[11] the MLE for the RR is simply

$$\hat{\gamma} = \hat{p}_1/\hat{p}_0$$

Note that because there is a positive probability that $\hat{p}_0 = 0$, the MLE $\hat{\gamma}$ does not have a finite expectation $E(\hat{\gamma})$. Note also that because the sampling distribution of $\hat{\gamma}$, a ratio of two sample proportions, can be skewed, we may wish to apply the logarithmic transformation to improve its normal approximation when deriving the interval estimator for the RR. This leads us to obtain an asymptotic $100(1-\alpha)\%$ confidence interval for the RR to be^[12]

$$\left[\hat{\gamma} \exp(-Z_{\alpha/2} (\widehat{\text{Var}}(\log(\hat{\gamma})))^{1/2}), \right. \\ \left. \hat{\gamma} \exp(Z_{\alpha/2} (\widehat{\text{Var}}(\log(\hat{\gamma})))^{1/2}) \right]$$

where Z_α is the upper $100(\alpha)\%$ th percentile of the standard normal distribution and $\widehat{\text{Var}}(\log(\hat{\gamma})) = (1-\hat{p}_1)/X_1 + (1-\hat{p}_0)/X_0$. Katz et al.^[12] compared the above interval estimator with other interval estimators. They noted that the interval estimator using the logarithmic transformation can perform well in a variety of situations. Other transformation to alleviate the skewness of the sampling distribution for $\hat{\gamma}$ can be found elsewhere.^[13] Note that if we should obtain either of X_i 's to be 0, the above interval estimator would not be applicable. Whenever this occurs, we can apply the commonly used ad hoc procedure of adding 0.50 to each cell for sparse data in calculation of $\widehat{\text{Var}}(\log(\hat{\gamma}))$. Using such an adjustment, we cannot, however, eliminate the bias. Lui proposed the use of inverse sampling,^[14] in which we continue sampling patients until we obtain a predetermined number x_i of patients with adverse events.^[4,15–17] Under this sampling, Lui derived the uniformly minimum variance unbiased estimator as well as several interval estimators for the RR.

STRATIFIED BINOMIAL SAMPLING

When patients are difficult to recruit, we may employ a multicenter design to facilitate the collection of data. Say, we have S centers participating in our study. For each center s ($s=1,2,\dots,S$), suppose that we randomly assign n_{is} patients to receive the experimental treatment ($i=1$) or the placebo ($i=0$), and obtain X_{is} number of patients with the adverse event of interest. Because patients from the center are likely to have similar characteristics and to be examined by the same laboratory, the baseline risks of patients

are likely to vary between centers. To account for this confounding effect, we may apply stratified analysis with strata formed by centers. Let p_{is} denote the corresponding probability of having an adverse event for patients receiving treatment i in center s . The number X_{is} then follows the binomial distribution with parameters n_{is} and p_{is} . The RR in stratum s between the two treatments is defined as $\gamma_s = p_{1s}/p_{0s}$. Note that the MLE of γ_s is $\hat{\gamma}_s = \hat{p}_{1s}/\hat{p}_{0s}$, where $\hat{p}_{is} = X_{is}/n_{is}$. In the following discussion, we assume that γ_s is constant across strata and denote this common value by γ_c . The likelihood is then given by

$$L(\gamma_c, p_0 | x_1, x_0) = \prod_{s=1}^S \binom{n_{1s}}{x_{1s}} (\gamma_s p_{0s})^{x_{1s}} (1 - \gamma_c p_{0s})^{n_{1s} - x_{1s}} \binom{n_{0s}}{x_{0s}} p_{0s}^{x_{0s}} (1 - p_{0s})^{n_{0s} - x_{0s}}$$

where $p_0 = (p_{01}, p_{02}, \dots, p_{0S})'$ and $x_i = (x_{i1}, x_{i2}, \dots, x_{iS})'$ for $i=1, 0$. The above likelihood depends on γ_c and S and other nuisance parameters p_{0s} , $s=1, 2, \dots, S$. To obtain the MLE $\hat{\gamma}_c$ requires us to employ an iterative numerical procedure^[18] to find the value γ_c which maximizes the above likelihood. Greenland and Robins^[19] note that the MLE $\hat{\gamma}_c$ can be biased in the sparse data, in which the number of patients within strata is small, but the number of strata is moderate or large.

Estimation in Large Stratum Size

First, we consider the situation in which the stratum size is large. Because it is asymptotically efficient and simple for use, the most commonly used point estimator for γ_c under stratified sampling is the weighted least-squares (WLS) point estimator.^[3,4,20] Because the asymptotic variance $\widehat{Var}(\hat{\gamma}_s)$ is proportional to $\widehat{Var}(\log(\hat{\gamma}_s))$, when γ_s is constant, to estimate γ_c , we may apply the WLS point estimator $\hat{\gamma}_{WLS} = \sum_s \hat{W}_s \hat{\gamma}_s / \sum_s \hat{W}_s$, where $\hat{W}_s = \widehat{Var}(\log(\hat{\gamma}_s))^{-1}$ and $\widehat{Var}(\log(\hat{\gamma}_s)) = [(1 - \hat{p}_{1s}) / (n_{1s} p_{1s}) + (1 - \hat{p}_{0s}) / (n_{0s} p_{0s})]$. When deriving the interval estimator for γ_c , we commonly employ the logarithmic transformation and obtain an asymptotic $100(1 - \alpha)\%$ confidence interval for γ_c to be

$$\left[\exp \left\{ \frac{\sum_s \hat{W}_s \log(\hat{\gamma}_s)}{\sum_s \hat{W}_s} - Z_{\alpha/2} / \sqrt{\sum_s \hat{W}_s} \right\}, \exp \left\{ \frac{\sum_s \hat{W}_s \log(\hat{\gamma}_s)}{\sum_s \hat{W}_s} + Z_{\alpha/2} / \sqrt{\sum_s \hat{W}_s} \right\} \right]$$

Note that if either $\hat{p}_{1s}$ or $\hat{p}_{0s}$ equals 0 in stratum s , the WLS estimators will be undefined. If this occurs, we may add 0.50 to each cell when calculating $\hat{\gamma}_{WLS}$ or the above interval estimator.

Estimation in Sparse Data

When the data are sparse (i.e., all n_{is} are small), however, using the WLS interval estimator even with the above ad hoc adjustment procedure of adding 0.50 to each cell can be misleading (i.e., the point estimator $\hat{\gamma}_{WLS}$ is seriously biased and the coverage probability of the interval estimator based on $\hat{\gamma}_{WLS}$ can be much less than the desired confidence level). These lead us to consider the Mantel-Haenszel (MH) type estimator:^[21–23]

$$\hat{\gamma}_{MH} = \left(\sum_{s=1}^S \bar{\omega}_s \hat{p}_{1s} \right) / \left(\sum_{s=1}^S \bar{\omega}_s \hat{p}_{2s} \right)$$

where $\bar{\omega}_s = n_{1s} n_{0s} / n_{.s} n_{.s} = n_{1s} + n_{0s}$. Note that the MH estimator $\hat{\gamma}_{MH}$ is defined as long as there is at least one $\hat{p}_{2s} > 0$. Greenland and Robins^[19] derived an estimated asymptotic variance of $\hat{\gamma}_{MH}$ for sparse data:

$$\widehat{Var}(\log(\hat{\gamma}_{MH})) = \left\{ \sum_s [n_{1s} n_{0s} (X_{1s} + X_{0s}) - n_{.s} X_{1s} X_{0s}] / n_{.s}^2 \right\} / \left\{ \left(\sum_s \bar{\omega}_s \hat{p}_{1s} \right) \times \left(\sum_s \bar{\omega}_s \hat{p}_{0s} \right) \right\}$$

Thus, we obtain an asymptotic $100(1 - \alpha)\%$ confidence interval for γ_c to be

$$\left[\hat{\gamma}_{MH} \exp \left\{ -Z_{\alpha/2} (\widehat{Var}(\log(\hat{\gamma}_{MH})))^{1/2} \right\}, \hat{\gamma}_{MH} \exp \left\{ Z_{\alpha/2} (\widehat{Var}(\log(\hat{\gamma}_{MH})))^{1/2} \right\} \right]$$

This interval is valid for use even for the case of one-to-one matching. For further information, we refer readers to publications^[24–26] that evaluate and compare various interval estimators of the RR in sparse data.

Test of Homogeneity of RR

Before we obtain a summary estimator for the underlying common γ_c , it is important for us to examine the assumption that $H_0: \gamma_1 = \gamma_2 = \dots = \gamma_s = \gamma_c$ to avoid overlooking the possible interaction effect between treatment and stratum effects on the outcome. When the stratum size is extremely large, we may apply the WLS test statistic.^[3,4]

$$T_{WLS} = \sum_s \hat{W}_s \left[\log(\hat{\gamma}_s) - \frac{\sum_s \hat{W}_s \log(\hat{\gamma}_s)}{\sum_s \hat{W}_s} \right]^2$$

and reject H_0 at α -level if $T_{WLS} > \chi_{\alpha, S-1}^2$, where $\chi_{\alpha, S-1}^2$ is the upper $100(\alpha)\%$ th percentile of the χ^2 -distribution with $S-1$ degrees of freedom. Lui and Kelly^[27] note that this WLS test procedure can be quite conservative if the stratum size is not large. To improve this test procedure, Lui and Kelly

develop several other alternative test procedures. For its simplicity, we present only the following simple test procedure, which is generally preferable to the above WLS test procedure.^[4,27] We define the test statistic

$$T_{WLS}^* = (T_{WLS} - (S-1)) / \sqrt{2(S-1)}$$

Note that when γ_s varies between strata, we expect T_{WLS} to be large. Thus, we reject H_0 at α -level if $T_{WLS}^* > Z_{\alpha}$. When the data are sparse, none of the test procedures considered by Lui and Kelly can perform well. Recently, Lui^[28] concentrates attention on sparse data and develop a simple test procedure based on the beta-binomial model by using the bootstrapping method^[29] for testing the homogeneity of relative difference. Lui demonstrates that his procedure can perform well even when the stratum size is as small as two patients per treatment in a variety of situations. Because the relative difference is equal to $1 - RR$ (when $RR \leq 1$), we may modify Lui's test procedure to test the homogeneity of the RR focused here. We refer readers to this article for details.

ESTIMATION UNDER STRATIFIED POISSON SAMPLING

When the treatment effect takes years to develop, many patients may drop out during the lengthy follow-up period. Thus, it is important to account for various person-time units at risk between the comparison groups. The Poisson model, which can easily incorporate this information into data analysis, is often employed.^[30,31]

Suppose that in an RCT, we want to assess the ratio of the adverse event rates between the experimental treatment ($i=1$) and the placebo ($i=0$) and suppose that we obtain X_{is} number of patients with the adverse event in T_{is} person-time units of follow-up from group i ($i=1,0$) in stratum s ($s=1,2,\dots,S$), we assume that X_{is} independently follows a Poisson distribution: $(\lambda_{is} T_{is})^{X_{is}} \exp(-\lambda_{is} T_{is}) / X_{is}!$, where λ_{is} is the adverse event rates. The rate ratio (which is, explicitly speaking, different from the RR discussed above) is defined as $\gamma_s^* = \lambda_{1s} / \lambda_{0s}$. The MLE of γ_s^* is simply given by $\hat{\gamma}_s^* = \hat{\lambda}_{1s} / \hat{\lambda}_{0s}$, where $\hat{\lambda}_{is} = X_{is} / T_{is}$. In the following discussion, we assume that γ_s^* is constant across strata and denote this common value by γ_c^* . If the expected number of cases per treatment in each stratum is large, we may use the WLS point estimator to estimate the underlying common rate ratio γ_c^*

$$\hat{\gamma}_{WLS}^* = \sum_s \hat{W}_s^* \hat{\gamma}_s^* / \sum_s \hat{W}_s^*$$

where $\hat{W}_s^* = \widehat{\text{Var}}(\log(\hat{\gamma}_s^*))^{-1}$ and $\widehat{\text{Var}}(\log(\hat{\gamma}_s^*)) = 1/X_{1s} + 1/X_{0s}$. Furthermore, when applying the WLS interval estimator with the logarithmic transformation, we obtain an asymptotic $100(1-\alpha)\%$ confidence interval for γ_c^* to be

$$\left[\exp \left\{ \sum_s \hat{W}_s^* \log(\hat{\gamma}_s^*) / \sum_s \hat{W}_s^* - Z_{\alpha/2} / \sqrt{\sum_s \hat{W}_s^*} \right\}, \right. \\ \left. \exp \left\{ \sum_s \hat{W}_s^* \log(\hat{\gamma}_s^*) / \sum_s \hat{W}_s^* - Z_{\alpha/2} / \sqrt{\sum_s \hat{W}_s^*} \right\} \right]$$

When the expected number of cases per treatment in each stratum is small, we may also consider the MH estimator^[18] for the rate ratio

$$\hat{\gamma}_{MH}^* = \sum_s X_{1s} T_{0s} / T_{1s} \bigg/ \sum_s X_{0s} T_{1s} / T_{0s}$$

where $T_{.s} = T_{1s} + T_{0s}$. The estimated asymptotic conditional variance of $\hat{\gamma}_{MH}^*$ is^[30]

$$\widehat{\text{Var}}(\hat{\gamma}_{MH}^*) = \hat{\gamma}_{MH}^* \left(\sum_s T_{1s} T_{0s} x_{.s} / T_{.s}^2 \right) / \left[\sum_s T_{1s} T_{0s} x_{.s} / (T_{.s} (\hat{\gamma}_{MH}^* T_{1s} + T_{0s})) \right]^2$$

These lead to obtain an asymptotic $100(1-\alpha)\%$ confidence interval for γ_c^* to be

$$\left[\hat{\gamma}_{MH}^* \exp \left\{ -Z_{\alpha/2} (\widehat{\text{Var}}(\log(\hat{\gamma}_{MH}^*)))^{1/2} \right\}, \right. \\ \left. \hat{\gamma}_{MH}^* \exp \left\{ Z_{\alpha/2} (\widehat{\text{Var}}(\log(\hat{\gamma}_{MH}^*)))^{1/2} \right\} \right]$$

where $\widehat{\text{Var}}(\log(\hat{\gamma}_s^*)) = \widehat{\text{Var}}(\hat{\gamma}_s^*) / (\hat{\gamma}_s^*)^2$. Recently, Lui^[31] has considered and compared eight interval estimators for the rate ratio, including the above two interval estimators presented here. Lui has noted that the interval estimator using Wald's statistic with the logarithmic transformation and the two other test-based confidence intervals based on the conditional distribution tend to be more efficient than the above interval estimator based on the MH estimator. However, using the former requires the use of iterative numerical procedures, while using the latter is much easy to calculate.

To test the homogeneity of the rate ratio between strata, we may employ the asymptotic likelihood ratio test proposed by Rothman and Boice^[18]

$$T_{LR} = 2 \left\{ \sum_s [x_{1s} \log(\hat{\gamma}_s) - x_{.s} \log(T_{1s} \hat{\gamma}_s + T_{0s})] \right. \\ \left. - \sum_s [x_{1s} \log(\hat{\gamma}_c) - x_{.s} \log(T_{1s} \hat{\gamma}_c + T_{0s})] \right\}$$

where $\hat{\gamma}_c$ is the conditional MLE and is the root γ_c satisfying the equation

$$x_{1.} = \sum_s x_{.s} T_{1s} \gamma_c / (T_{1s} \gamma_c + T_{0s})$$

where $x_{1.} = \sum_s x_{1s}$ and $x_{.s} = X_{1s} + X_{0s}$. If we obtain $T_{LR} > \chi_{\alpha, S-1}^2$, we will reject the null hypothesis $H_0: \gamma_1 = \gamma_2 = \dots = \gamma_S$ at α -level. Lui^[31] has given a sufficient and necessary condition for the existence of a unique finite conditional MLE $\hat{\gamma}_c$ and provided a numerical procedure to find $\hat{\gamma}_c$.

OTHER STUDY DESIGNS

When comparing an experimental treatment with a standard treatment (or a placebo), some patients assigned to the experimental treatment may decline and switch to the standard treatment. For example, consider the study of using vitamin A supplementation to reduce mortality in preschool children.^[32] Nearly 20% of children failed to take the vitamin A supplement. To assess the effect of vitamin A, we may need to account for this compliance information when estimating the RR of mortality among children who would be willing to take vitamin A.^[33] Readers may find a systematic discussion on interval estimation of the RR for this simple compliance study design (Lui, K.-J. Notes on interval estimation of risk ratio in the simple compliance randomized trial. Department of Mathematics and Statistics, San Diego, CA, unpublished manuscript.). Because the data structure for this type of simple compliance study is essentially parallel to the single consent-randomized design,^[34,35] the interval estimator of the RR for the former is applicable to the latter as well.

For administrative convenience, cluster randomization trials, in which the unit of randomization is a class, a household or a spouse pair, are widely used for evaluating an intervention of the prevention of a disease.^[36–39] As the responses on patients within clusters are likely correlated, it is important to incorporate this intraclass correlation into estimation to avoid overestimating the precision. Owing to the space limitations, we refer readers to publications^[4,37,39] in which some examples and discussions on interval estimation of the RR for cluster randomization trials can be found.

CONCLUSIONS

Because the RR, representing the times of the risk that a randomly selected patient has an adverse event in the experimental treatment relative to that in a placebo, can be easily understood, the RR (or equivalently, the relative risk reduction $1 - \text{RR}$) is one of the most useful indices to summarize the treatment effect in RCTs. By contrast, the OR, although which is frequently used to locate the potential cause of a disease in epidemiology, is more difficult to interpret itself when it cannot approximate the RR well. The discussion on the characteristics of the RR, the relations between the RR and other indices, and estimation of the RR under various study designs presented here should have use for biostatisticians and clinicians when they wish to employ the RR to present their clinical findings.

REFERENCES

1. Cornfield, J. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *J. Nat. Cancer Inst.* **1951**, *11*, 1269–1275.
2. Miettinen, O.S. Estimation of relative risk from individually matched series. *Biometrics* **1970**, *26*, 75–86.
3. Fleiss, J. *Statistical Methods for Rates and Proportions*, 2nd Ed.; Wiley: New York, 1981.
4. Lui, K.-J. *Statistical Estimation of Epidemiological Risk*; Wiley: New York, 2004.
5. Sheps, M.C. Shall we count the living or the dead? *N. Engl. J. Med.* **1958**, *259*, 1210–1214.
6. Sheps, M.C. An examination of some methods of comparing several rates or proportions. *Biometrics* **1959**, *15*, 87–97.
7. Laupacis, A.; Sackett, D.L.; Roberts, R.S. An assessment of clinically useful measures of the consequences of treatment. *N. Engl. J. Med.* **1988**, *318*, 1728–1733.
8. Cook, R.J.; Sackett, D.L. The number needed to treat: a clinically useful measure of treatment effect. *Br. Med. J.* **1995**, *310*, 452–454.
9. Tramèr, M.R.; Moore, A.; McQuay, H. Prevention of vomiting after paediatric strabismus surgery: a systematic review using the numbers-needed-to-treat method. *Br. J. Anaesth.* **1995**, *75*, 556–561.
10. Sackett, D.L.; Deeks, J.J.; Altman, D.G. Down with odds ratio! *Evid. Based Med.* **1996**, *1*, 164–166.
11. Casella, G.; Berger, R.L. *Statistical Inference*; Duxbury: Belmont, CA, 1990.
12. Katz, D.; Baptista, J.; Azen, S.P.; Pike, M.C. Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics* **1978**, *34*, 469–474.
13. Bailey, B.J.R. Confidence limits to the risk ratio. *Biometrics* **1987**, *43*, 201–205.
14. Haldane, J.B.S. On a method of estimating frequencies. *Biometrika* **1945**, *33*, 222–225.
15. Lui, K.-J. Confidence intervals for the risk ratio in cohort studies under inverse sampling. *Biom. J.* **1995**, *37*, 965–971.
16. Lui, K.-J. Notes on conditional confidence limits under inverse sampling. *Stat. Med.* **1995**, *14*, 2051–2056.
17. Lui, K.-J. Point estimation of relative risk under inverse sampling. *Biom. J.* **1996**, *38*, 669–680.
18. Rothman, K.J.; Boice, J.D. *Epidemiological Analysis with a Programmable Calculator*. NIH Publ. 79-1649. U.S. Government Printing Office: Washington, DC, 1979.
19. Greenland, S.; Robins, J.M. Estimation of a common effect parameter from sparse follow-up data. *Biometrics* **1985**, *41*, 55–68.
20. Grizzle, J.E.; Starmer, C.F.; Koch, G.G. Analysis of categorical data by linear models. *Biometrics* **1969**, *25*, 489–504.
21. Tarone, R.E. On summary estimators of relative risk. *J. Chronic Dis.* **1981**, *34*, 463–468.
22. Kleinbaum, D.G.; Kupper, L.L.; Morgenstern, H. *Epidemiologic Research: Principles and Quantitative Methods*; Lifetime Learning Publications: Belmont, CA, 1982.
23. Nurminen, M. Asymptotic efficiency of general noniterative estimators of common relative risk. *Biometrika* **1981**, *68*, 525–530.
24. Lui, K.-J. Interval estimation of the proportion ratio under multiple matching. *Stat. Med.* **2005**, *25*, 1275–1285.
25. Lui, K.-J. Interval estimation of the proportion ratio under multiple matching (Correction). *Stat. Med.* **2005**, *24*, 3682.
26. Lui, K.-J. A monte carlo evaluation of five interval estimators for the relative risk in sparse data. *Biom. J.* **2006**, *48*, 131–143.

27. Lui, K.-J.; Kelly, C. Tests for homogeneity of the risk ratio in a series of 2×2 tables. *Stat. Med.* **2000**, *19*, 2919–2932.
28. Lui, K.-J. Procedures for testing the homogeneity of relative difference in sparse data. *Stat. Med.* **2006**, *25*, 685–697.
29. Efron, B.; Tibshirani, R.Y. *An Introduction to the Bootstrap*; Chapman and Hall: New York, 1993.
30. Breslow, N.E. Elementary methods of cohort analysis. *Int. J. Epidemiol.* **1984**, *13*, 112–115.
31. Lui, K.-J. Eight interval estimators of a common rate ratio under stratified poisson sampling. *Stat. Med.* **2004**, *23*, 1283–1296.
32. Sommer, A.; Tarwotjo, I.; Djunaedi, E.; West, K.P.; Loedin, A.A.; Tilden, R.; Mele, L. Impact of vitamin A supplementation on childhood mortality: a randomized controlled community trial. *Lancet* **1986**, *1*, 1169–1173.
33. Sommer, A.; Zeger, S.L. On estimating efficacy from clinical trials. *Stat. Med.* **1991**, *10*, 45–52.
34. Zelen, M. A new design for randomized clinical trials. *N. Engl. J. Med.* **1979**, *300*, 1242–1245.
35. Ellenberg, S.S. Randomization designs in comparative clinical trials. *N. Engl. J. Med.* **1984**, *310*, 1404–1408.
36. Cornfield, J. Randomization by group: a formal analysis. *Am. J. Epidemiol.* **1978**, *108*, 100–102.
37. Donner, A.; Birkett, N.; Buck, D. Randomization by cluster sample size requirements and analysis. *Am. J. Epidemiol.* **1981**, *114*, 906–914.
38. Klar, N.; Donner, A. Current and future challenges in the design and analysis of cluster randomization trials. *Stat. Med.* **2001**, *20*, 3729–3740.
39. Lui, K.-J.; Mayer, J.A.; Eckhardt, L. Confidence intervals for the risk ratio under cluster sampling based on the beta-binomial mode. *Stat. Med.* **2000**, *19*, 2933–2942.

ROC Curve

Robert G. Lehr

Bayer Healthcare Pharmaceuticals, Montville, New Jersey, U.S.A.

Annpey Pong

Merck Research Laboratories, Summit, New Jersey, U.S.A.

Abstract

The receiver operating characteristic (ROC) curve has become a useful tool to distinguish differences in diagnostic performance. This entry is focused on breaking down the details of the ROC curve and provides statistical designs, methods, and analysis.

INTRODUCTION

In the recent development of diagnostic imaging, the receiver operating characteristic (ROC) curve has become a useful tool to distinguish differences in diagnostic performance. Use of this method enhances the traditional measures, and is now well established. The conventional approach of most diagnostic tests is to provide a dichotomous outcome (positive/negative of test result and presence/absence of disease status). The accuracy of test performance is then evaluated by “sensitivity” and “specificity.” The sensitivity is the probability that a test result is positive given the presence of disease. On the other hand, the specificity is the probability that a test result is negative given the absence of disease. However, very often a single pair of sensitivity and specificity measurements may provide a possibly misleading and even hazardous oversimplification of accuracy.^[1] This is why the ROC curve needs to be established.

For many diagnostic tests, the test data do not necessarily fall into one of two obviously defined categories. It is required that some confidence threshold be established in the mind of the decision maker, e.g., how strong the suspicion for the test result to be called as positive if the test suggests the possibility of disease. Therefore, the decision maker chooses between a positive and negative test result by using his/her confidence with an arbitrary confidence threshold, such as definitely negative, probably negative, questionable, probably positive, and definitely positive. Thereafter, an ROC curve is determined by a plot of sensitivity (usually on the vertical axis) vs. 1-specificity, which depicts sensitivity and specificity levels over the entire range of decision thresholds. Consequently, an ROC curve provides more information than just a single sensitivity and specificity pair to describe the accuracy of a set of test results.

ROC methodology was originally developed for electronic signal detection and evaluations of radar signals in

the early 1950s. In the mid-1960s, ROC curves were used in experimental psychology and other nonmedical fields. In the late 1960s, the radiologist Leo Lusted suggested using ROC analysis in medical decision making. The first ROC curve applied in studies of medical imaging devices was in 1969.^[2]

REGULATORY REQUIREMENTS

In 1998, the Food and Drug Administration (FDA) in the United States proposed a draft guidance to clarify the evaluation and approval of in vivo radiopharmaceuticals used in the diagnosis or monitoring of diseases. The purpose of the regulations is to respond to the requirements of the Food and Drug Administration Modernization Act of 1997 (FDAMA). In the regulations, certain types of indications for which the FDA may approve diagnostic radiopharmaceuticals are described. In addition, the criteria that the agency would use to evaluate the safety and effectiveness of a diagnostic radiopharmaceutical under the Federal Food, Drug, and Cosmetic Act (the act) and the Public Health Act (the PHS act) are included.

The FDA draft guidance^[3] requires special considerations in clinical design and analysis for medical imaging drugs and biologics. The ROC curve is recommended as one of several methods for the assessment in diagnostic validity, and ROC analysis can be used to evaluate diagnostic accuracy.

In 2004 a revised version of the guidance was issued, with no additional suggestions for the use of ROC analysis. In 2009, the European Medicines Agency (EMA) issued a revised version of a guidance originally adopted in 2001.^[4,5] The guidance states that although ROC curves are mainly considered to be a means for selecting appropriate cutoff points to be used prospectively in confirmatory trials, the area under the curve may also be used as a measure of diagnostic performance.

STATISTICAL DESIGN/METHOD/ANALYSIS

Construction of a Receiver Operating Characteristic Curve

As mentioned in the introduction, an ROC curve is a plot of sensitivity (or true positive fraction, TPF) vs. 1-specificity (false positive fraction, FPF). These two measures together quantify the accuracy of a test or method in classifying an experimental unit into one of two categories, usually referred to as “positive” or “negative.” The construction of the ROC curve depends on a continuous or an ordinal scale representing a series of “cut-off points” based on the confidence threshold from the less to the most confidence in a positive decision. An example of the cut-off points in an ROC curve is shown in Table 1.

Using Table 1, different pairs of the TPF and FPF derived from five cut-off confidence thresholds may be generated. The ROC curve is then constructed by plotting these pairs of (FPF, TPF). An example of the ROC curve is illustrated in Fig. 1, which is contained in the unique square with scale ranging from 0 to 1 in both *x*- and *y*-axis. The better performance of a diagnostic test is indicated by the ROC curve passes near the upper left corner at (0,1), where the FPF is 0 (100% of specificity) and the TPF is 1 (100% of sensitivity). If the test performance were almost random, then the ROC curve would lie near the 45-degree diagonal line.

Given this finite set of points that determine the ROC curve, generation of the remaining points and other measures associated with the curve may be done by either assuming some underlying distribution or not doing so.

Table 1	An Example of Cut-Off Points Scale in ROC Curve				
Scale	– 2	– 1	0	1	2
Confidence threshold	Definitely negative	Probably negative	Not sure	Probably positive	Definitely positive

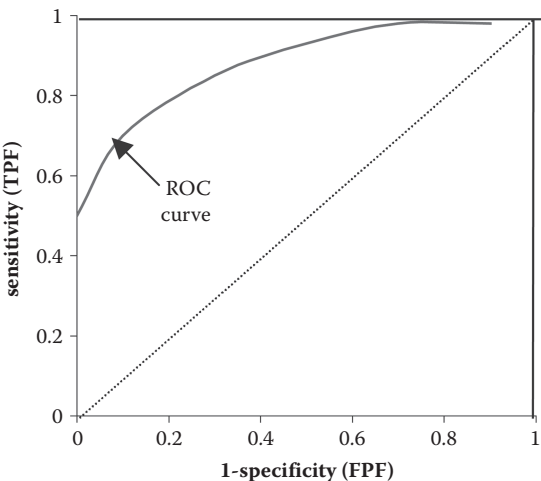


Fig. 1 Construction of an ROC curve

The most common distribution assumed for ROC curves is binormal distribution,^[2] which is described in the next section. For nonparametric approach, the methods are illustrated in the last section under biopharmaceutical applications of ROC curve.

Fitting a Binormal Receiver Operating Characteristic Curve

This family of curves results from assuming that the underlying data for both the positive and negative cases follow normal distributions, possibly with different parameters. This implies that the points on the ROC curves will fall on straight lines when plotted on axis transformed to normal scales. Each binormal ROC curve can be determined by two parameters. One possible way to parameterize these curves is by the slope and *y*-intercept of the line in “binormal space.”^[2] Given that μ_1 and σ_1 are the mean and standard deviation of the normal distribution of the negative or nondiseased units, and that μ_2 and σ_2 are the mean and standard deviation of the normal distribution of the positive or diseased units, then the “slope” parameter is $b = \sigma_1/\sigma_2$ and the intercept parameter is $a = (\mu_2 - \mu_1)/\sigma_2$.^[6]

The process of fitting the curve involves estimating the aforementioned parameters, *a* and *b*. Least squares methods are not appropriate for the estimation of *a* and *b*, but maximum likelihood techniques may be used because the assumptions that underlie the least squares methods are not valid for ROC data. The maximum likelihood estimation is used to produce parameter estimates as well as variability estimates.^[7]

Summary Measures for Receiver Operating Characteristic Curves

Despite the fact that no single number can define or characterize an ROC curve, questions may need to be answered that involve comparisons of two or more curves. These situations make desirable the calculation of meaningful summary statistics that best characterize the entire ROC curve. Several have been proposed, the most common of which are discussed subsequently.

Area Under the Receiver Operating Characteristic Curve

Although the ROC curve provides the diagnostic accuracy over the full range of decision thresholds, it would be helpful if the diagnostic performance could be quantified by a single number. One such measurement that can be derived from the ROC curve is the area under the curve (AUC). The AUC may be calculated by summing the areas of the individual trapezoids that comprise the region under the curve. The area under the ROC curve is a type of average of the ordinate values on the curve, and it therefore represents an average probability of correctly classifying a

case chosen at random. Also, the meaning of AUC has been proved to be the probability that a random pair of (FPF, TPF) individuals would be identified correctly by the diagnostic test.^[8] According to Hanley and McNeil in 1982,^[9] “The area under the ROC curve measures the probability, denoted by θ , that in randomly paired normal and abnormal images, the perceived abnormality of the two images will allow them to be correctly identified.” In other words, the greater the FPF and TPF values, the greater the AUC, which presents the better predictive ability. According to Hosmer and Lemeshow,^[10] the general rule of predictive ability for AUC is:

AUC	Predictive ability
0.5	No predictive ability
0.7 to <0.8	Acceptable predictive ability
0.8 to <0.9	Excellent predictive ability
≥0.9	Outstanding predictive ability

Based on the mathematical derivations, the area under the ROC curve is a function of the Wilcoxon rank-sum statistic.^[9] Based on this fact, therefore, the comparison of ROC curves may be calculated. The area under the binormal ROC curve is a simple function of the parameter estimates a and b mentioned earlier. In other words, the area under the binormal ROC curve equals the area under the standard normal curve from $-\infty$ to $a/\sqrt{1+b^2}$.^[6]

Partial Area Under the Receiver Operating Characteristic Curve

The greatest disadvantage of the full area under the ROC curve is the equal weight given to all false positive error rates. Practically, there are points and/or regions of the ROC curve that are of more interest than others when evaluating the performance of one or more diagnostic procedures. For example, if tests with very low sensitivity or specificity are considered undesirable, then the regions at or near the extremes of the curve are of no practical interest. Given minimum acceptable values for sensitivity or specificity (or some other criteria by which to partition the curve), a sector of the ROC curve is defined, and the area of the region under this sector may be used as a measure of the accuracy. For a binormal ROC curve, this area can be expressed as a simple closed form integral involving the standard normal probability function.^[6] This may then be “standardized” to a measure more like an average accuracy by dividing the area by the absolute difference of the horizontal (false positive rate) coordinates.

Single Points on the Receiver Operating Characteristic Curve

An extreme variation on using a region of the ROC curve would be to use just some single predefined point as the

basis for evaluation. Typically, points are chosen on their respective ROC curve where either the sensitivity and specificity are equal or some predetermined point where either sensitivity or specificity is a given value, such as 90%. For example, some medical applications may require the diagnostic with high specificity, Linnet^[11] suggested comparing two sensitivities under a single fixed level of specificity, p_0 .

Logistic Regression and Receiver Operating Characteristic Curve

The probability of correctly predicting the event in ROC analysis, measured by the AUC, is similar to the predicted probability in the logistic regression model. For a single test with no covariate, the classification table based on the logistic regression is the same as the conventional diagnostic tests, which provide a dichotomous outcome. By changing the cut-off points in the classification table of the logistic regression model, the ROC curve can be constructed accordingly. The comparisons of different curves are then analyzed by the methods described before, such as AUC.

Biopharmaceutical Applications of Receiver Operating Characteristic Curve

Choosing a Cut-Off Point for Positive Result for Confirmatory Trials

If a given diagnostic test yields a continuous or near-continuous result that is used to classify a patient into one of two categories, the ROC curve is very useful in identifying the optimal value to use as a cut-off point for the classification. For example, if a biochemical marker is a candidate for use in diagnosing a medical condition, samples may be taken in subjects with and without the condition. The ROC curve produced by choosing cut-off points across the full range of values for the marker provides a full picture of test performance, and should be able to provide guidance as to the best choice of standard for classifying a subject as a positive or negative case.

Similarly, if a confidence scale is associated with the classification of patients into “diseased” or “nondiseased” categories, the trade-off between sensitivity and specificity by using various confidence levels to conclude disease state can be effectively compared using the ROC curve.

Characterizing Accuracy with Summary Measures

One desirable feature of the ROC curves is that they provide a measure of accuracy, which incorporates both sensitivity and specificity. Clearly they have to be viewed as a pair when they are used to characterize the accuracy of a diagnostic method. Merely providing one of these two rates as evidence is generally not acceptable. Each of the

previously mentioned summary measures for ROC curves is a single number that incorporates both sensitivity and specificity. This collapsing of a two-dimensional situation to a single dimension serves an important statistical purpose, as it makes possible the specification of simple decision rules based on a single parameter. For example, if a given method is known to have a sensitivity of 90% and a specificity of 60%, and an alternate method has a sensitivity of 85% and specificity 80%, there is not a clear ordering of the performance of the two methods. Because ROC curve indices will be single numbers, a natural comparison by order is evident.

It is interesting to note the relationship between one of the most commonly used ROC summary measures, the entire area under the ROC curve, and the sometimes criticized measure of overall accuracy. If an ROC analysis is performed using a single sensitivity and specificity, the resulting empirical curve would be a quadrilateral whose area is equal to the average of the sensitivity and specificity. This can be seen by considering the three points in the unit square that would result: (0,0), (FPF, TPF), and (1,1). The areas of the resulting triangle and trapezoid are $1/2(\text{FPF})(\text{TPF})$ and $1/2(\text{TPF} + 1)(1 - \text{FPF})$, respectively. By adding both areas would result $1/2(\text{TPF} + 1 - \text{FPF})$, which is the average of sensitivity and specificity. Given equal numbers of positive and negative cases, this quantity is exactly the same as the overall accuracy.

Combining the Separate Measures of Accuracy and Diagnostic Confidence

In some studies, both the correctness of diagnosis and a separate measurement of confidence for the diagnosis are available. ROC analysis can be used to combine the accuracy and the diagnostic confidence, provided that the diagnosis and standard of reference are dichotomies (positive/negative or diseased/nondiseased). For example, if a diagnostic method yields results of + and – for a target disease and a rating physician provides a response of 0 = uncertain, 1 = fairly confident, 2 = very confident with his/her diagnosis, the pair of responses may be transformed into the following scale:

Combining assessment	Very confident case of –	Fairly confident case of –	Uncertain	Fairly confident case of +	Very confident case of +
Scale	– 2	– 1	0	1	2

With this transformation, the area under the ROC curve could be used to analyze accuracy and confidence simultaneously. Higher areas result from having high confidence when diagnoses are correct, and lower confidence when diagnoses are incorrect. These characteristics provide a reasonable variable for a combined assessment of diagnostic accuracy and diagnostic confidence.

Comparing Treatments or Methods Using Receiver Operating Characteristic Summary Measures

ROC curves provide an easy way to compare the accuracy of two or more different diagnostic tests or methods. The direct comparisons for treatments or methods are by plotting the ROC curves on the same set of axes. If superiority or equivalence is the question of interest, one or more summary measures from the curves may be compared quantitatively using a statistical method appropriate for the design of the experiment producing the data.

Hanley and McNeil^[9] derived a nonparametric method for comparing the areas under two ROC curves by calculating asymptotic standard errors for the areas assuming independent approximately continuous distributions for the curves. They use the relationship between the AUC and the Wilcoxon statistic, w , to calculate the standard error of the area under the curve without distributional assumptions as below:

$$SE(w) = \sqrt{\frac{\theta(1-\theta) + (n_A - 1)(Q_1 - \theta^2) + (n_N - 1)(Q_2 - \theta^2)}{n_A n_N}}$$

where θ is the estimate of AUC; n_A and n_N are the number of abnormal and normal cases, respectively; Q_1 is the probability that two randomly chosen abnormal images (or patients) are both ranked with greater abnormal test results than a randomly chosen normal image (patient); and Q_2 is the probability that one randomly chosen abnormal image (patient) is ranked with greater abnormal result than two randomly chosen normal images (patients).

The standard errors may be used to test the difference between two areas (AUC_1 and AUC_2) that are derived based on two independent groups. The test statistic z that follows a standard normal distribution is given as below:

$$z = \frac{AUC_1 - AUC_2}{\sqrt{[SE(AUC_1)]^2 + [SE(AUC_2)]^2}}$$

A method for comparing two ROC curves arising from correlated data was proposed by Hanley and McNeil in 1983.^[12] They suggested calculating the area and standard error estimates for the individual curves in the same way as the independent case described earlier. The standard error for the difference of the two curves is then calculated as

$$SE(AUC_1 - AUC_2) = \sqrt{SE^2(AUC_1) + SE^2(AUC_2) - 2rSE(AUC_1)SE(AUC_2)}$$

The estimation of the correlation r between the two areas is a function of the correlations of the individual ratings calculated separately for diseased and nondiseased patients.

The test statistic for the paired case is therefore

$$z = \frac{AUC_1 - AUC_2}{\sqrt{[SE(AUC_1)]^2 + [SE(AUC_2)]^2 - 2r[SE(AUC_1)][SE(AUC_2)]}}$$

where r is the Pearson correlation between AUC_1 and AUC_2 .

DeLong et al.^[13] proposed a different nonparametric method for comparing two correlated ROC curves. Their method exploits the properties of the Mann-Whitney statistics, whose relationship to ROC areas has been mentioned previously. They pointed out that the Mann-Whitney statistics are part of the set of generalized U -statistics. They further derived a test statistics for the difference of two correlated ROC curves based on the generalized U -statistics.

Thompson and Zucchini^[6] proposed an analysis of variance (ANOVA) model to compare accuracy indices between different modalities. Here, the accuracy index may be defined as the true positive rate, the partial area under the ROC curve, the whole area under the curve, or some other predefined measure. Their model allows for multiple readers and includes components for modalities, between readers, and within readers.

Campbell^[14] recommended a different alternative to ROC area for use as an index in comparing two curves, which result from continuous measurements on the same set of subjects. The proposed test is based on the maximum distances between the two curves, similar to the Kolmogorov-Smirnov tests for agreement of distributions. Bandos et al.^[15] proposed an exact non-parametric permutation test for comparing AUCs from 2 correlated ROC curves. They also give an asymptotic version of the procedure.

Another alternative approach to analyze ROC areas was proposed by Venkatraman and Begg.^[16] They presented a distribution-free permutation test for comparing ROC curves based on continuous data from a paired design. Their method attempts to assess the equality of the actual curves rather than some derived index. Liu et al. proposed methods for assessing equivalence and non-inferiority of ROC curves from paired data.^[17] They propose using the “two one-sided tests” approach and employing bootstrapping techniques to approximate the sampling distributions. Li et al.^[18] developed a non-inferiority test for paired ROC curves based on partial areas. They use the concept of a generalized p -value to construct their test procedure.

Combining Sensitivities and Specificities from Several Clinical Trials—Meta-Analysis Using a Summary Receiver Operating Characteristic Curve

If several estimates for the sensitivity and specificity of a given test are available from different sources, it may be necessary to derive a single estimate from these values. Recognizing the interrelationship between these two quantities, the use of the averages of the sensitivity and specificity values is generally not recommended. The approach proposed by Moses et al.^[19] to handle this problem is the summary ROC curve. They proposed some methods to estimate the summaries of ROC curve and how to characterize their adequacy. The underlying assumptions of their methods are the independence of studies, and each study

should provide an estimated FPR and an estimated TPR. These rates are then converted into their logistic transforms, U and V , respectively. Then for each study $V - U$, the log odds ratio, and $V + U$ are calculated and plotted. A type of regression line is then fitted to these points, and the line is backtransformed into ROC space. This produces a curve in the unit square, which looks like a traditional ROC curve. It is then suggested that any estimates that are to be made from the summary ROC curve be based only on “relevant” values of TPR and FPR (TPR > 0.5 and FPR < 0.5 is put forth as a possibility).

Walter^[20] explored the use of partial areas under the SROC curve and concluded that this measure was more sensitive to heterogeneity than full AUCs. Chappell et al.^[21] review several models for performing SROC, and provide some guidance for when these methods are appropriate.

Some of the aforementioned methods in the previous sections and other more advanced methods are included in a text by Zhou, Obuchowski, and McClish.^[22]

Sample Size Calculation

Hanley and McNeil^[9] provide a method for estimating sample size for comparing the areas under two independent ROC curves, given that equal numbers of normal and abnormal can be obtained. The formula is given as below

$$n_N = n_A = \left[\frac{z_\alpha \sqrt{2V_1} + z_\beta \sqrt{V_1 + V_2}}{\delta} \right]^2$$

where $V_1 = Q_1 + Q_2 - 2\theta_1^2$, $V_2 = Q_1 + Q_2 - 2\theta_2^2$, $\delta = \theta_1 - \theta_2$, z_α and z_β are the upper 100 α and 100 β percentile points of a standard normal distribution. In V_1 , Q_1 and Q_2 are defined as $Q_1 = \theta_1/(2 - \theta_1)$, $Q_2 = 2\theta_1^2/(1 + \theta_1)$. In V_2 , $Q_1 = \theta_2/(2 - \theta_2)$, $Q_2 = 2\theta_2^2/(1 + \theta_2)$.

Obuchowski and McClish^[23] proposed a method for calculating sample size assuming a binormal ROC curve. In their method, the variance and covariance of the parameters a and b of the ROC curve are estimated based on the Taylor expansions and the derivation of large sample theory.

Obuchowski^[24] proposed a method for assessing the equivalence of diagnostic tests based on ROC curve methodology. Following ideas from bioequivalence in pharmacokinetic studies, the criteria of “population equivalence” and “individual equivalence” were developed for diagnostic equivalence. The estimates for the equivalency parameters were established and the confidence intervals were constructed by using the bootstrap methods for both criteria. Based on the results from a Monte Carlo simulation study, Obuchowski^[24] concluded that the power for detecting equivalence of ROC areas was reasonably close to the power for detecting a difference of δ in ROC areas. Therefore, similar numbers of patients for testing the equivalence and detecting a difference would be needed for the comparison of diagnostic tests.

An Example

In this example, the data set from Hanley and McNeil^[12] was used to illustrate the ROC analysis. The data set consists of ratings (from 1 = definitely normal to 6 = definitely abnormal) given to two modalities used on each of 112 phantoms. The purpose of this analysis is to study the accuracy of two computer algorithms (modalities) used in image reconstruction for computed tomography (CT). The data is illustrated (see table below) by the phantom status (normal, abnormal):

	Ratings with modality 2						
Ratings with modality 1	1	2	3	4	5	6	Total
Frequency counts of ratings by modality for normal CT phantoms							
1	9	3	0	0	0	0	12
2	17	9	2	0	0	0	28
3	3	4	1	0	0	0	8
4	3	2	2	1	0	0	6
5	3	1	0	2	0	0	4
6	0	0	0	0	0	0	0
Total	31	19	5	3	0	0	58
Frequency counts of ratings by modality for abnormal CT phantoms							
1	0	0	1	0	0	0	1
2	1	0	2	0	0	0	3
3	1	1	1	3	0	0	6
4	1	1	1	9	1	0	13
5	0	0	0	7	10	5	22
6	0	0	0	0	4	5	9
Total	3	2	5	17	15	10	54

The results of ROC analysis are summarized in the following table:

Modality	Slope	Intercept	Z_A	AUC	SE(AUC)	r	z
Modality 1	0.945	1.72	1.25	0.8945	0.03	0.40	1.41
Modality 2	0.467	1.70	1.54	0.9382	0.026		

In the table, r is the correlation of AUCs from two modalities, and z is the test statistic to test the null hypothesis that the difference between two areas is zero. Based on a significance level of 5%, it concluded that there is no difference in accuracy of two different computer algorithms used in image reconstruction for CT. The ROC curves are shown in Fig. 2.

SCIENTIFIC ISSUES

The utility of ROC curve analysis to demonstrate or illustrate the trade-off between sensitivity and specificity of diagnostic tests or instruments for signal detection has rarely if ever been questioned. It is the application of ROC methods for other purposes, such as demonstrating or comparing diagnostic efficacy of tests or methods, which has evoked criticism.

One source of controversy alluded to previously is the lack of agreement on a single summary measure. Although the total AUC uses all the information (points) on the ROC curve, it includes sensitivity, specificity pairs that are of little or no interest. It has been pointed out that tests that are clearly superior may have similar total AUCs to inferior tests.^[14] Variations such as partial AUC alleviate this feature to some extent, but the bounds of the area must be somewhat subjective. Likewise, single points based on a predetermined sensitivity or specificity value require some subjective choice of standard, and this choice for a given test sample may prove inappropriate (if it falls far out of the range of observed values, for example). Unfortunately,

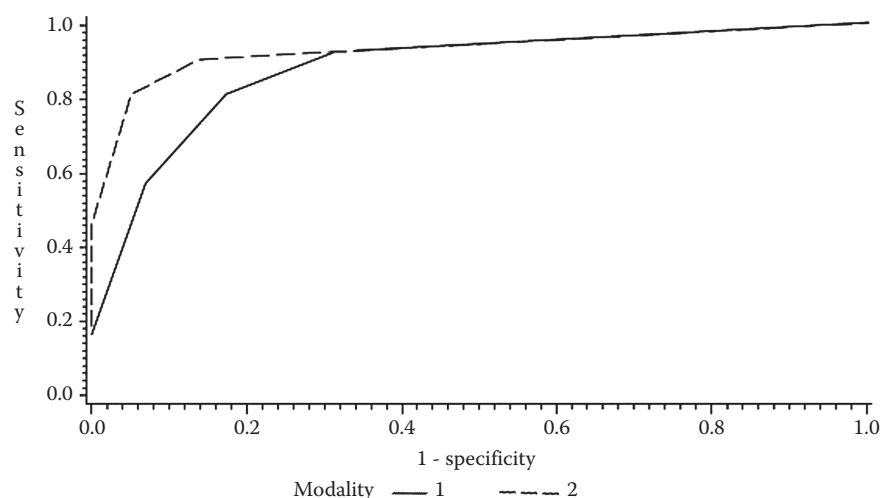


Fig. 2 ROC curves for Hanley and McNeil data

there is no perfect summary measure that can characterize an entire ROC curve.

Other problems arising from the use of ROC curves to quantify the accuracy of diagnostic tests arise from the definition of sensitivity and specificity rather than from the construction of the curve. Whenever there is not a perfect dichotomy for the classification of test subjects, inconsistencies may arise in the calculation of these and/or other related predictive rates. Although a simple solution to this problem may seem to be just to force a dichotomy, it is not always practical to do this. In studies of diagnostic devices or contrast agents for these devices, a result such "not assessable" is often difficult to avoid even in the best of circumstances. Although it is possible to alter the classical definitions of sensitivity and specificity to reasonably accommodate this extra category, these altered rates invariably lead to problems with the related predictive rates and with the construction of the usual ROC curves.

As with all statistical methods, missing data create problems in ROC curve analysis that do not have perfect solutions. Leaving out partial observations (those where either the test value or the standard of reference are missing) is always open to criticism because it leads to possibly biased results. Any attempt to include partial data utilizes direct or implied imputations, which must rely on some set of assumptions. Because the validity of these assumptions is always open to debate for any given situation, any method to handle missing data in ROC curve analysis may be criticized. This is not to say that they should not be used, because when there is a substantial amount of missing data, better estimates will most likely result from the methods using all cases. One method, proposed by Zhou and Gatsonis^[25] for handling a certain type of missing ROC data utilizes a derived covariance matrix of nonparametric estimators of ROC AUCs for incomplete paired data. They then use this estimator to test the difference of the areas using all available cases.

CONCLUSION

Two assumptions had been used for evaluating and comparing sensitivity and specificity in diagnostic tests. They are the following: (i) diagnostic tests are independent given the disease status and (ii) the gold standard (reference) is error free. For cases where the assumptions are not valid, several approaches have been investigated. Qu and Hadgu^[26] suggested some statistical methods for these cases that: diagnostic tests are correlated, gold standard is imperfect, and diagnostic tests are correlated when the gold standard is imperfect. Alternatively, Baker et al.^[27] proposed the methods for estimating sensitivity and specificity without a gold standard.

Another clinical practice in diagnostic imaging is that the accuracy of an imaging study is very often to be evaluated by multiple readers or/and patients to be examined

by multiple modalities. This was investigated by Toledano and Gatsonis,^[28] where the application of analysis of ROC data from multiple interpretations of the same diagnostic study and from the examination of the same patients with multiple diagnostic modalities was to be performed. They proposed the use of ordinal regression models in conjunction with generalized estimating equations to analyze the correlated ROC data.

Pepe^[29] interprets each point on the ROC curve as a conditional probability of a test result from a random diseased subject exceeding that from a random nondiseased subject. The ROC curves were then estimated by generalized linear model procedures. It has been observed that many disease outcomes are time dependent. Therefore, it is more appropriate to consider the ROC curve as a function of time. Heagerty et al.^[30] proposed the statistical method for time-dependent ROC curves for survival data.

The overview of ROC analysis for evaluation of the performance of diagnostic medical products can be found in Chow and Shao.^[31]

REFERENCES

1. Zweig, M.H.; Campbell, G. Receiver operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clin. Chem.* **1993**, *39*(4), 561–577.
2. Metz, C.E. ROC methodology in radiologic imaging. *Invest. Radiology* **1986**, *21*, 720–733.
3. FDA. *Draft Guidance for Industry: Development Medical Imaging Drugs and Biologics*; FDA: Rockville, MD, 1998.
4. FDA. *Draft Guidance for Industry: Development Medical Imaging Drugs and Biologics*; FDA: Rockville, MD, 2004.
5. EMEA. *Guideline on Clinical Evaluation of Diagnostic Agents*; European Medicines Agency (EMA): London, 2009.
6. Thompson, M.L.; Zucchini, W. On the statistical analysis of ROC curves. *Stat. Med.* **1989**, *8*, 1277–1290.
7. Metz, C.E.; Herman, B.A.; Shen, J.H. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. *Stat. Med.* **1998**, *17*, 1033–1053.
8. Green, D.M.; Swets, J.A. *Signal Detection Theory and Psychophysics*; John Wiley & Sons, Inc.: New York, 1966.
9. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 129–133.
10. Hosmer, D.W.; Lemeshow, S. *Applied Logistic Regression*, 2nd Ed.; John Wiley & Sons, Inc.: New York, 2000.
11. Linnet, K. Comparison of quantitative diagnostic tests: Type I error, power, and sample size. *Stat. Med.* **1987**, *6*, 147–155.
12. Hanley, J.A.; McNeil, B.J. A method of comparing areas under receiver operating characteristic curves derived from the same cases. *Radiology* **1983**, *148*, 839–843.
13. Delong, E.; Delong, D.; Clarke-Pearson, D. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **1988**, *44*, 837–845.

14. Campbell, G. General methodology I—Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Stat. Med.* **1994**, *13*, 499–508.
15. Bandos, A.I.; Rockette, H.E.; Gur, D. A permutation test sensitive to differences in areas for comparing ROC curves from a paired design. *Stat. Med.* **2005**, *24*, 2872–2893.
16. Venkatraman, E.S.; Begg, C.B. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika* **1996**, *83* (4), 835–848.
17. Liu, J.-P.; Ma, M.-C.; Wu, C.-Y.; Tai, J.-Y. Tests of equivalence and non-inferiority for diagnostic accuracy based on paired areas under ROC curves. *Stat. Med.* **2006**, *25*, 1219–1238.
18. Li, C.-R.; Liao, C.-T.; Liu, J.-P. A non-inferiority test for diagnostic accuracy based on paired partial areas under ROC curves. *Stat. Med.* **2008**, *27*, 1762–1776.
19. Moses, L.E.; Shapiro, D.; Littenberg, B. Combining independent studies of a diagnostic test into a summary ROC curve: Data-analytic approaches and some additional considerations. *Stat. Med.* **1993**, *12*, 1293–1316.
20. Walter, S.D. The partial area under a summary ROC curve. *Stat. Med.* **2005**, *24*, 2025–2040.
21. Chappell, F.M.; Raab, G.M.; Wardlaw, J.M. Where are summary ROC curves appropriate for diagnostic meta-analyses? *Stat. Med.* **2009**, *28*, 2653–2668.
22. Zhou, X.H.; Obuchowski, N.A.; McClish, D.K. *Statistical Methods in Diagnostic Imaging*, Wiley Interscience: New York, 2002.
23. Obuchowski, N.; McClish, D. Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. *Stat. Med.* **1997**, *16*, 1529–1542.
24. Obuchowski, N. Can electronic medical images replace hard-copy film? Defining and testing the equivalence of diagnostic tests. *Stat. Med.* **2001**, *20*, 2845–2863.
25. Zhou, X.H.; Gatsonis, C.A. A simple method for comparing correlated ROC curves using incomplete data. *Stat. Med.* **1996**, *15*, 1687–1693.
26. Qu, Y.; Hadgu, A. A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. *J. Am. Stat. Assoc.* **1998**, *93*, 920–928.
27. Baker, S.G.; Cannor, R.J.; Kessler, L.G. The partial testing design: A less costly way to test equivalence for sensitivity and specificity. *Stat. Med.* **1998**, *17*, 2219–2232.
28. Toledano, A.; Gatsonis, C. Ordinal regression methodology for ROC curves derived from correlated data. *Stat. Med.* **1996**, *15*, 1807–1826.
29. Pepe, M.S. An interpretation for the ROC curve and inference using GLM procedures. *Biometrics* **2000**, *56* (2), 352–359.
30. Heagerty, P.J.; Lumley, T.; Pepe, M.S. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **2000**, *56* (2), 337–344.
31. Chow, S.C.; Shao, J. *Statistics in Drug Research—Methodology and Recent Development*; Marcel Dekker, Inc.: New York, 2002.

Sample Size Calculation: Nonparametrics

Hansheng Wang

Peking University, Beijing, P. R. China

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

The objective of sample size calculation is to select an appropriate sample size so that not only the Type I error probability can be controlled under the specified level when the null hypothesis is true, but also to assure sufficient power for detection of a clinically meaningful difference when the alternative hypothesis is true. This entry reviews the sign test for comparing median and tests for independence and also provides simulation results and numerical examples.

INTRODUCTION

Sample size calculation plays an important role in clinical research.^[1] The objective of sample size calculation is to select an appropriate sample size so that not only the Type I error probability can be controlled under the specified level when the null hypothesis is true, but also to assure sufficient power for detection of a clinically meaningful difference when the alternative hypothesis is true.

In biopharmaceutical research, various nonparametric tests are often used to serve different scientific purposes when parametric assumptions are not met.^[2] Four non-parametric tests including sign test, one-sample Wilcoxon's test, two-sample Wilcoxon's test, and Kendall's test for independence are commonly used.

OVERVIEW

Under the null hypothesis, these test statistics are asymptotically distributed as a standard normal distribution. The knowledge of their distribution under the alternative hypothesis is essential to derive formula for sample size calculation. To achieve this objective, two different approaches are usually considered.

One is to derive the asymptotic distribution of the test statistics under the so-called local alternatives when the null hypothesis is false.^[3] Local alternative assumes that an appropriately defined difference between the true parameter and the reference value (or control) vanish to 0 at the speed of $n^{-1/2}$. A practical consequence of this assumption is that the test statistic's asymptotic variance under the local alternative stays the same as the asymptotic variance under the null hypothesis. In other words, the effect of the local alternative is just changing the mean of the

asymptotic distribution of the testing statistic but not its variance. The merit of this method is that the test statistic has a nondegenerate asymptotic distribution under the alternative hypothesis. Hence a formula for sample size calculation can be easily derived. In practice, the alternative hypothesis is usually fixed and they do not vanish to 0 as sample size goes larger. As a result, the idea of local alternative may not be easily understood and accepted by scientists. This induces another approach based on fixed alternatives.

Fixed alternative approach assumes that the appropriately defined difference between the true parameter and the reference (or control) stays fixed under the alternative hypothesis.^[4,5] With appropriate normalization and approximation, explicit formula for sample size calculation based on the four nonparametric tests can be obtained. The merit of this fixed alternative approach is that it is easily understood and accepted in practice. When the alternative hypothesis is fixed, one may expect the fixed-alternative-based formulae to have better precision than the local-alternative-based formula. The drawback of this approach is that it usually requires the knowledge of more parameters than the local-alternative-based formula. In practice, the parameters used to estimate the sample size are usually unknown. As a result, it has to be estimated from pilot studies, historical data, or literatures. As a result, the fact that more parameters need to be estimated for fixed-alternative-based formulae may imply less reliable sample size estimation if the parameter estimates themselves are not accurate enough.

In the next section, the sign test for comparing median is studied. In the succeeding sections, one-sample Wilcoxon's test, two-sample Wilcoxon's test, and Kendall's test for independence are discussed. In each section, formulae for sample size calculation based on both local

and fixed alternatives are derived. For illustration purposes, simulation results and numerical examples are also presented. For technical details, the readers should refer to Refs. [3–5].

SIGN TEST

Introduction

Let $x_1, \dots, x_n$ be independent and identically distributed continuous random samples with median θ , i.e., $P(x_i \leq \theta) = 0.5$. In biopharmaceutical research, it is often of interest to test the following hypothesis:

$$H_0: \theta = \theta_0 \quad \text{vs.} \quad H_1: \theta \neq \theta_0$$

Define $Z_i = I(x_i > \theta_0)$, where $I(\cdot)$ is the indicator function. Let $p = P(Z_i = 1)$, then the above hypotheses become

$$H_0: p = 0.5 \quad \text{vs.} \quad H_1: p \neq 0.5$$

As a result, we may consider the following test statistic

$$T = 2\sqrt{n}(\bar{z} - 0.5)$$

where $\bar{z}$ is the sample mean of $z_1, \dots, z_n$. Under the null hypothesis, T is asymptotically distributed as a standard normal distribution. As a result, we reject the null hypothesis at the α level of significance if $|T| > z_{\alpha/2}$.

Fixed Alternative

Under the fixed alternative hypothesis that $p \neq p_0$, the power of the above test is approximately given by

$$\Phi\left(\frac{\sqrt{n}|p - 0.5| + 0.5z_{\alpha/2}}{\sqrt{p(1-p)}}\right)$$

As a result, the sample size needed for achieving a desired power of $1 - \beta$ can be obtained by solving the following equation

$$\frac{\sqrt{n}|p - 0.5| + 0.5z_{\alpha/2}}{\sqrt{p(1-p)}} = -z_{\beta}$$

This leads to

$$n_f = \left(\frac{0.5z_{\alpha/2} + z_{\beta}\sqrt{p(1-p)}}{p - 0.5}\right)^2$$

Local Alternative

Under a local alternative, it is assumed that the difference between p and 0.5 is sufficiently small so that

$$\sqrt{p(1-p)} \approx 0.5$$

Thus n_f can be approximated by

$$n_l = \frac{(z_{\alpha/2} + z_{\beta})^2}{4(p - 0.5)^2}$$

which leads to the formula given in Ref. [3].

A Simulation Study

A simulation study was conducted to evaluate the performance of the derived sample size formulae for both fixed and local alternatives. x_i 's are simulated as independent and identically continuous responses with $P(x_i \leq \theta_0) = p$. For each value of p , sample sizes needed to achieve 80% power at the 5% level of significance are estimated. Then, using the calculated sample size, the true power is simulated based on the 10,000 simulations. Table 1 summarizes the results from the simulation. As it can be seen, the difference between the sample sizes determined based on fixed and local alternatives are very small. When the estimated sample size is relatively large, they have very similar performance while the local alternative approach is a little bit more conservative. When the estimated sample size is relatively small, the fixed alternative approach performs slightly better than the local alternative approach.

An Example

To illustrate the use of sample size formulae derived above, we consider an example concerning a clinical trial for evaluation of the effect of a test drug with some continuous clinical endpoint. Suppose that the investigator is interested in testing whether 0 is the median of this distribution. Let p denote the true probability of this continuous response to be less than 0. Then, it is equivalent to test whether $p = 0.5$. Suppose that, based on historical data, it is estimated that $p = 0.7$; hence the sample size needed to achieve an 80% power for detection of a clinically meaningful improvement. On the other hand, the local alternative approach gives

$$\begin{aligned} n_f &= \left(\frac{0.5z_{\alpha/2} + z_{\beta}\sqrt{p(1-p)}}{p - 0.5}\right)^2 \\ &= \left(\frac{0.5 \times 1.96 + 0.84\sqrt{0.7(1-0.7)}}{0.7 - 0.5}\right)^2 \\ &= 46.58 \approx 47 \end{aligned}$$

Thus a total of 47 subjects are needed to have an 80% power to reject the null hypothesis. On the other hand, using a local alternative formula, it can be obtained that

$$n_l = \frac{(z_{\alpha/2} + z_{\beta})^2}{4(p - 0.5)^2} = \frac{(1.96 + 0.84)^2}{4(0.7 - 0.5)^2} \approx 49$$

Table 1 Sample Size n and Actual Power for Sign Test with Estimated p with Power -0.80 {10,000 Simulations}

P	Fixed alternative		Local alternative		p	Fixed alternative		Local alternative	
	n	True power	n	True power		n	True power	n	True power
0.602	187	0.814	189	0.821	0.662	73	0.832	75	0.840
0.604	180	0.785	182	0.796	0.664	71	0.823	73	0.842
0.606	173	0.796	175	0.802	0.666	69	0.817	72	0.800
0.608	166	0.807	169	0.792	0.668	68	0.772	70	0.792
0.610	160	0.794	163	0.789	0.670	66	0.835	68	0.788
0.612	155	0.818	157	0.819	0.672	64	0.825	67	0.824
0.614	149	0.799	151	0.808	0.674	63	0.789	65	0.809
0.616	144	0.814	146	0.827	0.676	61	0.777	64	0.846
0.618	139	0.786	141	0.793	0.678	60	0.810	62	0.822
0.620	134	0.795	137	0.828	0.680	59	0.768	61	0.800
0.622	130	0.789	132	0.799	0.682	57	0.835	60	0.837
0.624	126	0.781	128	0.794	0.684	56	0.793	58	0.826
0.626	122	0.823	124	0.829	0.686	55	0.823	57	0.843
0.628	118	0.812	120	0.819	0.688	54	0.784	56	0.805
0.630	114	0.797	117	0.791	0.690	52	0.765	55	0.851
0.632	111	0.825	113	0.832	0.692	51	0.808	54	0.801
0.634	107	0.813	110	0.799	0.694	50	0.837	53	0.834
0.636	104	0.829	107	0.820	0.696	49	0.799	52	0.789
0.638	101	0.793	104	0.837	0.698	48	0.829	51	0.831
0.640	98	0.811	101	0.802	0.700	47	0.783	50	0.854
0.642	95	0.764	98	0.825	0.702	46	0.819	49	0.825
0.644	93	0.828	95	0.780	0.704	45	0.759	48	0.851
0.646	90	0.794	93	0.845	0.706	44	0.807	47	0.800
0.648	88	0.780	90	0.795	0.708	43	0.837	46	0.840
0.650	85	0.800	88	0.803	0.710	43	0.845	45	0.792
0.652	83	0.801	85	0.813	0.712	42	0.789	44	0.827
0.654	81	0.793	83	0.814	0.714	41	0.827	43	0.863
0.656	79	0.794	81	0.810	0.716	40	0.781	43	0.866
0.658	77	0.780	79	0.803	0.718	39	0.817	42	0.821
0.660	75	0.835	77	0.785	0.720	39	0.823	41	0.850

ONE-SAMPLE WILCOXON'S TEST

Introduction

Let $x_1, \dots, x_n$ be independent and identically distributed samples generated from the following model

$$x_i = \theta + e_i, \quad i = 1, \dots, n$$

where θ is the unknown location parameter and the e_i 's are random error with a distribution symmetric about 0. The hypothesis of interest is given by

$$H_0: \theta = 0 \quad \text{vs.} \quad H_1: \theta \neq 0$$

To test the above hypotheses, one-sample Wilcoxon's test can be used. Let R_i be the rank of $|x_i|$ in $\{|x_1|, \dots, |x_n|\}$. Define

$$\psi_i = \begin{cases} 1 & \text{if } x_i > 0 \\ 0 & \text{if } x_i < 0 \end{cases}$$

where $i = 1, \dots, n$. Then, the statistic

$$T^+ = \sum_{i=1}^n R_i \psi_i$$

is the sum of the positive signed ranks. It can be verified that under the null hypothesis,

$$T = \frac{T^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

is asymptotically distributed as a standard normal distribution. Therefore we reject the null hypothesis at the α level of significance if $|T| > z_{1-\alpha/2}$.

Fixed Alternative

To determine the sample size under a fixed alternative, we need to evaluate the mean and variance of T^+ under an alternative hypothesis. It can be verified⁴ that the mean and variance of T^+ are given by

$$\begin{aligned} E(T^+) &= np_1 + n(n-1)p_2 \\ \text{var}(T^+) &= np_1(1-p_1) + n(n-1) \\ &\quad \times (p_1^2 - 4p_1p_2 + 3p_2 - 2p_2^2) \\ &\quad + n(n-1)(n-2) \\ &\quad \times (p_3 + 4p_4 - 4p_2^2) \end{aligned}$$

Where

$$\begin{aligned} p_1 &= P(x_1 > 0) \\ p_2 &= P(x_1 \geq |x_2|) \\ p_3 &= P(x_1 \geq |x_2|, x_1 \geq |x_3|) \\ p_4 &= P(x_1 \geq x_2 \geq |x_3|) \end{aligned}$$

Let $\sigma_+^2 = \text{var}(T^+)$. Then, under the alternative hypothesis, T^+ can be approximated by a normal random variable with mean $E(T^+) = np_1 + n(n-1)p_2$ and variance σ_+^2 . As a result, the power of the one-sample Wilcoxon's test can be approximated by

$$\Phi\left(\frac{z_{\alpha/2}/\sqrt{12} + \sqrt{n}|1/4 - p_2|}{\sqrt{p_3 + 4p_4 - 4p_2^2}}\right)$$

Hence the sample size required for achieving a desired power of $1 - \beta$ can be obtained by solving the following equation

$$\frac{z_{\alpha/2}/\sqrt{12} + \sqrt{n}|1/4 - p_2|}{\sqrt{p_3 + 4p_4 - 4p_2^2}} = -z_\beta$$

This gives

$$n_f = \left(\frac{z_{\alpha/2}/\sqrt{12} + z_\beta\sqrt{p_3 + 4p_4 - 4p_2^2}}{(1/4 - p_2)^2}\right)^2$$

Local Alternative

Under a local alternative, the difference between the null hypothesis and the alternative hypothesis is sufficiently small. As a result, the following approximations hold:

$p_2 \approx 1/4$, $p_3 \approx 1/6$, $p_4 \approx 1/24$. It follows that n_f can be approximated by

$$n_f = \frac{(z_{\alpha/2} + z_\beta)^2}{3(0.5 - 2p_2)^2}$$

If we define $p' = P(x_1 + x_2 \geq 0)$, then it follows that

$$\begin{aligned} p' &= P(x_1 + x_2 \geq 0) \\ &= 2P(x_1 + x_2 \geq 0 \text{ and } |x_1| \geq |x_2|) \\ &= 2P(x_1 \geq |x_2|) = 2p_2 \end{aligned}$$

As a result, n_f can be expressed as

$$n_f = \frac{(z_{\alpha/2} + z_\beta)^2}{3(p' - 0.5)^2}$$

which leads to the formula given in Ref. [3].

A Simulation Study

A simulation study was conducted to evaluate the performance of the derived sample size formulae for both fixed and local alternatives, x_i 's are generated from normal population with mean θ and variance 1. The p_i 's are estimated by the Monte Carlo method based on a sample of size 100,000. The estimated values of p_i 's are used to determine the sample size based on both fixed and local alternatives. Then, using the calculated sample size, the true power is simulated based on the 10,000 simulations. Table 2 summarizes the results from the simulation. As the table shows, the difference between the sample sizes determined based on fixed and local alternatives are very small. When the estimated sample size is relatively large, they have very similar performance while the local alternative approach is a little bit more conservative. When the estimated sample size is relatively small, the fixed alternative approach performs slightly better than the local alternative approach.

An Example

To illustrate the use of sample size formulae derived above, we consider an example concerning a clinical trial for evaluation of the effect of a test drug with some continuous clinical endpoint. Each patient participating in the trial will have two measurements. One is before the treatment, which serves as a baseline value, and the other one is after the treatment. The difference of those two measurements will serve as the primary efficacy endpoint and hence will be used to evaluate the possible treatment effect of the test drug. When there is no treatment effect, the distribution of the pretreatment and posttreatment should be the same. As a result, it is reasonable to assume that under the null hypothesis,

Table 2 Sample Size n and Actual Power for One-Sample Location with θ and Estimated p with Power = 0.80 (10,000 Simulations)

θ	p_1	p_2	p_3	p_4	Fixed alternative		Local alternative	
					n	True power	n	True power
0.350	0.636	0.345	0.240	0.076	68	0.799	73	0.824
0.355	0.636	0.345	0.240	0.078	71	0.825	74	0.838
0.360	0.643	0.348	0.243	0.077	64	0.788	68	0.817
0.365	0.641	0.346	0.239	0.077	67	0.813	71	0.847
0.370	0.643	0.351	0.244	0.078	60	0.787	65	0.824
0.375	0.646	0.351	0.245	0.077	59	0.789	65	0.828
0.380	0.648	0.351	0.244	0.080	62	0.816	65	0.835
0.385	0.654	0.358	0.249	0.081	52	0.762	57	0.799
0.390	0.650	0.354	0.246	0.079	56	0.796	61	0.834
0.395	0.655	0.356	0.247	0.081	55	0.808	59	0.828
0.400	0.656	0.356	0.249	0.081	55	0.810	59	0.838
0.405	0.656	0.359	0.249	0.082	51	0.794	56	0.828
0.410	0.660	0.361	0.252	0.083	50	0.792	54	0.825
0.415	0.661	0.360	0.252	0.083	52	0.814	55	0.831
0.420	0.660	0.359	0.251	0.082	51	0.818	56	0.851
0.425	0.664	0.362	0.251	0.084	48	0.798	52	0.834
0.430	0.666	0.363	0.254	0.083	48	0.804	52	0.840
0.435	0.669	0.367	0.256	0.086	45	0.795	48	0.821
0.440	0.671	0.369	0.258	0.087	43	0.790	47	0.814
0.445	0.673	0.368	0.257	0.085	43	0.789	48	0.839
0.450	0.674	0.368	0.257	0.086	43	0.804	47	0.831
0.455	0.675	0.368	0.256	0.087	44	0.816	48	0.855
0.460	0.677	0.371	0.258	0.086	40	0.789	45	0.830
0.465	0.678	0.372	0.260	0.089	41	0.815	44	0.836
0.470	0.682	0.373	0.261	0.088	39	0.789	44	0.850
0.475	0.684	0.377	0.265	0.091	37	0.780	41	0.823
0.480	0.683	0.375	0.263	0.090	38	0.803	42	0.836
0.485	0.685	0.378	0.263	0.090	36	0.797	41	0.845
0.490	0.686	0.376	0.262	0.089	37	0.808	42	0.854
0.495	0.691	0.379	0.265	0.091	35	0.793	40	0.849
0.500	0.692	0.378	0.263	0.091	36	0.804	40	0.857

0 is the median of our primary efficacy endpoint. Historical data show that the distribution of the endpoint cannot be well approximated by normal distribution or other standard parametric distribution. On the other hand, it is also noted that the treatment effect (if exists) usually do not change the shape of the distribution of the clinical endpoint; instead, it only causes certain shift in the relative location of the distribution. As a result, the one-sample Wilcoxon's test can be used to evaluate the possible treatment effect. Suppose that based on historical data, it has been estimated that $p_2 = 0.360$, $p_3 = 0.252$, and $p_4 = 0.083$. Hence the sample size needed to achieve an 80% power for detection of a clinically meaningful improvement is given by

$$\begin{aligned}
 n_f &= \frac{\left(z_{\alpha/2}/\sqrt{12} + z_{\beta}\sqrt{p_3 + 4p_4 - 4p_2^2}\right)^2}{(1/4 - p_2)^2} \\
 &= \frac{(1.96/\sqrt{12} + 0.84\sqrt{0.252 + 4 \times 0.083 - 4 \times 0.360^2})^2}{(0.360 - 0.252)^2} \\
 &= 52.29 \approx 54
 \end{aligned}$$

Thus a total of 53 subjects are needed. On the other hand, the local alternative approach gives

$$n_l = \frac{(z_{\alpha/2} + z_{\beta})^2}{3(p' - 0.5)^2} = \frac{(1.96 + 0.84)^2}{3(2 \times 0.36 - 0.5)^2} = 53.99 \approx 54$$

TWO-SAMPLE WILCOXON'S TEST

Introduction

Let $x_i, i=1, \dots, n_1$ and $y_j, j=1, \dots, n_2$ be two independent and identically distributed random samples, it is assumed that $x_i (i=1, \dots, n_1)$ and $y_j (j=1, \dots, n_2)$ come from the same type of symmetric distribution but with different location parameter θ . The primary objective is to investigate whether the location parameter θ equals to 0, i.e.,

$$H_0: \theta = 0 \quad \text{vs.} \quad H_1: \theta \neq 0$$

To test the above hypotheses by Wilcoxon's rank sum test, we first order the $N = n_1 + n_2$ observations from least to greatest and let R_i denote the rank of y_i in this ordering. Let

$$W = \sum_{i=1}^{n_2} R_i$$

where R_i denotes the order of y_i in the total samples. Consider the following test statistic

$$W^* = \frac{W - E(W)}{\sqrt{\text{var}(W)}} = \frac{W - 0.5n_2(n_2 + n_1 + 1)}{\sqrt{n_1n_2(n_1 + n_2 + 1)/12}}$$

When sample sizes n_1 and n_2 are both sufficiently large, W^* is asymptotically distributed as a standard normal distribution when the null hypothesis is true. Therefore we reject H_0 at the α asymptotic level of significance if $|W^*| \geq z_{\alpha/2}$.

Fixed Alternative

Under a fixed alternative, the power of the two-sample Wilcoxon's test can be approximated by

$$\Phi \left(\frac{z_1 - \alpha/2 \sqrt{\kappa(1+\kappa)/12} + \sqrt{n_2\kappa|1/2 - p_1|}}{\sqrt{\kappa^2(p_2 - p_1^2) + \kappa(p_3 - p_1^2)}} \right)$$

where $\kappa = \lim(n_1/n_2)$ and

$$\begin{aligned} p_1 &= P(y_1 \geq x_1) \\ p_2 &= P(y_1 \geq x_1 \text{ and } y_1 \geq x_2) \\ p_3 &= P(y_1 \geq x_1 \text{ and } y_2 \geq x_1) \end{aligned}$$

For technical details, see Refs. [4,5]. As a result, the sample size needed to achieve the desired power of $1 - \beta$ can be obtained by solving the following equation

$$\frac{z_{\alpha/2} \sqrt{\kappa(1+\kappa)/12} + \sqrt{n_2\kappa(1/2 - p_1)}}{\sqrt{\kappa^2(p_2 - p_1^2) + \kappa(p_3 - p_1^2)}} = -z_\beta$$

This leads to (see below)

Local Alternative

Under a local alternative, the difference between the null hypothesis and the alternative hypothesis is sufficiently small. Thus the following approximations hold: $p_1 \approx 1/2$, $p_2 \approx 1/3$, and $p_3 \approx 1/3$. As a result, it follows that n_f can be approximated by

$$n_f = \frac{(1+\kappa)^2 (z_{\alpha/2} + z_\beta)^2}{12\kappa(p_1 - 1/2)^2}$$

Define $c = n_1/(n_1 + n_2) = 1/(1 + \kappa)$, it follows that

$$n_f = \frac{(z_{\alpha/2} + z_\beta)^2}{12c(1-c)(p_1 - 1/2)^2}$$

which leads to the formula given in Ref. [3].

A Simulation Study

A simulation study was conducted to evaluate the performance of the derived sample size formulae for both fixed and local alternatives. x_i 's are generated from normal population with mean 0 and variance 1, y_j 's are generated from normal population with mean θ and variance 1. The sample size ratio κ is chosen to be 1. The p_i 's are estimated by the Monte Carlo method based on a sample of size 100,000. The estimated values of p_i 's are used to determine the sample size based on both fixed and local alternatives. Using the calculated sample size, the true power is simulated based on 10,000 simulations. The results are given in Table 3. From the table, we can see that the difference between the sample sizes determined based on fixed and local alternatives are very small. When the estimated sample size is relatively large, they have very similar performance while the local alternative approach is a little bit more conservative. When the estimated sample size is relatively small, the fixed alternative approach performs slightly better than the local alternative approach.

An Example

To illustrate the use of sample size formulae derived above, we consider an example concerning a clinical trial for evaluation of the effect of a test drug with some continuous clinical endpoint. The historical data shows that the distribution of the endpoint cannot be well approximated by normal distribution or other standard parametric distribution. On the other hand, it is also noted that the treatment effect (if exists) usually do not change the shape of the distribution of the clinical endpoint; instead, it only causes certain shift in the relative location of the distribution. As a result, the two-sample Wilcoxon's test can be used to evaluate the possible treatment effect.

Table 3 Sample Size n_1 and Actual Power for Two-Sample Location with θ and Estimated p with Power 0.80 (10,000 Simulations)

θ	p_1	p_2	p_3	Fixed alternative		Local alternative	
				n_2	True power	n_2	True power
0.550	0.651	0.498	0.498	56	0.801	58	0.816
0.555	0.653	0.499	0.502	55	0.805	56	0.805
0.560	0.654	0.501	0.500	54	0.811	56	0.816
0.565	0.655	0.501	0.502	53	0.807	55	0.812
0.570	0.655	0.503	0.501	53	0.810	55	0.829
0.575	0.656	0.506	0.505	52	0.801	54	0.826
0.580	0.658	0.508	0.507	51	0.814	53	0.823
0.585	0.661	0.508	0.508	49	0.804	51	0.815
0.590	0.661	0.511	0.509	49	0.805	51	0.816
0.595	0.662	0.512	0.512	48	0.804	50	0.821
0.600	0.665	0.512	0.514	46	0.798	49	0.812
0.605	0.667	0.515	0.516	45	0.793	48	0.818
0.610	0.668	0.517	0.518	45	0.798	47	0.812
0.615	0.669	0.519	0.519	44	0.791	46	0.809
0.620	0.671	0.519	0.521	43	0.792	45	0.818
0.625	0.670	0.521	0.520	44	0.806	46	0.827
0.630	0.670	0.520	0.519	44	0.810	46	0.832
0.635	0.673	0.522	0.525	42	0.800	44	0.818
0.640	0.675	0.527	0.526	41	0.797	43	0.816
0.645	0.676	0.526	0.528	41	0.802	43	0.822
0.650	0.676	0.529	0.528	41	0.805	43	0.823
0.655	0.678	0.529	0.530	40	0.809	42	0.825
0.660	0.680	0.533	0.532	39	0.805	41	0.815
0.665	0.679	0.532	0.532	39	0.807	41	0.821
0.670	0.684	0.538	0.536	37	0.799	39	0.813
0.675	0.680	0.534	0.532	39	0.813	41	0.836
0.680	0.681	0.535	0.535	38	0.816	40	0.829
0.685	0.685	0.538	0.538	37	0.808	39	0.835
0.690	0.688	0.542	0.540	35	0.798	37	0.812
0.695	0.689	0.543	0.542	35	0.797	37	0.817
0.700	0.690	0.544	0.545	34	0.790	37	0.829

$n_1 = \kappa n_2$ and

$$n_2 = \frac{\left(z_{\alpha/2} \sqrt{\kappa(\kappa+1)/12} + z_{\beta} \sqrt{\kappa^2(p_2 - p_1^2) + \kappa(p_3 - p_1^2)} \right)^2}{\kappa^2(1/2 - p_1)^2}$$

Let $n_f = n_1 + n_2$, it follows that

$$n_f = \frac{(1 + \kappa) \left(z_{\alpha/2} \sqrt{\kappa(\kappa+1)/12} + z_{\beta} \sqrt{\kappa^2(p_2 - p_1^2) + \kappa(p_3 - p_1^2)} \right)^2}{\kappa^2(1/2 - p_1)^2}$$

The null hypothesis of interest is the one of no treatment difference. Suppose that based on historical data, it has

been estimated that $p_1 = 0.40$, $p_2 = 0.25$, and $p_3 = 0.25$. Hence the sample size needed to achieve an 80% power can be estimated by

$$\begin{aligned} n_f &= \frac{2 \left(z_{\alpha/2} / \sqrt{6} + z_{\beta} \sqrt{p_2 + p_3 - 2p_1^2} \right)^2}{(1/2 - p_1)^2} \\ &= \frac{2 \left(1.96 / \sqrt{6} + 0.84 \sqrt{0.25 + 0.25 - 2 \times 0.40^2} \right)^2}{(0.50 - 0.40)^2} \\ &= 267.52 \approx 268 \end{aligned}$$

Hence a total of 268 subjects are needed to have an 80% power for detection of a clinically meaningful

difference. On the other hand, the local alternative approach gives

$$\begin{aligned} n_l &= \frac{(z_{\alpha/2} + z_\beta)^2}{12c(1-c)(p_1 - 1/2)^2} \\ &= \frac{(1.96 + 0.84)^2}{12 \times 0.5(1 - 0.5)(0.4 - 0.5)^2} \\ &= 261.33 \approx 262 \end{aligned}$$

KENDALL'S TEST OF INDEPENDENCE

Introduction

In practice, it is often of interest to test if two continuous responses received from each subject (e.g., the baseline value and the posttreatment value) are actually independent of each other. For such data, if a bivariate normal assumption can be made, then testing independence becomes the problem of testing whether the correlation coefficient equals to 0. However, in many situations, this assumption does not hold. As a result, a certain nonparametric test for independence becomes desirable.

Let $(x_i, y_i), i = 1, \dots, n$, be the n independent and identically distributed bivariate observations from the n subjects involved in a clinical trial. To obtain a nonparametric test for independence between x_i and y_i , define

$$\tau = 2P((x_1 - x_2)(y_1 - y_2) > 0) - 1$$

where τ is the so-called Kendall coefficient. A necessary requirement for the independence of x_i and y_i is that $\tau = 0$. As a result, whether x_i and y_i are independent can be tested by considering the hypothesis $H_0: \tau = 0$. Under the null hypothesis, a nonparametric test can be obtained as follows. First, for $1 \leq i < j \leq n$, calculate $\zeta(x_i, x_j, y_i, y_j)$, where

$$\zeta(a, b, c, d) = \begin{cases} 1 & \text{if } (a - b)(c - d) > 0 \\ -1 & \text{if } (a - b)(c - d) < 0 \end{cases}$$

Consider

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \zeta(x_i, x_j, y_i, y_j)$$

Consider the following testing statistics

$$K^* = \frac{K - E(K)}{\sqrt{\text{var}(K)}} = K \left[\frac{n(n-1)(2n+5)}{18} \right]^{-1/2}$$

It can be shown that when sample size n is sufficiently large, K^* is asymptotically distributed as a standard normal distribution. Hence we would reject the null hypothesis at the α asymptotic level of significance for large samples if $|K^*| \geq z_{\alpha/2}$.

Fixed Alternative

Define

$$\begin{aligned} p_1 &= P((x_1 - x_2)(y_1 - y_2) > 0) \\ p_2 &= P((x_1 - x_2)(y_1 - y_2)(x_1 - x_3)(y_1 - y_3) > 0) \end{aligned}$$

Under the fixed alternative hypothesis that $p_1 \neq 0.5$ and assuming n is sufficiently large, the power of the above test can be approximated by

$$\Phi \left(\frac{z_{\alpha/2}/3 + \sqrt{n} |p_1 - 1/2|}{\sqrt{2p_2 - 1 - (2p_1 - 1)^2}} \right)$$

For the details of the above approximation, please refer to Ref. [4]. Hence the sample size needed to achieve a desired power of $1 - \beta$ can be obtained by solving the following equation

$$\frac{z_{\alpha/2}/3 - \sqrt{n} (p_1 - 1/2)}{\sqrt{2p_2 - 1 - (2p_1 - 1)^2}} = -z_\beta$$

This yields

$$n_f = \frac{\left[z_{\alpha/2}/3 + z_\beta \sqrt{2p_2 - 1 - (2p_1 - 1)^2} \right]^2}{(p_1 - 1/2)^2}$$

Local Alternative

Under a local alternative, the difference between the null hypothesis and the alternative hypothesis is sufficiently small. As a result, the following approximations hold: $p_1 \approx 1/2$ and $p_2 \approx 5/9$. It follows that n_l can be approximated by

$$n_l = \frac{(z_{\alpha/2} + z_\beta)^2}{9(p_1 - 1/2)^2}$$

which leads to the formula given in Ref. [3].

A Simulation Study

A simulation study was conducted to evaluate the performance of the derived sample size formulae for both fixed and local alternatives. (x_i, y_i) 's are generated in the following way: For any given correlation coefficient p , let

$$x_i = u_i \text{ and } y_i = \frac{p}{\sqrt{1-p^2}} u_i + v_i$$

Table 4 Sample Size n and Actual Power for Independence with Correlation Coefficient p and Estimated p with Power = 0.80 (10,000 Simulations)

p	p_1	p_2	Fixed alternative		Local alternative	
			n	True power	n	True power
0.300	0.597	0.570	90	0.793	93	0.798
0.305	0.597	0.571	90	0.799	91	0.806
0.310	0.600	0.569	83	0.780	86	0.802
0.315	0.601	0.571	83	0.798	85	0.807
0.320	0.605	0.573	76	0.775	78	0.782
0.325	0.608	0.573	72	0.772	74	0.778
0.330	0.606	0.573	74	0.785	77	0.805
0.335	0.607	0.570	73	0.794	76	0.813
0.340	0.610	0.573	69	0.791	72	0.811
0.345	0.611	0.576	69	0.801	71	0.811
0.350	0.616	0.576	62	0.764	64	0.774
0.355	0.616	0.574	61	0.773	64	0.792
0.360	0.621	0.579	57	0.756	59	0.768
0.365	0.620	0.576	57	0.761	60	0.790
0.370	0.617	0.576	61	0.809	64	0.825
0.375	0.621	0.576	56	0.781	59	0.802
0.380	0.625	0.580	54	0.780	56	0.794
0.385	0.627	0.581	51	0.766	54	0.792
0.390	0.628	0.578	49	0.759	53	0.803
0.395	0.628	0.582	51	0.785	53	0.806
0.400	0.634	0.582	45	0.742	48	0.778
0.405	0.631	0.583	48	0.788	50	0.798
0.410	0.637	0.584	44	0.758	46	0.786
0.415	0.637	0.582	43	0.767	46	0.789
0.420	0.639	0.585	42	0.753	44	0.788
0.425	0.637	0.586	44	0.784	46	0.819
0.430	0.640	0.588	42	0.780	44	0.803
0.435	0.645	0.586	38	0.751	41	0.783
0.440	0.646	0.588	38	0.766	40	0.791
0.445	0.645	0.587	38	0.773	41	0.803
0.450	0.651	0.588	35	0.752	38	0.780

where u_i and v_i are random samples generated from the standard normal distribution. The p_i 's are estimated by the Monte Carlo method based on a sample of size 100,000. The estimated values of p_i 's are used to determine the sample size based on both fixed and local alternatives. Then, using the calculated sample size, the true power is simulated based on 10,000 simulations. Table 4 summarizes the results. From the table, we can see that the difference between the sample sizes determined based on fixed and local alternatives are very small while the sample size determined by local alternative seems to be

slightly better than the sample size determined by fixed alternative.

An Example

Suppose x and y are two continuous responses in a clinical trial, where x represents the pretreatment measurement and y represents the posttreatment measurement. Past experiences show that these data cannot be fitted well by any standard parametric distribution. As a result, Kendall's test for independence is recommended to test whether there is

an association between the pretreatment and posttreatment measurement. Based on historical data, it is estimated that $p_1 = 0.60$ and $p_2 = 0.58$. Hence the sample size required for achieving an 80% power can be obtained as follows

$$\begin{aligned} n_f &= \frac{\left(z_{\alpha/2}/3 + z_\beta \sqrt{2p_2 - 1 - (2p_1 - 1)^2}\right)^2}{(p_1 - 1/2)^2} \\ &= \frac{\left(1.96/3 + 0.84 \sqrt{2 \times 0.58 - 1 - (1.20 - 1)^2}\right)^2}{(0.60 - 0.50)^2} \\ &= 89.17 \approx 90 \end{aligned}$$

Thus a total of 90 subjects are needed to achieve an 80% power to confirm the association between x and y . On the other hand, the local alternative approach gives

$$\begin{aligned} n_l &= \frac{\left(z_{\alpha/2} + z_\beta\right)^2}{9(p_1 - 0.5)^2} \\ &= \frac{(1.96 + 0.84)^2}{9(0.60 - 0.50)^2} \\ &= 87.11 \approx 88 \end{aligned}$$

CONCLUSION

In biopharmaceutical research, nonparametric methods are often used when the underlying distribution

assumption for parametric methods are not met. In this article, formulae for sample size calculation based on four commonly used nonparametric tests are derived. These tests include sign test, one-sample Wilcoxon's test, two-sample Wilcoxon's test, and Kendall's test of independence. Sample size formulae are derived under both local and fixed alternatives. Extensive simulations were performed to study the empirical performance of sample size derived under both fixed alternative and local alternative. Examples are given to illustrate the use of the derived formulae.

REFERENCES

1. Chow, S.C.; Liu, J.P. *Design and Analysis of Clinical Trials*, 2nd ed.; John Wiley & Sons: New York, 2003.
2. Hollander, M.; Wolfe, D.A. *Nonparametric Statistical Methods*; John Wiley & Sons: New York, 1973.
3. Noether, G.E. Sample size determination for some common nonparametric tests. *J. Am. Stat. Assoc.* **1987**, *82* (398), 645–647.
4. Chow, S.C.; Shao, J.; Wang, H. *Sample Size Calculation in Clinical Research*, 2nd Ed.; Chapman and Hall/CRC, Taylor & Francis: New York, 2008.
5. Wang, H.; Cheng, B.; Chow, S.C. Sample size determination based on rank tests in clinical research. *J. Biopharm. Stat.* **2003**, *13*, 731–747.

Sample Size Calculation: Survival Data

Qi Jiang and Steven Snapinn

Merck Research Laboratories, West Point, Pennsylvania, U.S.A.

Boris Iglewicz

Temple University, Philadelphia, Pennsylvania, U.S.A.

Abstract

Sample size calculation is an important aspect of the design of any clinical trial. This entry describes basic approaches for sample size calculation in survival studies and also discusses sample size calculation for group sequential survival trials.

INTRODUCTION

Accurate determination of the required sample size for survival studies involves various factors that either are unique to these studies or are not typically considered in non-survival studies. These include the survival distributions, censoring distributions, patient accrual patterns, rates of loss to follow-up and noncompliance, and lag in the treatment effect. In addition, the sample size calculation should be specific to the planned analysis approach. This article reviews these factors and the methods that have been proposed to account for them.

OVERVIEW

Sample size calculation is an important aspect of the design of any clinical trial. If the sample is too small there may be insufficient information to detect a true treatment difference, whereas if the sample is too large exposure to an experimental treatment and cost will be unnecessarily high. Accurate determination of sample size requires specification of an appropriate study endpoint, a statistical model and design parameters, the clinically meaningful difference, Type I error, and power. In addition, there are some unique factors affecting sample size in trials involving a time-to-event endpoint, as described below.

Patients typically enroll in a survival trial during a recruitment period. This process is referred to as “staggered entry” to distinguish it from simultaneous enrollment of all patients, and it can have an important impact on sample size requirements. Principal interest during follow-up concerns the time of occurrence of a particular “event” or “endpoint,” which could be death or some other clinical observation. In general, patients are followed from randomization through termination of the trial or an endpoint, whichever occurs first. Those patients who complete the study without experiencing an event

are said to be “right-censored.” The event and censoring distributions also play an important role in sample size determination.

Often some patients do not follow the prescribed protocol for the trial duration. Patients may be withdrawn from follow-up despite the investigators’ efforts; for example, they may refuse to be followed or they may die from an unrelated cause. These “lost to follow-up” patients are not under surveillance for the remainder of the trial. Other patients discontinue the experimental treatment (“drop-outs”) or switch from the control to the experimental treatment (“drop-ins”). We assume that drop-outs and drop-ins (“noncompliant” patients) continue in follow-up. One possible analysis approach is to censor these patients at the time of noncompliance; however, this results in effective derandomization and can bias the treatment comparison. The intention-to-treat analysis avoids this bias by including complete follow-up of all patients, regardless of premature discontinuation of treatment. However, by including complete follow-up in the analysis the treatment effect may be diluted. Careful assessment of the proportion of noncompliant patients is an important component of sample size calculation.

In some clinical trials the treatments reach full effectiveness immediately, whereas in other cases the full treatment effect requires a period of time, known as “lag time.” Several models have been considered, including a linear increase and a threshold model in which the full treatment effect begins after a delay. While the concept of lag usually refers to the onset of the treatment effect (“onset lag”), it may also refer to the loss of treatment effect following discontinuation of study treatment (“decay lag”).

There are three typical types of study termination criteria: 1) “fixed follow-up duration” for each patient; 2) “common closing date,” chosen based on predicted event rates used to calculate power; 3) “fixed number of events,” where the study continues until a prespecified

number of events are observed. For trials with staggered entry, a common closing date or fixed number of events will result in variable durations of follow-up. The fixed endpoint approach has the advantage of guaranteeing the desired power regardless of any discrepancy between the assumed and actual event rates. Given the critical importance of the number of observed events in survival studies, calculation of sample size is actually a matter of finding the right balance between sample size and study duration to produce the required number of events.

In the next sections, we describe three basic approaches for sample size calculation in survival studies, approaches based on the binomial distribution, exponential survival distributions, and semiparametric models such as those that are the basis for the logrank test and Cox proportional hazards regression. Within each section, we describe adjustments to these models proposed to account for the design issues described above. We also discuss sample size calculation for group sequential survival trials and commonly used software packages.

SAMPLE SIZE CALCULATION BASED ON THE BINOMIAL DISTRIBUTION

Although the analysis of survival studies involves methods designed for time-to-event data, the sample size calculation for a study of duration T years is sometimes based on the anticipated T -year event rates in the control group (p_c) and the experimental group (p_e), without consideration of the survival distributions. Assuming equally-sized groups and a one-tailed test of $H_0: p_e = p_c$, the total sample size N can be calculated using the following asymptotic formula:

$$N = \frac{2 \times [Z_\alpha \sqrt{2\bar{p}(1-\bar{p})} + Z_\beta \sqrt{p_e(1-p_e) + p_c(1-p_c)}]^2}{(p_e - p_c)^2} \quad (1)$$

where $\bar{p} = (p_e + p_c)/2$, and Z_α and Z_β are normal deviates associated with α and β .^[1] (In general, throughout this article, formulas involving Z_α refer to one-tailed tests, and formulas involving $Z_{\alpha/2}$ refer to two-tailed tests.) Adjustments for noncompliance and other factors typically involve first estimating effective event rates, and then plugging them into the above formula.

Schork and Remington^[2] estimated effective event rates by taking into account yearly event rates and noncompliance patterns, assuming time invariance and no decay lag. Halperin, Rogot, Gurian, and Ederer^[1] (hereinafter HRGE) estimated the effective event rate in the experimental group by allowing for a linear onset lag and for drop-outs, based on modeling the distribution of time to drop-out. Decay lag and drop-in were not considered. Wu, Fisher, and DeMets^[3] (hereinafter WFD) generalized HRGE by dividing the annual time intervals into subintervals and allowing for time-dependent drop-out and event rates. They also allowed for drop-ins (which could return to the

control treatment at a later date), unequal sample sizes, and linear onset and decay lags.

In 1986, Lakatos^[4] estimated effective event rates by using a discrete nonstationary Markov process that allows for stratification, time-dependent event rates, noncompliance, loss to follow-up, staggered entry, and onset and decay lag. A typical transition matrix for the experimental group is given below:

	L	E	C_1	C_2	C_3	D_0	D_1	D_2
L	1		l	l	l	l	l	l
E		1	p_{e_1}	p_{e_2}	p_{e_3}	p_{e_0}	p_{e_1}	p_{e_2}
C_1						b		
C_2			$1 - \Sigma$				b	
C_3				$1 - \Sigma$	$1 - \Sigma$			b
D_0			c			$1 - \Sigma$	$1 - \Sigma$	
D_1				c				$1 - \Sigma$
D_2					c			

State L denotes lost to follow-up, state E denotes an event, states C_i denote patients in follow-up with event rate p_{e_i} ($i = 1, 2, 3$), and states D_i denote patients in follow-up subsequent to drop-out with event rate p_{e_i} ($i = 0, 1, 2$). The multiple states C_i and D_i are used to model onset and decay lag, respectively. (The number of such states is flexible; this example was chosen for illustration only.) The yearly loss to follow-up rate is l , and the yearly drop-out and drop-in rates are c and b , respectively. Note that $1 - \Sigma$ represents 1 minus the sum of the remainder of the column, and zeroes are represented as blank elements. This model allows for patients to continually switch from compliance to noncompliance. The model accounts for staggered entry by assuming patients are administratively censored according to the study's accrual pattern, and can accommodate a wide range of distributional assumptions, either parametric or nonparametric. Furthermore, event, loss, and noncompliance rates within subintervals can be obtained. Under a common set of assumptions, Lakatos's^[4] method approaches WFD as the number of subintervals increases.

Other sample size calculations based on the binomial distribution include Shuster^[5] (without adjustment for noncompliance) and Palta and McHugh^[6] (with adjustment for noncompliance).

SAMPLE SIZE CALCULATION FOR EXPONENTIAL MODEL

The simplest parametric survival model assumes that event times are exponentially distributed, which is equivalent to the assumption of constant hazard rates (designated by λ_e

and λ_c for the two groups). Sample size calculation can be based on the likelihood ratio test (exponential MLE test) for equality of hazards.

Schoenfeld and Richter^[7] used large-sample normal theory to generate nomograms to convert accrual and follow-up durations into the required sample size under the assumption of uniform entry. These nomograms can also be used (with iteration) to determine the accrual and follow-up durations for a fixed rate of entry. Two separate figures are given for 80% and 90% power; a more complicated nomogram can be used for any desired power and to find power as a function of the treatment difference.

Rubinstein et al.^[8] discussed sample size for the Mantel-Haenszel test. They assumed: (1) Poisson accrual over a period, T_a , and an additional follow-up period, T_f , and (2) loss to follow-up times independently and exponentially distributed with parameters, ϕ_e and ϕ_c . This resulted in the total sample size for a one-sided test of

$$N = \left(\frac{Z_\alpha + Z_\beta}{\ln(\theta)} \right)^2 \left[\frac{1}{E(P_e)} + \frac{1}{E(P_c)} \right] \quad (2)$$

where

$$E(P_j) = \left[\frac{1 - e^{-\lambda_j^* T_f} (1 - e^{-\lambda_j^* T_a})}{\lambda_j^* T_a} \right] \left(\frac{\lambda_j}{\lambda_j^*} \right) \quad (3)$$

θ is the hazard ratio, and $\lambda_j^* = \lambda_j + \phi_j$, $j = e, c$.

Assume that the minimal relevant difference is $|\lambda_e - \lambda_c|$, patients will be recruited at a uniform rate over an accrual period T_a , and T_f is an additional follow-up period after completion of recruitment. Then the total sample size for a one-sided test is given by Lachin^[9] as

$$N = \frac{(Z_\alpha \sqrt{\phi(\bar{\lambda})(Q_e^{-1} + Q_c^{-1})} + Z_\beta \sqrt{\phi(\lambda_e)Q_e^{-1} + \phi(\lambda_c)Q_c^{-1}})^2}{(\lambda_e - \lambda_c)^2} \quad (4)$$

where Q_e and Q_c are the proportions of patients in each group,

$$\bar{\lambda} = Q_e \lambda_e + Q_c \lambda_c \quad (5)$$

and

$$\phi(\lambda) = \lambda^2 \left[1 - \frac{e^{-\lambda T_f} - e^{-\lambda(T_a + T_f)}}{\lambda T_a} \right]^{-1} \quad (6)$$

Assuming the drop-out rate c , Lachin's^[9] formula for the effective value, p_e^* , of the T -year event rate, p_e , was

$$p_e^* = p_e(1 - c) + p_c c \quad (7)$$

so that the effective treatment difference is

$$\delta^* = p_e^* - p_c = (1 - c)(p_e - p_c) \quad (8)$$

Donner^[10] compared Lachin's method^[9] with Schork and Remington's^[12] and HRGE, and concluded that Lachin's^[9] is the most conservative because it characterizes drop-outs completely by the control group event rate.

Lachin and Foulkes^[11] extended Lachin's^[9] method to allow for nonuniform entry, loss to follow-up, noncompliance, and a stratified analysis. Considering nonuniform patient entry, they derived

$$\phi(\lambda) = \lambda^2 \left\{ 1 + \frac{\gamma e^{-\lambda(T_a + T_f)} [1 - e^{(\lambda - \gamma)T_a}]}{(1 - e^{-\gamma T_a})(\lambda - \gamma)} \right\}^{-1} \quad (9)$$

where γ is the parameter in a truncated exponential entry distribution. Sample size is then evaluated by substituting $\phi(\lambda)$ into Eq. 6. When losses to follow-up are considered,

$$\phi(\lambda, \eta) = \lambda^2 \left(\frac{\lambda}{\lambda + \eta} \left\{ 1 - \frac{e^{-(\lambda + \eta)T_f} - e^{-(\lambda + \eta)(T_a + T_f)}}{T_a(\lambda + \eta)} \right\} \right)^{-1} \quad (10)$$

where η is the loss to follow-up rate. The noncompliance adjustment simply and conservatively adjusts for the noncompliance rates:

$$N_w = \frac{N}{(1 - w_e - w_c)^2} \quad (11)$$

where N is the total sample size required with full compliance, and w_e and w_c are the noncompliance rates in the experimental and control groups, respectively.

Hwang and Chen^[12] extended Lachin and Foulkes's method^[11] to consider onset lag. Other sample size calculations based on exponential models include those of Pasternack and Gilbert,^[13] Pasternack,^[14] George and Desu,^[15] Bernstein and Lagakos,^[16] Makuch and Simon,^[17] and Taulbee and Symons,^[18] although none adjust for noncompliance.

SAMPLE SIZE CALCULATION FOR SEMIPARAMETRIC TEST

The logrank test is a commonly used procedure that does not assume a specific survival distribution but does assume proportional hazards, and has excellent power when that assumption is true. Gail^[19] suggested that a sample size method based on the logrank test be used if hazards are proportional, and that the binomial method be used if hazards are nonproportional and duration is comparable to mean survival. Assuming that the number of events at the i th event time is small relative to the number at risk, the true log-hazard ratio θ_R is small, and recruitment to each treatment group proceeds at a similar rate, then the required number of events is given by Collett^[20] as

$$d = \frac{4(Z_{\alpha/2} + Z_\beta)^2}{\theta_R^2} \quad (12)$$

The required total number of patients is found from

$$N = \frac{d}{P(\text{death})} = \frac{d}{1 - \frac{1}{6}\{\bar{S}(T_f) + 4\bar{S}(0.5T_a + T_f) + \bar{S}(T_a + T_f)\}} \quad (13)$$

where

$$\bar{S}(t) = \frac{S_c(t) + S_e(t)}{2} \quad (14)$$

and $S_c(t)$ and $S_e(t)$ are the estimated values of the survivor functions for individuals in the control and experimental groups, respectively, at time t .

Freedman^[21] provided a derivation of Lachin's^[9] method based on the asymptotic conditional expectation and variance of the logrank statistic, and published a set of tables of required sample sizes. The approach allowed arbitrary entry distributions but assumed that the expected event rate was well approximated by the event rate at the average follow-up. To calculate the sample size, instead of using $4/\theta_R^2$ as in formula 12, he suggested using $[(1 + \theta)/(1 - \theta)]^2$, where θ is the true hazard ratio. The total sample size can then be calculated as

$$N = \frac{2d}{p_e + p_c} \quad (15)$$

Freedman suggested accounting for staggered entry by using the values of p_e and p_c at a time equal to the average length of follow-up, and he proposed inflating the sample size by a factor of $1/(1-x)$, where x is the anticipated fraction lost to follow-up.

Similar to his 1986 method, in 1988 Lakatos^[22] derived the required sample size by using a discrete nonstationary Markov process that allows for stratification and time-dependent event rates, noncompliance, loss to follow-up, staggered entry, and onset and decay lag. This method calculates the expected value and variance of the logrank statistic using the ratios of the hazards and proportions at risk within each discrete interval. The required number of events can then be calculated as follows:

$$d = \frac{[Z_{\alpha/2}(\sum w_i^2 \rho_i \eta_i)^{1/2} + Z_\beta(\sum w_i^2 \rho_i \eta_i)^{1/2}]^2}{(\sum w_i \rho_i \gamma_i)^2} \quad (16)$$

where w_i is the i th Tarone–Ware weight (logrank is obtained if $w_i = 1$),^[23] $\rho_i = d_i / \sum d_i$ with d_i the number of events in the i th interval, $\eta_i = \phi_i / (1 + \phi_i)^2$ with ϕ_i the ratio of patients in the two treatment groups at risk in the i th interval, and $\gamma_i = [\phi_i \theta_i / (1 + \phi_i \theta_i)] - [\phi_i / (1 + \phi_i)]$ with θ_i the hazard ratio in the i th interval. The quantities ρ_i , η_i , and γ_i can be determined using the Markov model. The required total sample size is:

$$N = 2d / (p_e + p_c) \quad (17)$$

where p_e and p_c are the cumulative event rates.

Ahnn and Anderson^[24] extended Lakatos's^[22] formula to the case of testing the equality of K survival distributions using the Tarone–Ware class of statistics in the presence of time-dependent event rates and losses, noncompliance, staggered entry, and lag, and to the stratified logrank test.

Shih^[25] published a public domain SAS IML macro version of Lakatos's^[4,22] methods, called SIZE, which can be used to calculate sample size, power, and study duration. SIZE allows for nonproportional hazards, noncompliance, loss to follow-up, lag, and uncertainties in treatment benefit. HRGE, WFD, and Lakatos^[4,22] all considered linear lag functions; in contrast, SIZE uses a more flexible step function. In addition, SIZE considers the impact on power of uncertainties in the parameter values, which are expressed in terms of a discrete prior distribution. The sample size and power calculated using SIZE were found to be close to those using Lakatos's method.

Zucker and Lakatos^[26] described an analysis method accounting for onset lag, and presented two weighted logrank-type statistics designed to have good efficiency for various lags. They also pointed out that the sample size calculation should be specific to the future analysis method, something that can easily be accommodated by simulation-based methods.

Lakatos and Lan^[27] compared the approaches of Rubinstein et al.,^[8] Freedman,^[21] and Lakatos^[22] using simulations. Their conclusions were the following: (1) The formula-based methods of Rubinstein et al.^[8] and Freedman^[21] are attractive; however, because they are restrictive regarding the survival distributions and other time-dependent factors, the calculated sample sizes might not be accurate; (2) Lakatos's method^[22] has several advantages, including greater accuracy, adaptability to a broad range of conditions (noncompliance, staggered entry, loss to follow-up, and treatment lag), and ability to derive sample sizes for all statistics in the Tarone–Ware family and the G^p family of Harrington and Fleming.

Other sample size calculations based on semiparametric tests without adjustment for noncompliance include those of Palta and Amini,^[28] Yateman and Skene,^[29] Shuster,^[30] Schoenfeld,^[31,32] Hsieh and Lavori,^[33] Halpern and Brown,^[34] Cantor,^[35] Henderson et al.,^[36] Gail,^[37] and Lu and Pajak.^[38] Other sample size calculations with adjustment for noncompliance include Goldman and Hillman.^[39]

SAMPLE SIZE CALCULATION IN GROUP SEQUENTIAL TRIALS

Kim and Tsiatis^[40] proposed a group-sequential design utilizing the spending function described by Lan and DeMets.^[41] Although this method guarantees a desired significance level without requiring specification of the number and timing of

interim analyses in advance, design of the trial requires the number and timing of the interim analyses to be temporarily fixed. Kim and Tsiatis assumed exponential survival distributions, but their arguments extend naturally to the proportional hazards model after suitable transformation of the time scale. They did not account for competing risks or losses to follow-up apart from administrative censoring at the end of the study, and only considered one-sided tests and early termination for rejection of the null hypothesis.

Halpern and Brown^[42] extended their 1987 procedure^[34] for the logrank or Gehan test, incorporating group-sequential testing using either the Pocock^[43] or the O'Brien–Fleming^[44] boundary, and accounting for staggered entry. They presented a program implementing this method, called “mhghseq.” Gu and Lai^[45] expanded Halpern and Brown's^[42] approach; their method enables one to account for factors such as noncompliance, loss to follow-up, patient accrual patterns, and nonproportional hazards alternatives.

STATISTICAL SOFTWARE FOR SAMPLE SIZE CALCULATION

Dupont and Plummer^[46] developed a public domain computer program called POWER for studies with dichotomous, continuous, or time-to-event responses. POWER implements the Schoenfeld/Richter^[7] method for exponential survival distributions and uniform recruitment. As described above, Shih's^[25] public domain SAS IML macro implementation of Lakatos's^[4,22] methods, called SIZE, calculates sample size, power, or study duration. Natarajan et al.^[47] presented a computer program for use in the design of multiple-arm and factorial clinical trials.

Iwane et al.^[48] summarized and reviewed six commercial software packages for sample size and power calculations, although none account for noncompliance. They were:

1. Nsurv,^[49] based on Lachin and Foulkes's^[11] method.
2. Clinical Trials Design Program,^[50] based on Rubinstein et al.'s,^[8] Freedman's,^[21] and Zelen and Dannemiller's^[51] methods.
3. EGRET SIZ,^[52] based on Self et al.'s^[53] method.
4. POWER,^[46] based on Schoenfeld and Richter's^[7] method.
5. Power Analysis and Sample Size (PASS,^[54]), based on Lachin and Foulkes's^[11] method.
6. Ex-Sample,^[55] based on Schoenfeld and Richter's^[7] method.

The authors commented that differences in specifying the model make it difficult to compare across programs.

Lane^[56] gave an overview of several software packages for sample size, including nQuery Advisor,^[57] PASS,

Power and Precision, PS, Statistica Power Analysis, and STPlan, etc. Whitehead^[58] reviewed several software packages including EGRET SIZ, Nsurv, nQuery Advisor, PASS, Power and Precision, PS, Sampsize, Statistica Power Analysis, and STPlan (DSTPlan). Whitehead suggested choosing a software package based on data type; formulae; accuracy; quality of the output, tables, and graphs; and review articles.

nQuery Advisor is extensively used in many research settings and for many types of studies. For the logrank test sample size is based on Freedman^[21] and for test based on exponential survival sample size is based on Lakatos and Lan.^[27] Lane^[56] indicated that nQuery Advisor 4.0 has wide coverage, has a spreadsheet-like interface that is easy to use, and easily produces graphs of power vs. sample size, but customization options are limited.

Emerson^[59] reviewed two commercially available programs, EaSt version 2^[60] and PEST3 version 3 (Planning and Evaluation of Sequential Trials), which aid in the design and conduct of group-sequential trials. Emerson indicated that whereas the PEST3 version 3 provides greater flexibility than EaSt version 2, the use of either program can greatly increase the efficiency with which a clinical trialist plans and monitors a clinical trial.

CONCLUSION

This article described basic methodology and issues affecting sample size calculation for survival trials. Three basic approaches have been used, based on binomial proportions, exponential survival distributions, and tests based on semiparametric models. We reviewed the literature addressing the issues of staggered entry, loss to follow-up, onset and decay lag and noncompliance, and discussed related software/programs.

In survival trials, interest is not just in sample size—the design must consider the trade-offs among sample size, study duration, and event rate. Issues such as staggered entry, lag, and noncompliance can have a major impact on power and should be considered in the design. The choice of a model for sample size calculation should be in agreement with the planned analysis method. Methods based on the binomial model should be avoided when event rates are high. Methods proposed by Lachin and Foulkes^[11] are useful, but may not always have sufficient flexibility to account for all desired factors. Lakatos's^[22] Markov model gives great flexibility, but also places great demands on the user with respect to the number of parameters that must be specified. Because some features, such as the noncompliance rate and the lag function, might not be known with certainty, it may be best to explore the impact of these features on sample size in a sensitivity analysis.

REFERENCES

- Halperin, M.; Rogot, E.; Gurian, J.; Ederer, F. Sample sizes for medical trials with special reference to long-term therapy. *J. Chron. Dis.* **1968**, *21*, 13–24.
- Schork, M.A.; Remington, R.D. The determination of sample size in treatment-control comparisons for chronic disease studies in which drop-out or non-adherence is a problem. *J. Chronic. Dis.* **1967**, *20*, 233–239.
- Wu, M.; Fisher, M.; DeMets, D. Sample sizes for long-term medical trial with time-dependent dropout and event rates. *Control. Clin. Trials* **1980**, *1*, 109–121.
- Lakatos, E. Sample size determination in clinical trials with time-dependent rates of losses and noncompliance. *Control. Clin. Trials* **1986**, *7*, 189–199.
- Shuster, J.J. Fixing the number of events in large comparative trials with low event rates: a binomial approach. *Control. Clin. Trials* **1993**, *14*, 198–208.
- Palta, M.; McHugh, R. Planning the size of a cohort study in the presence of both losses to follow-up and non-compliance. *J. Chron. Dis.* **1980**, *33*, 501–512.
- Schoenfeld, D.A.; Richter, J.R. Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics* **1982**, *38*, 163–170.
- Rubinstein, L.V.; Gail, M.H.; Santner, T.J. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *J. Chronic. Dis.* **1981**, *34*, 469–479.
- Lachin, J.M. Introduction to sample size determination and power analysis for clinical trials. *Control. Clin. Trials* **1981**, *2*, 93–113.
- Donner, A. Approaches to sample size estimation in the design of clinical trials—A review. *Stat. Med.* **1984**, *3*, 199–214.
- Lachin, J.M.; Foulkes, M.A. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics* **1986**, *42*, 507–519.
- Hwang, C.C.; Chen, B.L. Evaluations of sample size and power for analyses of survival with allowance for lag in treatment effect. *ASA Proc. Biopharm. Sect.* **1996**, 174–179.
- Pasternack, B.S.; Gilbert, H.S. Planning the duration of long-term survival time studies designed for accrual by cohorts. *J. Chronic. Dis.* **1971**, *24*, 681–700.
- Pasternack, B.S. Sample sizes for clinical trials designed for patient accrual by cohorts. *J. Chronic. Dis.* **1972**, *25*, 673–681.
- George, S.L.; Desu, M.M. Planning the size and duration of a clinical trial studying the time to some critical event. *J. Chronic. Dis.* **1974**, *27*, 15–24.
- Bernstein, D.; Lagakos, S.W. Sample size and power determination for stratified clinical trials. *J. Statist. Comput. Simul.* **1978**, *8*, 65–73.
- Makuch, R.W.; Simon, R.M. Sample size requirements for comparing time-to-failure among k treatment groups. *J. Chronic. Dis.* **1982**, *35*, 861–867.
- Taulbee, J.D.; Symons, M.J. Sample size and duration for cohort studies of survival time with covariables. *Biometrics* **1983**, *39*, 351–360.
- Gail, M.H. Applicability of sample size calculations based on a comparison of proportions for use with the logrank test. *Control. Clin. Trials* **1985**, *6*, 112–119.
- Collett, D. Sample Size Requirements for a Survival Study. *Modelling Survival Data in Medical Research*; Chapman & Hall/CRC: Boca Raton, FL, 1999.
- Freedman, L.S. Tables of the number of patients required in clinical trials using the logrank test. *Stat. Med.* **1982**, *1*, 121–129.
- Lakatos, E. Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics* **1988**, *44*, 229–241.
- Tarone, R.E.; Ware, J.H. On distribution-free tests for equality for survival distributions. *Biometrika* **1977**, *64*, 156–160.
- Ahnn, S.; Anderson, S.J. Sample size determination in complex clinical trials comparing more than two groups for survival endpoints. *Stat. Med.* **1998**, *17*, 2525–2534.
- Shih, J.H. Sample size calculation for complex clinical trials with survival endpoints. *Control. Clin. Trials* **1995**, *16*, 395–407.
- Zucker, D.M.; Lakatos, E. Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika* **1990**, *77* (4), 853–864.
- Lakatos, E.; Lan, K.K.G. A comparison of sample size methods for the logrank statistic. *Stat. Med.* **1992**, *11*, 179–191.
- Palta, M.; Amini, S.B. Consideration of covariates and stratification in sample size determination for survival time studies. *J. Chronic. Dis.* **1985**, *38* (9), 801–809.
- Yateman, N.A.; Skene, A.M. Sample sizes for proportional hazards survival studies with arbitrary patient entry and loss to follow-up distributions. *Stat. Med.* **1992**, *11*, 1103–1113.
- Shuster, J.J. Power and sample size for clinical trials of survival. *ASA Proc. Sect. Stat. Educ.* **1998**, 9–17.
- Schoenfeld, D. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* **1981**, *68*, 316–319.
- Schoenfeld, D.A. Sample-size formula for the proportional-hazards regression model. *Biometrics* **1983**, *39*, 499–503.
- Hsieh, F.Y.; Lavori, P.W. Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Control. Clin. Trials* **2000**, *21*, 552–560.
- Halpern, J.; Brown, Jr., B.W. Designing clinical trials with arbitrary specification of survival functions and for the log rank or generalized Wilcoxon test. *Control. Clin. Trials* **1987**, *8*, 177–189.
- Cantor, A.B. Power estimation for rank tests using censored data: Conditional and unconditional. *Control. Clin. Trials* **1991**, *12*, 462–473.
- Henderson, W.G.; Fisher, S.G.; Weber, L.; Hammermeister, K.E.; Sethi, G. Conditional power for arbitrary survival curves to decide whether to extend a clinical trial. *Control. Clin. Trials* **1991**, *12*, 304–313.
- Gail, M.H. Sample size estimation when time-to-event is the primary endpoint. *Drug Inf. J.* **1994**, *28*, 865–877.
- Lu, J.; Pajak, T.F. Statistical power for a long-term survival trial with a time-dependent treatment effect. *Control. Clin. Trials* **2000**, *21*, 561–573.
- Goldman, A.I.; Hillman, D.W. Exemplary data: Sample size and power in the design of event-time clinical trials. *Control. Clin. Trials* **1992**, *13*, 256–271.

40. Kim, K.; Tsiatis, A.A. Study duration for clinical trials with survival response and early stopping rule. *Biometrics* **1990**, *46*, 81–92.
41. Lan, K.K.G.; DeMets, D.L. Discrete sequential boundaries for clinical trials. *Biometrika* **1983**, *70*, 659–663.
42. Halpern, J.; Brown, Jr., B.W. A computer program for designing clinical trials with arbitrary survival curves and group sequential testing. *Control. Clin. Trials* **1993**, *14*, 109–122.
43. Pocock, S.J. Group sequential methods in the design and analysis of clinical trials. *Biometrika* **1977**, *64*, 191–199.
44. O'Brien, P.C.; Fleming, T.R. A multiple testing procedure for clinical trials. *Biometrics* **1979**, *35*, 549–556.
45. Gu, M.; Lai, T.L. Determination of power and sample size in the design of clinical trials with failure-time endpoints and interim analyses. *Control. Clin. Trials* **1999**, *20*, 423–438.
46. Dupont, W.D.; Plummer, Jr., W.D. Power and sample size calculations—a review and computer program. *Control. Clin. Trials* **1990**, *11*, 116–128.
47. Natarajan, R.; Turnbull, B.W.; Slate, E.H.; Clark, L.C. A computer program for sample size and power calculations in the design of multi-arm and factorial clinical trials with survival time endpoints. *Comput. Methods Programs Biomed.* **1996**, *49*, 137–147.
48. Iwane, M.; Palensky, J.P.; Plante, K. A user's review of commercial sample size software for design of biomedical studies using survival data. *Control. Clin. Trials* **1997**, *18*, 65–83.
49. Frick, H.; Jauch, S.; Rahlfs, V.W. *Nsurv. Version 2.2*; idv-Institute für Datenanalyse und Versuchsplanung: Munich, Germany, 1994.
50. Piantadosi, S. *Clinical Trials Design Program. Version 1*; Biosoft: Ferguson, MO, 1990.
51. Zelen, M.; Dannemiller, M. The robustness of life testing procedures derived from the exponential distribution. *Technometrics*. **1961**, *3*, 29–49.
52. *EGRET, SIZ Version 1.0*; Statistics and Epidemiology Research Corporation: Seattle, WA, 1993.
53. Self, S.; Mauritsen, R.; Ohara, J. Power calculations of likelihood ratio tests in generalized linear models. *Biometrics* **1992**, *48*, 31–39.
54. Hintze, J.L. *Power Analysis and Sample Size. Version 1.0*; Number Cruncher Statistical System: Kaysville, UT, 1993.
55. Brent, E.E.; Mirielli, E.; Thompson, A. *Ex-Sample™. Version 3.0*; Idea Works, Inc: Columbia, MO, 1993.
56. Lane, P. *An Overview of Software for Sample Size*; PSI talk, 2000.
57. Elashoff, J.D. *nQuery Advisor Version 4.0 User's Guide*; Statistical Solutions: Saugus, MA, 2000.
58. Whitehead, J. *Sample Size Determination of Clinical Trials*; Professional Development Courses in Statistics for Clinical Trials: Philadelphia, PA, 2001.
59. Emerson, S.S. Statistical packages for group sequential methods. *Am. Stat.* **1996**, *50* (2), 183–192.
60. Mehta, C. *EaSt—A Software Package for the Design and Interim Monitoring of Group Sequential Clinical Trials*; CyTEL Software Corporation: Cambridge, MA, 1993.

Sample Size Determination

Lilly Q. Yue

Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

David Li

Solvay Pharmaceuticals, Inc., Marietta, Georgia, U.S.A.

Shan Bai

Eli Lilly and Company, Indianapolis, Indiana, U.S.A.

Abstract

Sample size determination can be based on either the desired power of a particular statistical test for a population parameter or the desired precision of a confidence interval estimator of a population parameter when sampling from a large population. This entry reviews methods of determining sample sizes for statistical tests.

INTRODUCTION

Sample size determination refers to the evaluation of the sample size desired or required for a study during the design stage, before data are collected. A fundamental rule of sample size determination is that the method of evaluation should be based on the planned method of analysis. In sample size evaluation, one must strike a balance between enrolling sufficiently many patients to detect a clinically important difference and not wasting important resources.

Sample size determination can be based on either the desired power of a particular statistical test for a population parameter or the desired precision of a confidence interval estimator of a population parameter when sampling from a large population. A hypothesis test tells us whether the observed data are consistent with the null hypothesis; a confidence interval tells us which hypotheses are consistent with the observed data. Most of the literature on sample size determination discusses the power function of a statistical test rather than the precision of an estimator. However, it is well known that there is a simple correspondence between the two approaches.

OVERVIEW

Statistical tests can be conducted in many ways, but the setting most frequently encountered in clinical trials is comparative, in which two or more treatments are compared. In most cases, the null hypothesis, H_0 , is that the experimental treatment has no different effect on the outcome compared

with the control. When testing the experimental treatment, it may be assumed that one has introduced this new treatment because it is thought to be superior to the control, and the trial is being carried out to “prove” this. On the other hand, experimental treatments are also introduced that are equivalent (or noninferior) to the control, i.e., expected to have no better efficacy, or minimally poorer efficacy than control. Such a treatment may be introduced because this new treatment may have other benefits; for example, it is less toxic, less expensive, or easier to administer, takes less time to treat, or has fewer side effects. One key difference between testing for superiority and equivalence is that the former usually needs the specification of the smallest clinically meaningful and achievable treatment difference in the alternative hypothesis, H_a , while the latter requires the specification of the largest acceptable treatment difference in the null hypothesis, H_0 .

When conducting a statistical test, two types of error must be considered: Type I (false positive, i.e., rejecting the null hypothesis given that the null hypothesis is true) and Type II (false negative, i.e., failing to reject the null hypothesis given that the alternative hypothesis is true), with the probabilities, α and β , respectively. The power of a statistical test refers to the probability of rejecting the null hypothesis when the alternative is true, i.e., $1 - \beta$. Between the two types of error, the Type I error has traditionally been thought of as more critical. However, a smaller Type I error rate results in a lower power. On balance, primary importance is still given to the Type I error, and a properly higher sample size is needed to achieve higher power when both possible and necessary. In any statistical test, one assesses the probability of observing values as extreme or more extreme than the data observed under the null hypothesis, commonly termed the p value. Usually “the

The views expressed in this paper are those of the authors and not necessarily of the U.S. Food and Drug Administration.

data” are summarized in the form of an estimator or a sufficient statistic for an interesting parameter, termed a test statistic, and statistical significance is declared if the resulting p value is less than the preselected significance level, α .

In clinical trials, it is often the case that there are many interesting outcome or endpoint variables. In order to strike a balance between the desirability of assessing the effect of treatment on many variables and maintaining adequate power for specific variables, usually one primary variable is selected and used to determine the sample size. Otherwise, a multiple comparison adjustment is often applied to the sample size calculations through the significance levels assigned to the individual primary outcomes. The sample size is then chosen so that there is adequate power for each primary variable. While multiple comparisons are appropriate only if one will claim significance when any of the primary variables is significant, nominal significance level can be used to assure the significance for all of the primary variables.

Another critical factor in sample size calculation is the selection of the assumed parameters in the chosen population. The adequacy of the sample size determined depends on the accuracy of the initial specification of the parameters. For the purpose of planning a study, therefore, it is always advisable to consider a range of population parameters over which the planned analysis is conducted for a specific sample size.

Numerous articles and reference textbooks provide comprehensive reviews of sample size determination from the perspective of either the precision of an estimator or the power of a test. There is no one response to the general question of how many patients must be included in the treatment and the control groups. In general, sample size determination varies over the type of patient population (e.g., a single population vs. two populations), the type of response measurement (e.g., actual measurement vs. proportion successful), the type of statistical analysis (e.g., time-specific actuarial life table probability vs. crude probability), the type of study design (e.g., paired vs. independent or parallel vs. crossover), and other factors. The following is a compendium of exact and asymptotic formulas for estimating the sample size in clinical trials. Before the general formulas are reviewed, additional consideration is given to the factors that should be accounted for in sample size calculation.

SAMPLE SIZE ADJUSTMENT

A Simple Adjustment for the t -Test Statistic

If an estimator or a sufficient statistic of an interesting parameter is normally distributed, at least asymptotically, the sample size formula frequently takes the form

$$N_{\text{tot}} = \frac{c^2(z_{\alpha/2}\sigma_0 + z_\beta\sigma_1)^2}{\delta^2} \quad (1)$$

Here, N_{tot} is the total sample size for a two-sided test, c is a constant, z_α is the standard normal deviate with probability α of being exceeded, σ_0 and σ_1 are the standard deviations under the null and alternative hypotheses, respectively, and δ is the treatment effect. In one-sided test, $z_{\alpha/2}$ is replaced by z_α in Eq. 1.

If the test statistic has Student's t -distribution instead, then the use of Eq. 1 will underestimate the required sample size, although this effect is increasingly negligible for increasing degrees of freedom, df , of the t statistic. An adequate adjustment is obtained by the correction factor $f = (df + 3)/(df + 1)$, where the actual total sample size is fN and N is obtained from Eq. 1 (see Ref. [1]).

A Simple Adjustment for Missing Values

In some cases, part of the information in the data is missing completely at random (purely by chance). Then the sample size required should be adjusted upwards because the missing information would affect the power of a test and the precision of a confidence interval of a parameter. If M is the rate of missing information, then a simple adjustment is (see Ref. [2], p. 0000)

$$N_{\text{adj}} = \frac{N}{1 - M} \quad (2)$$

A Simple Adjustment for Unequal Allocation

It is well known that power is maximized and total sample size minimized for two equal-sized independent treatment groups if the population variances are equal. However, due to ethical considerations, unequal-sized groups may be desirable. Let subscripts “e” and “c” denote the experimental and control group, respectively. Suppose that the unequal sizes are reflected in the sample fractions, Q_e and Q_c , where $n_e = Q_e N$, $n_c = Q_c N$, $Q_e + Q_c = 1$, and $N = n_e + n_c$. Obviously, $Q_e = Q_c = 0.5$ is for the balanced case. An adjustment for the unbalanced case is

$$N_{\text{unbal}} = \frac{(Q_e^{-1} + Q_c^{-1})}{4} N_{\text{bal}} \quad (3)$$

where N_{bal} and N_{unbal} are the total sample size for balanced case and unbalanced case, respectively.^[2] That is, the unbalanced design requires $[(Q_e^{-1} + Q_c^{-1})/4 - 1]100\%$ larger sample size to provide the same level of power as the balanced design.

SAMPLE SIZE BASED ON MEANS

In this section, the outcome variable is continuous or quantitative. It is convenient to assume that the estimator or test statistic of an interesting parameter is normally distributed, at least asymptotically.

A Single Mean

In this case, there is no independent control group included in the clinical trials. The clinical decision is often based on the subjective evaluation of the outcome.

Hypothesis Test Approach

For the test $H_0: \mu = \mu_0$ with some a priori specified mean value μ_0 and variance σ_0^2 against an alternative $H_a: \mu = \mu_1 \neq \mu_0$ with variance σ_1^2 , the test statistic is

$$Z = \frac{\bar{Y} - \mu_0}{\sigma_0 / \sqrt{N}}$$

where $\bar{Y}$ is the mean of a single sample of observations. Then for specified α , β , μ_0 , μ_1 , σ_0 , and σ_1 , the total sample size required to ensure power $1 - \beta$ of detecting the difference $\delta = \mu_1 - \mu_0$ with a test at the significance level α is given by

$$N = \frac{(z_\alpha \sigma_0 + z_\beta \sigma_1)^2}{\delta^2} \quad (4)$$

In a clinical trial, σ_0 and σ_1 can be specified based on prior experiments using the same measurements or estimated from a small preliminary or “pilot” study. In these cases, it is best to use the largest value expected. If $\sigma_0 = \sigma_1 = \sigma$, Eq. 4 reduces to

$$N = \frac{\sigma^2 (z_\alpha + z_\beta)^2}{\delta^2} \quad (5)$$

Confidence Interval Approach

Suppose that one wishes to estimate the mean μ of a large population with variance component σ^2 using an estimator $\bar{Y}$ with variance σ^2/N . Then the sample size required to provide an interval estimator with the precision e at confidence level $1 - \alpha$ for a specified variance component σ^2 is given by

$$N = \left(\frac{z_{\alpha/2} \sigma}{e} \right)^2 \quad (6)$$

Correspondence of the Two Approaches

If $\sigma_0 = \sigma_1 = \sigma$, the N derived using a $1 - \alpha$ confidence interval to ensure a precision e is approximately the same as the N required to detect a difference $\mu_1 - \mu_0 = e$ with power 0.5 using a two-sided test. In this case, $e = z_{\alpha/2} \sigma / \sqrt{N}$. However, when $\sigma_0 \neq \sigma_1$, e is actually $z_{\alpha/2} \sigma_1 / \sqrt{N}$ because the variance under the alternative should be used to construct a confidence interval.^[2]

Two Independent Groups

In this case, two treatments are compared. Assume that the outcome variable has the means μ_1 and μ_2 , with the standard deviations σ_1 and σ_2 , respectively, for the two treatments. One wishes to estimate the difference of two population means to test for either superiority of one treatment to another or equivalence of the two treatments, in terms of the population mean.

Parallel Design

The simplest and probably most frequently employed design for a clinical trial is the parallel groups design. Eligible patients who have given their informed consent are randomly assigned to receive one and only one of two treatments.

Confidence Interval Approach Suppose that one wishes to estimate the difference of means, $\mu_1 - \mu_2$, with the common standard deviation $\sigma_1 = \sigma_2 = \sigma$. Based on normal theory, the sample size required to provide an interval estimate with the precision e at confidence level $1 - \alpha$ for a given variance component σ^2 is given by

$$N = \frac{2(z_{\alpha/2} \sigma)^2}{e^2} \quad (7)$$

This equation is for a two-sided confidence interval and is also applicable to the one-sided case by simply replacing $z_{\alpha/2}$ by z_α .

Hypothesis Test for Superiority Approach Consider the statistical model $Y_{ij} = \mu_i + e_{ij}$, $i = 1, 2$; $j = 1, 2, \dots, n$, where Y_{ij} denotes the measurement of outcome response on the j th subject for the i th treatment, μ_i is the underlying true mean response for the i th treatment, and e_{ij} 's denote the independently identically distributed measurement errors with common variance σ^2 . For the test $H_0: \mu_1 - \mu_2 = 0$ against an alternative $H_a: \mu_1 - \mu_2 \neq 0$ with common variance σ^2 , the test statistic is

$$Z = \frac{\bar{Y}_1 - \bar{Y}_2}{\sigma \sqrt{2/n}}$$

where $\bar{Y}_1$ and $\bar{Y}_2$ are the means of two independent samples and n is the sample size for each treatment group. Then, for specified α , β , $\delta = \mu_1 - \mu_2$, and σ^2 , the required sample size for each treatment group is

$$n = \frac{2\sigma^2 (z_{\alpha/2} + z_\beta)^2}{\delta^2} \quad (8)$$

and then the total sample size is $N = 2n$. For a one-sided test, the required sample size for each treatment group is obtained by substituting $z_{\alpha/2}$ by z_α in Eq. 8.

2 × 2 Crossover Design

The two-period crossover design consists of two treatments, which are conventionally labeled as A and B. Each subject (patient) receives the two treatments, with half the subjects randomly assigned to receive the treatments in the sequence AB and the rest in the sequence BA. Let $\bar{Y}_{ijk}$ be the measurement for the j th subject in the i th sequence of treatment at the k th period. Each of n patients in sequence 1 receives treatment A in period 1 and treatment B in period 2; each of n patients in sequence 2 receives the treatments in reverse order. If carryover effects are not present, an adequate model is

$$Y_{ijk} = \mu + \zeta_i + \xi_{ij} + \pi_k + \phi_l + e_{ijk} \quad (9)$$

$i = 1, 2$; $j = 1, 2, \dots, n$; $k = 1, 2$; $l = i - (-1)^i \delta_{2k}$; and $\delta_{2k} = 1$ if $k = 2$, $\delta_{2k} = 0$ otherwise. Here, μ is the overall mean; ζ_i the effect of the i th sequence; ξ_{ij} the effect of the j th patient within the i th sequence, which is the between-subject error; π_k the effect of the k th period; ϕ_l the effect of the l th treatment; and e_{ijk} the effect of the j th subject within the i th sequence at the k th period, which is the within-subject error. For the fixed source of variation, ζ_i , π_k , and ϕ_l , the standard constraints are assumed, i.e., the main effects sum to zero. For the random effects, ξ_{ij} and e_{ijk} , it is assumed that each set of error terms $\{\xi_{ij}\}$ and $\{e_{ijk}\}$ is independent and identically normally distributed with mean 0 and variance σ_s^2 and σ_e^2 , respectively, and the set of two error terms (ξ_{ij}, e_{ijk}) be independently distributed. The assumptions imply that the measurements have the same variance $\sigma_s^2 + \sigma_e^2$; the measurements on different subjects are independent; the measurements on the same subject within different treatment periods have covariance σ_s^2 .

For the test $H_0 : \phi_1 = \phi_2$ against an alternative $H_a : \phi_1 \neq \phi_2$, the difference of the measurements between periods 1 and 2 on each subject is considered by defining $d_{ij} = Y_{ij1} - Y_{ij2}$. Let $\phi_d = 2(\phi_1 - \phi_2)$ and $\hat{\phi}_d = \bar{d}_1 - \bar{d}_2$, then the estimator $\hat{\phi}_d$ is such that

$$E(\hat{\phi}_d) = \phi_d \text{ and } \text{Var}(\hat{\phi}_d) = \frac{2\sigma_d^2}{n^*} \quad (10)$$

Where $\sigma_d^2 = 2\sigma_e^2$ and n^* is the required number of subjects for each of the two sequences. Under the null hypothesis $\phi_1 = \phi_2$, and using the pooled estimator, $\hat{\sigma}_d^2$, of σ_d^2 , the test statistic

$$T_d = \frac{\hat{\phi}_d}{(\hat{\sigma}_d^2 / n^*)^{1/2}} \quad (11)$$

has Student's t -distribution with $2(n^* - 1)$ df . Note that for the two-period crossover design, the variance of the intra-subject differences (d_{ij}) is $2\sigma_e^2$ and the expected value of the difference between the two mean intrasubject differences $(\bar{d}_1 - \bar{d}_2)$ is $2(\phi_1 - \phi_2)$. Therefore, the required number of subjects for each of the two sequences is

$$n^* = \frac{2(2\sigma_e^2)(z_{\alpha/2} + z_\beta)^2}{[2(\phi_1 - \phi_2)]^2} = \frac{\sigma_e^2(z_{\alpha/2} + z_\beta)^2}{\delta^2} = \frac{n(1 - R)}{2} \quad (12)$$

where

$$n = \frac{2(\sigma_s^2 + \sigma_e^2)(z_{\alpha/2} + z_\beta)^2}{\delta^2} \quad (13)$$

would be the number of subjects required for each treatment group if the study were performed using two parallel groups, $\delta = \phi_1 - \phi_2$, and $R = \sigma_s^2 / (\sigma_s^2 + \sigma_e^2)$, the intraclass correlation coefficient. Even if R is as low as 0.5, the total number of subjects required for a two-period crossover study with a specified power is one quarter of the total number required for a parallel groups study with the same power. Because each subject is measured twice in a two-period crossover study, the total number of measurements on the $2n^*$ subjects in it is

$$2 \times 2n^* = 2n(1 - R) \quad (14)$$

a fraction $1 - R$ of the total of $2n$ measurements in the parallel groups study.^[3]

Parallel Design with Baseline and Repeated Measurements

Before randomization, repeated measurements can be used to identify patients who are eligible for the trial. In the long run, an outcome variable may be measured frequently to assure that the outcome is adequately controlled and accurate measurement is obtained. It also could be measured frequently to study the time course of response. Furthermore, because the variability of the measurement is a crucial factor for sample size, one can use repeated measurements to reduce the variability. For example, one can average all of the postrandomization measurements for an individual. In addition, it may reduce the study cost if enrolling more patients is not easy and more costly than taking repeated measurements on each patient for the desired power.

It is preferred to combine baseline (pretreatment) information with posttreatment assessment in the comparison of two treatments if the correlation between baseline and posttreatment is substantially large (usually greater than 0.5). Otherwise, only the posttreatment measurement is used.

Frison and Pocock^[4] explored the use of simple summary statistics for analyzing repeated measurements with baseline measurements in randomized clinical trials with two treatments and provided the methods for determining sample sizes in repeated measurements as follows.

Suppose a randomized clinical trial has two treatment groups with n_i patients per group, and all patients have p pretreatment visits, $m = -(p - 1), \dots, 0$ and r posttreatment visits, $m = 1, \dots, r$. A quantitative measurement is observed

at every visit for every patient, and the following simple model is considered:

$$\begin{aligned} Y_{ijm} &= \mu_{im} + e_{ijm} \\ i &= 1, 2; j = 1, 2, \dots, n_i; \\ m &= -(p-1), \dots, 0, 1, \dots, r \end{aligned} \quad (15)$$

Here, μ_{im} is the underlying true mean response for treatment i at time m . As a result of randomization, it is reasonable to assume $\mu_{1m}^{\text{pre}} = \mu_{2m}^{\text{pre}}$ for the pretreatment visits $m \leq 0$. e_{ijm} is the individual j th patient “error” or variation around the underlying mean μ_{im} , and these errors will not be independent within patients. For simplicity, compound symmetry structure over time points is assumed with equal variance σ^2 for all time points and both treatments and also equal correlation ρ between all pairs of time points on each patient. The correlation ρ is also expected to be substantial (typically greater than 0.5 in most trials) because it reflects the consistency of patient effects over time. Because in many clinical trials the primary objective is to assess the average response to treatment over time, often (but not necessarily) in anticipation that treatment response is liable to occur quickly and to remain reasonably steady over time, the following three possible methods of analysis are considered:

1. Posttreatment means: a simple analysis using the mean for each patient’s posttreatment measurements as the summary measure. For each individual, the summary statistic is

$$\bar{Y}_{ij.}^{\text{post}} = \frac{1}{r} \sum_{m=1}^r Y_{ijm} \quad (16)$$

Then, the overall posttreatment mean difference is

$$\frac{1}{n_1} \sum_{j=1}^{n_1} \bar{Y}_{1j.}^{\text{post}} - \frac{1}{n_2} \sum_{j=1}^{n_2} \bar{Y}_{2j.}^{\text{post}} = \bar{Y}_{1..}^{\text{post}} - \bar{Y}_{2..}^{\text{post}} \quad (17)$$

which has the expected value

$$\frac{1}{r} \sum_{m=1}^r (\mu_{1m} - \mu_{2m}) = \bar{\mu}_{1.}^{\text{post}} - \bar{\mu}_{2.}^{\text{post}} \quad (18)$$

2. Mean changes: a simple analysis of each patient’s difference between the mean of posttreatment measurements and the mean of baseline measurements, the latter often consisting of just a single baseline value per patient. For each individual, the summary statistic is mean change,

$$\frac{1}{r} \sum_{m=1}^r Y_{ijm} - \frac{1}{p} \sum_{m=-(p-1)}^0 Y_{ijm} = \bar{Y}_{ij.}^{\text{post}} - \bar{Y}_{ij.}^{\text{pre}} \quad (19)$$

Then the overall treatment difference in these mean changes is

$$\begin{aligned} \frac{1}{n_1} \sum_{j=1}^{n_1} (\bar{Y}_{1j.}^{\text{post}} - \bar{Y}_{1j.}^{\text{pre}}) - \frac{1}{n_2} \sum_{j=1}^{n_2} (\bar{Y}_{2j.}^{\text{post}} - \bar{Y}_{2j.}^{\text{pre}}) \\ = (\bar{Y}_{1..}^{\text{post}} - \bar{Y}_{1..}^{\text{pre}}) - (\bar{Y}_{2..}^{\text{post}} - \bar{Y}_{2..}^{\text{pre}}) \end{aligned} \quad (20)$$

which has expected value again equal to $\bar{\mu}_{1.}^{\text{post}} - \bar{\mu}_{2.}^{\text{post}}$, since the pretreatment expected values are the same for both treatments.

3. Analysis of covariance: between-patient variations in baseline measurements are taken into account by using the mean baseline measurement for each patient as a covariate in a linear model for treatment comparison of posttreatment means. The model is as follows:

$$\bar{Y}_{ij.}^{\text{post}} = \bar{\mu}_{i.}^{\text{post}} + \beta (\bar{Y}_{ij.}^{\text{pre}} - \bar{\mu}_{..}^{\text{pre}}) + \varepsilon_{ij} \quad (21)$$

where ε_{ij} ’s are independent random errors with assumed constant variance. With estimate $\hat{\beta}$ obtained by least squares and by defining

$$\bar{Y}_{ij.}^{\text{cov}} = \bar{Y}_{ij.}^{\text{post}} - \hat{\beta} (\bar{Y}_{ij.}^{\text{pre}} - \bar{Y}_{..}^{\text{pre}}) \quad (22)$$

the estimated mean treatment difference is

$$\frac{1}{n_1} \sum_{j=1}^{n_1} \bar{Y}_{1j.}^{\text{cov}} - \frac{1}{n_2} \sum_{j=1}^{n_2} \bar{Y}_{2j.}^{\text{cov}} = \bar{Y}_{1..}^{\text{cov}} - \bar{Y}_{2..}^{\text{cov}} \quad (23)$$

which again has expected value $\bar{\mu}_{1.}^{\text{post}} - \bar{\mu}_{2.}^{\text{post}}$.

It is convenient to assume that sample size is sufficiently large that the normal approximation to the t -distribution can be applied. For two equal sized treatment groups with null hypothesis $H_0: \bar{\mu}_{1.}^{\text{post}} = \bar{\mu}_{2.}^{\text{post}}$ and alternative $H_a: \bar{\mu}_{1.}^{\text{post}} \neq \bar{\mu}_{2.}^{\text{post}}$, the required sample size for each treatment group is approximately

$$n = \frac{2\sigma^2}{\delta^2} \left[\frac{1 + (r-1)\rho}{r} - \frac{p\rho^2}{1 + (p-1)\rho} \right] (z_{\alpha/2} + z_\beta)^2 \quad (24)$$

for the method of covariate analysis,

$$n = \frac{2\sigma^2}{\delta^2} \left[\frac{1 + (r-1)\rho}{r} \right] (z_{\alpha/2} + z_\beta)^2 \quad (25)$$

for the posttreatment means method, and

$$n = \frac{2\sigma^2}{\delta^2} \left[\frac{1 + (r-1)\rho}{r} - \frac{(p+1)\rho - 1}{p} \right] (z_{\alpha/2} + z_\beta)^2 \quad (26)$$

for the approach of mean changes, respectively, for given power $1 - \beta$, significance level α and detectable difference $\delta = \bar{\mu}_{1.}^{\text{post}} - \bar{\mu}_{2.}^{\text{post}}$. For the same power and when $\rho > 0.5$, the method of covariate analysis always requires a smaller number of patients.

Suppose only r posttreatment measurements are taken, a constant mean over time points for each treatment is expected, and the effect of time points and the interaction

of time point with treatment are not significant. Then the statistical model of Eq. 15 can be rewritten as

$$Y_{ijm} = \mu_i + f_j + e_{ijm} \quad i = 1, 2; \\ j = 1, 2, \dots, n; \quad m = 1, 2, \dots, r \quad (27)$$

where f_j denotes the effect of the j th subject with mean 0 and variance σ_s^2 , e_{ijm} 's denote the independently, identically distributed errors of measurement with mean 0 and common variance σ_e^2 , and these are uncorrelated. The assumptions imply compound symmetry structure over time points with equal variance $\sigma^2 = \sigma_s^2 + \sigma_e^2$ for all time points and both treatments and also equal correlation $\rho = \sigma_s^2 / (\sigma_s^2 + \sigma_e^2)$ between all pairs of time points for each subject.

Then, for the test $H_0: \mu_1 = \mu_2$ against alternative $H_a: \mu_1 \neq \mu_2$, Eq. 25 becomes (see Ref. [3])

$$n = \frac{2(\sigma_s^2 + \sigma_e^2 / r)(z_{\alpha/2} + z_\beta)^2}{\delta^2} \quad (28)$$

Appealing to a quasi-score test statistic for a generalized linear model based on the generalized estimating equation (GEE) method, Liu and Liang^[5] presented a method to compute sample size for studies involving correlated observations. The case of repeated measurements discussed earlier is an example of correlated data. Because of the potential association for observations within a subject, the intrasubject correlation should be taken into consideration in both the design and the analysis stages. Common choices for correlation structure over repeated measurements include uncorrelated structure, compound symmetry structure, autoregressive structure, and unstructured correlation. The first three correlation structures may be used for sample size calculations in most practical situations. There are no explicit sample size formulas for a generalized linear model, and numerical methods are required. However, for a typical two-sample problem with repeated measurements and compound symmetry correlation structure, an explicit formula for the required total sample size is available:

$$N = \frac{\sigma^2}{\pi_1 \pi_2 \delta^2} \left[\frac{1 + (r-1)\rho}{r} \right] (z_{\alpha/2} + z_\beta)^2 \quad (29)$$

where π_1 and π_2 are proportions of patients in the two treatment groups, respectively. Note that for the equal-sized case, i.e., $\pi_1 = \pi_2 = 0.5$, Eq. 29 is the same as Eq. 25.

2 × 2 Crossover Design with Repeated Measurements

Yue and Roach^[6] considered the extension of the model Eq. 9 for a 2 × 2 repeated measurements crossover design with mean summary:

$$Y_{ijkm} = \mu + \zeta_i + \xi_{ij} + \pi_k + \phi_l + e_{ijk} + \tau_m + (\zeta\tau)_{im} \\ + \omega_{ijm} + (\pi\tau)_{km} + (\phi\tau)_{lm} + f_{ijkm} \quad (30)$$

$i = 1, 2; j = 1, 2, \dots, n; k = 1, 2; l = i - (-1)^i \delta_{2k}; \delta_{2k} = 1$ if $k = 2, \delta_{2k} = 0$ otherwise; and $m = 1, 2, \dots, r$. In addition to the notation and assumptions in the model Eq. 9, τ_m is the effect of the m th time, $(\zeta\tau)_{im}$ the interaction effect of the m th time with the i th sequence, ω_{ijm} the effect of the j th subject within the i th sequence at the m th time, $(\pi\tau)_{km}$ the interaction effect of the m th time with the k th period, $(\phi\tau)_{lm}$ the interaction effect of the m th time with the l th treatment, and f_{ijkm} the random fluctuation of the j th subject within the i th sequence at the k th period and m th time.

For the fixed sources of variation, $\tau_m, (\zeta\tau)_{im}, (\pi\tau)_{km}$, and $(\phi\tau)_{lm}$, the main effect τ_m sums to zero, and the interaction effects sum to zero over both the r levels of time and the two levels of the main effect. For simplicity, it is assumed that each set of $\{\omega_{ijm}\}$ and $\{f_{ijkm}\}$ is independently and identically normally distributed with mean zero and variances σ_ω^2 and σ_f^2 , respectively, and the set of four error terms $(\xi_{ij}, e_{ijk}, \omega_{ijm}, f_{ijkm})$ be independently distributed. The assumptions imply compound symmetry structure over time points within each treatment period, i.e., equal variances for all time points and both treatments and also equal correlation between all pairs of time points. They also imply that the measurements have the same variance $\sigma_s^2 + \sigma_e^2 + \sigma_\omega^2 + \sigma_f^2$; the measurements on different subjects are independent; the measurements on the same subject within the same treatment period but at different time points have covariance $\sigma_s^2 + \sigma_e^2$; and the measurements on the same subject within different treatment periods have covariance $\sigma_s^2 + \sigma_\omega^2$ if at the same time point and σ_s^2 if at different time points.

Similarly, define

$$\phi_d = 2(\phi_1 - \phi_2) \\ \sigma_d^2 = 2(\sigma_e^2 + \sigma_f^2 / r) \\ \hat{\phi}_d = \bar{d}_{1..} - \bar{d}_{2..} \quad (31)$$

Then, using the similar t -test statistic, the required number of subjects for each of the two sequences for given power $1 - \beta$, significance level α , and detectable difference $\delta = (\phi_1 - \phi_2)$ with respect to Y_{ijkm} is approximately

$$n^* = \frac{(\sigma_e^2 + \sigma_f^2 / r)(z_{\alpha/2} + z_\beta)^2}{\delta^2} \quad (32)$$

If $r = 1$ (that is, there is no repeated measurement within each treatment period), then the model in Eq. 30 reduces to the model in Eq. 9, and n^* in Eq. 32 becomes n^* in Eq. 12.

If the interaction effects of the time with other factors and the main effect of the time are not significant, the model Eq. 30 can be simplified to

$$Y_{ijkm} = \mu + \zeta_i + \xi_{ij} + \pi_k + \phi_l + e_{ijk} + f_{ijkm} \quad (33)$$

where f_{ijkm} can be regarded as subsampling error. However, Eq. 32 is still valid for model Eq. 33.

Noninferiority Trials

In noninferiority (i.e., one-sided equivalence) trials, an objective is to show that an experimental treatment is as effective as a standard treatment (an active control) with respect to appropriate primary endpoint(s), from which it is then inferred that the experimental treatment is effective. In fact, in the comparison of the experiment to the standard treatment, it is really showing inferiority of no more than a prespecified noninferiority margin, delta, with respect to parameter(s) of interest, so that it is really “not-too-much-inferior.” The margin delta is the allowable reduction in therapeutic response and chosen based on the historical evidence of the effectiveness of the standard treatment and other clinical and statistical considerations relevant to the experimental treatment and the current study. The trials may be justified when considering the experimental treatment that is thought to be as effective, but perhaps not more effective than the standard. The experimental treatment may be preferred due to some reason other than effectiveness, in curing illness. However, a crucial issue for the noninferiority trials is agreement upon how much expense of reduced efficacy could be accepted as necessary for other benefits, such as reduced toxicity.^[7–10]

A sample size formula for this situation is

$$N_{\text{tot}} = \frac{c^2(z_\alpha\sigma_0 + z_\beta\sigma_1)^2}{(\delta - \delta_0)^2} \quad (34)$$

Here c is a constant, δ is the noninferiority margin, and δ_0 is the expected or true treatment difference. For noninferiority, one usually sets $\delta_0 = 0$. Eq. 34 is valid for confidence interval approaches.^[7]

For simplicity, it is assumed that higher means are more desirable, and μ_e and μ_s denote experimental and standard treatment means, respectively. For the hypothesis $H_0: \mu_s - \mu_e \geq \delta$ against alternative hypothesis $H_a: \mu_s - \mu_e < \delta$, where $\delta > 0$, the required sample size for each treatment group is

$$n = \frac{2\sigma^2(z_\alpha + z_\beta)^2}{(\delta - \delta_0)^2} \quad (35)$$

for given significance level α , power $1 - \beta$, common variance σ^2 , true mean difference $\delta_0 = \mu_s - \mu_e$, and sufficiently small difference δ that the two treatments are considered equivalent for practical purposes if $\delta_0 < \delta$.

Eq. 35 can be developed in terms of a one-sided $100(1 - \alpha)\%$ confidence interval for the difference $\mu_s - \mu_e$, with a specified probability $1 - \beta$ that the interval will not include a difference δ .

Paired Observations

In the case that the measurements in the two groups are linked together by pairing or repeated measures at

times a and b on the same patient, the t -test is conducted using the mean difference $\bar{d} = \bar{Y}_b - \bar{Y}_a$, with variance $\Sigma^2 = \sigma_d^2/n$, where $\sigma_d^2 = 2\sigma^2(1 - \rho)$ if $\sigma_a^2 = \sigma_b^2 = \sigma^2$, ρ being the correlation between the paired measurements. For the hypothesis $H_0: \mu_a - \mu_b = 0$ against an alternative $H_a: \mu_a - \mu_b = \delta \neq 0$, where δ is a clinically significant pre-specified value, the required number of sample pairs is approximately

$$n = \frac{\sigma_d^2(z_\alpha + z_\beta)^2}{\delta^2} \quad (36)$$

Here, σ_d^2 can be obtained from prior experience or estimated from a previous pilot study. Note that pairing is efficient only if $\rho > 0$, i.e., if there is positive correlation between the paired measurements. If no estimate of ρ is available, it is safe to assume $\rho = 0$ or, nominally, $\rho = 0.10$.^[11]

Note that Eq. 36 is applicable for an alternative $H_a: \mu_a - \mu_b \neq 0$ with $z_{\alpha/2}$ substituted for z_α and a detectable difference δ .

SAMPLE SIZE BASED ON PROPORTIONS

When the outcome measure in a clinical trial is a dichotomous variable, such as success versus failure, the data are usually expressed as a proportion of “success” p . The probability distribution of such a proportion is the binomial distribution $B(\pi, N)$ with parameters π (the true population proportion) and N (sample size). For large N , the binomial distribution asymptotically approaches a normal distribution $N(\mu, \sigma^2)$ with mean $\mu = \pi$ and variance $\sigma^2 = \pi(1 - \pi)/N$. In this case, the basic equations of tests for means may be employed as the approximate equations of tests for proportions.

A Single Proportion

Hypothesis Test Approach

For the one treatment group problem involving a single proportion, the hypothesis $H_0: \pi = \pi_0$ is tested to detect a clinically relevant alternative hypothesis $H_a: \pi = \pi_1$ where either $\pi_1 > \pi_0$ or $\pi_1 < \pi_0$. Based on a proportion p and large sample size N , the test statistic is $Z = (p - \pi)/\sigma_0$, where $Z \sim N(0, 1)$ and $\sigma_0 = \pi_0(1 - \pi_0)/N$, if H_0 is true. For the given significance level α and power $1 - \beta$, the equation for sample size N is (see Ref. [1])

$$N = \left[\frac{Z_\alpha \sqrt{\pi_0(1 - \pi_0)} + Z_\beta \sqrt{\pi_1(1 - \pi_1)}}{\pi_1 - \pi_0} \right]^2 \quad (37)$$

McNemar's χ^2 Test of Equality of Paired Proportions

When two groups of observation are linked together in a way such that matching on the sample individuals at

categories A and B. The basic data can be expressed as where m_{-+} , for instance, is the number of pairs with (−) for observation A and (+) for observation B, m_A and m_B are the total numbers for the observations A and B, respectively. Thus, the proportions (p 's)

		B		
		+	−	
A	+	m_{++}	m_{+-}	m_A
	−	m_{-+}	m_{--}	
		m_B		N

are defined as the ratios of frequencies (m 's) to the total number of pairs, N .

Analogous to the problem of t -tests for paired means, the null hypothesis for testing paired proportions is $H_0: \mu_d = (\pi_A - \pi_B) = 0$. Because the difference between π_A and π_B equals $(\pi_{++} - \pi_{--})$, the question can be expressed in terms of the discordant proportions π_{-+} and π_{+-} . This means that the hypothesis may be expressed as $H_0: \pi_{-+} = \pi_{+-} = \pi$. The test statistic for hypothesis H_0 is

$$Z = \frac{(p_{-+} - p_{+-})}{s}$$

where $s^2 = 2\bar{p}/N$ is the estimate of variance; $\bar{p} = (p_{-+} + p_{+-})/2$ is the estimate of π ; and $Z \sim N(0, 1)$ if H_0 is true. Note that Z^2 is equivalent to McNemar's χ^2 statistic usually used; thus, for given a significance level α and a power $1 - \beta$, the equation for sample size N is expressed as (see Ref. [1])

$$N = \left[\frac{Z_{\alpha/2} \sqrt{2\bar{\pi}} + Z_{\beta} \sqrt{\frac{2\pi_{-+}\pi_{+-}}{\bar{\pi}}}}{\pi_{-+} - \pi_{+-}} \right]^2 \quad (38)$$

where $\bar{\pi} = (\pi_{-+} + \pi_{+-})/2$.

Two Independent Proportions

Asymptotically Normal Method

A common clinical design is to employ two treatment groups in sample sizes n_1 , n_2 , and $n_2 = kn_1$ where k is the ratio of two sample sizes. For the null hypothesis $H_0: \pi_1 - \pi_2 = 0$, the exact sample size n_1 is determined by

$$n_1 = \min \left\{ n_1 : \prod (\pi_1, \pi_2 > 1 - \beta) \right\} \quad (39)$$

where

$$\prod (\pi_1, \pi_2) = \sum_{(x,y) \in C} \binom{n_1}{x} \binom{kn_1}{y} \pi_1^x (1 - \pi_1)^{n_1 - x} \pi_2^y (1 - \pi_2)^{kn_1 - y}$$

with the domain

$$C = \left\{ (x, y) : \frac{\sqrt{n_1 |\pi_1 - \pi_2|}}{\sqrt{\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)/k}} > Z_{\alpha} \right\}$$

and the critical value (see equation below)

With asymptotic normality assumption, the sample sizes n_1 and n_2 are expressed as

$$n_1 = \left[\frac{Z_{\alpha/2} \sqrt{\bar{\pi}(1 - \bar{\pi}) \left(1 + \frac{1}{k}\right)} + Z_{\beta} \sqrt{\pi_1(1 - \pi_1) + \frac{\pi_2(1 - \pi_2)}{k}}}{\pi_1 - \pi_2} \right]^2 \quad (40)$$

$$n_2 = kn_1$$

where $\bar{\pi} = (\pi_1 + k\pi_2)/(1 + k)$. For the equal sample size designs ($k = 1$), Eq. 40 becomes

$$n_1 = n_2 = \left[\frac{Z_{\alpha/2} \sqrt{2\bar{\pi}(1 - \bar{\pi})} + Z_{\beta} \sqrt{\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)}}{\pi_1 - \pi_2} \right]^2 \quad (41)$$

where $\bar{\pi} = (\pi_1 + \pi_2)/2$. Note that Eqs. 40 and 41 are without any continuity correction (see Ref. [11]).

Continuity Correction χ^2 Test

Based on Pearson χ^2 statistic and Yates's continuity correction, a formula for sample size calculation is given as

$$n_1 = \frac{n_1^*}{4} \left[1 + \sqrt{1 + \frac{(k+1)|\pi_1 - \pi_2|}{k(Z_{\alpha/2} + Z_{\beta})^2 \bar{\pi}(1 - \bar{\pi})}} \right]^2 \quad (42)$$

$$n_2 = kn_1$$

where n_1^* is determined by Eq. 40. For the equal sample size designs, Eq. 42 is

$$n_1 = \frac{n_1^*}{4} \left[1 + \sqrt{1 + \frac{2|\pi_1 - \pi_2|}{(Z_{\alpha/2} + Z_{\beta})^2 \bar{\pi}(1 - \bar{\pi})}} \right]^2 \quad (43)$$

where n^* is determined by Eq. 41 (see Ref. [12]).

Continuity-Corrected χ^2 Test (Fisher's Exact)

An improved equation for Fisher's exact test is given as

$$n_1 = \frac{n_1^*}{4} \left[1 + \sqrt{1 + \frac{2(k+1)}{kn_1^* |\pi_1 - \pi_2|}} \right]^2 \quad (44)$$

$$n_2 = kn_1$$

where n_1^* is determined by Eq. 40. For the equal sample size designs, Eq. 44 becomes

$$n = \frac{n^*}{4} \left[1 + \sqrt{1 + \frac{4}{n^* |\pi_1 - \pi_2|}} \right]^2 \quad (45)$$

where n^* is determined by formula Eq. 41 (see Ref. [12]).

Test for Relative Risk

In some cases of clinical trials, an estimate of anticipated event rate among the control group patients, π_c , has been estimated from past experience. One may then want to estimate the anticipated event rate of the experiment group, π_e , by relative risk

$$R = \frac{\pi_e}{\pi_c} \quad (46)$$

$$Z_u^* = \inf \left\{ Z_u : \sup \left\{ \sum_{(x,y) \in C} \binom{n_1}{x} \binom{kn_1}{y} \pi^{x+y} (1-\pi)^{(1+k)n_1 - (x+y)} \right\} \leq \alpha \right\}$$

instead of risk difference. In this case the null hypothesis is $H_0: R = 1$. The sample size of each group in an equal size design is determined by

$$n = \left\lceil \frac{Z_{\alpha/2} \sqrt{2\pi_R(1-\pi_R)} + Z_\beta \sqrt{\pi_c(1+R-\pi_c(1+R^2))}}{\pi_c(1-R)} \right\rceil^2 \quad (47)$$

where $\pi_R = \pi_c(1+R)/2$. Note that Eq. 47 is algebraically equivalent to Eq. 41. For the use of continuity correction, Eq. 47 is adjusted by adding $2/\pi_c |1-R|$ to the calculated value of sample size from Eq. 47 (see Ref. [13]).

Equivalence Approach

A question sometimes addressed in clinical trials is whether a new therapy is as effective as a standard therapy. Although it may be as effective or as nearly effective as the standard, the new therapy may provide some other meaningful advantages, such as being less expensive or less toxic than the standard therapy. Unlike the traditional concepts detecting the significant difference between two treatments, the statistical issue here is to test the equivalence. The therapies are considered equivalent for practical purpose if the difference is less than a given small value δ (assume that a larger proportion is more desirable and $\delta > 0$), i.e., if $\pi_s - \pi_e < \delta$. The null hypothesis and alternative hypothesis are $H_0: \pi_s \geq \pi_e + \delta$, $H_a: \pi_s < \pi_e + \delta$. The statistic for testing the null hypothesis is

$$z = \frac{p_s - p_e - \delta}{\sigma} \quad (48)$$

where

$$\sigma = \sqrt{\frac{p_s(1-p_s)}{n_s} + \frac{p_e(1-p_e)}{n_e}}$$

Therefore, one should set the sample sizes sufficiently large so that the confidence interval will not exceed some specified value δ with a desired power $1 - \beta$. This means that

$$\Pr \left(\frac{(p_s - p_e) - (\pi_s - \pi_e)}{\sigma} > \frac{\delta - (\pi_s - \pi_e)}{\sigma} - Z_\alpha \right) = \beta \quad (49)$$

For an equal-sample-size design, the sample size in each group is determined by (see Ref. [8])

$$n = \left\lceil \frac{\left(\frac{Z_\alpha + Z_\beta}{\pi_s - \pi_e - \delta} \right) \sqrt{\pi_s(1-\pi_s) + \pi_e(1-\pi_e)}}{1} \right\rceil^2 \quad (50)$$

More Than Two Independent Proportions

Test for Ordinal Data

When the patient response in two treatment groups is measured on an ordered categorical scale such as very good, good, moderate, poor, the data can be analyzed using techniques of logistic regression. Under the assumption of proportional odds, a formula for total sample size is

$$N = n_e + n_c = \frac{3(A+1)^2 (Z_{\alpha/2} + Z_\beta)^2}{A\theta_R^2 \left(1 - \sum_{i=1}^k \bar{\pi}_i^3 \right)} \quad (51)$$

where

n_e and n_c are the sample sizes for the experimental group and control group, respectively; A is the allocation ratio, $n_e = An_c$;

$\bar{\pi}_i$ is referred to as the anticipated proportion for category C_i ,

$$\bar{\pi}_i = \frac{(\pi_{ie} + \pi_{ic})}{2} \quad i = 1, 2, \dots, k \quad (52)$$

θ_R denotes the reference improvement, a point estimate of log-odds-ratio of the outcome C_i or better for an experimental subject relative to a control subject:

$$\theta_i = \log \left\{ \frac{Q_{ie}(1-Q_{ic})}{Q_{ic}(1-Q_{ie})} \right\} \quad (53)$$

$$Q_{iG} = \sum_{j=1}^i \pi_{jG} \quad G = e, c; \quad i = 1, 2, \dots, k-1$$

The sample size determination is illustrated as the follows. In most cases of clinical trials, the category probabilities in the control group π_{ic} have been estimated from past studies. Based on the previous experience, the

reference improvement θ_R has been also estimated in any given circumstance. Since the definition

$$\theta_i = \log \left\{ \frac{Q_{ie}(1-Q_{ic})}{Q_{ic}(1-Q_{ie})} \right\}$$

is true for all $i = 1, 2, \dots, k-1$, the cumulative probabilities of improvement Q_{ieR} in the experimental group can be anticipated

$$Q_{ieR} = \frac{Q_{ic}}{Q_{ic} + (1-Q_{ic})e^{-\theta_R}} \quad i = 1, 2, \dots, k-1 \quad (54)$$

That gives the category probabilities in the experimental group π_{ie} and the anticipated proportion $\bar{\pi}_i$, $i = 1, 2, \dots, k$. Then the sample size can be determined from Eq. 51 for the given significance level α , power $1-\beta$, and allocation ratio A (see Ref. [14]).

Test for Linear Trend in $G \times 2$ Tables

Let y_i be k mutually independent binomial variates based on sample sizes of n_i at dose level d_i and Y , U , and N be the relevant summations

$$Y = \sum_i y_i, \quad U = \sum_i y_i d_i, \quad N = \sum_i n_i, \quad i = 0, 1, \dots, k-1$$

Define $p = Y/N$, $q = 1 - p$, and $\bar{d} = \sum_i n_i d_i / N$. Assuming that the probability of response follows a linear trend on the logistic scale, $p_i = \exp(\gamma + \lambda d_i) / (1 + \exp(\gamma + \lambda d_i))$, an approximate test with continuity correction asymptotically approaches the normal distribution; that is,

$$z = \frac{U' - \Delta/2}{\sqrt{\text{var}(U'|Y)_{H_0}}} \sim N(0,1) \quad (55)$$

Where

$$U' = U - E(U|Y)_{H_0} = \sum_i y_i (d_i - \bar{d})$$

$$\text{var}(U'|Y)_{H_0} = pq \sum_i n_i (d_i - \bar{d})^2$$

and $\Delta/2$ is the continuity correction. When the Type I and Type II error probabilities are given by α and β , respectively, the following equation can be obtained:

$$E(U') - \frac{\Delta}{2} = Z_\alpha \sqrt{\text{var}(U')_{H_0}} + Z_\beta \sqrt{\text{var}(U')} \quad (56)$$

where

$$E(U') = \sum_i n_i p_i (d_i - \bar{d})$$

$$\text{var}(U')_{H_0} = pq \sum_i n_i (d_i - \bar{d})^2$$

$$p = \frac{\sum_i n_i p_i}{N}$$

and

$$\text{var}(U') = \sum_i n_i p_i q_i (d_i - \bar{d})^2$$

Let $r_i = n_i/n_0$ be the ratio of sample sizes of the i th group to that of the control group ($i = 0, 1, 2, \dots, k-1$) that may be determined in advance. Thus, sample size determination is focused on how to estimate the sample size of the control group. The sample size n_i of the i th group can be calculated directly via the ratio $r_i = n_i/n_0$ whenever n_0 is obtained. The sample size of the control group n_0 can be solved from the following quadratic equation with respect to $\sqrt{n_0}$:

$$n_0 = \left(\frac{n_0^*}{4} \right) \left[1 + \sqrt{1 + \frac{2\Delta}{n_0^* \sum_i \gamma_i p_i (d_i - \bar{d})}} \right]^2 \quad (57)$$

where n_0^* is the sample size of the control group without continuity correction (i.e., $\Delta = 0$) and is given as

$$n_0^* = \frac{\left[Z_\alpha \sqrt{pq \sum_i \gamma_i (d_i - \bar{d})^2} + Z_\beta \sqrt{\sum_i \gamma_i p_i q_i (d_i - \bar{d})^2} \right]^2}{\left[\sum_i \gamma_i p_i q_i (d_i - \bar{d}) \right]^2} \quad (58)$$

For the equal-sample-size design with $d_i = i$, for $i = 0, 1, 2, \dots, k-1$, the sample size per group is approximately expressed as (see Ref. [15])

$$n = \left(\frac{n^*}{4} \right) \left[1 + \sqrt{1 + \frac{2}{n^* \sum_i [i - 0.5(k-1)] p_i}} \right]^2 \quad (59)$$

where

$$n^* = \frac{\left[Z_\alpha \sqrt{\frac{k(k^2-1)pq}{12}} + Z_\beta \sqrt{\sum_i [i - 0.5(k-1)]^2 p_i q_i} \right]^2}{\left[\sum_i [i - 0.5(k-1)] p_i \right]^2} \quad (60)$$

SAMPLE SIZE IN SURVIVAL ANALYSIS

Survival analysis methods are designed for studies in which patients are followed until any event of interest occurs (e.g., death, heart attack, etc.), the patients are censored, or

the study ends. If the survival proportions of patients are simply compared, or all patients are followed for the same fixed time period, the sample size can be determined by the methods related to comparison of proportions. However, most survival studies are designed for comparing treatment groups on mean survival times rather than survival rates. For this reason, the sample size determinations for survival studies are differently considered from those for means or proportions. Similar to the methods for means and proportions, the sample size calculations for survival studies are derived under certain assumptions: proportional-hazard assumption or distribution assumption.

Log-Rank Test for Equality of Survival Curves

Schoenfeld^[16] and Freedman^[17] derived similar sample size formulas for the total number of events needed to be observed and the total number of patients per group needed to be recruited. Suppose that there are two groups of individuals, the experimental group and the control group. Assuming a proportional-hazards model for the survival times, the hazard event at time t for an individual in the experimental group, $h_e(t)$, can be expressed as

$$h_e(t) = \psi h_c(t) \quad (61)$$

where $h_c(t)$ is the hazard function at t for an individual in the control group and ψ is the unknown hazard ratio. Let θ , $\theta \in \Theta$, be the log-hazard ratio, $\theta = \log \psi$, then a zero value of θ indicates no treatment difference; negative value of θ means that survival is longer on the new treatment, and positive value of θ means that survival is longer on the standard treatment.

The required number of events is calculated in such a case to reject the null hypothesis that the observed value of θ is equal to zero at a significance level of α and with the desired power of $1 - \beta$. Both α and β are given for any design of survival study. The required number of events, d , can be calculated using

$$d = \frac{4(z_{\alpha/2} + z_{\beta})^2}{\theta^2} \quad (62)$$

where $z_{\alpha/2}$ and z_{β} are the upper $\alpha/2$ and the upper β points of the standard normal distribution, and θ is an estimate of Θ that might be chosen on the basis of the increase in median survival time, or other estimating method in the survival study. Freedman^[17] suggested using $\left[\frac{1 + e^{\theta}}{1 - e^{\theta}} \right]^2$ instead of $4/\theta^2$ for more accurate sample size. When θ is small, however, the results of these two methods are very close because

$$\left(\frac{1 + e^{\theta}}{1 - e^{\theta}} \right)^2 \approx \left(\frac{2 + \theta}{\theta + \theta^2/2} \right)^2 = \frac{4}{\theta^2} \quad (63)$$

The calculation of Eq. 62 assumes that an equal number of individuals is to be assigned to each treatment group. In the case of unequal sizes, for example, π_c and $(1 - \pi_c)$ as the proportion of individuals in the control and experimental groups, respectively, the required total number of events becomes

$$d = \frac{(z_{\alpha/2} + z_{\beta})^2}{\pi_c(1 - \pi_c)\theta^2} \quad (64)$$

An imbalance in number of individuals in the two groups will increase the total number of events required.

If the study has to be continued until a given percentage of the patients has had the events of interest, p , the total number of patients can be simply calculated

$$N = \frac{d}{p} \quad (65)$$

Otherwise, the probability of event over the duration of study needs to be taken instead of simple percentage p . Typically, individuals are recruited over an accrual period T_a . After recruitment is finished, there is an additional follow-up period T_f . The probability of an event over the duration of study can be calculated as

$$\begin{aligned} \text{Pr(event)} &= 1 - \frac{1}{6} \\ &\times \left[\bar{S}(T_f) + 4\bar{S}(0.5T_a + T_f) + \bar{S}(T_a + T_f) \right] \end{aligned} \quad (66)$$

where

$$\bar{S} = \frac{S_e(t) + S_c(t)}{2} \quad (67)$$

is the average survival function and $S_e(t)$ and $S_c(t)$ are the Kaplan–Meier estimate values of the survival functions for individuals on the experimental and control groups, respectively, at time t . Notice that there is no distribution assumption in Eqs. 66 and 67. Assuming that survival times are exponentially distributed, the average survival function is given by

$$\bar{S} = \frac{e^{-\lambda_e t} + e^{-\lambda_c t}}{2} \quad (68)$$

where estimates λ_e and λ_c can be estimated by means of the corresponding median survival times t_m (see Ref. [18]):

$$\lambda = \frac{\log 2}{t_m} \quad (69)$$

A variant on Eq. 69 for accrual and follow-up periods is given by Schoenfeld and Richter,^[19] who developed nomograms for calculating sample size. The nomograms are valid when survival is exponential and patients enter the study uniformly.

In the general case, Lakatos^[20] derived a method of sample size calculation by using a discrete nonstationary Markov process that allows any pattern of survival, non-compliance, loss to follow-up, drop-in, and lag in the effectiveness of treatment during the study period.

Test Based on Exponential Survival

Rubinstein et al.^[21] considered sample size estimation by assuming exponentially distributed survival duration. Their assumptions include that: (1) patients are accrued during the interval $[0, T]$ according to Poisson process and total trial length is $T + \tau$; (2) time from entry to event are independently and exponentially distributed in each of experimental and control groups, with parameters λ_e and λ_c , respectively; and (3) the losses to follow-up times are also independently and exponentially distributed in each group with parameters ϕ_e and ϕ_c , respectively. Based on the exponential maximum likelihood test, the sample size formula is given as

$$n = \left(\frac{z_\alpha + z_\beta}{\ln(\theta)} \right)^2 \left[\frac{1}{E(P_e)} + \frac{1}{E(P_c)} \right] \quad (70)$$

where

$$E(P_i) = \left[\frac{1 - e^{-\lambda_i^* \tau} (1 - e^{-\lambda_i^* T})}{\lambda_i^* T} \right] \left(\frac{\lambda_i}{\lambda_i^*} \right) \quad i = e, c \quad (71)$$

and where n is the sample size in each group, θ is the hazard ratio, and $\lambda_i = \lambda_i + \phi_i$, $i = e, c$. Several similar works on sample size and power determination under the exponential distribution are available in the literature.^[22]

SAMPLE SIZE BASED ON TREATMENT VARIANCE

Sometimes, a clinical trial will be proposed to test the hypothesis that an experimental drug reduces variability of the key outcome variable as compared to a control. It is convenient to assume that the key outcome variable is normally distributed, or at least asymptotically. The F -test method, which is commonly used to assess whether two variances are equal, compares the two treatments in terms of the ratio $\lambda = \sigma_e / \sigma_c$, where e and c denote experiment and control, respectively; or equivalently, the percentage difference $\Delta = (\lambda - 1) \times 100\%$. Therefore, the hypothesis becomes $H_0: \lambda = 1$ versus $H_a: \lambda \neq 1$. For a significance level of α , a power of $(1 - \beta)$, and a two tailed test, the sample size per arm, n , can be computed by the following quick and satisfactory approximation formula if a clinically important percentage difference Δ is specified

$$n \equiv 2 + \left((z_{\alpha/2} + z_\beta) / \log_e \lambda \right)^2 \quad (72)$$

where z_γ is the $(1 - \gamma)$ percentile of the standard normal distribution. Since $\log_e \lambda = -\log_e (1/\lambda)$, the formula gives the same estimate n whether λ is expressed as σ_e / σ_c or as σ_c / σ_e (see Ref. [23]).

SAMPLE SIZE BASED ON CORRELATION

In observational studies, correlation can be the subject of analysis. We assume the outcome variable is normally distributed, or at least asymptotically. Usually there are two types of hypotheses for testing correlation: (1) whether a true correlation actually exists, $H_0: \rho = 0$ versus $H_a: \rho = \rho_1 \neq 0$; (2) whether the two correlations are significantly different, $H_0: \mu_0 = (\rho_e - \rho_c) = 0$ versus $H_a: \mu_1 = (\rho_e - \rho_c) \neq 0$, where e and c denote experiment and control, respectively. The simplest approach to such problems is to employ Fisher's arctanh transformation:

$$C(r) = 1/2 \log_e \left(\frac{1+r}{1-r} \right) \quad (73)$$

Given a sample correlation r based on N observations, which is distributed about an actual correlation value ρ , then $C(r)$ is normally distributed with mean $C(\rho)$ and variance $\sigma^2 = 1/(N-3)$. The transformation of r to C (and vice versa) is widely tabulated.

A Single Correlation

To test hypothesis $H_0: \rho = 0$ versus $H_a: \rho = \rho_1 \neq 0$, we use the test statistic $Z = C(r) \sqrt{N-3}$. Under H_0 , Z is $N(0,1)$. The sample size then can be obtained by using the following equation:

$$\sqrt{N-3} C(\rho_1) = z_\alpha + z_\beta \quad (74)$$

Two Independent Correlations

In detecting a difference in correlations obtained from two independent samples, the null hypothesis $H_0: \mu_0 = 0$ is tested by using the statistic: $Z = [C(r_e) - C(r_c)] / \Sigma_0^2$, where $\Sigma_0^2 = N^{-1}(Q_e^{-1} + Q_c^{-1})$, $n_e - 3 = Q_e N$ and $n_c - 3 = Q_c N$. And under H_0 , Z is $N(0,1)$. The correlation r_e and r_c are obtained from two samples of size n_e and n_c such as $r_e = r_{e(uv)}$ and $r_c = r_{c(uv)}$ for variables u and v in groups e and c . The total sample size can be obtained by solving the following equation:

$$\frac{\sqrt{N} |C(\rho_e) - C(\rho_c)|}{\sqrt{Q_e^{-1} + Q_c^{-1}}} = z_\alpha + z_\beta \quad (75)$$

Note that N from the above equation will actually be 6 less than that actually needed because $n_e + n_c - 6 = N$ (see Ref. [1]).

REESTIMATING SAMPLE SIZE WITHOUT UNBLINDING

Sample size estimation is a key step for a successful clinical trial. Estimation of sample size in clinical trials depends on the level of significance, α , the power of the trial, $1-\beta$, the minimum detectable difference, δ , all of which are known, and on the knowledge of the variability of the primary response variable, which is usually uncertain to medical researchers. The values of the variability measures usually are guessed or taken from the results of previous trials. In either case, they are subject to uncertainty and may be materially smaller or larger than the true values. Overstating the true variability causes the trial to be larger than it should be, implying unnecessary expenditure of resources or a potential ethical problem due to exposure of more patients than necessary to an inferior treatment. Understating the true variability inflates the Type II error rate, β , which means decreasing the power of the trial, and therefore increasing the likelihood of an inconclusive trial when the treatment effect is real. The resource cost and unproductive patient exposure resulting from an inconclusive trial are much greater than when the variability is overstated. Even with previous data available, one must exercise caution regarding possible differences between the trials in terms of patient population, disease severity, diagnostic criteria, medical procedures, and other study conditions. It is therefore desirable to reestimate the sample size using interim data of the trial under study. It is also important to keep the treatment code blind in an interim estimation to avoid the introduction of any potential (conscious or unconscious) bias during the conduct and monitoring of a clinical trial, especially for those trials with which there is no independent, external data monitoring committee involved.

Normally Distributed Outcomes

Consider a clinical trial that compares two treatments, the null hypothesis $H_0: \mu_1 = \mu_2$ ordinarily would be tested against the alternative $H_a: \mu_1 \neq \mu_2$ using a Student t -test. Given the significance level, α , the power of the trial, $1-\beta$ and the minimum detectable difference, δ , the total sample size would be determined from

$$N = \frac{4\tilde{\sigma}^2(z_{\alpha/2} + z_\beta)^2}{\delta^2} \quad (76)$$

where $\tilde{\sigma}^2$ is the estimate of σ^2 , the within-group variance, and z_γ is the $(1-\gamma)$ percentile of the standard normal distribution.

Suppose that the sample size will be reconsidered after n ($<N$) observations without knowing the treatment assignments, that means we know the observation come from a treatment group (1 or 2), but we do not know the treatment identification of the treatment group. With the reestimate,

$\hat{\sigma}^2$, of σ^2 the within-group variance, one can determine the actual sample size as

$$N' = N\hat{\sigma}^2/\tilde{\sigma}^2 \quad (77)$$

where $\tilde{\sigma}^2$ is the estimate of σ^2 at the beginning of the study.

The Expectation Maximization (EM) algorithm will be used to reestimate σ^2 without unblinding. Since the treatment identifications are unknown, any of the interim observations x_i , $i = 1, \dots, n$ could be in either treatment group, so that the treatment assignments are “missing at random.” Let τ_i denote the treatment group membership indicator

$$\tau_i = 1 \text{ (0)}$$

if sample number i is in treatment group 1 (group2) $\tau_1, \dots, \tau_n$ are independent random variables with $\Pr(\tau_i=1)=0.5$. Given a τ_i value, x_i ($i=1, \dots, n$) has a normal distribution with density

$$f(x_i|\tau_i, \mu_1, \mu_2, \sigma) \propto \frac{1}{\sigma} \exp \left\{ -\frac{1}{\sigma^2} [\tau_i(x_i - \mu_1)^2 + (1 - \tau_i)(x_i - \mu_2)^2] \right\} \quad (78)$$

The expression for the conditional probability (or expectation) of τ_i given x_i therefore is

$$\Pr(\tau_i = 1|x_i) = 1/\{1 + \exp[(\mu_1 - \mu_2) \times (\mu_1 + \mu_2 - 2x_i)/2\sigma^2]\} \quad (79)$$

The log-likelihood of the interim observations follows from Eq. 78

$$l = n \log \sigma + \left\{ \sum_{i=1}^n [\tau_i(x_i - \mu_1)^2 + (1 - \tau_i)(x_i - \mu_2)^2] \right\} / 2\sigma^2 \quad (80)$$

The “E” step of the EM algorithm for estimating σ consists of substituting “current” estimates of μ_1 , μ_2 , and σ into Eq. 79 to obtain provisional values for the expectations of the τ_i . The “M” step consists of obtaining maximum likelihood estimates of μ_1 , μ_2 , and σ after replacing the τ_i in Eq. 80 with their provisional expectations. The “E” and “M” steps are repeated until the value of σ^2 stabilizes; the resulting value is the estimate, $\hat{\sigma}^2$, of σ^2 required in Eq. 77. This process estimates the variance satisfactorily, but does not provide a reliable estimate of the difference between the treatment group means (see Ref. [24]).

Binomial Distributed Outcomes

Consider a clinical trial that compares the effect of two treatments on a disease based on a binary outcome (y), for which we use the generic term “event” ($y = 1$) versus

“nonevent” ($y=0$). Denote p_1 and p_2 as the true “event” rates for the two treatment groups, p_1^* and p_2^* , as the estimates of p_1 and p_2 , respectively, $\sigma(p_1, p_2) = (p_1 + p_2)/2$ as the average event rate, and $\sigma(p_1, p_2) = (p_1 + p_2)/2$ as the treatment difference. The hypotheses to test are: $H_0: p_1 = p_2$ versus $H_a: p_1 \neq p_2$. For power $(1-\beta)$ and significant level α , the sample size per treatment group is determined by

$$n = \frac{2(z_{\alpha/2} + z_\beta)^2 \sigma(1-\sigma)}{\delta^2} \equiv 2(z_{\alpha/2} + z_\beta)^2 \lambda(\sigma, \delta) \quad (81)$$

where z_γ is the $(1-\gamma)$ percentile of the standard normal distribution, and $\lambda(\sigma, \delta) = \sigma(1-\sigma)\delta^2$ is the noncentrality parameter, $\delta \neq 0$. The estimated sample size is obtained by substituting p_1^* and p_2^* in Eq. 81.

We now describe a method that updates the estimation of p_i using interim data from the trial, without breaking the treatment code of the individual patients. We design the trial with a simple, random stratification scheme. For a multicenter trial, this stratification is within each center. Patients are first randomly assigned to either stratum A or B. Since this stratification is not based on any of the patient’s baseline characteristics, it is a “dummy” stratification. We use it only at the interim stage for reestimation and at the final stage we simply ignore it without affecting the regular inference. In stratum A, patients are randomly allocated to treatment group of $i = 1$ with a probability π and to treatment group of $i = 2$ with a probability $1-\pi$; $0 < \pi < 1$. In stratum B, we do the opposite; patients are randomly allocated to treatment group of $i = 1$ with a probability $1-\pi$ and to treatment group of $i = 2$ with a probability π . Hence, we maintain the overall balance of the treatment allocation. It is clear that we should avoid $\pi = 0.5$ because it defeats the purpose of the stratification.

We conduct an interim analysis after the trial is half-way completed as originally planned, that is, when the outcome data (y) are available from a total of n^* patients. Note that among these n^* patients, presumably we should have about $n^*/2$ from each treatment group and also $n^*/2$ from each stratum due to the randomization procedure applied to treatment and stratum, but these n^* patients’ treatment codes remain masked. We report only the pooled event rates for each stratum, without knowing treatment codes, and we estimate the event rates p_1 and p_2 as follows.

We use the result of stratum A to estimate $\theta_1 = P(y_j = 1 | \text{patient} \in \text{stratum A}) = \pi p_1 + (1-\pi)p_2$, and that of stratum B to estimate $\theta_2 = P(y_j = 1 | \text{patient} \in \text{stratum B}) = (1-\pi)p_1 + \pi p_2$. Let the observed event rates in the two strata be $\hat{\theta}_k$, which are unbiased estimators of θ_k , $k=1,2$. Thus, we have the following pair 2 of equations:

$$\begin{aligned} \pi \hat{p}_1 + (1-\pi) \hat{p}_2 &= \hat{\theta}_1 \\ (1-\pi) \hat{p}_1 + \pi \hat{p}_2 &= \hat{\theta}_2 \end{aligned} \quad (82)$$

Solving Eq. 82 for p_1 and p_2 , we have the unbiased estimators of the true event rates for the treatment group,

being weighted averages of the pooled rates from the two strata. We then update the sample size by substituting p_i for p_i^* , $i = 1, 2$ in Eq. 81. In practice, we generally recommend setting π near the middle of 0 and 0.5, e.g., $\pi = 0.2$, to provide reasonable protection of power, and, at the same time, for easy conduct of a trial (see Ref. [25]).

CONCLUSION

Although this entry covers many formulas and/or procedures for sample size determination in clinical research, there are still many areas such as sample size requirement in genomic studies and clinical studies utilizing adaptive trial design, which are not covered in this entry. More details can be found in Chow et al.^[26]

At the interim examination, the estimated variability could turn out to be much less than was anticipated in designing the trial. When this happens, one can terminate the trial without affecting α materially if the interim sample size is adequate. However, this ordinarily would not be advisable in practice. Confirmatory trials seldom have solely the objective of detecting a difference between treatments with respect to a single measurement. Safety and tolerability areas important as efficacy, which itself often needs to be described with more than one measurement. Reducing the sample size could imperil the sensitivity of the trial for detecting clinically meaningful differences with respect to “secondary,” but still important, measurements, and would reduce the sensitivity of the trial for assessing safety and tolerability (see Ref. [24]).

ACKNOWLEDGEMENTS

Minor updates have been made by the Editor-in-Chief in order to reflect developments since publication of the *Encyclopedia of Biopharmaceutical Statistics, Third Edition*.

REFERENCES

1. Lachin, J.M. Introduction to sample size determination and power analysis for clinical trials. *Control. Clin. Trials* **1981**, 2, 93–113.
2. Lachin, J.M. Armitage, P.; Colton, T. Sample size determination. In *Encyclopedia of Biostatistics*; John Wiley & Sons: New York, 1998, Vol. 5, 3892–3903.
3. Fleiss, J.L. *The Design and Analysis of Clinical Experiments*; John Wiley & Sons: New York, 1986.
4. Frison, L.; Pocock, S.J. Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Stat. Med.* **1992**, 11, 1685–1704.
5. Liu, G.; Liang, K.Y. Sample size calculations for studies with correlated observations. *Biometrics* **1997**, 53, 937–947.

6. Yue, L.Q.; Roach, P. A note on the sample size determination in two-period repeated measurements crossover design with application to clinical trial. *J. Biopharm. Stat.* **1998**, *8* (4), 577–584.
7. Lakatos, E. Armitage, P.; Colton, T. Sample size determination for clinical trials. In *Encyclopedia of Biostatistics*, John Wiley & Sons: New York, Vol. 5, 3903–3910, 1998.
8. Blackwelder, W.C. Proving the null hypothesis in clinical trials. *Control. Clin. Trials* **1982**, *3*, 345–353.
9. Farrington, C.P.; Manning, G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Stat. Med.* **1990**, *9*, 1447–1454.
10. Hwang, I.K.; Morikawa, T. Design issues in non-inferiority//equivalence trials. *Drug Inf. J.* **1999**, *33*, 1205–1218.
11. Fleiss, J.L. *Statistical Methods for Rates and Proportions*; John Wiley & Sons: New York, 1981.
12. Fleiss, J.L.; Tytun, A.; Ury, S.H.K. A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics* **1980**, *36*, 343–346.
13. Schlesselman, J.J. Sample size requirements in cohort and case control studies of disease. *J. Epidemiol.* **1974**, *99*, 381–384.
14. Whitehead, J. Sample size calculation for ordered categorical data. *Stat. Med.* **1993**, *12*, 2257–2271.
15. Nam, J. A simple approximation for calculating sample sizes for detecting linear trend in proportions. *Biometrics* **1987**, *43*, 701–705.
16. Schoenfeld, D.A. Sample-size formula for the proportional-hazards regression model. *Biometrics* **1983**, *39*, 499–503.
17. Freedman, L.S. Tables of the number of patients required in clinical trials using the logrank test. *Stat. Med.* **1982**, *1*, 121–129.
18. Collett, D. *Encyclopedia of Biostatistics*, 5th Ed.; Wiley & Sons, New York, 1998.
19. Schoenfeld, D.A.; Richter, J.R. Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics* **1982**, *38*, 163–170.
20. Lakatos, E. Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics* **1988**, *44*, 229–241.
21. Rubinstein, L.V.; Gail, M.H.; Santner, T.J. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *J. Chron. Dis.* **1981**, *34*, 469–479.
22. Lakatos, E.; Lan, G.A. Comparison of sample size methods for the logrank statistic. *Stat. Med.* **1992**, *11*, 179–191.
23. Jin, J.; Hsuan, F.; Hwang, D.S. Reestimation of Sample Size for Variability Comparison in Clinical Trials. In *Proceedings of the Annual ASA Meeting, San Francisco, CA, August 8–12, 1993*; American Statistical Association, Alexandria, VA, 1993; 328–333.
24. Gould, L.A. Planning and revising the sample size for a trial. *Stat. Med.* **1995**, *14*, 1039–1051.
25. Shih, W.J.; Zhao, P.L. Design for sample size re-estimation with interim data for double-blind clinical trials with binary outcomes. *Stat. Med.* **1997**, *16*, 1913–1923.
26. Chow, S.C.; Shao, J.; Wang, H.; Lokhnygina, Y. *Sample Size Calculation in Clinical Research*, 3rd Ed.; Taylor & Francis, New York, 2017.

Sample Size Determination: Three-arm Ratio of Mean Differences Equivalence Test

Yu-Wei Chang

Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, Connecticut, U.S.A.

Yi Tsong and Xiaoyu Dong

Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Zhigen Zhao

Department of Statistics, Temple University, Philadelphia, Pennsylvania, U.S.A.

Abstract

The equivalence assessment may be conducted through a three-arm clinical trial (namely test, reference and placebo) and it usually consists of three tests. The first two tests are to demonstrate the superiority of the test and the reference treatment to placebo, and they are followed by the equivalence assessment between the test and the reference treatment. With continuous response, equivalence is commonly defined in terms of mean difference, mean ratio or ratio of mean differences, i.e. the mean difference of the test and placebo to mean difference of the reference and placebo. These equivalence tests can be performed with both hypothesis testing approach and confidence interval approach. The advantage of applying the equivalence test by ratio of mean differences is that it can test both superiority of the test treatment over placebo and equivalence between the test and the reference simultaneously through a single hypothesis. In this article, we derive the test statistics and the power function for the ratio of mean differences hypothesis and solve the required sample size for a three-arm clinical trial. Examples of required sample size are given in this article, and are compared with the required sample size by the traditional mean difference equivalence test. After a careful examination, we suggest to increase the power of the ratio of mean differences approach by appropriately adjusting the lower limit of the equivalence interval.

Keywords: Equivalence test; Superiority; Three-arm clinical trial; Power; Bioequivalence.

INTRODUCTION

When a clinical trial is designed to assess the equivalence of the test and reference products based on therapeutic endpoints, it often involves three arms, namely, placebo, test, and reference arms. This type of three-arm trial is also called an active control trial with a placebo arm. The purpose of including the placebo arm is to establish the efficacy of the test treatment, as well as to validate the assay sensitivity through the effect of the reference treatment (Pigeot et al., 2003). When the response is continuous, equivalence is commonly established by showing that either the difference or the ratio of two treatment means is bounded within a prespecified equivalence margin, as discussed in many papers and FDA bioequivalence guidance (Kieser, 1999; Food and Drug Administration [FDA], 2001; Chow and Liu, 2008; Kang and Chow, 2013).

Conventionally, an equivalence test between the test and reference treatments is conducted only after showing

the efficacy of both the test and the reference treatments in the study's population (Koch and Rohmel, 2004). Such a three-arm trial has been discussed in Tsong et al. (2004) with the equivalence defined by the mean ratio. It can be shown that passing the equivalence test alone does not necessarily imply the positive effect of the test treatment or the reference treatment against the placebo as defined in a superiority trial. In addition, the results of a two-arm equivalence assessment by the mean difference or by the mean ratio provide little information about the relative efficacy of the test treatment and the reference treatment. To address these issues, this article describes an equivalence assessment based on the ratio of mean differences, that is, the mean difference of the test and the placebo divided by the mean difference of the reference and the placebo. Furthermore, when assay sensitivity is well established prior to the trial, it implies that the reference treatment is better than the placebo. In this case, one may simultaneously determine the equivalence between the test and the reference and the superiority of the test treatment over the

placebo by postulating a single hypothesis using the ratio of mean differences test.

BACKGROUND

Two-Arm Equivalence Test Based on Mean Difference or Mean Ratio

In a two-arm parallel design, equivalence is commonly established by showing that either the mean difference or the mean ratio of two treatments is bounded within a pre-specified equivalence interval when the response are normally distributed and the margin is fixed (Chow and Liu, 2008; Kang and Chow, 2013). In this article, the test arm and the reference arm are denoted by X_T and X_R respectively. X_T and X_R are assumed to be normally distributed with $X_T \sim N(\mu_T, \sigma_T^2)$ and $X_R \sim N(\mu_R, \sigma_R^2)$, where X_T and X_R are independent. The corresponding hypotheses of mean difference and mean ratio equivalence tests are given by,

$$\begin{aligned} H_0^D : \delta_1^* \geq \mu_T - \mu_R \quad \text{or} \quad \mu_T - \mu_R \geq \delta_2^* \quad \text{vs.} \\ H_a^D : \delta_1^* < \mu_T - \mu_R < \delta_2^* \end{aligned} \quad (1)$$

or

$$H_0^R : \frac{\mu_T}{\mu_R} \leq \delta_1 \quad \text{or} \quad \frac{\mu_T}{\mu_R} \geq \delta_2 \quad \text{vs.} \quad H_a^R : \delta_1 < \frac{\mu_T}{\mu_R} < \delta_2 \quad (2)$$

where (δ_1^*, δ_2^*) and (δ_1, δ_2) are the pre-specified equivalence intervals. These two-sided hypotheses (1) and (2) are equivalent to the following two one-sided hypotheses,

$$H_{01}^D : \mu_T - \mu_R \leq \delta_1^* \quad \text{vs.} \quad H_{a1}^D : \mu_T - \mu_R > \delta_1^* \quad (3)$$

$$H_{02}^D : \mu_T - \mu_R \geq \delta_2^* \quad \text{vs.} \quad H_{a2}^D : \mu_T - \mu_R < \delta_2^* \quad (4)$$

and

$$H_{01}^R : \frac{\mu_T}{\mu_R} \leq \delta_1 \quad \text{vs.} \quad H_{a1}^R : \frac{\mu_T}{\mu_R} > \delta_1 \quad (5)$$

$$H_{02}^R : \frac{\mu_T}{\mu_R} \geq \delta_2 \quad \text{vs.} \quad H_{a2}^R : \frac{\mu_T}{\mu_R} < \delta_2 \quad (6)$$

Schuurmann (1987) has demonstrated that a two-sided 5% level test of hypothesis (1) has the same decision rule as two one-sided 5% level tests of hypotheses (3) and (4) under normality assumption. Meanwhile, two-sided 5% level test of hypothesis (2) has the same decision rule as two one-sided 5% level tests for hypotheses (5) and (6) under normality or log-normality assumption.

To assess the equivalence by mean difference (1), the test statistics can be derived as

$$T_i^D = \frac{\bar{X}_T - \bar{X}_R - \delta_i}{se(\bar{X}_T - \bar{X}_R)} = \frac{\bar{X}_T - \bar{X}_R - \delta_i}{\sqrt{\frac{s_T^2}{n_T} + \frac{s_R^2}{n_R}}} \sim t_v, \quad i = 1 \text{ or } 2, \quad (7)$$

where $\bar{X}_T$ and $\bar{X}_R$ denote the sample mean of the test and the reference arm respectively. Under null hypotheses of $H_{0i}^D : \mu_T - \mu_R - \delta_i = 0$, T_i^D follows a central t -distribution with degrees of freedom, v , where v is approximated by Satterthwaite approach (Satterthwaite, 1946),

$$v = \frac{\left(\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R} \right)^2}{\left(\frac{\sigma_T^4/n_T^2}{n_T - 1} + \frac{\sigma_R^4/n_R^2}{n_R - 1} \right)}.$$

For equal variances with $\sigma_T = \sigma_R = \sigma$, the test statistics are

$$T_i^D = \frac{\bar{X}_T - \bar{X}_R - \delta_i}{se(\bar{X}_T - \bar{X}_R)} = \frac{\bar{X}_T - \bar{X}_R - \delta_i}{\sqrt{S_{pool}^2 \left(\frac{1}{n_T} + \frac{1}{n_R} \right)}} \sim t_v, \quad i = 1 \text{ or } 2 \quad (8)$$

where $S_{pool}^2 = \frac{(n_T - 1)S_T^2 + (n_R - 1)S_R^2}{n_T + n_R - 2} \sim \frac{\sigma^2 \chi_v^2}{v}$ with degrees

of freedom $v = n_T + n_R - 2$. Equivalence in terms of mean difference can be concluded if $T_1^D \geq t_{\alpha, v}$ and $T_2^D \leq -t_{\alpha, v}$ with $t_{\alpha, v}$ to be $(1 - \alpha)100$ th-percentile of the central t -distribution with degrees of freedom v .

Using the same notations, for testing the mean ratio (2), assuming positive treatment effect of the reference treatment (i.e. $\mu_R > 0$), one can multiply μ_R on both sides, then the two hypotheses can be expressed as a linear combination of treatment means, μ_R and μ_T :

$$H_{01}^R : \mu_T - \delta_1 \mu_R \leq 0 \quad \text{vs.} \quad H_{a1}^R : \mu_T - \delta_1 \mu_R > 0$$

$$H_{02}^R : \mu_T - \delta_2 \mu_R \geq 0 \quad \text{vs.} \quad H_{a2}^R : \mu_T - \delta_2 \mu_R < 0$$

The following test statistics can be applied for testing each null hypothesis, (5) and (6).

$$T_i^R = \frac{\bar{X}_T - \delta_i \bar{X}_R}{se(\bar{X}_T - \delta_i \bar{X}_R)} = \frac{\bar{X}_T - \delta_i \bar{X}_R}{\sqrt{\frac{s_T^2}{n_T} + \delta_i^2 \frac{s_R^2}{n_R}}} \sim t_{v_i}, \quad i = 1 \text{ or } 2. \quad (9)$$

Under null hypotheses $H_{0i}^R : \mu_T - \delta_i \mu_R = 0$, T_i^R follow central t -distributions with degrees of freedom

$$v_i = \frac{\left(\frac{\sigma_T^2}{n_T} + \frac{\delta_i^2 \sigma_R^2}{n_R} \right)^2}{\left(\frac{\sigma_T^4/n_T^2}{n_T - 1} + \frac{\delta_i^4 \sigma_R^4/n_R^2}{n_R - 1} \right)}, \quad \text{which is approximated by}$$

Satterthwaite approach (Satterthwaite, 1946). For equal variances, when $\sigma_T = \sigma_R = \sigma$, the test statistics are

$$T_i^R = \frac{\bar{X}_T - \delta_i \bar{X}_R}{se(\bar{X}_T - \delta_i \bar{X}_R)} = \frac{\bar{X}_T - \delta_i \bar{X}_R}{\sqrt{S_{pool}^2 \left(\frac{1}{n_T} + \delta_i^2 \frac{1}{n_R} \right)}} \sim t_v \quad i = 1 \text{ or } 2 \quad (10)$$

where $S_{pool}^2 = \frac{(n_T - 1)S_T^2 + (n_R - 1)S_R^2}{n_T + n_R - 2} \sim \frac{\sigma^2 X_v^2}{v}$ with degrees of freedom $v = n_T + n_R - 2$. Equivalence in terms of mean ratio can be concluded if $T_1^R \geq t_{\alpha, v}$ and $T_2^R \leq -t_{\alpha, v}$.

Under alternative hypothesis and with unequal variances, both T_i^D and T_i^R given in (7) and (9) follow bivariate non-central t-distribution approximately (Owen, 1965). For the mean difference test (1), under the alternative space that $\delta_1^* < \mu_T - \mu_R < \delta_2^*$, the non-centrality parameters for T_i^D are $\theta_i = \frac{\mu_T - \mu_R - \delta_i}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_R^2}{n_R}}}$, $i=1$ or 2 with correlation

coefficient 1 between T_i^D . For the mean ratio test (2), under the alternative space that $\delta_1 < \frac{\mu_T}{\mu_R} < \delta_2$, the non-

centrality parameters for T_i^R are $\theta_i = \frac{\mu_T - \delta_i \mu_R}{\sqrt{\frac{\sigma_T^2}{n_T} + \delta_i^2 \frac{\sigma_R^2}{n_R}}}$, $i=1$ or 2 with correlation coefficient $\frac{\frac{\sigma_T^2}{n_T} + \delta_1 \delta_2 \frac{\sigma_R^2}{n_R}}{\sqrt{\frac{\sigma_T^2}{n_T} + \delta_1^2 \frac{\sigma_R^2}{n_R}} \sqrt{\frac{\sigma_T^2}{n_T} + \delta_2^2 \frac{\sigma_R^2}{n_R}}}$

between T_i^R . The degrees of freedoms for both tests are derived earlier using Satterthwaite approach.

Under the assumption of equal variance and equal sample size per arm, the non-centrality parameters and the correlation coefficient of T_i^D and T_i^R can be simplified as $\theta_i = \frac{\mu_T - \mu_R - \delta_i}{\sqrt{\frac{2\sigma^2}{n}}}$ and $\theta_i = \frac{\mu_T - \delta_i \mu_R}{\sqrt{\frac{\sigma^2}{n}(1 + \delta_i^2)}}$ with correlation

coefficient 1 and $\frac{1 + \delta_1 \delta_2}{\sqrt{1 + \delta_1^2} \sqrt{1 + \delta_2^2}}$ respectively. The power function for both tests is given as

$$P[T_1 \geq t_{\alpha, 2n-2} \text{ and } T_2 \leq -t_{\alpha, 2n-2} | H_a] \\ = Q(\infty, -t_{\alpha, 2n-2}, \theta_1, \theta_2, \rho) - Q(t_{\alpha, 2n-2}, -t_{\alpha, 2n-2}, \theta_1, \theta_2, \rho)$$

$$\text{where } Q(a, b, c, d, e) = \frac{1}{\Gamma(n-1) \cdot 2^{n-2}} \int_0^\infty \Phi_2\left(\frac{ay}{\sqrt{2n-2}} - c, \frac{by}{\sqrt{2n-2}} - d, e\right) y^{2n-3} e^{-\frac{y^2}{2}} dy$$

$$\Phi_2(x, y, e) = \frac{1}{2\pi\sqrt{1-e^2}} \int_{-\infty}^x \int_{-\infty}^y \exp\left(-\frac{u^2 - 2euv + v^2}{2(1-e^2)}\right) dv du$$

Hauschke et al. (1999) derived the above power function for the mean ratio test.

THREE-ARM EQUIVALENCE TEST BASED ON MEAN DIFFERENCE OR MEAN RATIO

Equivalence between the test and reference treatments may be claimed if the null hypotheses of (1) or (2) are rejected

at $\alpha=5\%$ level and knowing that $\mu_R > \mu_p$ and $\mu_T > \mu_p$, where X_p is defined as the response from placebo arm, and $X_p \sim N(\mu_p, \sigma_p^2)$. Assume a higher response score indicates a more favorable outcome in this article. In order to establish assay sensitivity of the reference treatment (i.e. $\mu_R > \mu_p$) and efficacy of the test treatment (i.e. $\mu_T > \mu_p$) before equivalence assessment between the test and reference treatments, one may often have to carry out a parallel clinical trial with three arms including the test, the reference and placebo arms. In such a three-arm trial, conventionally, prior to test for equivalence hypotheses (1) or (2), one needs to test the following two superiority hypotheses at 2.5% level of α ,

$$H_0: \mu_R - \mu_p \leq 0 \quad \text{vs.} \quad H_a: \mu_R - \mu_p > 0 \quad (11)$$

$$H_0: \mu_T - \mu_p \leq 0 \quad \text{vs.} \quad H_a: \mu_T - \mu_p > 0. \quad (12)$$

The first superiority test is to validate assay sensitivity of the reference treatment. The second superiority test is to establish efficacy of the test treatment. After rejecting both superiority tests of (11) and (12), one may move on to the equivalence assessment based on mean difference (1) or mean ratio (2). From the historical data, if the reference treatment is consistently superior to placebo, one may assume that $\mu_R - \mu_p > 0$, and skip the test in (11). Thus the conventional three-arm equivalence assessment can be reduced to the following test (A) or test (B).

$$\text{A. } H_0: \mu_T - \mu_p \leq 0 \quad \text{vs.} \quad H_a: \mu_T - \mu_p > 0 \quad \alpha_1 = 2.5\% \\ H_0^D: \delta_1^* \geq \mu_T - \mu_R \quad \text{or} \quad \mu_T - \mu_R \geq \delta_2^* \quad \text{vs.} \\ H_{a1}^D: \delta_1^* < \mu_T - \mu_R < \delta_2^* \quad \alpha_2 = 5\%$$

$$\text{B. } H_0: \mu_T - \mu_p \leq 0 \quad \text{vs.} \quad H_a: \mu_T - \mu_p > 0 \quad \alpha_1 = 2.5\% \\ H_0^R: \frac{\mu_T}{\mu_R} \leq \delta_1 \quad \text{or} \quad \frac{\mu_T}{\mu_R} \geq \delta_2 \quad \text{vs.} \\ H_a^R: \delta_1 < \frac{\mu_T}{\mu_R} < \delta_2 \quad \alpha_2 = 5\%$$

THREE-ARM EQUIVALENCE TEST BASED ON RATIO OF MEAN DIFFERENCES

When the interest of an equivalence assessment involves whether the efficacy of the test treatment remains within given percentages of the efficacy of the reference treatment, one may consider assessing the equivalence based on ratio of mean differences, i.e. relative efficacy. With this objective, we can test the following hypothesis instead of (1) or (2).

$$H_0: \frac{(\mu_T - \mu_p)}{(\mu_R - \mu_p)} \leq \delta_1 \quad \text{or} \quad \frac{(\mu_T - \mu_p)}{(\mu_R - \mu_p)} \geq \delta_2 \quad \text{versus} \\ H_a: \delta_1 < \frac{(\mu_T - \mu_p)}{(\mu_R - \mu_p)} < \delta_2 \quad (13)$$

To simplify the notation, let M be the sample mean difference between the test treatment and placebo, and Q be the sample mean difference between the reference treatment and placebo. Assuming the responses from the test arm, reference arm and placebo arm are mutually independent, the joint distribution of M and Q is given by

$$\begin{pmatrix} M \\ Q \end{pmatrix} \sim MN_2 \left(\begin{pmatrix} \mu_T - \mu_P \\ \mu_R - \mu_P \end{pmatrix}, \begin{pmatrix} \frac{\sigma_T^2}{n_T} + \frac{\sigma_P^2}{n_P} & \frac{\sigma_P^2}{n_P} \\ \frac{\sigma_P^2}{n_P} & \frac{\sigma_R^2}{n_R} + \frac{\sigma_P^2}{n_P} \end{pmatrix} \right)$$

The correlation coefficient between M and Q is

$$\rho^{MQ} = \frac{\sigma_P^2}{n_P \sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_P^2}{n_P}} \sqrt{\frac{\sigma_R^2}{n_R} + \frac{\sigma_P^2}{n_P}}}.$$

The two one-sided hypotheses of the equivalence test (13) are therefore

$$H_{01} : \frac{\mu_M}{\mu_Q} \leq \delta_1 \quad \text{vs.} \quad H_{a1} : \frac{\mu_M}{\mu_Q} > \delta_1 \quad (14)$$

$$H_{02} : \frac{\mu_M}{\mu_Q} \geq \delta_2 \quad \text{vs.} \quad H_{a2} : \frac{\mu_M}{\mu_Q} < \delta_2 \quad (15)$$

where $\mu_M = \mu_T - \mu_P$, $\mu_Q = \mu_R - \mu_P$. Assuming that $\mu_Q > 0$ and multiplying by μ_Q on both sides, (14) and (15) can be further rewritten as

$$H_{01} : \mu_M - \delta_1 \mu_Q \leq 0 \quad \text{vs.} \quad H_{a1} : \mu_M - \delta_1 \mu_Q > 0 \quad (16)$$

$$H_{02} : \mu_M - \delta_2 \mu_Q \geq 0 \quad \text{vs.} \quad H_{a2} : \mu_M - \delta_2 \mu_Q < 0 \quad (17)$$

The test statistic for each hypothesis is

$$T_i = \frac{M - \delta_i Q}{Se(M - \delta_i Q)} = \frac{\bar{X}_T - \delta_i \bar{X}_R - (1 - \delta_i) \bar{X}_P}{\sqrt{\frac{S_T^2}{n_T} + \delta_i^2 \frac{S_R^2}{n_R} + (1 - \delta_i)^2 \frac{S_P^2}{n_P}}}, \quad i = 1 \text{ or } 2.$$

Under the null hypotheses, when $H_{0i} : \mu_M - \delta_i \mu_Q = 0$, T_1 and T_2 follow bivariate t-distribution with the degrees of freedom ν_i approximated by

$$\nu_i = \frac{\left(\frac{\sigma_T^2}{n_T} + \frac{\delta_i^2 \sigma_R^2}{n_R} + \frac{(1 - \delta_i)^2 \sigma_P^2}{n_P} \right)^2}{\left(\frac{\sigma_T^4 / n_T^2}{n_T - 1} + \frac{\delta_i^4 \sigma_R^4 / n_R^2}{n_R - 1} + \frac{(1 - \delta_i)^4 \sigma_P^4 / n_P^2}{n_P - 1} \right)}, \quad i = 1 \text{ or } 2 \quad (\text{Satterthwaite, 1946}).$$

To claim the equivalence of the test and the reference treatments, one needs to reject both null hypotheses, H_{01} and H_{02} , with $T_1 \geq t_{\alpha, \nu_1}$ and $T_2 \leq -t_{\alpha, \nu_2}$.

Under the alternative hypothesis, T_1 and T_2 approximately follow bivariate non-central t-distribution with the non-centrality parameters $\theta_i = \frac{\mu_M - \delta_i \mu_Q}{Sd(M - \delta_i Q)} = \frac{(\delta^* - \delta_i) \mu_Q}{Sd(M - \delta_i Q)}$

where $\delta^* = \frac{\mu_M}{\mu_Q}$ and $\delta_1 < \delta^* < \delta_2$. The covariance of numerators, $M - \delta_1 Q$ and $M - \delta_2 Q$ is

$$Cov(M - \delta_1 Q, M - \delta_2 Q) = \frac{\sigma_T^2}{n_T} + \delta_1 \delta_2 \frac{\sigma_R^2}{n_R} + (1 + \delta_1 \delta_2 - \delta_1 - \delta_2) \frac{\sigma_P^2}{n_P}$$

and the correlation coefficient is

$$\rho = \frac{\frac{\sigma_T^2}{n_T} + \delta_1 \delta_2 \frac{\sigma_R^2}{n_R} + \frac{\sigma_P^2}{n_P} (1 + \delta_1 \delta_2 - \delta_1 - \delta_2)}{\sqrt{\frac{\sigma_T^2}{n_T} + \delta_1^2 \frac{\sigma_R^2}{n_R} + \frac{\sigma_P^2}{n_P} (1 - \delta_1)^2} \sqrt{\frac{\sigma_T^2}{n_T} + \delta_2^2 \frac{\sigma_R^2}{n_R} + \frac{\sigma_P^2}{n_P} (1 - \delta_2)^2}}. \quad (18)$$

With equal sample variance and equal sample size, the test statistics are given by

$$T_i = \frac{M - \delta_i Q}{Se(M - \delta_i Q)} = \frac{\bar{X}_T - \delta_i \bar{X}_R - (1 - \delta_i) \bar{X}_P}{\sqrt{\frac{S_{pool}^2}{n} [1 + \delta_i^2 + (1 - \delta_i)^2]}}, \quad i = 1 \text{ or } 2 \quad (19)$$

where $S_{pool}^2 = \frac{S_T^2 + S_R^2 + S_P^2}{3} \sim \frac{\sigma^2 \chi_v^2}{v}$ and degrees of freedom $\nu = 3n - 3$. Under the null hypothesis, when $H_{0i} : \mu_M - \delta_i \mu_Q = 0$, T_1 and T_2 follow bivariate t-distribution with degrees of freedom $\nu = 3n - 3$. The equivalence of the test and reference treatments can be claimed by rejecting both H_{01} and H_{02} by showing that $T_1 \geq t_{\alpha, 3n-3}$ and $T_2 \leq -t_{\alpha, 3n-3}$. Under the alternative hypothesis, T_1 and T_2 follow bivariate non-central t distribution and the non-centrality parameters can be simplified as

$$\theta_i = \left(\frac{\mu_M - \mu_P}{\sigma} \right) \left(\frac{(\delta^* - \delta_i) \sqrt{n}}{\sqrt{1 + \delta_i^2 + (1 - \delta_i)^2}} \right), \quad (20)$$

$$\text{where } \delta^* = \frac{\mu_M}{\mu_Q} \text{ and } \rho = \frac{(2 + 2\delta_1 \delta_2 - \delta_1 - \delta_2)}{\sqrt{[1 + \delta_1^2 + (1 - \delta_1)^2][1 + \delta_2^2 + (1 - \delta_2)^2]}}.$$

The power function of this ratio of mean differences test under the assumption of equal variances and equal sample sizes is given below,

$$P(T_1 \geq t_{\alpha, 3n-3} \text{ and } T_2 \leq -t_{\alpha, 3n-3} | \delta^*, \delta_1, \delta_2, \sigma, \mu_R - \mu_P) = Q(\infty, -t_{\alpha, 3n-3}, \theta_1, \theta_2, \rho) - Q(t_{\alpha, 3n-3}, -t_{\alpha, 3n-3}, \theta_1, \theta_2, \rho)$$

where

$$Q(a, b, c, d, e) = \frac{1}{\Gamma(\frac{3n-3}{2}) \cdot 2^{\frac{3n-5}{2}}} \int_0^\infty \Phi_2 \left(\frac{ay}{\sqrt{3n-3}} - c, \frac{by}{\sqrt{3n-3}} - d, e \right) y^{3n-4} e^{-\frac{y^2}{2}} dy$$

$$\Phi_2(x, y, e) = \frac{1}{2\pi\sqrt{1-e^2}} \int_{-\infty}^x \int_{-\infty}^y \exp\left(-\frac{u^2 - 2euv + v^2}{2(1-e^2)}\right) dv du$$

The required sample size for a given power can be solved by assigning the values of δ^* , δ_1 , δ_2 , σ , $\mu_R - \mu_P$ and α .

Alternatively, the confidence interval approach can be applied and the decision rule becomes $\delta_1^{MR} \geq \delta_1$ and $\delta_2^{MR} \leq \delta_2$

$$\text{for } \delta_i^{MR} = \frac{(MQ - A) \pm \sqrt{2AM^2 + 2AQ^2 - 2AMQ - 3A}}{Q^2 - 2A}, i = 1$$

or 2

$$\text{where } A = t_{\alpha, v}^2 \left(\frac{S_{Pool}^2}{n} \right), v = 3n - 3, M = \bar{X}_T - \bar{X}_P, Q = \bar{X}_R - \bar{X}_P$$

$$\text{if and only if } (\bar{X}_T - \bar{X}_R)^2 + (\bar{X}_T - \bar{X}_P)^2 + (\bar{X}_R - \bar{X}_P)^2 - 3A \geq 0.$$

NUMERICAL EXAMPLE

In this section, the required sample size of the ratio of mean differences test (13) at a 5% significance level under the condition of equal variance and equal sample size per arm is given. Based on the non-centrality parameters, the power function depends on the values of σ and $\mu_R - \mu_P$, and from the historical data, the confidence interval of $\mu_R - \mu_P$ can be derived. In this example, we assume that the 95% confidence interval for $\mu_R - \mu_P$ is (7, 13) with sample mean difference equal to 10. The required sample size is generated by solving the sample size per arm in the power function by assigning $\mu_R - \mu_P$ equal to 7, 10 or 13 and σ equal to 3, 4, 5 or 6. The choice of values for the mean difference and the standard deviation is made by assuming that the proportion of $\frac{\sigma}{\mu_R - \mu_P}$ lies between 0.2 and 0.9. Also, for

illustration purpose, the equivalence margin (δ_1, δ_2) in the following example is given as (0.8, 1.25), which is a common equivalence margin for mean ratio bioequivalence test based on bioavailability data. Biological products are subject to larger variability, and thus different margins may be needed. Margin selection should be carefully discussed within the trial team and also agreed by medical investigators. Table 1 shows the required sample size per arm (test, reference and placebo) to attain a power of 0.8 while $\delta^* = \frac{(\mu_T - \mu_P)}{(\mu_R - \mu_P)} = 0.95, 1$, or $1/0.95$.

From the above table, one may notice that the required sample size is not symmetric about $\delta^* = 1$. When the test treatment mean equals the reference treatment mean, that is $\delta^* = 1$, the required sample sizes are smaller than the required sample size under the condition of $\delta^* = 1/0.95$ and 0.95. However, when the test treatment mean is larger than the reference treatment mean, that is when $\delta^* > 1$, the required sample size is much smaller than the condition in which the reference treatment mean is larger than the test treatment mean, that is when $\delta^* < 1$.

In practice, the assumption of equal sample size per arm may be hard to achieve, as in most three-arm trials, fewer

Table 1 Sample size per arm for the ratio of mean differences test to attain 80% power at a given mean differences ratio δ^* with equal variance and equal sample size and $(\delta_1, \delta_2) = (0.80, 1.25)$ at $\alpha = 0.05$

$\alpha^* = 0.05$		$\delta^* = 0.95$	$\delta^* = 1.00$	$\delta^* = 1 / 0.95$
σ	$\mu_R - \mu_P$			
3	7	87	67	79
	10	43	33	39
	13	26	20	23
4	7	155	118	140
	10	76	58	69
	13	45	35	41
5	7	242	184	218
	10	119	90	107
	13	70	54	64
6	7	348	265	314
	10	171	130	154
	13	101	77	91

sample size may be assigned to placebo arm for ethical reason. In addition, the variances may also be different within each arm. In the example, for simplicity purpose, only sample size requirements under the condition of equal sample size and equal variance are given. However, as one may expect that it is the most optimistic case which would give the highest power. Under the condition of unequal sample size between arms and higher variance in any arm would result in a higher total sample size.

SAMPLE SIZE COMPARISON

As discussed in the previous section, assuming that $\mu_R - \mu_P > 0$, equivalence between the test and reference treatments may be claimed if both null hypotheses in (A) are rejected. Replace the mean difference equivalence test in (A) by the ratio of mean differences test, and it becomes

$$\begin{aligned} C. \quad H_{01} : \mu_T - \mu_P \leq 0 \quad \text{vs.} \quad H_{a1} : \mu_T - \mu_P > 0 \quad \alpha_1 = 2.5\% \\ H_{02} : \frac{(\mu_T - \mu_P)}{(\mu_R - \mu_P)} \leq \delta_1 \quad \text{or} \quad \frac{(\mu_T - \mu_P)}{(\mu_R - \mu_P)} \geq \delta_2 \quad \text{vs.} \\ H_{a2} : \delta_1 < \frac{(\mu_T - \mu_P)}{(\mu_R - \mu_P)} < \delta_2 \quad \alpha_2 = 5\% \end{aligned}$$

To compare the sample size of test (A) and test (C), one may give the specific equivalence limit of (δ_1, δ_2) for (C) and the corresponding values of (δ_1^*, δ_2^*) for (A) can be derived. The relationship between the margins δ_i and δ_i^* is given by $\delta_i^* = (\delta_i - 1) \cdot (\mu_R - \mu_P)$. Thus, (A) tests whether the difference of the test treatment mean and the reference treatment mean lies within the interval of $(\delta_1 - 1 \cdot 100\%)$ of $\mu_R - \mu_P$ and $(\delta_2 - 1 \cdot 100\%)$ of $\mu_R - \mu_P$. Table 2 gives the required sample sizes for both test (A) and test (C) under

Table 2 Sample size per arm for the mean difference equivalence test (A) and the ratio of mean differences equivalence test (C) to attain 80% power at a given mean differences ratio δ^* with equal variance and equal sample size where $(\delta_1, \delta_2) = (0.8, 1.25)$ and $(\delta_1^*, \delta_2^*) = (-0.2 \cdot (\mu_R - \mu_P), 0.25 \cdot (\mu_R - \mu_P))$.

$\alpha_1 = 0.025$ $\alpha_2 = 0.05$		$\delta^* = 0.95$		$\delta^* = 1.00$		$\delta^* = 1/0.95$	
σ	$\mu_R - \mu_P$	(A)	(C)	(A)	(C)	(A)	(C)
3	7	102	88	66	67	67	79
	10	51	44	33	33	33	39
	13	30	26	20	20	20	24
4	7	181	156	117	119	118	140
	10	89	77	58	59	58	69
	13	53	46	35	35	35	41
5	7	282	243	182	184	184	219
	10	138	119	90	91	91	108
	13	82	71	53	54	54	64
6	7	404	348	262	265	265	315
	10	199	171	129	130	130	154
	13	118	102	77	78	77	92

the condition of $(\delta_1, \delta_2) = (0.80, 1.25)$ and Table 3 gives the required sample sizes under the condition of $(\delta_1, \delta_2) = (0.5, 2.0)$.

As the results show, the ratio of mean differences test (C) requires a smaller sample size than the mean difference test (A) when $\delta^* < 1$. When $\delta^* = 1$, the sample sizes for both methods are almost identical. However, when $\delta^* > 1$, the sample size for test (C) is larger than test (A). The relationship of power for both test (A) and test (C) is illustrated in Figure 1.

Table 3 Sample size per arm for the mean difference equivalence test (A) and the ratio of mean differences equivalence test (C) to attain 80% power at a given mean differences ratio δ^* with equal variance and equal sample size where $(\delta_1, \delta_2) = (0.5, 2.0)$ and $(\delta_1^*, \delta_2^*) = (-0.5 \cdot (\mu_R - \mu_P), (\mu_R - \mu_P))$.

$\alpha_1 = 0.025$ $\alpha_2 = 0.05$		$\delta^* = 0.75$		$\delta^* = 0.85$		$\delta^* = 0.95$		$\delta^* = 1.00$		$\delta^* = 1/0.95$		$\delta^* = 1/0.85$		$\delta^* = 1/0.75$	
σ	$\mu_R - \mu_P$	(A)	(C)	(A)	(C)	(A)	(C)	(A)	(C)	(A)	(C)	(A)	(C)	(A)	(C)
3	7	37	28	20	15	12	11	10	10	9	10	7	11	7	16
	10	19	14	10	8	7	6	6	6	5	6	4	6	4	8
	13	12	9	6	5	4	4	4	4	3	4	3	4	3	5
4	7	66	49	34	26	21	19	17	18	14	17	11	19	11	28
	10	33	24	17	13	11	10	9	9	8	9	6	10	6	14
	13	20	15	11	8	7	6	6	6	5	6	4	6	4	9
5	7	102	77	52	40	32	29	26	27	22	27	17	26	17	43
	10	50	38	26	24	16	15	13	14	11	13	9	15	9	22
	13	30	23	16	12	10	9	8	9	7	8	6	9	6	13
6	7	146	110	75	58	46	42	37	39	31	38	24	42	24	62
	10	72	54	37	29	23	21	19	19	16	19	12	21	12	31
	13	41	32	22	17	14	13	12	12	10	12	8	13	8	19

This figure shows that the power of test (C) is higher than test (A) when $\delta^* < 1$, but test (A) is more powerful than test (C) when $\delta^* > 1$ for the given sample size 20, $\mu_R - \mu_P = 13$ and $\sigma = 3$. This result can be generalized for different parameter values.

Also, one may notice that in Table 2 the required sample size of test (C) is almost identical to the required sample size given in Table 1, which means that adding one more superiority test of the test treatment over placebo to (13) does not have a great impact on the power of the ratio of mean differences equivalence test. This may be because

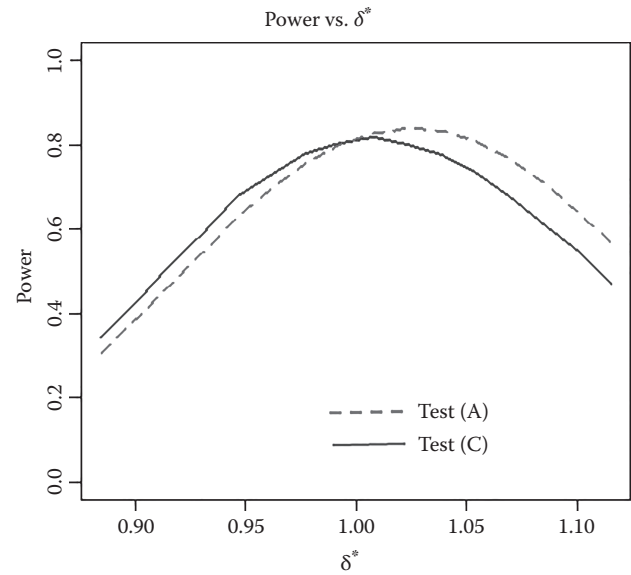


Figure 1 Power versus mean differences ratio δ^* ranging from 0.85 to 1.15 under the condition of $\mu_R - \mu_P = 13$, $\sigma = 3$ with sample size per arm of 20

by rejecting the null hypothesis of (13), it is concluded that $\delta_1 < \frac{(\mu_T - \mu_P)}{(\mu_R - \mu_P)} < \delta_2$, and this result also implies $\mu_R - \mu_P > 0$ because of the assumption that $\mu_R - \mu_P > 0$ while using a positive δ_1 . This implication is made at a 5% level of α , which is not as stringent as the superiority test at 2.5%. The power of ratio of mean differences test (13) also depends on the value of δ_1 . Thus, if the ratio of mean differences test is conducted with a smaller δ_1 , then the power of the test will increase and the 2.5% superiority test may have some impact on the ratio of mean differences test. One should always choose an appropriate δ_1 with care; otherwise, the equivalence test may not be validated because the equivalence interval may be too wide to conclude equivalence.

We further examined the sample size power relationship for $(\delta_1, \delta_2) = (0.5, 2.0)$. The relationship between the sample size requirement and δ^* remains the same as for $(\delta_1, \delta_2) = (0.8, 1.25)$ as shown in Table 3.

In order to test both superiority of the test treatment over the placebo and equivalence by using only one hypothesis test, one may remove the first superiority test from (C) and change the α level of $H_0 : \frac{(\mu_T - \mu_P)}{(\mu_R - \mu_P)} \leq \delta_1$ vs.

$H_0 : \frac{(\mu_T - \mu_P)}{(\mu_R - \mu_P)} > \delta_1$ from 5% to 2.5%. By doing so, the

lower equivalence margin could also serve as a more stringent superiority test margin, i.e. test (D) below. Not only does the following test (D) serve as the equivalence test, but it also serves as the superiority test of the test treatment over the placebo.

$$\begin{aligned} \text{D. } H_{01} : \frac{(\mu_T - \mu_P)}{(\mu_R - \mu_P)} \leq \delta_1 \quad \text{vs.} \quad H_{a1} : \frac{(\mu_T - \mu_P)}{(\mu_R - \mu_P)} > \delta_1 \quad \alpha_1 = 2.5\% \\ H_{02} : \frac{(\mu_T - \mu_P)}{(\mu_R - \mu_P)} \geq \delta_2 \quad \text{vs.} \quad H_{a2} : \frac{(\mu_T - \mu_P)}{(\mu_R - \mu_P)} < \delta_2 \quad \alpha_2 = 5\% \end{aligned}$$

Table 4 gives the sample size of test (C) vs. test (D). Test (D) requires a larger sample size than test (C) to attain the same power of 0.8. However, the sample size is comparable when the true ratio of mean differences is larger than 1.

CONCLUSION

An equivalence test may be carried out with a clinical trial of three parallel arms (test, reference, and placebo). The conventional approach consists of testing superiority and equivalence. The superiority test part of the process involves analyzing the test treatment and the reference treatment over placebo, while the equivalence test part of the process is to demonstrate whether the test treatment and the reference treatment are equivalent. With continuous responses, the equivalence of the treatments are commonly assessed in terms of either mean difference, mean ratio, or ratio of mean differences, where the test mean and the reference mean are adjusted by the placebo mean.

Table 4 Sample size per arm of test (C) and test (D) to attain 80% power at a given mean differences ratio δ^* with equal variance and equal sample size where $(\delta_1, \delta_2) = (0.8, 1.25)$

$\alpha_1 = 0.025$		$\delta^* = 0.95$		$\delta^* = 1.00$		$\delta^* = 1/0.95$	
$\alpha_2 = 0.05$							
σ	$\mu_R - \mu_P$	(C)	(D)	(C)	(D)	(C)	(D)
3	7	88	109	67	75	79	81
	10	44	54	33	37	39	40
	13	26	32	20	23	24	24
4	7	155	193	118	133	140	143
	10	77	95	59	66	69	71
	13	46	56	35	39	41	42
5	7	242	301	185	207	219	224
	10	119	148	91	102	108	110
	13	71	88	54	61	64	65
6	7	349	433	265	298	315	322
	10	171	213	130	146	154	158
	13	102	126	77	87	92	94

When assay sensitivity is validated by showing superiority of the reference over placebo, one may only focus on assessment of superiority of the test treatment over placebo and equivalence between the test and the reference treatments. It is also interesting to know that testing for equivalence using the ratio of mean differences may be more efficient than mean difference between the test and the reference treatments when the true ratio of mean differences is less than 1, i.e. when $(\mu_T - \mu_P) < (\mu_R - \mu_P)$.

When conducting the ratio of mean differences test, under the condition of $(\mu_T - \mu_P) / (\mu_R - \mu_P) > 0$ and a higher response score indicates a more favorable outcome, it directly implies that the test treatment is superior to placebo. In this case, one may waive the superiority test and apply a single ratio of mean differences test, but in order to satisfy the superiority test requirement, one has to perform two one-sided tests with 2.5% level for the lower side and 5% level for the upper side. As indicated in the numerical result, adding the superiority test of test over placebo to the ratio of mean differences test does not affect the power of the test much. By selecting a lower equivalence margin for the ratio of mean differences test which still satisfy the equivalence test criteria can then improve the power of the ratio of mean differences test.

ACKNOWLEDGEMENT

Part of results of this article is based on the FDA ORISE project completed in 2012 when Yu-Wei Chang served as a summer intern and Xiaoyu Dong was an FDA statistical reviewer.

REFERENCES

- Chow, S.-C.; Liu, J.-P. *Design and Analysis of Bioavailability and Bioequivalence Studies*; 3rd Ed.; Chapman and Hall/CRC Press: New York, 2008.
- FDA. Guidance on Statistical Approaches to Establishing Bioequivalence, Center for Drug Evaluation and Research, the US Food and Drug Administration, Rockville, MD, 2001.
- Hauschke, D.; Kieser, M.; Diletti, E.; Burke, M. Sample Size Determination for Proving Equivalence Based on the Ratio of Two Means for Normally Distributed Data. *Statistics in Medicine* **1999**, *18*, 93–105.
- Kang, S.H.; Chow, S.-C. Statistical assessment of biosimilarity based on relative distance between follow-on biologics. *Statistics in Medicine* **2013**, *32*, 382–329.
- Pigeot, I.; Schafer, J.; Rohmel, J.; Hauschke, D. Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. *Statistics in Medicine* **2003**, *22*, 883–899.
- Owen, D. B. A Special Case of a Bivariate Non-Central t-Distribution. *Biometrika* **1965**, *52*, 437–446.
- Satterthwaite, F. E. An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin* **1946**, *2*, 110–114.
- Schirmann, D. J. A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* **1987**, *15*, 657–680.
- Tsong, Y.; Zhang, J.; Wang, S. J. Group sequential design and analysis of clinical equivalence assessment for generic nonsystematic drug products. *Journal of Biopharmaceutical Statistics* **2004**, *14*, 359–373.

Sample Size Estimation: Generalized Estimating Equations (GEE) Method

Sin-Ho Jung

Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, U.S.A.

Abstract

This entry compares the change rate of repeated measurements over time between two treatment groups. It proposes a closed-form, sample size formula for between-group comparison that can be used when designing a randomized longitudinal study with either continuous or binary outcome variables.

Keywords: AR(1); Compound symmetry; Independent working correlation; Independent missing; Logit link; Missing completely at random; Monotone missing.

INTRODUCTION

Often, in controlled clinical trials, subjects are randomized to two treatment groups and evaluated at a baseline and at intervals across a treatment period. For example, in a study on labor pain, 83 women in labor were randomly assigned to a pain medication group (43 women) or placebo group (40 women). At 30-min intervals, the self-reported amount of pain was marked on a 100mm line, where 0 = no pain and 100 = extreme pain. The maximum number of measurements for each woman was $m = 6$, but there were numerous missing values at later measurement times with a monotone missing pattern. A simple approach to such a study objective may be to estimate and compare the slopes of pain scores over time in two treatment groups. In this study, the outcome variable is continuous.

Genetics vs Environment In Scleroderma Outcome Study (GENISOS) is an observational study designed in collaboration between the University of Texas-Houston Health Science Center, the University of Texas Medical Branch at Galveston, and the University of Texas-San Antonio Health Science Center.^[8] Scleroderma, or systemic sclerosis, is a multi-system disease of unknown etiology characterized by cutaneous and visceral fibrosis, small blood vessel damage, and autoimmune features.^[7] The study subjects are regularly followed to check the occurrence of pulmonary fibrosis. In this case, the outcome is a binary variable.

Typically, repeated measurements from each subject are correlated. Furthermore, the subjects may skip some visits, such that the resulting data are subject to missing. The Generalized Estimating Equations (GEE) method^[6] has been one of the most popular methods to relate repeatedly measured outcome variables with covariates, including

measurement time. If the missing is completely random,^[9] the GEE method provides a consistent regression estimator, to determine whether the working correlation structure specified for the repeated measurements is the true one.

This entry compares the change rate of repeated measurements over time between two treatment groups. We propose a closed-form, sample size formula for between-group comparison that can be used when designing a randomized longitudinal study with either continuous or binary outcome variables. Missing data attenuate the power of tests on the regression coefficients. Our formula accounts for the missing probabilities under an independent or monotone missing pattern. We assume that the working independent correlation model is used in the final data analysis, but the sample size formula is derived to incorporate the influence of the true correlations, especially under AR(1) or compound symmetry.

GENERALIZED ESTIMATING EQUATIONS

Suppose that n_k subjects are allocated to treatment group $k (= 1, 2)$, $n_1 + n_2 = n$. Let, for group k , $N_k = \sum_{i=1}^{n_k} m_{ki}$ denote the total number of observations and $r_k = n_k/n$ the allocation proportion ($r_1 + r_2 = 1$).

For subject i ($i = 1, \dots, n_k$) in group k , let y_{kij} denote the outcome variable at measurement time t_{kij} ($j = 1, \dots, m_{ki}$) with $\mu_{kij} = E(y_{kij})$ that is expressed as

$$g(\mu_{kij}) = a_k + b_k t_{kij},$$

where $g(\cdot)$ is a link function. To simplify the discussions, we use the identity link $g(\mu) = \mu$ for continuous outcome variables and the logit link $g(\mu) = \log\{\mu/(1-\mu)\}$ for binary

outcome variables, but the generalization to the use of other links is simple.

The coefficient b_k represents the change rate per unit time in mean response if y is continuous and in log-odds if y is binary. The measurement times may vary, subject by subject, due to missing measurements, patients' visits for measurements at unscheduled times, loss to follow-up, or other causes. In this article, we assume the missing is completely random by Rubin.^[9]

The repeated measurements (y_{kij} , $1 \leq j \leq m_{ki}$) within each subject tend to be correlated. However, the true correlation structure is usually unknown or of secondary interest. Liang and Zeger^[6] proposed a consistent estimator, called the GEE estimator, based on a working correlation structure. Using either the identity or logit link, the GEE estimator $(\hat{a}_k, \hat{b}_k)$, based on the working independent structure, solves $U_k(a, b) = 0$, where

$$U_k(a, b) = \frac{1}{\sqrt{n_k}} \sum_{i=1}^{n_k} \sum_{j=1}^{m_{ki}} \{y_{kij} - \mu_{kij}(a, b)\} \begin{pmatrix} 1 \\ t_{kij} \end{pmatrix}$$

and $\mu_{kij}(a, b) = g^{-1}(a + bt_{kij})$.

Continuous Outcome Variable Case

In the continuous outcome variable case, we have a closed form solution to $U_k(a, b) = 0$:

$$\hat{b}_k = \frac{\sum_{i=1}^{n_k} \sum_{j=1}^{m_{ki}} (t_{kij} - \bar{t}_k) y_{kij}}{\sum_{i=1}^{n_k} \sum_{j=1}^{m_{ki}} (t_{kij} - \bar{t}_k)^2}$$

and $\hat{a}_k = \bar{y}_k - \hat{b}_k \bar{t}_k$, where $\bar{t}_k = N_k^{-1} \sum_{i=1}^{n_k} \sum_{j=1}^{m_{ki}} t_{kij}$ and $\bar{y}_k = N_k^{-1} \sum_{i=1}^{n_k} \sum_{j=1}^{m_{ki}} y_{kij}$.

According to Liang and Zeger,^[6] as $n \rightarrow \infty$, $\sqrt{n_k}(\hat{b}_k - b_k) \rightarrow N(0, v_k)$ in distribution. Here, v_k is consistently estimated by

$$\hat{v}_k = \frac{\sum_{i=1}^{n_k} \left\{ \sum_{j=1}^{m_{ki}} (t_{kij} - \bar{t}_k) \hat{\varepsilon}_{kij} \right\}^2}{\left\{ \sum_{i=1}^{n_k} \sum_{j=1}^{m_{ki}} (t_{kij} - \bar{t}_k)^2 \right\}^2}$$

where $\hat{\varepsilon}_{kij} = y_{kij} - \hat{a}_k - \hat{b}_k t_{kij}$.

Binary Outcome Variable Case

Let $p_{kij} = \mu_{kij}$ in the binary outcome variable case. We have to solve the equation using a numerical method, such as the Newton–Raphson algorithm: at the l -th iteration,

$$\begin{pmatrix} \hat{a}_k^{(l)} \\ \hat{b}_k^{(l)} \end{pmatrix} = \begin{pmatrix} \hat{a}_k^{(l-1)} \\ \hat{b}_k^{(l-1)} \end{pmatrix} + n_k^{-1/2} A_k^{-1} \begin{pmatrix} \hat{a}_k^{(l-1)} \\ \hat{b}_k^{(l-1)} \end{pmatrix} \times U_k \begin{pmatrix} \hat{a}_k^{(l-1)} \\ \hat{b}_k^{(l-1)} \end{pmatrix},$$

where

$$\begin{aligned} A_k(a, b) &= -n_k^{-1/2} \frac{\partial U_k(a, b)}{\partial(a, b)} \\ &= \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{m_{ki}} p_{kij}(a, b) q_{kij}(a, b) \begin{pmatrix} 1 & t_{kij} \\ t_{kij} & t_{kij}^2 \end{pmatrix} \end{aligned}$$

$$p_{kij}(a, b) = g^{-1}(a + bt_{kij}) = \frac{e^{a+bt_{kij}}}{1 + e^{a+bt_{kij}}} \quad (1)$$

and $q_{kij}(a, b) = 1 - p_{kij}(a, b)$.

According to Liang and Zeger,^[6] for large n , $\sqrt{n_k}(\hat{a}_k - a_k, \hat{b}_k - b_k)^T$ is asymptotically normal, with mean 0 and variance V_k that can be consistently estimated by

$$\hat{V}_k = A_k^{-1}(\hat{a}_k, \hat{b}_k) \hat{\Sigma}_k A_k^{-1}(\hat{a}_k, \hat{b}_k)$$

where

$$\hat{\Sigma}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \left\{ \sum_{j=1}^{m_{ki}} \hat{\varepsilon}_{kij} \begin{pmatrix} 1 \\ t_{kij} \end{pmatrix} \right\}^{\otimes 2},$$

$\hat{\varepsilon}_{kij} = y_{kij} - p_{kij}(\hat{a}_k, \hat{b}_k)$, and $c^{\otimes 2} = cc^T$ for a vector c . Let $\hat{v}_k$ be the (2,2) component of $\hat{V}_k$.

SAMPLE SIZE CALCULATION

Suppose that we want to test on the rate of change between two groups, i.e. $H_0: b_1 = b_2$. Based on the asymptotic results from the previous section, we can reject $H_0: b_1 = b_2$, in favor of $H_1: b_1 \neq b_2$ when

$$\left| \frac{\hat{b}_1 - \hat{b}_2}{\sqrt{\hat{v}_1/n_1 + \hat{v}_2/n_2}} \right| > z_{1-\alpha/2}, \quad (2)$$

where $z_{1-\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of a standard normal distribution.

In this section, we derive a sample size formula for the two-sided test^[2] to detect $H_1: |b_2 - b_1| = d(>0)$ with power $1 - \beta$. When designing a study, we usually schedule fixed visit times $t_1 < \dots < t_m$ for m repeated measurements from each subject. We often set $t_1 = 0$ for the baseline measurement time. When the study is conducted, however, the subjects may skip some visit times due to various causes, which results in missing values, or may not follow the visit schedule correctly so that the observed visit times may be variable. Jung and Ahn^[3] show, through simulations

in continuous outcome variable cases, that the sample size formula based on fixed measurement times is very accurate, even when the observed measurement times are widely distributed around the scheduled times.

By simple algebra, we can derive a sample size formula to detect the specified difference $|b_2 - b_1| = d$ with power $1 - \beta$,

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 (v_1/r_1 + v_2/r_2)}{d^2}, \quad (3)$$

where $v_k = \lim_{n \rightarrow \infty} \hat{v}_k$. The expression of v_k is slightly different between continuous and binary outcome variable cases, as shown in the following subsections.

In order to allow for missing, let δ_j denote the proportion of patients with observation at t_j and $\delta_{jj'}$ and the proportion of patients with observations at both t_j and $t_{j'}$. Also, let $\rho_{jj'} = \text{corr}(y_{kij}, y_{kj'})$. We assume that the missing pattern and correlation structure are common in two treatment groups. "Modeling Missing Pattern and Correlation Structure" discusses specific models for missing pattern and correlation structure.

Continuous Outcome Variable Case

We assume that the continuous outcome variable has a constant variance $\text{var}(y_{kij}) = \sigma^2$ over time, which is common between two treatment groups. Jung and Ahn^[3] show that $v_1 = v_2 = v$ is expressed as

$$v = \frac{\sigma^2(s^2 + c)}{s^4},$$

where

$$s^2 = \sum_{j=1}^m \delta_j (t_j - \tau)^2$$

$$c = \sum_{j \neq j'} \delta_{jj'} \rho_{jj'} (t_j - \tau)(t_{j'} - \tau)$$

and

$$\tau = \frac{\sum_{j=1}^m \delta_j t_j}{\sum_{j=1}^m \delta_j}.$$

Hence, Eq. 3 is simplified to

$$n = \frac{v(z_{1-\alpha/2} + z_{1-\beta})^2}{d^2 r_1 r_2}. \quad (4)$$

Note that we do not have to specify the true values for $\{(a_k, b_k), k = 1, 2\}$, but only the difference $|b_2 - b_1| = d$ in

sample size calculation. The sample size is proportional to the variance of measurement error σ^2 , and decreases as r_1 approaches 1/2.

Chow et al.^[1] propose a sample size formula for ANOVA with repeated measures. Their formula is similar to the formula for K-sample by Jung and Ahn^[4] under compounding symmetry (CS) and no missing.

Binary Outcome Variable Case

Let $p_{kj} = a_k + b_k t_k$ denote the success probabilities under H_1 . Jung and Ahn^[5] show that

$$v_k = \frac{s_k^2 + c_k}{s_k^4},$$

where

$$s_k^2 = \sum_{j=1}^m \delta_j p_{kj} q_{kj} (t_j - \tau_k)^2$$

$$c_k = \sum_{j \neq j'} \delta_{jj'} \rho_{jj'} \sqrt{p_{kj} q_{kj} p_{kj'} q_{kj'}} (t_j - \tau_k)(t_{j'} - \tau_k),$$

and

$$\tau_k = \frac{\sum_{j=1}^m \delta_j p_{kj} q_{kj} t_j}{\sum_{j=1}^m \delta_j p_{kj} q_{kj}}.$$

As shown here, the sample size formula under the binary outcome variable case depends on the probabilities $(p_{kj}, j = 1, \dots, m)$ under H_1 , so that we need to specify all regression parameters $\{(a_k, b_k), k = 1, 2\}$. Let $d = b_2 - b_1$ denote the difference in slope between two groups under H_1 . If there exist pilot data for the control group (group 1), then we may use the estimates as the parameter values, a_1 and b_1 . If $t_1 = 0$ is the time of randomization, the intercepts in the two groups are set the same, i.e. $a_1 = a_2$. By setting $b_2 = b_1 + d$, we can specify all regression coefficients under H_1 .

If there are no pilot data available, we may specify the binary probabilities at the baseline, p_{11} , and at the end of the follow-up, p_{1m} , for the control group. Then, we obtain (a_1, b_1) by

$$b_1 = \frac{g(p_{1m}) - g(p_{11})}{t_m - t_1}, \quad (5)$$

$$a_1 = g(p_{11}) - b_1 t_1 = g(p_{11}).$$

And we set $a_2 = a_1$ (since $p_{11} = p_{21}$ at the baseline) and $b_2 = b_1 + d$.

MODELING MISSING PATTERN AND CORRELATION STRUCTURE

The calculation of n (or v_k) requires projection of the missing probabilities and the true correlation structure.

Missing Pattern

In order to specify $\delta_{jj'}$, we need to estimate the missing pattern. If missing at time t_j is independent of missing at time $t_{j'}$ for each patient, then we have $\delta_{jj'} = \delta_j \delta_{j'}$ and call this type *independent missing*.

In some studies, subjects missing at a measurement time may be missing at all following measurement times, as in the labor pain study discussed in the Introduction. This type of missing is called *monotone missing*. In this case, we have $\delta_{jj'} = \delta_{j'}$ for $j < j'$ ($\delta_1 \geq \dots \geq \delta_m$). In the monotone missing case, one may want to specify the proportion of patients who will contribute exactly the first j observations, say η_j . Then, noting that $\eta_j = \delta_j - \delta_{j+1}$ for $j = 1, \dots, m-1$ and $\eta_m = \delta_m$, we can obtain δ_j recursively starting from δ_m . Note also that $\sum_j \eta_j = \delta_1$, which equals 1 if all patients have measurements at the first measurement time t_1 .

In summary, we consider two missing patterns as candidates to the true one: (a) independent missing, where $\delta_{jj'} = \delta_j \delta_{j'}$ and (b) monotone missing, where $\delta_{jj'} = \delta_{j'}$ for $j < j'$ ($\delta_1 \geq \dots \geq \delta_m$).

Correlation Structure

Now, we specify the “true” correlation structure $\rho_{jj'}$. In the continuous outcome case, a reasonable model for $\varepsilon_{ij} = y_{ij} - g^{-1}(a + bt_{ij})$ may be

$$\varepsilon_{ij} = u_i + e_{ij},$$

where u_i are subject-specific error terms when variance σ_u^2 and e_{ij} are serially correlated, within-subject error terms with variance σ_e^2 and correlation coefficients $\tilde{\rho}_{jj'}$. Assuming that u_i and e_{ij} are independent, we have

$$\text{cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = \sigma_u^2 + \sigma_e^2 \tilde{\rho}_{jj'}.$$

Often, the variation between subjects (σ_u^2) is much larger than that within subject (σ_e^2). In this case, we have

$$\text{cov}(\varepsilon_{ij}, \varepsilon_{ij'}) \approx \sigma_u^2$$

and a CS correlation structure, $\rho_{jj'} = \rho$ for $j \neq j'$, that may be a reasonable approximation to the true one.

On the other hand, if the variation within subject (σ_e^2) dominates over the variation between subjects (σ_u^2), then we will have

$$\text{cov}(\varepsilon_{ij}, \varepsilon_{ij'}) \approx \sigma_e^2 \tilde{\rho}_{jj'},$$

so that a serial correlation structure may be a reasonable approximation to the true one. One of the most popular serial correlation structures, especially when measurement times are not equidistant, is a continuous autocorrelation model with order 1, AR(1), for which $\rho_{jj'} = \rho^{|t_j - t_{j'}|}$. We can make similar arguments in “discrete outcome case.”

We consider two correlation structures as candidate approximations to the true one: (i) CS, where $\rho_{jj'} = \rho$ and (ii) AR(1), where $\rho_{jj'} = \rho^{|t_j - t_{j'}|}$.

EXAMPLES

For sample size calculation, the following parameters need to be specified commonly in both continuous and discrete outcome variable cases.

- Type I and II errors α and β , respectively.
- Clinically significant difference, $|b_2 - b_1| = d$.
- Allocation proportion for group k , r_k .
- Correlation structure and the associated correlation parameter ρ , i.e. (i) $\rho_{jj'} = \rho$ for CS or (ii) $\rho_{jj'} = \rho^{|t_j - t_{j'}|}$ for AR(1).
- Proportion of patients with an observation at t_j , δ_j .
- Missing pattern: (a) independent missing ($\delta_{jj'} = \delta_j \delta_{j'}$) or (b) monotone missing ($\delta_{jj'} = \delta_{j'}$ for $j < j'$).

In addition, we need to specify $\sigma^2 = \text{var}(\varepsilon_{ij})$ in the continuous outcome variable case, and the regression coefficients $\{(a_k, b_k), k = 1, 2\}$ under H_1 in the binary outcome variable case.

We demonstrate our sample size formula with real longitudinal studies.

Labor Pain Study (Continuous Outcome Case)

Suppose that we want to design a new study on labor pain based on the data reported by Davis.^[2] The original study was briefly discussed in the Introduction. As in the original study, we assume monotone missing. In this study, the measurement times were equispaced, so that we set $t_j = j-1$ ($j = 1, \dots, 6$) for convenience. From the data, we obtained $\sigma^2 = 815.84$. Suppose we want to detect a difference of σ in mean pain score between two groups at t_6 . So, we project $d = \sigma/(t_6 - t_1) = 5.71$ in a new study. We consider a balanced randomization, i.e., $r_1 = r_2 = 1/2$. Also, from the data, the proportion of observed measurements are

$$(\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6) = (1, 0.90, 0.78, 0.67, 0.54, 0.41).$$

From these results, we obtain $\tau = 2.02$ and $s^2 = 11.42$.

Suppose that we want to detect a difference of $d = 5.71$ with 80% of power ($z_{1-\beta} = 0.84$) using a two-sided $\alpha = 0.05$ ($z_{1-\alpha/2} = 1.96$) test. Under CS, we obtain $\rho = 0.64$ and $c = -3.13$ from the data, so that we have $v = 815.84 \times (11.42 - 3.13)/11.42^2 = 0.0635$.

Hence, from Eq. 4, the required sample size is calculated as

$$n = \left\lceil \frac{0.0635 \times (1.96 + 0.84)^2}{5.71^2} \right\rceil + 1 = 50$$

where $\lceil x \rceil$ is the largest integer not exceeding x .

Under AR(1), we obtain from the data $\rho = 0.80$ and $c = 2.31$, so that the required sample size for detecting $d = 5.71$ with 80% of power using a two-sided $\alpha = 0.05$ test is given as $n = 83$. The sample size under AR(1) is larger than that under CS, by about 66%, in this example.

GENISOS (Binary Outcome Case)

As stated in the Introduction, GENISOS is an observational study. Suppose that we want to develop a new clinical trial to examine the effect of a new drug in preventing the occurrence of pulmonary fibrosis in subjects with scleroderma based on the observations from the observational study. The parameter values specified below for the sample size calculation are approximated by the estimates from the current data set of GENISOS.

We want to estimate the sample size for the new clinical trial using $\alpha = 0.05$ and $1 - \beta = 0.8$. As in GENISOS, the presence or absence of pulmonary fibrosis will be assessed at the baseline, and at months 6, 12, 18, 24, and 30. Since the measurement times are equidistant, we set $t_j = j-1$ ($j = 1, \dots, 6$) for the $m = 6$ time points. The within-group correlation structure of the repeated measurements conforms to AR(1) with the adjacent correlation equal to $\rho = 0.8$, that is, $\rho_{jj'} = 0.8^{|j-j'|}$. We consider assigning an equal number of scleroderma patients in each of two groups, i.e., $r_1 = r_2 = 1/2$.

Approximately 75% of scleroderma patients do not have pulmonary fibrosis at the baseline in the ongoing GENISOS. We project that the proportion of subjects without pulmonary fibrosis is 75% at the baseline, i.e., $p_{1,1} = 0.75$ and 50% at 30 months, i.e., $p_{1,6} = 0.50$, in a placebo group. We assume that a new therapy will prevent or delay further occurrence of pulmonary fibrosis. That is, the proportion of subjects without pulmonary fibrosis will remain 75% during a 30-month study in a new therapy group, i.e., $p_{2,1} = p_{2,6} = 0.75$. From these values and Eq. 5, we obtain

$$b_1 = \frac{g(0.5) - g(0.75)}{5 - 0} = -0.220,$$

and $a_1 = g(0.75) = 1.099$. Similarly, we obtain $(a_2, b_2) = (1.099, 0)$, so that $d = 0 - (-0.220) = 0.220$. By Eq. 1, the probabilities of no pulmonary fibrosis at the 6 time points

are obtained as (0.750, 0.707, 0.659, 0.608, 0.555, 0.500) for the placebo group, and (0.750, 0.750, 0.750, 0.750, 0.750, 0.750) for the treatment group.

The proportions of observed measurements are expected to be

$$(\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6) = (1, 0.95, 0.90, 0.85, 0.80, 0.75).$$

Suppose that we expect independent missing. Then, we obtain $\delta_{jj'}$ from the specified δ_j values using $\delta_{jj'} = \delta_j \delta_{j'}$. Now we have all the parameter values required, and obtain $v_1 = 0.305$ and $v_2 = 0.353$. Finally, from Eq. 3, we obtain

$$n = \left\lceil \frac{(1.96 + 0.84)^2 (0.305/0.5 + 0.353/0.5)}{0.220^2} \right\rceil + 1 = 215.$$

If we assume monotone missing ($\delta_{jj'} = \delta_{j \vee j'}$), we obtain $v_1 = 0.324$, $v_2 = 0.308$, and $n = 229$.

ARTICLES OF FURTHER INTEREST

Analysis of Repeated Measures Data with Missing Values: An Overview of Method; Clinical Trials; GEE; Logistic Regression; Randomization.

REFERENCES

1. Chow, S.C.; Shao, J.; Wang, H. *Sample Size Calculations in Clinical Research*; Taylor & Francis: New York, 2003.
2. Davis, C.S. Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials. *Stat. Med.* **1991**, *10*, 1959–1980.
3. Jung, S.H.; Ahn, C. Sample size estimation for GEE method for comparing slopes in repeated measurements data. *Stat. Med.* **2003**, *22*, 1305–1315.
4. Jung, S.H.; Ahn, C. K-sample test and sample size calculation for comparing slopes in data with repeated measurements. *Biometrical J.* **2004**, *46*, 554–564.
5. Jung, S.H.; Ahn, C. Sample size for a two-group comparison of repeated binary measurements using GEE. *Stat. Med.* **2005**, *24*, 2583–2596.
6. Liang, K.Y.; Zeger, S.L. Longitudinal data analysis using generalized linear models. *Biometrika* **1986**, *73*, 13–22.
7. Medsger, T.A. Jr. Systemic sclerosis (scleroderma): clinical aspects. In *Arthritis and Allied Conditions*; Koopman, W.J., Ed., 13th Ed.; Williams and Wilkins: Baltimore, MD, 1997; 1433–1464.
8. Reveille, J.; Fischbach, M.; McNearney, T.; et al. Systemic sclerosis in 3 US ethnic groups: a comparison of clinical, sociodemographic, serologic and immunologic determinants. *Semin. Arthritis Rheum.* **2001**, *30*, 332–346.
9. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592.

Sample Size for Multiregional Clinical Trials

Yuh-Jenn Wu

Department of Applied Mathematics, Chung Yuan Christian University, Taoyuan, Taiwan

Chieh Chiang, Chi-Tian Chen, Hsiao-Hui Tsou, and Chin-Fu Hsiao

Institute of Population Health Sciences, National Health Research Institutes, Miaoli, Taiwan

Abstract

To speed up drug development to allow faster access to medicines for patients globally, the design of multiregional clinical trials incorporates subjects from many countries around the world under the same protocol. The aim of a multiregional clinical trial is to show the efficacy of a drug in various global regions, and concurrently to evaluate the possibility of applying the overall trial results to each region so that simultaneous drug development, submission, and approval in the world are possible. In this entry, we reviewed some recent developments for sample size determination for MRCT. In the first part, we reviewed the methods for sample size calculation based on the assumptions that the effect size is uniform across regions and treatment effect is heterogeneous across regions respectively. In the second part, we reviewed statistical criteria for assessing the consistency between the region of interest and overall results.

Keywords: Bridging study; Consistent Trend; Ethnic effect; Multi-regional clinical trials.

1 INTRODUCTION

Biotechnology has been identified as one of the key technologies in the 21st century. Especially, biopharmaceutical industry is the most vital and important industry. In recent years, many pharmaceutical companies and regulatory agencies have spent efforts to promote global drug development and international clinical study. The aim of a MRCT is to show the efficacy of a drug in various global regions, and concurrently to evaluate the possibility of applying the overall trial results to each region. Doing so may enable simultaneous drug development, submission, and approval in the world so that the drug lag or the time lag for drug approval can be shortened in key markets.

Recently, the Japanese Ministry of Health, Labour and Welfare (MHLW) published the “Basic Principles on Global Clinical Trials” (1) guidance to promote Japan’s participation in global drug development. By Japanese guidance, global clinical studies refer to studies planned with the objective of world-scale development and approval of new drugs in which study sites of a multiple number of countries and regions participate in a single study based on a common protocol and conducted at the same time in parallel. The ICH E5 11th Q&A (2) also discusses the concept of a MRCT and states that “It may be desirable in certain situations to achieve the goal of bridging by conducting a multi-regional trial under a common protocol that includes sufficient numbers of patients from each of

multiple regions to reach a conclusion about the effect of the drug in all regions...” In 2016, the ICH released the draft guidance E17 of MRCT entitled “General Principles for Planning and Design of Multi-Regional Clinical Trials.” (3) The purpose of E17 is to describe general principles for the planning and design of MRCTs with the aim of increasing the acceptability of MRCTs in global regulatory submissions. All guidelines provide a framework on how to demonstrate the efficacy of a drug in all participating regions while also evaluating the possibility of applying the overall trial results to each region in a MRCT. Some practical issues that are commonly encountered when conducting multiregional clinical trials including criteria for assessment of similarity between regions and sample size calculation/allocation are also discussed in the review article by Chow and Hsiao (4).

When planning MRCTs, the differences in race, diet, environment, culture, and medical practice among regions may have an impact, and should be identified during the planning stage. In the first part of this entry, we will review some recent developments for sample size determination for MRCT. Another important issue is to evaluate the possibility of applying the overall trial results in a MRCT to each region. The Japanese guidance has provided two methods as examples for establishing consistency in treatment effect between the Japanese group and the entire group. In the second part of this entry, we will review statistical criteria for assessing the consistency between the region of interest and overall results.

2 SAMPLE SIZE FOR MRCT

Assume that the MRCT is conducted to compare a test product and a placebo control on a continuous efficacy endpoint. Let M be the number of regions.

2.1 Effect size is uniform across regions

Let X_{ij} and Y_{ik} be the efficacy responses for the j^{th} subject and the k^{th} subject in the i^{th} region receiving the test product and the placebo control respectively, $i = 1, \dots, M$, $j = 1, \dots, n_i^T$, and $k = 1, \dots, n_i^C$. Assume that $X_{ij} \sim N(\mu^T, \sigma^2)$ and $Y_{ik} \sim N(\mu^C, \sigma^2)$, where $N(\mu, \zeta^2)$ represents a normal distribution with mean μ and variance ζ^2 . Here a higher value of the population mean represents a better outcome. For convention, we assume that σ^2 is known, although in actual practice, σ^2 is unknown and must be estimated from some data. Let $\Delta = \mu^T - \mu^C$. It is assumed that effect size (Δ/σ) is uniform across regions. The hypothesis of testing for the overall treatment effect is given as

$$H_0: \Delta \leq 0 \text{ vs. } H_A: \Delta > 0. \quad (1)$$

Although the hypothesis is one-sided, the method proposed below can be straightforwardly extended to the two-sided hypothesis. Let Z the test statistic for (1). It can be derived that

$$Z = \frac{\left(\sum_{i=1}^M \sum_{j=1}^{n_i^T} X_{ij} \right) / \sum_{i=1}^M n_i^T - \left(\sum_{i=1}^M \sum_{k=1}^{n_i^C} Y_{ik} \right) / \sum_{i=1}^M n_i^C}{\sigma \sqrt{\frac{1}{\sum_{i=1}^M n_i^T} + \frac{1}{\sum_{i=1}^M n_i^C}}}.$$

Let N denote the total sample size for each group planned for detecting an expected treatment difference $\Delta = \delta$ at the desired significance level α and with power $1 - \beta$. Thus,

$$N = \frac{2(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{\delta^2},$$

where $z_{1-\alpha}$ is the $(1 - \alpha)^{\text{th}}$ percentile of the standard normal distribution (5, 6). For example, by considering $\alpha = 0.025$, $\beta = 0.1$, $\delta = 6$, and $\sigma = 14$, the sample size required for each group is 115.

2.2 Heterogeneous treatment effect across regions

Since the design of MRCTs incorporates subjects from many countries around the world, it might be reasonable to anticipate heterogeneity in treatment effect due to ethnic differences. In this section, much of the notation

developed in Section 2.1 will continue to be used. Different from Section 2.1, here we assume that $X_{ij} \sim N(\mu_i^T, \sigma_i^2)$ and $Y_{ik} \sim N(\mu_i^C, \sigma_i^2)$. Let θ_i be the treatment difference in the i^{th} region. That is, $\theta_i = \mu_i^T - \mu_i^C$. To address the heterogeneous treatment effect across regions, θ_i is considered as a random effect. More specifically, Chen et al. (7) assumed that $\theta_i \sim N(\theta, \tau^2)$, $i = 1, \dots, M$. Here θ can be thought of as the overall treatment effect for the MRCT. The hypothesis of testing for the overall treatment effect in the MRCT can thus be given as

$$H_0: \theta \leq 0 \text{ vs. } H_A: \theta > 0. \quad (2)$$

Let $\hat{\theta}_i$ represent the estimator of θ_i . It leads to $\hat{\theta}_i = \bar{X}_i - \bar{Y}_i$, where $\bar{X}_i = \left(\sum_{j=1}^{n_i^T} X_{ij} \right) / n_i^T$ and $\bar{Y}_i = \left(\sum_{k=1}^{n_i^C} Y_{ik} \right) / n_i^C$. Consequently, the general random effect model can be given by

$$\hat{\theta}_i = \theta + \nu_i + \varepsilon_i,$$

where $\nu_i \sim N(0, \tau^2)$, $\varepsilon_i \sim N(0, \xi_i^2)$, and ν_i and ε_i are assumed to be independent. It follows that $\hat{\theta}_i \sim N(\theta, \xi_i^2 + \tau^2)$ where

$$\xi_i^2 = \frac{\sigma_i^2}{n_i^T} + \frac{\sigma_i^2}{n_i^C} = \sigma_i^2 \left(\frac{1}{n_i^T} + \frac{1}{n_i^C} \right).$$

Let $\hat{\sigma}_i^2$, $\hat{\tau}^2$, and $\hat{\xi}_i^2$ be the estimators of σ_i^2 , τ^2 , and ξ_i^2 respectively. We obtain that

$$\hat{\sigma}_i^2 = \frac{(n_i^T - 1)s_{iX}^2 + (n_i^C - 1)s_{iY}^2}{n_i^T + n_i^C - 2} \text{ and } \hat{\xi}_i^2 = \hat{\sigma}_i^2 \left(\frac{1}{n_i^T} + \frac{1}{n_i^C} \right),$$

where $s_{iX}^2 = \sum_{j=1}^{n_i^T} (X_{ij} - \bar{X}_i)^2 / (n_i^T - 1)$ and $s_{iY}^2 = \sum_{k=1}^{n_i^C} (Y_{ik} - \bar{Y}_i)^2 / (n_i^C - 1)$. Set $w_i^{-1} = \xi_i^2$ and let $\hat{w}_i$ be the estimator of w_i . By DerSimonian and Laird (8), the method of moments estimator $\hat{\tau}^2$ for τ^2 is given by

$$\hat{\tau}^2 = \frac{\hat{w}}{\hat{w}^2 - \sum_{i=1}^M \hat{w}_i^2} \left\{ \sum_{i=1}^M \hat{w}_i \left(\hat{\theta}_i - \sum_{i=1}^M \frac{\hat{w}_i \hat{\theta}_i}{\hat{w}} \right)^2 - (M - 1) \right\},$$

where $\hat{w} = \sum_{i=1}^M \hat{w}_i$. Note that the estimate $\hat{\tau}^2$ will be set to 0 if a negative estimate for τ^2 is derived. With use of the distributional assumption, we can obtain

$$\hat{\theta}_i \sim N(\theta, w_i^{-1} + \tau^2).$$

Set $w_i^* = (w_i^{-1} + \tau^2)^{-1}$. Let $\hat{w}_i^*$ be the estimate of w_i^* . Consequently,

$$\hat{w}_i^* = (\hat{w}_i^{-1} + \hat{\tau}^2)^{-1}.$$

Note that the maximum likelihood estimate of θ is given by $\hat{\theta}^*$, where

$$\hat{\theta}^* = \sum_{i=1}^M \hat{\theta}_i w_i^* / \sum_{i=1}^M w_i^*,$$

by treating the term w_i^{*-1} as if it were the true variance of $\hat{\theta}_i$. It can be seen that $\hat{\theta}^*$ is asymptotically unbiased for θ , with variance approximately equal to $1 / \sum_{i=1}^M \hat{w}_i^*$. Let $S = \left[\sum_{i=1}^M \hat{w}_i^* (\hat{\theta}_i - \hat{\theta}^*)^2 \right] / \left(\sum_{i=1}^M \hat{w}_i^* \right)$. Since the number of regions, M , should be small, by Hartung (9), $\left(\sum_{i=1}^M \hat{w}_i^* \right) S$ can be approximated by a (central) χ^2 -distribution with $(M-1)$ degrees of freedom and $\hat{\theta}^*$ and S are stochastically independent. Consequently, the test statistic for (2) under H_0 is given as

$$T = \frac{\hat{\theta}^* \sqrt{\sum_{i=1}^M \hat{w}_i^*}}{\sqrt{\left(\sum_{i=1}^M \hat{w}_i^* \right) S / (M-1)}} = \frac{\hat{\theta}^*}{\sqrt{S / (M-1)}} \sim t_{M-1}, \quad (3)$$

where t_n represents the t distribution with degrees of freedom n .

Under the alternative hypothesis that $\theta = \Delta$, the test statistic T in (3) has a noncentral t distribution with $(M-1)$ degrees of freedom and noncentrality parameter $\sqrt{(M-1) \sum_{i=1}^M w_i^* \Delta}$. Let N be the total sample size required per group planned for detecting an expected treatment difference $\theta = \Delta$ at the desired significance level α and with power $1 - \beta$ for the MRCT, and p_i denote the proportion of patients out of $2N$ in the i^{th} region, $i = 1, \dots, M$, where $\sum p_i = 1$. Consequently, given σ_i^2 and τ^2 , N needs to satisfy

$$\begin{aligned} 1 - \beta &= P(T > t_{1-\alpha, M-1} \mid \theta = \Delta) \\ &= 1 - F_{M-1} \left(t_{1-\alpha, M-1} \mid \sqrt{(M-1) \left(\sum_{i=1}^M w_i^* \right) \Delta} \right) \\ &= 1 - F_{M-1} \left(t_{1-\alpha, M-1} \mid \Delta \sqrt{(M-1) \sum_{i=1}^M \left(\frac{2\sigma_i^2}{p_i N} + \tau^2 \right)} \right), \end{aligned} \quad (4)$$

where $t_{1-\alpha, n}$ denotes the upper α_{th} percentile of the t distribution with degrees of freedom n , and $F_{M-1}(\cdot \mid \delta)$ is the cumulative distribution function of the noncentral t -distribution with $(M-1)$ degrees of freedom and noncentral parameter δ . If $\sigma_i^2 = \sigma^2$, the equation (4) reduces to

$$\beta = F_{M-1} \left(t_{\alpha, M-1} \mid \Delta \sqrt{(M-1) \sum_{i=1}^M \left(\frac{2\sigma^2}{p_i N} + \tau^2 \right)} \right),$$

Given τ^2 , σ^2 , p_i and Δ , N can thus be derived by the above equation with specification of α and β .

Another alternative approach is proposed by Hung et al. (10). They derive that

$$\hat{\theta}_i \mid \theta_i \sim N(\theta_i, \xi_i).$$

If θ is interpretable for each region, we can estimate θ by $\hat{\theta}$, where

$$\hat{\theta} = \sum_{i=1}^M p_i \hat{\theta}_i.$$

It follows that

$$E(\hat{\theta}) = \theta$$

and

$$\text{Var}(\hat{\theta}) = 2\sigma^2 \left[\frac{1}{N} + \left(\frac{\tau}{2\sigma} \right)^2 \sum_{i=1}^M p_i^2 \right],$$

Where $\sigma_i^2 = \sigma^2$, $i = 1, \dots, M$. Let N^* be the total sample size required per group planned for detecting an expected treatment difference $\theta = \Delta$ at the desired significance level α and with power $1 - \beta$ for the MRCT. Assuming $\tau^2 \neq 0$, we can thus derive that

$$N^* = \left[\left(\frac{\Delta}{2\sigma(z_\alpha + z_\beta)} \right) - \left(\frac{\tau}{2\sigma} \right)^2 \sum_{i=1}^M p_i^2 \right]^{-1}.$$

By considering $\alpha = 0.025$, $\beta = 0.2$, $M = 3$, $\Delta = 13$, and $\sigma = 20$, Tables 1 and 2 exhibit the total sample size required

Table 1 The Sample Sizes of N and N^* when $\Delta = 13$, $\tau^2 = 1$, $\sigma = 20$

p_1	p_2	p_3	N	N^*
$p_1 = p_2 < p_3$				
0.1	0.1	0.8	81	77
0.15	0.15	0.7	80	77
0.2	0.2	0.6	79	76
0.25	0.25	0.5	79	76
0.3	0.3	0.4	79	76
$p_1 < p_2 = p_3$				
0.1	0.45	0.45	79	76
0.15	0.425	0.425	79	76
0.2	0.4	0.4	79	76
0.25	0.375	0.375	79	76
0.3	0.35	0.35	79	76
$p_1 < p_2 < p_3$				
0.1	0.25	0.65	80	77
0.15	0.3	0.55	79	76
0.2	0.35	0.45	79	76

Table 2 The Sample Sizes of N and N^* when $\Delta = 13$, $\tau^2 = 4$, $\sigma = 20$

p_1	p_2	p_3	N	N^*
$p_1 = p_2 < p_3$				
0.1	0.1	0.8	100	85
0.15	0.15	0.7	94	83
0.2	0.2	0.6	91	81
0.25	0.25	0.5	88	80
0.3	0.3	0.4	87	80
$p_1 < p_2 = p_3$				
0.1	0.45	0.45	90	81
0.15	0.425	0.425	89	81
0.2	0.4	0.4	88	80
0.25	0.375	0.375	87	80
0.3	0.35	0.35	87	80
$p_1 < p_2 < p_3$				
0.1	0.25	0.65	93	82
0.15	0.3	0.55	90	81
0.2	0.35	0.45	88	80

per group by Chen et al. (8) and Hung et al. (10) for different combinations of design parameters with (p_1, p_2, p_3) , $\tau^2 = 1$, and $\tau^2 = 4$ respectively.

3 CONSISTENCY CRITERIA IN MRCT

When planning a MRCT, the definitions of all regions are very critical (11). In ICH E5, the ethnic factors are classified into two categories: intrinsic and extrinsic factors. Intrinsic ethnic factors are more genetic and physiologic in nature (e.g., genetic polymorphism, age, gender, etc.) that can define and recognize the population in each region and perhaps influence the ability to extrapolate clinical data between regions. On the contrary, extrinsic ethnic factors are more social and cultural in nature (e.g., medical practice, diet and practices in clinical trials and conduct) which are associated with the environment and culture. Most importantly, ICH E17 suggests that “ethnic factors are a major point of consideration when planning MRCTs and should be identified during the planning stage, and information about them should also be collected and evaluated when conducting MRCTs.” Here region may refer to a geographical region, country or regulatory region in MRCT. ICH E17 also suggests that some regions may be pooled at the design stage.

In a MRCT, after showing the overall efficacy of a drug in all global regions, we can simultaneously evaluate the possibility of applying the overall trial results to each region and consequently support registration in each region. Therefore, how to bridge the results of the MRCT to each region is another important issue. In this regard,

ICH E17 has provided some approaches. Briefly, the first approach is to show that similar trends in treatment effects across regions can be demonstrated. The second approach is to show that the region-specific treatment effect preserves some pre-specified proportion of the overall treatment effect. The third approach is to show that significant results within one or more regions can be achieved. Similar to the second approach suggested by ICH E17, the Japanese MHLW guidance (1) has provided two methods as examples for deciding on the number of Japanese subjects in a MRCT for establishing the consistency of treatment effects between the Japanese group and the entire group.

For Method 1, it is suggested that given that the overall result is significant at α level, we will judge whether the treatment is effective in the specific region by the following criterion:

$$D_s \geq \rho D_{All}, \quad (5)$$

for some pre-specified $0 < \rho < 1$, where D_s and D_{All} are the treatment effects for the specific region and the entire group, respectively. Selection of the magnitude, ρ , of consistency trend may be critical. All differences in ethnic factors between the specific region and other regions should be taken into account. The Japanese MHLW suggests that ρ be 0.5 or greater. However, the determination of ρ will be and should be different from product to product, and from therapeutic area to therapeutic area. The design of such a study involves determination of sample size for the specific region to achieve the required power for the consistency assessment. That is, the selected sample size should satisfy

$$P(D_s/D_{All} > \pi) > \gamma.$$

Here γ should be selected to be 0.8 or greater. Other statistical criteria for consistency between the specific region and overall results can also be found in Uesaka (12) and Ko et al. (5), and are shown as follows.

- $D_s \geq \rho D_{sc}$ for some $0 < \rho < 1$,
- $D_s \geq \rho D_{All}$ for some $0 < \rho < 1$,
- $\rho \leq D_s/D_{sc} \leq 1/\rho$, for some $0 < \rho < 1$,
- $\rho \leq D_s/D_{All} \leq 1/\rho$, for some $0 < \rho < 1$,

where D_{sc} be the observed mean difference from regions other than the specific region. Note that the consistency criterion (ii) is equivalent to (5). These consistency criteria can be thought of as that the estimate of treatment effect from the specific region is noninferior/equivalent to the estimate of treatment effect from overall or from regions other than the specific region.

Let p_s denote the proportion of patients out of the overall sample size in the specific region. For convention, by considering the assumption in Section 2.1 that the effect size (Δ/σ) is uniform across regions, we now describe the method of determination of p_s to ensure that the assurance

probabilities of criteria (i)–(iv) given $\Delta = \delta$ are maintained at a desired level, say 80%.

The assurance probability of criterion (i) given $\Delta = \delta$ can be expressed as

$$AP_1 = P_\delta(D_s > \rho D_{sc} | Z > Z_{1-\alpha}) \\ = \left[\int_{a_1}^{\infty} \left(\Phi\left(\frac{u-c_2}{c_1}\right) - \Phi\left(\frac{c_5-c_3u}{c_4}\right) \right) \varphi(u) du \right] / (1-\beta),$$

where P_δ is the probability measure with respect to $\Delta = \delta$,

$$c_1 = \rho \sqrt{\frac{p_s}{1-p_s}}, \quad c_2 = (\rho-1)\sqrt{p_s}(Z_{1-\alpha} + Z_{1-\beta}), \quad c_3 = \sqrt{p_s},$$

$$c_4 = \sqrt{1-p_s}, \quad c_5 = -Z_{1-\beta}, \quad \text{and} \quad a_1 = c_2 + \frac{c_1(c_5 - c_2c_3)}{c_1c_3 + c_4}.$$

Similarly, the assurance probabilities of criteria (ii), (iii), and (iv) can be represented by

$$AP_2 = P_\delta(D_s > \rho D_{All} | Z > Z_{1-\alpha}) \\ = \left[\int_{a_2}^{\infty} \left(\Phi\left(\frac{u-c_7}{c_6}\right) - \Phi\left(\frac{c_5-c_3u}{c_4}\right) \right) \varphi(u) du \right] / (1-\beta),$$

$$\text{where } c_6 = \frac{\rho\sqrt{p_s(1-p_s)}}{1-\rho p_s}, \quad c_7 = \frac{(\rho-1)\sqrt{p_s}}{1-\rho p_s}(Z_{1-\alpha} + Z_{1-\beta}),$$

$$\text{and } a_2 = c_7 + \frac{c_6(c_5 - c_7c_3)}{c_6c_3 + c_4};$$

$$AP_3 = P_\delta\left(\rho < \frac{D_s}{D_{sc}} < \frac{1}{\rho} | Z > Z_{1-\alpha}\right) \\ = AP_1 - \left[\int_{a_3}^{\infty} \left(\Phi\left(\frac{u-c_9}{c_8}\right) - \Phi\left(\frac{c_5-c_3u}{c_4}\right) \right) \varphi(u) du \right] / (1-\beta),$$

$$\text{where } c_8 = \frac{1}{\rho} \sqrt{\frac{p_s}{1-p_s}}, \quad c_9 = \left(\frac{1}{\rho} - 1\right)\sqrt{p_s}(Z_{1-\alpha} + Z_{1-\beta}), \quad \text{and}$$

$$a_3 = c_9 + \frac{c_8(c_5 - c_9c_3)}{c_8c_3 + c_4};$$

$$AP_4 = P_\delta\left(\rho < \frac{D_s}{D_{All}} < \frac{1}{\rho} | Z > Z_{1-\alpha}\right) \\ = P_\delta\left(\rho < \frac{D_s}{D_{All}} | Z > Z_{1-\alpha}\right) - P_\delta\left(\frac{1}{\rho} \leq \frac{D_s}{D_{All}} | Z > Z_{1-\alpha}\right) \\ = P_\delta(D_s > \rho D_{All} | Z > Z_{1-\alpha}) - P_\delta\left(D_s > \frac{1}{\rho} D_{All} | Z > Z_{1-\alpha}\right) \\ = AP_2 - P_\delta\left(D_s > \frac{1}{\rho} D_{All} | Z > Z_{1-\alpha}\right).$$

Given $\alpha = 0.025$, $\beta = 0.1$, and $\rho = 0.5$, Table 3 exhibits the assurance probabilities of criteria (i), (ii), (iii), and (iv) respectively for various values of p_s . It can be seen from

Table 3 The Assurance Probability for Observing Criteria (i), (ii), (iii), and (iv) given $\alpha = 0.025$, $\beta = 0.1$, and $\rho = 0.50$

p_s	AP_1	AP_2	AP_3	AP_4
0.10	0.71	0.72	0.52	0.58
0.15	0.74	0.76	0.59	0.68
0.20	0.78	0.80	0.64	0.75
0.25	0.80	0.83	0.67	0.80
0.30	0.82	0.86	0.70	0.84
0.35	0.83	0.89	0.71	0.87
0.40	0.85	0.91	0.72	0.90
0.45	0.86	0.93	0.73	0.92
0.50	0.86	0.94	0.73	0.94
0.55	0.87	0.96	0.73	0.96
0.60	0.88	0.97	0.72	0.97
0.65	0.88	0.98	0.71	0.98
0.70	0.88	0.99	0.69	0.99
0.75	0.87	0.99	0.67	0.99
0.80	0.86	1.00	0.64	1.00

Table 2 that the assurance probabilities of criterion (iii) will never reach 80% regardless of the value of p_s . However, the sample sizes for the first region have to be around 25% out of the total sample size respectively to ensure the assurance probabilities of criteria (i), (ii) and (iv) to be at least 80%.

Alternatively, based on the consistency criterion (7), Quan et al. (13) derived closed form formulas for the sample size calculation for the specific region for normal, binary and survival endpoints. For example, if we choose $\rho = 0.5$, $\gamma = 0.2$, $\alpha = 0.025$, and $\beta = 0.9$, then the Asian sample size has to be at least 22.4% of the overall sample size for the MRCT.

On the other hand, suppose there are three regions participated in the MRCT. For Method 2, the sample size should be set so that

$$P(D_1 > 0, D_2 > 0, D_3 > 0) > \gamma,$$

where D_i represents the observed treatment effect for region i , $i = 1, \dots, 3$. Again, γ should be chosen to be 0.8 or greater. Based on Method 2, Kawai et al. (4) proposed an approach to rationalize partitioning the total sample size among the regions so that a high probability of observing a consistent trend under the assumed treatment effect across regions can be derived, if the treatment effect is positive and uniform across regions in a MRCT.

4 CONCLUSION

A MRCT may incorporate subjects from many countries around the world under the same protocol. After showing the overall efficacy of a drug in all global regions, we

can simultaneously evaluate the possibility of applying the overall trial results to each region and consequently support registration in each region. In Asian countries, the trend for simultaneous clinical development being undertaken simultaneously with clinical trials conducted in Europe and the United States has been speedily rising. In particular, Taiwan, Korea, Hong-Kong, and Singapore have already had much experience in planning and conducting the MRCTs.

In this entry, we reviewed some recent developments for sample size determination for MRCT. In the first part, we reviewed the methods for sample size calculation based on the assumptions that the effect size is uniform across regions and treatment effect is heterogeneous across regions respectively. In the second part, we reviewed statistical criteria for assessing the consistency between the region of interest and overall results.

ACKNOWLEDGMENTS

The authors would like to thank the Editor of EBS, Professor Shein-Chung Chow, for inviting us to contribute this entry.

REFERENCES

1. Ministry of Health, Labour and Welfare of Japan (MHLW). Basic Principles on Global Clinical Trials. Tokyo, Japan, 2007.
2. ICH, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Q&A for the ICH E5 Guideline on Ethnic Factors in the Acceptability of Foreign Data, 2006. Available at: <http://www.ich.org/LOB/media/MEDIA1194.pdf>.
3. ICH, International Conference on Harmonisation. Draft Guidance E17 General Principles for Planning and Design of Multi-Regional Clinical Trials, 2016. Available at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E17/E17_Final_Concept_Paper_July_2014.pdf.
4. Chow, S.C.; Hsiao, C.F. Bridging diversity: extrapolating foreign data to a new region. *Pharmaceutical Medicine* **2010**, *24* (6), 349–362.
5. Kawai, N.; Stein, C.; Komiyama O.; Li, Y. An approach to rationalize partitioning sample size into individual regions in a multiregional trial. *Drug Information Journal* **2008**, *42*, 139–147.
6. Ko, F.S.; Tsou, H.H.; Liu, J.P.; Hsiao, C.F. Sample size determination for a specific region in a multi-regional trial. *Journal of Biopharmaceutical Statistics* **2010**, *20*, 870–885.
7. Chen, C.T.; Hung, H.M.J.; Hsiao, C.F. Design and evaluation of multi-regional trials with heterogeneous treatment effect across regions. *Journal of Biopharmaceutical Statistics* **2012**, *22* (5), 1037–1050.
8. DerSimonian, R.; Laird, N. Meta-analysis in clinical trials. *Controlled Clinical Trials* **1986**, *7*, 177–188.
9. Hartung, J. An alternative method for meta-analysis. *Biometrical Journal* **1999**, *41* (8), 901–916.
10. Hung, H.M.J.; Wang, S.J.; O'Neill, R. Consideration of regional difference in design and analysis of multi-regional trials. *Pharmaceutical Statistics* **2010**, *9* (3), 173–178.
11. Tsou, H.H.; Chow, S. C.; Lan, K.K.G.; Liu, J.P.; Wang, M.; Chern, H.D.; Ho, L.T.; Hsiung, C.A.; Hsiao CF. Proposals of statistical consideration to evaluation of results for a specific region in multi-regional trials—Asian perspective. *Pharmaceutical Statistics* **2010**, *9*, 201–206.
12. Uesaka, H. Sample size allocation to regions in a multiregional trial. *Journal of Biopharmaceutical Statistics* **2007**, *19*, 580–594.
13. Quan, H.; Zhao, P.L.; Zhang, J.; Roessner, M.; Aizawa, K. Sample size considerations for Japanese patients in a multi-regional trial based on MHLW guidance. *Pharmaceutical Statistics* **2010**, *9*, 100–112.

Sample Size Re-estimation Based on Observed Treatment Difference

Lu Cui

Aventis Pharmaceuticals, Inc., Bridgewater, Connecticut, U.S.A.

Abstract

This entry discusses statistical issues of sample size re-estimation as well as the benefits it presents to certain studies.

INTRODUCTION

Sample size re-estimation based on an interim treatment difference is an adaptive clinical trial strategy that allows modification of the initially projected total sample size based on an observed treatment difference at an interim stage of a clinical trial. Such an adaptive design has drawn much attention from clinical trial practitioners as a flexible alternative to the traditional fixed sample size design.

FIXED SAMPLE SIZE DESIGN

For a fixed sample size trial, the total sample size may be determined using the prior knowledge on the true treatment difference, the variability of the response variable, and the desired Type I and Type II error rates. Under the two-sample normal mean test setting, the dependency of the total sample size on the mentioned parameters is given by the equation

$$N = (Z_{\alpha} + Z_{\beta})^2 \sigma^2 / \Delta^2 \quad (1)$$

where Δ denotes the true treatment difference, σ denotes the standard deviation of the response variable, and α and β denote the desired Type I and Type II error rates, respectively. With the given error rates and known σ , the planned sample size for a clinical trial is obtained by replacing the unknown quantity Δ in Eq. 1 with its initial estimate, say δ . The accuracy of δ or its closeness to Δ , therefore, is crucially important for an accurate projection of the total sample size needed. Improperly chosen δ may result in an insufficient total sample size, potentially leading to the failure of the trial, or an overestimated sample size, unnecessarily increasing the cost and duration of the trial.

For fixed sample size clinical trials, the δ is primarily obtained based on previous trial data. There are, however, often situations in clinical trials where a reliable initial estimate of the true treatment difference is difficult or

impossible to obtain. For example, in clinical trials to study drugs in new classes or drugs for new indications in new patient populations, typically, little or no prior information on possible effect size of the drugs is available.

SAMPLE SIZE MODIFICATION BASED ON OBSERVED TREATMENT DIFFERENCE

The adaptive design allowing sample size re-estimation based on observed interim treatment difference is to target the situation where there is no available prior information to allow a reliable initial estimate of the true treatment difference and an accurate projection of the needed total number of patients is difficult to obtain if not impossible. Such a design is intended to provide the trial necessary flexibility to adjust the sample size in the midcourse of a clinical trial if the initially planned sample size is discovered improper.

For clinical trials in which the choice of δ may also be based on some other considerations, such as the clinically meaningful treatment difference (if it exists and is well defined), the constrain of available resource, the drug marketing strategy, etc., a minimum size of the treatment difference to be detected, denoted by $\delta_{\min}$, may be specified. In this case, a fixed sample size design with the total sample size projected based on $\delta = \delta_{\min}$ may be the choice because of its simplicity and efficiency. Even in this case, it might be wise to put a sample size adaptive plan into this chosen design just in case that the initially planned total sample size is significantly wrong. If no indication of an improper total sample size, the trial can go on according to the plan.

ISSUES ON SAMPLE SIZE MODIFICATION

Several issues need special attention when sample size re-estimation based on observed treatment difference is used.

Inflation of Type I Error Rate

It has been known that Type I error rate can be inflated when the total sample size is modified based on observed interim treatment difference but such an adaptation is not properly taken into account in the analysis.^[1–4] After the outcome-dependent sample size modification, the new total sample size is no longer a fixed but random quantity depending on the magnitude of the observed interim treatment difference. Consequently, if a traditional test statistic and critical value are used as if the sample size were fixed, the corresponding Type I error rate may not be at the target level. For example, if no interim testing is performed, with the sample size modification, the Type I error rate is $E(P(T > C|M))$ instead of usual $P(T > C)$ under the null hypothesis, where T and C are test statistic and critical value, respectively, and M is the new total random sample size. For a group sequential clinical trial,^[5] the Type I error rate calculation can be further complicated due to the change of information time scale after the change in total sample size.

Method for Projecting New Sample Size

A widely used method for sample size re-estimation based on observed treatment difference is the conditional power approach.^[6] To describe this approach, consider a two-arm clinical trial with one interim analysis performed at the interim time t . For testing whether or not the parameter of interest, say Δ , is zero, let T_1 and T_2 be the test statistics in the interim and the final analyses, respectively. Let Δ_1 be the interim estimate of Δ . The estimate Δ_1 is a function of T_1 . Let $CP(\Delta) = \Pr(T_2 > C_2 | T_1, \Delta)$, where C_2 is the critical value for the final analysis. The conditional power $CP(\Delta_1)$ is obtained by replacing the unknown Δ with its interim estimate Δ_1 in the formulation of the conditional probability $CP(\Delta)$. With the conditional power, a decision as to whether or not the sample size needs to be modified can be made. For example, the total sample size can be increased if the conditional power is less than a certain percent. Once this decision is made, the new total sample size per group, denoted by M , can be obtained by solving the equation

$$CP(\Delta_1) = 1 - \beta \quad (2)$$

where $1 - \beta$ is the level of conditional power to achieve. Note that M reduces to the original total sample size if there is no sample size change after the interim analysis.

Many factors such as magnitude of the observed treatment difference, policy for sample size re-estimation, and modifications of test statistic or critical value for controlling Type I error rate after sample size change may affect the value of the conditional power, consequently the projected new total sample size.

To illustrate the use of conditional power, consider the method suggested by Cui et al.^[3] With this method, the

final test statistic after the interim sample size change is modified to

$$U = W_1 T_1 + W_2 T^* \quad (3)$$

where T_1 and T^* are the standardized original test statistics constructed using the data collected before and after the interim analyses, respectively, and W_1 and W_2 are fixed weights. The corresponding conditional power can be written as

$$CP(\Delta_1) = \Pr[T^* > (C_2 - W_1 T_1)/W_2 | \Delta_1, T_1] \quad (4)$$

For normal mean test, T_1 and T^* are two independent standardized normal mean tests. For many often used large sample tests, T_1 and T^* are mutually independent and asymptotically normal with variance of 1.^[3] The conditional power (Eq. 4) can be easily calculated by treating T^* as the test statistic for the final analysis, Δ_1 as the true treatment difference, and $C^* = (C_2 - W_1 T_1)/W_2$ as the critical value for the final analysis. For example, for two sample normal mean test, the close form expression of the conditional power (Eq. 4) can be written as

$$CP(\Delta_1) = 1 - \Phi[C^* - T_1(M/n - 1)^{1/2}]$$

where T_1 is the Z-test statistic in the interim analysis, n is the interim sample size, and $\Phi(\cdot)$ is the standard normal distribution function. For two-sample proportion test, the conditional power has the same expression but T_1 is the standardized two-sample proportion test in the interim analysis. Similarly, conditional power can be calculated for logrank test using Eq. 4 in which T^* is the standardized difference of the values of Brownian motion process derived from the logrank test based on the subjects available before and after the interim analysis.^[3] Eqs. 4 and 2 imply that the additional number of subjects after the interim analysis can be projected for a specific level of the conditional power by treating T^* as the test statistic for the final analysis, Δ_1 as the true treatment difference, and $C^* = (C_2 - W_1 T_1)/W_2$ as the critical value for the analysis. It should be pointed out that in practice, the total sample size can never be increased unlimitedly, and therefore an upper bound is always needed. This upper bound is usually set up based on the knowledge on possible magnitude of the treatment difference and some practical considerations such as those mentioned earlier.

Efficiency

The apparent advantage associated with a sample size adaptive design is the flexibility in changing maximum information. As a trade off, the design may lose some efficiency as compared to a fixed sample size design. The amount of loss of efficiency depends on several factors, such as the policy for changing sample size, the closeness

of the initial sample size to the true sample size needed, and others. To determine which sample size strategy to use, the fixed or adaptive, it is advisable to compare the candidate strategies in terms of efficiency first through, say, computer simulation. The decision then can be made based on the result of comparison and the assessment of the worth of the extra flexibility at the price of some efficiency.

Operational Issue

Because sample size re-estimation based on observed treatment difference requires unblinding patient treatment assignment codes, it subjects to the usual concerns for unblinded interim analysis and proper operational measures need to be taken to prevent possible occurrence of operational biases. In general, the trial sponsor should not have access to the information revealed in the analysis for sample size re-estimation. The sample size re-estimation is ideally performed by an independent party, e.g., an independent data safety monitoring board, rather than the sponsor. After the analysis for sample size re-estimation, only the new sample size is conveyed to the sponsor without disclosing the details of the interim analysis outcomes.

In general, the study protocol needs to include justification for performing sample size re-estimation using unblinded data. There are, however, clinical trials for which interim unblinding data is necessary due to reasons other than re-estimating sample size. For example, in most of cancer trials, complete blinding treatment assignments is almost impossible because of, for instance, the different drug delivery methods associated with the testing treatment and control treatment. For long-term mortality trials, because of ethic concerns, it is necessary to monitor the safety and efficacy of study drugs in an unblinded fashion. Because for such trials unblinding treatment codes is inevitable, it is natural for one to use the unblinded treatment information to adjust the total sample size if needed.

For a trial with sample size re-estimation based on unblinded data, in order to maintain the integrity of the trial, a high standard for trial conduct must be maintained. If the trial is confirmatory, the timing, the method for the sample size re-estimation, the method for later data analysis, and the standard operational procedure related to the sample size re-estimation need to be provided in the protocol and complied with.

STATISTICAL TEST BASED ON MODIFIED SAMPLE SIZE

Over the years, statistical methods have been developed for sample size re-estimation based on interim estimate of the variation of the response variable (see, e.g., Refs. [7–11]). Recently, several methods have been developed to allow sample size re-estimation based on interim treatment difference. The main objective of these new methods is

to utilize the updated knowledge on the treatment effect during a clinical trial to calibrate the total sample size to achieve the desired level of statistical power and to maintain the Type I error rate at the target level. The proposed procedures can be categorized into several groups by the way of approaching the problem.

Conditional Type I Error Function

Proschan and Hunsberger^[2] proposed a two-stage trial design. With this approach, the treatment difference is observed and tested at the end of the first stage of the trial. If the null hypothesis is not rejected at the time, the trial is extended to the second stage with more patients recruited and the hypothesis is then tested at the end of the second stage. The number of the patients needed for the second stage of the trial is calculated based on the conditional power approach using the observed treatment difference at the first stage of the trial. A key to their approach is the conditional Type I error function that regulates the critical value used for the second-stage testing and the Type I error rate associated with the procedure.

Liu and Chi^[12] later modified Proschan and Hunsberger's procedure by introducing a more general, conditional Type I error function. The total sample size, the critical values for rejection and acceptance of the null hypothesis for the first stage and the second stage of the trial can be chosen so that a set of design constraints including those on Type I and Type II error rates for the individual stages and for the overall are satisfied.

Combining Normal Scores Through Weighting

Cui et al.,^[3] as mentioned before, proposed a flexible statistical test allowing sample size adaptation for group sequential designs. This adaptive test procedure is also readily for use in the multistage design setting.^[13] With this procedure, the test statistic after the sample size modification is a weighted summation of the original normal score statistics for the data collected before and after the sample size modification, respectively. This test procedure increases the statistical power when the sample size increases and always maintains the Type I error rate at the target level regardless of how the new sample size is projected. This adaptive strategy is easy to use and readily applicable to many common large-sample tests, e.g., logrank test for survival analysis, which approximately follows a Brownian motion process.^[3] Other methods of combining normal scores for the grouped data are also available in the literature. For instance, Shen and Fisher^[14] proposed a different weighting scheme based on the concepts of variance-spending and self-designing clinical trial.^[15] The total sample size and weights are adjusted as the trial progresses based on the data available at the time of adjustment. Their test statistic is asymptotically normal when the random weights depend only on the sample path

of the test statistic and satisfy the condition that the sum of the squared weights at the time of the termination of the trial is unity. This leads to easy calculation of p value and a controlled Type I error rate but no interim testing is allowed by the procedure.

Combining p Values

Based on Fisher's combination test, Bauer and Köhne^[16] proposed a two-stage procedure. Let p_1 and p_2 be the p values of the test based on the subsamples at the first stage and the second stage of the trial, and let α_1 and α be the significance levels for the first stage and for the overall, respectively. With their approach, after extending the trial to the second stage, following the first stage at which the null hypothesis is not rejected, the null hypothesis is tested again by comparing the product $p_1 p_2$ with $\exp[-0.5\chi^2_{4,1-\alpha}]$, where $\chi^2_{4,1-\alpha}$ is the $(1-\alpha)$ -quantile of the χ^2 distribution with 4 degrees of freedom. With a proper choice of α_1 , a desired overall total Type I error rate can be maintained. Bauer and Köhne's method was modified by several investigators. For example, Lehman and Wassmer^[17] proposed to use an inversed normal score to combine p values from different stages of the trial.

Penalizing p Value

For the methods mentioned above, criteria can be implemented to allow termination of the trial early for futility. Early termination of a trial for futility has been used to avoid further wasting of resources as well as to reduce Type I error rate to maintain it at a desired level.^[4,18]

When a sample size adaptation strategy is used, the control of Type I error rate is achieved, e.g., by either modifying the critical value or imposing a p value penalty (see, e.g., Ref. [2]), or by modifying the test statistics (see, e.g., Refs. [3,14]). In the latter case, the test may be more flexible but less efficient than a fixed sample size approach when the initially projected sample size is accurate.

CONCLUSION

One major challenge nowadays to clinical trials stems from uncertainties about important trial design parameters, e.g., the true treatment difference, as a result of rapid advancement in discovery of new drugs or new indications for existing drugs. A flexible design allowing sample size modification based on the observed interim treatment difference appears to be a reasonable solution to this problem. With a sample size adaptation approach, a portion of the trial is used as an internal pilot part^[1] of the trial, and its outcome is used to reassure or modify initially projected total sample size.

Sample size adjustment based on the observed treatment difference is a part of the broader context in adaptation of clinical trial design. The way to use such a strategy starts with a careful sample size plan based on the best available information about the treatment difference and variation of the response variable and all other important trial considerations. A sample size adaptation design, on the top of a carefully planned initial sample size, provides an ability for further fine tuning of the total sample size based on updated treatment information and an extra assurance on the total sample needed. These two benefits are particularly important for large-scale clinical trials, involving significant complexity and investment. At present, research in sample size re-estimation is continually active, focusing on new methodologies, new applications, and the properties of the existing methods.^[19–23]

REFERENCES

1. Wittes, J.; Britain, E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Stat. Med.* **1990**, *9*, 65–72.
2. Proschan, M.; Hunsberger, S. Design extension of studies based on conditional power. *Biometrics* **1995**, *51*, 1315–1324.
3. Cui, L.; Hung, H.M.J.; Wang, S. Modification of sample size in group sequential clinical trials. *Biometrics* **1999**, *55*, 853–857.
4. Shun, Z.; Yang, W.; Brady, W.E.; Hsu, H. Type I error in sample size re-estimation based on observed treatment difference. *Stat. Med.* **2001**, *20*, 497–513.
5. Lan, K.K.G.; DeMets, D. Discrete sequential boundaries for clinical trials. *Biometrika* **1983**, *70*, 659–663.
6. Lan, K.K.; Simon, R.; Halperin, M. Stochastically curtailed tests in long-term clinical trials. *Seq. Anal.* **1982**, *1*, 207–219.
7. Stein, C. A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Stat.* **1945**, *16*, 243–258.
8. Gould, A.L. Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. *Stat. Med.* **1992**, *11*, 55–66.
9. Gould, A.L.; Shi, W.J. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Commun. Stat., Part A* **1992**, *21*, 2833–2853.
10. Gould, A.L.; Shi, W.J. Modifying the design of ongoing trials without unblinding. *Stat. Med.* **1998**, *17*, 89–100.
11. Denne, J.; Jennison, C. Estimating the sample size for a t -test using an internal pilot. *Stat. Med.* **1999**, *18*, 1575–1585.
12. Liu, Q.; Chi, G.Y.H. On sample size and inference for two-stage adaptive designs. *Biometrics* **2001**, *57*, 172–177.
13. Cui, L.; Hung, H.M.J. Adaptive testing with multiple sample size adjustment. *ASA Proc. Biopharm. Sect.* **1999**, 101–105.

14. Shen, Y.; Fisher, L.D. Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* **1999**, *55*, 190–197.
15. Fisher, L.D. Self-designing clinical trials. *Stat. Med.* **1998**, *17*, 1551–1562.
16. Bauer, P.; Köhne, K. Evaluation of experiments with adaptive interim analyses. *Biometrics* **1994**, *50*, 1029–1041.
17. Lehman, W.; Wassmer, G. Adaptive sample size calculations in group sequential trials. *Biometrics* **1999**, *55*, 1286–1290.
18. Lan, K.K.G.; Trost, D.C. Estimation of parameters and sample size re-estimation. *ASA Proc. Biopharm. Sect.* **1997**.
19. Posch, M.; Bauer, P. Adaptive two stage designs and the conditional error function. *Biom. J.* **1999**, *41*, 689–696.
20. Wassmer, G. A comparison of two methods for adaptive interim analyses in clinical trials. *Biometrics* **1998**, *54*, 696–705.
21. Wassmer, G. Multistage adaptive test procedures based on Fisher's product criterion. *Biom. J.* **1999**, *41*, 279–293.
22. Wang, S.; Hung, H.M.J.; Tsong, Y.; Cui, L. Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. *Stat. Med.* **2001**, *18*, 1903–1912.
23. Müller, H.; Schäfer, H. Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential procedures. *Biometrics* **2001**, *57*, 886–891.

Sample Size Re-estimation Design with Robust Power

Lu Cui

Data and Statistical Science, AbbVie Inc., North Chicago, Illinois, U.S.A.

Abstract

Flexible sample size design has been used as an alternative to fixed sample size design to achieve robust power and better efficiency of a clinical study. In this entry, the objective and performance evaluation of flexible sample size designs are discussed. Group sequential design and sample size re-estimation design—two kinds of popular flexible sample size designs—are introduced and compared. The implementation of flexible sample size design with design optimization is illustrated with examples.

Keywords: Sample size; Power; Group sequential design; Sample size re-estimation.

INTRODUCTION

Flexible sample size design has been used in clinical trials as an alternative to traditional fixed sample size design to achieve robust power and trial efficiency. Two kinds of flexible sample size designs often used are group sequential (GS) design and sample size re-estimation (RE) design. Both designs allow adjustment of the final sample size of a trial, N^* , based on unblinded interim treatment information. The key difference is that the former has fixed maximum sample size with a few interim time points for potential early stopping, but the latter allows varying final stopping time or final sample size after interim analysis.

This entry focuses on concepts and issues related to flexible sample size design. With the introduced objective of flexible sample size design and the corresponding performance evaluation criteria, GS and RE designs are compared. Implementation with design optimization is illustrated with examples.

LITERATURE REVIEW

The idea of statistical inference based on trial with flexible sample size can be traced back to Stein's two-stage procedure proposed in the 1940s.^[1] The interest surged due to increasing challenges in modern clinical trials in the early 1990s. To use interim data to improve clinical trial quality, Wittes and Brittain proposed the idea of internal piloting in clinical studies.^[2] The concept triggered waves of research in adaptive clinical trial designs, including flexible sample size design based on unblinded interim data. A method to handle the obtained stage-wise data via combining p values was given by Bauer and Köhne.^[3] Proschan and Hunsberger proposed a two-stage sample size RE design^[4] based on unblinded interim treatment difference with recalculated

final critical value to control type I error rate. Cui, Hung, and Wang^[5,6] and Lehman and Wassmer^[7] proposed methods to combine unblinded sample size RE with GS design via a re-weighting scheme. Different opinions advocating the use of GS approach rather than sample size RE design were given by Tsiatis and Mehta^[8] and Jennison and Turnbull^[9,10] based on point-wise efficiency consideration. Liu, Zhu, and Cui^[11] studied the design problem from different perspectives aiming at a robust design performance. Unlike Tsiatis and Mehta's work looking for a point-wise optimal flexible sample size design, Liu, Zhu, and Cui focused on optimal flexible sample size design with robust power for a range of the underlying parameter of interest. Wu and Cui showed^[12] that an RE design optimally chosen under the definition of Liu, Zhu, and Cui generally outperformed GS design. In fact, the optimal RE design can be obtained through global search of the best combination of the timing of the interim analysis, sample size of interim analysis, and weight used in the combination test.^[13] Among other works, an approach based on the concept of promising zone (PZ) became popular.^[14,15] A more general adaptive design method was proposed by Müller and Schäfer.^[16] By re-calculating critical values to control conditional type I error rate, the method allows modification of sample size as well as other trial parameters. Gao, Ware, and Mehta^[17] established the equivalence relationship among the methods by Proschan and Hunsberger; Cui, Hung, and Wang; and Müller and Schäfer. Because of its potential to improve quality of industry-sponsored clinical trials, flexible sample size design is of significant interest of drug developers and regulatory agencies. General regulatory guidances were issued by the U.S. Food and Drug Administration^[18] and the European Medicine Agency^[19] on adaptive clinical trials including flexible sample size design.

This short entry is to introduce flexible sample size design via sample size RE design and to discuss the related

issues. For this purpose, the design proposed by Cui, Hung, and Wang^[5,6] is particularly chosen to represent RE design because of its computational simplicity and maximum flexibility. Comparison of RE design with the GS design is provided to illustrate ideals. More references related to flexible sample size design can be found in literatures.^[20–29]

OBJECTIVE OF FLEXIBLE SAMPLE SIZE DESIGN

First consider a case of fixed sample size design under two-arm clinical trial setting. Assume that the outcomes of the primary efficacy measurement, X for placebo and Y for the new treatment, follow normal distributions with mean μ_0 and μ_1 , respectively, and with a common variance $0.5\sigma^2$. The experiment is to demonstrate that the new treatment is better than placebo based on testing the one-sided hypothesis,

$$H_0 : \delta = 0 \text{ vs. } H_1 : \delta > 0 \quad (1)$$

where $\delta = \mu_1 - \mu_0$ is the treatment difference. For fixed sample size design, the number of subjects per group needed to detect the treatment difference δ with power $1 - \beta$ at the significance level α is calculated as:

$$N = N(\delta) = \frac{(z_\alpha + Z_\beta)^2 \sigma^2}{\delta^2} \quad (2)$$

In the above, z_α is the upper α -quantile of the standard normal distribution. For simplicity, we further assume that $\sigma = 1$ is known. In the practice, σ is often readily estimated based on historical data or blinded interim data.

Since δ is unknown, to calculate the sample size, δ in Eq. 2 is replaced by a targeted treatment difference δ_0 . Determination of δ_0 can be a difficult task since it should be in line with the unknown true treatment difference δ essentially. Otherwise the projected sample size N based on δ_0 will be off the target leading to an either underpowered or overpowered study.

Flexible sample size design is to address the aforementioned difficulty. Unlike fixed sample size design, without an upfront accurate projection of the true treatment difference δ , flexible sample size design is to target the unknown treatment difference in a likely range, say, $[\delta_L, \delta_U]$. While fixed sample size design shoots a statistical power of $1 - \beta$ at a point value of δ_0 , the objective of flexible sample size design is to ensure a desirable study power of $1 - \beta$ for all true $\delta \in [\delta_L, \delta_U]$ through interim sample size assessment.

GROUP SEQUENTIAL AND SAMPLE SIZE RE-ESTIMATION DESIGNS

GS and RE approaches to flexible sample size design are briefly described below. Suppose that hypothesis (Eq. 1) is

to be repeatedly tested in a total of $K - 1$ interim analyses and in the final analysis. With a given α -spending function and the corresponding critical value C_k , $k = 1, \dots, K$, the test statistic of the usual (likelihood ratio) GS test in the k^{th} analysis is:

$$S_k = \frac{1}{\sqrt{n_k}} \sum_{i=1}^{n_k} (Y_i - X_i), \quad k = 1, \dots, K \quad (3)$$

where n_k is the sample size per group in the k^{th} interim analysis and $n_K = N$. The null hypothesis H_0 is rejected if $S_k > C_k$, and otherwise the trial proceeds to the next analysis and the sample size is incrementally increased. At the final analysis, H_0 is rejected if $S_K > C_K$ or the trial is inconclusive. Since the trial can be terminated early, the actual or the final sample size of the trial N^* can be smaller than the planned sample size or $N^* \leq N$.

For illustration, consider the RE design proposed by Cui, Hung, and Wang^[5,6] which is built on a K -analysis GS design with test statistics S_k and maximum sample size N . The hypothesis (Eq. 1) is then tested with test statistics $U_k = S_k$ as before for $k \leq l$. Assume that in the l^{th} interim analysis, the null hypothesis is not rejected, and subsequently the final sample size is re-estimated as $N^* = M$ per group. Let $b = (M - n_l) / (N - n_l)$, where $b < 1$ if $M < N$, $b > 1$ if $M > N$, or $b = 1$ if $M = N$. The remaining interim analyses after the l^{th} interim analysis can be performed with, say, m_{l+j} subjects per group, where $m_{l+j} = b(n_{l+j} - n_l) + n_l$, $j = 1, \dots, K - l$ and $m_K = M$. However, to control type I error rate, test statistic U_k is modified as:

$$U_k = S_l \sqrt{\frac{n_l}{n_k}} + \frac{\sum_{i=n_l+1}^{m_k} (Y_i - X_i)}{\sqrt{m_k - n_l}} \sqrt{\frac{n_k - n_l}{n_k}} \quad (4)$$

$k = l + 1, \dots, l + j, \dots, K$, and the hypothesis (Eq. 1) is further tested using U_k against C_k , $k = l + 1, \dots, l + j, \dots, K$. In general, the final sample size $N^* = M$ can be obtained via simulations using conditional power approach based on the interim data.

Test statistic U , proposed by Cui, Hung, and Wang, has been termed as CHW test or weighted RE method. Like GS method, it has controlled type I error rate. If the initial sample size remains unchanged, or $N^* = N$, the test statistic U_k reduces to the GS test S_k . This means that all GS tests form a subset of RE tests. The importance of this fact is that RE design in theory dominates GS design in term of achieving optimality.^[10]

Zhang, Cui, and Yang^[13] pointed out that the weight in Eq. 4 can be set as arbitrary fixed numbers satisfying $0 < w_1 < \dots < w_{K-1} < w_K = 1$. The test statistic U in Eq. 4, after the interim sample assessment, then becomes:

$$U_k = S_l \sqrt{\frac{w_l}{w_k}} + \frac{\sum_{i=n_l+1}^{m_k} (Y_i - X_i)}{\sqrt{w_k - w_l}} \sqrt{\frac{w_k - w_l}{w_k}} \quad (5)$$

$k=1+1,\dots,1+j,\dots,K$. The statistical testing at interim analysis can be conducted based on U_k as before with a given α -spending function evaluated at w_k as if w_k was information fraction. In the analysis, properly chosen n_k ($k \leq 1$) or m_k ($1 < k \leq K$) subjects per group are to be involved.

MEASURE OF ADAPTIVE PERFORMANCE

Recall the sample size calculation for the fixed sample size design. If the true δ is known, the sample size of the trial $N = N(\delta)$ is calculated at δ according to Eq. 2. The sample size $N(\delta)$ can be viewed as an ideal or a gold standard sample size since it is the perfect sample size that one may get to meet the power and type I error rate specifications. The notation $N(\delta)$ will particularly denote this gold standard sample size as a function of δ .

The above method motivates the performance measure of flexible sample size design based on the deviation of its final trial sample size N^* from the gold standard sample size $N(\delta)$. Such a deviation at each δ can be measured via sample size ratio (SR)^[11] or sample size difference (SD)^[12] defined as the following:

$$SR = \frac{N^*}{N(\delta)} - 1 \quad (6)$$

and

$$SD = N^* - N(\delta) \quad (7)$$

The average performance score (APS) then can be defined as the average of SR (APS_r) or SD (APS_d) over the range of δ :

$$APS_r = \int_{\delta_L}^{\delta_U} \left| \frac{E_{\delta}(N^*)}{N_{\delta}} - 1 \right| f(\delta) d\delta \quad (8)$$

$$APS_d = \int_{\delta_L}^{\delta_U} |E_{\delta}(N^*) - N_{\delta}| f(\delta) d\delta \quad (9)$$

In the above equations, $f(\delta) > 0$ is a tuning function. It can be chosen as a constant over $[\delta_L, \delta_U]$, if there is no detailed information about δ . Otherwise, $f(\delta)$, for example, can be chosen as a truncated normal density if one believes that likely values of δ are centered around a particular point in the interval.

The definition of APS allows to summarize the adaptive performance of a flexible sample size design over a range of δ by a single numerical value. Since APS penalizes both oversizing and undersizing the trial, a small value of APS generally means a smaller deviation from the true sample size required and consequently a smaller deviation from

the targeted power. Therefore, minimizing APS is desirable. This leads to the following definition of optimal flexible sample size design by Liu, Zhu, and Cui.^[11]

Definition

An optimal flexible sample size design is one for which the associated APS is minimized as compared to all flexible sample size designs.

This definition generally leads to a flexible sample size design with a close-to-target final sample size, and thus a robust power over $[\delta_L, \delta_U]$, achieving the objective of flexible sample size design. Since practically only a finite number of different sample size designs are considered at the design stage, the best flexible sample size design with the minimum APS value can always be found by comparing calculated APS values of the candidate designs.

PERFORMANCE COMPARISON AND DESIGN OPTIMIZATION

The APS provides an easy way to compare flexible sample size designs, including GS, RE, and PZ designs. The following examples show how to implement flexible sample size design with design optimization.

Example 1

Since fixed sample design can be viewed as a special case of flexible sample size design, the definition of APS optimality applies to fixed sample case. It can be shown that the optimal fixed size design among all fixed sample designs is the one with the sample size N calculated to target at the center of the interval $[\delta_L, \delta_U]$ or $\delta_0 = (\delta_L + \delta_U)/2$. In other words, when only knowledge available at the trial design stage is the range of δ , the best or a balanced fixed sample size strategy is to target at the midpoint of the range, which agrees with intuition.^[12]

Example 2

Liu, Zhu, and Cui^[11] considered the hypothesis testing problem (Eq. 1) and a set of candidate flexible sample size designs. The set consisted of four-look GS designs with O'Brien-Flemming (GSO) or Pocock (GSP) α -spending functions and with initial sample size $N = 40, 80, \dots, 1,000$ and four-look RE designs with the same spending functions (REO and REP designs) and initial sample sizes. There were 100 candidate flexible designs (25 of each type). In the example, the interim analyses were equally spaced. For GS design, an initial sample size was also the maximum sample size. For RE design, sample size increase was allowed after the sample size re-estimation in the second interim analysis with 50% of the initial sample size but the maximum sample size was capped at 1,000

subjects per treatment group for the sake of a fair comparison. The true treatment difference δ was assumed in the range $[\delta_L, \delta_U] = [0.1, 0.5]$. The objective was to find the best design among the candidate designs with a robust power over the range.

The performance score APS_r was calculated based on simulated data for all designs. The values of APS_r for the best GSO, GSP, as well as REO and REP designs were obtained as 75.4, 55.5, 50.3, and 37.2, respectively. The corresponding initial sample sizes were 160, 280, 160, and 240, respectively. Therefore, by Liu's Definition, the best GS design was the GS design with Pocock α -spending function and the maximum sample size $N = 280/\text{group}$ ($APS_r = 55.5$) which outperformed the best GS design with O'Brien-Fleming α -spending function and $N = 160/\text{group}$ ($APS_r = 75.4$). Even with the conservative O'Brien-Fleming α -spending function, the RE design with the initial sample size $N = 160$ ($APS_r = 50.3$) outperformed the best GS design ($APS_r = 55.5$). The winner among all candidate designs was the RE design with Pocock α -spending function and the initial sample size $N = 240$. This design had the smallest APS_r value of 37.2. Additional simulations showed that this design maintained the targeted statistical power about 90% or higher for $\delta \in [0.14, 0.5]$ or no less than 80% of power for the entire range $[0.1, 0.5]$. The design apparently achieved its objective with robust power and efficient average sample size for nearly all $\delta \in [\delta_L, \delta_U] = [0.1, 0.5]$.

Example 3

Zhang, Cui, and Yang^[13] studied the general CHW RE design with freely chosen weight w . Under a similar setting of example 2 with a four-look RE design and Pocock α -spending function, fixing $1 = 2$, the optimal RE design based on APS criterion was identified with $n_1 = 80$ and $w = (0.05, 0.1, 0.35, 1.0)$. The design was obtained through global search based on APS among possible choices of n_1 and w . The simulations in this study showed that even for the GS design, the APS optimization could help to find out robust GS design. Simulations also showed that RE design in general outperformed GS and PZ designs, while the latter two appeared to have comparable performance.

CONCLUSION

The main objective of flexible sample size design is to achieve a robust high statistical power of a clinical study. Based on the two commonly used flexible sample size designs, namely, GS and RE designs, the design performance evaluation and design optimization are illustrated. The examples show that an optimally chosen flexible design can achieve a robust statistical power over a wide range of likely values of the targeted parameter.

ACKNOWLEDGMENT

This work was supported by AbbVie. AbbVie participated in the interpretation of data, writing, review, and approval of the content of this work. Lu Cui is an employee of AbbVie.

REFERENCES

1. Stein, C. A two sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Stat.* **1945**, 16 (3), 243–258.
2. Wittes, J.; Brittain, E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Stat. Med.* **1990**, 9 (1–2), 65–72.
3. Bauer, P.; Köhne, K. Evaluation of experiments with adaptive interim analyses. *Biometrics* **1994**, 50 (4), 1029–1041.
4. Proschan, M.A.; Hunsberger, S.A. Designed extension of studies based on conditional power. *Biometrics* **1995**, 51 (4), 1315–1324.
5. Cui, L.; Hung, H.M.; Wang, S.J. Impact of changing sample size in a group sequential clinical trial. *ASA Proc. Biopharm. Sect.* **1997**, 52–57.
6. Cui, L.; Hung, H.M.; Wang, S.J. Modification of sample size in group sequential clinical trials. *Biometrics* **1999**, 55 (3) 853–857.
7. Lehmacher, W.; Wassmer, G. Adaptive sample size calculations in group sequential trials. *Biometrics* **1999**, 55 (4), 1286–1290.
8. Tsiatis, A.A.; Mehta, C. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* **2003**, 90 (2), 367–378.
9. Jennison, C.; Turnbull, B.W. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Stat. Med.* **2003**, 22 (6), 971–993.
10. Jennison, C.; Turnbull, B. Efficient group sequential designs when there are several effect sizes under consideration. *Stat. Med.* **2006**, 25 (6), 917–932.
11. Liu, G.F.; Zhu, G.R.; Cui, L. Evaluating the adaptive performance of flexible sample size designs with treatment difference in an interval. *Stat. Med.* **2008**, 27 (4), 584–596.
12. Wu, X.; Cui, L. Group sequential and discretized sample size re-estimation designs: A comparison of flexibility. *Stat. Med.* **2012**, 31 (24), 2844–2857.
13. Zhang, L.; Cui, L.; Yang, B. Optimal flexible sample size design with robust power. *Stat. Med.* **2016**, 35 (19), 3385–3396.
14. Chen, Y.H.; DeMets, D.L.; Lan, K.K. Increasing the sample size when the unblinded interim result is promising. *Stat. Med.* **2004**, 23 (7), 1023–1038.
15. Mehta, C.R.; Pocock, S.J. Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Stat. Med.* **2011**, 30 (28), 3267–3284.
16. Müller, H.H.; Schäfer, H. Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential procedures. *Biometrics* **2001**, 57 (3), 886–891.

17. Gao, P.; Ware, J.; Mehta, C. Sample size re-estimation for adaptive sequential design in clinical trials. *J. Biopharm. Stat.* **2008**, *18* (6), 1184–1196.
18. U.S. Food and Drug Administration. *Draft Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics*; U.S. Food and Drug Administration: Washington, DC, 2010.
19. European Medicine Agency. *Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with An Adaptive Design*; European Medicine Agency: London, 2007.
20. Shun, Z.; Yuan, W.; Brady, W.E.; Hsu, H. Type I error in sample size re-estimation based on observed treatment difference. *Stat. Med.* **2001**, *20* (4), 497–513.
21. Friede, T.; Kieser, M. A comparison of methods for adaptive sample size adjustment. *Stat. Med.* **2001**, *20* (24), 3861–3873.
22. Denne, J.S. Sample size re-calculation using conditional power. *Stat. Med.* **2001**, *20* (17–18), 2645–2660.
23. Desseaux, K.; Porcher, R. Flexible two-stage design with sample size reassessment for survival trials. *Stat. Med.* **2007**, *26* (27), 5002–5013.
24. Chow, S.C.; Chang, M. *Adaptive Design Methods in Clinical Trials*. 2nd Ed.; CRC Press: New York, 2011.
25. Pong, A.; Chow, S.C. *Handbook of Adaptive Designs in Pharmaceutical and Clinical Development*; CRC Press: New York, 2010.
26. Wang, S.J.; Hung, H.M.; O'Neil, R.T. Impacts on type I error rate with inappropriate use of learn and confirm in confirmatory adaptive design trials. *Biometrical J.* **2010**, *52* (6), 798–810.
27. Berry, S.; Carlin, B.; Lee, J.; Muller, P. *Bayesian Adaptive Methods for Clinical Trials*; CRC Press: New York, 2010.
28. Hung, H.M.; Wang, S.J.; Yang, P. Some challenges with statistical inference in adaptive designs. *J. Biopharm. Stat.* **2014**, *24* (5), 1059–1072.
29. Bauer, P.; Bretz, F.; Dragalin, V.; Konig, F.; Wassmer, G. Twenty-five years of confirmatory adaptive designs: Opportunities and pitfalls. *Stat. Med.* **2016**, *35* (3), 325–347.

Sample Size: Negative Binomial With Unequal Follow-Up Times

Yongqiang Tang

Shire, Lexington, Massachusetts, U.S.A.

Abstract

Recurrent events are common in clinical trials. Examples include exacerbations in chronic obstructive pulmonary disease, relapses in multiple sclerosis, seizures in epileptics, and hospitalizations. In late-phase confirmatory trials, the primary endpoint is generally the total number of events experienced by the subjects. The recurrent event counts often exhibit overdispersion, in the sense, that the variance exceeds the mean. Negative binomial (NB) regression has become widely used to analyze recurrent event data in clinical trials since it provides a natural way to account for overdispersion. Several previously developed methods for calculating sample size for the NB regression assume equal follow-up time for all subjects by ignoring early dropouts or by setting the follow-up time for all individuals to be the mean follow-up time. Ignoring variability in the duration of the follow-up generally leads to underpowered studies. In this entry, we review the sample size calculation methods for comparing two NB rates in superiority, non-inferiority, and equivalence trials with unequal follow-up times.

Keywords: Equivalence trial; Fixed margin approach; Negative binomial distribution; Non-inferiority margin; Non-inferiority trial; Overdispersion; Power and sample size; Superiority trial.

INTRODUCTION

Recurrent events are common in clinical trials. Examples include exacerbations in chronic obstructive pulmonary disease, relapses in multiple sclerosis, seizures in epileptics, and hospitalizations. In late-phase confirmatory trials, the primary endpoint is generally the total number of events experienced by the subjects. The recurrent event counts often exhibit overdispersion, in the sense, that the variance exceeds the mean.

Negative binomial (NB) regression has become widely used to analyze recurrent event data in clinical trials since it provides a natural way to account for overdispersion. Several previously developed methods for calculating sample size for the NB regression assume equal follow-up time for all subjects by ignoring early dropouts or by setting the follow-up time for all individuals to be the mean follow-up time. Ignoring variability in the duration of the follow-up generally leads to underpowered studies. In this entry, we review the sample size calculation methods for comparing two NB rates in superiority, non-inferiority (NI), and equivalence trials with unequal follow-up times.

NB DISTRIBUTION

The NB distribution [denoted by $Y \sim \text{NB}(\mu, \kappa)$] is the probability distribution of the number of failures Y before

$k = 1/\kappa$ successes in a series of independent Bernoulli trials with the same probability $p = 1/(1 + \kappa \mu)$ of success.

$$\Pr(Y = y) = \frac{\Gamma(y + 1/\kappa)}{y! \Gamma(1/\kappa)} p^{1/\kappa} (1 - p)^y \quad (1)$$

where μ is the mean and κ is the dispersion parameter. For count data, the NB distribution is often interpreted as a gamma mixture of Poisson distribution. That is, if Y follows a Poisson distribution with mean $\varepsilon\mu$ and ε is gamma distributed with mean 1 and variance κ , the marginal distribution of Y is $\text{NB}(\mu, \kappa)$. This representation does not require $k = 1/\kappa$ to be an integer. The random effect ε captures the between-subject heterogeneity in event rates, and κ measures the degree of heterogeneity. Including important risk factors in the model may reduce heterogeneity. The NB distribution tends to fit the overdispersed count data better than the Poisson distribution,^[1,2] and its mean is always less than its variance.^[3]

$$\begin{aligned} E(Y) &= E[E(Y | \varepsilon)] = \mu, \\ \text{var}(Y) &= E[\text{var}(Y | \varepsilon)] + \text{var}[E(Y | \varepsilon)] = \mu + \kappa\mu^2 \end{aligned} \quad (2)$$

The NB distribution belongs to the family of mixed Poisson distributions. The gamma distribution is commonly used as the mixing distribution because the corresponding marginal distribution has a closed form, which is NB.^[3,4]

ANALYSIS OF COUNT DATA

Suppose in a clinical trial, n subjects are randomized to the experimental ($g = 1$) or control ($g = 0$) treatment. Let t_{gj} be the follow-up time and y_{gj} be the number of observed events for subject $j = 1, \dots, n_g$ in group g . We assume constant event rate over time, and $y_{gj} \sim \text{NB}(\mu_{gj}, \kappa)$, where $\lambda_g = \exp(\gamma_g)$ is the event rate for group g , and $\mu_{gj} = \lambda_g t_{gj}$. In SAS, the NB regression can be fitted using the code,

```
proc genmod data =xxx;
class trt;
model y= trt /offset=logt link=log dist=nb noLOGNB;
run;
```

The offset term $\log(t_{gj})$ accounts for different follow-up periods for different subjects. When the *noLogNB* option is employed, the dispersion parameter is estimated based on κ rather than $\log(\kappa)$. The real data may not exhibit overdispersion^[1] if nearly all subjects have no or only one event (e.g., due to mild disease condition or short follow-up time), and the estimation based on $\log(\kappa)$ may fail to converge because of the restriction $\kappa > 0$.

Throughout, we assume a lower event rate is desirable. The treatment effect is often measured by the rate ratio $\exp(\beta) = \lambda_1 / \lambda_0$ or equivalently by the relative rate reduction $\text{RRR} = 1 - \lambda_1 / \lambda_0 = 1 - \exp(\beta)$. Let parameters with $\hat{\cdot}$ denote the maximum likelihood estimate (MLE). By equations (2.7a), (2.7b), and (2.8) of Lawless,^[4] the Fisher information matrix for $(\gamma_0, \gamma_1, \kappa)$ is diagonal $\text{diag}(e_0, e_1, e_*)$, where $e_g = \sum_{j=1}^{n_g} \lambda_g t_{gj} / (1 + \kappa \lambda_g t_{gj})$ and e_* is irrelevant in the sample size calculation. Therefore, $\hat{\gamma}_0, \hat{\gamma}_1$, and $\hat{\kappa}$ are asymptotically independent, $\text{var}(\hat{\lambda}_g) = e_g^{-1}$, and the variance of $\hat{\beta} = \hat{\gamma}_1 - \hat{\gamma}_0$ is:

$$\text{var}(\hat{\beta}) = \text{var}(\hat{\gamma}_0) + \text{var}(\hat{\gamma}_1) = e_0^{-1} + e_1^{-1} \quad (3)$$

The variance of $\hat{\beta}$ can also be obtained from the observed information matrix. The two variance estimates are asymptotically equivalent. The two-sided 100(1 - α)% confidence interval (CI) for β is:

$$[c_l, c_u] = \left[\hat{\beta} - Z_{1-\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\beta})}, \hat{\beta} + Z_{1-\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\beta})} \right] \quad (4)$$

where $\widehat{\text{var}}(\hat{\beta})$ is the variance of $\hat{\beta}$ evaluated at the MLE, and z_p is the p^{th} percentile of the standard normal distribution. The CI for $\exp(\beta) = \lambda_1 / \lambda_0$ is $[\exp(c_l), \exp(c_u)]$.

In a superiority trial, the control treatment is placebo or an active treatment, and the objective is to demonstrate that the experimental treatment can lower the event rate. The hypothesis can be expressed as:

$$H_0 : \lambda_0 = \lambda_1 \text{ or } \beta = 0 \text{ vs } H_1 : \lambda_0 \neq \lambda_1 \text{ or } \beta \neq 0 \quad (5)$$

if $c_u < 0$ or equivalently $\exp(c_u) < 1$, one can claim that the experimental treatment can significantly reduce the event rate compared to the control treatment.

In a NI trial, the control is an active treatment whose effect is well defined, and the objective is to show that the experimental treatment is not worse than the active control by M_0 , where $M_0 > 1$ is a prespecified NI margin. The NI trials could be analyzed using a fixed margin approach or the synthesis method.^[5] We focus on the fixed margin approach since it is recommended by the regulatory guidelines.^[5,6] The margin M_0 is generally chosen to be close to 1 in order to demonstrate that the new treatment is not materially inferior to the active comparator (refer to the Guidelines by United States Food and Drug Administration and the European Medicines Agency's Committee for Medicinal Products for Human Use^[5,6] for more details on the specification and interpretation of the NI margin). In NI trials, the hypothesis can be written as:

$$H_0 : \frac{\lambda_1}{\lambda_0} \geq M_0 \text{ versus } H_1 : \frac{\lambda_1}{\lambda_0} < M_0 \quad (6)$$

or equivalently as:

$$H_0 : \beta \geq \log(M_0) \text{ versus } H_1 : \beta < \log(M_0) \quad (7)$$

If the upper limit of the CI for $\exp(\beta) = \lambda_1 / \lambda_0$ satisfies $\exp(c_u) < M_0$, one can claim that the experimental treatment is not significantly inferior to the active comparator.

In an equivalence trial, the objective is to demonstrate that the two treatments are clinically equivalent. If the CI for $\exp(\beta)$ lies completely within the interval $[M_l, M_u]$, we can claim clinical equivalence of the two treatments, where $M_l < 1$ and $M_u > 1$ are two prespecified margins. The hypothesis can be written as:

$$H_0 : \frac{\lambda_1}{\lambda_0} \geq M_u \text{ or } \frac{\lambda_1}{\lambda_0} \leq M_l \text{ versus } H_1 : M_l < \frac{\lambda_1}{\lambda_0} < M_u \quad (8)$$

or equivalently as:

$$H_0 : \beta \geq \log(M_u) \text{ or } \beta \leq \log(M_l) \text{ versus } H_1 : \log(M_l) < \beta < \log(M_u) \quad (9)$$

One may refer to Chow, Wang, and Shao^[7] for more detailed discussions on superiority, NI, and equivalence trials, where they also provide methods for sample size determination for different types of endpoints.

SAMPLE SIZE FOR SUPERIORITY TRIALS

In a superiority trial, the event rates are statistically different between the two treatment groups if the CI of $\hat{\beta}$ excludes 0. Both $Z = (\hat{\beta} - \beta) / \sqrt{\widehat{\text{var}}(\hat{\beta})}$ and $-Z$ asymptotically follow the standard normal distribution $N(0, 1)$. Then:

$$\begin{aligned} \Pr(c_u < 0) &= \Pr(\hat{\beta} + z_{1-\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\beta})} < 0) \\ &= \Pr\left(Z < \frac{-z_{1-\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\beta})} - \beta}{\sqrt{\widehat{\text{var}}(\hat{\beta})}}\right) \end{aligned}$$

$$\begin{aligned} &\approx \Pr\left(Z < -z_{1-\alpha/2} - \frac{\beta}{\sqrt{\text{var}(\beta)}}\right) \\ &\approx \Phi\left(\frac{-\sqrt{n}\beta}{\sqrt{(d_0p_0)^{-1} + (d_1p_1)^{-1}}} - z_{1-\alpha/2}\right) \end{aligned} \quad (10)$$

where p_g is the proportion of subjects randomized to treatment group g and $d_g = E[\lambda_g t_{g|j} / (1 + \kappa \lambda_g t_{g|j})]$. The derivation of Eq. 10 relies on the large sample approximations $\widehat{\text{var}}(\hat{\beta}) / \text{var}(\beta) \approx 1$, and $e_g \approx n_g d_g$. Similarly, we have:

$$\Pr(c_1 > 0) \approx \Phi\left(\frac{\sqrt{n}\beta}{\sqrt{(d_0p_0)^{-1} + (d_1p_1)^{-1}}} - z_{1-\alpha/2}\right) \quad (11)$$

if $c_1 > 0$, one would conclude that the control treatment is better than the experimental treatment, and this is undesired. But the probability $\Pr(c_1 > 0)$ is negligible compared to $\Pr(c_u < 0)$ if the experimental treatment is truly effective in lowering the event rate (i.e., $\lambda_1 < \lambda_0$, and λ_1 is not too close to λ_0 to be of practical interest). The power is:

$$\begin{aligned} P &= \Pr(c_1 > 0) + \Pr(c_u < 0) \approx \Pr(c_u < 0) \\ &\approx \Phi\left(\frac{\sqrt{n}|\beta|}{\sqrt{(d_0p_0)^{-1} + (d_1p_1)^{-1}}} - z_{1-\alpha/2}\right) \end{aligned} \quad (12)$$

Inverting Eq. 12 yields the required total sample size.^[3]

$$n_{\text{tar}} = [(d_0p_0)^{-1} + (d_1p_1)^{-1}] \frac{(z_{1-\alpha/2} + z_p)^2}{\beta^2} \quad (13)$$

Under equal allocation ($p_0 = p_1 = 1/2$), the total size is:

$$n_{\text{tar}} = 2(d_0^{-1} + d_1^{-1}) \frac{(z_{1-\alpha/2} + z_p)^2}{\beta^2} \quad (14)$$

Eqs. 12–14 are also valid if $\lambda_1 > \lambda_0$ (i.e., $\beta > 0$), and the objective is to demonstrate that the experimental treatment can increase the event rate.

The power and sample size formulae reviewed in this entry are derived based on the Wald test from the NB regression, and they generally work well in moderate to large samples. Their performance deteriorates when the sample size becomes small. For example, Eq. 12 may slightly underestimate the power of the Wald test in small samples^[3] possibly because the dispersion parameter κ is underestimated^[8] in the MLE procedure (this is analogous to the analysis of variance). Partially for the same reason, the Wald-type test may not be able to control the type I error at the nominal level in small samples, and more robust tests shall be used to analyze small trials.

SAMPLE SIZE FOR NI TRIALS

In a NI trial, the experimental treatment is not materially worse than the active control if the whole CI of β is below $\log(M_0)$. The power is:^[9]

$$\begin{aligned} \Pr(c_u < \log(M_0)) &= \Pr(\hat{\beta} + z_{1-\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\beta})} < \log(M_0)) \\ &= \Pr\left(Z < \frac{-z_{1-\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\beta})} - \beta + \log(M_0)}{\sqrt{\text{var}(\beta)}}\right) \\ &\approx \Phi\left(\frac{\sqrt{n}\beta^*}{\sqrt{(d_0p_0)^{-1} + (d_1p_1)^{-1}}} - z_{1-\alpha/2}\right) \end{aligned} \quad (15)$$

where $\beta^* = \log(M_0 \lambda_0 / \lambda_1)$. The sample size is:^[9]

$$n_{\text{tar}} = [(d_0p_0)^{-1} + (d_1p_1)^{-1}] \frac{(z_{1-\alpha/2} + z_p)^2}{\beta^{*2}} \quad (16)$$

Eq. 16 looks similar to Eq. 13 except that $\beta^2 = \log^2(\lambda_0 / \lambda_1)$ is replaced by $\beta^{*2} = \log^2(M_0 \lambda_0 / \lambda_1)$. However, it would be cautious to calculate the sample size for a NI trial using the sample size calculation software for superiority trials because the event rates could be implicitly modified as $\lambda_1 = \lambda_0 \exp(\beta^*)$ if the software requires only the values for λ_0 and $\exp(\beta^*)$.^[9]

SAMPLE SIZE FOR EQUIVALENCE TRIALS

In an equivalence trial, the two treatments are not clinically different if the whole CI of β falls completely within the interval $[\log(M_1), \log(M_u)]$. The power is:

$$\begin{aligned} P &= \Pr(c_u = \hat{\beta} + z_{1-\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\beta})} < \log(M_u) \text{ and } \\ &\quad \hat{\beta} - z_{1-\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\beta})} > \log(M_1)) \\ &= \Pr\left(\frac{z_{1-\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\beta})} - \beta + \log(M_1)}{\sqrt{\text{var}(\beta)}} < Z \right. \\ &\quad \left. < \frac{-z_{1-\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\beta})} - \beta + \log(M_u)}{\sqrt{\text{var}(\beta)}}\right) \\ &\approx \Phi\left(\frac{\sqrt{n} \log(M_u \lambda_0 / \lambda_1)}{\sqrt{(d_0p_0)^{-1} + (d_1p_1)^{-1}}} - z_{1-\alpha/2}\right) \\ &\quad + \Phi\left(\frac{-\sqrt{n} \log(M_1 \lambda_0 / \lambda_1)}{\sqrt{(d_0p_0)^{-1} + (d_1p_1)^{-1}}} - z_{1-\alpha/2}\right) - 1 \end{aligned} \quad (17)$$

In the special case when $\lambda_0 = \lambda_1$ and $-\log(M_1) = \log(M_u)$ (i.e., $M_u M_1 = 1$), the sample size is given by:^[9]

$$n_{\text{tar}} = [(d_0p_0)^{-1} + (d_1p_1)^{-1}] \frac{(z_{1-\alpha/2} + z_{(1+p)/2})^2}{\log^2(M_u)} \quad (18)$$

In general, a closed-form sample size formula does not exist although some approximations can be derived when $\beta = \log(\lambda_1 / \lambda_0)$ is sufficiently close to 0^[7,9] or far from 0.^[9,10] We would recommend using a numerical method (e.g., bisection method) to find the sample size at which the power in Eq. 17 is not less than the target power.

SAMPLE SIZE BOUNDS

By using Jensen's inequality and Cauchy-Schwarz inequality, Tang^[3] derived the lower (d_{gl}) and upper (d_{gu}) bounds on d_g :

$$d_{gl} = \frac{\lambda_g v_g^2}{v_g + \kappa \lambda_g E(t_{gj}^2)} \leq d_g \leq d_{gu} = \frac{\lambda_g v_g}{1 + \kappa \lambda_g v_g} \quad (19)$$

where $v_g = E(t_{gj})$ is the mean follow-up time in group g . Replacing d_g by d_{gu} in the power and sample size formulae leads to a lower sample size bound and an upper power bound and replacing d_g by d_{gl} leads to an upper sample size bound and a lower power bound.

The lower size bound is the required size when all subjects in treatment group g are followed for the same time $t_{gj} = v_{tg}$ and can be decomposed into two terms:

$$n_l = \left[\frac{1}{p_0 \lambda_0 v_{t_0}} + \frac{1}{p_1 \lambda_1 v_{t_1}} \right] f + (p_0^{-1} + p_1^{-1}) \kappa f \quad (20)$$

where $f = (z_{1-\alpha/2} + z_p)^2 / \beta^2$ in a superiority trial, and $f = (z_{1-\alpha/2} + z_p)^2 / \beta^2$ in a NI trial. In an equivalence trial, the sample size bounds can be obtained by inverting the power bounds numerically, and $f = (z_{1-\alpha/2} + z_{(1+p)/2})^2 / \log^2(M_u)$ in the special case when $\lambda_0 = \lambda_1$ and $-\log(M_l) = (M_u)$. In Eq. 20, the first term is the required size ignoring overdispersion (i.e., the count data follow Poisson distribution), and the second term corrects for overdispersion. In the upper size bound, another term is added to account for variation in the duration of the follow-up.

$$n_u = n_l + \kappa f \left[\frac{1}{p_0} CV_0^2 + \frac{1}{p_1} CV_1^2 \right] \quad (21)$$

where $CV_g = \sqrt{\text{var}(t_{gj})} / v_{tg}$ is the coefficient of variation for t_{gj} . Note that n_u can be further bounded by:

$$n_u \leq \tilde{n}_u = n_l + \kappa f \left[\frac{t_{m_0} - v_{t_0}}{v_{t_0}} + \frac{t_{m_1} - v_{t_1}}{v_{t_1}} \right] \quad (22)$$

where t_{m_g} is the maximum follow-up time in group g .

Tang^[3] suggested using n_u as a conservative sample size estimate for several reasons: 1) it only slightly overestimates the required size in the NB regression especially for design 1 considered in the next section; 2) it is the required size for the robust Wald test from the Andersen-Gill^[11] model, and thus robust against the misspecification of the mixing distribution in the mixed Poisson regression; and 3) it requires information only on the mean and variance of t_{gj} , which are easier to compute than d_g .

The dispersion parameter κ is generally not reported in the medical literature. The inequalities given in the section can be used to back-calculate κ based on the reported point estimates of the event rate and rate ratio and their CI (refer to Tang^[9] for a numerical illustration).

ESTIMATION OF d_g , $E(t_{gj})$, AND $E(t_{gj}^2)$

Analytic expressions for d_g , $E(t_{gj})$, and $E(t_{gj}^2)$ are given for two types of designs studied by Tang.^[9] Note that $v_{tg} = E(t_{gj})$ and $E(t_{gj}^2)$ are needed in calculating the lower and upper size bounds, and d_g is required for computing n_{tar} . In general, evaluating d_g requires numerical integration and can be done using the SAS QUAD subroutine.^[3]

In design 1, the planned treatment duration is τ_c years for all patients, and the loss to follow-up is assumed to be exponentially distributed with mean δ^{-1} years and independent of the event process. The overall dropout rate is $w_c = 1 - \exp(-\delta \tau_c)$ and:

$$\begin{aligned} d_g &= E \left[\frac{\lambda_g t_{gj}}{1 + \kappa \lambda_g t_{gj}} \right] = \frac{1}{\kappa} - E \left(\frac{1}{\kappa + \kappa^2 \lambda_g t_{gj}} \right) \\ &= \frac{1}{\kappa} - \frac{\exp(-\delta \tau_c)}{\kappa + \kappa^2 \lambda_g \tau_c} - \int_0^{\tau_c} \frac{\delta \exp(-\delta t)}{\kappa + \kappa^2 \lambda_g t} dt \end{aligned} \quad (23)$$

$$E(t_{gj}) = \frac{1 - \exp(-\delta \tau_c)}{\delta} = \frac{w_c}{\delta},$$

$$E(t_{gj}^2) = \frac{2[1 - (1 + \delta \tau_c) \exp(-\delta \tau_c)]}{\delta^2}$$

In design 2, we assume patients are enrolled during an accrual period of τ_a years, and after the closure of recruitment, subjects are followed for an additional τ_c years.^[12] That is, all subjects will be administratively censored at time $\tau_a + \tau_c$. The loss to follow-up distribution is the same as design 1. Suppose the patient entry time e_{gj} is distributed as $g(e_{gj}) = \eta \exp(-\eta e_{gj}) / (1 - \exp(-\eta \tau_a))$.^[13] The entry distribution is convex (faster patient entry at the beginning) if $\eta > 0$, concave (lagging patient entry) if $\eta < 0$, and uniform $g(e_{gj}) = 1/\tau_a$ if $\eta \rightarrow 0$. Then:

$$\begin{aligned} d_g &= \frac{1}{\kappa} - \int_0^{\tau_c} \frac{\delta \exp(-\delta t)}{\kappa + \kappa^2 \lambda_g t} dt - \int_0^{\tau_a} \frac{\exp(-\delta(t + \tau_c))}{\kappa + \kappa^2 \lambda_g(t + \tau_c)} h(t) dt \\ E(t_{gj}) &= \frac{1}{\delta} - \frac{\exp(-\delta \tau_c)}{\delta} h_1 \\ E(t_{gj}^2) &= \frac{2}{\delta^2} \left\{ 1 - \exp(-\delta \tau_c) \left[(\delta \tau_c + 1) h_1(t) + \delta h_2(t) \right] \right\} \end{aligned} \quad (24)$$

where,

$$\begin{aligned} h(t) &= \frac{\delta + (\eta - \delta) \exp(\eta(t - \tau_a))}{1 - \exp(-\eta \tau_a)} \\ h_1 &= \frac{\eta \exp(-\delta \tau_a) - \exp(-\eta \tau_a)}{\eta - \delta} \\ h_2 &= \frac{\eta \exp(-\eta \tau_a) [(\eta - \delta) \tau_a - 1] \exp((\eta - \delta) \tau_a) + 1}{1 - \exp(-\eta \tau_a) (\eta - \delta)^2} \end{aligned} \quad (25)$$

For uniform patient entry ($\eta \rightarrow 0$), Eq. 25 reduces to:

$$\begin{aligned} h(t) &= \delta \frac{\tau_a - t}{\tau_a} + \frac{1}{\tau_a} \\ h_1 &= \frac{1 - \exp(-\delta \tau_a)}{\delta \tau_a} \\ h_2 &= \frac{1 - (\delta \tau_a + 1) \exp(-\delta \tau_a)}{\delta^2 \tau_a} \end{aligned} \quad (26)$$

ALTERNATIVE APPROACHES

Sample Size Assuming Equal Follow-Up Times

The sample size formulae comparing two NB rates have been developed for superiority trials^[14–16] (method 2 of Zhu and Lakkis^[15]), NI trials^{[16][17]} (method 1 of Zhu^[17]), and equivalence trials^[17] under the assumption of equal follow-up. This could be achieved by ignoring information from early dropouts for design 1 or by setting the follow-up time for all individuals to be the mean follow-up time. If data from early dropouts are ignored, the sample size after adjusting for dropout rate is overestimated by about 8%–45% in the numerical examples given by Tang.^[3] If the length of the follow-up for all individuals is set to be the mean follow-up time, the resulting sample size is in fact identical to the lower size bound n_l . The sample size formula for superiority trials developed by Cook^[12] is also identical to the lower size bound n_l .^[3]

Sample Size Using Variance Under Null Hypothesis

Zhu and Lakkis^[15] (method 3 in their study) proposed one variation of the sample size formula for superiority trials. The approach uses the variance derived under $H_0: \lambda_0 = \lambda_1$ by setting $t_{gj} = v_t = v_0 = v_1$. Specifically,

$$n_{zl} = \frac{[z_{1-\alpha/2} \sqrt{V_0} + z_p \sqrt{V_1}]^2}{\beta^2} \quad (27)$$

where $\tilde{\lambda}_0 = \tilde{\lambda}_1 = p_0 \lambda_0 + p_1 \lambda_1$ is the expected value of the MLE under H_0 ,

$$\begin{aligned} V_1 &= (p_0^{-1} + p_1^{-1})\kappa + (p_0 \tilde{\lambda}_0 v_t)^{-1} + (p_1 \tilde{\lambda}_1 v_t)^{-1}, \\ V_0 &= (p_0^{-1} + p_1^{-1})\kappa + (p_0 \tilde{\lambda}_0 v_t)^{-1} + (p_1 \tilde{\lambda}_1 v_t)^{-1} \end{aligned} \quad (28)$$

The numerical examples in Zhu and Lakkis^[15] indicate that Eq. 27 generally produces similar sample size estimate to the lower size bound given in Eq. 20 because $V_0 \approx V_1$ and:

$$n_l = \frac{[z_{1-\alpha/2} + z_p]^2 V_1}{\beta^2} = \frac{[z_{1-\alpha/2} \sqrt{V_1} + z_p \sqrt{V_1}]^2}{\beta^2} \quad (29)$$

Under equal allocation ($p_0 = p_1 = 1/2$), $V_0 < V_1$ and Eq. 27 always yields slightly smaller sample size estimate than the lower sample size bound given in Eq. 20, i.e., $n_{zl} < n_l$.

A similar variation was proposed by Zhu^[17] (method 3 in his study) for NI trials:

$$n_{zl} = \frac{[z_{1-\alpha/2} \sqrt{\tilde{V}_0} + z_p \sqrt{V_1}]^2}{\log^2(M_0 \lambda_0 / \lambda_1)} \quad (30)$$

where $\tilde{V}_0 = (p_0^{-1} + p_1^{-1})\kappa + (p_0 \tilde{\lambda}_0 v_t)^{-1} + (p_1 \tilde{\lambda}_1 v_t)^{-1}$, $\theta = p_1 / p_0$, $a = -\kappa v_t M_0 (1 + \theta)$, $b = \kappa v_t (\lambda_0 M_0 + \theta \lambda_1) - (1 + \theta M_0)$, $c = \lambda_0 + \theta \lambda_1$, and $\tilde{\lambda}_0 = (-b - \sqrt{b^2 - 4ac}) / (2a)$ and $\tilde{\lambda}_1 = M_0 \tilde{\lambda}_0$ are the approximate MLE under $H_0: \lambda_1 = M_0 \lambda_0$.

Zhu^[16] (method 3 in his study) also considered the equivalence trials when $\log(M_u) = -\log(M_l)$. The power can be calculated as:

$$\begin{aligned} &\Phi \left(\frac{\sqrt{n} \left| \log \frac{\lambda_1}{\lambda_0} + \log(M_u) \right| - Z_{\alpha/2} \sqrt{V_0^-}}{\sqrt{V_1}} \right) \\ &+ \Phi \left(\frac{\sqrt{n} \left| \log \frac{\lambda_1}{\lambda_0} + \log(M_u) \right| - Z_{\alpha/2} \sqrt{V_0^+}}{\sqrt{V_1}} \right) - 1 \end{aligned} \quad (31)$$

where $a = -\kappa v_t (1 + \theta) M_l$, $b = \kappa v_t [\lambda_0 M_l + \theta \lambda_1] - (1 + \theta M_l)$, $c = \lambda_0 + \theta \lambda_1$, $\tilde{\lambda}_0 = (-b - \sqrt{b^2 - 4ac}) / (2a)$, $V_0^- = [1 / p_0 + 1 / (p_1 M_l)] / (v_t \tilde{\lambda}_0) + \kappa / (p_0 p_1)$ and V_0^+ can be calculated in the same way with M_l replaced by M_u . Sample size can be calculated by inverting Eq. 31 numerically.

In general, the sample size estimate based on Eqs. 30 and 31 is close to the lower size bound given in Eq. 20 except under extreme parameter configurations.

Cook's Formulae for NI Trials

Cook, Lee, and Li^[18] derived the power formulae for NI trials under the assumption of equal follow-up. The margin is treated as a random quantity in the power calculation.

Sample Size for NI Trials Based on Rate Difference

So far, we assume that the treatment effect is measured by the event rate ratio between two arms. Treatments can also be compared based on the absolute difference in event rates. The choice of the treatment effect metric may depend on the therapeutic areas.^[5] By delta method, Tang^[9] derived the sample size for NI trials:

$$n = \left(\frac{\lambda_0^2}{p_0 d_0} + \frac{\lambda_1^2}{p_1 d_1} \right) \frac{(z_{1-\alpha/2} + z_p)^2}{(\lambda_1 - \lambda_0 - M_{0d})^2} \quad (32)$$

where M_{od} is the NI margin for the absolute rate difference. When the two margins satisfy $M_{od} = \bar{\lambda} \log(M_0)$, Eqs. 32 and 16 generally give close sample size estimates, where $\bar{\lambda} = \sqrt{\lambda_0 \lambda_1}$ is the geometric mean of λ_0 and λ_1 .

Sample Size Under Heterogeneous Dispersion

The power and sample size formulae described in early sections assume a common dispersion parameter across treatment groups. If this assumption does not hold and the dispersion parameter is assumed to vary between groups in the analysis, the power and sample size formulae remain almost unchanged except that one needs to replace κ in d_g , d_{su} , and d_{gl} by the treatment-specific dispersion parameter κ_g .^[9]

Table 1 provides a summary of existing sample size calculation methods for the NB regression by the type of trials and the assumption on the follow-up time.

NUMERICAL ILLUSTRATIONS

We illustrate the performance of several methods using simulation studies. In all cases, the difference in sample size between two arms is 0 or 1 depending on whether the total size n_{tar} is even or odd, and 40,000 trials are generated so that there is about 95% chance that the simulated power is within 0.4% of the true power (standard error $\approx \sqrt{0.8 * 0.2 / 40000} = 0.2\%$).

Table 2 displays the result for superiority trials. For both designs, we set $(\lambda_0, \kappa) = (0.6, 1)$ or $(0.8, 1.2)$ and the RRR to be $1 - \lambda_1 / \lambda_0 = 30\%$ or 50% . The loss to follow-up is exponentially distributed. The simulated power at n_{tar} is close to the 80% nominal power in all cases. Zhu–Lakkis's Eq. 27 yields smaller size estimate than the lower size bound, i.e., $n_{zl} < n_l$. The use of the lower size bound n_l or Zhu–Lakkis estimate by ignoring the variability in the follow-up duration underestimates the required size, and the amount of

Table 1 Summary of existing sample size calculation procedures for the NB regression by the type of trials and the assumption on the follow-up time

Trial	Equal follow-up		Unequal follow-up
	Completer cases ^a	Average follow-up ^b	
Superiority	[14,16]	[12,15]	[3]
NI	[16]	[17,18] ^c	[9] ^d
Equivalence		[17]	[9]

^aInformation from early dropouts is ignored.

^bThe follow-up time for each individual is set to be the average follow-up time in the treatment group. In Zhu and Lakkis^[15] and Zhu,^[17] the average follow-up time is assumed to be the same across treatment groups.

^cNI margin is not treated as fixed in Cook, Lee, and Li^[18]

^dMethods in Tang^[9] allow the dispersion parameter κ to vary across treatment groups.

Table 2 Estimated sample size needed to achieve 80% power and empirical power at n_{tar} for superiority trials

(λ_0, λ_1)	κ	Total sample size				Power (%) at n_{tar}
		n_{zl}	n_l	n_{tar}	n_u	
Design 1: $\tau_c = 2$, dropout 30% ($\delta = 0.178$)						
(0.6, 0.30)	1.0	155	163	169	171	80.89
(0.6, 0.42)	1.0	538	544	568	573	80.07
(0.8, 0.40)	1.2	146	152	159	161	80.84
(0.8, 0.56)	1.2	514	519	545	554	80.23
Design 2: $\tau_a = \tau_c = 2$, $\delta = 0.3$						
(0.6, 0.30)	1.0	143	149	164	170	80.92
(0.6, 0.42)	1.0	497	502	557	582	80.50
(0.8, 0.40)	1.2	137	142	157	167	80.59
(0.8, 0.56)	1.2	484	488	545	584	80.24

1) n_{zl} , n_l , n_{tar} , and n_u are evaluated using Eqs. 27, 20, 13, 21, respectively.

2) Data in design 2 are simulated under the assumption of uniform entry.

3) Empirical power is evaluated based on 40,000 simulated trials.

Table 3 Estimated sample size needed to achieve 80% power and empirical power at n_{tar} for NI trials

(λ_0, λ_1)	κ	Total sample size				Power (%) at n_{tar}
		n_{zl}	n_l	n_{tar}	n_u	
Design 1: $\tau_c = 2$, dropout 30% ($\delta = 0.178$)						
(0.6, 0.36)	1.0	148	152	158	160	80.76
(0.6, 0.54)	1.0	776	777	812	821	80.00
(0.8, 0.48)	1.2	140	144	151	153	80.93
(0.8, 0.72)	1.2	753	753	791	807	79.92
Design 2: $\tau_a = \tau_c = 2$, $\delta = 0.3$						
(0.6, 0.36)	1.0	136	140	155	161	80.57
(0.6, 0.54)	1.0	720	721	801	843	79.95
(0.8, 0.48)	1.2	132	135	150	160	80.60
(0.8, 0.72)	1.2	711	711	795	858	80.08

1) $M_0 = 1.2$

2) n_{zl} , n_l , n_{tar} , and n_u are evaluated using Eqs. 30, 20, 16, and 21, respectively.

3) Data in design 2 are simulated under the assumption of uniform entry.

4) Empirical power is evaluated based on 40,000 simulated trials.

Table 4 Estimated sample size needed to achieve 80% power and empirical power at n_{tar} for equivalence trials

(λ_0, λ_1)	κ	Total sample size				Power (%) at n_{tar}
		n_{zl}	n_l	n_{tar}	n_u	
Design 1: $\tau_c = 2$, dropout 30% ($\delta = 0.178$)						
(0.6, 0.6)	1.0	1219	1216	1272	1288	79.68
(0.8, 0.8)	1.2	1189	1187	1248	1274	80.14
Design 2: $\tau_a = \tau_c = 2$, $\delta = 0.3$						
(0.6, 0.6)	1.0	1133	1131	1259	1328	79.84
(0.8, 0.8)	1.2	1124	1123	1255	1360	79.89

1) $M_0 = 1/M_1 = 1.3$

2) n_{zl} , n_l , n_{tar} , and n_u are evaluated based on Eqs. 31, 20, 18, and 21, respectively.

3) Data in design 2 are simulated under the assumption of uniform entry.

4) Empirical power is evaluated based on 40,000 simulated trials.

underestimation reaches about 10% in all four cases for design 2. The upper bound n_u only slightly overestimates the required size (by $\leq 2\%$ in all four cases) for design 1.

Table 3 reports the result for NI trials. For both designs, we set $(\lambda_0, \kappa) = (0.6, 1)$ or $(0.8, 1.2)$ and the rate ratio to be $\lambda_1 / \lambda_0 = 60\%$ or 90% . We fix the NI margin for the rate ratio at $M_0 = 1.2$. The performance is similar to that for superiority trials.

Table 4 reports the result for equivalence trials. In both designs, we set $(\lambda_0, \lambda_1, \kappa) = (0.6, 0.6, 1)$ or $(0.8, 0.8, 1.2)$. We fix the margin at $M_u = 1/M_l = 1.3$. The use of the lower size bound n_l or Zhu–Lakkis estimate underestimates the required size.

CONCLUSION

Both theoretical results and simulations confirm that it is necessary to adjust for the between-subject variation in the follow-up time in the power and sample size calculation for comparing two NB rates. Ignoring such variation leads to underpowered trials. One may also use the upper size bound n_u as a slightly conservative size estimate for design 1.

REFERENCES

- Glynn, R.J.; Buring, J.E. Ways of measuring rates of recurrent events. *BMJ* **1996**, *312* (7027), 364–366.
- Wang, Y.; Meyerson, L.; Tang, Y.; Qian, N. Statistical methods for the analysis of relapse data in MS clinical trials. *J. Neurol. Sci.* **2009**, *285* (1–2), 206–211.
- Tang, Y. Sample size estimation for negative binomial regression comparing rates of recurrent events with unequal follow-up time. *J. Biopharm. Stat.* **2015**, *25* (5), 1100–1113.
- Lawless, J.F. Negative binomial and mixed poisson regression. *Can. J. Stat.* **1987**, *15* (3), 209–225.
- United States Food and Drug Administration. *Guidance for Industry Non-Inferiority Clinical Trials*, 2010.
- EMA Committee for Medicinal Products for Human Use (CHMP). *Guideline on the Choice of the Noninferiority Margin*; EMA: London, 2005.
- Chow, S.C.; Wang, H.; Shao, J. *Sample Size Calculations in Clinical Research*. 2nd Ed.; Chapman & Hall/CRC: Boca Raton, FL, 2008.
- Saha, K.; Paul, S. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics* **2005**, *61* (1), 179–185.
- Tang, Y. Sample size for comparing negative binomial rates in noninferiority and equivalence trials with unequal follow-up times. *J. Biopharm. Stat.* **2017**, 1–17.
- Chow, S.C.; Wang, H. On sample size calculation in bioequivalence trials. *J. Pharmacokinet. Phar.* **2001**, *28* (2), 155–169.
- Andersen, P.K.; Gill, R.D. Cox's regression model for counting processes: A large sample study. *Ann. Stat.* **1982**, *10* (4), 1100–1120.
- Cook, R.J. The design and analysis of randomized trials with recurrent events. *Stat. Med.* **1995**, *14* (19), 2081–2098.
- Lachin, J.M.; Foulkes, M.A. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics* **1986**, *42* (3), 507–519.
- Keene, O.N.; Jones, M.R.; Lane, P.W.; Anderson, J. Analysis of exacerbation rates in asthma and chronic obstructive pulmonary disease: example from the TRISTAN study. *Pharm. Stat.* **2007**, *6* (2), 89–97.
- Zhu, H.; Lakkis, H. Sample size calculation for comparing two negative binomial rates. *Stat. Med.* **2014**, *33* (3), 376–387.
- Friede, T.; Schmidli, H. Blinded sample size reestimation with negative binomial counts in superiority and non-inferiority trials. *Method. Inform. Med.* **2010**, *49* (6), 618–624.
- Zhu, H. Sample size calculation for comparing two Poisson or negative binomial rates in non-inferiority or equivalence trials. *Stat. Biopharm. Res* **2017**, *9* (1), 107–115.
- Cook, R.J.; Lee, K.A.; Li, H. Non-inferiority trial design for recurrent events. *Stat. Med.* **2007**, *26* (25), 4563–4577.

Scaled Average Bioequivalence (SABE)

Laszlo Endrenyi

Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada

Laszlo Tothfalusi

Department of Pharmacodynamics, Semmelweis University, Budapest, Hungary

Abstract

The approach of scaled average bioequivalence (SABE) standardizes, by the within-subject standard deviation, the more frequently applied method of unscaled average bioequivalence. The US Food and Drug Administration recommends that SABE be applied for determining bioequivalence (BE) for drugs that have large within-subject variation and also for those that have a narrow therapeutic index. The European Medicines Agency suggests another procedure for the evaluation of BE of highly variable drug products. Average BE with expanding limits (ABEL) is closely related to SABE and has similar characteristics. Both SABE and ABEL require fewer subjects than unscaled average BE with highly variable drugs. The contribution discusses their background as well as their statistical and regulatory features and presents procedures for their implementation.

Keywords: Bioequivalence; Highly variable drugs; Narrow therapeutic index; Scaled average bioequivalence; Two one-sided tests.

INTRODUCTION

Scaled average bioequivalence (SABE) is an approach for the determination of bioequivalence (BE) of two drug products, which modifies the more frequently applied method of (unscaled) average BE (ABE). Both approaches will be presented in detail. In essence, SABE standardizes the procedure of unscaled ABE by the within-subject standard deviation.

The application of SABE was first recommended for dealing with difficulties encountered when the BE of highly variable (HV) drugs was evaluated.^[1–3] These difficulties arise because the assessment of BE of HV drug products with the regulatory requirements for unscaled ABE cannot in practice be satisfied unless an unethically large number of subjects is enrolled in an investigation.^[4]

The use of SABE has been recently suggested also for the determination of BE of drug products that have a narrow therapeutic index (NTI), those for which there is only small difference between the doses and/or concentrations giving rise to therapeutic and adverse clinical effects.^[5]

Regulatory requirements for unscaled ABE will first be reviewed in this report. It will be followed by the definition and some features of SABE and of a closely related method. Procedures for the implementation of SABE will be summarized. Requirements of various regulatory authorities will then be presented together with their respective merits and disadvantages.

UNSCALED ABE

It is typically expected that the difference between the logarithmic means of relevant pharmacokinetic parameters for the test and reference drug products (μ_T and μ_R) should be between two regulatory limits (θ_1 and θ_2):

$$\theta_1 \leq \mu_T - \mu_R \leq \theta_2 \quad (1)$$

The BE limits are usually symmetrical in the logarithmic scale, $-\theta_1 = \theta_2 = \theta_A$, and therefore,

$$-\theta_A \leq \mu_T - \mu_R \leq \theta_A \quad (2)$$

Regulatory authorities set the value of the BE limit most frequently at $\theta_A = \ln(1.25)$. This corresponds to a geometric mean ratio (*GMR*) between the two formulations:

$$GMR = \exp(\mu_T / \mu_R) = 1.25, \quad (3)$$

i.e., to a 25% difference between the two untransformed means.

BE is tested by two one-sided null hypotheses. They assume the deviation between the logarithmic means is either too low:

$$H_{01} : \mu_T - \mu_R \leq -\theta_A \quad (4a)$$

or too high:

$$H_{02} : \mu_T - \mu_R \geq \theta_A. \quad (4b)$$

The alternative hypotheses represent BE as follows:

$$H_{a1} : \mu_T - \mu_R > -\theta_A \quad (5a)$$

and

$$H_{a2} : \mu_T - \mu_R < \theta_A. \quad (5b)$$

Therefore, BE can be declared if both null hypotheses are rejected in favor of the alternative hypotheses. An operational equivalent of this two one-sided tests (TOST) procedure is

$$-\theta_A \leq \mu_T - \mu_R \leq \theta_A. \quad (2)$$

Consequently, BE can be stated if the confidence interval for the difference between logarithmic means is between the BE limits.^[6,7] Usually, the 90% confidence limits are expected to be between $-\ln(1.25)$ and $+\ln(1.25)$. Thus, not more than a 5% probability is assumed as the Type I error for both null hypotheses.

In practice, the estimated values of the logarithmic means (m_T and m_R) are obtained from the data, and their 90% confidence interval is calculated. For example, in the most frequently performed crossover investigations with two periods and two sequences, the regulatory expectation is

$$-\theta_A \leq (m_T - m_R) \pm t(2s^2/N)^{1/2} \leq \theta_A \quad (6)$$

Here, t is the upper fifth percentile of Student's statistic with $N - 2$ degrees of freedom. N is the number of subjects. s^2 is the residual variation calculated in a relevant analysis of variance; it is related to, but generally not identical with, the within-subject variance.

SABE FOR HV DRUGS

The regulatory model for ABE (Eqs 2 and 5a,b) can be modified by applying a scaling factor (s_w) as follows^[1-3]:

$$-\theta_s \leq (m_T - m_R)/s_w \leq \theta_s. \quad (7)$$

The magnitude of the BE limit (θ_s) is set by the regulatory authorities. s_w is the standard deviation related to the within-subject variation. In two-period crossover investigations, it can be obtained from the residual variation. In studies with three or more periods, the intrasubject variation of one or both drug products can be directly evaluated.

The regulatory expectation of SABE, represented by Eq. 7, is again that its 90% confidence limits should lie within the BE limits of $\pm\theta_s$. Calculation of the BE limits will be summarized later.

The main advantage of SABE is that, for HV drugs, it requires meaningfully fewer subjects for the demonstration of BE compared to unscaled ABE. This is illustrated in Fig. 1. Details of the figure will be presented

later. It shows, under various conditions, the percentage of simulated three-period crossover studies with 24 subjects in which BE was accepted. The power curves were obtained by gradually increasing the *GMR*. At *GMR* = 1.00, the two products are bioequivalent and the highest acceptance is recorded (producer risk). Power curves are presented which are calculated using unscaled ABE and SABE. Acceptances by SABE are substantially higher than those yielded by ABE especially at larger variations. Other conditions shown in Fig. 1 will be described later. The producer risk does not depend on the variation.

SABE has interesting interpretations.^[4] It is an index of standardized effect size^[8] that can measure the importance and significance of clinical effects. It is also widely used in diverse areas of science. SABE is also a distance metric in the function space of probability distributions.^[9] SABE can be derived, with simplifying assumptions, from the regulatory model for individual BE and therefore provides a measure of therapeutic switchability.^[2,10-12]

REGULATORY IMPLEMENTATIONS OF SABE

Regulatory authorities have pursued differing approaches for the implementation of SABE. The approaches of the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA) of the European Union have been discussed and compared^[3,13] in Sections "Food and Drug Administration" and "European Medicines Agency."

Both agencies require that the estimated within-subject standard deviation of the reference product (s_{wR}) should be used as the measure of intra-subject variation. Consequently, the regulatory model for the implementation of SABE is

$$-\theta_s \leq (m_T - m_R)/s_{wR} \leq \theta_s. \quad (8)$$

The BE limit (θ_s) is set by the regulatory authorities. It is related to a regulatory constant (σ_0) by

$$\theta_s = \ln(1.25)/\sigma_0. \quad (9)$$

It was recently found^[14] that direct evaluation of Eq. 8 yields biased estimates. The bias could be corrected with Hedges' procedure.^[15] An exact procedure was proposed which is efficient and can be computed easily.^[14]

Food and Drug Administration

The regulatory requirements of the US FDA for the determination of BE of HV drug formulations were presented in a scientific publication.^[3] They were confirmed in more recent publications^[13,14] and a draft guidance.^[13]

The FDA recommends the application of SABE (Eq. 8). In order to obtain the estimated within-subject standard deviation (s_{wR}), the reference formulation should be measured (at least) twice in each subject. Therefore, replicate design studies with either three or four periods should be performed.

GMR limit = 1.25

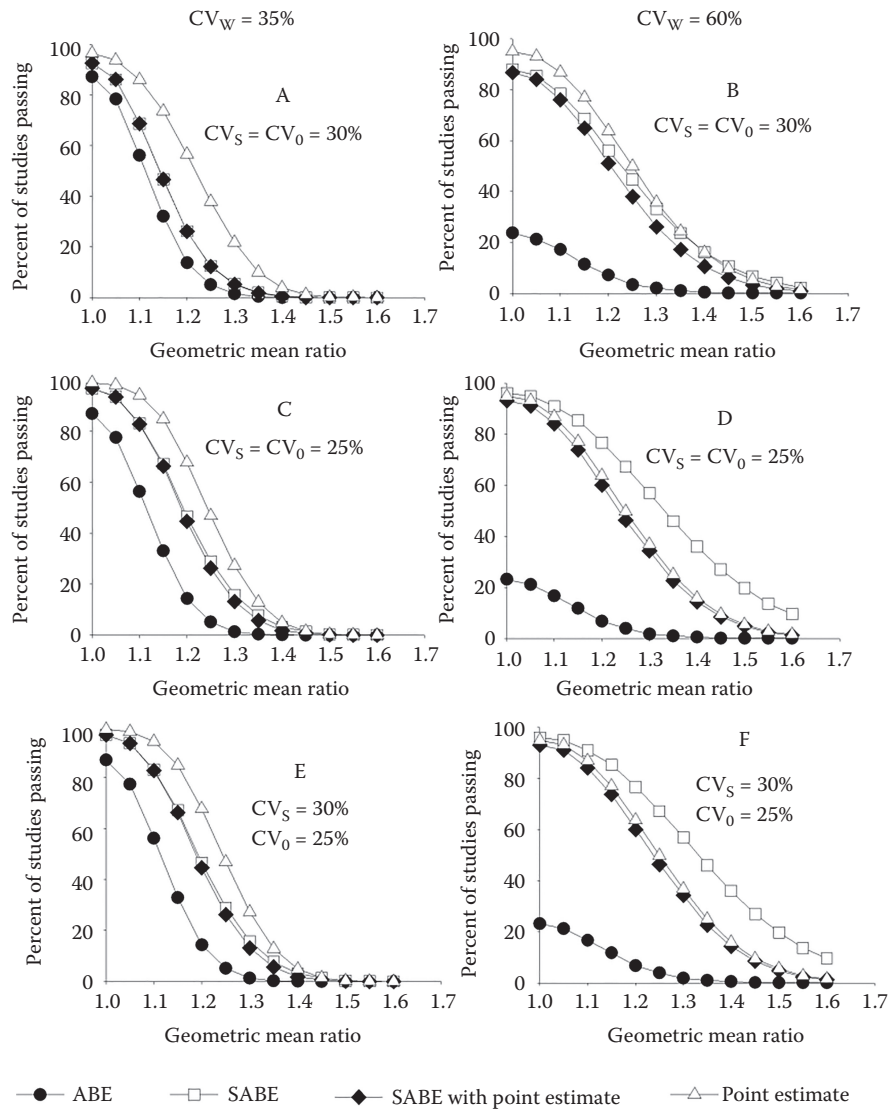


Fig. 1 Power curves for BE studies, which are evaluated by four methods: unscaled ABE, SABE, the point estimate criterion for *GMR*, and the combination of the SABE and point estimate criteria. The power curves contrast the proportion (in %) of simulated studies in which BE is accepted with increasing *GMR*. The within-subject variation is either 35% (A, C, D) or 60% (B, D and F). EMA requires $CV_s = 30\%$ and $CV_0 = 30\%$ (A and B) whereas FDA expects $CV_s = 30\%$ and $CV_0 = 25\%$ (E and F). Consequences of an unusual switching variation of $CV_s = 25\%$ are illustrated in (C and D)
 Source: [19], with the permission of the publisher.

The FDA suggests $\sigma_0 = 0.25$ (corresponding to a coefficient of variation of $CV_0 = 25\%$) as the value of the regulatory constant, and therefore, $\theta_s = 0.893$.^[3,13,16,17]

By international agreement, a drug is considered to be HV if its within-subject variation is $>30\%$.^[18] Therefore, the FDA recommends that SABE be applied if the observed variation is $>30\%$, whereas ABE be used at lower variations. This is a *mixed regulatory strategy* which uses a switching variation of 30% ($CV_s = 30\%$).

Figure 2 illustrates the strategy. When the recorded intra-subject variation exceeds 30%, the BE limits,

evaluated with ABE, widen with increasing s_{WR} . When the within-subject variation is not $>30\%$, the BE limits remain constant at $\pm \ln(1.25)$. Figure 2B shows that the strategy results in discontinuous (implied) BE limits because the switching variation is 30%, different from the stated regulatory constant of $\sigma_0 = 0.25$.^[2,19]

FDA suggests that, in order to calculate the confidence limits, the SABE model (Eq. 8) be squared and then linearized:

$$(m_T - m_R)^2 - \theta_s^2 \cdot s_{WR}^2 \leq 0. \quad (10)$$

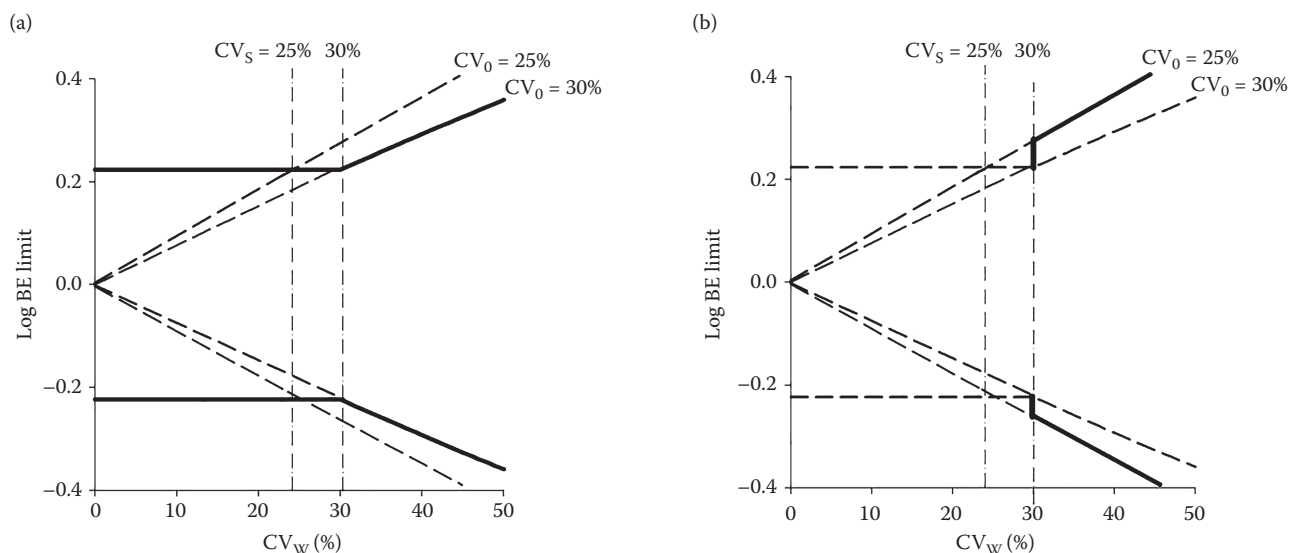


Fig. 2 Mixed regulatory model for the determination of BE of HV drugs. Thick lines show the logarithmic BE limits. They have a constant level of $\pm \ln(1.25)$, and unscaled ABE is applied, when the within-subject variation (CV_w) does not exceed 30%. The limits widen at higher variations and either SABE or ABEL is applied. The slope of the expansion (in the logarithmic scale) is determined by a regulatory constant which corresponds to a standardized variation (CV_0). (A) According to the requirement of EMA, $CV_0 = 30\%$; the BE limits are continuous. (B) Following the expectation of FDA, $CV_0 = 25\%$; the BE limits are discontinuous around $CV_w = 30\%$. Source: Endrenyi and Tothfalusi,^[19] with the permission of the publisher.

The 95% upper bound can be calculated from the distributions of the two components.^[20,21] BE is accepted if the upper bound is zero or negative and rejected if it is positive. The procedure was implemented in a SAS program and published in a draft guidance of FDA.^[22] However, the procedure has a small bias, which should be taken into account in regulatory considerations.^[14] The computer program also has some difficult features including instability with the partial replicate design and the need for balanced data; the latter requirement was analyzed and criticized.^[23]

For the determination of the BE of HV drug products, the FDA requires also that the point estimate of *GMR* be between 0.80 and 1.25.^[3,13,16,17] This expectation is additional to that based on SABE and introduces awkward complications.^[19]

European Medicines Agency

The approach of EMA for the determination of BE for HV drugs^[24] differs from that of FDA, but the two procedures are closely related. The SABE model (Eq. 8) is rearranged:

$$-\theta_A = -\theta_s * s_{WR} \leq m_T - m_R \leq \theta_s * s_{WR} = \theta_A. \quad (11)$$

Consequently, the BE limits widen in proportion to the estimated within-subject variation of the reference product. Boddy et al.^[25] first recommended this approach of ABE with expanding limits (ABEL).

EMA recommends the BE limit of $\theta_s = 0.760$, and, therefore, the regulatory constant of $\sigma_0 = 0.294$, in applying the ABEL approach.^[24] These regulatory definitions enable

the continuity of the BE limits at the switching variation of $CV_s = 30\%$ (Fig. 2A).^[19]

The approaches of SABE and ABEL are functionally identical and yield closely similar results. ABEL is actually ABE, and, therefore, the confidence limits can be calculated by the simple TOST procedure (as for Eqs 2 and 6); for a statement of BE, the 90% confidence limits are expected to be between the BE limits of $\pm \theta_A$. However, the BE limits of ABEL are random variables; this could be of limited concern which diminishes as the sample size increases. Even a typical study sample of 24 subjects yields satisfactorily robust estimates.

EMA also requires, in addition to the ABEL criterion, that the point estimate of *GMR* be between 0.80 and 1.25. Moreover, the BE limits may be widened only up to a variation of 50%. At higher variations, the limits should remain restricted to $\pm \ln(1.43)$. Furthermore, application of ABEL is limited to only one of the main pharmacokinetic parameters (maximum concentration, C_{max}) but not allowed for the other (area under the concentration–time curve or AUC).^[24]

Comparison of the Approaches of FDA and EMA

The procedures recommended by FDA and EMA for the determination of BE of HV drug products are similar in many respects. Most importantly, they utilize either SABE or an approach closely related to it, ABEL. The two procedures (Eqs 8 and 11) transform into each other. They apply the mixed regulatory strategy which starts to apply SABE or ABEL at a variation of 30%. They impose also

a secondary regulatory criterion that constrains the point estimate of *GMR*.

However, there are meaningful differences in the implementations of the two approaches. The confidence limits for SABE are evaluated from its squared and linearized form (Eq. 10).^[21] A specialized computer program, such as presented by the FDA,^[22] is needed for its calculation. In contrast, the confidence limits for ABEL are obtained by the customary TOST procedure. This appears to be simpler and easier to visualize and interpret.

The FDA and EMA recommend differing values of the regulatory constant σ_0 , 0.25 and 0.294, respectively. The resulting BE limits are $\theta_s = 0.893$ and 0.759, respectively. The differing conditions have important effects.

Figure 2 illustrates that the BE limits are continuous with the constant suggested by EMA around the variation of $CV = 30\%$ (Fig. 2A) but discontinuous with the value proposed by the FDA (Fig. 2B).

Discontinuity has regrettable consequences. First, it gives rise to regulatory uncertainty: the BE limits are strongly different depending on whether the estimated within-subject variation is, e.g., 29% or 31%. Second, the sample size required for pursuing a BE investigation with sufficient power is continuous with the regulatory constant suggested by EMA but discontinuous when the recommendation of FDA is followed.^[19]

The Type I error (i.e., the “consumer risk” of accepting BE when it is actually incorrect) is inflated around the switching variation ($CV_s = 30\%$). The inflation is much higher, almost 15%, with $\sigma_0 = 0.25$ as the value of the regulatory constant than with $\sigma_0 = 0.294$ which yields an inflation to about 7% (Table 1), thereby preferring the recommendation of EMA to that of the FDA.^[19] This observation was recently confirmed.^[26]

The inflation of the Type I error around $CV = 30\%$ is a consequence of the mixed regulatory strategy. Assume, as an illustration, that the true within-subject variation of

a drug is 29%, i.e., that it is not HV. Nevertheless, its variation could be by chance, say, 32% in a given investigation, thereby classifying the drug as being HV. Figure 2 illustrates that the BE limits would then be wider and the probability of acceptance higher. Thus, the risk of false acceptance would be $>5\%$. The risk is seen to be larger when $\sigma_0 = 0.25$ than when $\sigma_0 = 0.294$ (Fig. 2A vs. 2B).^[19]

The issue of the inflation of the Type I error has been controversial, but the observation has been further confirmed.^[27,28]

Moreover, when $\sigma_0 = 0.25$ and at high variation, the joint decision based on both the SABE confidence criterion and the *GMR* constraint is essentially that of the *GMR* constraint and not of the SABE criterion (Fig. 1D and 1F). This is a misleading and unfortunate situation. With $\sigma_0 = 0.294$, the influence of the SABE criterion is maintained much more effectively (Fig. 1B).

Application of SABE lowers the sample sizes needed for the determination of BE in comparison with the use of the usual, unscaled approach.^[29] Sample sizes required for the expectations of the FDA and EMA and with three- and four-period study designs are available at various within-subject variations.^[29] Fewer subjects are required by the FDA than by the EMA procedure. The additional requirements by the two regulatory agencies raise the required number of subjects.

SABE FOR NTI DRUGS

The FDA has recently recommended that SABE be used for the evaluation of BE for some drugs such as warfarin and tacrolimus which have an NTI.^[30] According to one definition, with NTI drugs “small differences in dose or blood concentration may lead to serious therapeutic failures and/or adverse drug reactions.”^[5]

The FDA suggests that the SABE model (Eq. 8) be applied with a regulatory constant of $\sigma_0 = 0.10$. The BE limit is defined now as

$$\theta_s = \ln(1.1111)/\sigma_0 = 1.0536, \quad (12)$$

noting that $1/0.90 = 1.1111$.^[30,31]

In order to obtain the estimated standard deviation of the reference product, four-way, fully replicated crossover investigations should be performed.

The 95% upper confidence bound should be calculated again from the transformation in Eq. 10.

The FDA recommends also additional conditions.^[30,31] First, the BE limits should not be wider than the usual range of 80%–125%. Therefore, the BE limits gradually widen with increasing within-subject variation until this reaches $CV = 21.4\%$ but remain at their constant range thereafter (Fig. 3).^[32] Therefore, the FDA proposes that the confidence interval be estimated by unscaled ABE, and that the results of a study be presented by either unscaled

Table 1 Consumer risk for the determination of BE

Mixed strategy	Regulatory standardized variation (%)	Consumer risk (%)		
		Unscaled ABE	SABE	
			Without <i>GMR</i> constraint	With <i>GMR</i> constraint
No	30	4.95	5.56	5.56
No	25	4.98	16.50	16.34
Yes	30	5.01	6.98	6.98
Yes	25	4.94	14.78	14.61

Switching variation: $CV_s = 30\%$;

Within-subject variation: $CV_w = 30\%$;

Constraint on point estimate = 25% ($GMR = 1.25$);

For the regulatory standardized variation (CV_0), corresponding to the assumed regulatory constant (σ_0), either 25% (the expectation of FDA) or 30% (the requirement of EMA) was considered. The computations were performed with or without the use of the mixed regulatory strategy, and with or without the constraint of $GMR = 1.25$. *GMR*: geometric mean ratio.

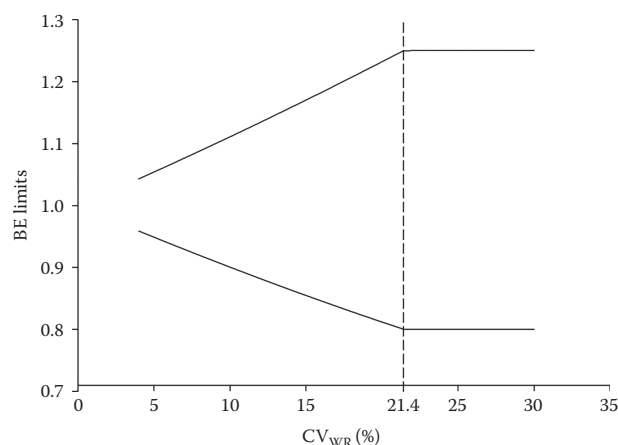


Fig. 3 BE limits for drugs with NTI. SABE is applied and the corresponding implied BE limits widen when the within-subject variation of the reference product (CV_{WR}) increases until it reaches 21.4%. The BE limits remain constant between 80% and 125%, and unscaled ABE is applied when the variation is higher

Source: Endrenyi and Tothfalusi,^[19] with the permission of the publisher.

ABE or SABE depending on whether calculated BE does not or does exceed the 80%–125% range.

The FDA suggests also that a second test be performed that compares the within-subject variances of the two drug products. By applying an F -test, the 90% upper limit for the test/reference variance ratio should not be higher than 6.25.^[30,31]

It is not known how the additional conditions suggested by the FDA affect the performance of primary test involving the confidence limits for SABE.

CONCLUSIONS

SABE is a valuable and useful approach for evaluating the BE of drug products under certain special circumstances. When a drug has large within-subject variation, SABE enables a substantial increase in power and a reduction in the number of subjects required for an effective investigation (Fig. 1).

The FDA recommends the model of SABE (Eq. 8) with a regulatory constant of $\sigma_0 = 0.25$ for the analysis of HV drugs. However, this constant results in a discontinuity of the BE limits and a higher Type I error around $CV = 30\%$ (Fig. 2B) as well as in larger regulatory uncertainty. The confidence bound for the relative pharmacokinetic parameters is calculated from a transformed form of the model (Eq. 10) by using specialized computer programs. Evaluation of SABE yields biased estimates which is sample size dependent for the FDA method and can be easily corrected for the exact approach.^[14]

EMA proposes that ABEL (Eq. 11), a procedure closely related to and a transformation of SABE, be used

with a regulatory constant of $\sigma_0 = 0.294$. The BE limits are continuous around $CV = 30\%$ and the usual TOST procedure can be utilized for calculating the confidence limits.

Both the FDA and EMA impose additional regulatory expectations that may modify the performance of the primary requirement, which involves the confidence limits for the comparative parameters.

The FDA also suggests that SABE be applied to the determination of BE for drugs which have an NTI. In this case, $\sigma_0 = 0.10$ is proposed. The expansion of the BE limits is recommended until a range of 80%–125% is reached (Fig. 3).

REFERENCES

- Schall, R. A unified view of individual, population, and average bioequivalence. In *Bio-International 2, Bioavailability, Bioequivalence and Pharmacokinetics*; Blume, H., Midha, K., Eds; Medpharm: Stuttgart, Germany, 1995; 91–106.
- Tothfalusi, L.; Endrenyi, L.; Midha, K.K.; Rawson, M.J.; Hubbard, J.W. Evaluation of the bioequivalence of highly-variable drugs and drug products. *Pharm. Res.* **2001**, *18* (6), 728–733.
- Haidar, S.H.; Davit, B.; Chen, M.-L.; Conner, D.; Lee, L.M.; Li, Q.H.; Lionberger, R.; Makhoul, F.; Patel, D.; Schuirmann, D.J.; Yu, L.X. Bioequivalence approaches for highly variable drugs and drug products. *Pharm. Res.* **2008**, *25*, 237–241.
- Tothfalusi, L.; Endrenyi, L.; Garcia Areta, A. Evaluation of bioequivalence for highly-variable drugs with scaled average bioequivalence. *Clin. Pharmacokinet.* **2009**, *48*, 725–743.
- Yu, L.X. *Approaches to Demonstrate Bioequivalence of Narrow Therapeutic Index Drugs*; FDA Advisory Committee for Pharmaceutical Science and Clinical Pharmacology: Silver Spring, MD, July 26, 2011. <http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeeting-Materials/Drugs/Drugs/AdvisoryCommitteeForPharmaceuticalScienceandClinicalPharmacology/UCM266777.pdf> (accessed in February 14, 2017).
- Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* **1987**, *15*, 657–680.
- Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*, 3rd Ed.; Chapman & Hall/CRC Press, Taylor & Francis: New York, NY, 2008.
- Dragalin, V.; Fedorov, V.; Patterson, S.; Jones, B. Kullback-Leibler divergence for evaluating bioequivalence. *Stat Med.* **2003**, *22* (6), 913–930.
- Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*; Academic Press: New York, NY, 1969.
- Schall, R.; Williams, R.L. Towards a practical strategy for assessing individual bioequivalence. Food and Drug Administration Individual Bioequivalence Working Group. *J. Pharmacokinet. Biopharm.* **1996**, *24* (1), 133–149.

11. Patnaik, R.N.; Lesko, L.J.; Chen, M.L.; Williams, R.L. Individual bioequivalence: New concepts in the statistical assessment of bioequivalence metrics. *Clin. Pharmacokinet.* **1997**, *33*, 1–6.
12. Tothfalusi, L.; Endrenyi, L. Limits for the scaled average bioequivalence of highly variable drugs and drug products. *Pharm. Res.* **2003**, *20*, 382–389.
13. Food and Drug Administration. *Draft Guidance for Industry: Bioequivalence Studies with Pharmacokinetic Endpoints for Drugs Submitted under an ANDA*; Center for Drug Evaluation and Research (CDER): Silver Spring, MD, 2013. <https://www.fda.gov/downloads/drugs/guidances/ucm377465.pdf>. Accessed in February 14, 2017.
14. Tothfalusi, L.; Endrenyi, L. An exact procedure for the evaluation of reference scaled average bioequivalence. *AAPS J.* **2016**, *18* (2), 476–489.
15. Hedges, L.V. Distribution theory for glass's estimator of effect size and related estimator. *J. Educ. Stat.* **1981**, *6* (2), 107–128.
16. Davit, B.M.; Patel, D.T. Bioequivalence of highly variable drugs. In *FDA Bioequivalence Standards*; Yu, L.X., Li, B.V., Eds; AAPS Advances in the Pharmaceutical Sciences Series 13; AAPS Press, Springer: New York, NY, 2014, 139–164.
17. Davit, B.M.; Chen, M.-L.; Conner, D.P.; Haidar, S.H.; Kim, S.; Lee, C.H.; Lionberger, R.A.; Makhoul, F.T.; Nwakama, P.E.; Patel, D.T.; Schuirmann, D.J.; Yu, L.X. Implementation of a reference-scaled average bioequivalence approach for highly variable generic drug products by the US Food and Drug administration. *AAPS J.* **2012**, *14*, 915–924.
18. Shah, V.P.; Yacobi, A.; Barr, W.H.; Breimer, D.; Dobrinska, M.R.; Endrenyi, L.; Fairweather, W.; Gillespie, W.; Gonzalez, M.A.; Hooper, J.; Jackson, A.; Lesko, L.J.; Midha, K.K.; Noonan, P.K.; Patnaik, R.; Williams, R.L. Evaluation of orally administered highly variable drugs and drug formulations. *Pharm. Res.* **1996**, *13*, 1590–1594.
19. Endrenyi, L.; Tothfalusi, L. Regulatory conditions for the determination of bioequivalence of highly variable drugs. *J. Pharm. Pharm. Sci.* **2009**, *12*, 138–149.
20. Howe, W. Approximate confidence limits on the mean of $X+Y$ where X and Y are two tabled independent random variables. *J. Am. Stat. Assoc.* **1974**, *69*, 789–794.
21. Hyslop, T.; Hsuan, F.; Holder, D.J. A small sample confidence interval approach to assess individual bioequivalence. *Stat. Med.* **2000**, *19*, 2885–2897.
22. Food and Drug Administration. *Draft Guidance for Industry: Bioequivalence Recommendations for Progesterone Oral Capsules*; Center for Drug Evaluation and Research (CDER): Silver Spring, MD, 2011. <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM209294.pdf>. Accessed in February 14, 2017.
23. Patterson S.D., Jones B. Viewpoint: Observations on scaled average bioequivalence. *Pharm. Stat.* **2012**, *11* (1), 1–7.
24. European Medicines Agency. *Guideline on the Investigation of Bioequivalence*; London, January 20, 2010. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/01/WC500070039.pdf. Accessed in February 14, 2017.
25. Boddy, A.W.; Snikeris, F.C.; Kringle, R.O.; Wei, G.C.G.; Opperman, J.A.; Midha, K.K. An approach for widening the bioequivalence acceptance limits in the case of highly variable drugs. *Pharm. Res.* **1995**, *12*, 1865–1868.
26. Munoz, J.; Alcaide, D.; Ocana, J. Consumer's risk in the EMA and FDA regulatory approaches for bioequivalence in highly variable drugs. *Stat. Med.* **2016**, *35*, 1933–1943.
27. Labes, D. "Alpha" of Scaled ABE. *Bioequivalence and Bioavailability Forum*; BEBAC Consultancy Services for Bioequivalence and Bioavailability Studies: Vienna, Austria, 2013. http://forum.bebac.at/mix_entry.php?id=10202.
28. Wonnemann, M.; Frömke, C.; Koch, A. Inflation of the type I error: Investigations on regulatory recommendations for bioequivalence of highly variable drugs. *Pharm. Res.* **2015**, *32* (1), 135–143.
29. Tothfalusi, L.; Endrenyi, L. Sample sizes for designing studies for highly variable drugs. *J. Pharm. Pharm. Sci.* **2011**, *15* (1), 73–84.
30. Food and Drug Administration. *Draft Guidance on Warfarin Sodium*; Center for Drug Evaluation and Research (CDER): Silver Spring, MD, 2012. <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201283.pdf>. Accessed in February 14, 2017.
31. Jiang, W.; Makhoul, F.; Schuirmann, D.J.; Zhang, X.; Zheng, N.; Conner, D.; Yu, L.X.; Lionberger R.A. Bioequivalence approach for generic narrow therapeutic index drugs: Evaluation of the reference-scaled approach and variability comparison criterion. *AAPS J.* **2015**, *17* (4), 891–901.
32. Endrenyi, L.; Tothfalusi, L. Determination of bioequivalence for drugs with narrow therapeutic index: Reduction of the regulatory burden. *J. Pharm. Pharm. Sci.* **2013**, *16* (5), 676–682.

Screening Design

John R. Murphy

Eli Lilly and Company, Indianapolis, Indiana, U.S.A.

Abstract

A “screening design” refers to an experimental design that is applicable when a large number of potential causative factors need to be examined to find the most important few that may have an effect on one or more responses of interest. This brief entry provides an illustrated definition of screening design.

INTRODUCTION

A “screening design” refers to an experimental design that is applicable when a large number of potential causative factors need to be examined to find the most important few that may have an effect on one or more responses of interest. Screening designs may be derived from highly fractionated factorial designs and are called “orthogonal main effect plans” or “orthogonal arrays,” because main effects and only a few two-factor interactions are unconfounded with one another. Effects are confounded when the contrasts for estimating them are not orthogonal, and effects are completely confounded when their contrasts are identical. Plackett and Burman^[1] devised orthogonal arrays useful for screening that yield unbiased estimates of all main effects in the smallest design possible. For example, up to 11 factors can be screened in a 12-run Plackett–Burman design. Examples of 8-run and 12-run Plackett–Burman designs are provided in [Tables 1](#) and [2](#), respectively. Interest in using screening designs in industrial settings has increased in recent years due partly to the work of Taguchi,^[2] a Japanese engineer who promoted the use of

Table 1 Eight-Run Plackett–Burman Screening Design

Trial	X1	X2	X3	X4	X5	X6	X7
1	–	–	–	–	–	–	–
2	–	–	+	+	–	+	+
3	–	+	–	+	+	+	–
4	–	+	+	–	+	–	+
5	+	–	–	+	+	–	+
6	+	–	+	–	+	+	–
7	+	+	–	–	–	+	+
8	+	+	+	+	–	–	–

Table 2 Twelve-Run Plackett–Burman Screening Design

Trial	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
1	+	+	–	+	+	+	–	–	–	+	–
2	+	–	+	+	+	–	–	–	+	–	+
3	–	+	+	+	–	–	–	+	–	+	+
4	+	+	+	–	–	–	+	–	+	+	–
5	+	+	–	–	–	+	–	+	+	–	+
6	+	–	–	–	+	–	+	+	–	+	+
7	–	–	–	+	–	+	+	–	+	+	+
8	–	–	+	–	+	+	–	+	+	+	–
9	–	+	–	+	+	–	+	+	+	–	–
10	+	–	+	+	–	+	+	+	–	–	–
11	–	+	+	–	+	+	+	–	–	–	+
12	–	–	–	–	–	–	–	–	–	–	–

designs he termed orthogonal arrays, many of which are similar to Plackett–Burman designs.

The most obvious criticism of screening designs is that they allow at most a small subset of the interactions to be estimated. Interactions can sometimes be more important than main effects for understanding or modeling the process under study. For this reason, many experts in the field recommend sequential experimentation, in which screening is followed by one or more experimental designs using fewer factors, such as larger fractions of factorial designs, central composite designs, or Box–Behnken designs, that permit estimation of interaction and/or curvature. Proponents of the Taguchi designs have favored a slightly different philosophy about screening designs, where the experiment is considered more of a “one-shot,” “pick-the-winner” approach to identify the optimal settings of the factors, followed up by several confirmatory runs at the predicted optimum.

CONCLUSION

In the pharmaceutical industry, the most concentrated use of screening designs, factorial designs, and response surface designs is found in the product and process development functions, including analytical method development, because of the increasing complexity in the optimization and validation of analytical methods. In the evolving environment of rapid and shortened development cycles, screening designs and experimental design in general

will continue to gain acceptance as valuable development tools.

REFERENCES

1. Plackett, R.L.; Burman, J.P. The design of optimal multifactorial experiments. *Biometrika* **1946**, *33*, 305–325.
2. Taguchi, G. Reports of statistical application research. *JUSE* **1960**, *6*, 1–52.

Seamless Adaptive Trial Design: Dose Selection Criteria

Jiayin Zheng and Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

In pharmaceutical/clinical development, commonly considered two-stage seamless adaptive designs include a two-stage phase I/II or phase II/III adaptive trial that combines one phase IIb study for dose finding or treatment selection and one phase III study for efficacy confirmation into a single study. At the end of stage 1, promising dose(s) will be selected based on pre-specified selection criteria. In practice, since there is little power with limited subjects available at interim, commonly considered selection criteria for critical decision-making include (i) conditional power, (ii) precision analysis, (iii) predictive probability of success, and (iv) probability of being the best dose or treatment. The selected promising dose(s) will then proceed to the next stage for efficacy confirmation. In this article, we introduce, compare, and evaluate these criteria. Besides, we attempt to address some concerns for two-stage seamless adaptive clinical trial.

Keywords: Dose selection; Efficacy; Interim analysis; Probability of being the best dose; Seamless adaptive design.

INTRODUCTION

In pharmaceutical/clinical development, a two-stage seamless adaptive trial design that combines two individual studies into a single study is often considered. Commonly considered two-stage seamless adaptive designs include a two-stage phase I/II or phase II/III adaptive trial that combines one phase IIb study for dose finding or treatment selection and one phase III study for efficacy confirmation into a single study. Thus, a two-stage seamless adaptive trial design consists of two stages: dose finding or treatment selection (stage 1) and efficacy confirmation (stage 2). The study objective at the first stage is for dose finding or treatment selection, whereas the study objective at stage 2 is to confirm the efficacy of the dose or treatment selected from the first stage compared to a control (dose or treatment). If the study endpoints employed at different stages are the same, the two-stage seamless adaptive trial design lead to a typical two-stage adaptive design with different study objectives at different stages as described in the work of Chow and Lin.^[1]

For a two-stage seamless adaptive trial design, at the first stage, qualified subjects are usually randomly assigned to receive a dose or treatment at a 1:1 ratio. At the end of stage 1, promising dose(s) will be selected based on pre-specified selection criteria. In practice, since there is little power with limited subjects available at interim, commonly considered selection criteria for critical decision-making include (i) precision analysis,

(ii) predictive probability of success, and (iii) probability of being the best dose or treatment.^[1,2] The selected promising dose(s) will then proceed to the next stage for efficacy confirmation. Based on the results observed at interim, some adaptations such as adaptive randomization or sample size re-estimation are often applied for achieving the study objectives and/or increasing the probability of success.

In practice, for clinical studies utilizing two-stage seamless adaptive trial designs, the following questions are commonly asked by the regulatory agencies such as the United States Food and Drug Administration. First, it is a concern whether the overall type I error rate is well controlled at a pre-specified level of significance. Second, it is a concern that the selection criteria at stage 1 may wrongly select inferior dose groups based on the limited number of subjects available at interim. Third, it is a concern that the interim data may deviate far from expected (i.e., the assumptions made at the beginning of study). In this research, we attempt to address these concerns for two-stage seamless adaptive clinical trial.

Criteria that are commonly employed for dose selection at the first stage are briefly described in Section “Criteria for Dose Selection.” The performances of these selection criteria in terms of the coverage probability of correctly selecting the promising dose(s) are evaluated through the conduct of extensive clinical trial simulation in Section “Practical Implementation and Example.” Section “Simulation Studies” illustrates the application of these selection

criteria based on single primary study endpoint as well as co-primary study endpoints. Concluding remarks are given in Section “Conclusion.”

CRITERIA FOR DOSE SELECTION

Consider the dose-finding stage in a clinical trial. Assume there are k test doses and one control dose to investigate in this stage. Denote the measurement means corresponding to the k doses by $D_1, D_2, \dots, D_k$ and that of the control dose by D_0 . We are interested in the difference $d_i = D_i - D_0$, $i = 1, \dots, k$. Without generality, assume the larger of this difference, the better effect of the corresponding dose. Denote X_{ij} , $i = 0, 1, \dots, k$, $j = 1, \dots, N_i$, where N_i is the sample size of the i th dose, as the measurement of the j th subject with the i th dose. For each dose, assume X_{ij} , $j = 1, \dots, N_i$ are independent and identically distributed, and follow a normal distribution with an unknown mean D_i and an unknown

where $q_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution and the pooled sample standard deviation

$$\tilde{S}_i' = \sqrt{\frac{(n_i - 1)S_i'^2/n_i + (n_0 - 1)S_0'^2/n_0}{n_i + n_0 - 2}}.$$

Conditional Power

At interim, denote the observed effect size by

$$\hat{\delta}_i' = \frac{\bar{X}_i' - \bar{X}_0'}{\sqrt{S_i'^2 + S_0'^2}} \text{ for dose } i.$$

Take $\hat{\delta}_i'$ (or $(\bar{X}_i', \bar{X}_0', S_i'^2, S_0'^2)$) as the true value of the effect size (or means of dose i and control dose, and variances of dose i and control dose, respectively), and calculate the power with the pre-specified sample size, denoted as p_i^{power} . p_i^{power} can be expressed as

$$pr \left\{ \frac{\bar{X}_i' - \bar{X}_0'}{\sqrt{\frac{(N_i - 1)S_i'^2}{N_i} + \frac{(N_0 - 1)S_0'^2}{N_0}}} > q_{\alpha} \mid (D_i, D_0, \sigma_i^2, \sigma_0^2) = (\bar{X}_i', \bar{X}_0', S_i'^2, S_0'^2) \right\},$$

variance σ_i^2 . Denote $\bar{X}_i$ as the sample mean and S_i^2 as the sample variance for dose i . Based on the total sample size of $N = \sum_{i=1}^k N_i$, we can construct the $1 - \alpha$ confidence interval of the mean d_i as follows:

$$(L_i, U_i) = (\bar{X}_i - \bar{X}_0 - q_{\alpha/2} \tilde{S}_i, \bar{X}_i - \bar{X}_0 + q_{\alpha/2} \tilde{S}_i),$$

where $q_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution and the pooled sample standard deviation

$$\tilde{S}_i = \sqrt{\frac{(N_i - 1)S_i^2/N_i + (N_0 - 1)S_0^2/N_0}{N_i + N_0 - 2}}.$$

Assume that at interim, the data from $n = \sum_{i=1}^k n_i$ subjects are available, with n_i subjects available for each dose. Denote $\bar{X}_i'$ as the sample mean and $S_i'^2$ as the sample variance for dose i at interim. Then similarly, we have the $1 - \alpha$ confidence interval of the mean d_i :

$$(L_i', U_i') = (\bar{X}_i' - \bar{X}_0' - q_{\alpha/2} \tilde{S}_i', \bar{X}_i' - \bar{X}_0' + q_{\alpha/2} \tilde{S}_i'),$$

where q_{α} is the upper α quantile of the standard normal distribution. Noncentral t distribution can be used to calculate this probability.

Select the dose with the largest p_i^{power} . If $N_i, i = 0, 1, \dots, k$ are equal, then this criterion is equivalent to choosing the one having the largest observed effect size $\hat{\delta}_i'$.

Precision Analysis Based on Confidence Interval

Denote (L_i^+, U_i^+) as the positive part of (L_i', U_i') . Define $p_i^+ = pr\{d_i \in (L_i^+, U_i^+)\}$ as the confidence level for achieving statistical significance. If $L_i' \geq 0$, $p_i^+ = 1 - \alpha$. If $U_i' < 0$, $p_i^+ = 0$. If (L_i', U_i') contains 0, then $p_i^+ = pr\{d_i \in (0, U_i')\}$, and we derive the exact expression of p_i^+ in this case. Denote

$w_i = \Phi\left(\frac{\bar{X}_i' - \bar{X}_0'}{\tilde{S}_i'}\right)$, where Φ is the cumulative distribution function of standard normal distribution. Then $p_i^+ = w_i - \alpha/2$.

Based on $p_i^+, i = 1, \dots, k$, select the dose having the highest p_i^+ .

Predictive Probability of Success

Assume that the difference observed at interim, $\bar{X}_i', i = 1, \dots, k$, preserved to the end of the study. We calculate

the predictive probability of success $pr\{d_i > d\}$ at interim, where d is a prespecified value. Denote $p_i^s = pr\{d_i > d\}$. We have the mean difference $\bar{X}_i - \bar{X}_0$ and the pooled sample standard deviation

$$\tilde{S}_i^p = \sqrt{\frac{(N_i - 1)S_i'^2/N_i + (N_0 - 1)S_0'^2/N_0}{N_i + N_0 - 2}}$$

based on the data obtained at interim. Then

$$p_i^s = \Phi\left(\frac{\bar{X}_i' - \bar{X}_0' - d}{\tilde{S}_i^p}\right),$$

$$p_i^{\text{power}} = pr\left\{\frac{\bar{X}_i - \bar{X}_0}{\sqrt{\frac{(N_i - 1)S_i^2}{N_i} + \frac{(N_0 - 1)S_0^2}{N_0}}} > q_\alpha \mid (D_i, D_0, \sigma_i^2, \sigma_0^2) = (\bar{X}_i', \bar{X}_0', S_i'^2, S_0'^2)\right\}.$$

where Φ is the cumulative distribution function of standard normal distribution. We may select the dose with the highest p_i^s .

Probability of Being the Best Dose

Calculate $p_i^b = pr\{\max_i(d_i) = d_i\}$, which is the probability of being the best dose, based on the interim data. Taking $X_{0j}, j = 1, \dots, N_0$, as fixed values, we can assume $(\bar{X}_i, S_i^2)$ is statistically independent with $(\bar{X}_{i_2}, S_{i_2}^2)$ for any $i_1 > i_2$. It also holds for

$$(\bar{X}_i', S_i'^2), i = 1, \dots, k.$$

We have

$$p_i^b = \int \frac{1}{\tilde{S}_i'} \phi\left(\frac{z - (\bar{X}_i' - \bar{X}_0')}{\tilde{S}_i'}\right) \prod_{g=1, i \neq g}^k \Phi\left(\frac{z - (\bar{X}_g' - \bar{X}_0')}{\tilde{S}_g'}\right) dz,$$

where ϕ and Φ are the probability density function and the cumulative distribution function of standard normal distribution, respectively. Similar to the predictive probability of success, $\tilde{S}_i'$ and $\tilde{S}_g'$ in the above expression of p_i^b can be replaced with $\tilde{S}_i^p$ and $\tilde{S}_g^p$.

Thus, we may select the dose with the largest p_i^b .

PRACTICAL IMPLEMENTATION AND EXAMPLE

In this section, we describe the practical implementation under the case of single primary endpoint and the case of co-primary endpoint. A real example will be given to illustrate these methods.

Single Primary Endpoint

As introduced in the previous section, we summarize the criteria for dose selection under the case of single primary endpoint as follows:

Conditional power—Select the dose having the largest p_i^{power} .

Precision analysis based on confidence interval—Select the dose having the highest p_i^* : if $L_i' \geq 0$, $p_i^* = 1 - \alpha$; if

$$U_i' < 0, p_i^* = 0; \text{ if } (L_i', U_i') \text{ contains } 0, p_i^* = \Phi\left(\frac{\bar{X}_i' - \bar{X}_0'}{\tilde{S}_i'}\right)$$

$$- \alpha/2, \text{ where } \tilde{S}_i' = \sqrt{\frac{(n_i - 1)S_i'^2/n_i + (n_0 - 1)S_0'^2/n_0}{n_i + n_0 - 2}}.$$

Predictive probability of success—Select the dose

with the highest p_i^s : $p_i^s = \Phi\left(\frac{\bar{X}_i' - \bar{X}_0' - d}{\tilde{S}_i^p}\right)$, where

$$\tilde{S}_i^p = \sqrt{\frac{(N_i - 1)S_i'^2/N_i + (N_0 - 1)S_0'^2/N_0}{N_i + N_0 - 2}} \text{ and } d \text{ is a pre-}$$

specified value.

Probability of being the best dose—Select the dose with the largest p_i^b :

$$p_i^b = \int \frac{1}{\tilde{S}_i'} \phi\left(\frac{z - (\bar{X}_i' - \bar{X}_0')}{\tilde{S}_i'}\right) \prod_{g=1, i \neq g}^k \Phi\left(\frac{z - (\bar{X}_g' - \bar{X}_0')}{\tilde{S}_g'}\right) dz.$$

Co-primary Endpoints

Assume under the case of co-primary endpoint, the type I error allocated to each endpoint is equal (i.e., $\alpha/2$). Denote the measurement means corresponding to the k doses by $D_1, D_2, \dots, D_k$, and that of the control dose by D_0 for the first

endpoint. Similarly, denote those for the second endpoint by $E_1, E_2, \dots, E_k$ and E_0 . We are interested in the differences $d_i = D_i - D_0$ and $g_i = E_i - E_0$, $i = 1, \dots, k$. Without generality, assume the larger this difference, the better effect of the corresponding dose. Denote X_{ij} , $i = 0, 1, \dots, k$, $j = 1, \dots, N_i$, where N_i is the sample size of the i th dose, as the measurement of the j th subject with the i th dose for the first endpoint. Similarly, denote Y_{ij} as the measurement of the j th subject with the i th dose for the second endpoint. For each dose, assume X_{ij} , $j = 1, \dots, N_i$ are independent and identically distributed, and follow a normal distribution with an unknown mean D_i and an unknown variance σ_{1i}^2 . Similarly, assume Y_{ij} , $j = 1, \dots, N_i$ are independent and identically distributed, and follow a normal distribution with an unknown mean E_i and an unknown variance σ_{2i}^2 . Denote $\bar{X}_i$ and $\bar{Y}_i$ as the sample means and S_i^2 and Q_i^2 as the sample variances for dose i . Based on the total sample size of $N = \sum_{i=1}^k N_i$, we can construct the $1 - \alpha$ confidence intervals of the mean d_i and g_i as follows:

$$(L_{1i}, U_{1i}) = (\bar{X}_i - \bar{X}_0 - q_{\alpha/2} \tilde{S}_i, \bar{X}_i - \bar{X}_0 + q_{\alpha/2} \tilde{S}_i),$$

$$(L_{2i}, U_{2i}) = (\bar{Y}_i - \bar{Y}_0 - q_{\alpha/2} \tilde{Q}_i, \bar{Y}_i - \bar{Y}_0 + q_{\alpha/2} \tilde{Q}_i),$$

Assume that at interim, the data from $n = \sum_{i=1}^k n_i$ subjects are available, with n_i subjects available for each dose. Denote $\bar{X}'_i$ and $\bar{Y}'_i$ as the sample means and $S_i'^2$ and $Q_i'^2$ as the sample variances for dose i at interim. Then similarly, we have the $1 - \alpha$ confidence interval of the means d_i and g_i

$$(L'_{1i}, U'_{1i}) = (\bar{X}'_i - \bar{X}'_0 - q_{\alpha/2} \tilde{S}'_i, \bar{X}'_i - \bar{X}'_0 + q_{\alpha/2} \tilde{S}'_i),$$

$$(L'_{2i}, U'_{2i}) = (\bar{Y}'_i - \bar{Y}'_0 - q_{\alpha/2} \tilde{Q}'_i, \bar{Y}'_i - \bar{Y}'_0 + q_{\alpha/2} \tilde{Q}'_i),$$

where $q_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution and the pooled sample standard deviations

$$\tilde{S}'_i = \sqrt{\frac{(n_i - 1)S_i'^2/n_i + (n_0 - 1)S_0'^2/n_0}{n_i + n_0 - 2}} \text{ and}$$

$$\tilde{Q}'_i = \sqrt{\frac{(n_i - 1)Q_i'^2/n_i + (n_0 - 1)Q_0'^2/n_0}{n_i + n_0 - 2}}.$$

Then we have the criteria for dose selection under the case of co-primary endpoint as follows:

Conditional power—Select the dose with the largest p_i^{power} .

$$p_i^{\text{power}} = \left\{ \begin{array}{l} \frac{\bar{X}_i - \bar{X}_0}{\sqrt{\frac{(N_i - 1)S_i^2 + (N_0 - 1)S_0^2}{N_i + N_0 - 2}}} > q_{\alpha/2}, \frac{\bar{Y}_i - \bar{Y}_0}{\sqrt{\frac{(N_i - 1)Q_i^2 + (N_0 - 1)Q_0^2}{N_i + N_0 - 2}}} > q_{\alpha/2} \\ \left| (D_i, D_0, \sigma_{1i}^2, \sigma_{10}^2) = (\bar{X}'_i, \bar{X}'_0, S_i'^2, S_0'^2), (E_i, E_0, \sigma_{2i}^2, \sigma_{20}^2) = (\bar{Y}'_i, \bar{Y}'_0, Q_i'^2, Q_0'^2) \right\} \end{array} \right.$$

where $q_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution and the pooled sample standard deviations

$$\tilde{S}_i = \sqrt{\frac{(N_i - 1)S_i^2/N_i + (N_0 - 1)S_0^2/N_0}{N_i + N_0 - 2}} \text{ and}$$

$$\tilde{Q}_i = \sqrt{\frac{(N_i - 1)Q_i^2/N_i + (N_0 - 1)Q_0^2/N_0}{N_i + N_0 - 2}}.$$

Precision analysis based on confidence interval—Select the dose having the highest p_i^+ :

$$\begin{aligned} p_i^+ &= \text{pr}\{d_i \in (L_{1i}^+, U_{1i}^+), g_i \in (L_{2i}^+, U_{2i}^+)\} \\ &= \text{pr}\{d_i \in (\max(0, L_{1i}^+), \max(0, U_{1i}^+)), \\ &\quad g_i \in (\max(0, L_{2i}^+), \max(0, U_{2i}^+))\}, \end{aligned}$$

where (L_{1i}^+, U_{1i}^+) and (L_{2i}^+, U_{2i}^+) as the positive parts of (L_{1i}', U_{1i}') and (L_{2i}', U_{2i}') , respectively. Under the assumption of the statistical independence between X_{ij} and Y_{ij} , we can express p_i^+ in detail as follows:

If $L_{2i}' \geq 0$, and $L_{1i}' \geq 0$, $p_i^+ = (1 - \alpha)^2$.

If $L_{2i}' \geq 0$ and (L_{1i}', U_{1i}') contains 0, $p_i^+ = (1 - \alpha)^* \left(\Phi \left(\frac{\bar{X}_i' - \bar{X}_0'}{\tilde{S}_i'} \right) - \alpha/2 \right)$, where

$$\tilde{S}_i' = \sqrt{\frac{(n_i - 1)S_i'^2/n_i + (n_0 - 1)S_0'^2/n_0}{n_i + n_0 - 2}}.$$

If $L_{1i}' \geq 0$ and (L_{2i}', U_{2i}') contains 0, $p_i^+ = (1 - \alpha)^* \left(\Phi \left(\frac{\bar{Y}_i' - \bar{Y}_0'}{\tilde{Q}_i'} \right) - \alpha/2 \right)$, where

$$\tilde{Q}_i' = \sqrt{\frac{(n_i - 1)Q_i'^2/n_i + (n_0 - 1)Q_0'^2/n_0}{n_i + n_0 - 2}}.$$

If both (L_{1i}', U_{1i}') and (L_{2i}', U_{2i}') contain 0, $p_i^+ = \left(\Phi \left(\frac{\bar{X}_i' - \bar{X}_0'}{\tilde{S}_i'} \right) - \frac{\alpha}{2} \right) * \left(\Phi \left(\frac{\bar{Y}_i' - \bar{Y}_0'}{\tilde{Q}_i'} \right) - \frac{\alpha}{2} \right)$.

Otherwise, $p_i^+ = 0$.

Predictive probability of success—Select the dose with the highest p_i^s : $p_i^s = pr\{d_i > d, g_i > g\}$. Under the assumption of the statistical independence between X_{ij} and Y_{ij} , we can express

$$p_i^s = \Phi \left(\frac{\bar{X}_i' - \bar{X}_0' - d}{\tilde{S}_i^p} \right) \Phi \left(\frac{\bar{Y}_i' - \bar{Y}_0' - g}{\tilde{Q}_i^p} \right)$$

where $\tilde{S}_i^p = \sqrt{\frac{(N_i - 1)S_i'^2/N_i + (N_0 - 1)S_0'^2/N_0}{N_i + N_0 - 2}}$, $\tilde{Q}_i^p = \sqrt{\frac{(N_i - 1)Q_i'^2/N_i + (N_0 - 1)Q_0'^2/N_0}{N_i + N_0 - 2}}$, and d, g are prespecified values.

Probability of being the best dose—Select the dose with the largest p_i^b :

$$\begin{aligned} p_i^b &= pr\{max_i(d_i) = d_i, max_i(g_i) = g_i\} \\ &= \int \frac{1}{\tilde{S}_i'} \phi \left(\frac{z - (\bar{X}_i' - \bar{X}_0')}{\tilde{S}_i'} \right) \prod_{g=1, i \neq g}^k \Phi \left(\frac{z - (\bar{X}_g' - \bar{X}_0')}{\tilde{S}_g'} \right) dz, \\ &\quad \int \frac{1}{\tilde{Q}_i'} \phi \left(\frac{w - (\bar{Y}_i' - \bar{Y}_0')}{\tilde{Q}_i'} \right) \prod_{h=1, i \neq h}^k \Phi \left(\frac{w - (\bar{Y}_h' - \bar{Y}_0')}{\tilde{Q}_h'} \right) dw, \end{aligned}$$

under the assumption of the statistical independence between X_{ij} and Y_{ij} .

An Example

A clinical study was intended to be designed as a two-stage phase II/III seamless, adaptive, randomized, double-blind, parallel-group, placebo-controlled, multicenter study to evaluate the efficacy, safety, and dose-response of a compound in treating postmenopausal women and adjuvant breast cancer patients with vasomotor symptoms (VMSs). VMSs are often referred to as hot flashes, flushes, and night sweats, sometimes accompanied by shivering and a sense of cold. The primary study endpoints would be frequency and severity of VMS evaluated at weeks 4 and 12 as co-primary endpoints. Since the real data for this study are unavailable, we provide this example with numerical simulated data to illustrate the use of the dose-selection methods, based on such studies.

Under this two-stage adaptive trial design, the first stage is a phase II trial for dose selection. The second stage is a phase III trial for confirming the efficacy and safety of the compound. An interim analysis will be conducted to evaluate the dose effect and to review the efficacy and safety effects when about 50% subjects at stage 1 complete week 4 of double-blind treatment period. One or two optimal dose(s) will be selected and the sample size will be reestimated. The second stage of the study starts after the interim analysis. Newly enrolled subjects will be randomized to receive the selected optimal dose(s) or placebo for 12 weeks. The second stage will serve as a pivotal study.

For frequency, at week 4, each study subject reported the number of VMSs each day, and those numbers were averaged for each subject. For severity, each subject was asked to rate how much they were bothered by VMS at the end of week 4 using scores from 1 to 10 with higher value indicating more severity. Three doses and one placebo were compared. Eight hundred subjects were recruited into the study with 200 in each group. At interim, each group had 100 subjects having completed week 4 of the treatment period. The logarithm transformation of frequency and the original scale of severity were treated as following normal distributions and were used for interim analysis. Using the method of probability of being the best dose, which was shown to be the best one in the simulation studies, we selected dose 3, which had the largest overall probability of being the best dose. The results were shown with means and variances of the two co-primary endpoints.

SIMULATION STUDIES

In this section, we conducted the simulation studies to investigate the performances of the four methods described in Section “Criteria for Dose Selection.”

Single Primary Endpoint

Assume the mean effects of four doses (dose 1 as the placebo) are to be compared, and all effects follow normal distributions. Different parameter settings were given to investigate the performances of the methods. Dose 4 was always set to be the most effective dose among the four

doses compared. We consider the following total sample sizes: 100, 150, and 200. The proportion of interim sample size was set to be 20%, resulting in interim sample sizes of 20, 30, and 40, respectively. For each case, 5000 repetitions were implemented. We summarized the simulation results as the frequency of being selected as the most effective dose in [Tables 1–3](#), according to difference sample sizes.

Table 1 Results of the example

		Placebo	Dose 1	Dose 2	Dose 3
Endpoint 1	Sample mean	3.37	3.42	3.07	2.56
	Sample std	1.61	2.1	1.62	1.53
	Probability of being the best		0	0.011	0.989
Endpoint 2	Sample mean	6.74	6.17	5.94	5.93
	Sample std	1.74	1.76	1.89	1.87
	Probability of being the best		0.07	0.45	0.479
Overall probability of being the best			0	0.005	0.474

std = standard deviation

Table 2 Results of simulations for single primary endpoint when the total sample size is 100

N = 100												
	Dose 1	Dose 2	Dose 3	Dose 4	Dose 1	Dose 2	Dose 3	Dose 4	Dose 1	Dose 2	Dose 3	Dose 4
Mean	0.5	0.51	0.52	0.53	0.5	0.51	0.525	0.53	0.5	0.505	0.51	0.53
Criterion	std	Probability of being selected			std	Probability of being selected			std	Probability of being selected		
C1	0.005	0.33	0.33	0.33	0.005	0.33	0.33	0.33	0.005	0.18	0.41	0.41
C2		0.33	0.33	0.33		0.33	0.33	0.33		0.32	0.34	0.34
C3		0.33	0.33	0.33		0.33	0.33	0.33		0.23	0.38	0.38
C4		0.00	0.00	1.00		0.00	0.01	0.99		0.00	0.00	1.00
C1	0.01	0.19	0.41	0.41	0.01	0.19	0.41	0.41	0.01	0.03	0.27	0.70
C2		0.33	0.34	0.34		0.33	0.34	0.34		0.22	0.38	0.40
C3		0.24	0.38	0.38		0.24	0.38	0.38		0.06	0.33	0.62
C4		0.00	0.00	1.00		0.00	0.06	0.94		0.00	0.00	1.00
C1	0.02	0.03	0.28	0.69	0.02	0.03	0.42	0.56	0.02	0.01	0.04	0.95
C2		0.22	0.38	0.40		0.22	0.39	0.39		0.13	0.28	0.59
C3		0.05	0.34	0.61		0.05	0.43	0.51		0.02	0.08	0.90
C4		0.00	0.06	0.94		0.00	0.21	0.79		0.00	0.00	1.00
C1	0.03	0.02	0.18	0.80	0.03	0.02	0.34	0.64	0.03	0.01	0.03	0.96
C2		0.16	0.35	0.49		0.16	0.40	0.44		0.10	0.21	0.69
C3		0.03	0.22	0.75		0.03	0.36	0.61		0.01	0.05	0.94
C4		0.01	0.14	0.85		0.01	0.30	0.69		0.00	0.02	0.98
C1	0.05	0.07	0.26	0.68	0.05	0.06	0.36	0.58	0.05	0.05	0.10	0.86
C2		0.13	0.31	0.56		0.13	0.39	0.48		0.11	0.17	0.72
C3		0.07	0.26	0.67		0.06	0.37	0.57		0.05	0.10	0.85
C4		0.06	0.25	0.69		0.05	0.36	0.58		0.04	0.10	0.86
C1	0.1	0.17	0.30	0.53	0.1	0.15	0.38	0.48	0.1	0.15	0.23	0.62
C2		0.19	0.31	0.50		0.17	0.38	0.46		0.17	0.24	0.59
C3		0.17	0.30	0.53		0.15	0.38	0.48		0.15	0.23	0.62
C4		0.17	0.30	0.53		0.14	0.38	0.48		0.15	0.22	0.63

(Continued)

Table 2 Results of simulations for single primary endpoint when the total sample size is 100 (*Continued*)

<i>N</i> = 100												
	Dose 1	Dose 2	Dose 3	Dose 4	Dose 1	Dose 2	Dose 3	Dose 4	Dose 1	Dose 2	Dose 3	Dose 4
Mean	0.5	0.51	0.52	0.53	0.5	0.51	0.525	0.53	0.5	0.505	0.51	0.53
Criterion	std	Probability of being selected			std	Probability of being selected			std	Probability of being selected		
C1	0.2	0.24	0.33	0.43	0.2	0.23	0.36	0.41	0.2	0.24	0.28	0.47
C2		0.25	0.33	0.42		0.23	0.36	0.41		0.25	0.29	0.46
C3		0.24	0.33	0.43		0.23	0.36	0.41		0.24	0.28	0.47
C4		0.24	0.33	0.43		0.22	0.37	0.41		0.24	0.28	0.48

C1: conditional power

C2: precision analysis based on confidence interval

C3: Predictive probability of success

C4: Probability of being the best dose

std = standard deviation

Table 3 Results of simulations for single primary endpoint when the total sample size is 150

<i>N</i> = 150												
	Dose 1	Dose 2	Dose 3	Dose 4	Dose 1	Dose 2	Dose 3	Dose 4	Dose 1	Dose 2	Dose 3	Dose 4
Mean	0.5	0.51	0.52	0.53	0.5	0.51	0.525	0.53	0.5	0.505	0.51	0.53
Criterion	std	Probability of being selected			std	Probability of being selected			std	Probability of being selected		
C1	0.005	0.33	0.33	0.33	0.005	0.33	0.33	0.33	0.005	0.27	0.37	0.37
C2		0.33	0.33	0.33		0.33	0.33	0.33		0.33	0.33	0.33
C3		0.33	0.33	0.33		0.33	0.33	0.33		0.30	0.35	0.35
C4		0.00	0.00	1.00		0.00	0.00	1.00		0.00	0.00	1.00
C1	0.01	0.26	0.37	0.37	0.01	0.26	0.37	0.37	0.01	0.05	0.37	0.58
C2		0.33	0.33	0.33		0.33	0.33	0.33		0.26	0.37	0.37
C3		0.30	0.35	0.35		0.30	0.35	0.35		0.09	0.41	0.51
C4		0.00	0.00	1.00		0.00	0.03	0.97		0.00	0.00	1.00
C1	0.02	0.05	0.37	0.58	0.02	0.05	0.46	0.49	0.02	0.01	0.08	0.91
C2		0.26	0.37	0.37		0.26	0.37	0.37		0.15	0.33	0.52
C3		0.09	0.40	0.51		0.08	0.45	0.47		0.02	0.13	0.85
C4		0.00	0.03	0.97		0.00	0.16	0.84		0.00	0.00	1.00
C1	0.03	0.02	0.20	0.78	0.03	0.02	0.36	0.62	0.03	0.00	0.03	0.97
C2		0.19	0.37	0.44		0.18	0.40	0.42		0.11	0.23	0.66
C3		0.04	0.25	0.71		0.04	0.39	0.57		0.01	0.05	0.94
C4		0.00	0.10	0.90		0.00	0.26	0.74		0.00	0.01	0.99
C1	0.05	0.04	0.22	0.73	0.05	0.03	0.35	0.63	0.05	0.02	0.07	0.91
C2		0.14	0.32	0.55		0.13	0.39	0.48		0.10	0.18	0.73
C3		0.05	0.23	0.72		0.03	0.35	0.62		0.02	0.08	0.90
C4		0.04	0.21	0.75		0.02	0.34	0.64		0.02	0.07	0.91
C1	0.1	0.15	0.29	0.56	0.1	0.12	0.38	0.50	0.1	0.12	0.19	0.69
C2		0.18	0.31	0.52		0.15	0.38	0.47		0.15	0.21	0.64
C3		0.15	0.29	0.56		0.12	0.38	0.50		0.12	0.19	0.69
C4		0.15	0.29	0.56		0.12	0.37	0.51		0.12	0.19	0.70

(Continued)

Table 3 Results of simulations for single primary endpoint when the total sample size is 150 (*Continued*)

<i>N</i> = 150												
	Dose 1	Dose 2	Dose 3	Dose 4	Dose 1	Dose 2	Dose 3	Dose 4	Dose 1	Dose 2	Dose 3	Dose 4
Mean	0.5	0.51	0.52	0.53	0.5	0.51	0.525	0.53	0.5	0.505	0.51	0.53
Criterion	std	Probability of being selected			std	Probability of being selected			std	Probability of being selected		
C1	0.2	0.23	0.31	0.46	0.2	0.20	0.38	0.42	0.2	0.22	0.26	0.51
C2		0.24	0.31	0.45		0.22	0.37	0.41		0.23	0.27	0.50
C3		0.23	0.31	0.46		0.20	0.38	0.42		0.23	0.26	0.51
C4		0.23	0.31	0.46		0.20	0.38	0.42		0.22	0.26	0.52

C1: conditional power

C2: precision analysis based on confidence interval

C3: Predictive probability of success

C4: Probability of being the best dose

std = standard deviation

From the simulation results, we observed that the last method “probability of being the best dose” performed better in all cases than other methods, with higher probability of choosing the right best dose and lower probability of choosing the wrong dose.

Co-primary Endpoints

Assume the mean effects of the co-primary endpoints of four doses (dose 1 as the placebo) are to be compared, and all effects follow normal distributions. Different parameter settings were given to investigate the performances of the methods. In the first scenario, dose 4 was set to be the most effective dose among the four doses compared, with both co-primary endpoints being the most effective.

In the second scenario, the first endpoint of dose 4 is most effective, whereas the second endpoint of dose 3 is most effective. We consider the following total sample sizes: 100 and 200. The proportion of interim sample size was set to be 20%, resulting in interim sample sizes of 20 and 40, respectively. For each case, 5000 repetitions were implemented. We summarized the simulation results as the frequency of being selected as the most effective dose in [Tables 4–6](#), according to difference sample sizes.

From the simulation results, we observed that the last method “probability of being the best dose” performed better in all cases than other methods, with higher probability of choosing the right best dose and lower probability of choosing the wrong dose.

Table 4 Results of simulations for single primary endpoint when the total sample size is 200

<i>N</i> = 200												
	Dose 1	Dose 2	Dose 3	Dose 4	Dose 1	Dose 2	Dose 3	Dose 4	Dose 1	Dose 2	Dose 3	Dose 4
Mean	0.5	0.51	0.52	0.53	0.5	0.51	0.525	0.53	0.5	0.505	0.51	0.53
Criterion	std	Probability of being selected			std	Probability of being selected			std	Probability of being selected		
C1	0.005	0.33	0.33	0.33	0.005	0.33	0.33	0.33	0.005	0.31	0.35	0.35
C2		0.33	0.33	0.33		0.33	0.33	0.33		0.33	0.33	0.33
C3		0.33	0.33	0.33		0.33	0.33	0.33		0.32	0.34	0.34
C4		0.00	0.00	1.00		0.00	0.00	1.00		0.00	0.00	1.00
C1	0.01	0.31	0.35	0.35	0.01	0.31	0.35	0.35	0.01	0.07	0.43	0.50
C2		0.33	0.33	0.33		0.33	0.33	0.33		0.28	0.36	0.36
C3		0.32	0.34	0.34		0.32	0.34	0.34		0.12	0.42	0.46
C4		0.00	0.00	1.00		0.00	0.01	0.99		0.00	0.00	1.00
C1	0.02	0.07	0.42	0.50	0.02	0.08	0.46	0.47	0.02	0.01	0.10	0.89
C2		0.28	0.36	0.36		0.28	0.36	0.36		0.17	0.35	0.48
C3		0.12	0.42	0.46		0.12	0.44	0.44		0.03	0.16	0.81
C4		0.00	0.01	0.99		0.00	0.13	0.87		0.00	0.00	1.00

(Continued)

Table 4 Results of simulations for single primary endpoint when the total sample size is 200 (*Continued*)

<i>N</i> = 200												
	Dose 1	Dose 2	Dose 3	Dose 4	Dose 1	Dose 2	Dose 3	Dose 4	Dose 1	Dose 2	Dose 3	Dose 4
Mean	0.5	0.51	0.52	0.53	0.5	0.51	0.525	0.53	0.5	0.505	0.51	0.53
Criterion	std	Probability of being selected			std	Probability of being selected			std	Probability of being selected		
C1	0.03	0.02	0.25	0.72	0.03	0.02	0.40	0.58	0.03	0.01	0.03	0.96
C2		0.21	0.38	0.41		0.21	0.39	0.40		0.12	0.26	0.61
C3		0.05	0.31	0.64		0.04	0.42	0.54		0.01	0.07	0.92
C4		0.00	0.07	0.93		0.00	0.23	0.77		0.00	0.00	1.00
C1	0.05	0.02	0.20	0.77	0.05	0.02	0.34	0.64	0.05	0.01	0.04	0.94
C2		0.14	0.33	0.53		0.14	0.39	0.47		0.09	0.18	0.73
C3		0.03	0.22	0.75		0.03	0.35	0.62		0.02	0.05	0.93
C4		0.02	0.18	0.80		0.01	0.32	0.67		0.01	0.04	0.95
C1	0.1	0.12	0.28	0.60	0.1	0.10	0.38	0.52	0.1	0.10	0.15	0.75
C2		0.16	0.30	0.53		0.14	0.39	0.47		0.14	0.19	0.67
C3		0.12	0.28	0.60		0.10	0.38	0.52		0.10	0.15	0.74
C4		0.12	0.28	0.60		0.10	0.38	0.52		0.10	0.15	0.75
C1	0.2	0.21	0.32	0.47	0.2	0.19	0.36	0.45	0.2	0.20	0.25	0.55
C2		0.23	0.32	0.46		0.20	0.36	0.44		0.22	0.25	0.53
C3		0.21	0.32	0.47		0.19	0.36	0.45		0.20	0.25	0.55
C4		0.21	0.32	0.47		0.19	0.37	0.45		0.20	0.25	0.55

C1: conditional power

C2: precision analysis based on confidence interval

C3: Predictive probability of success

C4: Probability of being the best dose

std = standard deviation

Table 5 Results of simulations for co-primary endpoints and scenario 1

Scenario 1									
	Dose 1	Dose 2	Dose 3	Dose 4		Dose 1	Dose 2	Dose 3	Dose 4
Mean1	0.5	0.51	0.52	0.53	Mean2	0.5	0.51	0.52	0.53
<i>N</i> = 100					<i>N</i> = 200				
std1	std 2	Probability of being selected			std1	std 2	Probability of being selected		
0.03	0.03	0.005	0.129	0.866	0.03	0.03	0.002	0.121	0.878
		0.052	0.294	0.654			0.099	0.399	0.502
		0.007	0.148	0.846			0.007	0.214	0.780
		0.002	0.067	0.931			0.000	0.019	0.981
0.03	0.05	0.030	0.202	0.767	0.03	0.05	0.009	0.148	0.843
		0.054	0.262	0.684			0.066	0.315	0.619
		0.031	0.205	0.764			0.011	0.170	0.819
		0.007	0.112	0.882			0.000	0.047	0.953
0.03	0.1	0.099	0.301	0.600	0.03	0.1	0.074	0.284	0.642
		0.104	0.314	0.582			0.091	0.314	0.595
		0.103	0.303	0.594			0.075	0.286	0.638
		0.016	0.155	0.828			0.002	0.098	0.900

(Continued)

Table 5 Results of simulations for co-primary endpoints and scenario 1 (*Continued*)

Scenario 1									
	Dose 1	Dose 2	Dose 3	Dose 4		Dose 1	Dose 2	Dose 3	Dose 4
Mean1	0.5	0.51	0.52	0.53	Mean2	0.5	0.51	0.52	0.53
N = 100					N = 200				
std1	std 2	Probability of being selected			std1	std 2	Probability of being selected		
0.05	0.05	0.044	0.221	0.734	0.05	0.05	0.013	0.149	0.838
		0.054	0.241	0.706			0.044	0.255	0.701
		0.045	0.223	0.732			0.014	0.155	0.831
		0.024	0.187	0.790			0.004	0.097	0.899
0.05	0.1	0.098	0.278	0.625	0.05	0.1	0.059	0.263	0.678
		0.100	0.281	0.619			0.072	0.277	0.651
		0.103	0.278	0.619			0.061	0.264	0.675
		0.055	0.233	0.712			0.017	0.181	0.802
0.1	0.1	0.144	0.292	0.564	0.1	0.1	0.082	0.273	0.645
		0.143	0.294	0.563			0.084	0.277	0.639
		0.145	0.296	0.559			0.084	0.276	0.640
		0.120	0.292	0.589			0.063	0.242	0.696

Mean1: mean of the first endpoint; mean2: mean of the second endpoint; std1: std of the first endpoint; std2: std of the second endpoint. std = standard deviation

Table 6 Results of simulations for co-primary endpoints and scenario 2

Scenario 1									
	Dose 1	Dose 2	Dose 3	Dose 4		Dose 1	Dose 2	Dose 3	Dose 4
Mean1	0.5	0.51	0.52	0.53	Mean2	0.5	0.51	0.52	0.53
N = 100					N = 200				
std1	std 2	Probability of being selected			std1	std 2	Probability of being selected		
0.03	0.03	0.005	0.129	0.866	0.03	0.03	0.002	0.121	0.878
		0.052	0.294	0.654			0.099	0.399	0.502
		0.007	0.148	0.846			0.214	0.780	
		0.002	0.067	0.931			0.981		
0.03	0.05	0.030	0.202	0.767	0.03	0.05	0.009	0.148	0.843
		0.054	0.262	0.684			0.066	0.315	0.619
		0.031	0.205	0.764			0.011	0.170	0.819
		0.007	0.112	0.882			0.000	0.047	0.953
0.03	0.1	0.099	0.301	0.600	0.03	0.1	0.074	0.284	0.642
		0.104	0.314	0.582			0.091	0.314	0.595
		0.103	0.303	0.594			0.075	0.286	0.638
		0.016	0.155	0.828			0.002	0.098	0.900
0.05	0.05	0.044	0.221	0.734	0.05	0.05	0.013	0.149	0.838
		0.054	0.241	0.706			0.044	0.255	0.701
		0.045	0.223	0.732			0.014	0.155	0.831
		0.024	0.187	0.790			0.004	0.097	0.899
0.05	0.1	0.098	0.278	0.625	0.05	0.1	0.059	0.263	0.678
		0.100	0.281	0.619			0.072	0.277	0.651
		0.103	0.278	0.619			0.061	0.264	0.675
		0.055	0.233	0.712			0.017	0.181	0.802

(Continued)

Table 6 Results of simulations for co-primary endpoints and scenario 2 (*Continued*)

Scenario 1									
	Dose 1	Dose 2	Dose 3	Dose 4		Dose 1	Dose 2	Dose 3	Dose 4
Mean1	0.5	0.51	0.52	0.53	Mean2	0.5	0.51	0.52	0.53
N = 100					N = 200				
std1	std 2	Probability of being selected			std1	std 2	Probability of being selected		
0.1	0.1	0.144	0.292	0.564	0.1	0.1	0.082	0.273	0.645
		0.143	0.294	0.563			0.084	0.277	0.639
		0.145	0.296	0.559			0.084	0.276	0.640
		0.120	0.292	0.589			0.063	0.242	0.696

Mean1: mean of the first endpoint; mean2: mean of the second endpoint; std1: std of the first endpoint; std2: std of the second endpoint. std = standard deviation

CONCLUSION

In this article, we discussed the problem of dose finding in two-stage adaptive clinical trials with the aim of selecting the most effective dose for further investigation of the next stage. Four criteria for dose selection were introduced and compared through extensive simulation studies. From the simulation results, the method of probability of being the best dose performed better than all other three methods. Thus, we recommend this method in practical applications as appropriate. In addition, an example of doseresponse of compound in treating postmenopausal women and

adjuvant breast cancer patients with vasomotor symptoms was given to illustrate the use of this method.

REFERENCES

1. Chow, S.C.; Lin, M. Analysis of two-stage adaptive seamless trial design. *Pharmaceutica Analytica Acta* **2015**, *6*, 3. doi:10.4172/2153-2435.1000341.

2. Lee, S.J.; Lin, M. Adaptive trial design—Case studies. Presented at the 2016 Duke-Industry Statistics Symposium (DISS), Durham, NC, September 14, 2016.

Semi-Parametric Time-Varying Regression Models

Thomas H. Scheike

Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark

Abstract

This entry reviews statistical models that are designed for estimating such time-varying effects and testing various hypotheses about them.

Keywords: Aalen model; Additive intensity model; Cox-Aalen model; Extended Cox model; Non-proportional hazards model; Semi-parametric modeling; Survival data; Time-varying effects.

INTRODUCTION

In biomedical applications, there will often be important time-varying effects. A typical example is a treatment effect that varies over time, and two important examples are where (i) treatment efficacy fades away over time owing to, for example, tolerance developed by the patient, or in the case of infectious diseases owing to resistance developed by the targeted bacteria and (ii) a treatment with many side effects will lead to an initial adverse effect that is compensated by a beneficial effect for those surviving the initial phase.

This entry considers statistical models that are aimed at estimating such time-varying effects and testing various hypotheses about them.^[1]

The class of models depends, of course, on the response considered, and this entry considers time to event data (survival analysis), although there also exist a similar development in the context of longitudinal data. For both classes of data, the aim is to estimate the effect of a possibly time-varying treatment effect while correcting for other covariates.

FLEXIBLE MODELING OF SURVIVAL DATA

In the survival context, we shall look at models where given a p dimensional covariate $X(t)$ that may vary over time (including treatment status) and an at risk indicator $Y(t)$ the intensity is

$$\lambda(t, X) = Y(t)h(X(t))\alpha(t) \quad (1)$$

where $h(t) = t$ or $h(t) = \exp(t)$. Here, $\alpha(t)$ gives the time-varying effects of the covariates on as excess risk when $h(t) = t$ or as time-varying multiplicative effect when $h(t) = \exp(t)$. We will illustrate the use of these models below.

Many traditional statistical models do not have the needed flexibility to estimate and make hypothesis testing about time-varying effects. In the survival analysis context, the standard regression model is Cox's regression model^[2] where effects are assumed to lead to constant relative risk, something that is violated for many biomedical studies. Even when a treatment effect does not lead to constant relative risk, the Cox model may still provide an interesting summary statistic. However, to provide a more detailed understanding of what is going on and the summary statistic provided by the Cox model, this entry aims to present a more detailed analysis.

There have been some attempts to extend the Cox model in order to deal with time-varying effects leading to Eq. (1) when $h(t) = \exp(t)$ as described, for example.^[3–6] There is still, however, some way to a fully satisfactory apparatus to deal with time-varying effects owing to the need for a smoothing parameter. Also, some covariates will have effects that are not well described as being multiplicative.

One important alternative to the proportional hazards model is Aalen's additive hazard model ($h(t) = t$).^[7–9] Aalen's additive hazard regression model is completely non-parametric and includes covariates additively in the model thus leading to an excess risk interpretation. The Aalen model is very flexible and will estimate all the covariate effects as time-varying effects.

The flexible models given by Eq. (1) will estimate the treatment effect and other prognostic factors as time varying. This is often to "big" a model that does not provide the simplest possible summary of the data. A model that only uses the flexibility where it is needed is the semi-parametric model

$$\lambda(t) = Y(t)(X^T(t)\beta(t) + Z^T(t)\gamma) \quad (2)$$

where $\beta(t)$ is a q_1 -dimensional time-varying regression coefficient, γ is a q_2 -dimensional regression coefficient,

and $(X(t), Z(t))$ is a set of $p = q_1 + q_2$ -dimensional covariate vector that is observable at time t . For this model, some effects are time-varying and other covariates lead to constant effects. In this model, important hypotheses are whether or not an effect is significantly time varying or if an effect is significant at all.

The semi-parametric version of the model additive risk model Eq. (2) was suggested by Refs. [10,11] for extensions and a procedure for hypothesis testing in the model. One advantage of the additive models is that time-varying effects are easy to estimate (with explicit estimators) and that inferential procedures for making conclusions about the time-varying effects exist.

The multiplicative version of Eq. (2) was suggested in Ref. [4]. In this model, it is also possible to test for time-varying effects,^[6] but one practical problem is that one needs to choose smoothing parameters to deal with the time-varying effects.

The additive and multiplicative models postulate different relationships between the hazard and covariates, and it is seldom clear which of the models should be preferred. The models may often be used to complement each other and to provide different summary measures. Sometimes, however, covariate effects are best modeled as multiplicative and other covariate effects are best modeled as being additive, and then one must combine the additive and multiplicative models. One useful model for combining additive and multiplicative effects is the Cox-Aalen model where

$$\lambda(t) = Y(t)(X^T(t)\beta(t)\exp(Z^T(t)\gamma)) \quad (3)$$

where some effects lead to multiplicative effects and other covariates lead to time-varying additive effects; this model is easy to fit just like the semi-parametric version of the additive risk model.^[12–13]

APPLICATIONS

Prognostic Effects for Acute Myocardial Infarction Data

Thetrandolapril cardiac evaluation trial (TRACE) study group^[14] has recorded information on 6,633 consecutive patients with myocardial infarction with the aim of studying the prognostic importance of various risk factors on mortality. We here consider a random sample of 2,000 of these patients. At the time of entry, the patients had various risk factors recorded such as age (median = 69 year and range = 23–99 year), gender [male = 1 (67%) and female = 0], congestive heart failure (CHF) [present = 1 (54%) and not present = 0], and ventricular fibrillation (VF) [present = 1 (7%) and not present = 0]. Some risk factors were expected to have strongly time-varying effects, in particular VF. The timescale used in our analysis is time since myocardial infarction. The total number of deaths in a three-year period

after entering the study was 2,351, and of these, 840 took place within the first month. Most of the cases within the first month resulted from conditions related to the presence of VF.

To illustrate the semi-parametric modeling approach, we start by fitting an additive risk model with the covariates age, gender, CHF, and VF. This leads to the following cumulative effects with 95% confidence bands (Hall-Wellner bands) (Fig. 1).

The cumulative coefficients clearly indicate that the excess risk is varying with time (the slope of the cumulatives). Formal testing for time-varying effects leads to p -values at 0.01 for age, 0.35 for sex, $p < 0.00001$ for CHF, and $p < 0.00001$ for VF. The VF and CHF conditions are strongly time varying. Age was borderline time varying, and comparing the cumulative to the cumulative in the end point divided with time (a simple estimate of its constant effect) leads to the following test process that is compared to resampled processes under the null in Fig. 2. A similar plot was produced for gender.

The plot reveals that although VF and CHF account for a great deal of time-varying effects, there is a slight effect left on age. Overall, a reasonable summary is to allow age and sex to have constant effects. Note that although age deviates considerably from the null performance in the initial phase, it has only quite minor effects even on the age scale. The plot indicates that the effect of age is slightly higher initially compared to later on.

Now, fitting a semi-parametric risk model with age and sex leading to constant excess risk (SD) yields 0.005 (0.00036) for age and 0.013 (0.009) for sex. Thus males have slightly higher risk than females, although a non-significant difference, and with a 10-year age difference resulting in an excess risk at 0.05.

By considering various additional goodness-of-fit statistics, we can see that the model provides a quite reasonable fit.

It might be preferable, however, to consider a slightly different model where age and sex lead to multiplicative effects, thus resulting in a modification of the excess risk for VF and CHF depending on age and gender. We therefore subtract the mean from age and fitted the Cox-Aalen survival model with VF and CHF as flexible additive effects and with age and gender as multiplicative effects. This leads to the additive effects depicted in Fig. 3. These are quite similar to the similar additive effects of the semi-parametric risk model, although CHF appears to lead to somewhat lower excess risk in the Cox-Aalen model.

The relative risk of were age 0.058 (0.004) and gender 0.141 (0.071), age leading to increased risk of VF and CHF and gender being borderline significant with males having increased risk. We have shown how flexible models can be applied and will give information additional to that of a standard survival model. The considered models are quite flexible, and one can obtain different summary measures from different models. Sometimes one needs to combine additive and multiplicative effects.

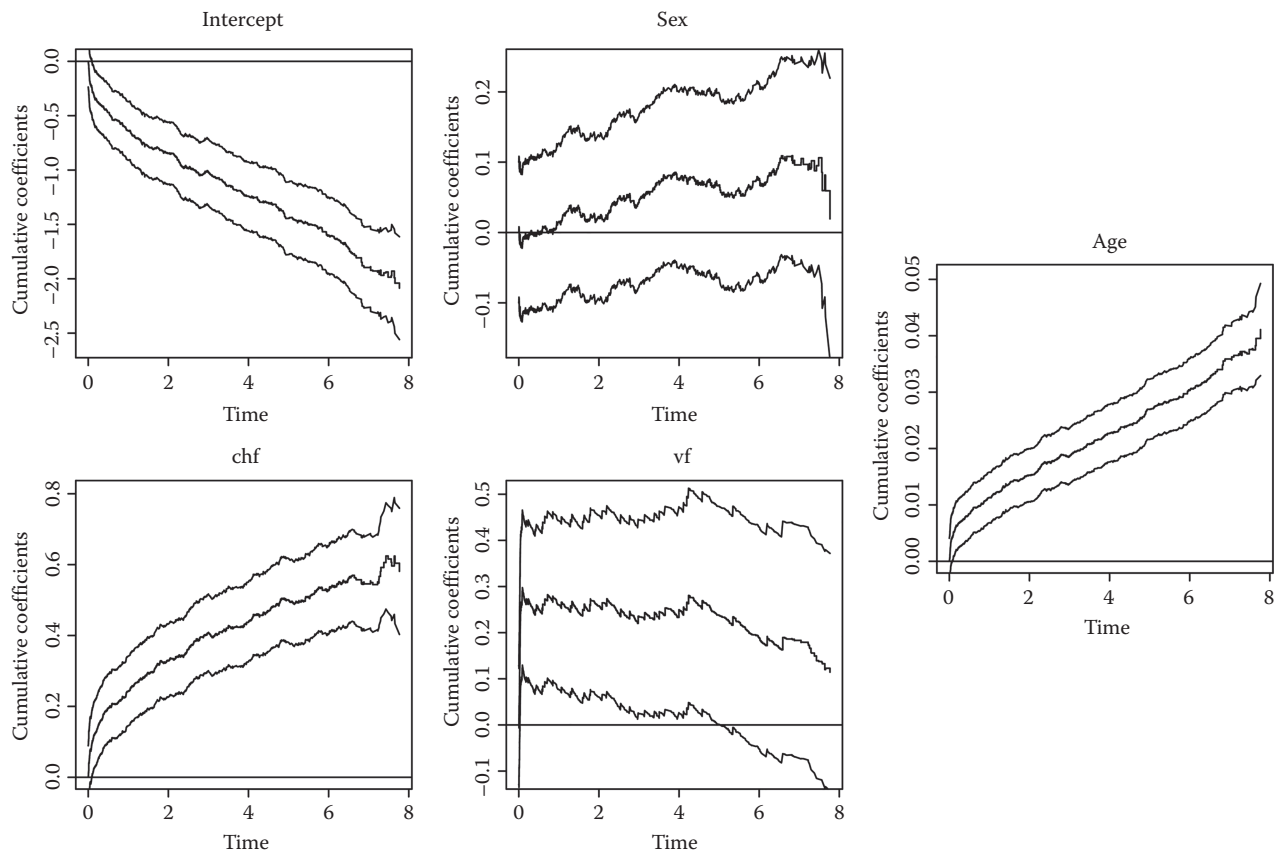


Fig. 1 Cumulative covariate effects for additive risk model. Cumulative excess risk estimates with 95 % Hall-Wellner confidence bands for TRACE data

Treatment Effect for CSL1 Study

In this section, we applied the semi-parametric model to data from the CSL1 study conducted by the Copenhagen Study Group for Liver Diseases (CSL).^[15] This was

a randomized clinical trial where the patients were given either prednisone or placebo. During the period from 1962 to 1969, 532 patients with histologically verified liver cirrhosis were included in the trial. The data is described in detail further in Ref. [16]. In 488 patients, the initial biopsy

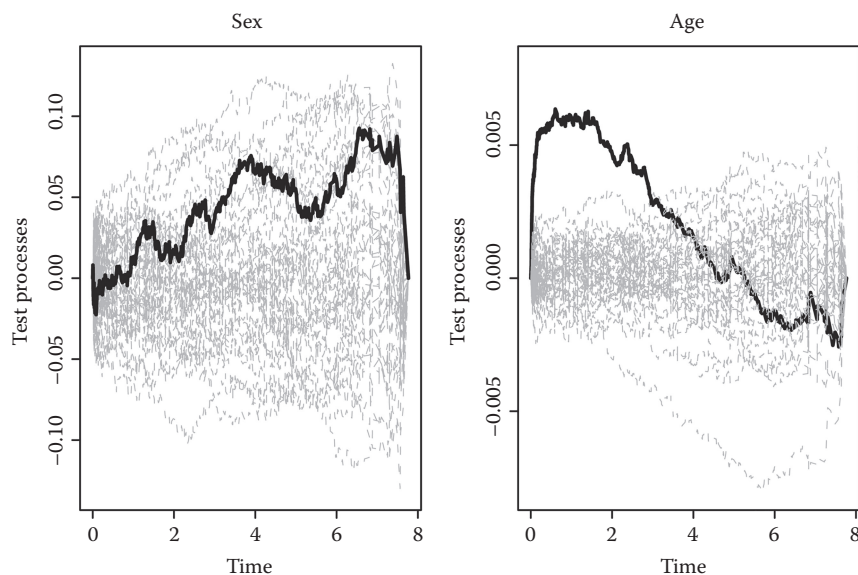


Fig. 2 Test processes for test of time-constant effects (thick line) and 50 random processes under the null of constant effects

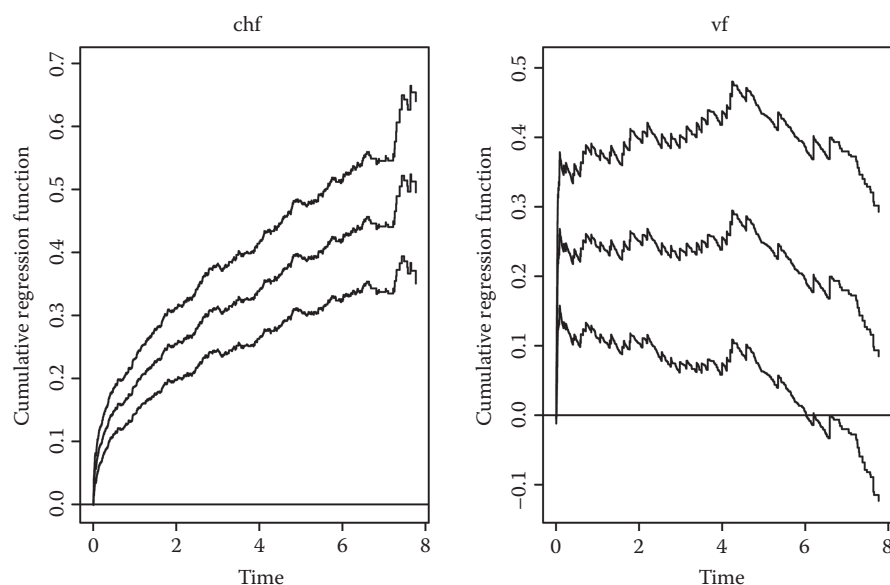


Fig. 3 Cumulative excess risk estimates with 95 % pointwise confidence intervals for Cox-Aalen model for TRACE data

could later be reevaluated, and we considered only data for these patients. The patients were followed up from the date of entry into the trial to death or the closing date of the study, September 1, 1974. A number of clinical and biochemical variables were registered during the study period. We considered only gender (males = 1/females = 0), age, and the treatment (prednisone = 1).

We started by considering the flexible multiplicative extension of the Cox regression model. This led to cumulative effects estimated using bandwidth equal to three years and local linear fitting to avoid edge effects. These are shown in Fig. 4.

Tests for time-varying effects indicated that all variables were well described as being constant. Prior to the study, it was suspected that the treatment effect might wear off during the study. After successive testing of time-varying effects, we therefore considered a model where age and gender had constant effect. These constant effects were age 0.032 (0.003) and gender 0.330 (0.059). This led to a time-varying effect of treatment similar to the one depicted in Fig. 4 that shows a treatment effect that has some initial effect that wears off. The estimate in this model depends on the smoothing parameter, and we provide below an alternative description of the

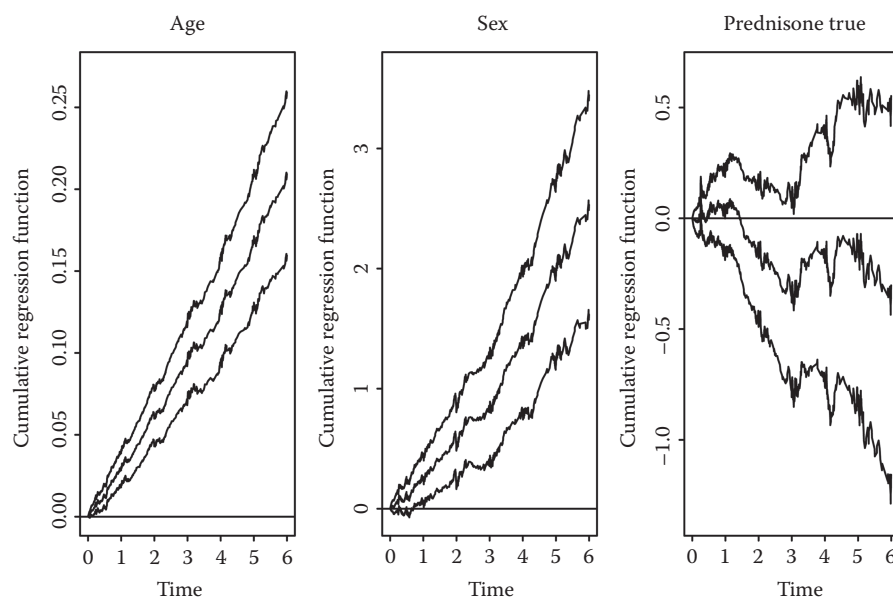


Fig. 4 Cumulative relative risk estimates for Cox regression model with time-varying effects for CSL data

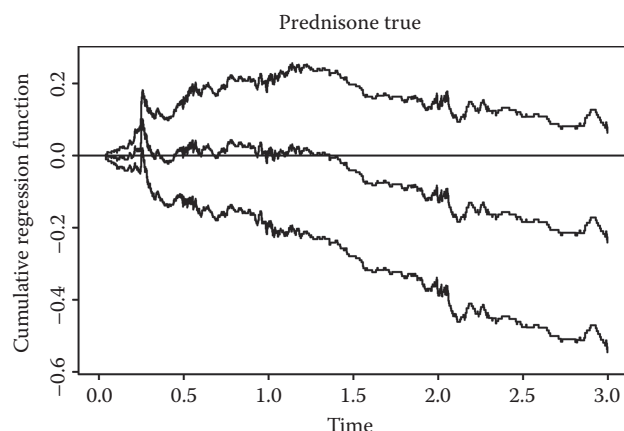


Fig. 5 Cumulative excess risk estimates for treatment with 95% pointwise confidence intervals based on Cox-Aalen model for CSL data

treatment effect using the Cox-Aalen model. A reasonable summary appears to be given by the standard Cox regression, and this leads to effects for age and gender, which are similar to those given just above, and a relative risk treatment effect at -0.045 (0.057)—in conclusion, a weakly non-significant protective effect.

Alternatively one may summarize the treatment effect by computing its excess risk in an additive model that is equal to -0.039 (0.051). Excess risk can be translated into increased survival as a function of time by $\exp(-0.039 \times t)$, thus implying, for example, that the five-year survival is 18 % better for subjects getting prednisone.

Finally, we take a look at the estimate of the time-varying effect of prednisone using the Cox-Aalen survival model where gender and sex have multiplicative effects. Fig. 5 shows the effect of prednisone on the excess risk scale. This gives pretty much the same overall impression as the estimate on the relative risk scale. One advantage, however, is that this estimate is independent of any smoothing parameter.

If we consider the time-interval from one to three years, only the treatment effect becomes borderline significant.

SOFTWARE

Software to fit the models using R (Splus) is available in form of the R-package *timereg* that can be downloaded from <http://cran.r-project.org/web/packages/timereg/>.

CONCLUSIONS

This entry demonstrated that there do exist models for studying more closely time dynamics of covariate effects. In the context of survival data, these flexible models can provide a more detailed and more biologically sensible

description of the underlying data. This is of practical relevance in many biopharmaceutical studies where treatment effects are expected to be time varying. We have considered three different model classes here for obtaining a description of the time dynamics: extended Cox models, additive risk models, and the Cox-Aalen model, which is a flexible combination of the Cox and additive Aalen model.

REFERENCES

1. Martinussen, T.; Scheike, T.H. *Dynamic regression models for survival data*; Springer-Verlag: New York, 2006.
2. Cox, D.R. Regression models and life tables (with discussion). *J. R. Stat. Soc. B.* **1972**, *34*, 187–220.
3. Zucker, D.M.; Karr, A.F. Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *Ann. Stat.* **1990**, *18*, 329–353.
4. Murphy, S.A.; Sen, P.K. Time-dependent coefficients in a Cox-type regression model. *Stochastic Proc. Appl.* **1991**, *39*, 153–180.
5. Martinussen, T.; Scheike, T.H.; Skovgaard, I.M. Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models. *Scand. J. Stat.* **2002**, *28*, 57–74.
6. Scheike, T.H.; Martinussen, T. On efficient estimation and tests of time-varying effects in the proportional hazards model. *Scand. J. Stat.* **2004**, *31*, 51–62.
7. Aalen, O.O. A Model for Non-parametric Regression Analysis of Counting Processes. In *Lecture Notes in Statistics-2: Mathematical Statistics and Probability Theory*; Springer-Verlag: New York, 1980, 1–25.
8. Aalen, O.O. A linear regression model for the analysis of life times. *Stat. Med.* **1989**, *8*, 907–925.
9. Aalen, O.O. Further results on the non-parametric linear regression model in survival analysis. *Stat. Med.* **1993**, *12*, 1569–1588.
10. McKeague, I.W.; Sasieni, P.D. A partly parametric additive risk model. *Biometrika* **1994**, *81*, 501–514.
11. Scheike, T.H. The additive nonparametric and semiparametric Aalen model as the rate function for a counting process. *Lifetime Data Anal.* **2002**, *8*, 247–262.
12. Scheike, T.H.; Zhang, M.-J. Extensions and applications of the Cox-Aalen survival model. *Biometrics* **2003**, *59*, 1033–1045.
13. Scheike, T.H.; Zhang, M.-J. An additive-multiplicative Cox-Aalen model. *Scand. J. Stat.* **2002**, *28*, 75–88.
14. Jensen, G.V.; Torp-Pedersen, C.; Hildebrandt, P.; Kober, L.; Nielson, F.E.; Melchior, T.; Joen, T.; Andersen, P.K. Does in-hospital ventricular fibrillation affect prognosis after myocardial infarction? *Eur. Heart J.* **1997**, *18*, 919–924.
15. Schlichting, P.; Christensen, E.; Andersen, P.; Fauerholdt, L.; Juhl, E.; Poulsen, H.; Tygstrup, N. The Copenhagen study group for liver diseases. *Hepatology* **1983**, *3*, 889–895.
16. Andersen, P.K.; Borgan, Ø; Gill, R.D.; Keiding, N. *Statistical Models Based on Counting Processes*; Springer-Verlag: New York, 1993.

Sequential Estimation: Additive Hazards Rate Model with Staggered Entry

Laurent Bordes

*Laboratoire de Mathématiques Appliquées, Université de Pau et des Pays de l'Adour,
Pau, France*

Christelle Breuils

Département STID, IUT de Metz, Metz, France

Abstract

This entry proposes estimators of the additive hazards rate model in the presence of right-censoring and staggered entries and also touches on the problems of sequential estimations.

Keywords: Additive hazards model; Continuous-time martingales; Empirical processes; Sequential estimation Staggered entry; Survival analysis.

INTRODUCTION

In semiparametric survival studies that include covariates, it is often assumed that covariates have a multiplicative effect on the hazard rate function leading to the famous Cox Proportional Hazard Model (PHM). However, several studies have proved that the hazard proportionality can fail, and thus, alternative models are necessary. The Additive Hazards Model (AHM), like the PHM, allows easy interpretation of covariate effects. It assumes that conditionally, on its covariate, an individual has a hazard function equal to an unknown baseline hazard rate function plus a function of a covariate, which depends on an unknown regression parameter, similarly for the PHM. Another interest of the AHM is that there exist explicit estimators for the unknown parameter of the model, which is no longer true for the PHM, and more generally, for almost all semiparametric lifetime regression models.

In this entry, we propose estimators of the AHM in the presence of right-censoring and staggered entries. That is, we observe lifetimes of two types. Either it can be the duration of interest or it can be a censoring time, in which case we only know that the duration of interest is larger than the duration observed. This is the right-censoring mechanism. Moreover, we want to note that individuals are recruited at different times (staggered entry). Therefore, at the time of study, the duration of interest can be censored by the time lapsed entry time and the calendar time.

Parameter estimation accounting staggered entry is often referred as sequential analysis in the survival analysis literature. In the last decade, attention has been paid to derive specific methods to evaluate large sample properties of sequential estimators. These works always assume that,

for a fixed study interval, the number of observations can be arbitrarily large. In this entry, we show that if we assume that the entry process is a Poisson process with parameter growing to infinity, then the large sample assumption is equivalent to a large value for the parameter of the entry process and that the asymptotic behavior of the regression parameter is unchanged, even if the estimator is indexed by a random sample size.

Finally, we discuss the problem of sequential estimations, that is, to define an earlier end to the study (stopping time) without deteriorating the quality of estimators (e.g., confidence regions with fixed width).

PARAMETERS ESTIMATION

Regression Models

The problem is knowing when to stop the study with a prior fixed precision. This implies that the model is known and that the sample is not determined by a clinical plan or a quantity of observed events. (These techniques do not find applications in comparative clinical trials.) Thus, data are collected to estimate the regression parameter of the model (in order to determinate the influence of the covariates in the lifetime variable). This is the case when we aim to estimate a parameter for a well-known model. It is worth noticing that the end of the study is fully determined by the sequential procedure and we have no a priori techniques for it, contrary to common planned clinical trials.

Because patients are recruited in clinical trials at different times with respect to the calendar time (staggered entry), and because the time of interest is the time in trial,

it may be important to perform analysis of some characteristics of the trial time monitored by the calendar time. Such sequential analysis could allow the clinical trials to stop at the earliest time where information seems definitively sufficient to give clinical conclusions. In the presence of covariates, such analysis can be done by using PHM (see Sellke and Siegmund,^[1] Biliias et al.^[2]). In the same spirit, Tsiatis^[3] considered a large class of time-sequential tests for two sample survival data (for large sample properties see Slud^[4] and Gu and Lai^[5]). As we said in the introduction, the PHM is the most commonly used model to account covariates effects on lifetime distributions. It assumes that a duration T_i observed with a covariate Z , admits conditionally on Z at time t , a hazard rate function $\lambda(t; Z)$ defined by:

$$\lambda(t; Z) = \lambda_0(t) \exp(\beta_0^T Z(t)), \quad \forall t \in R^+,$$

where $\lambda_0(t)$ is an unknown baseline hazard rate function and β_0 an unknown regression parameter in R^p where p is the number of covariates, that is the size of Z (which can depend on time). In presence of right censoring Cox^[6] proposed an estimator of the regression parameter β_0 based on partial likelihood. One of the main advantages of PHM is that simple information on the sign of β_0 allows easy interpretation of covariate effects. However, as noted by Huffer and McKeague,^[7] the proportionality assumption on hazards can failed and then it is necessary to propose alternative models. An alternative for the PHM is the AHM proposed by Cox and Oakes^[8] which assumes that:

$$\lambda(t; Z) = \lambda_0(t) + \beta_0^T Z(t), \quad \forall t \in R^+$$

where again $\lambda_0(t)$ and β_0 are, respectively, an unknown baseline hazard rate function and an unknown regression parameter. In this model, the covariates effect is additive (again the value of β_0 allows an easy interpretation of covariates effect). Another basic lifetime regression model is the Accelerated Life Model (ALM) generalized by Lin and Ying^[9] to time-dependent covariates. It assumes that the cumulative hazard rate function $\Lambda(t; Z)$, conditionally to a covariate Z , is defined by:

$$\Lambda(t; Z) = \int_0^t \lambda(s; Z) ds = \int_0^t \exp(\beta_0^T Z(s)) \lambda_0(s) ds, \quad \forall t \in R^+,$$

where $\lambda_0(t)$ and β_0 are still unknown baseline hazard rate function and regression parameter, respectively. In this model, the covariate effects are accounted by modifying the time scale factor. There are numbers of generalizations of these basic models, see, for example Bagdonavicius, and Nikulin,^[10] Lin and Ying^[21] and Lin et al.^[22], but most of these generalizations do not allow easy interpretations of covariate effects. In the sequel we will focus our interest on the AHM for which generalizations were also proposed in Bordes.^[11]

Data Description

Suppose that we observe a n -sample $(\tau_i, X_i, \delta_i, Z_i)_{1 \leq i \leq n}$, where i denotes the i th individual, τ_i is an entry time of individual i (the time at which the individual is recruited in the study, the time at which the treatment begins, etc.), X_i denotes the minimum between the duration of interest T_i (time elapsed between the entry date and the date of the event of interest) and a censoring time C_i (time elapsed between the entry date and the date of another event than the event of interest), δ_i is a binary variable equal to 1 if X_i is equal to the time of interest T_i and 0 otherwise, and Z_i is a vector of covariates that may be time-varying; if so, $Z_i = \{Z_i(s); 0 \leq s \leq X_i\}$. With such observations, two time scales have to be considered: the calendar time (t) and the duration on study (s). A typical way to represent staggered entry is to use a Lexis diagram as it is shown on Fig. 1 above.

However, when the time elapsed from the beginning of study is t , the information available for the i -th individual is:

$$\begin{cases} X_i(t) = \min(X_i, (t - \tau_i)^+) \\ \delta_i(t) = I(T_i = X_i(t)), \end{cases}$$

where $a^+ = \max(a, 0)$, $I(A) = 1$, if A is true and 0 otherwise. It follows that the duration of interest T_i may be censored either by C_i or by $(t - \tau_i)^+$ which is the duration between the date of entry and the date at which the study is done.

Estimators for the AHM

In this section we assume that for each individual, the hazard rate function of the i th individual can be written:

$$\lambda_i(t) = \lambda(t; Z_i) = \lambda_0(t) + \beta_0^T Z_i(t),$$

where λ_0 is an unknown baseline hazard rate function and β_0 is an unknown regression parameter. Quantities to be estimated are, β_0 (in order to evaluate the covariates effects) and e.g. $\Lambda(t; Z)$ the cumulative hazard function conditionally to the covariate Z and $S(t; Z)$ the survival function conditionally to a covariate Z . Following the methodology of Lin and Ying,^[12] we propose to estimate β_0 at calendar time t and after a duration of study of s by:

$$\hat{\beta}_n(t, s) = \left(\sum_{i=1}^n \int_0^{\min(s, X_i(t))} (Z_i(u) - \bar{Z}(t, u))^{\otimes 2} du \right)^{-1} \left(\sum_{i=1}^n (Z_i(X_i(t)) - \bar{Z}(t, X_i(t))) N_i(t, s) \right),$$

where if a is a column vector $a^{\otimes 2} = a^T a$, $N_i(t, s) = I(X_i(t) \leq s; \delta_i(t) = 1)$ and

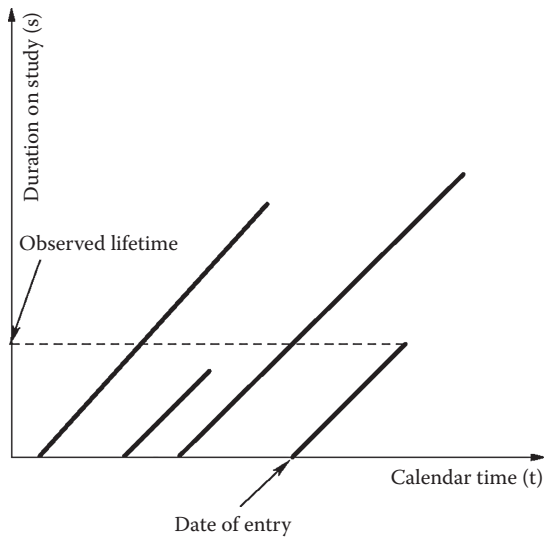


Fig. 1 Lexis diagram

$$\bar{Z}(t, s) = \frac{\sum_{i=1}^n \gamma_i(t, s) Z_i(s)}{\bar{Y}(t, s)} \quad \text{and} \\ \bar{Y}(t, s) = \sum_{i=1}^n \gamma_i(t, s).$$

If the origin of the calendar time is defined at the beginning of the study, at calendar time t , β_0 is estimated by $\hat{\beta}_n(t, t)$. In this case, for $s \in [0, t]$ the estimator of $\Lambda(s; Z)$ is defined by:

$$\hat{\Lambda}_n(t, s; Z) = \sum_{i=1}^n \frac{N_i(t, s)}{\bar{Y}(t, X_i(t))} - \hat{\beta}_n(t, t) \\ \times \int_0^s (\bar{Z}(t, u) - Z(u)) du.$$

Therefore, the estimator of the survival function $S(t; Z)$ is given for $s \in [0, t]$ by:

$$\hat{S}_n(t, s; Z) = \exp \left(\hat{\beta}_n^T(t, t) \int_0^s (\bar{Z}(t, u) - Z(u)) du \right) \\ \times \prod_{\substack{1 \leq i \leq n; \\ N_i(t, s) = 1}} \left(\frac{\bar{Y}(t, X_i(t)) - 1}{\bar{Y}(t, X_i(t))} \right).$$

Large Sample Properties

Several authors show that standard counting process theory (see Andersen et al.^[13] or Fleming and Harrington^[20]) cannot be applied to obtain large sample properties of the above estimators with respect to the two time scales (see, e.g., Sellke and Siegmund^[1]). Biliás et al.^[2] developed a general empirical process approach (see Pollard^[14]) that allows deriving large sample properties for the PHM

estimators with staggered entry. This approach had been applied to the AHM by Bordes^[15] proving that under standard regularity assumptions, quantities:

$$\sqrt{n}(\hat{\beta}_n(t, s) - \beta_0), \\ \sqrt{n}(\hat{\Lambda}_n(t, s; Z) - \Lambda(s; Z))$$

and

$$\sqrt{n}(\hat{S}_n(t, s; Z) - S(s; Z))$$

converge weakly to Gaussian random field indexed by $(t, s) \in D = \{(u, v) \in \mathbb{R}^+; t_* \leq v \leq u \leq t^*\}$ where t_* and t^* are, respectively, the time at which starts the study (not recruitments) and the maximum duration of the study. Let us now give estimators of asymptotic variances of each of the previous estimators. We have:

$$\text{Var}(\hat{\beta}_n(t, s)) \approx \hat{A}^{-1}(t, s) \times \hat{B}(t, s) \times \hat{A}^{-1}(t, s),$$

where

$$\hat{A}(t, s) = \left(\sum_{i=1}^n \int_0^{\min(s, X_i(t))} (Z_i(u) - \bar{Z}(t, u))^{\otimes 2} du \right),$$

and

$$\hat{B}(t, s) = \sum_{i=1}^n (Z_i(X_i(t)) - \bar{Z}(t, X_i(t)))^{\otimes 2} N_i(t, s).$$

$$\text{Var}(\hat{\Lambda}_n(t, s; Z)) \approx \sum_{i=1}^n \frac{N_i(t, s)}{\bar{Y}^2(t, X_i(t))} \\ + 2 \left\{ \sum_{i=1}^n \frac{(Z_i(X_i(t)) - \bar{Z}(t, X_i(t)))^T}{\bar{Y}(t, X_i(t))} N_i(t, s) \right\} \\ \times \hat{A}^{-1}(t, s) \hat{\mu}(t, s) + \hat{\mu}^T(t, s) \hat{A}^{-1}(t, s) \\ \times \hat{B}(t, s) \times \hat{A}^{-1}(t, s) \hat{\mu}(t, s).$$

and

$$\text{Var}(\hat{S}_n(t, s; Z)) \approx \hat{S}_n^2(t, s; Z) \text{Var}(\hat{\Lambda}_n(t, s; Z)).$$

SEQUENTIAL STOPPING RULES

Because the human or financial cost of a survival study can increase with the time, it is necessary to find procedures that allow having an early stop of the experiments. Thus, if we stop the study when a required precision level on estimation is reached, we have to define the so called stopping rules (see Ghosh et al.^[16]). All this stopping rules lead to obtain the best compromise between an accurate estimation and a early end of the study. Thus, the goal of the sequential stopping rule is to provide the required sample size N_α for which a fixed level of precision $\alpha > 0$

is obtained. Sequential estimation based on fixed width confidence region as defined e.g., by Dmitrienko and Govindarajulu^[17] are the most commonly used stopping rules. As a consequence of these methods, the sample size is generally random, and it increases with the precision level.

With such a stochastic sample size, large sample properties of the regression estimator need to be extended. In order to obtain these properties, we need to assume regular properties for the stopping variable, that is, we suppose the existence of a deterministic sequence $(n_\alpha)_{\alpha \rightarrow 0}$ such that we have:

$$N_\alpha/n_\alpha \xrightarrow{p} 1 \quad \text{and} \quad n_\alpha \rightarrow +\infty \text{ as } \alpha \rightarrow 0,$$

where N_α is the sample size for a level of precision α . Notice that here the maximum precision is therefore $\alpha=0$. A sequence of random sample size $(N_\alpha)_{\alpha \rightarrow 0}$ that satisfies the above convergence is said to be regular.

Using the above explicit formulation for the estimator of β_0 of the AHM, we need to obtain asymptotics for

$$\sqrt{N_\alpha}(\hat{\beta}_{N_\alpha}(t, t) - \beta_0).$$

The theory developed in Bordes and Breuils^[18, 19] allows us to prove that, under suitable regularity conditions given in Bordes,^[15] for a fixed time of study t , the estimator indexed by either a deterministic sample size n (with $n \rightarrow +\infty$) or by a regular random sample size N_α (with $\alpha \searrow 0$), has the same asymptotic behavior. It follows that the estimator of the variance of $\hat{\beta}_{N_\alpha}(t, t)$ is the same as the estimator of the variance of $\hat{\beta}_n(t, t)$ simply replacing n by N_α in the expression of $\text{Var}(\hat{\beta}_n(t, s))$

Example 1. Considering the case of 1-D covariates, we know from the previous paragraph that $\sqrt{n}(\hat{\beta}_n(t, s) - \beta_0)$ is approximately Gaussian with asymptotic variance estimated by $\hat{\sigma}_n^2(t, s)$. If $u_{\theta/2}$ satisfies, $2\Phi(u_{\theta/2}) - 1 = 1 - \theta$ where Φ is the cumulative distribution function of standard Gaussian distribution, then

$$\hat{I}_\theta = \left[\hat{\beta}_n(t, s) - \frac{u_{\theta/2} \hat{\sigma}_n(t, s)}{\sqrt{n}}, \hat{\beta}_n(t, s) + \frac{u_{\theta/2} \hat{\sigma}_n(t, s)}{\sqrt{n}} \right]$$

is approximately a $(1-\theta)$ -confidence interval for the unknown parameter β_0 . If we want a precision of estimation specified by length $(\hat{I}_\theta) \leq 2\alpha$ then n have to satisfy the inequality:

$$n \geq \frac{u_{\theta/2}^2 \hat{\sigma}_n^2(t, s)}{\alpha^2}.$$

Therefore, we decide to stop the study when the above inequality is satisfied and for $n \geq n_0$, where n_0 is the minimum size of the sample that ensures that we are not too far

from large sample properties. With such a stopping rule, the sample size at time t is random and equal to:

$$N_\alpha(t) = \inf \left\{ n \geq n_0; n \geq \frac{u_{\theta/2}^2 \hat{\sigma}_n^2(t, s)}{\alpha^2} \right\}.$$

Finally, by the preliminary considerations of this section, we obtain that for a small value of α and a large value of n , $\sqrt{N_\alpha(t)}(\hat{\beta}_{N_\alpha(t)}(t, s) - \beta_0)$ and $\sqrt{n}(\hat{\beta}_n(t, s) - \beta_0)$ have approximately the same Gaussian distribution.

Example 2. Here is a very simple example where we consider that entry dates follow a Poisson process with parameter $1/\alpha$. If we want to traduce the fact that on a given time interval $[0, t]$, the number of observations grows, we may consider that $\alpha \rightarrow 0$. Therefore, by classical results on Poisson processes, we have:

$$\frac{N_\alpha(t)}{t/\alpha} \xrightarrow{p} 1, \quad \text{as } \alpha \rightarrow 0$$

where

$$N_\alpha(t) = \sum_{i \geq 1} 1(\tau_i \leq t).$$

Thus, by preliminary results of this section, we know that for a fixed time t , the asymptotic Gaussian distribution of $\sqrt{N_\alpha(t)}(\hat{\beta}_{N_\alpha(t)}(t, t) - \beta_0)$ for α small enough, is the same as the asymptotic Gaussian distribution of $\sqrt{n}(\hat{\beta}_n(t, t) - \beta_0)$ for a large value of n .

CONCLUSION

In this paper we discussed inference procedures for the AHM with staggered entry. We showed that estimators are naturally indexed by two time scales: the calendar time and the duration of study. We proposed estimators of the unknown parameters of the AHM that account these two time scales and gave explicit formulae for their asymptotic variances. In “Sequential Stopping Rules” we discussed sequential methods of estimation of the regression parameter when an early stop of the study is necessary. We showed that such sequential methods lead to estimators indexed by random sample size and that whenever the random sample size is regular, the nonsequential and the sequential estimators have the same large sample properties.

The above staggered entry model assumes that at a fixed time of study s , the number of individuals is a deterministic value n . Moreover, it is supposed that the number of individuals that “fall” in the study interval $[0, s]$ increases and tends to infinity (this is the sense of the “large sample” theory developed e.g. in Biliás et al.^[21]). Because the entry times are random, it is more reasonable to consider that the number of individuals that are present in the study at fixed time of study is a random variable and to consider

time asymptotic for the estimators. In this case the random sample size would be

$$N_t = \sum_{i \geq 1} 1(\tau_i \leq t).$$

Therefore, estimators of the unknown parameters of the AHM model will be indexed by time t . In a further work we will discuss the asymptotic behavior of estimators when $t \rightarrow +\infty$. As in Example 1 of “Sequential Stopping Rules,” we can construct time stopping rule T_α but asymptotic behavior of estimators indexed by such random time are unknown and need further investigations.

ARTICLES OF FURTHER INTEREST

Confidence Interval and Hypothesis Testing; Proportional Hazards Regression Model; Sample Size Determination; Survival Analysis.

REFERENCES

1. Sellke, T.; Siegmund, D. Sequential analysis of the proportional hazards model. *Biometrika* **1983**, *70*, 315–326.
2. Biliyas, Y.; Gu, M.; Ying, Z. Towards a general asymptotic theory for Cox model with staggered entry. *Ann. Statist.* **1997**, *25* (2), 622–682.
3. Tsiatis, A.A. Repeated significance testing for a general class of statistics used in censored survival analysis. *J. Am. Statist. Assoc.* **1982**, *77*, 855–861.
4. Slud, E.V. Sequential linear tests for two-sample censored survival data. *Ann. Statist.* **1984**, *12*, 551–571.
5. Gu, M.G.; Lai, T.L. Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials. *Ann. Statist.* **1991**, *19*, 1403–1433.
6. Cox, D.R. Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **1972**, *34*, 187–220.
7. Huffer, F.W.; McKeague, I.W. Weighted least square estimation for Aalen’s additive risk model. *J. Am. Statist. Assoc.* **1991**, *86*, 114–129.
8. Cox, D.R.; Oakes, D. *Analysis of Survival Data*; Chapman & Hall: London, 1984.
9. Lin, D.Y.; Ying, Z. Semiparametric inference for the accelerated life model with time-dependent covariates. *J. Statist. Plann. Inf.* **1995**, *44*, 47–53.
10. Bagdonavicius, V.; Nikulin, M.S. Transfer functionals and semiparametric regression models. *Biometrika* **1996**, *84*, 365–378.
11. Bordes, L. Semiparametric additive accelerated life models. *Scand. J. Statist.* **1999**, *26*, 1–17.
12. Lin, D.Y.; Ying, Z. Semiparametric analysis of the additive risk model. *Biometrika* **1994**, *81* (1), 61–71.
13. Andersen, P.K.; Borgan, O.; Gill, R.D.; Keiding, N. *Statistical Models Based on Counting Processes. Springer Series in Statistics*; Springer-Verlag: New York, 1993.
14. Pollard, D. *Empirical Processes: Theory and Applications*. NFS–CMS Regional Conference Series in Probability and Statistics 2. 1990.
15. Bordes, L. Sequential analysis of the additive hazards model. Technical report 99–01 of the University of Technology of Compiègne, 1999.
16. Ghosh, M.; Mukhopadhyay, N.; Sen, P.K. *Sequential Estimation*; John Wiley and Sons, Inc.: New York, 1997.
17. Dmitrienko, A.; Govindarajulu, Z. Sequential confidence regions for maximum likelihood estimates. *Ann. Statist.* **2000**, *28* (5), 1472–1501.
18. Bordes, L.; Breuils, C. Sequential estimation for semiparametric models with application to the proportional hazards. *J. Statist. Plann. Inf.* **2006**, *136* (11), 3735–3759.
19. Bordes, L.; Breuils, C. Sequential estimation of the semiparametric additive hazards model, 2004 submitted.
20. Fleming, T.R.; Harrington, D.P. *Counting Processes and Survival Analysis*; John Wiley and Sons, Inc.: New York, 1990.
21. Lin, D.Y.; Ying, Z. Semiparametric analysis of general additive-multiplicative hazards models for counting process. *Ann. Statist.* **1996**, *23*, 1712–1734.
22. Lin, D.Y.; Wei, L.J.; Ying, Z. Accelerated failure time models for counting processes. *Biometrika* **1998**, *85*, 605–618.

Signal Detection in Pharmacovigilance

Patrick J. Farrell

*School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada;
McLaughlin Center for Population Health Risk Assessment, University of Ottawa, Ottawa,
Ontario, Canada; and Risk Sciences International, Ottawa, Ontario, Canada*

Christopher A. Gravel

*School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada;
McLaughlin Center for Population Health Risk Assessment, University of Ottawa, Ottawa,
Ontario, Canada; and Department of Epidemiology, Biostatistics, and Occupational Health,
McGill University, Montreal, Quebec, Canada*

Daniel Krewski

*School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada;
McLaughlin Center for Population Health Risk Assessment, University of Ottawa, Ottawa,
Ontario, Canada; Risk Sciences International, Ottawa, Ontario, Canada; and School of
Epidemiology, Public Health, and Preventive Medicine, University of Ottawa, Ottawa,
Ontario, Canada*

Abstract

Spontaneous reporting systems containing reports of adverse health outcomes associated with the use of pharmaceutical products represent a valuable source of information for drug safety evaluation. This article provides an overview of statistical methods used in determining whether elevated frequencies in the reporting of adverse events provide evidence of potential safety signals in passive pharmacovigilance. Both frequentist and Bayesian methods are used to examine disproportionality in the reporting rates of adverse events associated with the drug of interest relative to that for a reference drug or group of drugs. Statistical issues in the application of disproportionality methods are discussed, including the selection of a threshold level for signal identification and characterization of false discovery rates. Other issues relevant to the analysis and interpretation of adverse event data are also noted, including under-reporting and confounding. This discussion provides a guide for practitioners in the selection of appropriate statistical methods for signal detection in passive pharmacovigilance and in the interpretation of the analytic results.

Keywords: Adverse Drug Reactions; Adverse Events; Bayesian Approaches; Disproportionality Methods; False Discovery Rate; Frequentist Approaches; Passive Surveillance; Signal Threshold Level; Simultaneous Hypothesis Testing; Spontaneous Reporting Data.

1 INTRODUCTION

Prescription medications undergo rigorous preclinical and clinical testing before receiving market authorization by regulatory authorities (Institute of Medicine^[1,2]). Despite extensive premarket testing under controlled conditions, it is not possible to evaluate all possible adverse effects that might be associated with the use of a particular medication in diverse patient populations. Consequently, pharmaceutical products are subject to post-market evaluation following their introduction into the marketplace using statistical methods for pharmacovigilance (Huang, Guo et al.,^[3] Candore, Juhlin et al.^[4]).

Pharmacovigilance seeks to identify safety signals heralding potential risks of adverse drug reactions (ADRs) under real world conditions of use. The World Health

Organization (WHO) defines pharmacovigilance as “the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problems” (WHO^[5]). Reports of adverse drug reactions (ADRs) are compiled in databases such as Canada Vigilance maintained by Health Canada, the FDA Adverse Events Reporting System (FAERS) maintained by the US Food & Drug Administration (FDA), and the EudraVigilance maintained by the European Medicines Agency (EMA). The WHO consolidates reports from multiple jurisdictions internationally within the WHO Vigibase, which contains more than 12 million individual case safety reports (ICSRs) as of December 2015. Recent analyses of the Vigibase data have suggested possible associations between botulinum neurotoxin type A (BoNT-A) used as an anti-spastic agent in children with cerebral palsy

and mortality (Montrastruc, Marque et al.^[6]), and between the use of quinolones and suicidal behaviors (Samyde, Petit et al.^[7]).

Historically, passive pharmacovigilance has focused on the analysis of data from spontaneous reporting (SR) systems, such as the ICSRs in Vigibase (Shankar^[8]). Spontaneous reporting data typically includes limited patient information within the ICSR, often limited to age and sex. Further, due to the self-reporting nature of passive pharmacovigilance systems, SR data has been used primarily to motivate the formulation of drug safety hypotheses. More recently, increasing emphasis has been placed on active pharmacovigilance, using longitudinal observational datasets (LODs) to investigate the risk of adverse drug reactions within large patient populations with more detailed patient information, including co-morbidity and polypharmacy (Huang, Moon et al.,^[9] Zhuo, Farrell et al.^[10]). Within the last decade, major initiatives such as the establishment of the Canadian Network for Observational Drug Effect Studies (CNODES) in Canada (Suisse, Henry et al.^[11]) and the Sentinel Initiative in the United States (Ball, Robb et al.^[12]) have begun to harness the increasing availability of LODs for drug safety and effectiveness evaluation. Initiatives such as these are beginning to yield results that will serve to establish proof-of-principle of LODs as a key resource for drug safety evaluation. As an example, in a combined analysis of data from six participating sites in Canada, the United States, and the United Kingdom involving over 1.5 million patients, Azoulay, Fillion et al.^[13] reported no association between incretin-based drugs acute pancreatitis, in comparison with other oral antidiabetic drugs.

While active pharmacovigilance will play an increasing important role in drug safety evaluation in the future, passive pharmacovigilance remains an important tool for the early detection of safety signals (Chakravarty, Izem et al.^[14]). The signals identified using SR data can be investigated in more depth using LODs.

This article focuses on statistical methods for the early detection of adverse events (AEs) based on SR data (Candore, Juhlin et al.^[4]). Specifically, we provide a review of the signal detection methods that have been proposed to date, including frequentist and Bayesian approaches. These methods look to identify ‘signals of disproportionate reporting’ (SDRs) in spontaneous reporting data. They evaluate evidence in support of increased reporting rates rather than potential causal relationships (Hauben and Aronson^[15]). The statistical basis for each of the methods considered is presented, along with important issues in their application, such as under-reporting, and the control of false positive rates within a multiple testing context. This review will help practitioners working in the field of pharmacovigilance identify the most appropriate statistical methods to assist them in identifying safety signals in SR data that warrant intervention or further investigation.

In order to facilitate and motivate the presentation to follow, the [Section 2](#) describes the nature of spontaneous

reporting data, and provides an introduction to the passive pharmacovigilance methods for adverse event detection, along with a discussion of the challenges faced when analyzing SR data. [Section 3](#) presents the frequentist methods for signal detection, while [Section 4](#) describes the Bayesian approaches. [Section 5](#) discusses two important issues that arise in the analysis of SR data: decision criteria for signal detection, and adjustments for multiple testing. A conclusion and discussion is given in [Section 6](#), where we elaborate further on additional issues encountered in the application of disproportionality methods in passive pharmacovigilance.

2 PASSIVE PHARMACOVIGILANCE

Spontaneous reporting data can be conveniently summarized in the form of contingency tables as shown in [Table 1](#). Imagine that a particular data set includes information on I drugs and J AEs. We let D_i represent the i -th drug, and A_j the j -th adverse event. The entry n_{ij} in the table denotes the number of patients who were treated with drug D_i and experienced adverse event A_j .

The Observational Medical Outcomes Partnership (OMOP)^[16] aims to identify “the most reliable methods for analyzing huge volumes of data drawn from heterogeneous sources, and has compiled a methods library that includes methods for disproportionality analysis. These techniques compare the reporting rate of a particular AE for a given drug with that for all other drugs and AEs, either individually or combined. Alternatively, they compare the observed number of reports of the drug/AE combination with the expected number of reports under the assumption of no association between the drug and the AE. When the reporting rate of a particular AE for a given drug is compared with its counterpart over all other drugs and AEs combined, the analysis can be based on a collapsed 2×2 version of the above table as represented below in [Table 2](#).

The common disproportionality measures cited by OMOP include the proportional reporting ratio, PRR (Evans, Waller et al.^[17]), the reporting odds ratio, ROR (Van

Table 1 Contingency table summarizing the number of each of J adverse events to each of I different drugs

Drug	Adverse Event						Total
	A_1	A_2	...	A_j	...	A_J	
D_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1J}	$n_{1.}$
D_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2J}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
D_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iJ}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
D_I	n_{I1}	n_{I2}	...	n_{Ij}	...	n_{IJ}	$n_{I.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.J}$	n

Table 2 Collapsed 2×2 contingency table summarizing of spontaneous reporting data for comparing the reporting rate of the j -th adverse event for the i -th drug to the analogous rate over all other drugs combined

Drug	Adverse Event		Total
	A_j	$A_{j'}$	
D_i	n_{ij}	$n_{ij'}$	$n_{i\cdot}$
$D_{i'}$	$n_{i'j}$	$n_{i'j'}$	$n_{i'\cdot}$
Total	$n_{\cdot j}$	$n_{\cdot j'}$	n

Puijenbroek, Bate et al.^[18] Rothman, Lanes et al.^[19], the (multi) gamma-Poisson shrinker, MGPS (DuMouchel,^[20] DuMouchel and Pregibon^[21]), and the Bayesian Confidence Propagation Neural Network, BCPNN (Bate, Lindquist et al.^[22]). These methods rely exclusively on the number of reports of adverse events, and do not consider drug exposure-time information within the patient population.

The disproportionality methods for spontaneous AE reporting can be categorized as frequentist or Bayesian. Of those above cited above, the PRR and the ROR are included in the former category. Other frequentist approaches include the chi-square test for independence, with or without continuity correction (Yates,^[23] Yule and Kendall,^[24] Greenwood and Nikulin^[25]), and the reporting Fisher's exact test (Agresti,^[26] Ahmed, Dalmasso et al.^[27]). More recently, a 'frequentist' version of the information component (IC) and a likelihood ratio test (LRT) have also been proposed (Huang, Guo et al.^[3] Huang, Zalkikar et al.^[28]). The IC measure has previously been used in the BCPNN, which, along with the MGPS, are the Bayesian disproportionality methods identified by OMOP. They fall under a subcategory known as empirical Bayes techniques.

The general principle underlying signal detection in disproportionality methods involves the comparison of a 'score' for a particular drug/AE combination with a 'threshold' level. These scores can be viewed as a numerical summary measure of the relationship between a given AE and a certain drug obtained by comparing the frequency of occurrence of the AE with the drug to the frequency of the same AE with all other drugs. A test of the null hypothesis of no association between the drug and the AE is performed. A score exceeding the threshold results in rejection of this hypothesis, and suggests an association between the drug and AE under consideration. If the hypothesis is only concerned with a single combination, the Type I error is maintained at a specified level. However, it is often the case that there are many drug/AE combinations that are simultaneously of interest: in such situations, a number of the disproportionality methods do not control the overall Type I error rate across the multiple hypothesis tests conducted, resulting in too many false positive alarms. The number of false positives in a multiple testing context can be reduced by increasing the threshold level, although this

will increase the number of false negatives. A number of solutions to this issue have been proposed, including the adaptation of the False Discovery Rate, or FDR (Benjamini and Hochberg^[29]), that was originally applied in multidimensional gene and protein microarray analysis (Efron, Tibshirani et al.^[30]).

A number of authors have estimated false positive and false negative error rates for common signal detection methods using simulated data (Rolka, Bracy et al.,^[31] Roux, Thiessard et al.,^[32] Almenoff, LaCroix et al.,^[33] Matsushita, Kuroda et al.,^[34] Ahmed, Dalmasso et al.,^[27] Ahmed, Thiessard et al.,^[35] Huang, Zalkikar et al.^[28]). In addition to identifying issues relating to the selection of threshold levels, these studies have also highlighted challenges in comparing the performance of Bayesian and frequentist approaches. These and other issues encountered when analyzing SR data will be discussed in [Sections 5 and 6](#). The next two sections are devoted to frequentist and Bayesian approaches for signal detection in passive pharmacovigilance.

3 FREQUENTIST METHODS

Since SR data is a collection of AE reports with information on prescription drug utilization by the individual making the report, the data are essentially retrospective: conditional on AE status, the drugs that the individual was taking in advance of making the report are observed. Spontaneous reporting data therefore needs to be analyzed as such. Consider the collapsed 2×2 contingency table given in [Table 2](#). Let π_{ij} be the probability of jointly observing the i -th drug and the j -th AE. In addition, let $\pi_{i\cdot}$ and $\pi_{\cdot j}$ be the marginal probabilities of observing the i -th drug and the j -th AE, respectively. Therefore, $\pi_{i'} = 1 - \pi_{i\cdot}$ and $\pi_{\cdot j'} = 1 - \pi_{\cdot j}$ are the marginal probabilities of not observing the i -th drug and the j -th AE, respectively. Similarly, $\pi_{ij'} = \pi_{i\cdot} - \pi_{ij}$ is the joint probability of observing the i -th drug, but not the j -th AE, while $\pi_{i'j} = \pi_{\cdot j} - \pi_{ij}$ is the joint probability of observing the j -th drug, but not the i -th AE drug. Finally, $\pi_{i'j'} = \pi_{i'} - \pi_{i'j} = \pi_{\cdot j'} - \pi_{ij'}$ is the joint probability of observing neither the i -th drug nor the j -th AE. These joint and marginal probabilities are summarized in [Table 3](#), which takes on a form identical to that summarizing SR data in [Table 2](#).

The joint and marginal probabilities given in [Table 3](#) allow for the determination of four conditional probabilities of observing (or not observing) the i -th drug conditional on reporting (or not reporting) the j -th AE. The probability of observing the i -th drug conditional on reporting the j -th AE is given by $\pi_{i|j} = \pi_{ij} / \pi_{\cdot j}$, while the probability of not observing the i -th drug conditional on reporting the j -th AE is given by $\pi_{i'|j} = \pi_{i'j} / \pi_{\cdot j}$. Given that the j -th AE was not reported, the conditional probability of observing the i -th drug is $\pi_{i|j'} = \pi_{ij'} / \pi_{\cdot j'}$, while the conditional probability of not observing the i -th drug is $\pi_{i'|j'} = \pi_{i'j'} / \pi_{\cdot j'}$. Using this

Table 3 Collapsed 2×2 contingency table summary of joint and marginal probabilities for comparing the reporting rate of the j -th adverse event for the i -th drug to that for all other drugs combined

Drug	Adverse Event		Total
	A_j	$A_{j'}$	
D_i	π_{ij}	$\pi_{ij'}$	$\pi_{i\cdot}$
$D_{i'}$	$\pi_{i'j}$	$\pi_{i'j'}$	$\pi_{i'\cdot}$
Total	$\pi_{\cdot j}$	$\pi_{\cdot j'}$	1

notation, we now summarize frequentist methods for the analysis of SR data.

Reporting Odds Ratio, ROR

For the 2×2 table given in Table 3, the ROR for the i -th drug and j -th AE is given by

$$ROR_{ij} = \frac{\pi_{i|j}/\pi_{i'|j}}{\pi_{i|j'}/\pi_{i'|j'}} = \frac{\pi_{i|j}\pi_{i'|j'}}{\pi_{i|j'}\pi_{i'|j}} = \frac{\pi_{ij}\pi_{i'j'}}{\pi_{ij'}\pi_{i'j}}$$

Notice that in the second term, the numerator $\pi_{i|j}/\pi_{i'|j}$ is simply the odds of observing the i -th drug (the ratio of the probability of observing the i -th drug to the probability of not observing it), conditional on reporting the j -th AE. The denominator $\pi_{i|j'}/\pi_{i'|j'}$ is the odds of observing the i -th drug conditional on the j -th AE not being reported. Thus, the larger the value of ROR_{ij} , the higher the association between the i -th drug and the j -th AE. In fact, if $ROR_{ij} > 1$, the odds of observing drug i given the j -th AE is reported exceed the odds if the j -th AE is not reported, while the opposite is true if $ROR_{ij} < 1$.

The reporting odds ratio ROR_{ij} is estimated by

$$\widehat{ROR}_{ij} = \frac{n_{ij}n_{i'j'}}{n_{ij'}n_{i'j}}$$

While the distribution of $\widehat{ROR}_{ij}$ can be highly skewed, particularly if n_{ij} or $n_{i'j}$ are very small, the asymptotic distribution of $\log(\widehat{ROR}_{ij})$ is normal. Note that $\log(\widehat{ROR}_{ij})$ is not defined when either $n_{ij'}$ or $n_{i'j}$ are zero. Applying the delta method, the estimated variance of $\log(\widehat{ROR}_{ij})$ is

$$\hat{V}[\log(\widehat{ROR}_{ij})] \approx \frac{1}{n_{ij}} + \frac{1}{n_{i'j}} + \frac{1}{n_{ij'}} + \frac{1}{n_{i'j'}}$$

An approximate $100(1 - \alpha)\%$ confidence interval for ROR_{ij} is therefore given by

$$\left(e^{\log(\widehat{ROR}_{ij}) - z_{\alpha/2} \sqrt{\hat{V}[\log(\widehat{ROR}_{ij})]}}, e^{\log(\widehat{ROR}_{ij}) + z_{\alpha/2} \sqrt{\hat{V}[\log(\widehat{ROR}_{ij})]}} \right)$$

where $z_{\alpha/2}$ is the $\alpha/2$ upper critical value of the standard normal distribution. If the entire interval situates entirely

above 1, the j -th AE can be interpreted as statistically significant evidence of a signal of a disproportionate reporting rate for the i -th drug at significance level α .

Proportional Reporting Ratio, PRR

The PRR for the i -th drug and j -th AE is given by

$$PRR_{ij} = \frac{\pi_{j|i}}{\pi_{j|i'}} = \frac{\pi_{ij}/\pi_{i\cdot}}{\pi_{i'j}/\pi_{i'\cdot}}$$

It can be interpreted as the ratio of the conditional probability of observing the j -th AE given that the reporter has been treated with the i -th drug to the conditional probability of observing the j -th AE given that the i -th drug was not prescribed. Thus, if $PRR_{ij} > 1$, the probability of observing the j -th AE given that the i -th drug has been prescribed exceeds the corresponding probability when the i -th drug is not used.

The proportional reporting ratio PRR_{ij} is estimated by

$$\widehat{PRR}_{ij} = \frac{n_{ij}/n_{i\cdot}}{n_{i'j}/n_{i'\cdot}}$$

Since the asymptotic distribution of $\log(\widehat{PRR}_{ij})$ is approximately normal, its variance can be estimated by

$$\hat{V}[\log(\widehat{PRR}_{ij})] \approx \frac{1}{n_{ij}} + \frac{1}{n_{i\cdot}} + \frac{1}{n_{i'j}} + \frac{1}{n_{i'\cdot}}$$

An approximate $100(1 - \alpha)\%$ confidence interval for PRR_{ij} is therefore given by

$$\left(e^{\log(\widehat{PRR}_{ij}) - z_{\alpha/2} \sqrt{\hat{V}[\log(\widehat{PRR}_{ij})]}}, e^{\log(\widehat{PRR}_{ij}) + z_{\alpha/2} \sqrt{\hat{V}[\log(\widehat{PRR}_{ij})]}} \right)$$

While evidence supporting the conclusion of a disproportionate rate for the i -th drug is demonstrated when the lower limit of this interval exceeds one, there have been suggestions to modify the threshold (Evans, Waller et al.^[16]), raising it from one to two, and requiring that all $n_{ij} \geq 3$. The implications of the choice of criteria for signal detection are discussed in detail in Section 5.

Note that with the PRR, inference is conditional on the drug, and not on the AE, which is contrary to the nature of SR data. However, consider the case of a rare AE, regardless of the drug administered. In this case, we may assume that both n_{ij} and $n_{i'j}$ are very small relative to $n_{ij'}$ and $n_{i'j'}$, respectively. Since $n_{i\cdot} = n_{ij} + n_{i'j} \approx n_{ij'}$ and $n_{i'\cdot} = n_{i'j} + n_{i'j'} \approx n_{i'j'}$, in the case of rare events, we can estimate PRR_{ij} by

$$\widehat{PRR}_{ij} = \frac{n_{ij}/n_{i\cdot}}{n_{i'j}/n_{i'\cdot}} \approx \frac{n_{ij}/n_{ij'}}{n_{i'j}/n_{i'j'}} = \frac{n_{ij}n_{i'j'}}{n_{ij'}n_{i'j}} = \widehat{ROR}_{ij}$$

Care must nevertheless be taken in using the PRR and ROR for rare AEs since, if $n_{i\cdot}$ for the i -th drug is much smaller

than n_i for all other drugs, extremely small but similar values for n_{ij} and n_j may result in a conclusion that the j -th AE can be interpreted as a signal of a disproportional reporting rate for the i -th drug that is based on negligible occurrences of the AE, both for the i -th drug, as well as all other drugs.

Frequentist-Based Information Component, IC

Huang, Guo et al.^[3] propose a frequentist version of the IC for examining the relationship between the i -th drug and j -th AE. The IC is based on the relative reporting rate, or RRR. The RRR essentially compares the probability of observing the i -th drug and j -th AE with this same probability under the assumption of independence, or no association. Since the joint probability of observing the i -th drug and j -th AE is π_{ij} , while the analogous probability under the assumption of independence of drug i and the j -th AE is $\pi_i \cdot \pi_j$. The RRR is simply the ratio of these two probabilities, given by

$$RRR_{ij} = \frac{\pi_{ij}}{\pi_i \cdot \pi_j}$$

The estimate of the RRR_{ij} is

$$\widehat{RRR}_{ij} = \frac{n_{ij}/n}{(n_i/n)(n_j/n)} = \frac{n_{ij}}{(n_i \cdot n_j)/n} = \frac{n_{ij}}{E_{ij}}$$

where

$$E_{ij} = \frac{n_i \cdot n_j}{n}$$

is the expected number of reports of the i -th drug and j -th AE combination under the assumption of independence. Clearly, values of $\widehat{RRR}_{ij}$ much in excess of one would suggest that the j -th AE could be interpreted as a signal of a disproportionate AE rate for drug i .

The IC is defined as

$$IC_{ij} = \log_2(RRR_{ij}) = \log_2\left(\frac{\pi_{ij}}{\pi_i \cdot \pi_j}\right)$$

and is estimated by

$$\widehat{IC}_{ij} = \log_2\left[\frac{n_{ij}}{(n_i \cdot n_j)/n}\right] = \log_2\left(\frac{n_{ij}}{E_{ij}}\right)$$

provided that n_i , n_j , and n_{ij} are all greater than zero. Noting that

$$\log_2\left[\frac{n_{ij}}{(n_i \cdot n_j)/n}\right] = \frac{\log\left[\frac{n_{ij}}{(n_i \cdot n_j)/n}\right]}{\log 2}$$

and that the asymptotic distribution of $\widehat{IC}_{ij}$ is approximately normal, its variance can be estimated by

$$\widehat{V}(\widehat{IC}_{ij}) \approx \frac{1}{(\log 2)^2} \left(\frac{1}{n_{ij}} + \frac{1}{n_i} + \frac{1}{n_j} \right)$$

An approximate $100(1 - \alpha)\%$ confidence interval for IC_{ij} is therefore given by

$$\left(\widehat{IC}_{ij} - z_{\alpha/2} \sqrt{\widehat{V}(\widehat{IC}_{ij})}, \widehat{IC}_{ij} + z_{\alpha/2} \sqrt{\widehat{V}(\widehat{IC}_{ij})} \right)$$

If this interval lies entirely above zero, then the j -th AE has been reported at a disproportionate rate for the i -th drug, providing evidence of a signal of an AE.

Chi-Square Test of Independence

The chi-square (χ^2) test of independence formalizes the joint probability comparison of the RRR. However, the test requires that the SR data are obtained via multinomial sampling. Note that when conditioning on the j -th AE, the column totals, n_j , are fixed. This is the case when a set of n_j records with the j -th AE are examined to determine the incidence of the i -th drug. In multinomial sampling, only the total number of records examined, n , is fixed. For such a scheme, a set of n records would be selected to determine the prevalence of both the i -th drug and the j -th AE. Returning for the moment to the case of conditioning on the j -th AE, there is an analogous chi-square test for such data, called the chi-square test for homogeneity. In the chi-square test for homogeneity, the n_{ij} observed in the 2×2 table are compared to their expected values under the assumption that $\pi_{i|j} = \pi_{i|j'}$ (that is, the probability of observing the i -th drug conditional on reporting the j -th AE is the same as the analogous probability conditional on the j -th AE not being reported). In the chi-square test of independence, the n_{ij} are compared to their expected values under the assumption that $\pi_{ij} = \pi_i \cdot \pi_j$ (implying no association, or independence, between the i -th drug and the j -th AE). However, despite the fact that the derivation of the appropriate formula for investigating whether the j -th AE is a signal differs for the two chi-square tests, the resulting formula for the chi-square statistic turns out to be exactly the same for the two procedures! In what follows, we therefore describe the test for independence, bearing in mind that it is appropriate only for multinomial sampling, and not when data are collected conditional on AE status.

The chi-square test for independence requires the calculation of a test statistic that uses each cell of the 2×2 table. The test statistic is given by

$$\chi_{ij}^2 = \frac{(n_{ij} - E_{ij})^2}{E_{ij}} + \frac{(n_{ij'} - E_{ij'})^2}{E_{ij'}} + \frac{(n_{i'j} - E_{i'j})^2}{E_{i'j}} + \frac{(n_{i'j'} - E_{i'j'})^2}{E_{i'j'}}$$

Under the hypothesis of independence, this test statistic follows a chi-square distribution with 1 degree of freedom. Hence $P(\chi^2_1 > \chi^2_{ij})$ represents an approximate p-value for assessing whether the j -th AE can be interpreted as a signal of a disproportional rate for the i -th drug. Evans, Waller et al.^[17] suggest further conditions for concluding that the j -th AE is a signal, adding that it should be the case that $\chi^2_{ij} \geq 4$ and $n_{ij} \geq 3$.

It is important to note here that while the n_{ij} are counts and therefore discrete, the chi-square distribution followed by the test statistic χ^2_{ij} is continuous. A number of different continuity corrections have been applied to the test statistic to take this discrepancy into account. A common correction is the “0.5 continuity correction”, which leads to

$$\chi^2_{ij} = \frac{(n_{ij} - E_{ij})^2 - 0.5}{E_{ij}} + \frac{(n_{ij'} - E_{ij'})^2 - 0.5}{E_{ij'}} + \frac{(n_{i'j} - E_{i'j})^2 - 0.5}{E_{i'j}} + \frac{(n_{i'j'} - E_{i'j'})^2 - 0.5}{E_{i'j'}}$$

as the test statistic. Alternatively, Yates^[23] proposed a correction that yielded the test statistic

$$\chi^2_{ij} = \frac{(|n_{ij}n_{i'j'} - n_{ij'}n_{i'j}| - 0.5n^2)n}{n_i n_{i'} n_j n_{j'}}$$

while Yule and Kendall^[23] describe another measure involving a continuity correction called Yule's Q_{ij} , which is given by

$$Q_{ij} = \frac{n_{ij}n_{i'j'} - n_{ij'}n_{i'j}}{n_{ij}n_{i'j'} + n_{ij'}n_{i'j}}$$

It can be shown that $-1 \leq Q_{ij} \leq 1$, and that a value of $Q_{ij} = 0$ provides evidence of no association between drug i and the j -th AE. Since the variance of Q_{ij} can be estimated by

$$\hat{V}(Q_{ij}) = \frac{(1 - Q_{ij}^2)^2}{4} \left(\frac{1}{n_{ij}} + \frac{1}{n_{i'j}} + \frac{1}{n_{ij'}} + \frac{1}{n_{i'j'}} \right)$$

an approximate $100(1 - \alpha)\%$ confidence interval based on Q_{ij} as an estimate is given by

$$(Q_{ij} - z_{\alpha/2} \sqrt{\hat{V}(Q_{ij})}, Q_{ij} + z_{\alpha/2} \sqrt{\hat{V}(Q_{ij})})$$

The j -th AE would be considered as having been reported at a disproportionate rate for the i -th drug if the above interval lay entirely above zero.

Reporting Fisher's Exact Test, RFET, and the Urn Model

It was indicated above that the validity of the chi-square test of independence rests on sampling of records being multinomial, and not conditioned on a given AE. The reporting Fisher's exact test (Fisher,^[36] Ahmed, Dalmaso et al.^[27]) requires conditioning not only on a given AE, but also on a particular drug. Thus, not only is n_j fixed, but so is n_i . Here, the random variable, N_{ij} , is for the outcome n_{ij} ; it follows a hypergeometric distribution given n_i , n_j , and n . Therefore

$$P(N_{ij} = n_{ij}) = \frac{\binom{n_j}{n_{ij}} \binom{n_{j'}}{n_i - n_{ij}}}{\binom{n}{n_i}}$$

Here the quantity $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is the binomial coefficient, which represents the number of ways that x objects can be selected from n , without regards to order. A p-value for a test that the j -th AE is a signal for a disproportional rate for the i -th drug is given by $(N_{ij} \geq n_{ij})$.

The hypergeometric distribution is highly discrete for small samples, so when n_{ij} can only assume a few values, the p-value can only take on the same small number of possible values. It is therefore usually not possible to achieve the desired significance level to which the p-value is compared. To reduce conservatism, Agresti^[26] recommends the use of the mid-p-value given by

$$\text{mid-p-value} = \frac{1}{2} P(N_{ij} = n_{ij}) + P(N_{ij} > n_{ij})$$

While the probability of Type I error is no longer guaranteed to be no greater than the specified significance level, in practice it is rarely much greater (Agresti^[26]).

Hochberg, Reisinger et al.^[37] introduced the urn model under similar assumptions as those of the reporting Fisher's exact test. They define the quantity

$$U_{ij} = \frac{1}{P(N_{ij} \geq n_{ij})}$$

as a measure of statistical unexpectedness. The probability in the denominator, which is the p-value given above for Fisher's test, is evaluated conditional on the assumption of no association between the i -th drug and j -th AE. The quantity U_{ij} can therefore be thought of as a reciprocal p-value, and compared to $1/\alpha$ to assess the presence or absence of a signal.

Likelihood Ratio Test, LRT

The likelihood ratio test proposed by Huang, Zalkikar et al.^[28] compares the relative reporting rate, RRR_{ij} , of

different AEs, A_j , for $j = 1, \dots, J$ for a given drug i of interest. The test is therefore based on multiple 2×2 tables of the form given in Table 2. The null hypothesis is that for the i -th drug, $RRR_{ij} = 1$ for all $j = 1, \dots, J$, with the alternative hypothesis that $RRR_{ij} > 1$ for at least one $j = 1, \dots, J$. For drug i , the test is based on J test statistics given by

$$LRT_{ij} = \frac{\binom{n_{ij}}{n_{.j}} \binom{n_{ij'}}{n_{.j'}}}{\binom{n_{i.}}{n}} = \binom{n_{ij}}{E_{ij}} \binom{n_{ij'}}{E_{ij'}} \quad \text{for } j = 1, \dots, J$$

Let the estimates of the conditional probabilities of observing the i -th drug given that the j -th AE is/is not reported be $\hat{\pi}_{i|j} = n_{ij}/n_{.j}$ and $\hat{\pi}_{i|j'} = n_{ij'}/n_{.j'}$, respectively. For the i -th drug, the maximum likelihood test statistic for the alternative hypothesis $RRR_{ij} > 1$ for at least one $j = 1, \dots, J$ is

$$MLR_i = \max_j \left[LRT_{ij} I(\hat{\pi}_{i|j} > \hat{\pi}_{i|j'}) \right] \\ = \max_j \left[\binom{n_{ij}}{E_{ij}} \binom{n_{ij'}}{E_{ij'}} I(\hat{\pi}_{i|j} > \hat{\pi}_{i|j'}) \right]$$

where $I(\hat{\pi}_{i|j} > \hat{\pi}_{i|j'}) = 1$ if $\hat{\pi}_{i|j} > \hat{\pi}_{i|j'}$, and 0 otherwise.

The distribution of MLR_i under the null hypothesis that $RRR_{ij} = 1$ for all $j = 1, \dots, J$ is not analytically tractable, but can be estimated using Monte Carlo methods by generating a large number of simulated data sets; Huang, Zalkikar et al.^[28] create 9,999 such data sets. The simulation is based on the full $I \times J$ contingency table given in Table 1. A single simulated data set is obtained by generating values of $n_{i1}, n_{i2}, \dots, n_{iJ}$ from a multinomial distribution where the total number of trials is $n_{i.}$ and the probabilities are $n_{.j}/n$ for $j = 1, \dots, J$. Let the values obtained for the s -th simulated data set be $n_{i1}^{(s)}, n_{i2}^{(s)}, \dots, n_{iJ}^{(s)}$ for $s = 1, \dots, 9999$. For each of these simulated data sets, a value of $MLR_i^{(s)}$ is calculated. The 9,999 values obtained for MLR_i through the simulation can be used to define the empirical distribution of this statistic.

To test for AEs signaling disproportionate rates for drug i , the first step is to compare the value of MLR_i computed from the actual data to an appropriate percentile of the empirical distribution of MLR_i obtained via simulation. For illustrative purposes, suppose that we set $\alpha = 0.05$. Then if the value of MLR_i computed from the actual data exceeds the upper 5th percentile of the empirical distribution of MLR_i , then the AE corresponding to $\max_j \left[LRT_{ij} I(\hat{\pi}_{i|j} > \hat{\pi}_{i|j'}) \right]$ based on the actual data is deemed to be the most significant signal. To identify secondary signals, the AE that yielded the second largest value for $LRT_{ij} I(\hat{\pi}_{i|j} > \hat{\pi}_{i|j'})$ from the actual data is considered next. If this value also exceeds the upper 5th

percentile of the empirical distribution of MLR_i , there is evidence to suggest that this AE is also a signal. The AE associated with the next largest $LRT_{ij} I(\hat{\pi}_{i|j} > \hat{\pi}_{i|j'})$ from the actual data would be investigated next, and so on, until no further signals can be identified.

4 BAYES AND EMPIRICAL BAYES METHODS

Traditional Bayes techniques for signal detection in passive pharmacovigilance tend to focus on the relative reporting rate. In this section, both empirical and fully Bayesian methods for disproportionality analysis are presented.

Bayesian Confidence Propagation Neural Network, BCPNN

The information component

$$IC_{ij} = \log_2(RRR_{ij}) = \log_2 \left(\frac{\pi_{ij}}{\pi_{i.} \pi_{.j}} \right)$$

is the focus of the BCPNN developed by investigators at the World Health Organization (WHO) Collaborating Centre for International Drug Monitoring (Bate, Lindquist et al.^[22]). Starting from the $I \times J$ contingency table given in Table 1, $J - 2 \times 2$ tables are constructed for the i -th drug. For the i -th drug and j -th AE, beta prior distributions for the unconditional marginal and joint reporting probabilities π_{ij} , $\pi_{i.}$, and $\pi_{.j}$ are specified. Conditional on these distributions, the associated cell counts n_{ij} , $n_{i.}$, and $n_{.j}$ are assumed to be realizations from independent binomial distributions, each with a parameter n , for the number of trials. Specifically, it is assumed that

$$n_{ij} | \pi_{ij} \sim \text{Binomial}(n, \pi_{ij}) \text{ with } \pi_{ij} \sim \text{Beta}(\alpha_{ij}, \beta_{ij})$$

$$n_{i.} | \pi_{i.} \sim \text{Binomial}(n, \pi_{i.}) \text{ with } \pi_{i.} \sim \text{Beta}(\alpha_{i.}, \beta_{i.})$$

$$n_{.j} | \pi_{.j} \sim \text{Binomial}(n, \pi_{.j}) \text{ with } \pi_{.j} \sim \text{Beta}(\alpha_{.j}, \beta_{.j})$$

Given these specifications, it follows that the posterior distributions of π_{ij} , $\pi_{i.}$, and $\pi_{.j}$ are also beta, with

$$\pi_{ij} | n_{ij} \sim \text{Beta}(\alpha_{ij}^*, \beta_{ij}^*)$$

$$\pi_{i.} | n_{i.} \sim \text{Beta}(\alpha_{i.}^*, \beta_{i.}^*)$$

$$\pi_{.j} | n_{.j} \sim \text{Beta}(\alpha_{.j}^*, \beta_{.j}^*)$$

where $\alpha_{ij}^* = \alpha_{ij} + n_{ij}$, $\beta_{ij}^* = \beta_{ij} + n - n_{ij}$, $\alpha_{i.}^* = \alpha_{i.} + n_{i.}$, $\beta_{i.}^* = \beta_{i.} + n_{i.}$, $\alpha_{.j}^* = \alpha_{.j} + n_{.j}$, and $\beta_{.j}^* = \beta_{.j} + n_{.j}$. In some of the early studies of the BCPNN (Bate, Lindquist et al.^[21] Orre,

Lansner et al.^[38], starting values of one are specified for α_{ij} , α_i , and α_j , as well as for β_i and β_j . The hyperparameter β_{ij} is estimated by solving the equation that the prior mean of π_{ij} is equal to its posterior mean under the hypothesis of no association between the i -th drug and j -th AE, namely

$$\beta_{ij} = \frac{1 - E(\pi_i | n_i) E(\pi_j | n_j)}{E(\pi_i | n_i) E(\pi_j | n_j)} = \frac{1 - \frac{\alpha_i^*}{(\alpha_i^* + \beta_i^*)} \frac{\alpha_j^*}{(\alpha_j^* + \beta_j^*)}}{\frac{\alpha_i^*}{(\alpha_i^* + \beta_i^*)} \frac{\alpha_j^*}{(\alpha_j^* + \beta_j^*)}}$$

which is estimated by

$$\hat{\beta}_{ij} = \frac{(n+2)^2 - (n_i+1)(n_j+1)}{(n_i+1)(n_j+1)}$$

if α_{ij} , α_i , α_j , β_i , and β_j are all set to one.

Using the delta method, we can write

$$IC_{ij} \approx \frac{1}{\log 2} \left\{ \log \left[\frac{E(\pi_{ij} | n_{ij})}{E(\pi_i | n_i) E(\pi_j | n_j)} \right] + \frac{\pi_{ij} - E(\pi_{ij} | n_{ij})}{E(\pi_{ij} | n_{ij})} - \frac{\pi_i - E(\pi_i | n_i)}{E(\pi_i | n_i)} - \frac{\pi_j - E(\pi_j | n_j)}{E(\pi_j | n_j)} \right\}$$

which allows for the posterior mean and variance of IC_{ij} to be determined, yielding

$$\begin{aligned} E(IC_{ij} | n_{ij}) &= \log_2 \left[\frac{E(\pi_{ij} | n_{ij})}{E(\pi_i | n_i) E(\pi_j | n_j)} \right] \\ &= \log_2 \left[\frac{\frac{\alpha_{ij}^*}{(\alpha_{ij}^* + \beta_{ij}^*)}}{\frac{\alpha_i^*}{(\alpha_i^* + \beta_i^*)} \frac{\alpha_j^*}{(\alpha_j^* + \beta_j^*)}} \right] \end{aligned}$$

and

$$\begin{aligned} V(IC_{ij} | n_{ij}) &\approx \frac{1}{(\log 2)^2} \left\{ \frac{V[\pi_{ij} - E(\pi_{ij} | n_{ij})]}{[E(\pi_{ij} | n_{ij})]^2} + \frac{V[\pi_i - E(\pi_i | n_i)]}{[E(\pi_i | n_i)]^2} + \frac{V[\pi_j - E(\pi_j | n_j)]}{[E(\pi_j | n_j)]^2} \right\} \\ &\approx \frac{1}{(\log 2)^2} \left\{ \frac{\beta_{ij}^*}{\alpha_{ij}^* (\alpha_{ij}^* + \beta_{ij}^* + 1)} + \frac{\beta_i^*}{\alpha_i^* (\alpha_i^* + \beta_i^* + 1)} + \frac{\beta_j^*}{\alpha_j^* (\alpha_j^* + \beta_j^* + 1)} \right\} \end{aligned}$$

which become

$$E(IC_{ij} | n_{ij}) = \frac{1}{\log 2} \log \left[\frac{(n_{ij}+1)(n+2)^2}{(n+\beta_{ij}+1)(n_i+1)(n_j+1)} \right]$$

and

$$\begin{aligned} V(IC_{ij} | n_{ij}) &\approx \frac{1}{(\log 2)^2} \left\{ \frac{(n-n_{ij}-\hat{\beta}_{ij})}{(n_{ij}+1)(n+\hat{\beta}_{ij}+2)} + \frac{(n-n_i+1)}{(n_i+1)(n+3)} \right. \\ &\quad \left. + \frac{(n-n_j+1)}{(n_j+1)(n+3)} \right\} \end{aligned}$$

if α_{ij} , α_i , α_j , β_i , and β_j are all set to one. These quantities can be estimated by replacing β_{ij} with the estimate $\hat{\beta}_{ij}$, while allows for an approximate $100(1-\alpha)\%$ confidence interval for IC_{ij} to be specified as

$$\left(E(IC_{ij} | n_{ij}) - z_{\alpha/2} \sqrt{V(IC_{ij} | n_{ij})}, E(IC_{ij} | n_{ij}) + z_{\alpha/2} \sqrt{V(IC_{ij} | n_{ij})} \right)$$

If this interval lies entirely above zero, then the j -th AE can be interpreted as a signal of a disproportionate AE rate for the i -th drug.

Subsequent to this approximation proposed by Bate, Linquist et al.^[22], for small sample sizes, Gould^[39] presented exact expressions for the posterior mean and variance of the IC_{ij} using the moment generating function technique. These are given by

$$\begin{aligned} E(IC_{ij} | n_{ij}) &= \frac{1}{\log 2} \left\{ \Psi(\alpha_{ij}^*) - \Psi(\alpha_{ij}^* + \beta_{ij}^*) \right. \\ &\quad \left. - [\Psi(\alpha_i^*) - \Psi(\alpha_i^* + \beta_i^*) + \Psi(\alpha_j^*) - \Psi(\alpha_j^* + \beta_j^*)] \right\} \end{aligned}$$

and

$$\begin{aligned} V(IC_{ij} | n_{ij}) &= \frac{1}{(\log 2)^2} \left\{ \Psi'(\alpha_{ij}^*) - \Psi'(\alpha_{ij}^* + \beta_{ij}^*) \right. \\ &\quad \left. - [\Psi'(\alpha_i^*) - \Psi'(\alpha_i^* + \beta_i^*) + \Psi'(\alpha_j^*) - \Psi'(\alpha_j^* + \beta_j^*)] \right\} \end{aligned}$$

where $\Psi(k) = d \log \Gamma(k) / dk = \Gamma'(k) / \Gamma(k)$ is the digamma function, and $\Psi'(k) = d \Psi(k) / dk$ is the trigamma function,

both of which are extensively tabulated and available as built-in functions in many statistical software packages. The quantity $\Gamma(k)$ is the gamma function, defined as

$$\Gamma(k) = \int_0^{\infty} x^{k-1} e^{-x} dx$$

This function has the properties that $\Gamma(1) = 1$ and, provided that k is an integer, $\Gamma(k) = (k-1)!$. We note that the corresponding formulas based on the moment generating technique give approximately the same results for large samples.

Gould^[39] also suggested using a less concentrated prior density for π_{ij} that is independent of the observed values for π_i and π_j by replacing these two marginal probabilities with their a priori expected values $E(\pi_i) = \alpha_i / (\alpha_i + \beta_i)$ and $E(\pi_j) = \alpha_j / (\alpha_j + \beta_j)$. Under this replacement, the prior distribution for π_{ij} is beta with parameters α_{ij} and

$$\beta_{ij} = \frac{1 - E(\pi_i)E(\pi_j)}{E(\pi_i)E(\pi_j)} = \frac{1 - \frac{\alpha_i}{(\alpha_i + \beta_i)} \frac{\alpha_j}{(\alpha_j + \beta_j)}}{\frac{\alpha_i}{(\alpha_i + \beta_i)} \frac{\alpha_j}{(\alpha_j + \beta_j)}} = \frac{(\alpha_i + \beta_i)(\alpha_j + \beta_j) - \alpha_i \alpha_j}{\alpha_i \alpha_j}$$

If the same starting values of one as specified above for α_{ij} , α_i , α_j , β_i , and β_j are used, then $\beta_{ij} = 3$.

More recently, Noren, Bate et al.^[40] pointed out that π_{ij} , π_i , and π_j are not necessarily independent as the above analysis assumes, and proposed a full Dirichlet distribution as the multivariate prior distribution of the vector of joint probabilities $\pi' = (\pi_{ij}, \pi_{ij'}, \pi_{i'j}, \pi_{i'j'})$. For a particular 2×2 table, setting $i = 1$, $i' = 0$, $j = 1$, and $j' = 0$, yields $\pi' = (\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})$ with parameters $\alpha_{ij} = q_i q_j \alpha$ where $q_i = (n_i + 0.5) / (n + 1)$, $q_j = (n_j + 0.5) / (n + 1)$, and $\alpha = 0.5 / q_1 q_1$ for $i, j = 0, 1$. Since, along with the quantity n , the vector π' contains the parameters of the multinomial distribution characterizing the prior distribution of the random vector $(n_{11}, n_{10}, n_{01}, n_{00})'$, this leads to a posterior Dirichlet distribution for π' with parameters $\gamma_{ij} = \alpha_{ij} + n_{ij}$ for $i, j = 0, 1$, and beta distributions for the marginal distributions of π_{11} , $\pi_{1\cdot}$, and $\pi_{\cdot 1}$. For additional details, the reader is referred to Noren, Bate et al.^[40]

Multi Gamma Poisson Shrinker, MGPS

The MGPS approach, proposed by DuMouchel,^[20] and DuMouchel and Pregibon^[21], also considers the information component, $IC_{ij} = \log_2(RRR_{ij})$. It is based on the assumption that the number of reports, n_{ij} , of the j -th AE with the i -th drug, $i = 1, \dots, I$ and $j = 1, \dots, J$, are independent Poisson random variables with means $\mu_{ij} = \lambda_{ij} \times E_{ij}$,

where $\lambda_{ij} = RRR_{ij}$, the relative reporting ratio, and E_{ij} is the expected value of n_{ij} when $RRR_{ij} = 1$. As indicated in the previous discussion of frequentist methods, the relative reporting ratio is estimated by n_{ij} / E_{ij} ; thus an estimate for $\lambda_{ij} = RRR_{ij}$ in excess of one would imply that the observed frequency of the i -th drug and j -th AE pair exceeds what would be expected under independence. However, in the Bayesian approach, rather than treating the λ_{ij} as constants, each is assumed to be drawn independently from a common prior distribution, which is assumed to be a mixture of two gamma distributions. The family of gamma distributions have a conjugate relationship with the Poisson distribution, making them a convenient choice for this purpose. Specifically, the probability density function for a gamma random variable, λ , with mean α/β and variance α/β^2 is given by

$$f(\lambda; \alpha, \beta) = \beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda} / \Gamma(\alpha)$$

and the prior probability density function of λ_{ij} is

$$f(\lambda_{ij}; \alpha_0, \beta_0, \alpha_1, \beta_1, P) = Pf(\lambda_{ij}; \alpha_0, \beta_0) + (1 - P)f(\lambda_{ij}; \alpha_1, \beta_1)$$

According to DuMouchel,^[20] “the exact choice of the prior distribution for $[\lambda_{ij}]$ is not so important as that it have several free parameters so that the distribution of $[\lambda_{ij}]$ can be fit using observed data.”

Under these assumptions, the marginal distribution of n_{ij} is a mixture of negative binomial distributions; namely

$$f(n_{ij}; \alpha_0, \beta_0, \alpha_1, \beta_1, P, E_{ij}) = Pf\left(n_{ij}; \alpha_0, \frac{E_{ij}}{\beta_0 + E_{ij}}\right) + (1 - P)f\left(n_{ij}; \alpha_1, \frac{E_{ij}}{\beta_1 + E_{ij}}\right)$$

Where

$$f\left(n_{ij}; \alpha, \frac{E_{ij}}{\beta}\right) = \binom{\alpha + n_{ij} - 1}{n_{ij}} \left(\frac{E_{ij}}{\beta + E_{ij}}\right)^{n_{ij}} \left(1 - \frac{E_{ij}}{\beta + E_{ij}}\right)^{\alpha-1}$$

The posterior distribution of λ_{ij} remains a mixture of gamma distributions, but with modified parameters. Let Q be the posterior probability that λ_{ij} comes from the first component of the mixture distribution, given n_{ij} . Then, from Bayes rule,

$$Q = \frac{pf\left(n_{ij}; \alpha_0, \frac{E_{ij}}{\beta_0 + E_{ij}}\right)}{pf\left(n_{ij}; \alpha_0, \frac{E_{ij}}{\beta_0 + E_{ij}}\right) + (1 - P)f\left(n_{ij}; \alpha_1, \frac{E_{ij}}{\beta_1 + E_{ij}}\right)}$$

This yields

$$f(\lambda_{ij}|n_{ij}) = Qf(\lambda_{ij}; \alpha_0 + n_{ij}, \beta_0 + E_{ij}) \\ + (1-Q)f(\lambda_{ij}; \alpha_1 + n_{ij}, \beta_1 + E_{ij})$$

as the posterior probability distribution of λ_{ij} given n_{ij} .

Maximum likelihood estimates of the five parameters $(\alpha_0, \hat{\beta}_0, \alpha_1, \beta_1, P)$ can be obtained using numerical methods, which can be computationally intensive. Since the parameters are estimated, the approach is empirical Bayes in nature. Once these estimates have been obtained, the posterior mixing parameter Q can also be estimated. The posterior means and variances of λ_{ij} are

$$E(\lambda_{ij}|n_{ij}) = \hat{Q} \frac{\hat{\alpha}_0 + n_{ij}}{\hat{\beta}_0 + E_{ij}} + (1 - \hat{Q}) \frac{\hat{\alpha}_1 + n_{ij}}{\hat{\beta}_1 + E_{ij}}$$

and

$$V(\lambda_{ij}|n_{ij}) = \hat{Q} \frac{\hat{\alpha}_0 + n_{ij}}{(\hat{\beta}_0 + E_{ij})^2} + (1 - \hat{Q}) \frac{\hat{\alpha}_1 + n_{ij}}{(\hat{\beta}_1 + E_{ij})^2}$$

Bearing in mind the information component $IC_{ij} = \log_2(RRR_{ij}) = \log_2(\lambda_{ij})$, the delta method allows us to write

$$E[\log_2(\lambda_{ij}|n_{ij})] \approx \log_2 E(\lambda_{ij}|n_{ij})$$

and

$$V[\log_2(\lambda_{ij}|n_{ij})] \approx \frac{1}{(\log 2)^2} \frac{V(\lambda_{ij}|n_{ij})}{[E(\lambda_{ij}|n_{ij})]^2}$$

This leads to the $100(1 - \alpha)\%$ confidence interval for λ_{ij} given by

$$\left(\exp \left[EBGM(\lambda_{ij}) - z_{\alpha/2} \frac{\sqrt{V(\lambda_{ij}|n_{ij})}}{E(\lambda_{ij}|n_{ij})} \right], \exp \left[EBGM(\lambda_{ij}) + z_{\alpha/2} \frac{\sqrt{V(\lambda_{ij}|n_{ij})}}{E(\lambda_{ij}|n_{ij})} \right] \right)$$

where $EBGM(\lambda_{ij}) = \log_2 E(\lambda_{ij}|n_{ij})$ is the empirical Bayes geometric mean of λ_{ij} . To assess whether the j -th AE can be interpreted as a signal of a disproportionate rate for the i -th drug, DuMouchel and Pregibon^[21] suggest using the fifth percentile of the distribution of $EBGM(\lambda_{ij})$, known as EB05, to rank signals by order of ‘interestingness’. For $n_{ij} > 0$, Szarfman, Machado et al.^[41] recommend a comparison of this percentile to a value of two to assess whether the j -th AE can be interpreted as a signal of a disproportionate rate for the i -th drug. For commonly-encountered situations

in which a significant number of the n_{ij} in the $I \times J$ table presented in Table 1 are zero, to ease computing time DuMouchel and Pregibon^[21] provide a modified likelihood that results from replacing the unconditional densities in $f(n_{ij}; \alpha_0, \beta_0, \alpha_1, \beta_1, P, E_{ij})$ by densities that are conditional

on $n_{ij} \geq n^*$, say. Specifically, $f\left(n_{ij}; \alpha_k, \frac{E_{ij}}{\beta_k + E_{ij}}\right)$ for $k = 0, 1$ are replaced by

$$f\left(n_{ij}; \alpha_k, \frac{E_{ij}}{\beta_k + E_{ij}}\right) / \left(1 - \sum_{m=0}^{n^*-1} f\left(m; \alpha_k, \frac{E_{ij}}{\beta_k + E_{ij}}\right)\right)$$

This replacement also yields the correct likelihood function for deriving maximum likelihood estimates when a minimum level of n^* is placed on n_{ij} .

Noren and Edwards^[42] introduced another IC_{ij} measure by also treating the number of reports n_{ij} , of the j -th AE with the i -th drug, $i = 1, \dots, I$ and $j = 1, \dots, J$, as independent Poisson random variables with means $\mu_{ij} = \lambda_{ij} \times E_{ij}$, with $\lambda_{ij} = RRR_{ij}$. The λ_{ij} are assumed to be independent with a gamma prior distribution $f(\lambda_{ij}; \alpha, \beta)$ with $\alpha = 0.5$ and $\beta = 0.5$, namely

$$f(\lambda_{ij}; 0.5, 0.5) = \frac{1}{\Gamma(0.5)\sqrt{2}} \lambda_{ij}^{-0.5} e^{-0.5\lambda_{ij}}$$

The posterior distribution of λ_{ij} given n_{ij} is also gamma. However, instead of $\alpha = 0.5$ and $\beta = 0.5$, the parameters are $0.5 + n_{ij}$ and $0.5 + E_{ij}$, respectively, so that

$$E[\log_2(\lambda_{ij}|n_{ij})] \approx \log_2 E(\lambda_{ij}|n_{ij}) = \log_2 \frac{0.5 + n_{ij}}{0.5 + E_{ij}}$$

It is instructive to compare this result to the estimate of IC_{ij} obtained under the frequentist approach; namely

$$\widehat{IC}_{ij} = \log_2 \left(\frac{n_{ij}}{E_{ij}} \right)$$

Note that the result of Noren and Edwards^[42] simply adds 0.5 to the numerator and denominator of the frequentist expression, which results in shrinkage towards $IC_{ij} = 0$. These authors state that “the impact is equivalent to that of ½ additional expected events, and an exactly matching increase in the observed number.” Noren, Bate et al.^[40] and

Noren, Sundberg et al.^[43] successfully applied this shrinkage strength to pattern discovery in the analysis of large collections of individual case safety reports.

Using the gamma posterior distribution for λ_{ij} given n_{ij} , credibility interval limits for λ_{ij} can be found numerically as solutions to the following equation for appropriate posterior quantiles (p):

$$\int_0^{\lambda_{ij}(p)} \frac{(0.5 + E_{ij})^{0.5+n_{ij}}}{\Gamma(0.5 + n_{ij})} \lambda_{ij}^{(0.5+n_{ij})-1} e^{-(0.5+n_{ij})\lambda_{ij}} d\lambda_{ij} = p$$

Noren and Edwards^[42] suggest setting $p = 0.025$ and $p = 0.975$ and specify the resulting values of $\lambda_{ij}(0.025)$ and $\lambda_{ij}(0.975)$ as the bounds of a two-sided 95% credibility interval for λ_{ij} . Since $IC_{ij} = \log_2(\lambda_{ij})$, this interval can then be transformed to one for the information component.

More recently, Huang, Zalkikar et al.^[44] introduced a simplified Bayes method that, similar to Noren and Edwards,^[42] also imposes a gamma prior distribution on λ_{ij} . Huang, Zalkikar et al.^[44] set $\alpha = \beta$, after specifying $0 < \alpha < 1$. Three choices of α selected were 0.5, 0.01, and 0.0001. The case where $\alpha = 0.5$ corresponds to a continuous uniform prior over (0, 1), which the authors described as the informative case. The other two cases were depicted as less informative, and non-informative, respectively.

The approaches of Noren and Edwards^[42] and Huang, Zalkikar et al.^[44] are fully Bayesian since the parameters of the gamma prior distribution are specified. By contrast, the BCPNN and MGPS methods are empirical Bayes in nature as these parameters are estimated.

5 ISSUES IN THE USE OF DISPROPORTIONALITY METHODS FOR SIGNAL DETECTION

When disproportionality methods are used for signal detection, there are a number of important issues that require careful contemplation. Two of the most commonly considered ones are discussed here; the choice of a signal threshold level, and the control of the Type I error rate when a number of drug/adverse event pairs are simultaneously investigated.

Signal Threshold Level

There has been much discussion in the literature on passive pharmacovigilance regarding the criteria that have been proposed to identify signals of adverse drug reactions. Historically, such criteria have been based on judgment about whether the estimated risk of an AE is sufficiently large to conclude that a signal is present. A decision is often made by determining if the lower bound of a confidence or credible interval indicates a significant increase in reporting rate over what would be expected under no

association. For example, in the above discussion of the proportional reporting ratio as one of the frequentist measures of disproportionality for signal detection, an approximate $100(1-\alpha)\%$ confidence interval was presented for this measure, and it was proposed that evidence of a signal would prevail if the lower limit of this interval exceeded one.

While a threshold level of unity appears reasonable when risk is measured in relative terms (on a log scale this cutoff value would be transformed to zero), not all studies involving such disproportionality methods abide by this criteria. In fact, Hauben, Madigan et al.^[45] point out that “there is currently no theoretical basis or firm empirical support establishing universal thresholds [for] defining a potential signal.” This lack of a theoretical basis is apparent when methodologies are compared. Ahmed, Haramburu et al.^[46] reiterate this point, noting that “all the proposed methods are based on arbitrary thresholds for signal generation where decision rules are not defined by any error-limiting constraint.”

An interesting study conducted by Deshpande, Gogolak et al.^[47] provides further evidence of a lack of universal agreement regarding signal detection thresholds. These authors performed a literature review to determine what thresholds were being used for signal detection in disproportionality methods. While documenting that the PRR, ROR, MGPS, and BCPNN were the most commonly used approaches, there was substantial variation in the threshold level used for each method. For example, the PRR, the most frequently-employed measure in the literature that was identified, also demonstrated the greatest variation in threshold levels to define a signal. Critical values of the PRR ranged from 1 to 3, including values 1, 1.5, 2, and 3. A number of investigations that used the PRR also required a chi-square test statistic to exceed a certain value (typically 4), with some added conditions for minimum values of n_{ij} . This latter requirement was also inconsistently applied from study to study, with the minimum being specified by different authors to be 2, 3, or higher.

For studies that use more than one disproportionality method for signal detection, it is important to note that while each method maintains a specified Type I error rate, the overall error rate will be greater, since two criteria must simultaneously hold in order to conclude that there is evidence of a signal. This needs to be taken into account when setting the individual Type I error rates for each method in order to ensure that the overall Type I error for signal detection is at the desired level.

In a similar vein, Huang, Guo et al.^[3] warn that the disproportionality methods described above “are not meant for analyzing multiple AEs for signals simultaneously. If that is the goal, one may use [a] Bonferroni correction formula (a family-wise error rate procedure) for preserving the overall Type I error; however, [if the number of AEs in the $I \times J$ table summary of the spontaneous reporting

data of interest is extremely large; for example, exceeding ten thousand as in the FAERS database], the Bonferroni tests would be very conservative.” As such, this Bonferroni family-wise error rate (FWER) has been deemed overly stringent for many applications. Benjamini and Hochberg^[29] describe some of the difficulties with FWER that result in its underutilization, and propose to control instead the expected proportion of errors among rejected hypotheses of no signal. They refer to this expected error proportion as the false discovery rate, FDR, which is described below.

False Discovery Rate

Consider the problem of simultaneously testing m hypotheses. For example, focusing on the $I \times J$ table summary of spontaneous reporting data, suppose we use the chi-square test of independence to investigate whether the j -th AE can be interpreted as a signal of a disproportionate rate for the i -th drug, for $i = 1, \dots, I$ and $j = 1, \dots, J$. As such, the particular m hypotheses to be tested are known in advance. For each of the m hypotheses, unknown to us, either the null hypothesis of independence between the drug and AE is true or not. When we test, we either reject or do not reject this null hypothesis, and in doing so, we could be correct, or in error. These various outcomes can be summarized in the following 2×2 table, where it is assumed that m_0 of the null hypotheses are true.

Note that while r can be ascertained upon completion of testing the m hypotheses, the values of u , v , t , and s are unknown since we do not know the true state in any of the tests. Therefore, defining R to be a random variable with observed value r , and U , V , T , and S as unobservable random variables associated with u , v , t , and s in the table, then $\text{FWER} = P(V \geq 1)$. Testing each individual null hypothesis at level α guarantees that $(V/m) \leq \alpha$, while testing each at level α/m guarantees that $P(V \geq 1) \leq \alpha$.

The proportion of errors committed when the null hypothesis is rejected (the proportion of times that the null hypothesis is rejected when it is true) can be viewed through the random variable $W = V/(V + S)$. Benjamini and Hochberg^[29] define the false discovery rate, FDR, as the expected value of W , given by

$$\text{FDR} = E(W) = E\left(\frac{V}{V + S}\right) = E\left(\frac{V}{R}\right)$$

Table 4 Summary of the results of m simultaneous hypothesis tests

True State	Test Result		Total
	Null Not Rejected	Null Rejected	
Null Hypothesis True	u	v	m_0
Null Hypothesis False	t	s	$m - m_0$
Total	$m - r$	r	m

To avoid division by zero, Benjamini and Hochberg^[29] propose an alternative formulation of the FDR in which there is conditioning on at least one null hypothesis being rejected ($R > 0$), namely

$$\text{FDR} = E\left(\frac{V}{R} | R > 0\right) P(R > 0)$$

Note that when all the null hypotheses are true, and $m_0 = m$, the FDR is equivalent to the FWER. However, if $m_0 < m$, the $\text{FDR} \leq \text{FWER}$. Therefore in this situation, a multiple testing procedure that controls the FDR at some level w^* , say, will be less stringent, making it more likely that true signals where the null hypothesis is false will be correctly detected, with a concomitant gain in power.

Returning to the example where m multiple hypothesis chi-square tests of independence are to be conducted, the FDR can be controlled at some specified level w^* as follows. Let H_l represent the l -th hypothesis to be tested, $l = 1, \dots, m$. Consider testing $H_1, \dots, H_m$ based on the corresponding p-values, which we shall denote as $P_1, \dots, P_m$. Let $P_{(1)} \leq P_{(2)} \leq \dots \leq P_{(m)}$ be the ordered p-values, and $H_{(l)}$ be the null hypothesis corresponding to $P_{(l)}$. If k is defined as the largest integer l for which $P_{(l)} \leq \frac{l}{m} w^*$, then the decision to reject all (l) for $l = 1, \dots, k$ will control the FDR at w^* .

Note that although the illustration above was presented in the context of the chi-square test for independence, the FDR can be controlled in the same fashion for other frequentist and Bayesian approaches that rely on confidence or credible intervals. As an example, consider the ROR, with a corresponding confidence interval based on the standard normal distribution. Suppose we wish to conduct m tests of hypothesis involving I drugs and J AEs, where, for the i -th drug and j -th AE, we wish to test the null hypothesis that $\text{ROR}_{ij} = 1$ (or $\log(\text{ROR}_{ij}) = 0$). The p-value for this test is given by

$$\text{p-value} = 2P\left[Z \geq \log(\widehat{\text{ROR}}_{ij}) / \sqrt{V[\log(\widehat{\text{ROR}}_{ij})]}\right]$$

While it is possible to determine appropriate p-values for the other disproportionality methods in a similar manner, it is also worth noting that the LRT proposed by Huang, Zalkikar et al.^[28] is able to control the FDR, as well as the overall Type I error rate.

Ahmed, Haramburu et al.^[46] Ahmed, Dalmasso et al.^[27] and Ahmed, Thiessard et al.^[35] describe the use of the FDR in various pharmacovigilance applications. They also note that since the seminal paper of Benjamini and Hochberg,^[29] a number of procedures have been proposed to estimate the FDR, citing an approach based on assuming a mixture model for the marginal distribution of the p-values resulting from the m tests of hypothesis as a particularly popular approach. Techniques for estimating the FDR employing this strategy include the q-value approach (Storey,^[48]

Storey,^[49] and Storey and Tibshirani^[50]); the method of Pounds and Morris,^[51] which assumes a beta-uniform mixture (BUM) for the marginal distribution of the p-values (BUM); the spacings LOESS histogram (SPLOSH) introduced by Pounds and Cheng^[52] (2004); and an general class of estimators proposed by Dalmasso, Broet et al.^[53] known as location based estimators (LBE).

It is worth noting that clinical AE data can be affected by another form of multiplicity. Adverse events have certain biological relationships within the context of the body system. Ignoring such biological structure may result in a loss of control over the FDR. Mehrotra and Heyse^[54] propose a two-step application of adjusted p-values based on the Benjamini and Hochberg^[29] approach. Alternatively, Berry and Berry^[55] subdivide the set of all AEs into body systems and apply a three level fully Bayesian model to the log odds ratio of patients receiving the *i*-th drug who are experiencing the *j*-th AE relative to patients who experiencing a particular AE, but not receiving the *i*-th drug. While these two methods were developed with clinical rather than SR data in mind, such data are relevant to the overall weight of evidence in support of a possible AE.

6 DISCUSSION

This article provides an overview of the most commonly-used passive-surveillance disproportionality methods for the detection of adverse events associated with the use of pharmaceutical products in spontaneous reporting databases. Both frequentist and Bayesian approaches are described. For each method, we describe how the technique can be applied for the purpose of signal detection using spontaneous reporting data. Important considerations in such applications are the choice of the signal threshold level, and control of the Type I error and false discovery rate when testing multiple hypotheses.

It is worth noting that various software packages are available for the application of most of the disproportionality methods summarized above. These include the JMP Clinical software from SAS^[56], an R package entitled *PhViD*, which serves as an acronym for pharmacovigilance signal detection (Ahmed and Ponset^[57]), and the software from the Observational Medical Outcomes Partnership (OMOP)^[16] methods library. The approaches covered by these packages include the reporting odds ratio, the proportional reporting ratio, reporting Fisher's exact test, the Bayesian confidence propagation neural network, and the multi Gamma Poisson shrinker.

A number of issues may arise when applying the disproportionality methods summarized above. Spontaneous reporting data is retrospective in nature: conditional on AE status, the drugs that the individual was taking in advance of making the report are observed. Spontaneous reporting data should therefore be analyzed as such.

The voluntary nature of SR data can result in under-reporting of AEs experienced by patients taking the drug of interest. The effects of under-reporting have been discussed in the literature (van der Heijden, Van Puijenbroek et al.,^[58] Hazell and Shakir^[59]), and methods for bias correction when using the PRR and ROR with supplementary information have been suggested (Ghosh and Dewanji^[60]). When describing the underlying mechanism responsible for SR data, in addition to the drug/AE relationship, Roux, Thiessard et al.^[31] acknowledged the effects of a number of factors, including the background incidence of the AE, the background drug exposure frequency, and the probability that the event is reported. All of these factors contribute to the potential for underreporting of a safety issue.

Disproportionality methods focus solely on the occurrence of the *j*-th AE with the *i*-th drug within a spontaneously generated safety report, without taking into account external factors such as co-morbidity. An excellent summary of the challenges encountered in the interpretation of the results from disproportionality approaches is given by van Manen, Fram et al.^[61] They highlight Simpson's paradox (Simpson^[62]), signal leakage (also known as 'signal absorption', or 'confounding by association'), confounding by indication, and masking as important issues to be considered when interpreting analysis of SR databases.

Simpson's paradox may manifest if an AE is more prevalent in a subgroup of the population that is also more likely to be exposed to a particular drug. Application of disproportionality methods to the general population in this case may suggest that the drug/AE combination occurs more frequently than expected. Stratification (cf. Mantel and Haenszel^[63]) is one approach that has been proposed to address this issue, where disproportionality measures are applied to each subgroup (or stratum), and then combined afterwards. A number of authors have considered extensions to disproportionality methods that involve stratification. For example, Huang, Zalkikar et al.^[27] describe a stratified analysis for the likelihood ratio test.

Consider further the situation in which two drugs are often administered together when treating a particular condition, and that one of the drugs is known to be related to a specific AE. Confounding by association may occur if application of disproportionality methods suggest that the other drug is also related to the AE due to the higher-than-expected joint occurrence of the two drugs. The statistical methods proposed by Noren, Sundberg et al.^[42] for evaluating drug-drug interactions in a surveillance context may be useful in addressing this issue. Another form of confounding (by indication) may occur when the AE is not associated with another drug, but rather with the disease being treated (Psaty, Koepsell et al.,^[64] Abrahamowicz, Bjerre et al.^[65]). Finally, masking refers to a situation where the association between a particular drug and AE is artificially diminished as a result of the fact that the database consists of other drugs that are observed with the same AE, but in much larger numbers (Maignen,

Hauben et al.^[66,67]). Some simple strategies for unmasking in ADR surveillance are discussed by Juhlin, Ye et al.^[68].

While active pharmacovigilance will play an increasing important role in drug safety evaluation in the future, passive pharmacovigilance remains an important tool for the early detection of safety signals (Chakravarty, Izem et al.^[16]). Of late, disproportionality methods have been extended to SR data with a longitudinal component. For example, Noren, Hosptadius et al.^[69] applied the fully Bayesian IC approach to study the temporal pattern discovery in longitudinal electronic patient records, while Schuemie^[70] (2011) proposed a longitudinal gamma Poisson shrinker. For additional details, see Zorych, Madigan et al.^[71] for an in-depth evaluation of different disproportionality methods for longitudinal observational databases.

ACKNOWLEDGEMENTS

The authors are grateful to the editor and a referee for useful comments on this manuscript. The research program of Patrick J. Farrell is supported by a grant from the Natural Sciences and Engineering Research Council. Christopher A. Gravel holds a Mitacs postdoctoral fellowship at McGill University, in partnership with Risk Sciences International. Daniel Krewski is the Natural Sciences and Engineering Research Council of Canada Chair in Risk Science at the University of Ottawa.

REFERENCES

1. IOM: Institute of Medicine. *The Future of Drug Safety: Promoting and Protecting the Health of the Public*. National Academies Press: Washington, DC, 2007.
2. IOM: Institute of Medicine. *Ethical and Scientific Issues in Studying the Safety of Approved Drugs*. National Academies Press: Washington, DC, 2012.
3. Huang, L.; Guo, T.; Zalkikar, J.N.; Tiwari, R.C. A review of statistical methods for safety surveillance. *Therapeutic Innovation & Regulatory Science* **2014**, *48* (1), 98–108.
4. Candore, G.; Juhlin, K.; Manlik, K.; Thakrar, B.; Quarcoo, N.; Seabroke, S.; Wisniewski, A.; Slattery, J. Comparison of statistical signal detection methods within and across spontaneous reporting databases. *Drug Safety* **2015**, *38* (6), 577–587.
5. WHO: World Health Organization. *The Importance of Pharmacovigilance: Safety Monitoring of Medicinal Products*, 2002. Office of Publications, World Health Organization, Geneva. <http://apps.who.int/iris/bitstream/10665/42493/1/a75646.pdf> (accessed October 24, 2016).
6. Montastruc, J.; Marque, P.; Moulis, F.; Bourg, V.; Lambert, V.; Durrieu, G.; Montastruc, J.L.; Montastruc, F. Adverse drug reactions of botulinum neurotoxin type A in children with cerebral palsy: A pharmaco-epidemiological study in VigiBase. *Developmental Medicine and Child Neurology* **2016**, doi: 10.1111/dmcn.13286. [Epub ahead of print].
7. Samyde, J.; Petit, P.; Hillaire-Buys, D.; Faillie, J.L. Quinolone antibiotics and suicidal behavior: analysis of the World Health Organization's adverse drug reactions database and discussion of potential mechanisms. *Psychopharmacology (Berl)* **2016**, *233* (13), 2503–2511.
8. Shanker, S. VigiAccess: Promoting public access to VigiBase. *Indian Journal of Pharmacology* **2016**, *48* (5), 606–607.
9. Huang, Y.L.; Moon, J.; Segal, J.B. A comparison of active adverse event surveillance systems worldwide. *Drug Safety* **2014**, *37* (8), 581–596.
10. Zhuo, L.; Farrell, P.J.; McNair, D.; Krewski, D. Statistical methods for active pharmacovigilance, with applications to diabetes drugs. *Journal of Biopharmaceutical Statistics* **2014**, *25*, 856–873.
11. Suissa, S.; Henry, D.; Caetano, P.; Dormuth, C.R.; Ernst, P.; Hemmelgarn, B.; Leloir, J.; Levy, A.; Martens, P.J.; Paterson, J.M.; Platt, R.W.; Sketris, I.; Teare, G. CNODES: The Canadian Network for Observational Drug Effect Studies. *Open Medicine* **2016**, *6* (4), e134–e140.
12. Ball, R.; Robb, M.; Anderson, S.A.; Dal, P.G. The FDA's sentinel initiative - A comprehensive approach to medical product surveillance. *Clinical Pharmacology & Therapeutics* **2016**, *99* (3), 265–268.
13. Azoulay, L.; Fillion, K.B.; Platt, R.W.; Dahl, M.; Dormuth, C.R.; Clemens, K.K.; Durand, M.; Hu, N.; Juurlink, D.N.; Paterson, J.M.; Targownik, L.E.; Turin, T.C.; Ernst, P.; and the Canadian Network for Observational Drug Effect Studies (CNODES) Investigators, Suissa, S.; Dormuth, C.R.; Hemmelgarn, B.R.; Teare, G.F.; Caetano, P.; Chateau, D.; Henry, D.A.; Paterson, J.M.; LeLorier, J.; Levy, A.R.; Ernst, P.; Platt, R.W.; Sketris, I.S. Association between incretin-based drugs and the risk of acute pancreatitis. *JAMA Internal Medicine* **2016**, *176* (10), 1464–1473.
14. Chakravarty, A.G.; Izem, R.; Keeton, S.; Kim, C.Y.; Levenson, M.S.; Soukup, M. The role of quantitative safety evaluation in regulatory decision making of drugs. *Journal of Biopharmaceutical Statistics* **2016**, *26* (1), 17–29.
15. Hauben, M.; Aronson, J.K. Defining 'signal' and its subtypes in pharmacovigilance based on a systematic review of previous definitions. *Drug Safety* **2009**, *32* (2), 99–110.
16. OMOP: Observational Medical Outcomes Partnership. Disproportionality analysis, 2012. OMOP Research Team. Organization, Geneva. <http://omop.org> (accessed January 14, 2017).
17. Evans, S.J.; Waller, P.C.; Davis, S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety* **2001**, *10*, 483–486.
18. Van Puijenbroek, E.P.; Bate, A.; Leufkens, H.G.; Lindquist, M.; Orre, R.; and Egberts, A.C. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and Drug Safety* **2002**, *11*, 3–10.
19. Rothman, K.J.; Lanes, S.; Sacks, S.T. The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiology and Drug Safety* **2004**, *13*, 519–523.
20. DuMouchel, W. Bayesian data mining in large frequency tables, with an application to the FDA Spontaneous Reporting System. *The American Statistician* **1999**, *53*, 177–190.

21. DuMouchel, W.; Pregibon, D. Empirical bayes screening for multi-item associations, KDD '01: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; ACM: New York, 2001, 67–76.
22. Bate, A.; Lindquist, M.; Edwards, I.R.; Olsson, S.; Orre, R.; Lansner, A.; De Freitas, R.M. A Bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology* **1998**, *54*, 315–321.
23. Yates, F. Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society*, **1934**, *1*, 217–235.
24. Yule, G.V.; Kendall, M.G. *An Introduction to the Theory of Statistics*, 14th Edition; Griffin: London, 1950.
25. Greenwood, P.E.; Nikulin M.S. *A Guide to Chi-Squared Testing*; Wiley: New York, 1996.
26. Agresti, A. *Categorical Data Analysis*, 2nd Edition; John Wiley & Sons, Inc: Hoboken, NJ, 2002.
27. Ahmed, I.; Dalmasso, C.; Haramburu, F.; Thiessard, F.; Broet, P.; and Tubert-Bitter, P. False discovery rate estimation for frequentist pharmacovigilance signal detection methods. *Biometrics* **2010**, *66*, 301–309.
28. Huang, L.; Zalkikar, J.N.; Tiwari, R.C. A likelihood ratio test based method for signal detection with application to FDA's drug safety data. *Journal of the American Statistical Association* **2011**, *106*, 1230–1241.
29. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **1995**, *57*, 289–300.
30. Efron, B.; Tibshirani, R.; Storey, J.D.; Tusher, V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **2001**, *96*, 1151–1160.
31. Rolka, H.; Bracy, D.; Russell, C.; Fram, D.; Ball, R. Using simulation to assess the sensitivity and specificity of a signal detection tool for multidimensional public health surveillance data. *Statistics in Medicine* **2005**, *24*, 551–562.
32. Roux, E.; Thiessard, F.; Fourrier, A.; Begaud, B.; Tubert-Bitter, P. Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance. *IEEE Transactions on Information Technology in Biomedicine* **2005**, *9*, 518–527.
33. Almenoff, J.S.; LaCroix, K.K.; Yuen, N.A.; Fram, D.; DuMouchel, W. Comparative performance of two quantitative safety signalling methods: implications for use in a pharmacovigilance department. *Drug Safety* **2006**, *29*, 875–887.
34. Matsushita, Y.; Kuroda, Y.; Niwa, S.; Sonehara, S.; Hamada, C.; Yoshimura, I. Criteria revision and performance comparison of three methods of signal detection applied to the spontaneous reporting database of a pharmaceutical manufacturer. *Drug Safety* **2007**, *7*, 715–726.
35. Ahmed, I.; Thiessard, F.; Miremont-Salame, G.; Begaud, B.; Tubert-Bitter, P. Pharmacovigilance data mining with methods based on false discovery rates: a comparative simulation study. *Clinical Pharmacology & Therapeutics* **2010**, *88*, 492–498.
36. Fisher, R.A. *The Design of Experiments*, 8th Ed.; 1966; Oliver & Boyd: Edinburgh, 1935.
37. Hochberg, A.M.; Reisinger, S.J.; Pearson, R.K.; O'Hara, D.J.; Hall, K. Using data mining to predict safety actions from FDA adverse event reporting system data. *Drug Information Journal* **2007**, *41* (5), 633–643.
38. Orre, R.; Lansner, A.; Bate, A.; Lindquist, M. Bayesian neural networks with confidence estimations applied to data mining. *Computational Statistics and Data Analysis* **2000**, *34* (4), 473–493.
39. Gould, A.L. Practical pharmacovigilance analysis strategies. *Pharmacoepidemiology and Drug Safety* **2003**, *12*, 559–574.
40. Noren, G.N.; Bate, A.; Orre, R.; Edwards, I.R. Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. *Statistics in Medicine* **2006**, *25*, 3740–3757.
41. Szarfman, A.S.; Machado, S.G.; O'Neill, T. Use of screening algorithm and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. *Drug Safety* **2002**, *25* (6), 381–392.
42. Noren, G.N.; Edwards, I.R. Opportunities and challenges of adverse drug reaction surveillance in electronic patient records. *Pharmacovigilance Review* **2010**, *4*, 17–20.
43. Noren, G.N.; Sundberg, R.; Bate, A.; Edwards, I.R. A statistical methodology for drug-drug interaction surveillance. *Statistics in Medicine* **2008**, *27*, 3057–3070.
44. Huang, L.; Zalkikar, J.N.; Tiwari, R.C. Likelihood ratio test based method for signal detection in drug classes using FDA's AERS database. *Journal of Biopharmaceutical Statistics* **2013**, *21*, 178–200.
45. Hauben, M.; Madigan, D.; Gerrits, C.M.; Walsh, L.; Van Puijenbroek, E.P. The role of data mining in pharmacovigilance. *Expert Opinion on Drug Safety* **2005**, *4*, 929–948.
46. Ahmed, I.; Haramburu, F.; Fourrier-Reglat, A.; Thiessard, F.; Kreft-Jais, C.; Miremont-Salame, G.; Begaud, B.; Tubert-Bitter, P. Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting. *Statistics in Medicine* **2009**, *28*, 1774–1792.
47. Deshpande, G.; and Gogolak, V.; Smith, S.W. Data mining in drug safety. *Pharmaceutical Medicine* **2010**, *24* (1), 37–43.
48. Storey, D. A direct approach to false discovery rates. *Journal of The Royal Statistical Society, Series B* **2002**, *64*, 479–498.
49. Storey, D. The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics* **2003**, *31* (6), 2013–2035.
50. Storey, J.D.; Tibshirani, R. Statistical significance for genomewide studies, *Proceedings of the National Academy of Sciences; USA*, **2003**; *100*, 9440–9445.
51. Pounds, S.; Morris, S.W. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* **2003**, *19*, 1236–1242.
52. Pounds, S.; Cheng, C. Improving false discovery rate estimation. *Bioinformatics* **2004**, *20*, 1737–1745.
53. Dalmasso, C.; Broet, P.; Moreau, T. A simple procedure for estimating the false discovery rate. *Bioinformatics* **2005**, *21*, 660–668.

54. Mehrotra, D.V.; Heyse, J.F. Use of the false discovery rate for evaluating clinical safety data. *Statistical Methods in Medical Research* **2004**, *13*, 227–238.
55. Berry, S.M.; Berry, D.A. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. *Biometrics* **2004**, *60*, 418–426.
56. SAS: JMP Clinical Software. http://www.jmp.com/en_us/software/jmp-clinical.html#Pharmacovigilance (accessed January 14, 2017).
57. Ahmed, I.; Poncet, A. Pharmacovigilance signal detection: R package PhViD. November 3, 2016.
58. van der Heijden, P.G.; Van Puijenbroek, E.P.; van Buuren, S.; van der Hofstede, J.W. On the assessment of adverse drug reactions from spontaneous reporting systems: the influence of under-reporting on odds ratios. *Statistics in Medicine* **2002**, *21*, 2027–2044.
59. Hazell, L.; Shakir, S.A. Under-reporting of adverse drug reactions: A systematic review. *Drug Safety* **2006**, *29*, 385–396.
60. Ghosh, P.; Dewanji, A. Analysis of spontaneous adverse drug reaction (ADR) reports using supplementary information. *Statistics in Medicine* **2011**, *30*, 2040–2055.
61. van Manen, R.P.; Fram, D.; DuMouchel, W. Signal detection methodologies to support effective safety management. *Expert Opinion on Drug Safety* **2007**, *6* (4), 451–464.
62. Simpson, E.H. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B* **1951**, *13* (2), 238–241.
63. Mantel, N.; Haenszel, M.W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **1959**, *22*, 719–748.
64. Psaty, B.M.; Koepsell, T.D.; Lin, D.; Weiss, N.S.; Siscovick, D.S.; Rosendaal, F.R.; Pahor, M.; Furberg, C.D. Assessment and control for confounding by indication in observational studies. *Journal of the American Geriatric Society* **1999**, *47*, 749–754.
65. Abrahamowicz, M.; Bjerre, L.M.; Beauchamp, M.E.; LeLorier, J.; Burne, R. The missing cause approach to unmeasured confounding in pharmacoepidemiology. *Statistics in Medicine* **2016**, *35*, 1001–1016.
66. Maignen, F.; Hauben, M.; Hung, E.; Holle, L.V.; Dogne, J.M. Assessing the extent and impact of the masking effect of disproportionality analyses on two spontaneous reporting systems databases. *Pharmacoepidemiology and Drug Safety* **2014**, *23*, 195–207.
67. Maignen, F.; Hauben, M.; Hung, E.; Holle, L.V.; Dogne, J.M. A conceptual approach to the masking effect of measures of disproportionality. *Pharmacoepidemiology and Drug Safety* **2014**, *23*, 208–217.
68. Juhlin, K.; Ye, X.; Star, K.; Noren, G.N. Outlier removal to uncover patterns in adverse drug reaction surveillance - A simple unmasking strategy. *Pharmacoepidemiology and Drug Safety* **2013**, *22*, 1119–1129.
69. Noren, G.N.; Hopstadius, J.; Bate, A.; Star, K.; Edwards, I.R. Temporal pattern discovery in longitudinal electronic patient records. *Data Mining and Knowledge Discovery* **2010**, *20*, 361–387.
70. Schuemie, M.J. Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. *Pharmacoepidemiology and Drug Safety* **2011**, *20*, 292–299.
71. Zorych, I.; Madigan, D.; Ryan, P.; Bate, A. Disproportionality methods for pharmacovigilance in longitudinal observational databases. *Statistical Methods in Medical Research* **2013**, *22* (1), 39–56.

Slope Approach: Assessment of Dose Proportionality and Linearity Under a Crossover Design

Bin Cheng

University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

Slope Approach for Assessment of Dose Proportionality/Linearity Under a Crossover Design In this entry, we consider tests for assessment of dose proportionality/linearity under a general crossover design based on the slope approach idea.

INTRODUCTION

In pharmaceutical research and development, dose response studies are often conducted not only to assess drug tolerance and safety in Phase I clinical development but also to characterize drug–response curve with respect to efficacy in phase II clinical development. Rodda et al.^[1] indicated that the clinical investigation of the dose–response curve can be classified into categories of (1) dose ranging study, which estimates the minimum effective dose and maximum tolerable dose; (2) dose response existence; and (3) dose response characterization, which describes the shape of the dose–response relationship and decides whether there is a clinically meaningful increase in response between the minimum and maximum effective doses. Ruberg^[2,3] provided a comprehensive review of design considerations, analysis, and interpretations of dose–response studies.

OVERVIEW

In practice, it is of interest to determine the relationship between the pharmacokinetic parameters and dose levels within a clinical relevant dose range (e.g., the range between the minimum effective dose and maximum tolerable dose). If the relationship is linear, the rate of change in pharmacokinetic parameters, which are commonly used to measure the rate and extent of drug absorption over a given range, is a constant. In this case, the dose curve is said to have a linearity property. Under dose linearity, the pharmacokinetic effect of a dose change can be predicted and the dose can be adjusted accordingly to achieve the optimal/desired magnitude of effect. If, in addition to dose linearity, the dose curve also passes the origin, then it is said to have a proportionality property. Under dose proportionality, the pharmaceutical parameter changes in direct proportion to

the dose level. Dose proportionality is stronger and often more desirable than dose linearity. However, dose linearity has its own merits. First, it is often the first step toward dose proportionality, and is consequently easier to establish. Second, dose linearity is often sufficient enough for making prediction about the pharmacokinetic parameter at a given dose level. Thus dose linearity studies over a given dose range have become increasingly important in pharmaceutical research and development.

In practice, dose proportionality or linearity could be conducted under a parallel-group design, a crossover design, or a combination of a parallel-group and a crossover design. For a parallel-group design, subjects are randomly assigned to one of the L dose groups being studied. Each subject will only receive the assigned dose level once. For a $K \times J$ crossover design, subjects are randomly assigned to one of the K sequences of dose levels. Each subject will receive different dose levels at J dosing periods. A sufficient length of washout is usually applied between dose levels. In many cases, when L , the number of dose levels being studied, is greater than K , the number of dosing periods, an incomplete block crossover design is often considered. In other words, each subject will not receive all the L dose levels but J dose levels at different dosing periods. Unlike a parallel-group design, the assessment of dose linearity under a crossover design has the following complications. First, the responses from each subject are not independent. Second, the variabilities associated with the dose levels may proportionally increase with respect to the dose levels. For more information about the crossover designs and their applications, see Refs. [4–6] and the references therein.

In this entry, we consider tests for assessment of doseproportionality/linearity under a general crossover design based on the slope approach idea suggested by Chow and Liu.^[4] In the next section, a statistical model under a

$K \times J$ crossover design and the idea of slope approach are introduced. An asymptotically valid test is derived in “Tests for Dose Proportionality/Linearity.” Formulas for sample size calculation based on a prestudy power analysis under the $K \times J$ crossover design are derived in “Sample Size Calculation.” A test for departure from linearity is given in “Test for Departure from Linearity.” To illustrate the tests in this entry, an example is considered in “An Example.” And some concluding remarks are made in the last section.

STATISTICAL MODEL AND SLOPE APPROACH

Let Y_{ijk} be the response of the i th subject in the k th sequence in the j th dosing period, where $i = 1, \dots, n_k$, $j = 1, \dots, J$, and $k = 1, \dots, K$. And let $D_1, \dots, D_L$ be the dose levels being studied. Because the design is an *incomplete* block crossover design, we have to introduce the following indicators to describe the model. For each fixed j, k , and l , define δ_{jkl} to be 1 if the j th dosing period of the k th sequence has dose level D_l ; otherwise $\delta_{jkl} = 0$. Let $X_{jk} = \sum_{l=1}^L \delta_{jkl} D_l$, then X_{jk} is nothing but the actual dose level received by patients at the j th dosing period of the k th sequence. The following repeated measure model is considered:

$$Y_{ijk} = \mu_{jk} + X_{jk}(e_{ik} + \varepsilon_{ijk}) \quad (1)$$

where $X_{jk}e_{ik}$ is used to model the random effect of the k th sequence and $X_{jk}\varepsilon_{ijk}$ is used to model the random errors. We assume that e_{ik} 's are independent, identically distributed (i.i.d.) as a normal random variable with mean 0 and variance σ_s^2 , i.e., $N(0; \sigma_s^2)$, ε_{ijk} 's are i.i.d. normally distributed as $N(0; \sigma_e^2)$, and e_{ik} 's and ε_{ijk} 's are mutually independent. For simplicity, we assume that there are no carryover effects. Note that statistical models for assessment of dose proportionality/linearity as described by Chow and Liu^[4] are special cases of the above model. For example, when $J = 1$, the above model reduces to a model for parallel group design. One important assumption made here is that the standard deviation of the dose response is proportional to the dose level. This assumption is reflected in model 1 as we have $\text{Var}(Y_{ijk}) = X_{jk}^2(\sigma_s^2 + \sigma_e^2)$.

Let μ_l be the mean response at the dose level D_l . Define

$$\theta_l = \frac{\mu_{l+1} - \mu_l}{D_{l+1} - D_l}, \quad \text{for } l = 1, \dots, L-1$$

$$\mathbf{M} = \begin{pmatrix} D_3 - D_2 & D_1 - D_3 & D_2 - D_1 & & & \\ & D_4 - D_3 & D_2 - D_4 & D_3 - D_2 & & \\ & & \ddots & & & \\ & & & D_L - D_{L-1} & D_{L-2} - D_L & D_{L-1} - D_{L-2} \end{pmatrix}_{(L-2) \times L} \quad (4)$$

then it is easy to see that $\theta_1, \dots, \theta_{L-1}$ are the adjacent slopes of the dose-response curve at dose level $D_1, \dots, D_L$. The hypotheses of dose linearity can then be formulated as

$$H_0: \theta_1 = \dots = \theta_{L-1} \quad \text{vs.} \quad H_1: H_0 \text{ is not true} \quad (2)$$

Failure to reject the null hypothesis of equal slopes would lead to the conclusion of dose linearity within the dose range under study.

If we define $\theta_0 = \mu_1/D_1$, then the hypotheses of dose proportionality are

$$H_0: \theta_0 = \theta_1 = \dots = \theta_{L-1} \quad \text{vs.} \quad H_1: H_0 \text{ is not true} \quad (3)$$

TESTS FOR DOSE PROPORTIONALITY/LINEARITY

Under model 1 as described above, a test for assessment of dose proportionality/linearity within the dose range under study can be derived. In this section, we give the derivation for testing the dose linearity (Eq. 2). The derivation for testing the dose proportionality (Eq. 3) is quite similar.

Define

$$\mu = (\mu_1, \dots, \mu_L)'$$

and (see Eq. 4 below).

Then, the hypotheses in Eq. 2 is equivalent to

$$H_0: \mathbf{M}\mu = 0, \quad \text{vs.} \quad H_1: \mathbf{M}\mu \neq 0 \quad (5)$$

Define

$$\hat{\mu}_l = \sum_{j=1}^J \sum_{k=1}^K \delta_{jkl} \bar{Y}_{jk}, \quad \hat{\mu} = (\hat{\mu}_1, \dots, \hat{\mu}_L)' \quad (6)$$

where $\bar{Y}_{jk} = 1/(n_k) \sum_{i=1}^{n_k} Y_{ijk}$. It can be shown that $\hat{\mu}$ has a multivariate normal distribution with mean μ and covariance matrix Σ , where

$$\Sigma_{ll} = \sum_{k=1}^K \left(\sum_{j=1}^J \delta_{jkl} X_{jk} \right)^2 \frac{\sigma_s^2}{n_k} + \sum_{k=1}^K \sum_{j=1}^J \delta_{jkl} X_{jk}^2 \frac{\sigma_e^2}{n_k} \quad (7)$$

and for $l \neq l'$

$$\Sigma_{ll'} = \sum_{k=1}^K \left(\sum_{j=1}^J \sum_{j'=1}^J \delta_{jkl} \delta_{j'kl'} X_{jk} X_{j'k} \right) \frac{\sigma_s^2}{n_k} + \sum_{k=1}^K \sum_{j=1}^J (\delta_{jkl} \delta_{jkl'} X_{jk}^2) \frac{\sigma_e^2}{n_k} \quad (8)$$

By applying the techniques described in Section 3.2 of Ref. [7] to the ratios Y_{ijk}/X_{jk} , $i = 1, \dots, n_k$; $j = 1, \dots, J$; $k = 1, \dots, K$, consistent estimators of σ_s^2 and σ_e^2 , denoted as $\hat{\sigma}_s^2$ and $\hat{\sigma}_e^2$, respectively, can be obtained. This leads to the following asymptotic test for Eq. 5 and hence for Eq. 2.

Theorem A. Let $\hat{\Sigma}$ be the estimator of Σ by replacing σ_s^2, σ_e^2 with $\hat{\sigma}_s^2, \hat{\sigma}_e^2$, respectively. Define

$$T = \hat{\mu}' \mathbf{M}' (\mathbf{M} \hat{\Sigma} \mathbf{M}')^{-1} \mathbf{M} \hat{\mu} \quad (9)$$

then an asymptotic level α test rejects H_0 in Eq. 5 if and only if $T > \chi_{L-2, \alpha}^2$, where $\chi_{m, \alpha}^2$ is the $(1 - \alpha)$ th percentile of a chi-square random variable with degrees of freedom m .

Proof: Note that $\hat{\mu}$ defined in Eq. 6 has a multivariate normal distribution. It is seen that the mean of $\hat{\mu}$ is μ , and the components of its covariance matrix Σ are determined as Eqs. [7 and 8] based on model 1. Then $\mathbf{M} \hat{\mu}$ has a multivariate normal distribution with mean $\mathbf{M} \mu$ and covariance matrix $\mathbf{M} \Sigma \mathbf{M}'$; therefore $\hat{\mu}' \mathbf{M}' (\mathbf{M} \Sigma \mathbf{M}')^{-1} \mathbf{M} \hat{\mu}$ has a central chi-square distribution with degrees of freedom $L - 2$, the rank of $\mathbf{M}$. Because $\hat{\Sigma}$ is a consistent estimator of Σ , T defined in Eq. 9 has an asymptotic chi-square distribution by Slutsky Lemma.

Note that the above result can also be found in Ref. [8] It should be noted that for testing dose proportionality in Eq. 3, the T statistic in Eq. 9 has degrees of freedom $L - 1$ instead of $L - 2$.

SAMPLE SIZE CALCULATION

Sample size calculation can be obtained based on a pre-study power analysis for achieving a desired power of $1 - \beta$ for testing Eq. 2 under model 1. For simplicity, we assume that $n_k \equiv n$; that is, there are equal numbers of patients in each dose sequence (Figs.1–3). Under assumption that $n_k \equiv n$, define

$$\lambda = n^{-1} \mu \mathbf{M}' (\mathbf{M} \Sigma \mathbf{M}')^{-1} \mathbf{M} \mu \quad (10)$$

which is independent of n . As a result, formulas for sample size calculation based on a prestudy power analysis can be derived.^[8]

Theorem B. Under the assumption that $n_k \equiv n$, to achieve a desired power of $1 - \beta$ for testing Eq. 2, the sample size n satisfies the following equation

$$(L - 2) + n\lambda = \chi_{L-2, \alpha}^2 + z_\beta \sqrt{2(L - 2) + 4n\lambda} \quad (11)$$

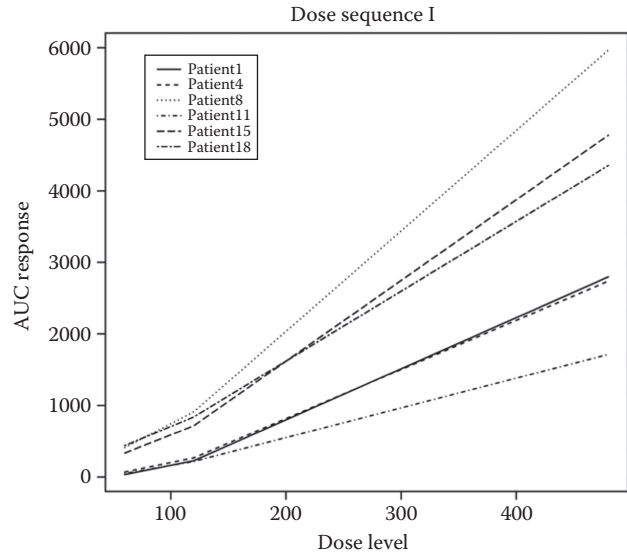


Fig. 1 Individual dose response curves for sequence I

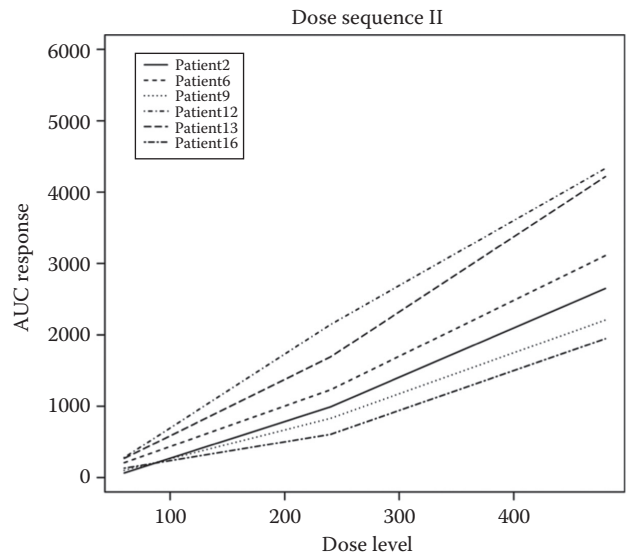


Fig. 2 Individual dose response curves for sequence II

or, approximately

$$n = 4 z_\beta^2 / \lambda$$

Proof: As indicated in Ref. [9] when n is large, the noncentral chi-square distribution can be approximated by a normal distribution as follows:

$$\chi_d^2(\lambda) = N(d + \lambda, 2(d + 2\lambda)) + O_p(\lambda^{-1/2}) \quad (12)$$

Then, under a fixed alternative in Eq. 5, we have

$$P(\chi_{L-2}^2(n\lambda) > \chi_{L-2, \alpha}^2) = 1 - \beta$$

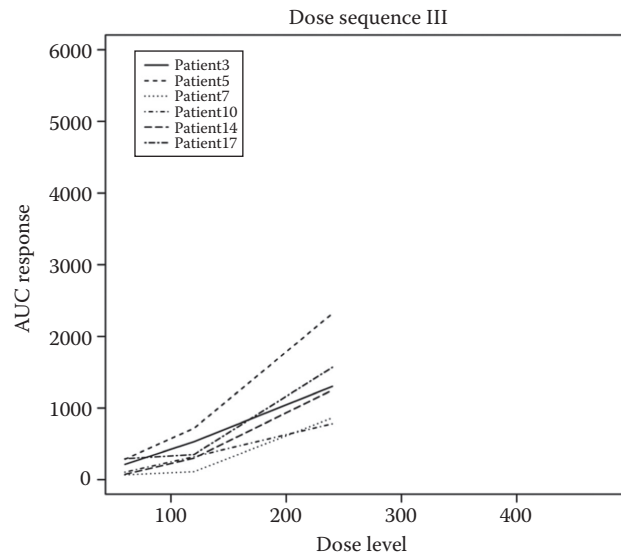


Fig. 3 Individual dose response curves for sequence III

By Eq. 12, the above equation is asymptotically equivalent to

$$\frac{\chi^2_{L-2,\alpha} - (L-2+n\lambda)}{\sqrt{2(L-2)+4n\lambda}} = -z_\beta$$

hence Eq. 11. Let $n \rightarrow \infty$, we get $n = 4z_\beta^2/\lambda$.

TEST FOR DEPARTURE FROM LINEARITY

As pointed out by Chow and Liu^[4] and Law,^[10] measuring departure from dose linearity is also of clinical importance. Law^[10] briefly described a way to test the departure from dose proportionality/linearity based on the parametric regression model. However, no analysis of variance (ANOVA)-based tests are available for testing departure from dose linearity. Following Cheng et al.,^[8] it can be seen that when $n_k \equiv n$, λ given in Eq. 10 is a measure of departure from dose linearity. In fact, $\lambda = 0$ corresponds to the dose linearity and a large value of λ indicates a serious departure from dose linearity. The hypotheses for testing departure from dose linearity can then be written as follows.

$$H_0: \lambda \geq \lambda_0 \text{ vs. } H_1: \lambda < \lambda_0 \quad (13)$$

where λ_0 is a prespecified critical relevant limit. An asymptotically valid test for testing Eq. 13 can be derived based on the T statistic given in Theorem A.

Theorem C. Assume $n_k \equiv n$. For testing Eq. 13, an asymptotic size α test rejected H_0 if and only if

$$T < \chi^2_{L-2,1-\alpha}(n\lambda_0)$$

where T is given in Eq. 9 and $\chi^2_{m,\alpha}(\theta)$ is the $(1 - \alpha)$ th percentile of a noncentral chi-square random variable with noncentrality parameter θ and degrees of freedom m .

Proof: The result follows from the fact that $\chi^2_{L-2}(n\lambda)$ is stochastically decreasing with respect to its noncentrality parameter $n\lambda$.

Note that more details can be found in Ref. [8].

AN EXAMPLE

To illustrate the proposed testing procedures for assessment of dose proportionality/linearity and departure from linearity, we consider the following example.

Example. A dose proportionality study was conducted on 18 subjects to evaluate dose proportionality or linearity of AUC between the dose range of 60 and 480 mg of a pharmaceutical compound. The study design is a 3×3 incomplete crossover design with 4 dose levels, 3 dose sequences, and 3 dosing periods. Subjects were randomly assigned to one of the 3 sequences, and each subject has AUC responses at 3 dosing periods. The AUC data are given in Table 1. The plots of individual dose response curves for the 3 dose sequences are given in Figs. 1–3. From the plots, we see that the dose curve is not a straight line. Based on the above AUC data, the estimated quantities are

$$\hat{\mu} = (190.57, 459.7, 1297.75, 3400.98),$$

Table 1 AUC Data

Sequence	Subject	Dose levels (mg)			
		60	120	240	480
1	1	35.25	227.95		2797.65
1	4	70.50	268.30		2738.00
1	8	412.05	911.90		5967.25
1	11	49.85	218.60		1714.10
1	15	334.60	717.00		4777.30
1	18	439.00	839.85		4354.10
2	2	63.45		990.70	2649.60
2	6	207.35		1229.65	3110.15
2	9	99.70		829.80	2207.95
2	12	280.30		2144.20	4332.15
2	13	268.70		1690.15	4217.55
2	16	130.55		605.85	1945.90
3	3	213.85	529.90	1302.15	
3	5	285.95	717.60	2318.40	
3	7	66.20	111.50	862.50	
3	10	105.20	321.55	780.65	
3	14	74.45	301.10	1249.60	
3	17	293.35	351.50	1569.30	

and

$$\hat{\sigma}_s^2 = 5.12, \quad \hat{\sigma}_e^2 = 0.69, \quad T = 7.74.$$

Since $T = 7.74 > \chi_{2,0.05}^2 = 5.99$, we reject the null hypothesis that the dose curve is linear. When the dose linearity hypothesis is rejected, a test for departure from linearity may be of interest. If we choose $\lambda_0 = 3.5$ in Eq. (13), then $T = 7.74 < \chi_{2,0.95}^2(21) = 9.44$. Therefore the null hypothesis in Eq. (13) is rejected and we conclude that the dose curves are not seriously different from a straight line under the chosen departure upper limit $\lambda_0 = 3.5$. It should be quickly noted that, in general, λ_0 should be chosen by authority group(s) based on the nature and particular purpose of the study.

DISCUSSION

In model 1, we assume that the standard deviation of the dose-response Y_{ijk} are proportional to the dose level X_{jk} . In practice, a logarithm transformation is usually suggested to achieve the constant variance. However, under model 1, the dose linearity means μ_{jk} follows a straight line. Therefore it may not be a good idea to consider a transformation of Y_{ijk} as the transformation will destroy the linearity of μ_{jk} .

The slope approach is essentially an ANOVA-based approach, which does not model the underlying dose curve as regression-based approaches do. Therefore the result for testing Eq. 2 need to be interpreted with caution. When H_0 in Eq. 2 is rejected, it can be concluded that the dose curve is not linear; while if H_0 is not rejected, to conclude that the dose curve is linear, some additional assumptions about the dose curve (e.g., that the dose curve is convex within the dose range under study) are generally understood.

As the dose linearity formulation in this entry is based on the adjacent slopes θ_l , it is also referred to as an adjacent slope approach, which is a special case of the slope approach first introduced by Chow and Liu.^[4] It should be noted that many other types of slopes could be considered. For example, we may define

$$\psi_l = \frac{\mu_{l+1} - \mu_l}{D_{l+1} - D_l} \text{ for } l = 1, \dots, L-1$$

which are the slopes with respect to the initial dose level D_1 . Thus the dose linearity test can be formulated as

$$H_0 : \psi_1 = \dots = \psi_{L-1} \text{ vs. } H_1 : H_0 \text{ is not true}$$

while the dose proportionality test can be formulated as

$$H_0 : \frac{\mu_1}{D_1} = \dots = \frac{\mu_L}{D_L} \text{ vs. } H_1 : H_0 \text{ is not true}$$

This approach is known as a baseline slope approach. The comparison of the baseline slope approach with the adjacent slope approach can be found in Ref. [11].

REFERENCES

1. Rodda, B.E.; Tsianco, M.C.; Bolognese, J.A.; Kersten, M.K. Clinical Development. In *Biopharmaceutical Statistics for Drug Development*; Peace, K.E., Ed.; Marcel Dekker, Inc.: New York, 1988.
2. Ruberg, S.J. Dose response studies I. Some design considerations. *J. Biopharm. Stat.* **1995**, *5*, 1–14.
3. Ruberg, S.J. Dose response studies II. Analysis and interpretation. *J. Biopharm. Stat.* **1995**, *5*, 15–42.
4. Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*, 2nd Ed.; Marcel Dekker, Inc.: New York, 2000.
5. Chow, S.C.; Shao, J. *Statistics in Drug Research—Methodologies and Recent Developments*; Marcel Dekker, Inc.: New York, 2002.
6. Jones, B.; Kenward, M.G. *Design and Analysis of Crossover Trials*, 2nd Ed.; Chapman & Hall/CRC: New York, 2003.
7. Vonesh, E.F.; Chinchilli, V.M. *Linear and Nonlinear Models for the Analysis of Repeated Measurements*; Marcel Dekker, Inc.: New York, 1997.
8. Cheng, B.; Chow, S.C.; Wang, H.S. *Testing Dose Linearity Under Incomplete Block Cross Over Design*; 2003. Manuscript submitted.
9. Johnson, N.L.; Kotz, S.; Balakrishnan, N. *Continuous Univariate Distributions*, 2nd Ed.; Wiley & Sons: New York, 1995.
10. Law, C.G. Dose Proportionality. In *Encyclopedia of Biopharmaceutical Statistics*; Chow, S.C., Ed.; Marcel Dekker, Inc.: New York, 2000, 167–170.
11. Cheng, B.; Chow, S.C. *Comparison of Adjacent Slope Approach and Baseline Slope Approach for Assessing Dose Proportionality/Linearity*; 2003. Manuscript in preparation.
12. Budde, M.; Bauer, P. Multiple test procedures in clinical dose finding studies. *J. Am. Stat. Assoc.* **1989**, *84*, 792–796.
13. Chow, S.C.; Liu, J.P. *Design and Analysis of Clinical Trials: Concepts and Methodologies*; Wiley & Sons: New York, 1998.

Spatio-Temporal Modeling

Ying C. MacNab

Department of Health Care and Epidemiology, University of British Columbia, Vancouver, British Columbia, Canada

Charmaine B. Dean

Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada

Abstract

This entry outlines models and methods for spatio-temporal analyses of spatially aggregated area rates, a topic of substantial development in recent years and summarizes several different processes of this modeling technique.

INTRODUCTION

Spatio-temporal models and related methods of inference have seen considerable development over the past decade primarily due to developments in statistical methodology for modeling spatial data and Bayesian computational methods, as well as accessibility of geographical information systems and mapping techniques. While there is a large body of work in the broad area of spatio-temporal models, this entry outlines models and methods for spatio-temporal analyses of spatially aggregated area rates, a topic of substantial development in recent years. A brief summary is also given of the models for spatial Gaussian processes, point pattern processes, and latent processes (reflecting hidden structure), respectively. The methods discussed herein have seen applications in disease epidemiology,^[1] disease surveillance and prevention,^[2,3] health outcomes studies,^[4,5] and environmental health research.^[1,4]

The broad focus of the spatial analysis of disease incidence, mortality, or health outcome rates is on visually describing the spatial distribution of small-area rates over space and time. The goal of such analyses is to produce smoothed maps for identifying spatial and temporal risk trends and to suggest factors that may be linked to various diseases or health outcomes for further epidemiological investigation or health outcome study. Often discussed under the subject title of “disease mapping,” the methodology has been developed to analyze spatially aggregated data, typically in the form of counts for the response, Y_{jt} , $j = 1, \dots, J$, $t = 1, \dots, T$, which represent the number of cases/events for each of the J subregions that constitute the region under study, with the data being collected over a consecutive set of time intervals such as months or years indexed by t . Also available is a set of counts N_{jt} , $j = 1, \dots, J$, $t = 1, \dots, T$, representing the “at-risk” population for the j th subregion in time period t . Age effects, and fixed or time-varying

covariates can also be incorporated into an analysis. For example, the methodology can be used to study geographical or regional variation in small-area prevalence rates of a drug prescription among a general population of a particular region under study, or differences in regional variation in utilization rates of a specific medication between males and females in a set of subregions.

Statistical models and methods for spatio-temporal analyses remain a relatively new arena of development. Some useful references in this area are Bernardinelli et al.,^[6] Waller et al.,^[7] Xia and Carlin,^[8] Knorr-Held and Besag,^[9] Sun et al.,^[10] Pickle,^[11] MacNab and Dean.^[12] Spatio-temporal models for spatially referenced point processes are common in environmental epidemiology and disease surveillance.^[13] Extensive treatment of models and methods for spatial data can be found in books by Cressie^[14] and Lawson et al.^[4] Edge effects in mapping studies are also discussed by Cressie.^[14] Wakefield and Morris^[15] discuss problems caused by aggregation and inaccuracies due to data collection, with specific reference to how these potentially lead to ecological bias and confounding.

SPATIO-TEMPORAL MODELS FOR AREA-AGGREGATED RATES

Spatial Models with Random Effects

Consider a common spatial model that involves a response $Y = (Y_1, \dots, Y_J)$ of independent Poisson random variables, corresponding to J local regions given a set of small-area random effects $\alpha = (\alpha_1, \dots, \alpha_J)$ and the “at-risk” population N_j , $j = 1, \dots, J$ for a particular time period considered (e.g., 1 or 5 years). The small-area random effects represent small-area relative risks which are generally of prime interest in these studies. Using a logarithmic link function in a generalized

linear mixed model framework, the conditional Poisson mean $\mu^\alpha = E(Y|\alpha)$ can be expressed as:

$$\log(\mu^\alpha) = X\beta + \alpha + \text{offset} \quad (1)$$

where X represents the design matrix and β the regression coefficients of fixed effects. The spatial random effects α reflect extra components of variability associated with the subregions under study, for example, variability arising from unobserved subregion covariates or latent effects. In a Poisson model, the offset represents a measure of exposure; here it is the logarithm of the population size. It is usual to account for age and gender in mapping studies, and model (1) can be extended in this manner.^[3,8] When the disease or mortality counts Y are aggregated over age and gender, an expected count is computed for each small area, usually based on the age- and sex-specific “at-risk” population sizes and standard rates for each age–sex stratum. This expected count is then used as the offset, and the relative risks α reflect a spatial comparison between observed and expected counts.^[1,4] When rates are not rare, the binomial mixed model is preferable.^[16]

A variety of distributional forms for the random effects have been proposed and discussed in the disease mapping literature. A simple spatial model proposed by Besag, York, and Mollié^[17] is the intrinsic Gaussian autoregressive (IGA) model, which accounts for what is termed locally structured spatial variation. Assuming a multivariate normal distribution for α , the IGA model describes the conditional distribution of α_j , given all other random effects, $[\alpha_i | \alpha_i, i \neq j]$, as depending only on those spatial random effects in a small local neighborhood of the j th subregion:

$$\alpha_j | \{\alpha_i, i \neq j\} \sim N(\bar{\alpha}_{(j)}, \sigma_\alpha^2 / \delta_j) \quad (2)$$

where $\bar{\alpha}_{(j)}$ represents the mean of the random effects corresponding to the subregions that belong to the “neighborhood” of the j th area and δ_j is the number of areas forming this neighborhood. The conditional variance in (2), $V(\alpha_j | \{\alpha_i, i \neq j\})$, is inversely proportional to the number of neighbors of the j th area. Neighborhoods may be defined in various ways, depending on the context of the analysis, but one common definition is simply the set of areas that border the j th area. Unconditionally, the IGA model becomes

$$\alpha \sim N(0, \Sigma) \quad (3)$$

where the generalized inverse of Σ is $\Sigma^{-1} = Q/\sigma_\alpha^2$; Q is a square matrix of dimension J determined by the neighborhood structure of the regions. Under the given specification (2) for example, the neighborhood matrix Q has j th diagonal element δ_j , and the off-diagonal elements Q_{jk} , $j \neq k$, equal to -1 if the subregions j and k are neighbors, and 0 otherwise.

More general formulations define the mean $E(\alpha_j | \{\alpha_i, i \neq j\})$ as a weighted average of all the other area-specific effects. To give more influence to “local” areas, the weights may vary with distance from area j :

$$\alpha_j | \{\alpha_i, i \neq j\} \sim N\left(\frac{\sum_i W_{ij} \alpha_i}{\sum_i W_{ij}}, \frac{\sigma_\alpha^2}{\sum_i W_{ij}}\right) \quad (4)$$

where $\{W_{ij}, i = 1, \dots, J, j = 1, \dots, J, i \neq j\}$ are user-specified weights linking subregions i and j .^[15,18]

As with any generalized linear mixed model (GLMM), various components of variability may also be included. A common extension to the IGA model uses:

$$\Sigma = \sigma_\alpha^2 Q^{-1} + \sigma_u^2 I_J \quad (5)$$

where I_J is an identity matrix of dimension J ; σ_α^2 models spatially structured variation as previously, while the additional component σ_u^2 incorporates unstructured variation.^[17] Leroux, Lei and Breslow^[19] offer an alternative formulation to (5), which builds upon the idea of conditional modeling, where it is the inverse of the variance that is modeled as having these two components, so that conditional distributions of random effects are as easily interpreted as in the IGA model. In this case

$$\Sigma = \sigma^2 D^{-1}, \quad D = \lambda Q + (1 - \lambda) I_J \quad (6)$$

Here, σ^2 is the dispersion parameter and λ , $1 \geq \lambda \geq 0$, measures the spatial autocorrelation. It should be noted that model specification (6) reduces to a Gaussian independence/exchange model if $\lambda = 0$ or to the IGA model if $\lambda = 1$. The Gaussian independence assumption, on one hand, models spatially “unstructured” variation and leads to “global” smoothing, and the IGA model, on the other hand, assumes purely neighborhood (spatially) “structured” effects and results in maximum “local smoothing.” By allowing λ to vary between 0 and 1, the conditional autoregressive (CAR) specification (6) facilitates spatial smoothing in accordance with the extent to which areas of close spatial proximity have similar underlying disease rates.^[5,16] MacNab^[5] presents a comparison of random spatial effect models under assumptions (5) and (6).

Spatio-Temporal Models

Let j denote the indices for area units, or area–age units, if age effects are included. General spatio-temporal models with Poisson intensity $\mu(j, t)$, conditional on random effects $\mathbf{b}$, can be expressed as:

$$\mu(j, t) = \rho \lambda_0(j, t) \lambda_j(j; \boldsymbol{\beta}^{(j)}, \mathbf{b}^{(j)}) \lambda_{j,t}(j, t; \boldsymbol{\beta}^{(j,t)}, \mathbf{b}^{(j,t)}), \quad (7)$$

where ρ represents the offset, $\lambda_0(j, t)$ the area–time (or area–age–time or age–time) conditional intensity of the “at-risk” population, $\lambda_j(j; \beta^{(j)}, b^{(j)})$ the spatial variability attributable to time-constant covariates (with regression coefficients $\beta^{(j)}$) and residual spatial effects ($b^{(j)}$), and, assuming a log-linear form for $\lambda_{j,t}$ to aid in exposition,

$$\lambda_{j,t}(j, t; \beta^{(j,t)}, b^{(j,t)}) = \exp[X(j, t)\beta^{(j,t)} + \gamma(j, t)] \quad (8)$$

with $X(j, t)$ denoting time–varying covariates and $\gamma(j, t)$ representing area–time (or area–age–time) residual variation, thus allowing for area–time interactions. Corresponding models for the binomial likelihood can be handled.^[20]

Specific spatio-temporal random effect Poisson models for the analysis of small-area disease rates have been proposed and discussed by Waller et al.,^[7] Sun et al.,^[10] MacNab and Dean,^[12] among others. In particular, the aforementioned authors extend spatial autoregressive models to include temporal components. In Ref. [7], the index j refers to area–stratum units, where the strata may be gender and race, for example. The area–stratum covariates are fixed over time, and the stratum covariates are fixed over area. This form would assume that sex and race effects are not affected by area and time. The logarithm of λ_j in (7) is written as a sum: $X_{j1}\beta_1 + X_{j2}\beta_2$, where X_{j1} is the design matrix for the stratum effects, including interactions, for example, race–gender interactions, and X_{j2} is the design matrix for the fixed area effects. The space–time interaction model defines the logarithm of $\lambda_{j,t}$ as including both spatial and unstructured random effects that vary over time. For example, the spatial random effects for given t may follow an IGA model. Simpler models are also considered in an analysis of Ohio lung cancer including one with $\lambda_{j,t}$ being a linear function of t and without area or temporal random effects. Other models mentioned in the paper include the temporal effect in an autoregressive AR(1) structure. Xia and Carlin^[8] extend Waller et al.^[7] to allow for errors in covariates.

In Ref. [10] where the index j refers to area–age units, temporal trends are modeled as linear functions of t . The logarithm of λ_j includes fixed age and random spatial effects, the latter following a modification of the IGA model. The spatio-temporal component permits a different trend effect over age groups, i.e., the fixed coefficients of the linear temporal effect differ over age groups. There are also random area coefficients of the linear temporal effect permitting random area trends over time.

Nonlinear forms for temporal trends are accommodated in MacNab and Dean^[12] via spline smoothing. Splines^[21–23] permit a broad class of arbitrary smooth functions of time. The logarithm of λ_0 includes a cubic spline that models the trend over the whole region considered, or, if age-group effects are included, a two-dimensional fixed spline to represent the overall trend over the age–time surface. Random temporal spline effects are included in $\lambda_{j,t}$, permitting the

random area relative risks to vary over time in a nonlinear fashion. Simpler spatio-temporal effects are also considered, for example, having the random area relative risks vary linearly over time as in the other spatio-temporal models above. Large positive and negative values of these linear area trend terms then identify areas whose rates are increasing the fastest, or decreasing the slowest with respect to the overall trend over the entire area.

ESTIMATION

Empirical Bayes Methods

The above mentioned models can be readily represented by the following GLMM framework:^[24]

$$g(\mu^b) = Xa + Zb + \text{offset} \quad (9)$$

where $\mu^b = E(Y|b)a$, and b indicate the fixed and random effects, and X and Z the corresponding design matrices. Estimation can be implemented using empirical Bayes techniques that make use of likelihood-based procedures, such as maximum likelihood estimation (MLE) [or restricted maximum likelihood estimation (RMLE)]^[25] to estimate hyperparameters, and a “plug-in approximation” to the posterior inference of the random effects.

Full maximum likelihood estimation for random effects Poisson or binomial models often requires numerical integrations that are intractable. For computational simplicity, the use of penalized quasilielihood (PQL), a second order Laplace approximation to the integrated quasilielihood function, may be used. The PQL method was popularized by Breslow and Clayton^[24] for inference within the GLMM framework; the algorithm has been used in empirical Bayes disease mapping estimation for Poisson count data.^[12,19] Specifically, by introducing a modified working response variable:

$$Y^* = \log y + (y - \mu^b)/\mu^b - \log N \quad (10)$$

PQL fitting proceeds by considering Y^* to be derived from an associated normal theory mixed model.^[24]

The PQL algorithm typically yields consistent and nearly unbiased point estimates of the relative risks under Poisson likelihood. For nonrare outcomes and random effects binomial models, the second order Laplace approximation to the corresponding integrated quasilielihood function may result in considerable bias in the PQL estimators. In such cases, a higher-order Laplace approximation to the corresponding integrated quasilielihood function might be explored for bias corrections. Potential alternatives for the GLMM inference also include double PQL^[26] and the so-called expectation–maximization (EM) algorithm.^[27,28]

Full Bayesian Methods

The disease mapping literature also presents numerous applications of Markov chain Monte Carlo (MCMC) methods for full Bayesian (FB) estimation of relative risks. Unlike the naive empirical Bayes' methods, a fully Bayesian approach permits inference for relative risks based on estimated posterior distributions where the uncertainty associated with the estimates is properly characterized.^[29] Fully Bayesian disease mapping inference is commonly implemented via Gibbs and adaptive rejection sampling. The methods, however, may suffer from slow mixing when high posterior correlations occur, or when a large number of geographical areas are investigated and the data are sparse.^[4,29] More recently, alternative MCMC algorithms for Bayesian inference of relative risks have been explored.^[16,30]

OTHER SPATIO-TEMPORAL MODELS

Models for Spatial Gaussian Processes

The basic spatial models for continuous data assume normality of the spatial responses. Let $Z(s_i) = Z_i$ and μ_i represent the responses and their means at location s_i . The conditional specification of the model assumes that conditional on $\{Z_j, j \neq i\}$, the distribution of $Z_i, i = 1, \dots, n$ is normal with mean of the form

$$E\{Z_i | (Z_j, j \neq i)\} = \mu_i + \sum_{j=1}^n c_{ij}(Z_j - \mu_j)$$

and variance

$$\text{Var}\{Z_i | (Z_j, j \neq i)\} = \sigma^2$$

Here, $c_{ii} = 0$, $c_{ij} = c_{ji}$, and the mean usually depends only on responses in a small neighborhood of site s_i with $c_{ij} = 0$ unless sites s_i and s_j are neighbors. When $c_{ij} > 0$, there is positive interaction between sites; $c_{ij} < 0$ implies negative interactions. If a regression form for the mean is assumed, so

$$\mu_i = \beta_0 + \beta_1 x_1(s_i) + \beta_2 x_2(s_i) + \dots + \beta_p x_p(s_i)$$

where $x_j(s_i)$ is the value of the j th explanatory variable at site i , then the joint distribution of the responses is of the multivariate regression form:

$$Z \sim N(X\beta, \sigma^2(I - C)^{-1})$$

where X is

$$\begin{pmatrix} 1 & x_1(s_1) & \dots & x_p(s_1) \\ & & \ddots & \\ 1 & x_1(s_n) & \dots & x_p(s_n) \end{pmatrix}$$

I is the identity matrix, and C is the matrix with ij th element c_{ij} . Simple models define $c_{ij} = 1$ if i and j are neighbors, and 0 otherwise. Other models permit c_{ij} to depend on a measure of distance between s_i and s_j and vary the structure of the conditional variance. If the sites do not occur as a regular lattice then a conditional variance that takes into account the cardinality of the neighborhood may be more appropriate. Another class of models for continuous data uses simultaneous (instead of conditional) autoregressive models. Under the simultaneous autoregressive model (known as the SAR model^[14]),

$$[Z(s_i) - \mu_i] - \sum_{j=1}^n b_{ij}[Z(s_j) - \mu_j] = \epsilon(s_i)$$

where $\epsilon(s_1), \dots, \epsilon(s_n)$ are independent normal variates with zero mean and variance σ^2 . The coefficients b_{ij} must satisfy: $b_{ii} = 0$, $b_{ij} = 0$ unless sites s_i and s_j are neighbors. The joint distribution of the responses becomes multivariate normal with mean μ and variance $\sigma^2(I - B)^{-1}(I - B')^{-1}$ (B' indicates transpose of B) where I is the identity matrix; B has elements b_{ij} . Analogs to time-series moving average models are also available in the spatial context. Extensions to consider spatio-temporal analysis follow similarly as for count data. For example, autoregressive models for the temporal dimension could be explored or models incorporating splines could be used.

Relatively well-developed models for spatial processes and continuous responses have been used in the analyses of geostatistical data.^[14] The models and methods have been applied to public health research such as the analysis of small-area sudden-infant death syndrome data.^[14,31]

Models for Point Pattern Processes

Models for point pattern processes are developed for case-event data, with risk intensity of cases and the corresponding locations of events being of particular interest.^[4,13] Investigation of disease risks through point process modeling requires cases and event specific data at the individual level, permitting the assessment of spatial risk patterns and risk clustering. Spatio-temporal models of case-event data remain relatively unexplored. Recently, Diggle et al.^[13] discussed and explored spatio-temporal point process models for spatial disease surveillance.

Discrete Mixture Models

Discrete mixture spatial models^[32-34] are also common for mapping small-area rare rates and for cluster analysis. This nonparametric mixture model approach assumes that the region-specific random effects have a discrete probability distribution taking k values $R_1, \dots, R_k$ with probabilities $p_1, \dots, p_k$, respectively. The mixture model represents k latent components or clusters in relative risks containing a

proportion p_j from the population. If k is fixed, estimation via the EM algorithm is straightforward.^[27,28] Flexibility in the number of clusters permits the isolation of extreme relative risks and therefore enables the assessment of whether a region is in a high-risk cluster.^[33] Spatio-temporal mixture models for mapping disease risks over space and time are explored and discussed in Böhning, Schlattmann, and Lindsay.^[34] Green and Richardson^[35] have recently proposed the use of hidden Markov random field models for the analysis of rare disease rates so that spatial correlation in the probabilities of belonging to various clusters is achievable.

AN EXAMPLE

We present here an illustrative spatio-temporal analysis of injury hospitalization rates among children of age 0–19 in British Columbia (BC), Canada. Specifically, we present here an analysis of child injuries due to accident falls. The data consist of annual incidence cases y_{ijt} of accident falls and the midyear “at-risk” population estimates n_{ijt} for young girls of age 0–4, 5–14, and 15–19 ($i = 1, 2, 3$) in 20 health regions in British Columbia ($j = 1, \dots, 20$), and from 1987 to 1996 ($t = 1987–1996$), respectively. Conditional on random effects $\mathbf{b}$, we assume a Poisson likelihood for the observations y_{ijt} with mean:^[3,12]

$$\mu_{ijt} | \mathbf{b} = n_{ijt} m \exp[S_0(t) + a_{0i} + a_{it}] \exp(\gamma_j) \times \exp(RS_j(t)) \quad (11)$$

or via log-link we assume:

$$\log(\mu_{ijt} | \mathbf{b}) = \log(n_{ijt}) + \log(m) + S_0(t) + a_{0i} + a_{it} + \gamma_j + RS_j(t) \quad (12)$$

where $\log(n_{ijt})$ is the offset; m represents the mean rate over all local regions at the midpoint of the time period considered (when t is centered, as is the case in the application); $\{a_{0i}\}$ and $\{a_{it}\}$ represent fixed age effects; $S_0(t)$ is a single spline representing the BC (“global”) trend and $\{RS_j(t)\}$ a family of random splines representing regional (“local”) trends relative to the “global” spline; γ_j is the random area effect and $\gamma = (\gamma_1, \dots, \gamma_{20}) \sim N(\mathbf{0}, \Sigma_\gamma)$. Specifically, model (11) corresponds to the broad spatio-temporal model framework in Eq. (7) such that λ_0 , the fixed age-time component in (7), is $m \exp[S_0(t) + a_{0i} + a_{it}]$; λ_j , the random spatial effects, are here $\exp(\gamma_j)$; and $\lambda_{j,t}$, the random spatio-temporal interaction, is $\exp(RS_j(t))$.

In this example, regression B-splines with one inner knot are used for the fitting of the fixed and random temporal effects $S_0(t)$ and $RS_j(t)$; the estimated model appears to produce adequate prediction of the rates and ratios using B-splines of order 2. Thus, in this example, we assume the following specifications for the spline terms:^[3,12]

$$S_0(t) = \sum_{k=1}^3 a_{0k} B_k(t) \quad (13)$$

and

$$RS_j(t) = \sum_{k=1}^3 \beta_{jk} B_k(t) \quad (14)$$

where $\{B_1(t), B_2(t), B_3(t)\}$ denote corresponding B-spline basis functions (without the intercept term) evaluated at time t (t is centered); $\beta_k = (\beta_{1k}, \dots, \beta_{jk})^T$ are random spline coefficients with respect to $B_k(t)$, $\beta_k \sim N(\mathbf{0}, \Sigma_k)$, $k = 1, 2, 3$.

The PQL analysis is presented here, assuming that the conditional expectation of y_{ijt} given the random effects follows (12–14), with variance–covariance matrix:

$$\Sigma_b = \begin{pmatrix} \Sigma_\gamma & \rho I_{20} & \rho I_{20} & \rho I_{20} \\ \rho I_{20} & \Sigma_1 & 0 & 0 \\ \rho I_{20} & 0 & \Sigma_2 & 0 \\ \rho I_{20} & 0 & 0 & \Sigma_3 \end{pmatrix} \quad (15)$$

where I_{20} is a 20×20 identity matrix. The variance–covariance matrices Σ_γ and Σ_{β_k} may be assumed to follow one of the CAR specifications given by Eqs. (4)–(6). Preliminary analyses indicated that the spatial correlations, among spatial random effects ensemble γ , as well as β_k , $k = 1, 2, 3$ are small and negligible. Here, we present the analysis assuming $\Sigma_\gamma = \sigma_\gamma^2 I_{20}$, $\Sigma_k = \sigma_{\beta_k}^2 I_{20}$, $k = 1, 2, 3$, $\text{Corr}(\gamma, \beta_k) = \rho$, $k = 1, 2, 3$, $\text{Corr}(\beta_r, \beta_k) = 0$, $k \neq r$. Note that the inclusion of age effects allows one to predict age- and region-specific rate and ratio trends as well as age-adjusted region-specific ratio estimates.

The data suggest some evidence of over dispersion; Table 1 lists the estimates of the variance components and the model coefficients. Fig. 1 plots provincial rate spline estimates by age group, showing mild curvature in provincial rates. The injury rates are considerably lower for girls in their mid-teens to mid-twenties while young girls between age 5 and 14 appear to be more likely to have such injuries. Fig. 2 displays age-adjusted injury ratio estimates for each of the 20 health regions, indicating regions of high injury risks: North Okanagan, Thompson, Coast Garibaldi, Upper Island/Central Coast, and Northern Interior, and those of low risks: Vancouver, Burnaby, North Shore, and Richmond. Increasing trends in injury risks were observed in health regions North Okanagan, Coast Garibaldi, Upper-Island/Central Coast, and Capital. Rate and ratio estimates can also be examined by health region by age group. To conserve space, Fig. 3 plots the health region rate splines for young girls of age 5–14, where temporal smoothing predicts moderate curvature for most of the health regions. Clearly,

Table 1 Spatio-Temporal Model Fit to BC Child Injury Hospitalization Rates, Accident Falls, Girls Age 0–19, 1987–1996

Parameter	Estimate	Std. error
$\log m$	−4.163	0.04613
α_{01}	−0.240	0.05085
α_{02}	−0.842	0.04062
α_{03}	−1.357	0.03197
α_{11}	−1.192	0.02160
α_{12}	−1.110	0.01926
α_{13}	−1.861	0.02140
β_1	0.087	0.00760
β_2	0.093	0.00582
β_3	0.118	0.00747
$\sigma\gamma^2$	0.0814	0.0304
$\sigma_{\beta_1}^2$	0.0142	0.0162
$\sigma_{\beta_2}^2$	0.0104	0.0123
$\sigma_{\beta_3}^2$	0.0266	0.0164
ρ	0.00423	0.01017

for young girls of age 5–14, injuries caused by accidental falls declined in all 20 regions, and similar trends, not shown here, are observed for young girls of other age groups. Fig. 4 maps age-adjusted injury ratios for health regions from 1987 to 1996; a change in spatial pattern

is observed, with high injury risk clusters shifting from northwest towards the southcentral.

CONCLUSIONS

This entry gives a cursory review of spatial and spatio-temporal models and related methods of inference. A recent monograph by Banerjee, Carlin, and Gelfand^[36] presents a current account on spatial and spatio-temporal models and related issues, methods, and applications. The book also presents recent developments of spatial and spatio-temporal frailty models for analysis of survival (or time-to-event) data. Implementations of spatio-temporal modeling remain considerably challenging primarily due to the model and the computational complexity, as related analyses essentially aim to explore and tackle spatial correlation, temporal correlation, and spatial–temporal interaction within a unified modeling framework. Currently available software packages do not readily facilitate complex spatio-temporal modeling. Most applications to date involve extensive computer programming and coding is often done on a case-to-case basis. Recent developments of statistical and computational software such as WinBUGS (<http://www.mrc-bsu.cam.ac.uk/bugs>), Splus (<http://www.insightful.com>), and others, as well as geographical information system (GIS) software, such as ArcView and ArcGIS (<http://www.esri.com>), provide the potential for more accessible software-facilitated spatio-temporal analysis procedures in the near future.

Scaled-SROC

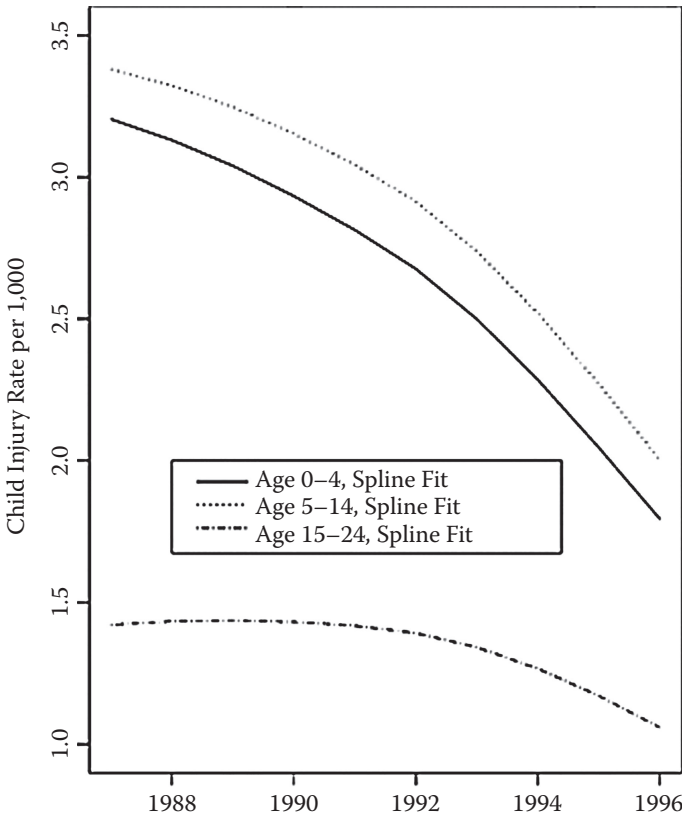


Fig. 1 Child injury due to accidental falls: fitted rate per 1000, 1987–1996, British Columbia, Canada

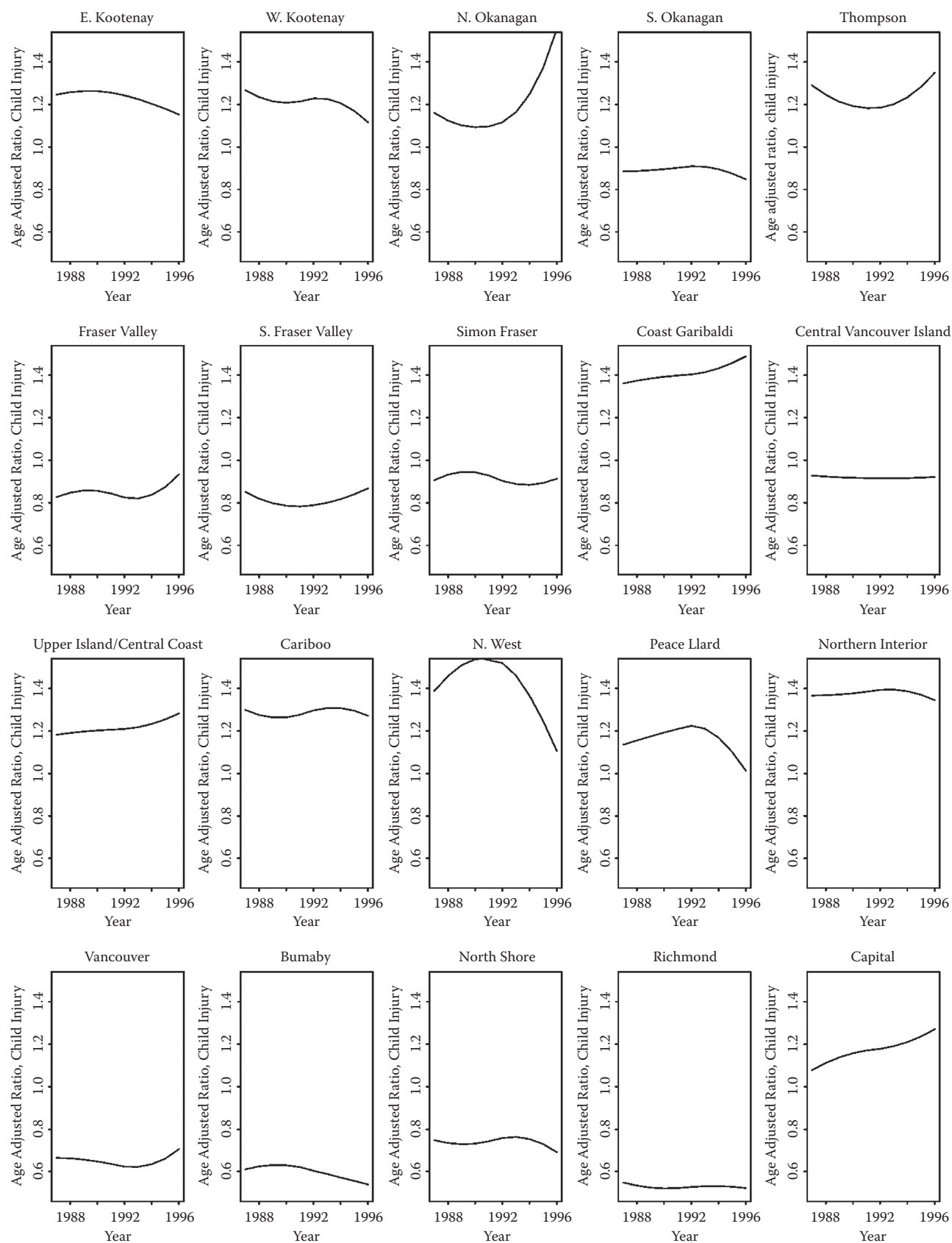


Fig. 2 Child injury due to accidental falls: age-adjusted ratios for health regions, 1987–1996, British Columbia, Canada

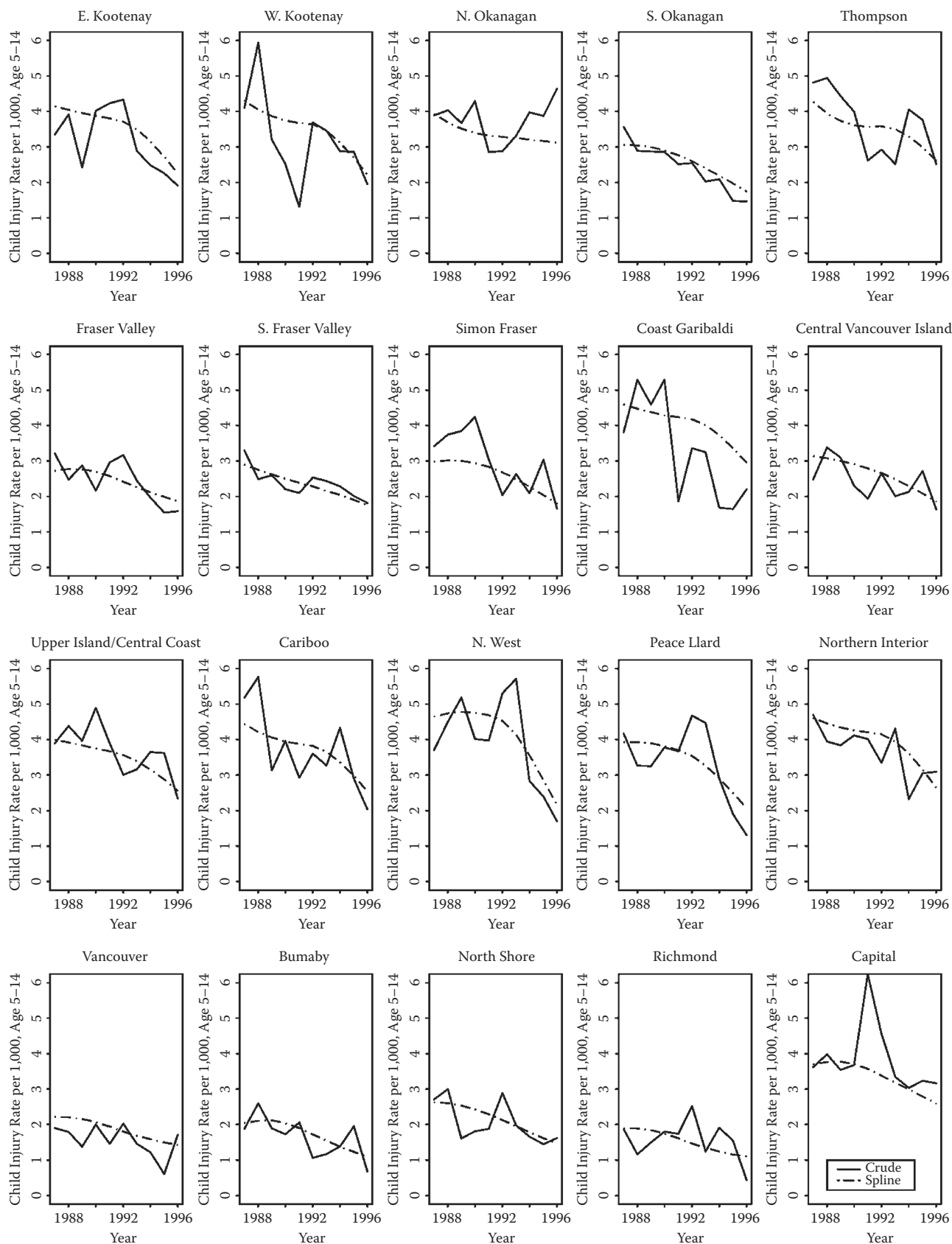


Fig. 3 Child injury due to accidental falls: fitted rate per 1000 for health regions, 1987–1996, girls age 5–14, British Columbia, Canada

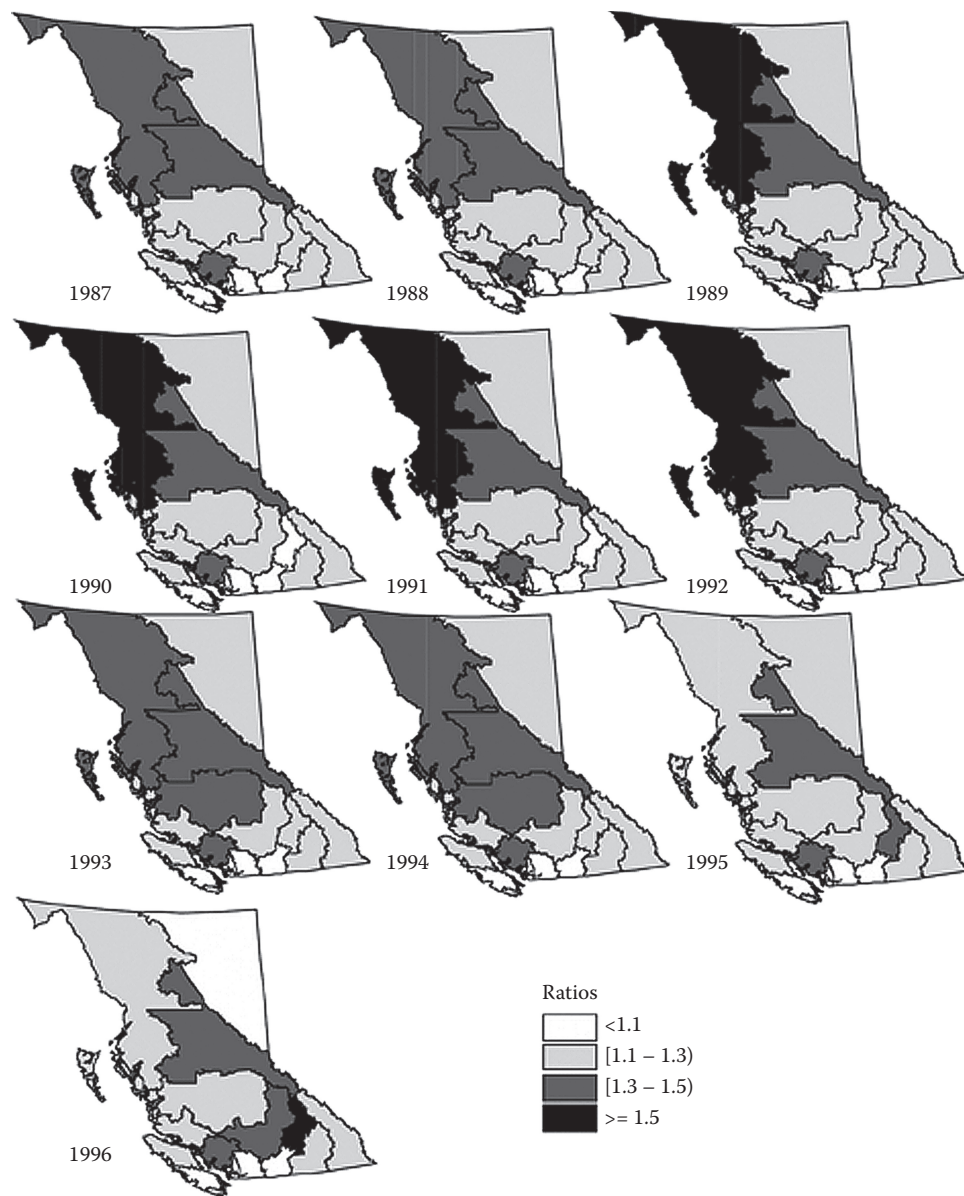


Fig. 4 Child injury due to accidental falls: fitted ratio for health regions, 1987–1996, girls age 5–14, British Columbia, Canada

ACKNOWLEDGMENTS

This work was supported through funds from the Natural Sciences and Engineering Research Council of Canada. DrMacNab's research is also supported through an Investigator Award from the British Columbia Research Institute for Children's and Women's Health.

REFERENCES

1. *Spatial Epidemiology: Methods and Applications*; Elliott, P. Wakefield, J.C., Best, N.G., Briggs, D.J., Eds.; Oxford University Press: Oxford, 2000.
2. Böhning, D.; Dietz, E.; Schlattmann, P. Space-time mixture modeling of public health data. *Stat. Med.* **2000**, *19* (17/18), 2333–2344.
3. MacNab, Y.C. A Bayesian hierarchical model for accident and injury surveillance. *Accid. Anal. Prev.* **2003**, *35* (1), 91–102.
4. *Disease Mapping and Risk Assessment for Public Health*; Lawson, A.; Biggeri, A.; Böhning, D.; Lesaffre, E.; Viel, J.-F. Bertollini, R., Eds.; John Wiley & Sons: New York, 1999.
5. MacNab, Y.C. Hierarchical Bayesian modelling of spatially correlated health service outcome and utilisation rates. *Biometrics*. **2003**, *59* (2), 305–316.
6. Bernardinelli, L.; Clayton, D.; Pascutto, C.; Montmoli, C.; Ghislandi, M. Bayesian analysis of space-time variation in disease risk. *Stat. Med.* **1995**, *14*, 2433–2443.

7. Waller, L.A.; Carlin, B.P.; Xia, H.; Gelfand, A.E. Hierarchical spatio-temporal mapping of disease rates. *J. Am. Stat. Assoc.* **1997**, *92*, 607–617.
8. Xia, H.; Carlin, C.P. Spatio-temporal models with errors in covariates: mapping Ohio lung cancer mortality. *Stat. Med.* **1998**, *17*, 2025–2043.
9. Knorr-Held, L.; Besag, J. Modelling risk from a disease in time and space. *Stat. Med.* **1998**, *17*, 2045–2060.
10. Sun, D.; Tsutakawa, R.K.; Kim, H.; He, Z. Spatio-temporal interaction with disease mapping. *Stat. Med.* **2000**, *19* (15), 2015–2036.
11. Pickle, L.W. Exploring spatio-temporal patterns of mortality using mixed effects models. *Stat. Med.* **2000**, *19* (17–18), 2251–2264.
12. MacNab, Y.C.; Dean, C.B. Autoregressive spatial smoothing and temporal smoothing B-splines for mapping rates. *Biometrics* **2001**, *57* (3), 949–956.
13. Diggle, P.; Knorr-Held, L.; Rowlingson, B.; Su, T.L.; Hawtin, P.; Bryant, T. Manuscript **2002**, <http://www.maths.lancs.ac.uk/diggle/papers/surveillance.ps>.
14. Cressie, N.A. *Statistics for Spatial Data*; John Wiley & Sons: New York, 1993.
15. Wakefield, J.; Morris, S. Spatial dependence and errors-in-variables in environmental epidemiology. In *Bayesian Statistics 6*; Bernardo, J.M.; Berger, J.O.; Dawid, A.P.; Smith, A.F., Eds.; Clarendon Press: Oxford, 1999, 657–684.
16. MacNab, Y.C. Hierarchical Bayesian spatial modelling of small-area rates of non-rare disease. *Stat. Med.* **2003**, *22* (10), 1761–1773.
17. Besag, J.; York, J.; Mollié, A. Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Stat. Math.* **1991**, *43*, 1–21.
18. Best, N.G.; Arnold, R.A.; Thomas, A.; Waller, L.A.; Conlon, E.M. Bayesian methods in the atmospheric sciences. In *Bayesian Statistics 6*; Bernardo, J.M.; Berger, J.O.; Dawid, A.P.; Smith, A.F., Eds.; Clarendon Press: Oxford, 1999, 131–156.
19. Leroux, B.G.; Lei, X.; Breslow, N. Estimation of disease rates in small areas: A new mixed model for spatial dependence. In *Statistical Models in Epidemiology, the Environment and Clinical Trials*; Halloran, M.E.; Berry, D., Eds.; Springer-Verlag: New York, 1999, 135–178.
20. Knorr-Held, L. Bayesian modeling of inseparable space-time variation in disease risk. *Stat. Med.* **2000**, *19* (17/18), 2555–2568.
21. Brumback, B.A.; Rice, J.A. Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *J. Am. Stat. Assoc.* **1998**, *93* (443), 961–994.
22. Hastie, T.J.; Tibshirani, R.J. *Generalized Additive Models*; Chapman & Hall/CRC: New York, 1990.
23. Green, P.P.; Silverman, B.W. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*; Chapman & Hall/CRC: New York, 2000.
24. Breslow, N.E.; Clayton, D.G. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* **1993**, *88*, 9–25.
25. Manton, K.G.; Woodbury, M.A.; Stallard, E.R.; et al. Empirical Bayes procedures for stabilizing maps of US cancer mortality rates. *J. Am. Stat. Assoc.* **1989**, *84*, 637–650.
26. Lin, X.; Zhang, D. Inference in generalized additive mixed models by using smoothing splines. *J. R. Stat. Soc. B* **1999**, *61*, 381–400.
27. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. B* **1977**, *39*, 1–38.
28. McLachlan, G.J.; Krishnan, T. *The EM Algorithm and Extensions*; Wiley: New York, 1997.
29. Bernardinelli, L.; Montomoli, M. Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Stat. Med.* **1992**, *11*, 983–1007.
30. He, Z.; Sun, D. Hierarchical Bayes estimation of hunting success rates with spacial correlations. *Biometrics* **2000**, *56*, 360–367.
31. Cressie, N.; Read, T.R.C. *Statistics and decisions* **1989**, (Suppl. 2,3), 333–349.
32. Schlattmann, P.; Böhning, D. Computer packages C.A.MAN (computer assisted mixture analysis) and Dis-map. *Stat. Med.* **1993**, *12*, 1943–1950.
33. Militino, A.F.; Ugarte, M.D.; Dean, C.B. The use of mixture models for identifying high risks in disease mapping. *Stat. Med.* **2001**, *20*, 2035–2049.
34. Böhning, D.; Schlattmann, P.; Lindsay, B.G. Computer-assisted analysis of mixtures (C.A.MAM): statistical algorithms. *Biometrics* **1992**, *48*, 283–303.
35. Green, P.J.; Richardson, S. Hidden Markov models and disease mapping. *J. Am. Stat. Assoc.* **2002**, *97* (460), 1055–1070.
36. Banerjee, S.; Carlin, B.P.; Gelfand, A.E. *Hierarchical Modeling and Analysis for Spatial Data*; Chapman and Hall/CRC: New York, 2004.

Specifications

John R. Murphy

Eli Lilly and Company, Indianapolis, Indiana, U.S.A.

Abstract

Specifications on the physical and chemical attributes of drug substances and drug products are the quality standards that provide the patient assurance that the drug will perform its intended purpose without unexpected side effects. This entry gives a brief overview of how these specifications are determined.

INTRODUCTION

In the pharmaceutical industry, “specifications” usually mean the same thing as “specification limits,” which refer to numerical tolerances within which the measured result for a quality attribute of a dosage form should fall. Specifications may also refer to more general and/or complex criteria, which can sometimes be found in a compendium such as the United States Pharmacopeia and National Formulary (USP/NF).^[1] An example of specification limits that are numerical tolerances might be where the purity of a drug substance must be no less than 97.0% and no more than 103.0%. An example of specifications based on more general criteria might be where a solid oral dosage form must meet the USP dissolution criteria using a Q of 75% at 45 minutes.

OVERVIEW

Specifications on the physical and chemical attributes of drug substances and drug products are the quality standards that provide the patient assurance that the drug will perform its intended purpose without unexpected side effects. In that sense, specifications may be thought of as defining the numerical tolerances or other criteria that make the drug safe, effective, and fit for use. Ideally, specification limits based on those considerations could be established, and the manufacturing process would be capable of providing units falling well within such bounds. In practice, however, many times the allowable limits based on safety, efficacy, or fitness for use seem too wide, stemming from a belief that narrow or tight specifications mean a higher-quality product. The net result is that the determination of specification limits is sometimes based almost entirely on judgments about manufacturing capability.

Because specifications are a legal commitment with regulatory authorities, their determination is an important undertaking, and the process of getting to agreed-upon

specifications is sometimes difficult. Part of the problem is the complexity of the measurement process, which can often comprise a significant amount of the variation in the measured results. As a result, many specifications cannot be set without tying them to the analytical methods that produce the result. Another feature of drug substances and drug products is that degradation be factored into the process of setting specifications. Stability estimation can present a special problem because the effect (change over time) is often small relative to the variation in the manufacturing and the measurement processes. Given these facts, it seems evident that determining rational specifications requires the use of statistical methods to estimate and account for all the variation that may be present.

Specifications mean different things to different people, and part of the reason for this is a failure to recognize the difference between several types of limits that are derived to meet different purposes.^[2] The type of specification limits that most often come to mind are “expiry specifications” or “regulatory specifications,” which are the limits a product must meet throughout its shelf life and which are documented in a regulatory submission or compendial monograph. The failure of units in commercial distribution to comply with expiry specifications is a serious matter that can result in regulatory censure and/or expensive and time-consuming product recall. In order to minimize the probability of out-of-specification results, manufacturers utilize “release specifications” or “acceptance limits,” which are applied at the time of product release. In the United States, release limits are still considered by the FDA to be mostly a matter of internal company policy, while in Europe the release limits must be formally registered in the regulatory dossier. The third type of limits are what may be informally regarded as “statistical control limits,” which reflect the capability of a manufacturing process in a state of statistical control and serve to alert us to unusual results that are likely to be the consequence of special causes.

One primary reason for disagreement about how wide or narrow expiry specifications should be is a misunderstanding

and lack of agreement about the function and the relative location of these three types of limits. The purpose of release limits is to prevent noncomplying products from reaching the marketplace, and they are best determined using statistical methods that take into account the sampling and measurement variation and the change in the product attribute over time. Ideally, they are set from the expiry limits inward such that if the batch result falls within the release limits, there is a specified level of confidence that future measurements will fall within the expiry limits. In and of themselves, release limits do not have any relation to the capability of the process to produce results that will fall within them. Statistical control limits, on the other hand, are related to the capability of the process, and when the release limits are at least as wide as the statistical control limits, the process will be capable of producing results that have a high probability of falling within the release limits. Unfortunately, sufficient data firmly to establish control limits and release limits may not be available at the point in time in the development process when the limits need to be set; it is therefore imperative that regulatory bodies worldwide recognize the need for expiry limits and release limits to be wider at first approval. There can be an expectation that specification limits will be appropriately tightened as more supporting process data becomes available.

Currently, there is no international consensus on the approach to specification setting, but International Conference on Harmonization (ICH) document Q6A on specifications^[3] is a large step forward. This guideline provides common definitions, discusses justification of specifications, and gives a listing of the analytical properties that may be considered universal for drug substances and the various dosage forms of drug products. The use of

statistical methods is not specifically mentioned or encouraged in the document, but the necessity of establishing release specifications separate and different from expiry specifications is recognized.

CONCLUSION

In summary, specifications define the standards of identity, strength, quality, and purity of drug products. The patient derives assurance that the drug will perform its intended function from the knowledge that its physical and chemical properties have been tested and have been determined to comply with those standards. Good specifications are those that guarantee that the product will be safe and effective while at the same time providing sufficient latitude for the manufacturer to produce and release product that is highly probable to remain within specification throughout the shelf life. Such specifications are best developed with the aid of statistical methods.

REFERENCES

1. USP/NF. *The United States Pharmacopeia XXIV and the National Formulary XIX*. The United States Pharmacopeial Convention, Inc: Rockville, MD, 2000.
2. Davis, G.E.; Murphy, J.R.; Weisman, D.A.; Andersen, S.W.; Hofer, J.D. *Pharm. Technol.* September **1996**, 100–118.
3. *International Conference on Harmonization. Specifications: Test Procedures and Acceptance Criteria for New Drug Substances and New Drug Products: Chemical Substances—Q6A (Step 2 of the ICH Process)*, July 18, 1997.

SROC Curve

S. D. Walter

*Department of Clinical Epidemiology and Biostatistics, Faculty of Health Sciences,
McMaster University, Hamilton, Ontario, Canada*

Petra Macaskill

*Screening and Test Evaluation Program (STEP), Sydney School of Public Health, University
of Sydney, Sydney, New South Wales, Australia*

Abstract

The summary receiver operating characteristic (SROC) curve has been recommended to represent the performance of a diagnostic test, based on data from a meta-analysis. This entry aims to detail the process of a meta-analysis of diagnostic test data and discuss issues with using the fixed-effect model.

Keywords: Area Under the Curve (AUC); Diagnostic test evaluation; Fixed effects; Hierarchical mixed effects; Meta-analysis; Sensitivity; Specificity; Summary receiver operating characteristic (SROC) curve.

INTRODUCTION

The summary receiver operating characteristic (SROC) curve has been recommended to represent the performance of a diagnostic test, based on data from a meta-analysis. Under a fixed-effect, logit-threshold model, the position of the SROC curve can be characterized in terms of the overall diagnostic odds ratio and the magnitude of inter-study heterogeneity. The Area Under the Curve (AUC) and an index Q^* are potentially useful summary measures. It is shown that AUC is maximized when the study odds ratios are homogeneous, and that it is quite robust to heterogeneity. An explicit upper bound is derived for AUC in the homogeneous situation, and a lower bound based on the limit case Q^* , defined by the point where sensitivity equals specificity: Q^* is invariant to heterogeneity. The standard error of AUC is derived for homogeneous studies, and is shown to be a reasonable approximation with heterogeneous studies. AUC and its standard error are easily computed in the homogeneous case, and avoid the need for numerical integration in the more general case. $SE(AUC)$ and $SE(Q^*)$ are numerically close, with $SE(Q^*)$ being larger if the odds ratio is very large. Motivation for the use of the AUC and Q^* measures as summaries of the SROC curve are discussed.

Multi-level mixed models are also described which provide a more general framework for SROC estimation that take account of within study sampling error for both sensitivity and specificity, and also include random study effects to take account of additional unexplained variability in test performance.

META-ANALYSIS OF DIAGNOSTIC TEST DATA

Systematic reviews of primary studies are becoming increasingly important for summarizing evidence about the accuracy of diagnostic tests. Guidelines for the conduct of such reviews^[1–4] include defining the objectives, retrieval of the relevant literature, data extraction, meta-analytic methods for obtaining summary estimates of test accuracy, and investigating reasons for variation in test accuracy across studies.

The receiver operating characteristic (ROC) curve is well established as a method of summarizing the performance of a diagnostic test within a single study.^[5–7] It indicates the relationship between the true positive rate (TPR) and the false positive rate (FPR) of the test, as the threshold used to distinguish disease cases from noncases varies. For instance, the threshold might be a defined level of serum cholesterol as a marker of cardiovascular disease, or a particular level of cellular abnormality as a marker of malignancy. The summary receiver operating characteristic (SROC) curve and the area under the curve (AUC) have been proposed as a way to describe diagnostic data in the context of a meta-analysis.^[1,2,4,8–11]

A schematic ROC curve is shown in Fig. 1. All of its data points come from a single study, and are defined by the results arising when one of several alternative thresholds (or cut points) on the test results is used to discriminate between disease cases and noncases. Liberal thresholds that give high values of TPR will tend to incorrectly label a relatively high fraction of the noncases as cases, so the false positive rate (FPR) will be high. Conversely, conservative

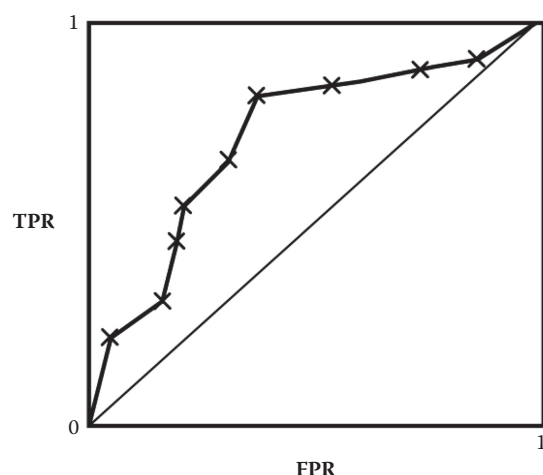


Fig. 1 Schematic ROC curve

thresholds yield incorrect labels for noncases rather infrequently, but at the cost of detecting a smaller fraction of the true cases; in this situation, the value of TPR and FPR are both relatively low. Thus one generally expects TPR and FPR to be positively associated. In the ROC space, points near the lower-left corner correspond to conservative thresholds (low TPR and low FPR values), and points near the upper-right corner correspond to liberal thresholds (high TPR and high FPR values).

In a single study, changing the threshold necessarily results in monotonic changes in TPR and FPR. Accordingly, the ROC curve can always be empirically represented, in its simplest form by connecting the data points as shown in Fig. 1. Alternatively, smoothed curves can also be fitted using a latent variable approach, based on a model that the case and noncase test values follow normal or logistic distributions.^[12–20]

In a meta-analysis, the units of analysis are separate studies. In the simplest case, each study contributes an estimate of TPR and FPR. The SROC curve represents the relationship between TPR and FPR across studies, recognizing they may have used different thresholds. In contrast to the ROC analysis, the set of (FPR, TPR) points does not necessarily yield a unique, monotonic curve. Several methods have been proposed for fitting an SROC curve;^[8,11,21,22] this paper will focus on the properties of the popular method developed by Moses et al.^[11] Alternative but more complex hierarchical models such as that proposed by Rutter and Gatsonis^[21,22] are also outlined and their potential advantages are discussed.

LOGIT-THRESHOLD FIXED EFFECT MODEL FOR SUMMARY RECEIVER OPERATING CHARACTERISTIC CURVE

The most popular method to obtain a smoothed fit of the SROC curve is to use the regression model proposed

by Moses et al.^[11] The dependent and independent variables are

$$D = \ln\left(\frac{\text{TPR}}{1 - \text{TPR}}\right) - \ln\left(\frac{\text{FPR}}{1 - \text{FPR}}\right) \quad (1)$$

and

$$S = \ln\left(\frac{\text{TPR}}{1 - \text{TPR}}\right) + \ln\left(\frac{\text{FPR}}{1 - \text{FPR}}\right) \quad (2)$$

respectively. D is equivalent to the diagnostic log odds ratio, $\ln(\text{OR})$, which conveys the test's accuracy in discriminating cases from noncases. S is a function of the diagnostic threshold in a study, with high values corresponding to liberal inclusion criteria for cases. $S = 0$ when $\text{TPR} = 1 - \text{FPR}$, i.e., on the antidiagonal from the top-left to bottom-right corners of the SROC space. The regression equation

$$D = a + bS \quad (3)$$

can be fitted by standard least squares methods, assuming that D is approximately normally distributed for a given value of S . Optionally, weights can be employed to reflect interstudy heterogeneity with respect to the sample variance of D , or a robust fitting procedure may also be used.^[11] The coefficient b in Eq. 3 represents the dependence of the test accuracy on threshold. If $b \approx 0$, then the studies will be referred to as *homogeneous*; they can then be summarized by an overall OR, noting that $a = \ln(\text{OR})$. If $b \neq 0$, then the studies are referred to as *heterogeneous* with respect to OR. In this case, a can be thought of as the value of $\ln(\text{OR})$ when $S = 0$.

This model assumes logistic distributions for the test values in the cases and noncases, but normal distributions may also constitute an adequate approximation. However, the model is generally fitted without directly appealing to the logistic distribution assumption. If the underlying test results actually have logistic distributions, then S for a given study can be expressed as a function of the cut point that gives rise to the sensitivity and specificity for that study.^[11] If the variances of the distributions of test results for the cases and noncases are equal, then the SROC is symmetric about the diagonal line where $\text{TPR} = 1 - \text{FPR}$. If the variances are unequal, the resulting SROC will be asymmetric. No assumptions are made about the distribution of S . Covariates may be added to the model to assess whether test accuracy varies systematically with other study-related factors.^[23,24]

If the estimate of test accuracy for each study is weighted by the inverse of its variance, one is assuming that sampling error is the only source of variability between studies. However, this assumption may be unrealistic, given that studies are likely to vary in a number of respects, including the spectrum of disease, conditions under which the test was administered, and

other study and patient characteristics that could affect test accuracy.^[25–27]

Although the parameters a and b are both fixed because they are assumed to be constant across studies, applying equal weights to studies (i.e., fitting an unweighted regression) has been empirically shown to give results that are consistent with assuming a random effect.^[9,11,21] This is because both the within- and between-study variances are taken into account, thereby giving relatively more weight to smaller studies, as would occur in a random effect model. Irwig et al.^[9] recommend the unweighted analysis, because weights based on sample variances may give too much emphasis to inaccurate studies, and hence give a biased SROC curve. Moses et al.^[11] also recommend the unweighted approach.

Once the regression has been fitted, one can reverse the transformations (Eqs. [1] and [2]) and hence deduce the relationship between TPR and FPR as

$$\text{TPR} = \frac{\exp\left(\frac{a}{1-b}\right)\left(\frac{\text{FPR}}{1-\text{FPR}}\right)^{(1+b)/(1-b)}}{1 + \exp\left(\frac{a}{1-b}\right)\left(\frac{\text{FPR}}{1-\text{FPR}}\right)^{(1+b)/(1-b)}} \quad (4)$$

Expression 4 gives TPR at any given value of FPR, and hence defines the entire SROC curve. While there may be interest in identifying particular points on the curve, it is also useful to have an overall summary measure of the curve's behavior. One appropriate measure is the Area Under the Curve (AUC), which can be calculated as

$$\text{AUC} = \int_0^1 \frac{\exp\left(\frac{a}{1-b}\right)\left(\frac{x}{1-x}\right)^{(1+b)/(1-b)}}{1 + \exp\left(\frac{a}{1-b}\right)\left(\frac{x}{1-x}\right)^{(1+b)/(1-b)}} dx \quad (5)$$

In general, AUC must be numerically obtained, because there is no closed form for expression 5.

Adoption of a summary index is helpful for succinct reporting of a given data set, especially when limited data preclude the reliable identification of particular points on the curve. The AUC measure is widely used in ROC analysis, where it can be interpreted as the probability that the test values for a random pair of diseased and nondiseased individuals would be correctly ranked; it also represents the (unweighted) average of TPR over all possible values of FPR.

The AUC is also a natural candidate summary for an SROC analysis. We will consider the alternative index Q^* , which will be defined as a point of indifference on the SROC curve, where the probabilities of an incorrect test result are equal for disease cases and noncases. The partial AUC has also been proposed as a summary measure, this being the area under some restricted portion of the curve corresponding to FPR values of clinical interest, or in which study data are located.

EMPIRICAL BEHAVIOR OF THE SUMMARY RECEIVER OPERATING CHARACTERISTIC CURVE

Fig. 2 shows a set of three symmetric SROC curves with $b = 0$, which occurs when the studies are homogeneous and thus exhibit no relationship between OR and threshold (S). The curves were obtained by numerical evaluation of Eq. 4 for $a = 1.5, 2$, and 3 , which correspond to $\text{OR} = 4.5, 7.4$, and 20.1 , respectively. As a increases, the SROC curve moves closer to its ideal position near the upper-left corner. If $a \rightarrow \infty$, then $\text{AUC} \rightarrow 1$, which would indicate a perfect test having 100% sensitivity and specificity, and no errors in distinguishing cases from noncases. In contrast, if $a \approx 0$ (or $\text{OR} \approx 1$), the curve lies close to the diagonal $\text{TPR} = \text{FPR}$ in the SROC space; then, $\text{AUC} = 1/2$ and the test performs no better than chance.

For completeness, we mention situations where $a < 0$. These correspond to $\text{OR} < 1$, when the test discriminates cases and noncases in the “wrong” direction, and *worse* than at random. As $\text{OR} \rightarrow 0$, $\text{AUC} \rightarrow 0$, and the curve lies close to the lower-right corner of the SROC space. Such situations are unlikely to occur in practice.

We now consider the case of heterogeneous studies under model 3, where diagnostic accuracy depends on threshold ($b \neq 0$). Fig. 3 shows three SROC curves derived from Eq. 4, all with $a=2$ but different values of b . Nonzero values of b give asymmetric curves. If $b > 0$, the curve initially rises less steeply than the symmetric curve (with $b=0$), but then it rises more steeply and crosses the symmetric curve to achieve relatively high values of TPR for high values of FPR. The SROC curves with $b < 0$ exhibit the opposite behavior—see, for example, the curve with $b=-0.5$ in Fig. 3.

Interestingly, as suggested by Fig. 3 and as shown by Moses et al.,^[11] the family of curves defined by a fixed value of a all pass through a common point, located on the

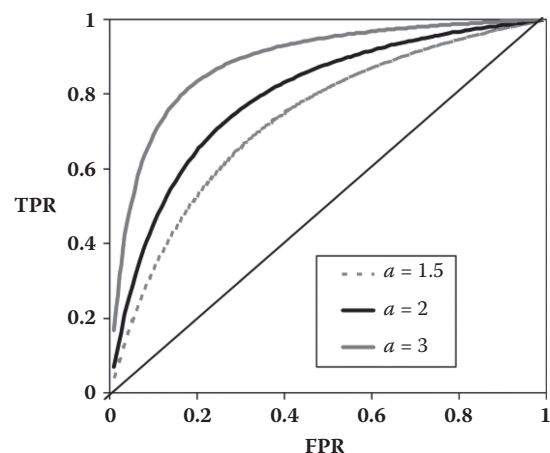


Fig. 2 Summary receiver operating characteristic with various values of a ($b = 0$)

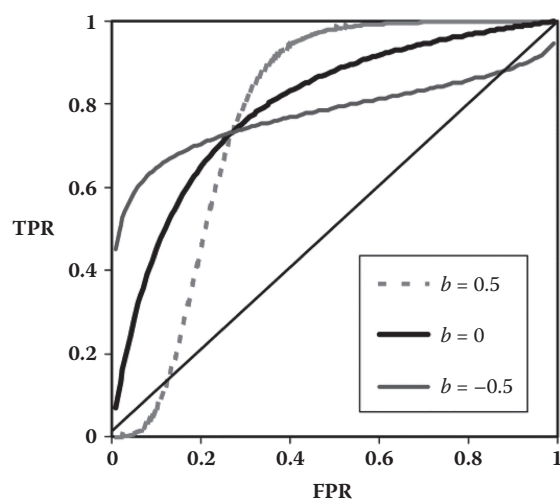


Fig. 3 Summary receiver operating characteristic with various values of b ($a = 2$)

antidiagonal, where $\text{TPR} = 1 - \text{FPR}$, or sensitivity equals specificity. That point has coordinates

$$\text{TPR} = \frac{\exp(a/2)}{1 + \exp(a/2)} \quad (6)$$

and

$$\text{FPR} = \frac{1}{1 + \exp(a/2)} \quad (7)$$

Moses et al.^[11] denote the value of TPR at this point by Q^* . In ROC analysis, Q^* has been suggested as a further summary measure. It corresponds to the point closest to the ideal top-left corner of the SROC space in symmetric curves.

The fact that all SROC curves with a given value of a pass through the same Q^* point means that Q^* conveys no additional statistical information beyond the odds ratio. However, use of Q^* can be motivated by its simple interpretability, given that Q^* is the point on the SROC curve where $\text{TPR} = 1 - \text{FPR}$; thus it represents the diagnostic threshold at which the probability of a correct diagnosis is constant for all subjects. Also, as will be shown later, the existence of a common value of Q^* permits the derivation of a useful lower bound for AUC.

As a consequence of assuming model 3, when $b \neq 0$, the SROC curve has a region where $\text{TPR} < \text{FPR}$, which lies below the main diagonal. This region may be seen, for example, near the lower-left corner in the curve for $b=0.5$ in Fig. 3. In this region, the test would theoretically be performing worse than at random. However, in practice, this region is very small. Even if relatively strong heterogeneity (with $b > 0$) is present, the “improper” part of the SROC curve (where $\text{TPR} < \text{FPR}$) includes very low values of TPR, and these values correspond to test thresholds that are unlikely to be acceptable in clinical practice.

On the other hand, if heterogeneity with $b < 0$ occurs, the improper part of the curve corresponds to very high values of FPR, which would again have no clinical relevance. See Ref. [28] for numerical details of this effect.

We now briefly discuss the behavior of model 3 for extreme values of b . As $b \rightarrow 1$, the fitted SROC curve becomes progressively steeper, and in the limit case at $b = 1$ it degenerates to a vertical line. This could occur, for example, if there was no variation in FPR between studies. However, the line still passes through the common Q^* point defined by Eqs. 6 and 7. Once $b > 1$, the curve inverts and shows a negative relationship of TPR to FPR. This is implausible in practice, except perhaps by chance in small samples.

Similar behavior is seen if b is near or below -1 . Near $b = -1$, the curve is horizontal, indicating no relationship of TPR to FPR. This could occur if there was no variation in TPR between studies. If $b < -1$, the curve again inverts (to a different shape) and suggests an implausible negative relationship between TPR and FPR. Despite this, all the curves for $b > 1$ possess the common value of Q^* given by Eqs. 6 and 7. In practice, data yielding $b > 1$ are unlikely. Empirical experience suggests that practical meta-analysis data sets often have b close to 0, and are rarely larger than 0.5 in absolute value.

SUMMARY MEASURES: AREA UNDER THE CURVE AND Q

The AUC is a popular index of the overall performance of a test.^[5–7,19,20,29] As mentioned earlier, AUC ranges from 1 for a perfect test that always correctly diagnoses, to 0 for a test that never correctly diagnoses. In single studies, AUC can be interpreted as the probability that the test will correctly rank a randomly chosen case/noncase pair with respect to their test values.^[30] The AUC is intended to fulfill the same function in meta-analyses, and in effect one assumes that the SROC curve accurately conveys test performance at the individual subject level.

Although numerical integration is required in general to obtain AUC, some special cases are of interest because they yield exact analytic expressions and comparative results, as we now discuss. As expected, AUC increases with a , for fixed b . By examining Eq. (5), one can prove that for a given value of a , AUC is maximized when $b=0$, implying that AUC is optimally large in homogeneous studies.^[28] Furthermore, one may show that AUC is symmetric in b , so that negative values of b yield the same value of AUC as the equivalent positive value, this despite the very different shapes of their associated SROC curves. If $a=b=0$, then from Eq. (5) $\text{AUC} = \int_0^1 x dx = \frac{1}{2}$. By symmetry, it is evident that $\text{AUC} = 1/2$ when $a=0$ even if $b \neq 0$, so $\text{AUC} = 1/2$ indicates random overall performance for any set of studies.

In the homogeneous case $b=0$, the general expression 5 becomes

$$AUC = \int_0^1 \frac{\exp(a) \left(\frac{x}{1-x} \right)}{1 + \exp(a) \left(\frac{x}{1-x} \right)} dx$$

In this case, we can obtain an exact solution

$$AUC_{\text{hom}} = \frac{OR}{(OR-1)^2} [(OR-1) - \ln(OR)] \quad (8)$$

where AUC_{hom} indicates the AUC for homogeneous studies, and $OR = \exp(a)$. If $a=0$ (or $OR=1$), then the special value $AUC_{\text{hom}} = 1/2$ should be used in place of Eq. (8), which is then degenerate. Expression 8 can be used to evaluate AUC for homogeneous studies, without the need for numerical integration. As demonstrated later, expression 8 is also a useful upper bound and close approximation for AUC in heterogeneous studies. For reference, Table 1 shows the value of AUC, Q^* , and their difference for a range of values of OR in the homogeneous case.

By noting that AUC declines with increasing b , and that the limit curve with $b \rightarrow 1$ passes through the common Q^* point, from Eqs. (6) and (7), we may deduce that a lower bound for AUC in curves with a given value of a is

$$Q^* = \frac{\exp(a/2)}{1 + \exp(a/2)} = \frac{\sqrt{OR}}{1 + \sqrt{OR}} \quad (9)$$

which is equivalent to the TPR value given in Eq. (6). Also, Q^* from Eq. (9) and the maximum value AUC_{hom} from Eq. (8) provide easily computable lower and upper bounds, respectively, for AUC with any given value of $a > 0$.

Numerical Tabulations

Numerical evaluations of expressions 8 and 9 demonstrate that in the homogeneous case, the difference between AUC_{hom} and Q^* increases for moderate values of OR, but

Table 1 AUC_{hom} , Q^* , and Their Difference for Various Values of the Diagnostic Odds Ratio: Homogeneous Case

Odds ratio	AUC_{hom}	Q^*	$AUC_{\text{hom}} - Q^*$
0.5	0.386	0.414	-0.028
1	0.500	0.500	0.000
1.5	0.567	0.551	0.017
2	0.614	0.586	0.028
3	0.676	0.634	0.042
4	0.717	0.667	0.051
5	0.747	0.691	0.056
10	0.827	0.760	0.067
20	0.887	0.817	0.069
30	0.913	0.846	0.068
40	0.929	0.863	0.065
50	0.939	0.876	0.063

is never more than about 7%.^[28] The maximum difference occurs at $a=2.85$ (or $OR=17.3$), which value is the solution to a transcendental equation. For larger values of a , the difference declines very slowly, with a limit value $AUC_{\text{hom}} - Q^* = 0$ at $a = \infty$.

The numerical integration of Eq. (5) for the heterogeneous case allows one to assess the impact of heterogeneity on the value of AUC. For a fixed value of OR, AUC declines slowly as b increases from 0 (homogeneous studies) to larger values (increasing heterogeneity). However, the dependence of AUC on b is weak, and the dominant effect on AUC is the value of OR (or a). For $|b| < 0.4$, the percentage change in AUC compared to the homogeneous case is less than 2%. Accordingly, AUC_{hom} provides a good approximation to AUC even in heterogeneous studies.^[28]

STANDARD ERRORS OF AREA UNDER THE CURVE AND Q

We first consider the sample variation in $\widehat{AUC}$. From Eq. (5), we see that AUC is a function of the regression parameters a and b , and hence the variability in $\widehat{AUC}$ is a function of the sample variation in $\hat{a}$ and $\hat{b}$. Using the delta method, an approximate variance for $\widehat{AUC}$ is

$$\begin{aligned} \text{var}(\widehat{AUC}) = & \left(\frac{\partial AUC}{\partial a} \right)^2 \text{var}(\hat{a}) + \left(\frac{\partial AUC}{\partial b} \right)^2 \text{var}(\hat{b}) \\ & + 2 \left(\frac{\partial AUC}{\partial a} \right) \left(\frac{\partial AUC}{\partial b} \right) \text{cov}(\hat{a}, \hat{b}) \end{aligned} \quad (10)$$

where, from Eq. 5

$$\begin{aligned} \frac{\partial AUC}{\partial a} = & \left(\frac{1}{1-b} \right) \exp \left(\frac{a}{1-b} \right) \\ & \times \int_0^1 \frac{\left(\frac{x}{1-x} \right)^p}{\left[1 + \left(\frac{x}{1-x} \right)^p \exp \left(\frac{a}{1-b} \right) \right]^2} dx \end{aligned} \quad (11)$$

$$\begin{aligned} \frac{\partial AUC}{\partial b} = & \left(\frac{1}{1-b} \right)^2 \exp \left(\frac{a}{1-b} \right) \\ & \times \int_0^1 \frac{\left(\frac{x}{1-x} \right)^p \left[a + 2 \ln \left(\frac{x}{1-x} \right) \right]}{\left[1 + \left(\frac{x}{1-x} \right)^p \exp \left(\frac{a}{1-b} \right) \right]^2} dx \end{aligned} \quad (12)$$

and $p = (1+b)/(1-b)$. The variances and covariance of $\hat{a}$ and $\hat{b}$ can be directly obtained from the standard regression software used to fit model 3 to the data.

In general, evaluation of $\text{var}(\widehat{AUC})$ again requires numerical integration. However, for the special case of homogeneous studies where $b=0$, an approximate, large-sample

expression is possible. Using the delta method, we may then obtain

$$SE(\widehat{AUC}_{\text{hom}}) = \frac{OR}{(OR-1)^3} [(OR+1)\ln OR - 2(OR-1)] SE(\hat{a}) \quad (13)$$

Eq. 13 implies $SE(\widehat{AUC}_{\text{hom}})$ is symmetric in $\ln(OR)$, although we would usually be concerned with values $OR > 1$. When $OR=1$, expression 13 is degenerate, but by using L'Hôpital's rule, one can show that in the neighborhood of $OR=1$,

$$SE(\widehat{AUC}) \approx SE(\hat{a})/6 \quad (14)$$

The delta method also yields an approximate standard error for $\hat{Q}^*$ as

$$SE(\hat{Q}^*) = \frac{\sqrt{OR}}{2(\sqrt{OR} + 1)^2} SE(\hat{a}) \quad (15)$$

Moses et al.^[11] give this result in an alternative form involving $\cosh(a/4)$. Numerical evaluations show that the standard errors of Q^* and AUC are both maximized when $OR = 1$, so the least precise situation for either index is for tests that have close to random performance. In the region of $OR = 1$, $SE(\widehat{AUC}) \approx 1/6$ and $SE(\hat{Q}^*) \approx 1/8$. Comparisons show $SE(\widehat{AUC}) > SE(\hat{Q}^*)$ for OR values between 1 and 17.3, the same value associated with the values of AUC and Q^* ; for $OR > 17.3$, AUC has the smaller standard error. Both standard errors approach 0 as $a \rightarrow \infty$.

For heterogeneous studies, $SE(\widehat{AUC})$ is not necessarily maximized when $a = 0$, because it involves $\text{var}(\hat{b})$ and $\text{cov}(\hat{a}, \hat{b})$ as well as $\text{var}(\hat{a})$. However, it is the last of these terms that dominates, with the other two making relatively small numerical contributions. Thus in practice $SE(\widehat{AUC})$ is maximized approximately when $OR = 1$, even for heterogeneous studies. For most values of OR , the standard error declines by up to about 10% at the most extreme level of b .

OTHER ISSUES WITH THE FIXED-EFFECT MODEL

So far, we have established some basic properties of the SROC curve under model 3, and we found that the value AUC_{hom} associated with homogeneous studies is a reasonable upper-bound approximation even for the general AUC with heterogeneous studies. The corresponding standard error also provides a good approximation for heterogeneous studies, and is conservatively large except in some cases of extreme heterogeneity. In the presence of strong heterogeneity, the AUC would be an inadequate summary of the data anyway, and it would then be preferable to examine the SROC curve in more detail, including specific TPR values for given FPR.

Q^* also provides an easily computed lower bound for AUC, but empirically appears to be not quite as good an approximation as AUC_{hom} . The motivation for Q^* as an index in its own right is that it is located where the SROC curve crosses the antidiagonal from (0, 1) to (1, 0) of the SROC space. Hence $TPR = 1 - FPR$ at Q^* , and so the probability of an incorrect result from the test is the same for cases and noncases. Q^* is therefore a point of "indifference" between false-positive and false-negative diagnostic errors. In homogeneous studies, Q^* is the point on the SROC curve lying closest to the optimal upper-left corner, but this is not true with heterogeneous studies.

Use of Q^* as the summary measure assumes implicitly that false-negative and false-positive test results are of equal value. In practice, there may be different costs associated with these two types of error: One wishes to minimize false-positive results because of the additional testing required to establish the correct diagnosis (noncase), and because the additional tests tend to be more costly, invasive, or risky. On the other hand, false-negative results lead to disease cases being missed, with possible deterioration in their prognosis. In general, one must weigh the false-positive and false-negative errors to balance the overall performance of the test in a population; the optimal diagnostic threshold does not then correspond in general to the Q^* point.

One can motivate the use of AUC as an index that represents the probability that the test will correctly rank a case/noncase pair of subjects.^[15,30] It can also be thought of as the average TPR over the entire range of FPR values. Because it summarizes the whole SROC curve, AUC has a symmetric interpretation with respect to either TPR or FPR. The AUC is affected by the whole SROC curve, including regions with limited or no data, or by sectors corresponding to TPR and FPR values that are unlikely to occur in practice. Accordingly, it has been suggested that partial SROC curves be adopted, by limiting attention to those portions of the SROC curve of clinical interest, or where data are actually observed.^[31-34]

There are some unresolved issues on the use of partial SROC curves and the corresponding partial AUC^[35] First, the partial AUC may be thought of as the average TPR within a restricted range of FPR, but not vice versa. Hence the partial AUC has an asymmetric interpretation with respect to TPR and FPR; on the other hand, the complete AUC enjoys a symmetric interpretation with respect to both types of test error. Second, there may be some arbitrariness in which regions of the curve to select. One might examine the SROC curve within a prespecified range of TPR values, for instance by defining a maximum acceptable level for clinical practice. Another approach would be to choose the region according to the observed range of data in the meta-analysis; the choice itself would then be affected by sampling variation, which might be substantial in meta-analyses involving small studies. Furthermore, a reasonable choice for one test may be unreasonable for another test, thus complicating their comparison.

The analytic methods for AUC presented in this paper can be extended to cover the partial AUC and its standard error.^[35] We may expect the effect of interstudy heterogeneity to be greater for the partial AUC than the weak effect seen for the full AUC. For instance, if attention is limited to values of FPR < 0.2 when $a=2$ (Fig. 3), the corresponding partial AUC will be greater when $b > 0$ than when $b < 0$. Hence the partial AUC lacks symmetry with respect to b . On the other hand, the complete AUC has some compensating decreases in contributions to the area at higher values of FPR, so there are only modest changes in the total AUC as b varies (Fig. 3). We may also recall that AUC is symmetric with respect to b , so that the same summary value will be obtained for a given strength of dependence of diagnostic accuracy on the test threshold. The partial AUC does not possess these properties, and hence it will show greater dependence on the degree of interstudy heterogeneity. Further work is needed to explore the properties of the partial AUC in more detail. Similar investigation is needed for other summary indices, such as the ASC (Area Swept out by the Curve), PLC (Projected Length of the Curve),^[36,37] and the Gini/Lorenz coefficients.^[38]

Sampling variability and dependence between test accuracy and test threshold are unlikely to fully account for the observed heterogeneity in test accuracy between studies. Additional sources of heterogeneity can be explored by examining the association between test accuracy and study level covariate information.^[3,9,39] However, given the limited data on patient and study design characteristics that are typically available, it is unlikely that study level variables will fully explain the “excess” variability in test accuracy.^[27] Inappropriately assuming a fixed effect can lead to spurious precision and result in covariates being incorrectly identified as significantly associated with test accuracy.^[27] A mixed model that includes random effects for test accuracy can take into account unexplained variability between studies. This approach is used in the hierarchical (mixed) models outlined briefly below, which, unlike the Moses approach, directly models the TPR and FPR for each study.

HIERARCHICAL (MIXED) EFFECTS MODELS

Rutter and Gatsonis Model

An alternative approach to fitting SROC curves has been proposed by Rutter and Gatsonis.^[21,22] As with the Moses model, each study (i) contributes an estimate of TPR and FPR to the meta-analysis. For each study, the number in the diseased group who have a positive test result (y_{i1}) and the number in the nondiseased group who have a positive test result (y_{i2}) are both assumed to follow a binomial distribution such that $y_{ij} \sim \beta(n_{ij}, \pi_{ij})$, where π_{ij} represents the probability of a positive test for group j ($j=1,2$) in study i , and n_{ij} represents the total number of positive and

negative test results in group j . The model is based on an ordinal logistic regression model^[40] and takes the form $\text{logit}(\pi_{ij}) = (\theta_i + \alpha_i \text{dis}_{ij}) \exp(-\beta \text{dis}_{ij})$ where dis_{ij} represents the true disease status (coded as -0.5 for the nondiseased and 0.5 for the diseased). The model parameters estimate a proxy for threshold (θ_i) that gives a positive test result in study i and the diagnostic accuracy (α_i) for study i . The parameter β allows for an association between test accuracy and test threshold. The estimated SROC is symmetric when $\beta=0$. This hierarchical (multilevel) model takes into account both the within- and between-study variability.

The within-study variability is modeled at the first level. For the i th study, $\text{logit}(\text{TPR}_i)$ and $\text{logit}(\text{FPR}_i)$ are modeled to estimate the threshold (θ_i) and diagnostic accuracy (α_i) for that study. Hence random effects are assumed for both threshold and accuracy as these may vary across studies. When the SROC is symmetric (i.e., $\beta=0$), the level 1 analysis for each study reduces to an ordinary logistic regression model where α_i is estimated by $\text{logit}(\text{TPR}_i) - \text{logit}(\text{FPR}_i)$ (which is equivalent to the observed value D_i in study i for the dependent variable in model 3) and θ_i is estimated by $(\text{logit}(\text{TPR}_i) + \text{logit}(\text{FPR}_i))/2$ (which is equivalent to $S_i/2$, where S_i is the observed value in study i for the independent variable in model 3). The sampling variability for both TPR_i and FPR_i is taken into account.

The variability between studies is modeled at level 2. The random effects for test threshold and accuracy are assumed to be independent (uncorrelated) and normally distributed with $\theta_i \sim N(\Theta, \tau_\theta^2)$ and $\alpha_i \sim N(\Lambda, \tau_\alpha^2)$. The hyperparameters Θ and Λ represent the expected threshold and accuracy, respectively, across all studies in the analysis. If $\tau_\theta^2 = 0$ and $\tau_\alpha^2 = 0$, the model reduces to a fixed-effect model. Because each study contributes only one point in ROC space to the analysis, a single study does not provide information on the shape of the SROC. Hence the shape parameter (β) is assumed to be a fixed effect, which is estimated from the study points jointly considered. The hierarchical model may be fitted using Markov Chain Monte Carlo (MCMC) Bayesian methods.^[21,22] This requires a third level to specify prior distributions for all model parameters.

A summary ROC curve can be constructed by choosing a range of values of FPR and using the estimated model parameters to compute the predicted values for sensitivity. The expected TPR at a chosen FPR value is given by $\text{TPR}(\text{FPR}) = 1 / (1 + \exp[-\{\Lambda \exp(-0.5\beta) + \text{logit}(\text{FPR}) \exp(-\beta)\}])$.

The expected operating point on the curve is estimated using $E(\text{TPR}) = 1 / (1 + \exp[-\{(\Theta + \Lambda/2) \exp(-\beta/2)\}])$ and $E(\text{FPR}) = 1 / (1 + \exp[-\{(\Theta - \Lambda/2) \exp(-\beta/2)\}])$. Covariates can be added to the model to assess whether threshold, accuracy, and/or the shape of the SROC vary with patient or study characteristics. Such terms are generally fitted as fixed effects, but could also be included as random effects.

Rutter and Gatsonis demonstrate how their model may be fitted using Markov Chain Monte Carlo (MCMC) methods.^[22] A more straightforward approach that utilizes

PROC NLMIXED in SAS has been shown to provide very similar results.^[41,42] PROC NLMIXED allows for a non-linear function of the model parameters, and for a non-normal error distribution, but the random effects are restricted to be normally distributed. Summary estimates of TPR, FPR, likelihood ratios can be computed and their asymptotic confidence intervals estimated using the delta method.^[42] The distribution of the random effects may be checked by examining a histogram and normal probability plot of their Empirical Bayes (EB) estimates. This follows the same approach adopted for checking distributional assumptions for linear mixed models.^[43,44]

Bivariate Model

A bivariate model has been proposed that, like the Rutter and Gatsonis model, assumes the number correctly diagnosed (y_{ij}) for both the diseased ($j = 1$) and non-diseased ($j = 2$) groups follow a binomial distribution $B(\pi_{ij}, n_{ij})$, where π_{ij} and n_{ij} are defined as above. The random effects distribution for $\text{logit}(\text{TPR}_i)$ is assumed to follow a normal distribution with mean μ_A and variance σ_A^2 , and the random effects distribution for $\text{logit}(1-\text{FPR}_i)$ is assumed to follow a normal distribution with mean μ_B and variance σ_B^2 . Correlation between sensitivity and specificity is assumed, and estimated as σ_{AB} .^[45] The parametrization of this model allows for straightforward estimation of the expected (summary) operating point for the test which is given by exponentiating μ_A and μ_B . A 95% confidence region for this summary point can also be derived from the model estimates. The model may also be fitted using PROC NLMIXED,^[46] or using other software that can fit generalized linear mixed models.

It has been demonstrated that the Rutter and Gatsonis model and the bivariate model are equivalent when no covariates are included.^[47] Hence, the parameter estimates from either of the two models can be used to estimate the underlying summary ROC curve and the estimated operating point. However, the addition of covariates to the Rutter and Gatsonis model allows inferences to be made readily about how the position and shape of the underlying summary ROC curve varies with study and patient characteristics, whereas the addition of covariates to the bivariate model allows inferences to be made readily about how the expected sensitivity and specificity vary with study and patient characteristics.

ADVANTAGES OF THE HIERARCHICAL MODELS

The hierarchical models described above provide a general framework for the meta-analysis of diagnostic test performance. The straightforward linear regression method proposed by Moses does not directly model the TPR and FPR values for each study.^[11] Hence, because D and S are functions of both TPR and FPR, the parameters for the

Moses model cannot be used to obtain the expected operating point on the SROC, the corresponding likelihood ratios at that point, or the standard errors of these estimates. The hierarchical models do not have this limitation, and also have the advantage that they take account of both the within- and between-study variability, thereby taking account of unexplained variability in diagnostic test performance between studies through the inclusion of random effects.

CONCLUSION

The hierarchical models provide a rigorous method of analysis that takes proper account of the sources of variability in test performance between studies. The resulting standard errors (and confidence intervals) reflect the unexplained heterogeneity that is likely to be present in many diagnostic meta-analyses. The hierarchical models also allow estimation of clinically relevant indicators of test performance such as the expected TPR, FPR and likelihood ratios. A range of alternative fixed-effect and random effects methods are available for obtaining these summary measures.^[23] However, they can lead to results that are potentially inconsistent within the same meta-analysis. For instance, for the same data, summary estimates of TPR and FPR could be computed that are not consistent with summary estimates of likelihood ratios or an SROC estimated using the Moses model. A useful feature of the hierarchical models is that fixed effect approaches can be regarded as special cases. Recent work has demonstrated that the parameters from the bivariate model can be used to generate alternative SROC curves that include not only the Rutter and Gatsonis SROC curve but also the random effects equivalent of the Moses SROC curve.^[48]

The Moses fixed-effect SROC model has the advantage of simplicity, and it makes no assumptions about the distribution of S . It is useful as a preliminary step before fitting a hierarchical model. Major differences in the variables found to be associated with test accuracy or the shape of the SROC would warrant investigation as this may be indicative of convergence problems or model instability.

An issue that potentially affects all of the methods discussed here is that reference test (or gold standard) has been assumed to be error-free. In fact, the reference standard is often itself subject to measurement error, as illustrated, for example, in the observable differences between pathologists and radiologists in their assessments of the same sample material. Several methods have been proposed to correct for referent errors, but they primarily pertain to data from single studies. Rather little has been done on this problem in the context of combining studies in a meta-analysis, but one proposed approach has been to use a latent class framework to extend the logit-threshold model (Eq. [3]), and to recognize the fact that both the candidate and referent tests are potentially subject to error.^[49]

The true disease state of an individual cannot be directly observed, but an estimate of the probability of an individual having disease can be obtained from the latent class model. This then permits a deattenuation of the errors in the data, and the resulting adjustment to the SROC curve tends to lead to an improved estimate of test performance.

Further development of the SROC methods is required to allow comparison of summary curves when not all component studies in a meta-analysis are independent of one another. For example, if the purpose of a meta-analysis is to compare the diagnostic performance of alternative tests some of the component studies may make direct comparisons of two tests while other studies only evaluate one. Methods to take such dependencies into account have been proposed for therapeutic studies,^[50] and extensions to diagnostic test comparisons would be useful.

The techniques discussed here apply to situations where only summary measures (TPR and FPR) from each study are available. Sometimes access to the individual level data in each study is possible, in which case a multilevel analysis, taking both inter- and intrastudy variation into account, as well as the effects of subject-specific covariates would be feasible. However, in practice, most diagnostic meta-analyses must rely on published data where the reporting of methodology and patient characteristics is often poor.^[1,39,51,52] The reporting of diagnostic studies is likely to improve in the future because of new standards of reporting that have been adopted by many journals.^[53] Accordingly, further work to understand the properties of the SROC curve based on summary data from each study seems warranted.

REFERENCES

1. Irwig, L.; Tosteson, A.N.A.; Gatsonis, G.; Lau, J.; Colditz, G.; Chalmers, T.C.; Mosteller, F. Guidelines for meta-analyses evaluating diagnostic tests. *Ann. Intern. Med.* **1994**, *120* (8), 667–676.
2. Vamvakas, E.C. Meta-analyses of studies of the diagnostic accuracy of laboratory tests. *Arch. Pathol. Lab. Med.* **1998**, *122* (8), 675–686.
3. Deville, W.L.; Buntinx, F. Guidelines for Conducting Systematic Reviews of Studies Evaluating the Accuracy of Diagnostic Tests. *The Evidence Base of Clinical Diagnosis*; BMJ Books: London, 2002, 145–165.
4. Leeflang, M.G.; Deeks, J.J.; Gatsonis, C.; Bossuyt, P.M.M. Systematic reviews of diagnostic test accuracy. *Ann. Intern. Med.* **2008**, *149* (12), 889–897.
5. Hanley, J.A. Receiver Operating Characteristic (ROC) Curves. *Encyclopedia of Biostatistics*; Wiley: Chichester, 1998, 3738–3745.
6. Beck, J.R.; Shultz, E.K. The use of relative operating characteristic (ROC) curves in test performance evaluation. *Arch. Pathol. Lab. Med.* **1986**, *110* (1), 13–19.
7. Metz, C.E.; Herman, B.A.; Shen, J.H. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Stat. Med.* **1998**, *17* (9), 1033–1053.
8. Kardaun, J.W.P.F.; Kardaun, O.J.W.F. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. *Methods Inf. Med.* **1990**, *29* (1), 12–22.
9. Irwig, L.; Macaskill, P.; Glasziou, P.; Fahey, M. Meta-analytic methods for diagnostic test accuracy. *J. Clin. Epidemiol.* **1993**, *48* (1), 119–130.
10. Midgette, A.S.; Stukel, T.A.; Littenberg, B. A meta-analytic method for summarizing diagnostic test performances: Receiver-operating-characteristic-summary point estimates. *Med. Decis. Making* **1993**, *13* (3), 253–256.
11. Moses, L.E.; Shapiro, D.E.; Littenberg, B. Combining independent studies of a diagnostic tests into a summary ROC curve: Data-analytic approaches and some additional considerations. *Stat. Med.* **1993**, *12* (14), 1293–1316.
12. Dorfman, D.D.; Alf, E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals—rating method data. *J. Math. Psychol.* **1969**, *6* (3), 487–496.
13. Hanley, J.A. The use of the binormal model for parametric ROC analysis of quantitative diagnostic tests. *Stat. Med.* **1996**, *15* (14), 1575–1585.
14. McCullagh, P. Regression models for ordinal data (with discussion). *J.R. Stat. Soc., Ser. B* **1980**, *42* (2), 109–142.
15. Ma, G.; Hall, W.J. Confidence bands for receiver operating characteristics curves. *Med. Decis. Making* **1993**, *13* (3), 191–197.
16. Green, D.M.; Swets, J. *Signal Detection Theory and Psychophysics*; Wiley: New York, 1996.
17. Hanley, J.A. Receiver operating characteristic (ROC) methodology: The state of the art. *Crit. Rev. Diagn. Imaging* **1989**, *29* (3), 307–335.
18. Ogilvie, J.C.; Creelman, C.D. Maximum likelihood estimation of ROC curve parameters. *J. Math. Psychol.* **1968**, *5* (3), 377–391.
19. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143* (1), 29–36.
20. Hilden, J. The area under the ROC curve and its competitors. *Med. Decis. Making* **1991**, *11* (2), 95–101.
21. Rutter, C.M.; Gatsonis, G. Regression methods for meta-analysis of diagnostic test data. *Acad. Radiol.* **1995**, *2* (suppl 1), S48–S56.
22. Rutter, C.M.; Gatsonis, G. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat. Med.* **2001**, *20* (19), 2865–2884.
23. Deeks, J.J. Systematic Reviews of Evaluations of Diagnostic and Screening Tests. *Systematic Reviews in Health Care: Meta-Analysis in Context*; BMJ Publishing Group: London, 2001.
24. Fahey, M.T.; Irwig, L.; Macaskill, P. Meta-analysis of Pap smear accuracy. *Am. J. Epidemiol.* **1995**, *14* (7), 680–689.
25. Irwig, L.; Bossuyt, P.; Glasziou, P.; Gatsonis, G.; Lijmer, J. Designing Studies to Ensure that Estimates of Test Accuracy Will Travel. *Designing Studies on Diagnostic Tests*; BMJ Books: London, 2001.
26. Ransohoff, D.F.; Feinstein, A.R. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N. Engl. J. Med.* **1978**, *299* (17), 926–930.

27. Lijmer, J.G.; Bossuyt, P.M.M.; Heisterkamp, S.H. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat. Med.* **2002**, *21* (11), 1525–1537.
28. Walter, S.D. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat. Med.* **2002**, *21* (9), 1237–1256.
29. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver-operating characteristic curves: A non-parametric approach. *Biometrics* **1988**, *44* (3), 837–845.
30. Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating graph. *J. Math. Psychol.* **1975**, *12* (4), 387–415.
31. McClish, D.K. Analyzing a portion of the ROC curve. *Med. Decis. Making* **1989**, *9* (3), 190–195.
32. Thompson, M.L.; Zucchini, W. On the statistical analysis of ROC curves. *Stat. Med.* **1989**, *8* (10), 1277–1290.
33. Wieand, S.; Gail, M.H.; James, B.R. A family of nonparametric statistics for comparing diagnostic tests with paired or unpaired data. *Biometrika* **1988**, *76* (3), 585–592.
34. Jiang, Y.; Metz, C.E.; Nishikawa, R.M. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* **1996**, *201* (3), 745–750.
35. Walter, S.D. The partial area under the summary ROC curve. *Stat. Med.* **2005**, *24* (13), 2025–2040.
36. Lee, W.C.; Hsiao, C.K. Alternate summary indices for the receiver operating characteristic curve. *Epidemiology* **1996**, *7* (6), 605–611.
37. Zhang, X.; Walter, S.D.; Agnihotram, R.V. Properties of the projected length of the curve (PLC) and area swept out by the curve (ASC) indices for the receiver operating characteristic (ROC) curve. *Int. J. Biostat.* **2009**, *5* (1).
38. Lee, W.C. Probabilistic analysis of global performances of diagnostic tests: Interpreting the Lorenz curve-based summary measures. *Stat. Med.* **1999**, *18* (4), 455–471.
39. Lijmer, J.G.; Mol, B.W.; Heisterkamp, S. Empirical evidence of design related bias in studies of diagnostic tests. *JAMA* **1999**, *282* (11), 1061–1066.
40. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd Ed.; Chapman and Hall: London, 1989.
41. SAS Institute Inc. *SAS OnlineDoc*, 9.1.3; SAS Institute Inc: Cary, NC, 2002–2005.
42. Macaskill, P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *J. Clin. Epidemiol* **2004**, *57* (9), 925–932.
43. Turner, R.M.; Omar, R.Z.; Yang, M.; Goldstein, H.; Thompson, S.G. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Stat. Med.* **2000**, *19* (24), 3417–3432.
44. Verbeke, G.; Molenberghs, G. *Linear Mixed Models for Longitudinal Data*; Springer: New York, 2000.
45. Reitsma, J.B.; Glas, A.S.; Rutjes, A.W.S.; Scholten, R.J.P.M.; Bossuyt, P.M.; Zwinderman, A.H. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J. Clin. Epidemiol.* **2005**, *58*, 982–990.
46. Hamza, T.H.; van Houwelingen, H.C.; Stijnen, T. The binomial distribution of meta-analysis was preferred to model within-study variability. *J. Clin. Epidemiol.* **2008**, *61* (1), 41–51.
47. Harbord, R.M.; Deeks, J.J.; Egger, M.; Whiting, P.; Sterne, J.A. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* **2007**, *8*, 239–251.
48. Arends, L.R.; Hamza, T.H.; van Houwelingen, J.C.; Heijenbrok-Kal, M.H.; Hunink, M.G.M.; Stijnen, T. Bivariate random effects meta-analysis of ROC curves. *Med. Decis. Making* **2008**, *28* (5), 621–638.
49. Walter, S.D.; Irwig, L.; Glasziou, P.P. Meta-analysis of diagnostic tests with imperfect reference standards. *J. Clin. Epidemiol* **1999**, *52* (10), 943–951.
50. Bucher, H.C.; Guyatt, G.H.; Walter, S.D.; Griffith, L. The results of direct and indirect treatment comparisons in meta-analysis of randomized clinical trials. *J. Clin. Epidemiol.* **1997**, *50* (6), 683–691.
51. Walter, S.D.; Jadad, A.R. Meta-analysis of screening data: A survey of the literature. *Stat. Med.* **1999**, *18* (24), 3409–3424.
52. Reid, M.C.; Lachs, S.; Feinstein, A.R. Use of methodological standards in diagnostic test research: getting better but still not good. *JAMA* **1995**, *274* (8), 645–651.
53. Bossuyt, P.M.; Reitsma, J.B.; Bruns, D.E.; Gatsonis, C.A.; Galsziou, P.P.; Irwig, L.M.; et al., Standards for reporting of diagnostic accuracy. Towards complete and accurate reporting of studies for diagnostic accuracy: The STARD Initiative. *Ann. Intern. Med.* **2003**, *138*, 40–44.

Stability Analysis: Frozen Drug Products

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Jun Shao

University of Wisconsin–Madison, Madison, Wisconsin, U.S.A.

Abstract

In practice, the shelf-life of a frozen drug product is usually determined based on a two-phase stability study. This entry examines statistical models and procedures for two-phase stability analysis and presents an example concerning the establishment of a shelf-life statement for a frozen drug product.

INTRODUCTION

For every drug product in the marketplace, the Food and Drug Administration (FDA) requires that an expiration dating period be indicated on the immediate container label. The expiration dating period is also known as shelf-life. The shelf-life of a drug product is defined as the time interval in which the characteristic of the drug product (e.g., potency) remains within an approved specification after being manufactured. The shelf-life of a drug product is usually established through a stability study under the appropriate storage conditions. (See, e.g., FDA,^[1] International Conference on Harmonization [ICH],^[2] Chow,^[3] Chow and Liu,^[4] and Chow and Shao.^[5]) However, some drug products must be stored at certain temperatures, such as -20°C (frozen temperature), 5°C (refrigerator temperature), and 25°C (room temperature), in order to maintain their stability until use.^[6] Drug products of this kind are referred to as *frozen drug products*. Unlike the usual drug products, a shelf-life statement for frozen drug products is usually composed of multiple phases with different storage temperatures. For example, a commonly adopted shelf-life statement for frozen products could be *24 months at -20°C followed by 2 weeks at 5°C* .

In practice, the shelf-life of a frozen drug product is usually determined based on a two-phase stability study. The first phase of the stability study is intended to determine the drug's shelf-life under a frozen storage condition, while the second phase of the stability study is intended to estimate the shelf-life of the drug product under a refrigerated condition. The first phase of the stability study is usually referred to as the *frozen study*, while the second phase of the stability study is known as the *thawed study*. The frozen study is usually conducted in a similar manner to a regular long-term stability study, except that the drug is stored in frozen conditions. In other words, stability testing will be normally conducted at 3-month intervals during the first year, 6-month intervals during the second year, and

annually thereafter. Stability testing for the thawed study is conducted following the stability testing for the frozen study. It may be performed at 1-week (or 1-day) intervals, for up to several weeks.

Mellon^[6] suggested that stability data from the frozen study and a thawed study be analyzed separately to obtain a combined shelf-life for the drug product. However, his method does not account for the fact that the stability at the second phase (thawed study) may depend upon the stability at the first phase (frozen study), i.e., an estimated shelf-life from a thawed study following 3 months of frozen storage may be longer than that of one following 6 months of frozen storage. Shao and Chow^[7] proposed a method for the determination of drug shelf-lives for the two phases using a two-phase linear regression based on the statistical principle described in both the FDA and ICH stability guidelines.^[1,2]

In the section “Statistical Models,” statistical models and procedures for two-phase stability analysis are described. An example concerning the establishment of a shelf-life statement for a frozen drug product is given in the section “An Example.” A brief discussion is given in the section “Conclusion.”

STATISTICAL MODEL

Without loss of generality, we consider the case where the drug characteristic decreases with time. The other case where the drug characteristic increases with time can be treated similarly. For continuous assay results (e.g., potency expressed in percent of label claim), the following linear model is usually considered to describe the degradation of a drug product in stability analysis:

$$y_{ij} = \beta'_i x_{ij} + e_{ij}, i = 1, \dots, k, j = 1, \dots, n$$

where i is the index for batch, x_{ij} is a $p \times 1$ vector of non-random covariates of the form $(1, t_{ij}, w_{ij})'$, $(1, t_{ij}, t_{ij} w_{ij})'$ or

$(1, t_{ij}, w_{ij}, t_{ij} w_{ij})'$ is the j th time point for the i th batch, w_{ij} is the j th value of a vector of nonrandom covariates (e.g., package type and strength), β_i is an unknown $p \times 1$ vector of parameters, and e_{ij} 's are independent random errors (in observing y_{ij} 's), which are assumed to be identically distributed as $N(0, \sigma_e^2)$.

When $\beta_i = \beta$ for all i , we state that there is no batch-to-batch variation. In such a case, the above model reduces to a fixed effects linear regression model. When β_i 's are dissimilar, the best way to interpret the batch-to-batch variation is to consider $\beta_i, i = 1, \dots, k$, as a random *sample* from a population of all future batches, which the established shelf-life is to be applied. In this entry, we assume that β_i 's are independent and identically distributed as $N(\beta, \Sigma)$, where β and Σ are the unknown mean vector and covariance matrix of β_i , respectively. Consequently, the above model becomes a mixed effects linear regression model. The fixed effects linear regression model is a special case of the general mixed effects model. It should be noted that the statistical procedures employed in these two cases may be different (see, e.g., Ref. [5]).

Two-Phase Regression Analysis

For simplicity and the purpose of illustration, we assume that the time interval is the only covariate and that there is no batch-to-batch variation. Results for general situations can be similarly obtained.

For the first phase of the stability study, we have data

$$y_{ij} = \alpha + \beta t_j + e_{ij}, \quad i = 1, \dots, k, \quad j = 1, \dots, n$$

where y_{ij} is the assay result from the i th batch at time t_j , α , and β are unknown parameters, and e_{ij} 's are independent random errors that are identically distributed as $N(0, \sigma_e^2)$. Because there is no batch-to-batch variation, α and β do not depend on i . Hence data from different batches at each time point can be treated as replicates.

At time t_j second-phase stability data are collected at time intervals $t_{jh}, h = 1, \dots, m \geq 2$. The total number of data for the second phase is knm . Data from the two phases are independent. Typically, $t_{jh} = t_j + s_h$, where s_h is time in the second phase and $s_h = 1, 2, 3$ days (or weeks), etc.

Because the degradation lines for the two phases intersect, the intercept of the second-phase degradation line at time t is $\alpha + \beta t$. Let $\gamma(t)$ be the slope of the second-phase degradation line at time t . Then, at time t_j , we have second-phase stability data

$$y_{ijh} = \alpha + \beta t_j + \gamma(t_j) s_h + e_{ijh}, \quad i = 1, \dots, k, \quad h = 1, \dots, m$$

where $j = 1, \dots, n$ and e_{ijh} 's are independent random errors identically distributed as $N(0, \sigma_{e2}^2)$.

It is assumed that $\gamma(t)$ is a polynomial in t with unknown coefficients, i.e.,

$$\gamma(t) = \sum_{l=0}^L \gamma_l t^l$$

where γ_l 's are unknown parameters and $L + 1 < km$ and $L < n$. Typically, $L \leq 2$. When $L = 0, \gamma(t) = \gamma_0$ is a common slope model. When $L = 1, \gamma(t) = \gamma_0 + \gamma_1 t$ is a linear trend model, while when $L = 2, \gamma(t) = \gamma_0 + \gamma_1 t + \gamma_2 t^2$ is a quadratic trend model.

The Second-Phase Shelf-Life

The first-phase shelf-life can be estimated based on the first-phase data $\{y_{ij}\}$ and the method as described in the FDA stability guideline. We now consider the second-phase shelf-life. Because the slope of the second-phase degradation line may vary with t , we estimate the slope at time t_j using the second-phase data at t_j .

$$b_j = \frac{\sum_{i=1}^k \sum_{h=1}^m (s_h - \bar{s}) y_{ijh}}{k \sum_{h=1}^m (s_h - \bar{s})^2}$$

where $\bar{s}$ is the average of s_h 's. Because $L < n$, the unknown parameters γ_l can be estimated by the least squares estimators under the following model:

$$b_j = \sum_{l=0}^L \gamma_l t_j^l + \text{error}$$

i.e., the parameter vector $(\gamma_0, \gamma_1, \dots, \gamma_L)'$ can be estimated by

$$(W'W)^{-1}W'b$$

where $b = (b_1, \dots, b_n)'$, $W' = [g(t_1), \dots, g(t_n)]$, and $g(t) = (1, t, \dots, t^L)'$. Consequently, the slope $\gamma(t)$ can be estimated by

$$\hat{\gamma}(t) = g(t)'(W'W)^{-1}W'b$$

The covariance matrix of $(W'W)^{-1}W'b$ is

$$V = \frac{\sigma_{e2}^2}{k \sum_{h=1}^m (s_h - \bar{s})^2} (W'W)^{-1}$$

where the unknown variance σ_{e2}^2 can be estimated by

$$\hat{\sigma}_{e2}^2 = \frac{1}{knm - n(L+2)} \sum_{i=1}^k \sum_{j=1}^n \sum_{h=1}^m (y_{ijh} - \bar{y}_j - b_i(s_h - \bar{s}))^2$$

and $\bar{y}_j$ is the average of y_{ijh} 's with a fixed j . Let be the same as $\hat{V}$ but with σ_{e2}^2 replaced by $\hat{\sigma}_{e2}^2$. Then $(W'W)^{-1}W'b$ and its estimated covariance matrix $\hat{V}$ can be used to form approximate t -tests.

The variance of $\hat{\gamma}(t)$ can be estimated by $g(t)' \hat{V}g(t)$. For fixed i and s , let

$$v(t, s) = v(t) + g(t)' \hat{V}g(t)s^2$$

and

$$L(t, s) = \hat{\alpha} + \hat{\beta}t + \hat{\gamma}(t)s - t_{0.95; kn+knm-2-n(L+2)} \sqrt{V(t, s)}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the least square estimators of α and β based on the data from the first phase. For any fixed t less than the first-phase true shelf-life, i.e., t satisfying $\alpha + \beta t > \eta$, the second-phase shelf-life can be estimated as

$$\hat{s}^*(t) = \inf \{s \geq 0 : L(t, s) \leq \eta\}$$

(if $L(t, s) < \eta$ for all s , then $\hat{s}^*(t) = 0$), where η is the given lower limit for the drug characteristic. That is, if the drug product is taken out of the first-phase storage condition at time t , then the estimated second-phase shelf-life is $\hat{s}^*(t)$. The justification for $\hat{s}^*(t)$ is that for any t satisfying $\alpha + \beta t > \eta$,

$$P\{\hat{s}^*(t) \leq \text{the true second phase shelf life}\} \approx 95\%$$

for large sample sizes or small error variances.

In practice, the time at which the drug product is taken out of the first-phase storage condition may be unknown. In such a case, we may apply the following method to assess the second-phase shelf-life. Select a set of time intervals $t_v < \hat{t}^*$, $v = 1, \dots, J$, where $\hat{t}^*$ is the estimated first-phase shelf-life, and construct a table (or a figure) for $(t_v, \hat{s}^*(t_v))$, $v = 1, \dots, J$ (see, for example, Table 1). If a drug product is taken out of the first-phase storage condition at time t_0 , which is between t_v and t_{v+1} , then its second-phase shelf-life is $\hat{s}^*(t_{v+1})$.

AN EXAMPLE

A pharmaceutical company wishes to establish a shelf-life for a drug product at a frozen condition of -20°C followed by a shelf-life at a refrigerated condition of 5°C . During the frozen phase, the product stored at -20°C was tested at 0, 3, 6, 9, 12, and 18 months. At each sampling time point of the frozen phase, the product was tested at 1, 2, 3, and 5 weeks at the refrigerated condition of 5°C . Three batches of the drug product were used at each sampling time point ($k = 3$). Assay results in the percent of labeled concentration (g/50 ml) are given in Table 1. The p -values from the F -tests of batch-to-batch variation for the two phases of data are larger than 0.25, and, thus, we combined three batches in the analysis.

Assay results in Table 1 indicate that the degradation of the drug product at the refrigerated phase is faster at later sampling time points of the frozen phase. Assuming the linear trend model $\gamma(t) = \gamma_0 + \gamma_1 t$, we obtain estimates

Table 1 Stability Data

Time		Stability data	
t (mo)	s (day)	Frozen condition	Refrigerated condition
0		100.0, 100.1, 100.1	
3		99.2, 99.0, 99.1	
	7		97.7, 97.4, 97.4
	14		96.5, 96.2, 95.9
	21		94.8, 94.5, 94.6
	35		91.7, 91.8, 91.4
6		98.2, 98.2, 98.1	
	7		96.7, 96.5, 96.7
	14		95.4, 95.2, 95.1
	21		93.6, 93.5, 93.6
	35		90.3, 90.6, 90.5
9		97.5, 97.4, 97.5	
	7		95.9, 95.6, 95.7
	14		94.5, 94.3, 94.1
	21		92.7, 92.5, 92.5
	35		89.3, 89.5, 89.3
12		96.4, 96.5, 96.5	
	7		94.5, 94.7, 94.8
	14		92.8, 93.3, 93.4
	21		91.3, 91.4, 91.5
	35		88.2, 88.2, 87.9
18		94.4, 94.6, 94.5	
	7		92.6, 92.6, 92.4
	14		91.0, 91.1, 90.6
	21		88.9, 89.0, 89.0
	35		85.1, 85.6, 85.6

$\hat{\gamma}_0 = -0.2025$ and $\hat{\gamma}_1 = -0.0031$ with estimated standard errors of 0.0037 and 0.0004, respectively. Hence the hypothesis that $\gamma_1 = 0$ (which leads to the common slope model $\gamma(t) = \gamma_0$) is rejected based on an approximate t -test with the significance level ≤ 0.0001 .

Table 2 lists the lower confidence bounds $L(t)$ and $L(t, s)$ and the second-phase data. For the purpose of comparison, the results are given under both common slope and linear trend models. The acceptable lower product specification limit in this example is $\eta = 90$. Thus the first-phase (frozen) shelf-life for this drug product is 32 months.

The results for $L(t, s)$ in Table 2 are given at s such that $L(t, s) > 90$ whereas $L(t, s+1) < 90$. Thus s values listed in Table 2 are the second-phase (refrigerated) shelf-lives for various t values. For example, if the drug product is taken out of the frozen storage condition at month 10, then its refrigerated shelf-life is 29 days under the common slope model or 28 days under the linear trend model.

If we are unable to determine which of the two models (common slope and linear trend) is significantly better than

Table 2 Lower Confidence Bounds and Estimated Shelf-Lives

Month <i>t</i>	<i>L(t)</i>	Common slope model		Linear trend model	
		Day <i>s</i>	<i>L(t, s)</i>	Day <i>s</i>	<i>L(t, s)</i>
1	99.70	40	90.14	42	90.21
2	99.41	39	90.05	41	90.08
3	99.11	38	90.21	39	90.19
4	98.82	37	90.15	38	90.06
5	98.52	36	90.09	36	90.15
6	98.22	35	90.02	35	90.01
7	97.93	33	90.19	33	90.10
8	97.63	32	90.13	31	90.19
9	97.34	31	90.06	30	90.03
10	97.02	29	90.23	28	90.12
11	96.72	28	90.17	26	90.21
12	96.42	27	90.10	25	90.05
13	96.11	26	90.04	23	90.16
14	95.81	24	90.20	22	90.01
15	95.50	23	90.14	20	90.14
16	95.20	22	90.07	19	90.01
17	94.89	20	90.23	17	90.17
18	94.59	19	90.16	16	90.06
19	94.28	18	90.09	14	90.25
20	93.97	17	90.02	13	90.17
21	93.67	15	90.18	12	90.09
22	93.36	14	90.11	11	90.03
23	93.05	13	90.04	9	90.28
24	92.75	11	90.20	8	90.24
25	92.44	10	90.13	7	90.20
26	92.13	9	90.05	6	90.18
27	91.82	7	90.21	5	90.16
28	91.52	6	90.14	4	90.15
29	91.23	5	90.06	3	90.15
30	90.90	3	90.22	2	90.15
31	90.60	2	90.14	1	90.15
32	90.30	1	90.07	0	90.30
33	89.98	—	—	—	—

the other, then we could adopt a conservative approach by selecting the shorter of the shelf-lives computed under the two models. Note that a higher-degree polynomial model does not always produce shorter shelf-lives at the second phase. In the example, the linear trend model produced shorter second-phase shelf-lives in later months, but longer second-phase shelf-lives in early months. This is because under the common slope model, the estimated common slope is the average of the five estimated slopes (in different months) that decrease with time (*t*), because $\hat{\gamma}_1 < 0$.

CONCLUSION

In practice, the assumption of simple linear regression in the second phase of a stability study may not be appropriate. For example, there may be an acceleration in decay with *s* in the second phase. The method described in this entry can be easily extended to the case where polynomial regression models are considered in the second phase. For example, suppose that in the second phase,

$$y_{ijh} = \alpha + \beta t_j + \gamma(t_j)s_h + \rho(t_j)s_h^2 + e_{ijh}$$

where both $\gamma(t)$ and $\rho(t)$ are polynomials, At month t_j , let b_j and c_j be the least squares estimates of the coefficients of the linear and quadratic terms in the second-phase quadratic model. Then, the estimators of $\gamma(t)$ and $\rho(t)$ can be obtained using

$$b_j = \gamma(t_j) + \text{error}$$

and

$$c_j = \rho(t_j) + \text{error}$$

Usual model diagnostic methods may be applied to select an adequate model.

The selection of sampling lime points at the second-phase stability study depends on how fast the degradation would be. It may be a good idea to have shorter time intervals for the second phase in later months; for example, we may test the drug product more frequently in later weeks. For a fixed total sample size, it is of interest to examine the relative efficiency for the estimation of shelf-lives using either more sampling time points in the first phase and less sampling time points in the second phase or less sampling time points in the first phase and more sampling time points in the second phase. The allocation of sampling lime points at each phase then becomes an interesting research topic for two-phase shelf-life estimation. Furthermore, because the degradation at the second phase is highly correlated with the degradation at the first phase, it may be of interest to examine such correlation for future design planning.

For the method described in this article, equal assay variabilities for the second-phase data at different t_j are assumed. If this assumption is not true, the variance estimator should be modified.^[8] Note that the method described in this entry can be extended to the case of multiple-phase shelf-life estimation in a straight forward manner.

REFERENCES

1. FDA. *Guideline for Submitting Documentation for the Stability of Human Drugs and Biologics*; Center for Drugs and Biologics, Office of Drug Research and Review, Food and Drug Administration: Rockville, MD, 1987.

2. ICH. Stability Testing of New Drug Substances and Products. *Tripartite International Conference on Harmonization Guideline*, 1993.
3. Chow, S.C. *Statistical Design and Analysis of Stability Studies*; Chapman & Hall/CRC, Taylor & Francis: New York, 2007.
4. Chow, S.C.; Liu, J.P. *Statistical Design and Analysis in Pharmaceutical Science*; Marcel Dekker, Inc: New York, 1995.
5. Chow, S.C.; Shao, J. *Statistics in Drug Research—Methodologies and Recent Development*; Marcel Dekker, Inc: New York, 2002.
6. Mellon, J.I. Design and Analysis Aspects of Drug Stability Studies When the Product is Stored at Several Temperatures. *Presented at the 12th Annual Midwest Statistical Workshop*, Muncie, IN, 1991.
7. Shao, J.; Chow, S.C. Shelf life estimation. *Stat. Sin.* **2001**, *11*, 737–745.
8. Shao, J.; Chow, S.C. Two-phase shelf-life estimation. *Stat. Med.* **2001**, *20*, 1239–1248.

Stability Matrix Designs

Earl Nordbrock

Nordbrock Consulting, Draper, Utah, U.S.A.

Abstract

For a new drug product, accelerated testing is required for 6 months, and long-term testing is required for the length of shelf life; thus the cost of the stability studies can be substantial. This leads quite naturally to statistically designed stability studies, which are called matrix designs. This entry provides several models of stability matrix designs and discusses the general rules when using each model.

Keywords: Matrix design; Stability matrix; Time design.

INTRODUCTION

Stability studies are conducted “to provide evidence on how the quality of a drug substance or drug product varies with time under the influence of a variety of environmental factors such as temperature, humidity, and light, and to establish a retest period for the drug substance or a shelf life for the drug product and recommended storage conditions.”^[1] Because a drug product is defined as the dosage form in the final package, stability studies must be applicable to all strengths and all packages. For a new drug product, accelerated testing is required for 6 months, and long-term testing is required for the length of shelf life; thus the cost of the stability studies can be substantial. This leads quite naturally to statistically designed stability studies, which are called *matrix* designs, or studies where “a selected subset of the total number of possible samples for all factor combinations is tested at a specified time point.”^[1]

HISTORY

Statistically designed stability studies were first used in the early 1980s (E. Nordbrock, 1981, stability protocols) and were accepted by the U.S. Food and Drug Administration (FDA) (W. Fairweather, 1982, personal communication). In 1989, Nordbrock^[2] and Wright^[3] made presentations to stability groups at two different conferences. Nakagaki,^[4] who was present at the first of these conferences, made a presentation using the terminologies *matrix* and *bracket*. Nordbrock^[5] made a presentation in 1991, and the first journal articles appeared in 1992.^[6–8] Since then, there have been a number of presentations and publications in this area.^[9–28]

A guidance on bracketing and matrixing was issued in 2003.^[29] A guidance on evaluation of stability data was issued in 2004.^[30]

DESIGNS

Background

Because the basic analysis applied to stability data is a linear regression of the parameter of interest on time, the selection of observations that give the minimum variance for the slope is to take one-half at the beginning of the study and one-half at the end. Thus statistically designed stability designs attempt to use this basic principle while being cognizant of the fact that there is a well-defined beginning (start of the stability study, usually called $t - 0$) but not a single end, because analyses are typically done at several different times (e.g., at the time a registration application for the new drug product is filed, or yearly for a marketed product). Because analyses are done at multiple times, there is no unique best design, and the choice of design must use the fact that analyses will be done after additional data are collected. Several designs are given next. These designs, presented via examples, can easily be applied to other studies.

Basic Matrix 2/3 on Time Design

A complete long-term study for one strength of a dosage form in one package has three batches, with all three tested every 3 months in the first year, every 6 months in the second year, and annually thereafter (see the entry on *Expiration Dating Period*). Thus if a 36-month shelf life is desired and the complete study is used, each of the three batches is tested at 0, 3, 6, 9, 12, 18, 24, and 36 months. The basic matrix 2/3 on time design has only two of the three batches tested at intermediate time points (other than time 0 and 36), as presented in [Table 1](#). If an analysis is to be done after 18 months (e.g., for a registration application), the basic matrix 2/3 on time design can be modified by testing all batches at 18 months.

Table 1 Basic Matrix 2/3 on Time Design

Batch	Test times
A	0, 3, 9, 12, 24, 36
B	0, 3, 6, 12, 18, 36
C	0, 6, 9, 18, 24, 36

Table 2 Matrix 2/3 on Time Design with Multiple Packages

Batch	Pkg 1	Pkg 2	Pkg 3
A	T1	T2	T3
B	T2	T3	T1
C	T3	T1	T2

Pkg 1 = Package 1, etc.

Table 3 Test Code Definitions

Code	Test times after time 0
T1	3, 9, 12, 24, 36
T2	3, 6, 12, 18, 36
T3	6, 9, 18, 24, 36

Batches are tested at time 0, consistent with the manufacturing process.

Matrix 2/3 on Time Design with Multiple Packages

The first extension of the basic design is when one strength is packaged into three packages (i.e., when each batch is packaged into each of three packages). The basic matrix 2/3 on time design is applied to each package in a balanced fashion, as presented in Tables 2 and 3. Balance means that each batch is tested twice at each intermediate time point, and each package is tested twice at each intermediate time point. If an analysis will be done after 18 months (e.g., for a registration application), this design can be modified by testing all batch-by-package combinations at 18 months.

Matrix 2/3 on Time Design with Multiple Packages and Multiple Strengths

The next extension is when three strengths (say, 10, 20, and 30) are manufactured using different weights of the same formulation, giving nine subbatches. It is further assumed that there are three packages for each strength. In this case, the basic matrix 2/3 on time design can be applied to each of the nine subbatches in a balanced fashion, as presented in Tables 4 and 5. Balance in this design means that each subbatch is tested twice at each intermediate time point, each package is tested twice at each intermediate time point for each batch, each batch is tested six times at each intermediate time point, and each package is tested six times at each intermediate time point. If an analysis will be done after 18 months (e.g., for a registration application), this design can be modified by testing all batch-by-strength-by-package combinations at 18 months.

Table 4 Matrix 2/3 on Time Design with Multiple Packages and Multiple Strengths

Batch	Strength	Pkg 1	Pkg 2	Pkg 3
A	10	T1	T2	T3
A	20	T2	T3	T1
A	30	T3	T1	T2
B	10	T2	T3	T1
B	20	T3	T1	T2
B	30	T1	T2	T3
C	10	T3	T1	T2
C	20	T1	T2	T3
C	30	T2	T3	T1

Pkg 1 = Package 1, etc.

Table 5 Test Code Definitions

Code	Test times after time 0
T1	3, 9, 12, 24, 36
T2	3, 6, 12, 18, 36
T3	6, 9, 18, 24, 36

Strength subbatches are tested at time 0, consistent with the manufacturing process.

Matrix 1/3 on Time Design

A further reduction in the amount of testing is accomplished by reducing the testing in each of the preceding designs from 2/3 to 1/3. For example, the basic 1/3 on time design has one of the three batches tested at each intermediate time point, as presented in Table 6. If an analysis will be done after 18 months (e.g., for a registration application), the basic matrix 1/3 on time design can be modified by testing all batches at 18 months.

Matrix on Batch $\times$ Strength $\times$ Package Combinations

If there are multiple strengths and multiple packages, one could also choose to test only a portion of the batch-by-strength-by-package combinations. An example of when this might be appropriate is when there are three batches, each made into two strengths, giving six subbatches. Although three packages will be used, the batch size is small and only two packages can be manufactured in each strength subbatch. A matrix design on batch $\times$ strength $\times$ package combinations is presented in Tables 7 and 8, with two packages selected for each of the six subbatches, and where time is also matrixed by the factor 1/2. This design is approximately balanced because two packages are tested per subbatch, one or two strengths are tested for each selected package by batch, four subbatches are tested for each package, etc. Similar statements for the balance on time can be made.

Table 6 Basic Matrix 1/3 on Time Design

Batch	Test times
A	0, 3, 12, 36
B	0, 6, 18, 36
C	0, 9, 24, 36

Table 7 Matrix 1/2 on Time and Matrix on Batch $\times$ Strength $\times$ Package

Batch	Strength	Pkg 1	Pkg 2	Pkg 3
A	10	T1	T2	–
A	20	T2	–	T1
B	10	T2	–	T1
B	20	–	T1	T2
C	10	–	T1	T2
C	20	T1	T2	–

Table 8 Test Code Definitions

Code	Test times after time 0
T1	3, 9, 18, 36
T2	6, 12, 24, 36

Strength subbatches are tested at time 0, consistent with the manufacturing process.

Uniform Matrix Design

Another approach to design is the uniform matrix design,^[16] for which “the same time protocol is used for all combinations of the other design factors.” The strategy is to delete certain times (e.g., the 3-month, 6-month, 9-month, and 18-month time points); therefore testing is done only at 12, 24, and 36 months. This design has the advantages of simplifying the data entry of the study design and eliminating time points that add little to reducing the variability of the slope of the regression line. The disadvantage is that if there are major problems with the stability, there is no early warning because early testing is not done. Further, it may not be possible to determine if the linear model is appropriate (e.g., it may not be possible to determine whether there is an immediate decrease followed by very little decrease). However, the major disadvantage is that this design is probably not acceptable to some regulatory agencies.

ANALYSIS

Background

Although there may be instances when a linear regression is not appropriate, the rest of this discussion assumes that a straight-line linear regression of the parameter

of interest on time is appropriate. Further, it is assumed that the parameter of interest is expected to decrease over time. For long-term data with a single package and a single strength, the ICH Q1E Guidance^[30] specifies that the 95% one-sided lower confidence bound for the mean regression line must be above the lower specification at all times prior to the shelf life. When there are multiple strengths and/or multiple packages, according to ICH Q1E^[30] there are three possible approaches for analysis. The first approach is to analyze each package-by-strength combination separately—in other words, to do multiple analyses. The second approach is to model all data with one analysis, with separate intercepts and separate slopes for each batch by strength by package, without testing for poolability, so there is no reduced model. The third approach is to model all data with one analysis and test for poolability and then select the appropriate reduced model.

SEPARATE ANALYSIS APPROACH

In the first approach, multiple analyses are conducted, with a separate analysis of each package-by-strength-by-batch. A shelf life is calculated for each of the separate analyses, and the product shelf life is the minimum of all the package-by-strength-by-batch shelf lives.

One Analysis Approach, without Testing Poolability

In the second approach, one analysis is done which includes all data. A shelf life is estimated using individual intercepts, individual slopes, and the pooled mean square error from the entire data set. Shelf life for each batch-by-package-by-strength is taken as the time the batch-by-package-by-strength remains within acceptable limits, i.e., the time that the 95% one-sided lower confidence bound for the mean regression line remains above specification.

If there are multiple batches but only one package and one strength, the SAS model is $Y = B A B^*A$, where B is a class term for batch and A is the covariate for age. The 95% one-sided lower confidence bound for the mean regression line is calculated for each batch. The shelf life of each batch is such that the confidence bound is within the specification at all times prior to the shelf life. The shelf life of the product is the minimum of the batch shelf lives.

If there are multiple batches and multiple packages but only one strength, and if the initials for a particular batch are applicable to all packages, then the SAS model is $Y = B A B^*A P^*A B^*P^*A$, where B is a class term for batch, P is a class term for package and A is the covariate for age. This model has separate slopes for each batch-by-package, separate intercepts for each batch, and a common intercept for all packages from the same batch. For example, if initials are tested using bulk product (before the product is packaged), and all packages are entered

into the study at the same time, then this model is typically appropriate. The 95% one-sided lower confidence bound for the mean regression line is calculated for each batch by package. The shelf life of each batch-by-package is such that the confidence bound is within the specification at all times prior to the shelf life. The shelf life of the product is the minimum of all batch-by-package shelf lives.

One Analysis: Testing Poolability

When using the model-building (test poolability) approach, it is very important that the full model reflect the manufacturing process. In this section, it is assumed that there are multiple strengths and multiple packages and that a batch is manufactured into multiple strengths by using different weights of the same exact formulation. It is assumed that a granulation batch is split into subbatches, where each subbatch is manufactured into a different-strength product using different weights of the granulation, and it is assumed that every strength is manufactured from every granulation batch. It is assumed that the time 0 samples are collected from each tablet subbatch, and every tablet subbatch is packaged into all packages.

The full model includes slope terms for all two-way interactions of package, strength, and batch, and it includes intercept terms that reflect the manufacturing process. The manufacturing process dictates that each tablet subbatch must have a separate intercept in the full model, but there is no need to allow packages in each tablet subbatch to have separate intercepts. (Process validation provides evidence that the entire tablet subbatch is uniform.)

Thus in the example, the full SAS model is: $Y = B \ S \ B*S \ A \ B*A \ P*A \ S*A \ B*P*A \ B*S*A \ P*S*A$ with B as a class term for granulation batch, P as a class term for package, S as a class term for strength, and A as a covariate for time. The model building begins by testing the slope two-way interactions to determine if any can be deleted, using a significance level of 0.25 when the batch is part of the term and 0.05 otherwise. Then the main-effect slope terms are tested, using the 0.25 (batch) or 0.05 (not batch) level. Terms that are not significant are deleted, except that any main-effect slope included in a nondeleted two-way slope term cannot be deleted.

After deleting slope terms, the intercept terms are tested (using the same criterion for significance as was used for slopes) and nonsignificant terms are deleted. Using the final model, the 95% one-sided lower confidence bound for the mean regression line(s) is (are) found, and shelf life is assigned for each package-by-strength combination such that the 95% one-sided lower confidence bound(s) is (are) within specification at all times prior to the shelf life.

Slightly different algorithms for deleting terms from the full model have been proposed.^[14,27] A correction to Ref. [27] is planned.^[28]

COMPARISON OF DESIGNS

Nordbrock^[6] compared designs based on the power approach. This approach computes the probability that a statistical test will be significant when there is a specified alternative slope configuration. Power can be computed easily in SAS.^[23] The strategy is to compute power for several designs and then to choose the design that has acceptable power and the smallest sample size (or cost). Acceptable power is not well defined at this stage. Note that in Ref. [2] the power of specific $1 - df$ contrasts were calculated using a $1 - df(t)$ test, whereas in Ref. [23] the power of specific contrasts is computed using the corresponding F -test from the analysis of variance.

Other methods of comparing designs are given in Ju and Chow,^[18] where the criterion is the precision for estimating shelf life. Murphy^[16] gives five criteria for comparing designs, based on moments, D -efficiency, uncertainty, G -efficiency, and power.

When evaluating designs, it is also important to answer the question "What is the probability of being able to defend the desired shelf life with the study?"^[23] In other words (assuming that the parameter is expected to decrease over time), what is the probability that the 95% one-sided lower confidence bound for the slope will be acceptable for specified values of the slope(s) for particular subsets of data, which may include, for example, only one strength and/or only one package? It is important to know at the design stage what the statistical penalty (with respect to shelf life) might be if differences among packages and/or strengths are found. Similarly, Nordbrock^[26] compares matrix designs to full designs using the probability of achieving the desired shelf life.

FACTORS ACCEPTABLE TO MATRIX

In the foregoing, examples have been used to present possible matrix designs. In this section, a summary of when it is acceptable to matrix is given based on a document prepared by the PhRMA Stability Working Group,^[20] on FDA presentations;^[21,22] and on ICH Q1D.^[29]

- It is acceptable to matrix at all stages of development for a drug product and also for a drug substance. It is acceptable for new drug application (NDA) studies, investigational new drug application (IND) studies, supplements, and marketed product studies.
- It is acceptable to matrix for all types of products, such as solids, semisolids, liquids, and aerosols.
- It is acceptable to matrix after bracketing.
- It is acceptable to matrix when there are multiple sources of raw material (e.g., drug product).
- It is acceptable to matrix if there are multiple sites of drug product manufacture.
- It is acceptable to matrix when identical formulations are manufactured into several strengths.

- It is acceptable to matrix if formulations are closely related (e.g., difference in colorant or flavoring).
- Matrixing is applicable to the orientation of container during storage.
- Matrixing may be acceptable in certain cases when closely related formulations are used for different strengths (e.g., if an inactive is replaced by an active).
- Matrix across container and closure systems may be applicable if justified.
- It is acceptable to matrix within a package composition type, e.g., of different sizes if fill (i.e., head space) is the same, or if of the same size but different fills (head space). It may be acceptable to matrix if container size and fill size change, if there is adequate explanation. It is not acceptable to matrix across package composition type (e.g., blisters and HDPE).
- It is not acceptable to matrix across storage conditions. However, it is acceptable to do a separate matrix design for each storage condition.
- It is not acceptable to matrix across parameters, such as dissolution and potency. However, it is acceptable to do a separate design for each parameter.
- Matrixing is applicable regardless of method precision; however, it should be remembered that when using a matrix design, the resulting shelf life is generally shorter than when a complete design is used and that when the method precision is larger, the difference between a complete design and a matrixed design will be larger (i.e., a larger penalty to the sponsor, resulting in a shorter shelf life for the matrix design than the complete design).
- Matrixing is applicable regardless of the stability of the product. However, comments similar to those in the preceding point apply, and it should be remembered that if a product has a poor stability profile (e.g., shelf life of 1 year), matrixing will usually result in an even shorter shelf life.
- The latest guideline should be consulted for applicability.

GENERAL RULES

Several general rules should be followed when designing studies:

- Matrix designs should be approximately balanced (i.e., for all one-way, two-way combinations of batch, package, and strength that are ever tested, approximately the same number of tests should be done cumulatively to every time point).
- When every batch-by-strength-by-package combination is not tested, every strength-by-package combination that is ever tested should be tested in at least two batches (i.e., for every package-by-strength combination that is ever tested, there should be at least two batches tested; W. Fairweather, 1995, personal communication).

- Unless there are manufacturing restrictions such as in the foregoing example, it is probably acceptable to matrix on batch $\times$ strength $\times$ package combinations only when there are more than three strengths or more than three packages.

CONCLUSION

Matrix designs are generally applicable to many situations and can result in significant savings, with the 1/3 matrix on time readily acceptable for stable products. There are two basic approaches when analyzing data from a matrixed design. There are several methods used to evaluate and to compare potential designs.

REFERENCES

1. *Guidance for Industry. Q1A(R2) Stability Testing of New Drug Substances and Products*; November, 2003.
2. Nordbrock, E. Statistical Study Design. *Presentation at National Stability Discussion Group*; October, 1989.
3. Wright, J. Use of Factorial Designs in Stability Testing. *Proceedings of Stability Guidelines for Testing Pharmaceutical Products: Issues and Alternatives. AAPS Meeting*; December, 1989.
4. Nakagaki, P. *AAPS Annual Meeting*; 1990.
5. Nordbrock, E. Statistical Comparison of NDA Stability Study Designs. *Midwest Biopharmaceutical Statistics Workshop*; May, 1991.
6. Nordbrock, E. Statistical comparison of stability study designs. *J. Biopharm. Stat.* **1992**, 2, 91–113.
7. Helboe, P. New designs for stability testing programs: Matrix or factorial designs. Authorities' viewpoint on the predictive value of such studies. *Drug Inf. J.* **1992**, 26, 629–634.
8. Carstenson, J.T.; Franchini, M.; Ertel, K. Statistical approaches to stability protocol design. *J. Pharm. Sci.* **1992**, 81, 303–308.
9. Chow, S.C. Statistical Design and Analysis of Stability Studies. *48th Annual Conference on Applied Statistics*; 1992.
10. Nordbrock, E. Design and Analysis of Stability Studies. *ASA Proceedings of Biopharmaceutical Section*; 1994, 291–294.
11. Nordbrock, E. Statistically Designed Stability Studies, an Industry Perspective. *Presentation at Biostatistics Subsection/Clinical Data Management Group of Pharmaceutical Research and Manufacturers of America*; October, 1994.
12. Lin, T.D. Applicability of Matrix and Bracket Approach to Stability Study Design. *ASA Proceedings of Biopharmaceutical Section*; 1994, 142–147.
13. Fairweather, W.R.; Lin, T.D.; Kelly, R. Regulatory and Design Aspects of Complex Stability Studies. *Presentation at Biostatistics Subsection/Clinical Data Management Group of Pharmaceutical Research and Manufacturers of America*; October, 1994.
14. Fairweather, W.R.; Lin, T.D.; Kelly, R. Regulatory, design, and analysis aspects of complex stability studies. *J. Pharm. Sci.* **1995**, 84, 1322–1326.

15. Golden, M.H.; Cooper, D.C.; Riebe, M.T.; Carswell, K.E. A matrixed approach to long term stability of pharmaceutical products. *J. Pharm. Sci.* **1996**, *85*, 240–245.
16. Murphy, J.R. Uniform matrix stability study designs. *J. Biopharm. Stat.* **1996**, *6*, 477–494.
17. DeWoody, K.; Raghavarao, D. Some optimal matrix designs in stability studies. *J. Biopharm. Stat.* **1997**, *7*, 205–213.
18. Ju, H.L.; Chow, S.C. On stability designs in drug shelf-life estimation. *J. Biopharm. Stat.* **1995**, *5*, 210–214.
19. CPMP. *Reduced Stability Testing Plan—Bracketing and Matrixing* (CPMP/QWP/157/96); October, 1997.
20. Nordbrock, E.; Valvani, S. PhRMA Stability Working Group. *Guideline for Matrix Designs of Drug Product Stability Protocols*; January, 1995.
21. Chen, C. FDA's Views on Bracketing and Matrixing. *EFPIA Symposium: Advanced Topics in Pharmaceutical Stability Testing—Building in the ICH Guideline*; October, 1996.
22. Lin, D. Stability Studies at the FDA. *PERI Course: Non-Clinical Statistics for Drug Discovery and Development*; March, 1997.
23. Nordbrock, E. Computing Power Details. *PERI: Training Course in Non-Clinical Statistics*; February, 1994.
24. Pong, A.; Raghavarao, D. Comparison of bracketing and matrixing designs for a two-year stability study. *J. Biopharm. Stat.* **2000**, *10*, 217–228.
25. Chow, S.-C. *Statistical Design and Analysis of Stability Studies*; Chapman & Hall: New York, 2007.
26. Nordbrock, E. Use of Statistics to Establish a Stability Trend: Matrixing. In *Pharmaceutical Testing to Support Global Markets*; Huynh-Ba, K., Ed.; Springer: New York, 2009.
27. Tsong, Y.; Chen, W.-J.; Chen, C.W. Ancova Approach for Shelf Life Analysis of Stability Study of Multiple Factor Designs. *J. Biopharm. Stat.* **2008**, *13* (3), 375–393.
28. Tsong, Y. Personal Communication, 2008.
29. *Guidance for Industry. Q1D Bracketing and Matrixing Designs for Stability Testing of New Drug Substances and Products*; November, 2003.
30. *Guidance for Industry Q1E Evaluation of Stability Data*; June, 2004.

Statistical Genetics

Rongling Wu

Departments of Public Health Science and Departments of Statistics, Pennsylvania State University, University Park, Pennsylvania, U.S.A.

Guifang Fu

Departments of Statistics, Pennsylvania State University, University Park, Pennsylvania, U.S.A.

Hongying Li

Departments of Statistics, University of Florida, Gainesville, Florida, U.S.A.

Abstract

Genetic mapping with molecular markers provides a powerful tool for identifying the genes underlying a complex trait. Molecular markers are distinctive segments of DNA that serve as landmarks for a target gene. This entry examines genetic mapping in terms of statistical genetics.

GENETIC MAPPING

Most traits in nature and of fundamental importance to agriculture, biology, and health sciences are complex in terms of their genetic underpinnings. First, these traits are usually controlled by a network of genes that act independently or interactively although an exact pattern of genetic action is never known. Second, the expression of the genes involved depends on the genetic background of an organism, the stage of its development, and the environment in which the organism is reared. Because of these characteristics, classic Mendelian genetic approaches have been little useful for studying the genetic architecture of complex traits.

Genetic mapping with molecular markers provides a powerful tool for identifying the genes underlying a complex trait. Molecular markers are distinctive segments of DNA that serve as landmarks for a target gene. Markers are genetically linked to the target gene on a chromosome, so that through linkage analysis the latter can be inferred from the former. With the availability of high-throughput molecular markers almost for any genome, the genetic analysis of complex traits has entered a new era in which genes that control a complex trait can be located individually on the genome and their genetic effects can be estimated. The motive force that allows the identification of individual genes for complex traits, or quantitative trait loci (QTLs), comes from Lander and Botstein^[1] for interval mapping of QTLs.

Because QTLs cannot be observed directly, Lander and Botstein constructed a mixture model-based likelihood, in which each individual is assumed to arise from one of all possible QTL genotypes. The likelihood is then solved by implementing the expectation-maximization

(EM) algorithm. The existence of significant QTLs can be tested and their genetic effects on a complex trait estimated. The past 20 years have witnessed a resurgence of interest in the development of statistical models and algorithms for genetic mapping of complex traits. Broadly speaking, these new methods have the following aims, among others: improving the precision of QTL separation from a chromosomal segment (composite interval mapping);^[2,3] extending genetic mapping to accommodating dominant markers;^[4] outbred crosses;^[5] complicated mating designs;^[6] and natural populations;^[7] mapping multiple interacting QTLs (multiple interval mapping);^[8] mapping QTLs for trait correlations^[9] and genotype-environment interactions;^[10] mapping QTLs for binary^[11] and categorical traits;^[12] and deriving QTL mapping procedures within the Bayesian paradigm.^[13,14] Specific statistical issues for QTL mapping have also been considered in several areas including the determination of critical thresholds,^[15,16] model selection,^[17] non-parametric mapping of QTLs,^[18] and asymptotic properties of QTL parameter estimates.^[19]

These statistical mapping models have played an important role in the characterization of the genetic architecture of complex traits. Thousands of thousands of QTLs have been reported in plants, animals, and humans. In several examples, the detection of QTLs by genetic mapping leads to the positional cloning of genes for quantitative traits, such as grain yield and its components in agriculture,^[20] life-history traits in evolutionary biology,^[21] and diseases in humans.^[22] As a statistical approach, the theory and methods for genetic mapping have continued to grow steadily. Especially, integrated with biological problems and genomic technologies, this approach is seeking new motive forces and expected to produce new breakthroughs.

FUNCTIONAL MAPPING

The integration of genetic mapping with growth and developmental processes leads to the birth of functional mapping. Many biological processes in real life are expected to arise as curves, such as growth curves, allometric scalings, hormone profiles, and norms of reaction. A growth curve or trajectory represents an individual as a function that relates the age of an individual to some measure of its size. Because ages can be considered as continuous, growth trajectories can be thought of as function-valued traits^[23,24] or infinite-dimensional characters.^[25] Other examples of function-valued traits include the continuous change of a morphological or physiological variable with body size (allometric scaling)^[26,27] and responsive phenotypes of a given genotype to a changing environment (reaction norm).^[28] The common property of these function-valued traits is that they can be described as a function (or stochastic process) of some independent and continuous variable consisting of an infinite number of points, such as age, temperature, light intensity, or biological size.

To determine the dynamic changes of genetic effects in development, one can measure trait values at a series of discrete time points or states and then extend the traditional interval mapping approach to accommodate the multivariate nature of time-dependent traits. However, this extension is limited in three aspects:

- i. Individual expected means of different QTL genotypes at all points and all elements in the covariance matrix need to be estimated, resulting in substantial computational difficulties when the vector and matrix dimensions are large;
- ii. The result from this approach may not be biologically meaningful because the underlying biological principle for growth is not incorporated;
- iii. This approach cannot be well deployed in a practical scheme because of (ii). Thus, some biologically interesting questions cannot be asked and answered.

More recently, a library of statistical methods implemented with growth model theories have been proposed to map QTL that govern growth trajectories using molecular linkage maps.^[29–34] The basic principle of this method, called functional mapping, is to express the genotypic means of a QTL at different time points in terms of a continuous growth function with respect to time t . Under this principle, the parameters describing the shape of growth curves, rather than the genotypic means as expected in traditional mapping strategies, are estimated within a maximum likelihood framework. Also unlike traditional mapping strategies, functional mapping estimates the parameters that model the structure of the covariance matrix among multiple different time points and, therefore, largely reduces the number of parameters being estimated for variances and

covariances, especially when the number of time points is large.

Although the idea of functional mapping was originally stimulated from a growth model for a controlled cross, this method has been extensively expanded to a wide spectrum of biomedical and evolutionary fields. Wang and Wu^[35] and Wang et al.^[36] combined the principle of functional mapping and mathematical models for HIV dynamics established by Ho et al.^[37] to map host QTL that govern dynamic changes of HIV viral loads in a human body. Functional mapping has been extended to study the genetic control of drug response^[38] by implementing well-developed pharmacodynamic and pharmacokinetic models.^[39] “Naturally occurring” or “programmed” cell death (PCD) in which the cell uses specialized cellular machinery to kill itself is a ubiquitous phenomenon that occurs early in organ development. A recently extended functional-mapping model has made it possible to identify and detect QTL that are responsible for PCD in any organism.^[40]

In theory, genetic information contained within any kind of longitudinally measured data can be extracted by functional mapping. The advantages of functional mapping in model flexibility, stability, and power result from its statistical formulation including the mathematical approximation of the mean vector and the modeling of the covariance structure by stationary^[29] or non-stationary models.^[41] Although original functional mapping was derived for the biological processes that can be described by parametric functions, functional mapping has been modified within the nonparametric context to accommodate the situation in which biologically meaningful mathematical equations do not exist.^[40] Analysis of longitudinal data needs a number of mathematical manipulations for the covariance matrix, such as the calculation of the matrix determinant and inverse. In modeling the structure of covariance matrix by parametric functions, however, these mathematical manipulations will not be made possible because of the matrix's sparse structure, especially when the dimension of the matrix is too large. To overcome this problem, Zhao and Wu^[42] have proposed a denoising approach based on wavelet transformation to reduce the dimensionality of longitudinal data. Preserving the favorable properties of functional mapping, wavelet thresholding approach creates a new direction in statistical genetics to manipulate high-dimensional multivariate dynamic data in both statistically and biologically meaningful ways.

FROM QTL TO QTN

Although QTL mapping has proven to be powerful for study the genetic control of complex traits at the molecular level, QTLs identified on the basis of one or a set of known polymorphic markers by this approach actually represent chromosomal segments in which the structure and organization of DNA sequences are unknown. A more accurate

and useful approach for the characterization of genetic variants contributing to quantitative variation is to directly analyze DNA sequences associated with a particular disease.^[43] DNA sequence variations between individuals are responsible for our biological and physical differences. Although approximately 99.9% of human DNA sequences are the same across a population, the remaining 0.1% variations in these sequences can have a major impact on how humans respond to disease, bacteria, viruses, toxins, chemicals, and drugs.^[44–46]

Many of these DNA sequence variants lie outside genes and may have no effect on the individual. Some of the variations lie within genes, so they may alter the structure of the proteins that are encoded by these genes and thus may affect the functioning of these proteins. Some variants may lie within the region responsible for regulating the expression of the gene and thus may alter the level of gene expression, the specific timing, or the location of expression.

SINGLE NUCLEOTIDE POLYMORPHISMS

Single nucleotide polymorphisms or SNPs (pronounced “snips”) are single base pair positions in genomic DNA sequence at which a single nucleotide (A, T, C, or G) in the genome differs, so different alleles exist, between members of a species (or between paired chromosomes

in an individual). For example, two DNA sequences on the same chromosome segment from two different individuals, AGCCT and AGCTT, contain a difference in a single nucleotide (C to T). In this case, we say that there are two alleles: C and T. Almost all common SNPs have only two alleles. The specific set of alleles observed on a single chromosome, or part of a chromosome, is called a haplotype. In Fig. 1, four DNA sequences of the same chromosome region from four different persons are compared. Most of the DNA sequence is identical; only three bases have variation, which are called single nucleotide polymorphisms or SNPs (Fig. 1a). By taking out the SNPs from original DNA sequence, a particular combination of bases at nearby SNPs can be constructed, which is called haplotype (Fig. 1b). SNPs are the most frequent form of sequence variation among individuals, accounting for around 90% of sequence differences.^[47] The rest may due to insertions or deletions, repeats, and rearrangements. These SNPs are present at high density in the genome, highly conserved and also evolutionarily stable making them powerful tools for the mapping and diagnosis of disease related alleles. SNPs are of great value for biomedical research and for developing pharmaceutical products or medical diagnostics.

SNP alleles that are closely located tend to be inherited together. This leads to associations between these alleles in the population, called linkage disequilibrium (LD). Many empirical studies found highly significant

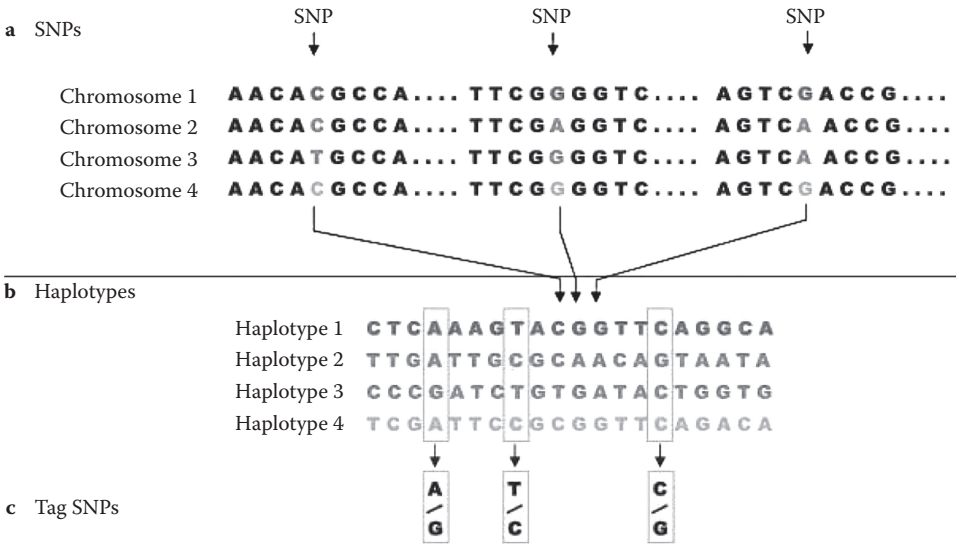


Fig. 1 SNPs, haplotypes and tag SNPs. (a) SNPs. Shown is a short stretch of DNA from four versions of the same chromosome region in different people. Most of the DNA sequence is identical in these chromosomes, but three bases are shown where variation occurs. Each SNP has two possible alleles; the first SNP in panel a has the alleles C and T. (b) Haplotypes. A haplotype is made up of a particular combination of alleles at nearby SNPs. Shown here are the observed genotypes for 20 SNPs that extend across 6,000 bases of DNA. Only the variable bases are shown, including the three SNPs that are shown in panel a. For this region, most of the chromosomes in a population survey turn out to have haplotypes 1–4. (c) Tag SNPs. Genotyping just the three tag SNPs out of the 20 SNPs is sufficient to identify these four haplotypes uniquely. For instance, if a particular chromosome has the pattern A-T-C at these three tag SNPs, this pattern matches the pattern determined for haplotype 1. Note that many chromosomes carry the common haplotypes in the population.

linkage disequilibrium between nearby SNPs in the human genome. Because of the strong associations between SNPs in a region, it is possible to use only a few carefully chosen SNPs to identify each of the common haplotypes in a region (Fig. 1c). Recent molecular surveys suggest that the human genome contains many discrete haplotype blocks that are sites of closely located SNPs.^[48] Each block may have a few common haplotypes that account for a large proportion of chromosomal variation. Between adjacent blocks are there large regions, called hotspots, in which recombination events occur with high frequencies. Several algorithms have been developed to identify a minimal subset of SNPs, i.e., “tagging” SNPs, that can characterize the most common haplotypes. The number and type of tagging SNPs within each haplotype block can be assumed to have determined prior to association studies.

QUANTITATIVE TRAIT NUCLEOTIDES

A haplotype represents a linear arrangement of nucleotides (alleles) at different SNPs on a single chromosome or part of a chromosome. The pair of haplotypes is called a diplotype. The observed phenotype of a diplotype is called a genotype. A diplotype is always constructed by two haplotypes, one from the maternal parent and the other from the paternal parent. Suppose there are two different SNPs on the same genomic region, one with two alleles *A* and *a* and the other with two alleles *B* and *b*, respectively. Allele *A* from SNP 1 and allele *B* from SNP 2 are located on the first homologous chromosome, whereas allele *a* from SNP 1 and allele *b* from SNP 2 located on the second homologous chromosome. Thus, *AB* is one haplotype and *ab* is a second haplotype, and both constitute a diplotype *ABlab*.

In a practical genetic analysis, we can only observe the genotype expressed as *AaBb*. However, the double heterozygote may be one (and only one) of two possible diplotypes *ABlab* and *AblaB*. But these two diplotypes cannot be directly observed and should be inferred from SNP genotype data. In practice, it is important to estimate haplotype effects on a quantitative trait based on the diplotypes and therefore genotypes. For example, if an animal carries haplotype *AB*, it will grow better than other animals that carry any other haplotypes, *Ab*, *aB* and *ab*. For this reason, the same genotype *AaBb* may perform differently, depending on what diplotype it carries. If this genotype is diplotype *ABlab*, then it will have a better growth. If the animal is diplotype *AblaB*, its growth will be poorer. The statistical model being developed will be used to determine which diplotype is associated with better growth in experimental crosses.

Genetic mapping with the diplotypes has power to detect particular DNA sequences that encode a complex trait. The nucleotide components of a haplotype or diplotype that affect a quantitative trait are called quantitative trait nucleotides (QTNs). As mentioned above,

the human genome project generates massive amounts of DNA sequence data across the human genome^[49] and this will facilitate the complete identification of QTNs responsible for complex traits, their number and distribution as well as the number and order of nucleotides that comprise the QTNs. Also, the frequencies of haplotypes for a QTN are also important for the understanding of the population structure and diversity.

As a series of DNA sequences, a QTN that encode a particular trait can be detected with SNPs genotyped from candidate genes or throughout the whole genome. The recent development of the human haplotype map (HapMap) has made it possible to detect genome-wide QTNs that are responsible for complex diseases.^[49] Liu et al.^[50] were among the first who proposed a general statistical model for the characterization of QTN variants that encode a complex phenotype in a natural population. A strategy based on QTL information has also been developed to identify QTNs for complex traits in controlled crosses.^[51]

GENETIC MAPPING BY INCORPORATING MECHANISTIC PATHWAYS

Traditional genetic-mapping approaches are based on a direct relationship between genotype and phenotype through statistics. Functional mapping relates DNA variants with phenotypic traits at different stages of development via mathematical models. All these connections are based on a simple genotype-phenotype correspondence. Given that a black box is not revealed from the genotype to phenotype, they are not able to unravel the mechanistic and regulatory process of how genes control a phenotype. Despite continuous addendums and modifications received in the recent past, the “Central Dogma” of biology, **DNA → mRNA → enzyme (inactive) → enzyme (active) → metabolite(s) → metabolism → cellular physiology → life**, is still thought to be a fundamental rule to form, control and direct every aspect of a living system (Fig. 2). Regulatory control is exerted to affect virtually all these levels.

These pathways and processes from DNA to life can be viewed as a biological system comprised of large populations of elements that function together within the system according to unique design principles. Thus, the life can be defined by the structure and dynamics of the system. These two aspects, structure and dynamics, provide a baseline that can be used to analyze an essential property of the biological system: robustness, i.e., single perturbations will rarely cause systemic failure.^[52,53] In such a biological system, different parts are assembled together, and they interact with each other and with the surrounding environment by two main types of digital information: genes, which encode the proteins through the intermediary of RNA, and regulatory networks, which specify how these genes

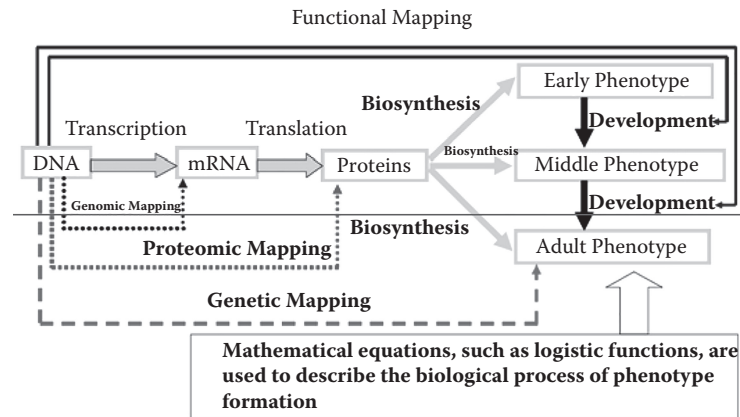


Fig. 2 Biological networks of the phenotypic formation and development of a trait. Genetic mapping is moving from a direct genotype-phenotype correspondence to regulatory networks that underlies the formation of a phenotypic trait

are expressed in time and space. Genetic information can operate across such different time spans: Millions to tens of years (evolution and conservation of species), tens of years to hours (life cycles), and weeks to milliseconds (physiology),^[54] whereas biological information operates on multiple levels of organization (molecular, cellular, tissular, organic, systemic, etc.) in complex networks.^[53]

Current biotechnologies can measure gene expression, a process by which inheritable information from the DNA sequence is converted into a functional gene product (such as RNA or protein). Aimed to gain an insight into the entire genome of an organism, genomics is the study of the whole set of genes of a biological system. Likewise, the object of the study of proteomics is the proteome, understood as the entire collection of proteins that are expressed in a system. In this way, the respective discipline arises from the study of the transcriptome (the set of RNA transcripts of a specific system) and the metabolome (the entire range of metabolites taking part in a biological process). Other omes (sets) that may also be of interest include: the interactome (complete set of interactions between proteins or between these and other molecules), the localizome (localization of transcripts, proteins, etc.), or even the phenome (complete set of phenotypes) of a given organism.^[52,55]

It is recognized that transcriptional, proteomic, and metabolomic expressions can be controlled by specific QTLs. Traditional genetic-mapping approaches can be used to map these QTLs, called eQTLs, pQTLs, and mQTLs, respectively, by viewing transcriptional, proteomic, and metabolomic data as general phenotypes. There is an increasing wealth of literature on the development of statistical models for mapping eQTLs, pQTLs, and mQTLs.^[56] Meanwhile, an increasing wealth of literature evidence has reported on the existence of these different types of QTLs.^[57] It is expected that genetic mapping can be more efficient and more biologically motivated when it is moved from a simple genotype-phenotype correspondence to the underlying regulatory network of trait formation (Fig. 2).

FUTURE DIRECTION: MODELING REGULATORY NETWORK

The information of omes enables scientists to measure multiple features of complex systems at various levels of biological organization from the cell to the whole organism within a defined, fully delineated framework. The challenge is therefore how this information is decoded according to the physiological function of the system studied.^[58]

To measure a system, even at the single-cell level, an understanding of how gene, protein, metabolic, and physiological events are expressed in time course is crucial. Attempts to correlate gene expression data with proteomic and metabolomic data have been made by using classic correlation methods or multivariate statistics. However, they have often proved to be unsatisfactory even for simple systems such as yeasts^[59] in large part because the timescales of various biological control functions are either very different or simply unknown. Fig. 3 shows such a problem in which two pairs of gene-protein couples are expressed in time course. It can be seen that the levels of mRNA transcripts are not consistent with protein levels on timescales.^[58] If the action of the stimulus, such as drug intervention, is at the genetic level, it will take a finite amount of time in a cell for the associated protein synthesis (or post-translational modification) to occur. Also, the duration of the gene events (t_g) and protein events (t_p) may be very different. All this will lead to the consequence that the observed covariance of the gene and protein events is highly dependent on sampling time point and frequency. As shown in Fig. 3, in some instances, a single sampling point (say τ_3) might result in the incorrect assumption that gene 2 co-varied with protein 1. The times of maximal activity or expression for protein may not be constant; for example, the t_p values and related turnover times are larger for protein 2 than protein 1.

A problem also exists when one attempts to correlate proteomic with metabolomic data. Currently, proteomic

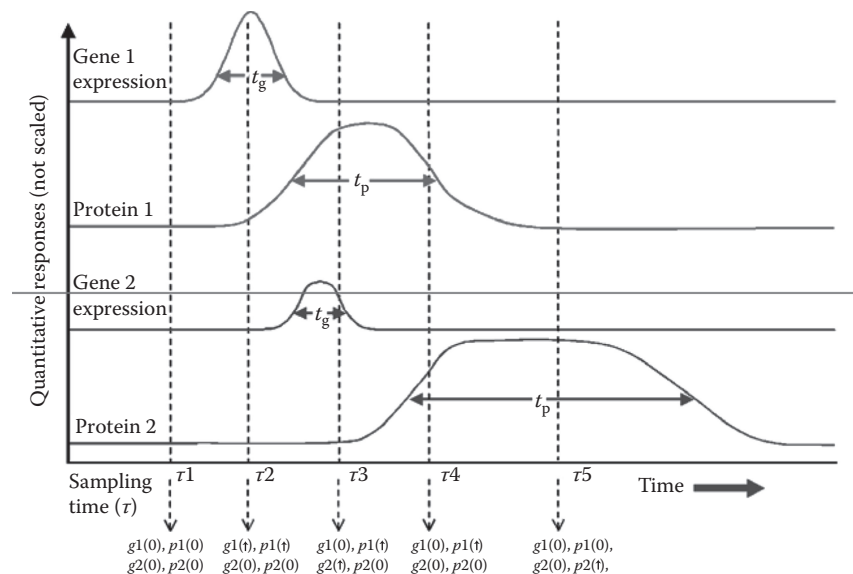


Fig. 3 Time courses for two hypothetical gene–protein couples (relating transcription activity to protein level) that are up-regulated following a system stimulus, such as a drug intervention at time zero. Key: $g1(0)$, $p1(0)$ etc. describe the relative condition with respect to up regulation of each gene and protein at a given time-point (e.g. τ_1 , τ_2 , etc.), where (0) indicates baseline level and ($\uparrow$) indicates upregulation. Thus the post intervention observation of relative state can be seen to be dependent on sampling time-point. (Adapted from Ref. [58].)

data sets contain no information on the activity of specific proteins, which is dependent on their location in the cell and the presence of co-factors or inhibitors.

To understand the true quantitative relationship between the variation in activity of every one of the thousands of gene-protein couples or protein-metabolite couples in a cellular system, it is important to consider several critical factors, that is, the time displacement of the genetic and protein synthetic and post-translational events, their different timescales and their half-lives. Ordinary differential equations (ODE) that model electronic networks in engineering have been proposed to describe regulatory networks of a biological system.^[60–65] These ODE methods were applied to model the regulatory network of *Halobacterium salinarum*, suggesting that the model could predict mRNA levels of 2,000 out of a total 2,400 genes found in the genome (a network with 1,431 predicted regulatory influences controlling 459 biclusters that represent 85% of the genes in the genome).^[66] A similar application was made to model human regulatory networks (mediating TLR-5-mediated stimulation of macrophages) and several other microbial networks.^[63] The accuracy and precision of a regulatory network can be increased when the model is combined with environmental and genetic perturbations by incorporating the principle of functional mapping.

ACKNOWLEDGMENTS

This work is partially supported by a Joint grant DMS/NIGMS-0540745 to RW.

REFERENCES

1. Lander, E.S.; Botstein, D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **1989**, *121*, 185–199.
2. Zeng, Z.B. Precision mapping of quantitative trait loci. *Genetics* **1994**, *136*, 1457–1468.
3. Jansen, R.C.; Stam, P. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **1994**, *136*, 1447–1455.
4. Wu, R.L.; Ma, C.X.; Casella, G. *Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL*; Springer-Verlag: New York, 2007.
5. Xu, S. Mapping quantitative trait loci using four-way crosses. *Genet. Res.* **1996**, *68*, 175–181.
6. Jannink, J.L.; Wu, X.L. Estimating allelic number and identity in state of QTLs in interconnected families. *Genet. Res.* **2003**, *81*, 133–144.
7. Wu, R.L.; Ma, C.X.; Casella, G. Joint linkage and linkage disequilibrium mapping of quantitative trait loci in natural populations. *Genetics* **2002**, *160*, 779–792.
8. Kao, C.H.; Zeng, Z.B.; Teasdale, R.D. Multiple interval mapping for quantitative trait loci. *Genetics* **1999**, *152*, 1203–1216.
9. Jiang, C.; Zeng, Z.B. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* **1995**, *140*, 1111–1127.
10. van Eeuwijk, F.A.; Malosetti, M.; Yin, X.Y.; Struik, P.C.; Stam, P. Statistical models for genotype by environment data: from conventional ANOVA models to eco-physiological QTL models. *Aust. J. Agric. Res.* **2005**, *56*, 883–894.
11. Xu, S.; Atchley, W.R. Mapping quantitative trait loci for complex binary diseases using line crosses. *Genetics* **1996**, *143*, 1417–1424.

12. Li, J.; Wang, S.; Zeng, Z.-B. Multiple interval mapping for ordinal traits. *Genetics* **2006**, *173*, 1649–1663.
13. Satagopan, J.M.; Yandell, B.S.; Newton, M.A.; Osborn, T.C. A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **1996**, *144*, 805–816.
14. Yi, N.; Xu, S.; Allison, D.B. Bayesian model choice and search strategies for mapping interacting quantitative trait loci. *Genetics* **2003**, *165*, 867–883.
15. Churchill, G.A.; Doerge, R.W. Empirical threshold values for quantitative trait mapping. *Genetics* **1994**, *138*, 963–971.
16. Zou, F.; Fine, J.P.; Hu, J.; Lin, D.Y. An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. *Genetics* **2004**, *168*, 2307–2316.
17. Broman, K.W.; Speed, T.P. A model selection approach for the identification of quantitative trait loci in experimental crosses (with discussion). *J. R. Stat. Soc. B* **2002**, *64*, 641–656.
18. Zou, F.; Nie, L.; Wright, F.A.; Sen, P.K. A robust QTL mapping procedure. *J. Stat. Plan. Infer.* **2009**, *139*, 978–989.
19. Chen, Z. The full EM algorithm for the MLEs of QTL effects and positions and their estimated variances in multiple interval mapping. *Biometrics* **2005**, *61*, 474–480.
20. Li, C.B.; Zhou, A.L.; Sang, T. Rice domestication by reducing shattering. *Science* **2006**, *311*, 1936–1939.
21. Anholt, R.R.H.; Mackay, T.F.C. Quantitative genetic analyses of complex behaviours in *Drosophila*. *Nat. Rev. Genet.* **2004**, *5*, 838–849.
22. Peltonen, L.; McKusick, V.A. Dissecting human disease in the postgenomic era. *Science* **2001**, *291*, 1224–1229.
23. Pletcher, S.D.; Geyer, C.J. The genetic analysis of age-dependent traits: Modeling the character process. *Genetics* **1999**, *153*, 825–835.
24. Jaffrezic, F.; Thompson, R.; Hill, W.G. Structured antedependence models for genetic analysis of repeated measures on multiple quantitative traits. *Genet. Res.* **2003**, *82*, 55–65.
25. Kirkpatrick, M.; Heckman, N. A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. *J. Math. Biol.* **1989**, *27*, 429–450.
26. West, G.B.; Brown, J.H.; Enquist, B.J. A general model for the origin of allometric scaling laws in biology. *Science* **1997**, *276*, 122–126.
27. West, G.B.; Brown, J.H.; Enquist, B.J. The fourth dimension of life: Fractal geometry and allometric scaling of organisms. *Science* **1999**, *284*, 1607–1609.
28. Via, S.; Gomulkiewicz, R.; de Jong, G.; Scheiner, S.E.; Schlichting, C.D.; van Tienderen, P. Adaptive phenotypic plasticity: consensus and controversy. *Trend. Ecol. Evol.* **1995**, *10*, 212–217.
29. Ma, C.X.; Casella, G.; Wu, R.L. Functional mapping of quantitative trait loci underlying the character process: A theoretical framework. *Genetics* **2002**, *161*, 1751–1762.
30. Wu, R.L.; Ma, C.X.; Lin, M.; Casella, G. A general framework for analyzing the genetic architecture of developmental characteristics. *Genetics* **2004**, *166*, 1541–1551.
31. Wu, R.L.; Ma, C.X.; Lin, M.; Wang, Z.H.; Casella, G. Functional mapping of growth quantitative trait loci using a transform-both-sides logistic model. *Biometrics* **2004**, *60*, 729–738.
32. Wu, R.L.; Wang, Z.H.; Zhao, W.; Cheverud, J.M. A mechanistic model for genetic machinery of ontogenetic growth. *Genetics* **2004**, *168*, 2383–2394.
33. Wu, R.L.; Lin, M. Functional mapping – how to study the genetic architecture of dynamic complex traits. *Nat. Rev. Genet.* **2006**, *7*, 229–237.
34. Li, Y.; Wu, R.L. Functional mapping of growth and development. *Biol. Rev.* **2009**, Nov 23. [Epub ahead of print].
35. Wang, Z.H.; Wu, R.L. A statistical model for high-resolution mapping of quantitative trait loci determining HIV dynamics. *Stat. Med.* **2004**, *23*, 3033–3051.
36. Wang, Z.H.; Li, Y.; Li, Q.; Wu, R.L. Joint functional mapping of quantitative trait loci for HIV-1 and CD4+ dynamics. *Int. J. Biostat.* **2009**, *5* (1), 1–Article 9.
37. Ho, D.D.; Neumann, A.U.; Perelson, A.S.; Chen, W.; Leonard, J.M.; Markowitz, M. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **1995**, *373*, 123–126.
38. Lin, M.; Aqvilonte, C.; Johnson, J.A.; Wu, R.L. Sequencing drug response with HapMap. *Pharmacogenomics J.* **2005**, *5*, 149–156.
39. Derendorf, H.; Meibohm, B. Modeling of pharmacokinetic/pharmacodynamic (PK/PD) relationships. Concepts and perspectives. *Pharm. Res.* **1999**, *16*, 176–185.
40. Cui, Y.H.; Zhu, J.; Wu, R.L. Functional mapping for genetic control of programmed cell death. *Physiol. Genom.* **2006**, *25*, 458–469.
41. Zhao, W.; Hou, W.; Littell, R.C.; Wu, R.L. Structured antedependence models for functional mapping of multivariate longitudinal quantitative traits. *Stat. Appl. Mol. Genet. Biol.* **2005**, *4*, 1–Article 33.
42. Zhao, W.; Wu, R.L. A wavelet-based functional mapping of longitudinal trajectories. *J. Am. Stat. Assoc.* **2008**, *103*, 714–725.
43. Cooper, R.S.; Psaty, B.M. Genomics and medicine: Distraction, incremental progress, or the dawn of a new age. *Ann. Intern. Med.* **2003**, *138*, 576–580.
44. Camerino, G.; Grzeschik, K.H.; Jaye, M.; De La Salle, H.; Tolstoshev, P.; Lecocq, J.P.; Heilig, R.; Mandel, J.L. Regional localization on the human X chromosome and polymorphism of the coagulation factor IX gene (hemophilia B locus). *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81* (2), 498–502.
45. Tatsuguchi, M.; Furutani, M.; Hinagata, J.; Tanaka, T.; Furutani, Y.; Imamura, S.; Kawana, M.; Masaki, T.; Kasanuki, H.; Sawamura, T.; Matsuoka, R. Oxidized LDL receptor gene (OLR1) is associated with the risk of myocardial infarction. *Biochem. Biophys. Res. Commun.* **2003**, *303*, 247–250.
46. Johansson, I.; Oscarson, M.; Yue, Q.Y.; Bertilsson, L.; Sjoqvist, F.; Ingelman-Sundberg, M. Genetic analysis of the Chinese cytochrome P4502D locus: characterization of variant CYP2D6 genes present in subjects with diminished capacity for debrisoquine hydroxylation. *Mol. Pharm.* **1994**, *46*, 452–459.
47. Collins, F.S.; Brooks, L.D.; Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome. Res.* **1998**, *8*, 1229–1331.
48. Dawson, E.; Abecasis, G.R.; Bumpstead, S.; Chen, Y.; Hunt, S.; Beare, D.M.; Pabial, J.; Dibling, T.; Tinsley, E.; Kirby, S.; Carter, D.; Papaspyridonos, M.; Livingstone, S.; Ganske, R.; Lohmussaar, E.; Zernant, J.; Tonisson, N.;

- Remm, M.; Magi, R.; Puurand, T.; Vilo, J.; Kurg, A.; Rice, K.; Deloukas, P.; Mott, R.; Metspalu, A.; Bentley, D.R.; Cardon, L.R.; Dunham, I. A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **2002**, *418*, 544–548.
49. The International HapMap Consortium. The International HapMap Project. *Nature* **2003**, *426*, 789–794.
 50. Liu, T.; Johnson, J.A.; Casella, G.; Wu, R.L. Sequencing complex diseases with HapMap. *Genetics* **2004**, *168*, 503–511.
 51. Hou, W.; Yap, J.S.; Wu, S.; Liu, T.; Cheverud, J.M.; Wu, R.L. Haplotyping a quantitative trait with a high-density map in experimental crosses. *PLoS. One.* **2007**, *2* (8), e1–e732.
 52. Ideker, T.; Galitski, T.; Hood, L. A new approach to decoding life: Systems biology. *Annu. Rev. Genom. Hum. Genet.* **2001**, *2*, 343–372.
 53. De Hoog, C.L.; Mann, M. Proteomics. *Annu. Rev. Genom. Hum. Genet.* **2004**, *5*, 267–293.
 54. Chong, L.; Ray, L.B. Whole-istic biology. *Science* **2002**, *295*, 1–1661.
 55. Kitano, H. Computational systems biology. *Nature* **2002**, *420*, 206–210.
 56. Perez-Enciso, M.; Quevedo, J.R.; Bahamonde, A. Genetical genomics: use all data. *BMC Genomics* **2007**, *8*, 1–69.
 57. Rockman, M.V.; Kruglyak, L. Genetics of global gene expression. *Nat. Rev. Genet.* **2006**, *7*, 862–872.
 58. Nicholson, J.K.; Holmes, E.; Lindon, J.C.; Wilson, I.D. The challenges of modeling mammalian biocomplexity. *Nat. Biotech.* **2004**, *22*, 1268–1274.
 59. Gygi, S.P.; Rochon, Y.; Franza, B.R.; Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **1999**, *19*, 1720–1730.
 60. Gardner, T.S.; di Bernardo, D.; Lorenz, D.; Collins, J.J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **2003**, *301*, 102–105.
 61. Tegner, J.; Yeung, M.K.; Hasty, J.; Collins, J.J. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 5944–5949.
 62. Yeung, M.K.; Tegner, J.; Collins, J.J. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 6163–6168.
 63. Gilchrist, M.; Thorsson, V.; Li, B.; Rust, A.G.; Korb, M.; Roach, J.C.; Kennedy, K.; Hai, T.; Bolouri, H.; Aderem, A. Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. *Nature* **2006**, *441*, 73–178.
 64. Bonneau, R.; Reiss, D.J.; Shannon, P.; Facciotti, M.; Hood, L.; Baliga, N.S.; Thorsson, V. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome. Biol.* **2006**, *7*, R1–R36.
 65. Bonneau, R.; Facciotti, M.T.; Reiss, D.J.; Schmid, A.K.; Pan, M.; Kaur, A.; Thorsson, V.; Shannon, P.; Johnson, M.H.; Bare, J.C.; Longabaugh, W.; Vuthoori, M.; Whitehead, K.; Madar, A.; Suzuki, L.; Mori, T.; Chang, D.E.; Diruggiero, J.; Johnson, C.H.; Hood, L.; Baliga, N.S. A predictive model for transcriptional control of physiology in a free living cell. *Cell* **2007**, *131*, 1354–1365.
 66. Bonneau, R. Learning biological networks: from modules to dynamics. *Nat. Chem. Biol.* **2008**, *4*, 658–664.

Statistical Principles for Clinical Trials

Allan Phillips

Wythe Research, Berkshire, U.K.

Abstract

This entry will focus on the key statistical principles of concern to regulators as identified in the most recent guidelines—namely, bias and robustness.

INTRODUCTION

The application of biostatistics in the drug development process is now recognized worldwide as being essential. Over the last 10–20 years, the increasing importance of biostatistical methodology has led to a rapid increase in the biostatistical resources involved in drug development, especially in the pharmaceutical industry in Europe, the United States, Japan, and elsewhere.

Before regulatory agencies approve a new drug, they require substantial evidence of effectiveness and safety through clinical trials. However, before it can be tested in humans, a number of animal studies need to be conducted to assess the pharmacological characteristics, efficacy, and safety of the drug in animals. To assist researchers in fulfilling the requirements, guidelines have been developed for the various stages of the development process, including Good Laboratory Practice (GLP), Good Clinical Practice (GCP), and Good Regulatory Practice (GRP). As discussed by Chow,^[1] adherence to these guidelines relies heavily upon the implementation of Good Statistical Practice (GSP). In essence, this demands that:

- a. The scientific or medical questions under investigation are addressed using a valid design.
- b. Representative samples are drawn at random for testing.
- c. Statistical inferences are based on appropriate statistical methods and provide an unbiased assessment.

One of the key stages of the drug development process is the clinical development phase, where the goal is to establish evidence of effectiveness in humans through clinical trials. Statistical principles are now commonly used in such trials to assist researchers in providing a more accurate and reliable assessment of any treatment effect.

KEY STATISTICAL PRINCIPLES

Most of the key statistical principles for clinical trials relate to minimizing bias and to ensuring the robustness of conclusions drawn from the data. As defined in the

International Conference on Harmonization (ICH) guideline E9 on “Statistical Principles for Clinical Trials,”^[2] bias is the systematic tendency of any factor associated with the design, conduct, analysis, and interpretation of the results of clinical trials to make the estimate of a treatment effect deviate from its true value. In clinical trials, there are usually two types of bias that are of concern—statistical bias and operational bias. Bias introduced during the conduct of a clinical trial is often referred to as “operational” bias. The other sources of bias are referred to as “statistical” bias. Because bias cannot be measured or even quantified, it is often impossible to adjust for it in the statistical analysis. Thus it is imperative that when designing a clinical trial, all potential sources of bias are identified and, wherever practical, eliminated. The presence of bias can seriously compromise the credibility of a clinical trial, including any inference that is made from a trial.

In many statistical textbooks, robustness relates to the vulnerability of a statistic to the presence of outliers and dirty data. A robust statistic is one that works well for a wide variety of population types from which the samples are drawn. In the clinical trial context, however, robustness relates to the reliability of the inference drawn from the clinical trial data. Because it is unlikely that all sources of bias can be identified and eliminated, it is important to show that the results are unaffected by any limitations of the data, assumptions, and analytical techniques used. Conclusions of a clinical trial that are not substantially affected when alternative approaches are employed are said to be robust.

Two other statistical principles for clinical trials, although less well recognized in the published literature, are measurement error and efficiency. Measurement error relates to the need for any measurement taken on a patient to be precise and reproducible so that observer differences do not bias treatment comparisons. Efficiency relates to the desire to not only provide an unbiased estimate of a treatment effect, but to use the estimator that has minimum error associated with it.

After a brief look at the regulatory guidelines relating to statistical principles for clinical trials, this entry will focus

on the key statistical principles of concern to regulators as identified in the most recent guidelines—namely, bias and robustness. Potential sources of bias in study design and analysis will be reviewed. Advice will be provided on how to minimize the impact of any bias on inferences to be drawn from clinical trials, especially during study design. Because bias can occur in subtle and unknown ways, it can never be totally eliminated. Thus the need to investigate the robustness of principal inferences will also be considered.

REGULATORY REQUIREMENTS

In 1988, the U.S. Food and Drug Administration (FDA) issued a guideline entitled “Guideline for the Format and Content of the Clinical and Statistical Sections of New Drug Applications.”^[3] The purpose of the guideline was to provide direction to sponsors in the design, conduct, analysis, and evaluation of clinical trials. The guideline proved historic in that it was the first regulatory document to describe some of the fundamental statistical principles for clinical trials.

Since 1988, other regulatory authorities have issued guidelines partially or entirely concerned with statistical principles for clinical trials. The most recent and most important biostatistical regulatory guideline is the ICH guideline E9 on “Statistical Principles for Clinical Trials.” An ICH expert working group developed the guideline. The group utilized the Committee for Proprietary Medicinal Products (CPMP) biostatistical guideline^[4] as a starting point, but was also influenced by guidelines from the U.S. Food and Drug Administration^[3] and the Japanese Ministry of Health and Welfare^[5] guideline. The guideline was adopted by the CPMP at its March 1998 meeting and is therefore now operational in Europe. It has also been adopted in United States and Japan. The importance of this guideline is best illustrated by the number of times it has been referenced in other articles within this encyclopedia.

More detailed guidance on the statistical principles relating to selected topics discussed in ICH E9 “Statistical Principles for Clinical Trials” has been issued in recent years by the CPMP in the form of “Points to Consider” documents. They include, for example, Switching Between Superiority and Non-Inferiority,^[6] Meta Analysis and One Pivotal Study,^[7] and Missing Data.^[8] Others are planned to be issued in the future.

STATISTICAL DESIGN

The main source of bias in study design usually arises from the allocation of treatments to patients. However, a variety of statistical techniques such as blinding, randomization, stratification, and minimization can often minimize bias.

Blinding

Pocock^[9] points out that in any clinical trial, the comparison of treatments may be distorted if the patient and those responsible for treatment and evaluation know which treatment is being used. Consequently, to limit the occurrence of conscious and unconscious bias arising from knowing which patients are assigned to which treatment, wherever possible, patients, clinicians, and evaluators should be blinded to treatment assignments. When this is achieved, the trial is known as a double-blind clinical trial.

Blinding also helps to ensure that investigators do not influence the randomization of patients to treatments. Doing so has the clear potential for introducing bias and clearly violates the statistical benefits of randomization, which is discussed next.

Although double-blind trials are the optimal design, they are usually only feasible when comparing treatments of a similar nature. The suitable use of matching placebos in some trials is helpful to make treatments indistinguishable. However, in some clinical trials, it is not possible to have identical treatments (e.g., consider an oncology trial to compare chemotherapy vs. surgery). Consequently, other types of design, such as open-label trials and single blind trials, are employed. Further details can be found in ICH E9 or in other entries in this encyclopedia. In all cases, the essential aim of “blinding” is to prevent the identification of treatments until such opportunities for bias have passed.

Randomization

Randomization helps to avoid possible bias in the selection and the allocation of subjects arising from the predictability of treatment assignments. It introduces a deliberate element of chance into the assignment process. In the simplest of trials, a statistician prepares a randomization list. The list is used to make up identical packages containing the different treatments to be studied in the clinical trial. Patients are then sequentially assigned to the next unidentifiable package on the randomization list. This later part of the process is either manually performed by the investigator at the study site, or, more commonly now, via telecommunication systems or computer software packages.

To improve the comparability of the treatment groups, patients are sometimes randomized in blocks or strata. This is especially helpful when the patient characteristics change over time (e.g., as a result of changes in recruitment policy). Another example where it is helpful to block is when environmental conditions may vary over time such as in an allergy study over several seasons. Most multicenter trials have a separate block or stratum for each center. Stratification by important prognostic factors measured at baseline (e.g., age, sex) may also be valuable at times in order to help achieve balanced treatment groups.

An alternative technique used in clinical trials to minimize bias in the allocation of treatments to patients is minimization. The technique involves the dynamic allocation of patients to treatments. A patient is allocated to a treatment based on the demographic characteristics of the given patient and of those already assigned to treatment. In a stratified trial, the allocation balances within the stratum to which the patient belongs.

Further details on randomization, stratification, and minimization to avoid selection and allocation bias can be found in ICH E9 or in other entries in this encyclopedia.

STATISTICAL ANALYSIS

There are many potential sources of bias that can occur while analyzing clinical trial data. Some of the more important and frequently occurring sources include the following:

1. Exclusion of patients from the analysis based upon the knowledge of the treatment that they received and their outcome.
2. Missing data.
3. Outliers.
4. Improperly conducted interim analyses.

Although statistical methods that minimize the potential impact of any bias are available, in reality, bias will always occur to some extent. Subsequently, it is important that inferences drawn from a clinical trial are robust.

Analysis Sets

The ICH E9 guideline makes it clear that regulatory statisticians are concerned about bias introduced as a result of excluding patients from analyses based upon the knowledge of the treatment that patients received and their outcome. To overcome this, in general, it is advisable to demonstrate a robustness of the principal clinical trial results to alternative choices of the set of patients analyzed, defined prior to unblinding a study.

Depending on the disease and context, it is valid to omit patients from an analysis, especially when the approach is specified in advance in the protocol. A number of reasons used include the following:

- a. Failure to take even one dose of medication.
- b. Absence of disease (e.g., oncology–nonmeasurable lesion, antibiotics–no bacteria detected).
- c. Fraudulent data from a center.
- d. Center measuring primary outcome incorrectly.

If patients are to be omitted from an analysis, then it is important to clearly document the reasons why. This is needed for regulators and others to be able to explore any

potential bias arising from the exclusion of the patients. If the number of excluded patients is substantial, then a sensitivity analysis should always be conducted. This is achieved via the use of different analysis populations such as the full analysis set, per protocol analysis set, etc. Further details are provided in the ICH E9 guideline.

As discussed earlier, randomization helps to avoid possible bias in the selection and the allocation of subjects arising from the predictability of treatment assignments. In some trials, patients can fail to take the randomized treatment but receive one of the other clinical trial treatments, which is another potential source of bias. As pointed out in Phillips et al.,^[10] in nearly all clinical trials, safety data should be analyzed according to the treatment actually taken. For efficacy analyses, however, the approach probably depends on how and why the subjects were misallocated. In situations where the misallocation may cause bias, the analysis should be conducted according to the original randomization (e.g., in open trials where patients may be allocated to surgical intervention).

Missing Data

A potential source of bias when analyzing clinical trial data is the missing data. Every effort needs to be directed toward ensuring that patients comply with the protocol concerning the collection and the management of data to minimize the problem. However, in reality, missing data are likely to always occur. This can be for a variety of reasons including, for example, patient refusal to participate in a clinical trial (e.g., withdrawal of consent, etc.), patient experiencing a treatment-related event (e.g., a treatment success, failure, or an adverse reaction), patient failure to attend a visit (e.g., missed appointment on vacation), or one of many other reasons.

As stated in the Committee for Proprietary Medicinal Products Points to Consider Document on Missing Data,^[8] bias arising from missing data may affect the estimate of a treatment effect, the comparability of the treatment groups, and the representativeness of the clinical trial sample in relation to the target population. The extent of the bias depends on many factors (e.g., the risk of bias in the estimation of a treatment effect depends upon the relationship among the true value of the missing data, the treatment, and the outcome). It is likely that an estimate of a treatment effect will be biased when a poor outcome is more likely to be missing in one treatment group because it is not as effective.

No universally accepted methods exist for handling missing data. In fact, a number of methods employed by statisticians to address the issue of bias arising from missing data in themselves provide a source of bias. For example, a technique frequently used in the pharmaceutical industry is to carry forward the last value observed for a patient for all other assessments if a patient discontinues a clinical trial. This, in itself, may introduce bias in that a

treatment effect is likely to be overestimated if the missing value relates to the treatment and the true value of the missing data (e.g., one treatment does not provide as rapid a response as another and a poor outcome is more likely to be missing). Subsequently, it is imperative that an investigation of the robustness of the principal inferences to be drawn from a clinical trial is explored using different data assumptions concerning any missing data.

Outliers

One of the main statistical principles for clinical trials relates to ensuring that inferences drawn from clinical trial data are robust. That is, there is a need to establish that results are unaffected by any limitations of the data, assumptions, and analytical techniques used. Subsequently, the influence of outliers or extreme values on inferences needs to be carefully investigated. As discussed in ICH E9, procedures for identifying both statistical and medical outliers, and how they will be handled in the statistical analysis, need to be defined prior to unblinding a clinical trial. The procedures should be such as not to favor one treatment or another; otherwise, the estimated treatment effect would be biased. As pointed out by Gardner and Altman,^[11] because outliers often have a pronounced effect, it is common to analyze the data both with and without such observations to assess how much any conclusion depends on these values.

Interim Analyses

An interim analysis of accumulating data in a clinical trial, either formally or informally, is an established practice for ethical and scientific reasons. However, a major concern to regulators and others involved in the design and conduct of clinical trials is the introduction of bias following an interim analysis. Any interim analysis that is not planned appropriately (with or without the consequence of stopping the trial early) may flaw the results of a trial and possibly weaken confidence in the conclusions drawn.

Broadly speaking, methodology exists to address statistical bias arising from interim analyses. Significance tests are usually corrected properly for the interim looks; however, treatment effect estimates are not (e.g., see Facey and Whitehead^[12]). End points correlated with the interim analysis end point are also not adjusted.

To minimize operational bias, Data Monitoring Committees are often used to assess the progress of a clinical trial, safety data, and critical efficacy data at specific time points, particularly multinational, multicenter trials. As discussed by Armitage,^[13] Data Monitoring Committees come in many forms. There is no standard formula suitable for all clinical trials. Each Data Monitoring Committee has to work out its own *modus operandi*. Nevertheless, the practices and experiences of various committees have been documented and provide useful guidance.

For some clinical trials, Data Monitoring Committees comprise solely of representatives from the sponsor organizations (i.e., people with a vested interest in the results of the trial). Better control may be demonstrated when an external monitoring committee is used. By establishing an independent external group, it is possible to protect the integrity of the trial from adverse impact resulting from access to trial information. However, whether an internal or external monitoring board is used, documentation is essential and needs to be available to regulators (roles and responsibilities, minutes, etc.).

The PMA Biostatistics and Medical Ad Hoc Committee has published a more comprehensive discussion on interim analyses in "Interim Analysis in Pharmaceutical Industry."^[14]

CONCLUSION

The application of biostatistics in clinical trial design and analysis is now recognized worldwide as being essential. Many of the key statistical principles attributed to statisticians in clinical trial design and analysis relate to minimizing bias and to ensuring robustness of conclusions drawn from the data. Because bias cannot be measured or even quantified, it is imperative that when designing a clinical trial, all potential sources are identified and, wherever practical, eliminated. Failure to do this can seriously affect the credibility of a clinical trial including any inferences that are drawn from the data. In addition, it is important that inferences drawn from clinical trials are shown to be robust. A more detailed discussion of the topics covered in this entry and other related issues can be found in the ICH E9 guideline entitled "Statistical Principles for Clinical Trials."

ACKNOWLEDGMENTS

The author wishes to thank Mr. V. Haudiquet, Dr. R. Starbuck, and Mr. P. Worthington for reading the manuscript and for providing useful suggestions.

REFERENCES

1. Chow, S. Good statistical practice in the drug development and regulatory approval process. *Drug Inf. J.* **1997**, *31*, 1157–1166.
2. Statistical Principles for Clinical Trials. In *International Conference on Harmonization Guideline. Statistical Principles for Clinical Trials*; February 5, 1998.
3. Food and Drug Administration. *Guideline for the Format and Content of the Clinical and Statistical Sections of New Drug Applications*; FDA, U.S. Department of Health and Human Services: Rockville, MD, 1988.
4. Lewis, J.A.; Jones, D.R.; Rohmel, J. Biostatistical methodology in clinical trials—A European guideline. *Stat. Med.* **1995**, *14*, 1655–1682.

5. Ministry of Health and Welfare. *Guideline for the Statistical Analysis of Clinical Trials*; MHW Pharmaceutical Affairs Bureau: Tokyo, Japan, 1992 (in Japanese).
6. *Committee for Proprietary Medicinal Products Points to Consider Document on Switching Between Superiority and Non-Inferiority*; www.eudra.org. (Adopted July 2000).
7. *Committee for Proprietary Medicinal Products Points to Consider Document on Meta Analysis and One Pivotal Study*; www.eudra.org (Adopted May 2001).
8. *Committee for Proprietary Medicinal Products Points to Consider Document on Missing Data*; 2001. Consultation draft.
9. Pocock, S.J. *Clinical Trials: A Practical Approach*; Wiley: Chichester, 1983.
10. Phillips, A.; Ebbutt, A.; France, L.; Morgan, D. The International Conference on Harmonization Guideline “Statistical Principles for Clinical Trials”: Issues in applying the guideline in practice. *Drug Inf. J.* **2000**, *34*, 337–348.
11. Gardner, M.J.; Altman, D.G. *Statistics with confidence*. *Br. Med. J. (London)* **1989**.
12. Facey, K.M.; Whitehead, J. An improved approximation for calculation of confidence intervals after a sequential clinical trials. *Stat. Med.* **1990**, *9*, 1277–1285.
13. Armitage, P. Interim analysis in clinical trials. *Stat. Med.* **1991**, *10*, 925–937.
14. PMA Biostatistics and Medical Ad Hoc Committee on Interim Analysis. Interim analysis in the pharmaceutical industry. *Control. Clin. Trials* **1993**, *14*, 160–173.

Statistical Process Control

William R. Myers and William A. Brenneman

The Procter & Gamble Co., Mason, Ohio, U.S.A.

Abstract

Statistical process control (SPC) is a collection of tools and techniques used to monitor, to control, and to improve a process. This entry discusses the details of statistical process control, including several examples of process flowcharts, control charts, and cause-and-effect diagrams to illustrate this concept.

Keywords: Charts; Histograms; Statistical process control.

INTRODUCTION

Statistical process control (SPC) is a collection of tools and techniques used to monitor, to control, and to improve a process. The primary goal of SPC is to achieve process stability and to improve process capability by reducing variability. A process is defined as being in *statistical control* when only a certain amount of natural or routine variability exists. This variability is intrinsic to the actual process and cannot be fully eliminated. On the other hand, a process may be exposed to sources of variation that are beyond the natural variability of the process. These sources of variation are defined as *assignable or special causes*. A process where assignable causes exist is defined as being out of control (or not in statistical control). A process that is not in a state of statistical control is one in which the variable that is being measured does not have a stable distribution. A few examples of assignable causes are nonconforming raw material, operator error, analytical error, and machine malfunction. A primary goal of SPC is to detect, as swiftly as possible, any assignable causes so that a proper investigation can take place in order to fix the problem. The ultimate objective of SPC is the elimination of variation in the process as a result of assignable or special causes.

Quality control and, more specifically, SPC have become a critical component within the pharmaceutical industry. Both batch release criteria—such as assay content uniformity and dissolution, as well as process control testing like tablet weight, hardness, and gauge—can be effectively monitored through SPC. The need for efficient and stable processes with improved capability within a highly regulated environment has made SPC an essential component within a pharmaceutical company's quality organization. Some results of poor process quality could include an excessive scrapping of entire lots and unnecessary costs and time to perform rework on poor quality product.

OVERVIEW

The potential success of a complete SPC program requires management involvement and the commitment to the quality improvement process throughout the organization. It must be clearly pointed out that SPC should not only be applied when trouble occurs and subsequently abandoned, but should be applied throughout the life of the process. Quality improvement should become a part of the culture and should be a proactive tool rather than a reactive tool.

The most important and broadly used tools in SPC are

1. Histograms.
2. Pareto charts.
3. Cause-and-effect diagram.
4. Flowchart.
5. Control chart.

Histograms

A histogram is a graphical representation of the frequency distribution of a continuous variable. Fig. 1 illustrates a histogram of tablet weight data. Histograms allow one to more easily observe the shape, location, and spread of the data. Histograms will point out skewness and heaviness in the tails, and may also give an indication of outliers. From Fig. 1, it is clear that the tablet weight has an approximate normal distribution, with an approximate location (mean) of 10 and a range of 1.5. The following are some general guidelines for constructing histograms:

- Select the number of bins (or groups) approximately equal to the square root of the sample size.
- The width of the bins should be uniform.
- Start the lower limit for the first bin just below the smallest data point.

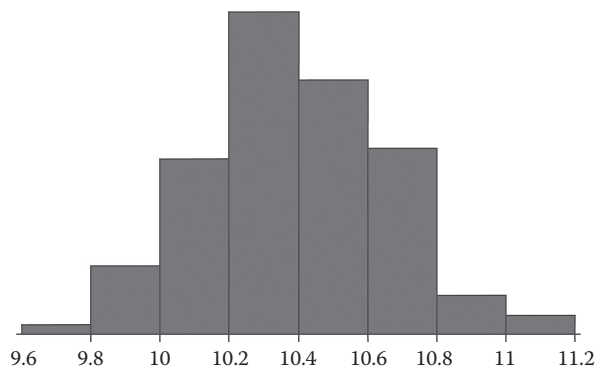


Fig. 1 A histogram of tablet weight (mg)

Histograms can also be used to evaluate process capability, by plotting the histogram along with the specification limits in order to show the portion of data that exceeds the specification limits. Furthermore, superimposing the best-fitting normal curve on the histogram provides a graphical way to check for normality and, if specification limits are plotted, will also graphically show the theoretical out-of-specification probability. Montgomery^[1] states that at least 100 or more observations should be used in order for the histogram to be moderately stable.

Pareto Charts

The Pareto chart (or diagram) is a frequency distribution of attribute data arranged by category (i.e., bar chart). A Pareto chart can determine those problems that occur most frequently and thus allows the efforts and resources to be focused on those particular concerns. However, the chart does not directly differentiate the severity of the problems. Fig. 2 is an example of a Pareto chart based on hypothetical data. In the example, 60% of problems found were related to cracked tablets; 20% were related to broken tablets; 10% were related to non-uniform tablet size, shape,

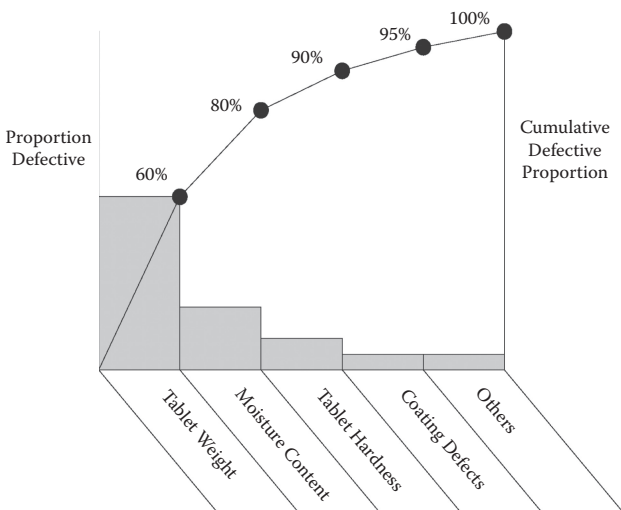


Fig. 2 An example of a Pareto chart

or color; 5% were related to uneven tablet coating; and 5% were related to other factors. Before starting to work on the cracked tablets, an analysis would need to be done to see whether fixing the cracked tablets would be more beneficial to the business than fixing the broken tablet problem. See Gopal^[2] for another example of a Pareto chart.

Cause-and-Effect Diagram

The cause-and-effect (or fishbone) diagram is a powerful tool used to determine the potential causes of undesirable (or desirable) results. After a problem has been identified, a team should be created to brainstorm on major potential cause categories of the undesirable effect. Some examples of major potential cause categories are raw materials, machines, analytical methods, and personnel. Next, possible causes within each major potential cause categories are identified and rank-ordered in terms of most likely to cause the undesirable result. A template for a cause-and-effect diagram is provided in Fig. 3. One critical outcome of such an exercise is to identify all possible causes of the problem and thus remove the personal bias individual team members have to dismiss causes they believe to be implausible. Another outcome is that the cause-and-effect diagram can be used by the team to monitor their progress in solving the problem by adding and by deleting possible causes along the way.

Process Flowchart

A process flowchart is a chronological sequence of process steps. They are very useful in visualizing and defining the process in order to better understand and to improve the process. Process flowcharts should include enough detail to be used as an effective resource when a problem occurs, or when an improvement effort begins but should not be so detailed as to cloud the overall picture. Process flowcharts are invaluable when new members are added to a team and need to learn quickly the main features of the manufacturing process. Process flowcharts can also aid in the brainstorming sessions that lead to the construction of

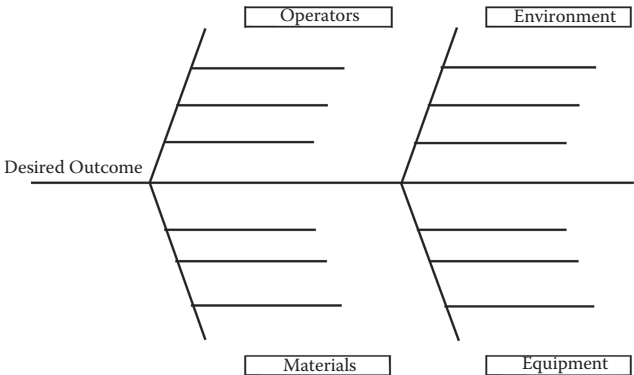


Fig. 3 A template of a cause-and-effect diagram

cause-and-effect diagrams. An example of a basic flowchart is given in Fig. 4, where the potential main components of a pharmaceutical process are graphed.

Control Charts

The control chart is the most widely used tool within the SPC tool bag. A control chart is used to detect the out-of-control state of a process. As previously stated, it is important that the process shift be detected quickly so that the problem can be corrected before many non-conforming (or out of specification) products are produced. The proper use of a control chart can often prevent an overreaction to changes that represent only random fluctuation. An example of a control chart ($\bar{X}$ chart) is presented in Fig. 5, where each point on the plot represents a sample average of a quality characteristic (tablet weight) taken over time. The sample values are divided into subgroups, which are often called rational subgroups. The rational subgroups are chosen so that the within-subgroup variability is small and the between-subgroup variability is maximized so that if assignable causes are present, they can be detected with high probability. The time order of production is a natural way to select subgroups. The control chart has a centerline that corresponds to the grand average of the quality measurement. The other two lines represent the lower control limit (LCL) and the upper control limit (UCL), which are constructed in such a way that if a sample point falls outside the limits, it is evidence that the process is out of statistical control.

In most cases, 3σ control limits are used in order to maintain a small probability of a sample point falling outside the control limits given that the process is in control (i.e., *false alarm* rate). If the variable is normally distributed and 3σ control limits are used, the probability is 0.0027 that the sample point (e.g., mean for the $\bar{X}$ chart) falls outside the control limits, given that the process is in control. It has become general practice to use 3σ control limits, even if the distribution of the quality characteristic is not normally distributed. One common misconception is that data must be normally distributed in order to be placed on a control chart. However, Burr^[3] and Schilling and Nelson^[4] have demonstrated that a control chart, with 3σ control limits, is reasonably robust to the normality assumption. With respect to an $\bar{X}$ chart, even if the individual values are not normally distributed, the distribution of $\bar{X}$ is approximately normal and thus there is fairly accurate control over the probability of a type I error for the case in which σ is known. It has been demonstrated over the many years of using the control chart that the 3σ control limits do very well in practice.



Fig. 4 The main components of a pharmaceutical flow chart

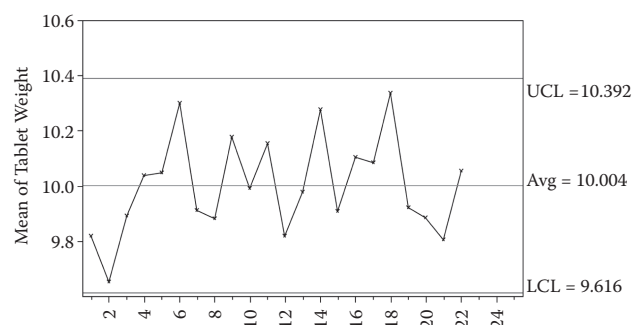


Fig. 5 An example of an $\bar{X}$ chart for tablet weight

Preliminary or trial control limits must be determined in order to evaluate the future output of a process. The control limits should be based on data that are in statistical control. The standard practice is to use at least 20 subgroups in order to compute the preliminary control limits. The sample size within each sample is typically small (e.g., three to eight). If the sample statistic (e.g., $\bar{X}$) for each sample is inside the control limits and no apparent systematic pattern exists, then it can be assumed that the process is in control and the control limits can be used to evaluate future output from the process. However, if one or more sample statistic is outside the trial control limits, then they must be investigated for an assignable cause. If an assignable cause is established for those points outside the control limits, then they should be removed and new control limits should be computed. There are two options if an assignable cause cannot be attributed to those points plotted beyond the control limits. One option is to keep the out-of-control point(s) and continue on with the preliminary control limits. The second option is to delete the out-of-control point and recompute the control limits. If only one or two points are beyond the control limits, these two options will often produce similar results. It is standard practice to routinely revise the control limits. If there is a change in the process, then the control limits should definitely be revised. For example, if there were improvements made in the process that reduced variation, then the original control limits would be too wide and would not readily detect a shift in either the mean or the variability.

Many drug characteristics, such as potency, tablet weight, content uniformity, and dissolution, are measured on a continuous scale. The $\bar{X}$ and R chart are used to monitor the mean and the variability of a continuous variable. Fig. 6 provides an example of a $\bar{X}$ chart for dissolution. The formulas for constructing the control limits on the $\bar{X}$ chart are as follows:

$$\text{Upper Control Limit} = \bar{\bar{X}} + \frac{3}{d_2\sqrt{n}} \bar{R}$$

$$\text{Center Line} = \bar{\bar{X}}$$

$$\text{Lower Control Limit} = \bar{\bar{X}} - \frac{3}{d_2\sqrt{n}} \bar{R}$$

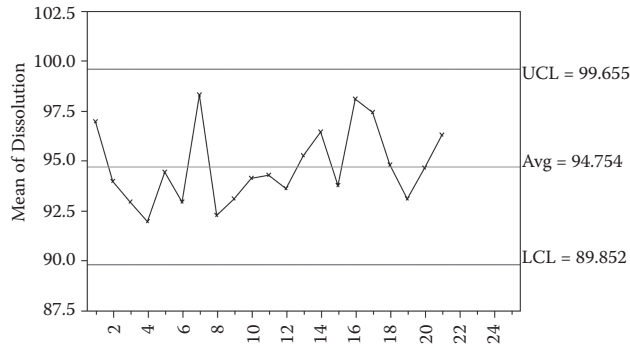


Fig. 6 An example of an $\bar{X}$ chart for dissolution

where $\bar{\bar{X}}$ represents the overall grand mean, d_2 is the constant depending on the sample size, n is the subgroup size, and $\bar{R}$ is the average within subgroup range.

An R chart can be used to monitor and to control process variability. R charts are constructed by plotting the sample ranges over time. Fig. 7 provides an R chart for the corresponding $\bar{X}$ chart mentioned above. The formulas for the R chart are as follows:

$$\begin{aligned}\text{Upper Control Limit} &= \bar{R}D_4 \\ \text{Center Line} &= \bar{R} \\ \text{Lower Control Limit} &= \bar{R}D_3\end{aligned}$$

where D_4 and D_3 are constants depending on the sample size. There are cases where the standard deviation can be used directly rather than using the range. If the sample size within a subgroup is relatively large (greater than 10), then the S chart is often recommended. The formulas for the S chart are as follows:

$$\begin{aligned}\text{Upper Control Limit} &= B_4\bar{S} \\ \text{Center Line} &= \bar{S} \\ \text{Lower Control Limit} &= B_3\bar{S}\end{aligned}$$

where B_3 and B_4 are constants depending on the sample size. The constants that are used to compute all the previously discussed control limits can be found in Montgomery,^[1] Ryan,^[5] or Wheeler.^[6]

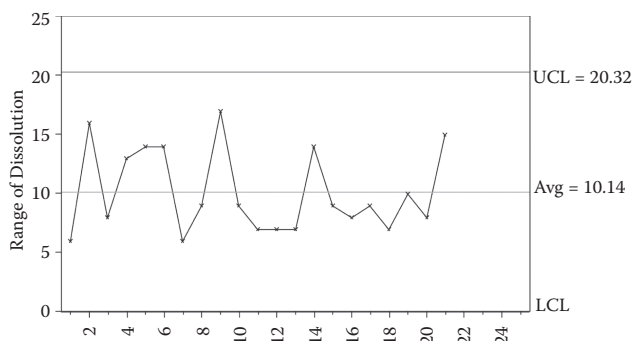


Fig. 7 An example of an R chart for dissolution

Two aspects of control chart design to be considered are sample size and frequency of sampling. The choice of sample size and frequency of sampling will certainly depend on the amount of available resources. Given that there is a certain level of resources available for sampling/inspection, it must be decided how the sampling effort is allocated. Should one sample more frequently with a smaller sample size or sample less frequently with a larger sample size? The most accepted practice is to sample more frequently with a smaller sample size.

One way to formally determine the appropriate sample size and sampling frequency is via the *average run length* (ARL). The ARL is the average number of samples that must be plotted on the control chart before an out-of-control point is found after the process mean has undergone a meaningful shift. The ARL is equal to $1/p$, where p is the probability that the point is beyond the control limits, when the process mean has shifted a certain distance from the original mean. The value for p can be determined by an operating characteristic curve. Montgomery^[1] and Walpole et al.^[7] provide a nice discussion on operating characteristic curves. A performance measure that is often used to determine the sample size and sampling frequency is the *average time to signal* (ATS). The ATS, as the name indicates, is defined as the average time it will take for the control chart to detect a signal (out-of-control condition) when a given shift in the mean has occurred. Assume it is determined that, based on available resources, one can sample and test 10 items per hour. The question arises whether it would be more beneficial to sample five items every 30 minute or 10 items every hour. If the user is interested in detecting a 1.5σ shift in the mean, having a sample size of 10 per hour would produce an ATS of 1 hour, while a sample size of five per half hour would produce an ATS of approximately 45 minute. Operating characteristic curves can also be computed for the R chart, in order to detect a shift in variation. Montgomery^[1] and Walpole et al.^[7] provide a brief discussion on the operating characteristic curve for the R chart.

There are circumstances where no rational subgroup is available and the sample size used for process monitoring is $n = 1$ (e.g., composite assay on a given batch for product release). An individual (X) and moving range (MR) chart can be used when a subgroup of one is required. Because only one measurement is available at each sample, a moving range of two successive observations is used to estimate the process variability. The moving range is defined as $MR = |x_i - x_{i-1}|$. The formulas for constructing the control limits for the individuals chart are as follows:

$$\begin{aligned}\text{Upper Control Limit} &= \bar{\bar{X}} + 3 \frac{\text{mean}(\text{MR})}{d_2} \\ \text{Center Line} &= \bar{\bar{X}} \\ \text{Lower Control Limit} &= \bar{\bar{X}} - 3 \frac{\text{mean}(\text{MR})}{d_2}\end{aligned}$$

The control limits for the MR chart are analogous to the R chart, with the mean of the moving ranges used in place of the mean of the ranges. However, the MR chart does not allow one to truly evaluate the process variation. An example of an individual (X) and moving range chart for composite assay is provided in Fig. 8.

There are many quality measurements (e.g., cracked tablet, defective coating, split capsule, etc.) that are not numerical in nature, but rather are defined as an attribute. Attribute characteristics are defined as being defective or non defective. There are several standard control charts that handle attribute data. The first and most often used is a p chart, which is based on the binomial distribution, where p is the probability that any unit will not conform within the specification limits. The formulas for constructing the control limits for the p chart are as follows:

$$\begin{aligned}\text{Upper Control Limit} &= \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \\ \text{Center Line} &= \bar{p} \\ \text{Lower Control Limit} &= \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\end{aligned}$$

Fig. 9 provides an example of a p chart for coating defects. In the case where the defect level has a very low incidence rate ($p = 0.001$ or 1000 ppm), the p chart will often be of little use because most of the samples will have zero defects. Montgomery^[1] proposes to use the time between successive defects as the response. If the number of defects follows a Poisson distribution, then the time between defects follows an exponential distribution. An

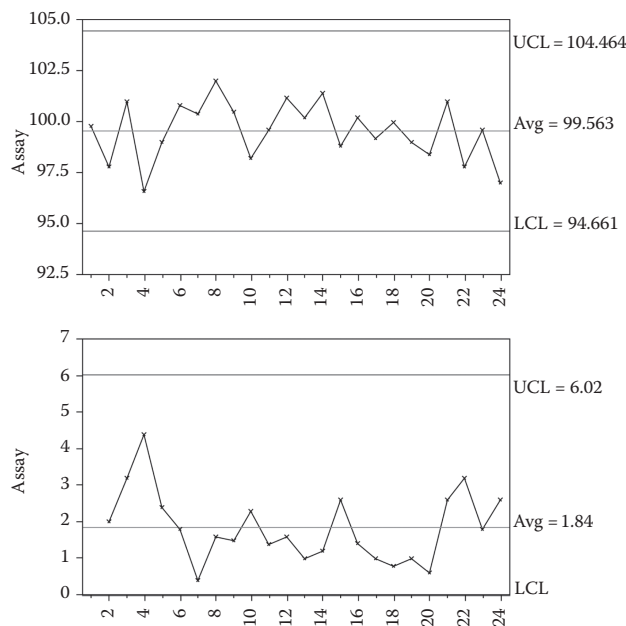


Fig. 8 An example of individual and moving range chart for composite assay

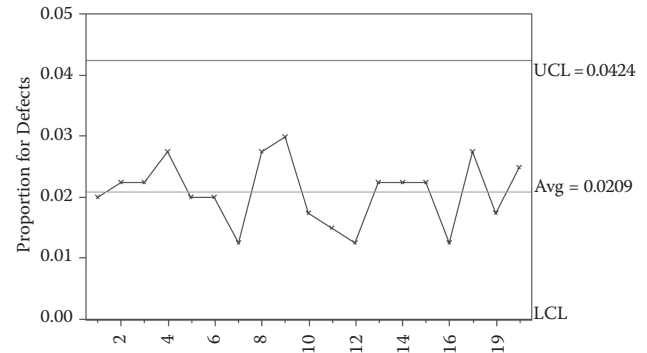


Fig. 9 An example of a p chart for coating defects

np chart, which considers the number of nonconforming defects, and the c chart, which considers the number of defects per defined unit, are two additional control charts that are used in practice for attribute data. Montgomery,^[1] Ryan,^[5] and Wheeler^[6] provide more discussions on control charts for attributes.

Probably the most important aspect of control charts is to evaluate whether the process under study is out of control. A general guideline is that a process can be defined as out of control if one (or more) subgroup value is outside the control limits, or when the control chart demonstrates a non-random pattern. For example, eight consecutive subgroup values below the centerline would, in most cases, represent a non-random pattern. There are numerous patterns, which would provide evidence that the process is out of control, that could exist. Figs. 10–12 provide examples of non-random patterns. Fig. 10 represents an example where a shift in the process mean has occurred. The shift appears to have occurred at the 15th subgroup. Fig. 11 represents an example where the mean exhibits a cyclical pattern. Fig. 12 shows a range control chart that indicates that the variance of the process has shifted around the 12th subgroup. There are many rules for determining a non-random pattern (i.e., a process is out of control). Montgomery^[1] provides an extensive list of rules that can be used. Chow and Liu^[8] discuss some non-random patterns and the corresponding probability of the event occurring under normality. A very popular set of decision rules is provided by the Western Electric Handbook.^[9] One should be aware

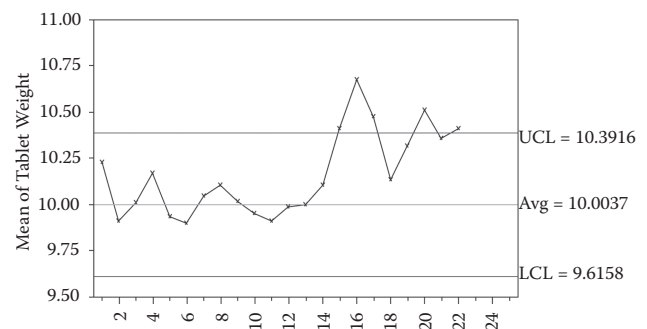


Fig. 10 An example of a shift in the process mean

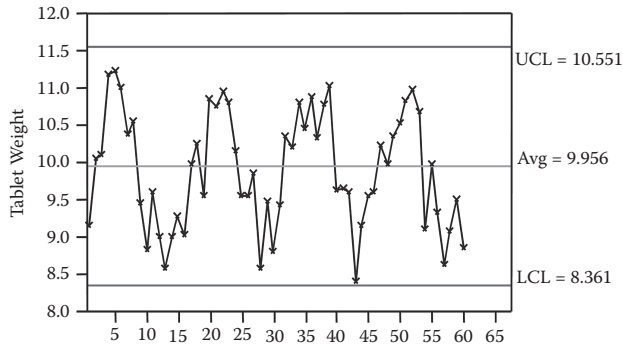


Fig. 11 An example of a cyclical pattern in the process mean for tablet weight

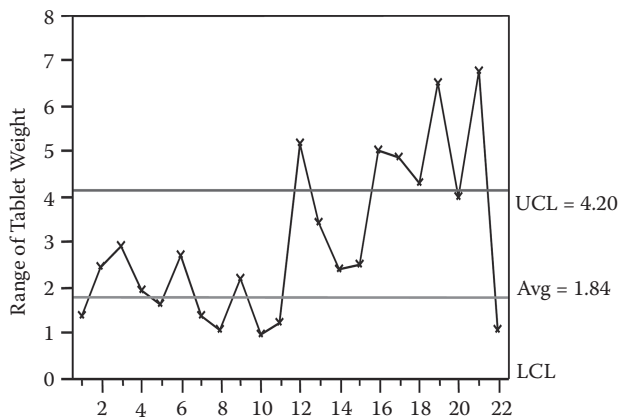


Fig. 12 An example of a shift in the process variance for tablet weight

that by using more decision rules, the probability of a false alarm significantly increases.

It is important to mention that both the process mean and the process variability need to be evaluated. There have been instances observed by the authors where the user considered only the $\bar{X}$ chart and mistakenly ignored the R chart. Often, one will evaluate the process variability before evaluating the process mean. If the R chart reveals an out-of-control condition, then the $\bar{X}$ chart should not be analyzed until the R chart has been rectified.

PROCESS CAPABILITY

Process capability is not an actual technique within the SPC tool bag; however, it is a critical component of statistical quality control. It is used to evaluate the process relative to specification limits. As previously mentioned, plotting a histogram along with the specification limits can be helpful in determining process capability. It has become common practice to use a single number (process capability indices) to quantify process capability. One of the most commonly used indices is the process capability ratio, which is defined as

$$C_p = \frac{USL - LSL}{6\sigma}$$

where USL and LSL are the upper specification limit and lower specification limit, respectively, and σ is the standard deviation computed from the control chart discussed in "Control Charts." The index C_p evaluates the spread of the distribution relative to the process range. A limitation of this index is that it does not address the location of the process mean relative to the specifications. A second commonly used capability index is the C_{pk} . This index accounts for the location of the process mean and provides a measure of how off-center the process is operating. It is defined as

$$C_{pk} = \min\left(\frac{USL - \mu}{3\sigma}, \frac{\mu - LSL}{3\sigma}\right)$$

It is our belief that confidence intervals of capability indices should become common practice when evaluating process capability. The confidence intervals should be used in order to account for the uncertainty because of sampling variability. Montgomery^[1] and Rodriguez^[10] discuss other capability indices, confidence intervals of capability indices, and procedure in assessing process capability when the quality measurement does not come from a normal distribution.

There are two very important assumptions that need to be made when presenting capability indices. They are that the quality measurement comes from a normal distribution, and that the process is in statistical control. If the quality measurement is not normally distributed, then the capability indices may not be very informative and could provide inaccurate estimates of defect levels. The normal distribution assumption can be evaluated by either histograms or goodness-of-fit tests. It is considered necessary that the process is in statistical control before evaluating process capability. If the process is not in statistical control, then the mean and/or variance is unstable and one may not be able to predict future process capability.

In the literature, there has been a discussion of the term process performance and how it differentiates from process capability. Breyfogle^[11] defines process performance as how well the process performs relative to the needs of the customer. Process performance indices use the standard deviation of the combination of all the data, which is used to estimate *long-term* variation. Long-term variation is a combination of both routine cause and special cause. So, instead of using the estimated standard from the control chart (e.g., $\bar{R}/d_2$), the sample standard deviation:

$$\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)}$$

is used to compute the process performance index. Montgomery^[1] mentions that some recommend using

process performance indices when the process is not in statistical control. However, there is much debate on the value of using process performance indices. Montgomery,^[1] Breyfogle,^[11] and Kotz and Lovelace^[12] provide some perspectives on the topic. However, if the performance index is used to quantify the performance of the process over a specified period of time and predicting future performance is not an immediate objective, then there appears to be some utility in process performance indices.

CONCLUSION

Statistical process control is an effective method for understanding and improving a process by finding assignable or special causes of variation and by removing them from the process. Histograms, Pareto charts, cause-and-effect diagrams, flowcharts, and control charts are SPC tools available to realize the goal of a stable and predictable process. Process capability indices are of use when a process is in control and future predictions about process capability need to be made. Process capability indices are of use when one wants to summarize the past performance of an in-control process or an out-of-control process. History has shown that SPC, when used diligently and effectively, can greatly enhance product quality and hence significantly contribute to a pharmaceutical company's success.

REFERENCES

1. Montgomery, D.C. *Introduction to Statistical Quality Control*; John Wiley: New York, 2009.
2. Gopal, C. Statistical process control (SPC). *Pharm. Technol* **1989**, *13* (2), 56–65.
3. Burr, I.J. The effect of nonnormality on constants for $\bar{X}$ and R charts. *Ind. Qual. Control* **1967**, *23*, 563–568.
4. Schilling, E.G.; Nelson, P.R. The effect of nonnormality on the control limits of $\bar{X}$ charts. *J. Qual. Technol.* **1974**, *8*, 183–188.
5. Ryan, T.P. *Statistical Methods for Quality Improvement*; John Wiley: New York, 1989.
6. Wheeler, D.J. *Advanced Topics in Statistical Process Control*; SPC Press: Knoxville, TN, 1995.
7. Walpole, R.E.; Myers, R.H.; Myers, S.L.; Ye, K. *Probability and Statistics for Engineers and Scientists*; Prentice-Hall: Upper Saddle River, NJ, 2002.
8. Chow, S.C.; Liu, J.P. *Statistical Design and Analysis in Pharmaceutical Science*; Prentice-Hall: Upper Saddle River, NJ, 1998.
9. Western Electric. *Statistical Quality Control Handbook*; Western Electric Corporation: Indianapolis, IN, 1956.
10. Rodriguez, R.N. Recent developments in process capability analysis. *J. Qual. Technol.* **1992**, *24* (4), 176–187.
11. Breyfogle, F.W. *Implementing Six Sigma*; John Wiley: New York, 1999.
12. Kotz, S.; Lovelace, C.R. *Process Capability Indices in Theory and Practice*; Arnold: London, 1998.

Statistical Significance

Dieter Hauschke

Byk Gulden Pharmaceuticals, Konstanz, Germany

Robert Schall and Herman G. Luus

Quintiles ClinData, Bloemfontein, South Africa

Abstract

This entry demonstrates that current statistical methodology can be applied to provide the necessary equivalence between statistical and clinical significance.

INTRODUCTION

Most controlled clinical trials are performed to demonstrate that a new investigational treatment is superior to a concurrent placebo or to an active control (superiority trials); that a test treatment is not inferior to an active treatment by more than a prespecified clinically irrelevant amount (noninferiority trials); or that two active treatments are therapeutically equivalent (equivalence trials). In the past, data from such trials have been usually analyzed by testing the traditional null hypothesis of no difference in the effect between the treatments. However, the most important drawback of this classical approach is the fact that looking only at the presence or absence of an effect ignores the issue of effect size. Hence, a statistically significant result gives only the information that the treatment difference is not zero but does not provide information as to the clinical relevance. Furthermore, failure to reject the null hypothesis does not imply the acceptance of the hypothesis of no difference. Therefore this practice has been criticized and the use of confidence intervals is advocated by regulatory guidelines.^[1,2] However, without the concept of the clinically significant (relevant) difference, the two methods are only technically different. Thus the adequate test problem should be formulated by incorporating the clinically significant difference, resulting in an equivalence of statistical and clinical significance. This direct approach will be demonstrated for superiority, noninferiority, and equivalence trials.

TESTING SHIFTED NULL HYPOTHESES

Proof of Superiority

A two-sample situation is considered. Let the clinical end point be denoted as X_1 for the test treatment and by X_0 for the control group. Suppose that these random variables are

mutually independent and have a continuous distribution function from a location family, that is,

$$X_{0j} \sim F(x - \mu_0) \quad \text{and} \quad X_{1j} \sim F(x - \mu_1), \\ j = 1, \dots, n_i, \quad i = 0, 1 \quad (1)$$

Without loss of generality, it is assumed that the population means are positive and that it is a priori known that if there is a favorable response to the treatment, it will increase in magnitude (i.e., $\mu_1 > \mu_0$), hence indicating the appropriateness of a one-sided alternative. It should be noted that the issue of a one-sided or a two-sided alternative is controversial as discussed in the literature, but this is outside our topic. Recently, the ICH guideline^[2] recommended the setting of the type I error for one-sided tests at half the conventional type I error used in two-sided testing. The traditional test problem of no difference is formulated as follows:

$$H_0 : \mu_1 - \mu_0 \leq 0 \\ H_1 : \mu_1 - \mu_0 > 0 \quad (2)$$

The rejection of the null hypothesis by a statistical test at level α (e.g., Student's t -test or Wilcoxon test) could provide evidence for the conclusion that there is a clinically relevant effect of the treatment. However, large sample sizes and little variation may lead to the problem that a clinically unimportant difference will be declared statistically significant. Suppose that a difference $\mu_1 - \mu_0 = \delta > 0$ is considered the maximum clinically irrelevant difference. If the true expected difference is less than or equal to this threshold value, then a statistically significant result for test problem 2 may be of little interest because the difference is not clinically significant. Consequently, the adequate test problem should be formulated as follows:

$$H_0^\delta : \mu_1 - \mu_0 \leq \delta \\ H_1^\delta : \mu_1 - \mu_0 > \delta \quad (3)$$

where $\delta, \delta > 0$ denotes the maximum clinically irrelevant difference. It should be noted that this definition implies that only differences greater than δ are regarded as clinically important. For example, the European guideline^[3] for chronic arterial occlusive disease requires that an irrelevant difference for the improvement in walking distance between the placebo and the new treatment should be excluded.

Assuming $F(x) = \Phi(x/\sigma)$, where $\Phi(\bullet)$ denotes the cumulative distribution function of the standard normal distribution, the shifted null hypothesis H_0^δ can be rejected if:

$$T^\delta = \frac{\bar{X}_1 - \bar{X}_0 - \delta}{S\sqrt{1/n_1 + 1/n_0}} \geq t_{\alpha, n_0+n_1-2} \quad (4)$$

where $t_{\alpha, \nu}$ is the $(1 - \alpha)$ percentile of the central t distribution with ν degrees of freedom; $\bar{X}_1$ and $\bar{X}_0$ denote the sample means of the treatment and control groups, respectively; and S_2 is the pooled estimator of σ^2 . Testing the nonzero null hypothesis (i.e., incorporating a shift parameter δ) permits the assessment of the clinical relevance of an observed treatment difference.^[4] Obviously, this decision procedure is equivalent to the condition that the lower limit of the one-sided $100(1 - \alpha)\%$ confidence interval is greater than δ , that is,

$$\left[\bar{X}_1 - \bar{X}_0 - t_{\alpha, n_0+n_1-2} S \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}, \infty \right) \subset (\delta, \infty) \quad (5)$$

The exact sample size to attain a prespecified power of $1 - \beta$ for a given level α , a standard deviation σ , and an important value $\mu_1 - \mu_0 = \delta_1 > \delta$, which should not be overlooked, can be directly obtained from the univariate noncentral t distribution or can be approximated for the balanced design (i.e., $n_0 = n_1 = n$) as follows:

$$n \geq 2 \left(\frac{(t_{\alpha, 2n-2} + t_{\beta, 2n-2}) \sigma}{\delta_1 - \delta} \right)^2 \quad (6)$$

The corresponding nonparametric method for testing the shifted hypothesis using the Wilcoxon rank sum statistic $R_1(\delta) = \sum_{j=1}^{n_1} R_{1j}(\delta)$ is presented in Hollander and Wolfe^[5] and is based on the ranks $R_{1j}(\delta)$ of the shifted observations in the treatment group in the combined sample:

$$X_{11} - \delta, \dots, X_{1n_1} - \delta, X_{01}, \dots, X_{0n_0} \quad (7)$$

H_0^δ is rejected if $R_1(\delta) \geq r(\alpha, n_0, n_1)$, where $r(\alpha, n_0, n_1)$ is the $(1 - \alpha)$ percentile of the Wilcoxon rank statistic $R_1(\delta)$. The calculation of this confidence interval according to Hollander and Wolfe^[5] can be performed as follows. Let $D_1 \leq \dots \leq D_{n_0 n_1}$ denote the ordered values of the $n_0 n_1$ differences $X_{1j} - X_{0j^*}$, $j = 1, \dots, n_1$ and $j^* = 1, \dots, n_0$. The median of these pairwise differences serves as the Hodges–Lehmann

estimator^[6] and the lower limit of the one-sided $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_0$ is given by:

$$L = D_{C_\alpha}, \quad \text{where } C_\alpha = \frac{n_1(2n_0 + n_1 + 1)}{2} + 1 - r(\alpha, n_0, n_1) \quad (8)$$

Appropriate methods for power calculation and necessary sample sizes for nonparametric procedures are given by Noether^[7] and Collings and Hamilton.^[8]

It should be noted that the use of p values and hypothesis tests in the analysis of medical data has been criticized, and point estimates and confidence intervals have been proposed. Nevertheless, it is possible to evaluate the results of clinical studies appropriately using hypothesis tests, as long as relevant hypotheses like those in Eq. (3) are tested, and not the meaningless hypothesis of Eq. (2). However, although there is nothing wrong with testing relevant hypotheses and reporting the associated p value, the use of point estimates and confidence intervals has two advantages. First, as already shown, the shifted hypotheses can conveniently be evaluated using confidence intervals. Second, point estimates and confidence intervals provide direct information about the magnitude of the treatment difference; this may be easier to interpret than p values, which express the magnitude on the probability scale.

The specification of δ requires that statisticians and clinicians think about what constitutes a maximum clinically irrelevant difference, and a common situation in practice is that the value δ is expressed as a proportion of the unknown population mean μ_0 . Therefore assuming that $\delta = f\mu_0$, $f > 0$, the test problem 3 can be formulated as:

$$\begin{aligned} H_0^\delta : \mu_1 - \mu_0 &\leq f\mu_0 \\ H_1^\delta : \mu_1 - \mu_0 &> f\mu_0 \end{aligned} \quad (9)$$

or as

$$\begin{aligned} H_0^\theta : \frac{\mu_1}{\mu_0} &\leq \theta \\ H_1^\theta : \frac{\mu_1}{\mu_0} &> \theta \end{aligned} \quad (10)$$

where $\theta = 1 + f$, $\theta > 1$, is the corresponding maximum clinically irrelevant value for μ_1/μ_0 . Under the assumption of normality and common variance, Sasabuchi^[9] demonstrated that the size- α likelihood ratio test rejects the shifted null hypothesis H_0^θ concerning the ratio of the two means if:

$$T^\theta = \frac{\bar{X}_1 - \theta \bar{X}_0}{S\sqrt{1/n_1 + \theta^2/n_0}} \geq t_{\alpha, n_0+n_1-2} \quad (11)$$

Algebraic rearrangement shows that condition 11 is equivalent to:

$$\begin{aligned} \theta_- \geq \theta \text{ and } \bar{X}_0^2 > a_0, \\ \text{where } \theta_- = \frac{\bar{X}_0 \bar{X}_1 - \sqrt{a_0 \bar{X}_1^2 + a_1 \bar{X}_0^2 - a_0 a_1}}{\bar{X}_0^2 - a_0}, \quad (12) \\ a_0 = \frac{S^2}{n_0} t_{\alpha, n_0 + n_1 - 2}^2, \quad \text{and} \quad a_1 = \frac{S^2}{n_1} t_{\alpha, n_0 + n_1 - 2}^2 \end{aligned}$$

Condition 12 can be interpreted as follows: If the lower limit of the one-sided $100(1 - \alpha)$ Fieller^[10] confidence set for μ_1/μ_0 is finite, the interval is given by $I_F = (\theta_-, \infty)$. Furthermore, Fieller's one-sided confidence set is a lower bounded interval if and only if $\bar{X}_0^2 > a_0$ holds true. Hence, the following two procedures for the assessment lead to the same decisions: (1) conclude superiority if and only if H_0^θ is rejected by the level α Sasabuchi test; and (2) conclude superiority if and only if the lower limit of the one-sided $100(1 - \alpha)\%$ Fieller confidence interval is above θ .

The following approximate formula for a balanced sample size determination to attain a prespecified power of $1 - \beta$ for a given level α , a coefficient of variation $CV_0 = \sigma/\mu_0$, and a value from the alternative $\mu_1/\mu_0 = \theta_1 > \theta$ was given by Hauschke et al.:^[11]

$$n \geq (1 + \theta^2) \left(\frac{(t_{\alpha, 2n-2} + t_{\beta, 2n-2}) CV_0}{\theta_1 - \theta} \right)^2 \quad (13)$$

It should be noted that a short discussion of corresponding nonparametric methods was provided by Vuorinen and Turunen,^[12] but further research on this issue still needs to be performed.

Proof of Noninferiority

The most convincing way to establish the efficacy of a new investigational treatment is to demonstrate clinically relevant superiority to a concurrent placebo group. In the case where an accepted treatment exists, the comparison against placebo may be considered unethical. Furthermore, when a new treatment has certain advantages, such as fewer side effects or simpler application, it might be enough to show that the test treatment is not relevantly inferior to the comparator. The choice of the extent of irrelevant inferiority should be based on clinical reasoning and will require researchers to think about what constitutes a clinically relevant difference. The corresponding hypothesis and alternative can be written as:

$$\begin{aligned} H_0^\delta : \mu_1 - \mu_0 \leq \delta^* \\ H_1^\delta : \mu_1 - \mu_0 > \delta^* \end{aligned} \quad (14)$$

where $\delta^*, \delta^* < 0$ denotes the maximum clinically irrelevant difference. Analogous to the previously given methodology for superiority trials, this shifted null hypothesis can be tested and the adequate sample size determination can be performed.

Proof of Equivalence

Many clinical trials aim to show equivalence between a test treatment under development and an existing reference treatment. In such studies, the issue is no longer to detect a difference but to demonstrate that the two active treatments are equivalent within a priori stipulated margins. A widespread application of this problem is the assessment of comparative bioavailabilities in a bioequivalence trial. Let the interval (δ_1, δ_2) denote the prespecified equivalence range; the corresponding test problem is formulated as follows:

$$\begin{aligned} H_0^\delta : \mu_1 - \mu_0 \leq \delta_1 \text{ or } \mu_1 - \mu_0 \geq \delta_2 \\ H_1^\delta : \delta_1 < \mu_1 - \mu_0 < \delta_2 \end{aligned} \quad (15)$$

Under normality and by assuming that the margins δ_1 and δ_2 are known, H_0^δ can be rejected at level α in favor of H_1^δ (equivalence) if the parametric two-sided $100(1 - 2\alpha)\%$ confidence interval for $\mu_1 - \mu_0$ is entirely included in the equivalence range:

$$\left[\bar{X}_1 - \bar{X}_0 \pm t_{\alpha, n_0 + n_1 - 2} S \sqrt{\frac{1}{n_0} + \frac{1}{n_1}} \right] \subset (\delta_1, \delta_2) \quad (16)$$

This procedure is equivalent to the two one-sided tests procedure by means of shifted two-sample t -tests.^[13] It should be noted that the corresponding nonparametric test procedure and the corresponding confidence interval construction were provided by Hauschke et al.^[14]

In the following, the equivalence limits δ_1 and δ_2 are again defined as proportions of the unknown mean μ_0 . Suppose that $\delta_1 = f_1 \mu_0$ and $\delta_2 = f_2 \mu_0$, where $-1 < f_1 < 0 < f_2$. For example, $f_1 = -f_2 = -0.20$ corresponds to the common $\pm 20\%$ criteria. The foregoing test problem can be formulated as:

$$\begin{aligned} H_0^\theta : \frac{\mu_1}{\mu_0} \leq \theta_1 \text{ or } \frac{\mu_1}{\mu_0} \geq \theta_2 \\ H_1^\theta : \theta_1 < \frac{\mu_1}{\mu_0} < \theta_2 \end{aligned} \quad (17)$$

where (θ_1, θ_2) , $\theta_1 = 1 + f_1$, $\theta_2 = 1 + f_2$, $0 < \theta_1 < 1 < \theta_2$, is the corresponding equivalence interval for the ratio μ_1/μ_0 . Testing this two-sided equivalence problem is equivalent

to the simultaneous testing of the following two one-sided hypotheses:

$$\begin{aligned} H_{01}^{\theta} : \frac{\mu_1}{\mu_0} \leq \theta_1 & \quad H_{02}^{\theta} : \frac{\mu_1}{\mu_0} \geq \theta_2 \\ \text{and} & \\ H_{11}^{\theta} : \frac{\mu_1}{\mu_0} > \theta_1 & \quad H_{12}^{\theta} : \frac{\mu_1}{\mu_0} < \theta_2 \end{aligned} \quad (18)$$

The size- α likelihood ratio test rejects H_{01}^{θ} and H_{02}^{θ} if:

$$\begin{aligned} T_1^{\theta} &= \frac{\bar{X}_1 - \theta_1 \bar{X}_0}{S \sqrt{1/n_1 + \theta_1^2/n_0}} \geq t_{\alpha, n_0+n_1-2} \quad \text{and} \\ T_2^{\theta} &= \frac{\bar{X}_1 - \theta_2 \bar{X}_0}{S \sqrt{1/n_1 + \theta_2^2/n_0}} \leq -t_{\alpha, n_0+n_1-2} \end{aligned} \quad (19)$$

The algebraic rearrangement shows that condition 19 is equivalent to:

$$\begin{aligned} \theta_- \geq \theta_1 \quad \text{and} \quad \theta_+ \leq \theta_2 \quad \text{and} \quad \bar{X}_0^2 > a_0, \\ \text{where } \theta_{\pm} &= \frac{\bar{X}_1 \bar{X}_0 \pm \sqrt{a_0 \bar{X}_1^2 + a_1 \bar{X}_0^2 - a_0 a_1}}{\bar{X}_1^2 - a_0}, \\ a_0 &= \frac{S^2}{n_0} t_{\alpha, n_0+n_1-2}^2, \quad \text{and} \quad a_1 = \frac{S^2}{n_1} t_{\alpha, n_0+n_1-2}^2 \end{aligned} \quad (20)$$

Condition 20 can be interpreted as follows: If the two-sided $100(1 - 2\alpha)\%$ Fieller confidence set for μ_1/μ_0 has finite length, it is given by the interval $I_F = (\theta_-, \theta_+)$. Furthermore, Fieller's confidence set is an interval if and only if $\bar{X}_0^2 > a_0$ holds true. Hence the following two procedures for assessing test problem 17 lead to the same decisions: (1) conclude equivalence if and only if H_{01}^{θ} and H_{02}^{θ} are rejected by the level α Sasabuchi tests; and (2) conclude equivalence if and only if the two-sided $100(1 - 2\alpha)\%$ Fieller confidence interval is included in the equivalence range.

The problem of exact power calculation and sample size determination under the assumption of normality was addressed by Hauschke et al.^[15] The following approximate formulas for sample size determination were given by Kieser and Hauschke,^[16] assuming $\theta_1 = 1/\theta_2$ and $\theta^* = \mu_1/\mu_2$:

$$\begin{aligned} \text{if } \theta^* = 1 \quad n &\geq (1 + \theta_2^2) (t_{\alpha, 2n-2} + t_{\beta/2, 2n-2})^2 \times \left(\frac{CV_0}{1 - \theta_2} \right)^2 \\ \text{if } 1 < \theta^* < \theta_2 \quad n &\geq (1 + \theta_2^2) (t_{\alpha, 2n-2} + t_{\beta, 2n-2})^2 \times \left(\frac{CV_0}{\theta_2 - \theta^*} \right)^2 \\ \text{if } \frac{1}{\theta_2} < \theta^* < 1 \quad n &\geq \left(1 + \frac{1}{\theta_2^2} \right) (t_{\alpha, 2n-2} + t_{\beta, 2n-2})^2 \times \left(\frac{CV_0}{1/\theta_2 - \theta^*} \right)^2 \end{aligned} \quad (21)$$

Wider Applications

In the preceding sections, the evaluation of study results using shifted hypothesis testing and confidence intervals was described when the focus of interest was a difference or a ratio of treatment means. A similar approach can be used in other cases. For example, when a difference between two proportions, such as the proportion of patients with a positive outcome for two treatments, is the focus of interest, then the shifted hypotheses can be modified by replacing μ_1 and μ_0 by π_1 and π_0 . The same approach can be adopted for the ratio of proportions (i.e., the relative risk).^[17]

CONCLUSION

Testing a nonzero null hypothesis instead of one of no treatment difference removes at a stroke some logical problems. For example, one problem of testing the classical null hypothesis is that even for the most minute true difference between two treatments, the p value will indicate statistical significance if the sample size is big enough. When data are analyzed by testing shifted hypotheses, or through confidence intervals as outlined, then an increase in sample size or in the precision of the data, or both, will lead to a narrowing of the confidence interval and thus to an increase in the probability of obtaining a conclusive result (i.e., a statistically and clinically relevant significance). This entry demonstrates that current statistical methodology can be applied to provide the necessary equivalence between statistical and clinical significance.

REFERENCES

1. CPMP Working Party on Efficacy of Medical Products. *Note for Guidance: Biostatistical Methodology in Clinical Trials in Applications for Marketing Authorizations for Medicinal Products*; 1994.
2. The European Agency for the Evaluation of Medicinal Products. ICH E9. In *Note for Guidance on Statistical Principles for Clinical Trials*; 1998.
3. The European Agency for the Evaluation of Medicinal Products. *The Clinical Investigation on Medicinal Products in the Treatment of Chronic Peripheral Arterial Occlusive Disease*; 1995; CPMP/EWP/235/95.
4. Victor, N. On clinically relevant differences and shifted null hypotheses. *Methods Inf. Med.* **1987**, 26, 109–116.
5. Hollander, M.; Wolfe, D.A. *Nonparametric Statistical Methods*; Wiley: New York, 1973.
6. Hodges, J.L.; Lehmann, E.L. Estimates of location based on rank tests. *Ann. Math. Stat.* **1963**, 34, 598–611.
7. Noether, G.E. Sample size determination for some common nonparametric tests. *J. Am. Stat. Assoc.* **1987**, 82, 645–647.
8. Collings, B.J.; Hamilton, M.A. Estimating the power of the two-sample Wilcoxon test for location shift. *Biometrics* **1988**, 44, 847–860.

9. Sasabuchi, S. A multivariate one-sided test with composite hypotheses determined by linear inequalities when the covariance matrix has an unknown scale factor. *Mem. Fac. Sci., Kyushu Univ., Ser. A, Math.* **1988**, *42*, 9–19.
10. Fieller, E. Some problems in interval estimation. *J. R. Stat. Soc., B* **1954**, *16*, 175–185.
11. Hauschke, D.; Kieser, M.; Hothorn, L.A. Proof of safety in toxicology based on the ratio of two means for normally distributed data. *Biom. J.* **1999**, *41*, 295–304.
12. Vuorinen, J.; Turunen, J. A simple three-step procedure for parametric and nonparametric assessment of bioequivalence. *Drug Inf. J.* **1997**, *31*, 167–180.
13. Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* **1987**, *15*, 657–680.
14. Hauschke, D.; Steinijans, V.W.; Diletti, E. A distribution-free procedure for the statistical analysis of bioequivalence studies. *Int. J. Clin. Pharmacol. Ther. Toxicol.* **1990**, *28*, 72–78.
15. Hauschke, D.; Kieser, M.; Diletti, E.; Burke, M. Sample size determination for proving equivalence based on the ratio of two means for normally distributed data. *Stat. Med.* **1999**, *18*, 93–105.
16. Kieser, M.; Hauschke, D. Approximate sample sizes for testing hypotheses about the ratio and difference of two means. *J. Biopharm. Stat.* **1999**, *9*, 641–650.
17. Farrington, C.P.; Manning, G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Stat. Med.* **1990**, *9*, 1447–1454.

Structural Equation Model

Joseph L. Carrasco

Bioestadística, Universitat de Barcelona, Barcelona, Spain

Abstract

This entry examines the simple linear structural equation with a dependent variable and an independent variable with measurement error. It also shows how the structural equation model can be applied to individual bioequivalence assay.

Keywords: Individual bioequivalence; Latent variable; Measurement error; Measurement method comparison; Orthogonal regression; Reliability; Structural relationship.

INTRODUCTION

The structural equation model is a regression approach that allows to estimate a linear regression when independent variables are measured with error. The approach is based on defining latent variables, which represent unobservable true values; thus the model considers that the observed data are the result of adding the measurement error to the *true values*. Moreover, the relationship between the true values of the independent and dependent variables is defined through the structural equation.

The classical regression models do not take into account the measurement error of the independent covariates, thus giving biased estimators of slopes.

Here, we examine the simple linear structural equation with a dependent variable and an independent variable with measurement error. We also show how the structural equation model can be applied to individual bioequivalence assay.

STRUCTURAL EQUATION MODEL

The classical regression model links a dependent variable or outcome to a vector of independent or explanatory variables. This model assumes that the explanatory variables are measured with no error and that the only random variation is in the outcome conditioned to the explanatory variables. However, there are situations where this assumption may not be valid. For example, let us suppose that we want to estimate the relationship between the levels of a hormone with respect to the intake of alcohol and age. In this situation, a sample of subjects is collected and the blood concentration of the hormone, the weekly intake of alcohol, and the age of the subjects are measured. It appears that the age of the subjects can be measured with no error, at least if it is rounded to years, but the same cannot be

said of the intake of alcohol. This variable will probably be measured by using a questionnaire where the subjects declare their alcohol consumption in the last week. Even if the subjects are completely sincere, the amount of alcohol declared will be an approximation of the real quantity; therefore the intake of alcohol will be measured with error. Thus if the relationship between the outcome and the explanatory variable is estimated using an approach that does not take into account the measurement error (such as ordinary least squares), such estimators will be biased and inconsistent.^[1]

The structural equation model is a regression approach that estimates the model by taking into account that the independent variables are measured with error. For that reason, the structural equation model is also called measurement error regression, or error-in-variables regression.

Here, only the simple linear structural equation with an explanatory variable will be considered, but there are extensions to more complex models, including several dependent variables^[2] and nonlinear relationships.^[3]

THE MODEL

Let Y_i be an outcome and let X_i be an explanatory variable measured on the i th subject, where $i = 1, \dots, n$. The simple structural equation model is defined as:^[4]

$$\begin{aligned} Y_i &= \eta_i + \varepsilon_i \\ X_i &= \xi_i + \delta_i \\ \eta_i &= \alpha + \beta \xi_i \end{aligned}$$

where Y and X are the observed data, η and ξ are the *true values* or latent variables, and ε and δ are the measurement errors. The parameters α and β characterize the structural equation, which is no more than the linear relationship between the latent variables.

It is usually assumed that the measurement errors do not covary with any other component of the model, and that they are distributed under normal distributions with mean 0 and standard deviations σ_ε and σ_δ , respectively. It is also assumed that the latent variable ξ is distributed under a normal distribution, with mean μ and standard deviation σ .

Thus the structural equation determines the relationship between the means of the variables and the covariance structure. The mean of Y can be expressed as:

$$E(Y) = E(\eta + \varepsilon) = E(\alpha + \beta\xi + \varepsilon) = \alpha + \beta E(\xi) + E(\varepsilon) = \alpha + \beta\mu$$

which is the same result as that achieved with the classical regression model. But there are differences in the covariance structure because:^[4]

$$Var(Y) = Var(\eta + \varepsilon) = Var(\alpha + \beta\xi + \varepsilon) = \beta^2\sigma^2 + \sigma_\varepsilon^2$$

$$Var(X) = Var(\xi + \delta) = Var(\xi) + Var(\delta) = \sigma^2 + \sigma_\delta^2$$

$$cov(Y, X) = cov(\eta + \varepsilon, \xi + \delta) = \beta\sigma^2$$

Thus the slope of the structural equation is not only involved in the mean but also in the variance and covariance. Moreover, the slope also defines the relationship between the standard deviations of the latent variables:

$$Var(\eta) = Var(\alpha + \beta\xi) = \beta^2 var(\xi)$$

$$\beta = \sqrt{\frac{Var(\eta)}{Var(\xi)}}.$$

Therefore there are six parameters that define the model and have to be estimated: μ , σ , α , β , σ_ε , and σ_δ .

The estimation of parameters can be carried out by maximum likelihood, but, unfortunately, there are more parameters to estimate than there are estimating equations (i.e., there is not enough information in the data to allow the estimation of parameters). This problem is known as a lack of identification of the model.

The model becomes identified by incorporating information on some of the parameters. Traditionally, five cases can be found in the literature:^[5]

1. The ratio of the error variances $\lambda = \sigma_\varepsilon^2 / \sigma_\delta^2$ is known.
2. Both of the error variances are known.
3. σ_δ^2 is known.
4. σ_ε^2 is known.
5. The intercept α is known.

Probably, the most commonly used assumption is the case where the ratio of error variances is known and is equal to 1, perhaps because it is the first assumption ever used in structural equation models.^[6] This case has been also called orthogonal regression or principal component model^[7] because the orthogonal distances from the pair of points to the regression line are minimized.

The expression of the estimates and their standard errors will rely on the assumption made. A summary of the estimates and the variance-covariance matrix for each case can be found in Ref. [8].

It is possible to achieve identification by taking replicates (measuring individuals more than once) of Y and/or X and by using them to estimate the measurement error variances.^[9] Once the error variances have been estimated, they can be considered as known, and the estimation of the remaining parameters becomes feasible.^[12,9] Depending on whether the replicates are taken on one or both variables, cases (2), (3), or (4) can be applied. This approach is known as partial likelihood analysis^[9] because full log likelihood is expressed as the sum of a conditional log likelihood and a marginal log likelihood. For the case of both variances known, the log likelihood can be expressed as:

$$\log L(\alpha, \beta, \sigma^2, X_{ij}, Y_{ij}, \sigma_\delta^2, \sigma_\varepsilon^2) = \log L(\alpha, \beta, \sigma^2 | \bar{X}_i, \bar{Y}_i, \sigma_\delta^2, \sigma_\varepsilon^2) + \log L(W_X, W_Y, \sigma_\delta^2, \sigma_\varepsilon^2)$$

where

$$\begin{aligned}\bar{X}_i &= \frac{1}{k} \sum_{j=1}^k X_{ij} \\ \bar{Y}_i &= \frac{1}{k} \sum_{j=1}^k Y_{ij} \\ W_X &= \sum_{i=1}^n \sum_{j=1}^k (X_{ij} - \bar{X}_i)^2 \\ W_Y &= \sum_{i=1}^n \sum_{j=1}^k (Y_{ij} - \bar{Y}_i)^2\end{aligned}$$

and $i = 1, \dots, n$ and $j = 1, \dots, k$, where n is the number of subjects and k is the number of replicates. Thus the conditional log likelihood depends on individual means and within-subject variances, whereas marginal log likelihood relies on within-subject deviations. Then within-subject variances are estimated by:

$$\sigma_\delta^2 = \frac{W_X}{n(k-1)}$$

$$\sigma_\varepsilon^2 = \frac{W_Y}{n(k-1)}$$

and incorporated in the conditional log likelihood as known, and the remaining parameters are estimated through the conditional log likelihood.

The estimates follow an asymptotic normal distribution;^[4] thus building confidence intervals or testing hypotheses can be performed by using a normal distribution, although Fuller^[2] recommend a t student distribution for small samples.

However, there is a problem with the confidence interval estimation of β , which is known as the Gleser-Hwang effect. Gleser and Hwang^[10] demonstrated that when the

number of subjects is fixed, every confidence interval for β , of finite length, has a confidence equal to 0. Nevertheless, Gleser^[11] showed that if the quantity known as signal-to-noise, $\kappa^2 = \sigma^2/\sigma_\delta^2$ and the number of subjects are large enough ($\kappa^2 \geq 1$ and more than 25 subjects), asymptotic confidence is reasonably close to the desired coverage probability. The heuristic rationale could be that we are trying to estimate the relationship between two variables that are *contaminated* by some noise (measurement error). Of course, the greater is the noise, the more difficult it will be to estimate this relationship and more *signal* will be required. The signal can be understood as the variability between subjects, so that, if the noise is large, more subjects with a greater spread of values will be required.

The maximum likelihood approach gives consistent and efficient estimates, but it relies on the normality assumption of the data. Other estimation procedures do not depend on such an assumption; thus they are more robust to departures from normality. However, if data are normally distributed, the maximum likelihood estimates are the most efficient.

Among these procedures, we would like to highlight the generalized least squares, the instrumental variables approach, and methods based on ranks.

The generalized least squares or weighted least squares^[12] consist of minimizing square distances from points to the regression line by using weights. These weights are taken to be proportional to the reciprocal of variance errors: $1/(\sigma_e^2 + \beta^2 \sigma_\delta^2)$. Nevertheless, it has been demonstrated that the generalized least squares approach is no more than the classical orthogonal regression.^[4]

The instrumental variables approach is useful in any regression model where the independent variable is correlated with the error. The structural model is an example of this kind of regression because the explanatory variable X is correlated with the error of the regression line. This fact can be easily seen if the structural model is expressed as: $Y = \alpha + \beta X + e$, $e = \varepsilon - \beta\delta$ then $cov(X, e) = cov(\xi + \delta, \varepsilon - \beta\delta) = -\beta cov(\delta, \delta) = -\beta\sigma_\delta^2$.

The instrumental variables approach^[13] consists of finding a variable, say Z , that is correlated with the independent variable X but uncorrelated with the error e . This variable can be continuous, discrete, or a factor. Then the estimator of the slope is:

$$\hat{\beta}_V = \frac{cov(Z, Y)}{cov(Z, X)}$$

However, the use of instrumental variables has been criticized because of the difficulty of finding such a variable,^[3,5] and in the words of Cheng and Van Ness,^[4] “the justification of using instrumental variables is based on consistency considerations only, and this does not guarantee good performance in applications involving large sample variances and moderate sample sizes, in which the consistency properties cannot be brought to bear.”

Moreover, the selection of instrumental variables might not be unique. Then, in the same analysis, it would be possible to choose different instrumental variables that would give different estimates.

The methods based on ranks^[14,15] can be viewed as a special case of instrumental variables, where the instrumental variable used is the order statistic. Although this is a robust approach, there is an underlying assumption about the distribution of the true values. The X values must be of the same order as the true values; therefore the true values must be more widely distributed than the error variance.

STRUCTURAL EQUATION MODEL APPLIED TO BIOEQUIVALENCE

Structural equation models have been applied to several fields. As examples, we would like to highlight reliability studies,^[16] analytical methods comparison,^[7,15] health sciences,^[17–9] and behavioral sciences.^[20,21]

In the pharmaceutical field, the structural equation model has also been applied to evaluate individual bioequivalence.^[22] The aim of a bioequivalence assay is to demonstrate that a reference (R) and a tested (T) formulation are equal with respect to their safety and efficacy—this objective is achieved if the two formulations have the same amount of bioavailabilities.

Bioequivalence assays are typically based on models of the logarithm of bioavailability measurements. The measurement model proposed for bioavailabilities (or its log-transformed values) is:^[23] $T_{ij} = \mu_T + u_{Ti} + e_{Tij}$; $R_{ij} = \mu_R + u_{Ri} + e_{Rij}$ where T_{ij} and R_{ij} are the bioavailabilities of tested (T) and reference (R) formulations; μ_T and μ_R are the overall averages of T and R; u_{Ti} and u_{Ri} are the mean deviations of individual i with respect to the overall mean; and e_{Tij} and e_{Rij} are the random deviations of individual i on j th determination, where $i = 1, \dots, n$ and $j = 1, \dots, k$. Hereinafter, we will refer to u_{Ti} and u_{Ri} as between-subject deviations, and e_{Tij} and e_{Rij} as within-subject deviations. It is assumed that both between-subject and within-subject deviations are distributed under normal distributions with parameters:

$$u_T \sim N(0, \sigma_{BT}), u_R \sim N(0, \sigma_{BR}), \\ e_T \sim N(0, \sigma_{WT}), e_R \sim N(0, \sigma_{WR})$$

It is also assumed that the within-subject deviations do not covariate with any parameter.

Individual bioequivalence requires a certain degree of agreement between the bioavailabilities of two formulations: the overall averages, the between-subject deviations, and the variability of the within-subject deviations have to be close enough.

From the measurement model for bioavailabilities, we can define a structural model equation. Because the between-subject deviations u_T and u_R are unknown and

unobservable, they can be considered as latent variables, and the within-subject deviations can be considered as some noise around the latent variable. Then, the structural equation is defined as: $u_{T_i} = \alpha + \beta u_{R_i}$.

The requirements to guarantee that the bioavailabilities are close enough can be summarized in three points:

1. The difference between overall means is within some interval, around $0\phi = \mu_T - \mu_R = \alpha + (\beta - 1)\mu_R \in [\theta_A, \theta_B]$
2. The slope of the structural equation is within some interval, around 1 $\beta \in [\theta_C, \theta_D]$
3. The ratio of within-subject variances is below some limit $\lambda = \sigma_{WT}^2 / \sigma_{WR}^2 \leq \theta_E$

where, depending on the values of θ_A , θ_B , θ_C , θ_D , and θ_E , the criteria will be more or less restrictive.

Therefore the hypothesis of individual bioequivalence will be tested in a disaggregated sense by evaluating five hypotheses about ϕ , β , and λ :

$$H_{01} : \phi < \theta_A, H_{02} : \phi > \theta_B, H_{03} : \beta < \theta_C, H_{04} : \beta > \theta_D, H_{05} : \lambda > \theta_E$$

$$H_{11} : \phi \geq \theta_A, H_{12} : \phi \leq \theta_B, H_{13} : \beta \geq \theta_C, H_{14} : \beta \leq \theta_D, H_{15} : \lambda \leq \theta_E$$

Thus the overall hypothesis of individual bioequivalence is expressed as:

$$H_0 : \text{No bioequivalence} = H_{01} \cup H_{02} \cup H_{03} \cup H_{04} \cup H_{05}$$

$$H_1 : \text{Bioequivalence} = H_{11} \cap H_{12} \cap H_{13} \cap H_{14} \cap H_{15}$$

This sort of mixed hypothesis is tested using the *intersection-union principle* of Roy.^[24] Following this principle, the null hypothesis of no bioequivalence is rejected with a significance level α if all null hypotheses are rejected at level α .

The hypotheses concerning β and ϕ can be tested in a *two one-sided way*^[25] by using interval estimation with a confidence of $1 - 2\alpha$, with α as the desired significance level. If the whole confidence interval lies within the bioequivalence limits, both null hypotheses will be rejected.

The hypothesis concerning within-subject variances can be tested by estimating the $1 - \alpha$ percentile of λ , with α as the desired significance level. This percentile is estimated as: $\hat{\lambda}F_{1-\alpha, v_1, v_2}$ where $F_{1-\alpha, v_1, v_2}$ is the $1 - \alpha$ percentile of F distribution with: $v_1 = v_2 = n(k - 1)$ degrees of freedom, where k is the number of measures by subject and formulation, and n is the number of subjects.

EXAMPLE

As an example to illustrate the structural equation model applied to bioequivalence data, we will use data from a β -adrenergic blocking agent (Table 1). The bioavailabilities were measured on 22 subjects using a four-period, four-sequence replicated crossover design. Therefore two measures of formulation and subject are available; thus it

is possible to estimate both within-subject variances. The bioavailability measure used was the area under the curve (AUC), and its logarithm was taken.

The estimates of the within-subject variances were:

$$\hat{\sigma}_{WT}^2 = 0.1453 \quad \hat{\sigma}_{WR}^2 = 0.0334$$

$$\hat{\lambda} = \frac{0.1453}{0.0334} = 4.3458$$

These variances are implemented as known parameters in the conditional log likelihood. The overall means, variances, and covariances of the individual means are:

$$\bar{X}_R = \frac{1}{22} \sum_{i=1}^{22} \bar{R}_i = 8.448$$

$$\bar{X}_T = \frac{1}{22} \sum_{i=1}^{22} \bar{T}_i = 8.3393$$

$$S_R^2 = \text{Var}(\bar{R}_i) = 0.1398$$

$$S_T^2 = \text{Var}(\bar{T}_i) = 0.1894$$

$$S_{RT} = \text{cov}(\bar{R}_i, \bar{T}_i) = 0.1237$$

The estimates of the remaining parameters are:^[19]

$$\hat{\beta} = \frac{S_T^2 - \hat{\lambda}S_R^2 + \left[(S_T^2 - \hat{\lambda}S_R^2)^2 + 4\hat{\lambda}S_{RT}^2 \right]^{1/2}}{2S_{RT}} = 0.9938$$

$$\hat{\alpha} = \bar{X}_T - \hat{\beta}\bar{X}_R = -0.0562$$

$$\hat{\sigma}_{BR}^2 = \frac{S_T^2 + \hat{\lambda}S_R^2 - \hat{\sigma}_{WT}^2 + [(S_T^2 - \hat{\lambda}S_R^2)^2 + 4\hat{\lambda}S_{RT}^2]^{1/2}}{2(\hat{\lambda} + \hat{\beta}^2)} = 0.1233$$

To contrast the hypotheses related to individual bioequivalence, it is necessary to define the bioequivalent limits, which depend on how restrictive the criterion should be. A proposal could be:

$$\theta_A = -\log 1.25 \quad \theta_B = \log 1.25 \quad \theta_C = 2/3$$

$$\theta_D = 1.5 \quad \theta_E = 2.25$$

The values of θ_A and θ_B come from the average bioequivalence criterion;^[26] θ_C and θ_D are set using the fact that β is the ratio of between-subject standard deviations. Thus any standard deviation cannot exceed the other by more than 1.5 times. This criterion is also used for the ratio of within-subject variances, but here, it is only necessary that the within-subject variance of the tested formulation does not exceed the within-subject variance of reference formulation by more than 2.25 times (1.5 in standard deviation scale).

First, the hypothesis relating to the difference of overall averages will be tested. In this order, a 90% confidence interval is built as: $\hat{\phi} \pm t_v SE(\hat{\phi})$ where $\hat{\phi} = \hat{\mu}_T - \hat{\mu}_R = \hat{\alpha} + (\hat{\beta} - 1)\hat{\mu}_R = -0.1088$ t_v means a t

Table 1 Bioavailabilities from 22 Subjects

Subject	Treatment	AUC _{inf}	Subject	Treatment	AUC _{inf}
1	T	4249.13	12	R	2977.81
1	R	6247.17	12	T	3447.86
1	R	6176.99	12	T	3685.11
1	T	5704.69	12	R	3133.16
2	T	4338.26	13	T	7515.12
2	R	3007.12	13	R	5470.67
2	T	3669.55	13	T	6689.67
2	R	3300.89	13	R	7312.90
3	R	7159.13	14	T	1805.74
3	T	6903.59	14	R	5632.05
3	T	6310.22	14	T	2599.29
3	R	8784.03	14	R	1397.73
4	R	4192.02	15	T	6363.07
4	T	6547.37	15	R	4191.13
4	R	4300.16	15	R	5408.03
4	T	5427.73	15	T	6595.91
5	R	4587.51	16	R	2311.37
5	T	4470.78	16	T	3823.20
5	T	4347.01	16	T	3080.52
5	R	5286.37	16	R	3468.52
6	T	2914.40	17	T	3257.30
6	R	3034.20	17	R	5633.36
6	R	2885.66	17	T	6127.34
6	T	3761.72	17	R	3368.14
7	R	4201.72	18	R	3803.58
7	T	4070.66	18	T	6040.94
7	R	5568.49	18	R	5183.14
7	T	4618.16	18	T	5977.77
8	T	2239.15	19	T	2475.68
8	R	4920.52	19	R	2402.68
8	T	7197.01	19	R	3398.51
8	R	4830.84	19	T	3653.72
9	T	2937.23	20	R	7057.29
9	R	3214.84	20	T	9100.23
9	R	3884.37	20	T	5687.27
9	T	3164.91	20	R	7135.98
10	R	4696.92	21	T	6631.13
10	T	3980.49	21	R	7236.01
10	R	4182.13	21	R	7990.74
10	T	4494.21	21	T	7001.24
11	R	9533.56	22	R	556.99
11	T	8070.16	22	T	4496.19
11	T	7190.22	22	R	2298.70
11	R	8095.60	22	T	2677.48

All subjects were measured twice in both reference (R) and tested (T) formulations.

distribution with $v = n - 2$ degrees of freedom, with n as the number of subjects; therefore $v = 20$. The standard error of the difference of means is:^[22]

$$Var(\hat{\phi}) = \frac{\sigma_{WT}^2 + \beta^2 \sigma_{WR}^2}{kn}$$

with k as the number of measures by subject and formulation (in this case, $k = 2$). Then: $SE(\hat{\phi}) = 0.0637$ and the 90% confidence interval becomes $[-0.2186; 0.0010]$, which is completely within the bioequivalent limits; thus both hypotheses concerning difference of means are rejected.

In relation to the slope, the 90% confidence interval is built using: $\hat{\beta} \pm t_v SE(\hat{\beta})$ and, in that case, the standard error of the slope is expressed by:

$$Var(\hat{\beta}) = \frac{1}{n\sigma_{BR}^4} \left(\theta \left(\sigma_{BR}^2 + \frac{\sigma_{WR}^2}{k} \right) - \beta^2 \left(\frac{\sigma_{WR}^2}{k} \right)^2 \right)$$

where $\theta = \frac{1}{k} (\sigma_{WT}^2 + \beta^2 \sigma_{WR}^2)$

The estimate of the standard error is 0.1910; thus the 90% confidence interval for β is $[0.6642; 1.3333]$. The lower limit exceeds the bioequivalent limit. Thus one of the hypotheses concerning β is not rejected.

Finally, the hypothesis on λ is tested by estimating the 95th percentile of the λ estimate. This percentile is estimated by using an F distribution with $n(k-1)$ and $n(k-1)$ degrees of freedom; therefore: $per95(\hat{\lambda}) = \hat{\lambda} per95(F_{22,22}) = 8.8991$ which is greater than the bioequivalent limit; thus the null hypothesis is rejected.

We conclude that the two formulations are not individually bioequivalent, but a structural equation model allows a disaggregate analysis, thereby making it possible to discuss the sources of nonbioequivalence.

In this example, we have shown that the difference in overall averages is low enough to conclude that they are equivalent. However, the between-subject variances are not equivalent. The large variability of the β estimate, combined with a point estimate close to 1, indicates a problem of lack of power: more subjects were needed to make a decision. Indeed, the number of 22 subjects is quite low to achieve a high degree of power. Carrasco and Jover^[22] carried out a power analysis by arriving at the conclusion that between 50 and 100 subjects is necessary to guarantee a certain degree of power. However, the within-subject variance of the tested formulation is clearly greater than the within-subject variance of the reference formulation, this being the main reason to reject the individual bioequivalence.

CONCLUSION

Structural equation modeling is a useful regression approach when the independent variables are measured

with a nonnegligible measurement error. If this measurement error is not taken into account, the estimates of the slopes and their standard errors will be biased.

Here the simple normal structural model has been dealt with. Thus only an independent variable is included in the model, and the true values as well as the measurement errors are assumed to be distributed under normal distributions. In this case, the maximum likelihood estimates can be obtained by applying the explicit expressions shown in Ref. [4] according to the different assumptions that make the model identifiable. Nevertheless, multivariate models with more than one independent (and/or dependent) variable could be considered. Then, it could be possible that the maximum likelihood estimates did not have an explicit expression. Thus numerical methods to compute maximum likelihood estimates are needed. These numerical methods are implemented in several statistical packages. Among these packages, we would like to highlight the PROC CALIS of SAS®, the AMOS of SPSS®, the *sem* function for R,^[27] and the *Mplus*® package.^[28]

The estimation procedures introduced here are based on a frequentist point of view where the estimation process completely relies on the sample data. Alternatively, Bayesian approach^[29] has been proposed to get the structural equation model estimates. Using Bayesian approach the identification of the model can be achieved by introducing previous knowledge through the prior distributions of the parameters and hyperparameters involved in the model. The prior distributions combined with the likelihood lead to the parameter posterior distributions that are used for inferential purposes. With large sample size, both frequentist and Bayesian estimates concur because the likelihood has much more weight than the prior distributions on the posterior distribution. However, if the sample size is small the estimates are highly influenced by the prior belief and their consistency depends upon the appropriateness of the prior distributions. The structural equation model introduced here assumes that the relationship between the latent variables is unique for all individuals. This assumption can be quite unrealistic (Ref. [3], pp. 30) and as a consequence the estimates may be biased. However it can be relaxed by adding a random error to the structural equation that allows the relationship between latent variables randomly changes from one individual to other. Thus, the model is named the structural error-in-equation model (SEEM),^[4] which has been also applied to the individual bioequivalence problem.^[30]

The use of the structural equation model allows estimations of the underlying model that generates the data. When this model is applied to bioequivalence trials, the disaggregate nature of the approach allows us to determine why the formulations are not bioequivalent, rather than to decide if the formulations are bioequivalent. Therefore this approach can be understood as a tool to analyze what is wrong with a tested formulation.

REFERENCES

- Gleser, L.J.; Carroll, R.J.; Gallo, P.P. The limiting distribution of least squares in an errors-in-variables linear regression model. *Ann. Stat.* **1987**, *15*, 220–233.
- Fuller, W.A. *Measurement Error Models*; John Wiley and Sons: New York, 1987.
- Carroll, R.J.; Ruppert, D.; Stefanski, L.A. *Measurement Error in Nonlinear Models*; Chapman and Hall: London, 1995.
- Cheng, C.L.; Van Ness, J.W. *Statistical Regression with Measurement Error*; Kendall's Library of Statistics; Arnold: London, 1999.
- Cheng, C.L.; Van Ness, J.W. On estimating linear relationship when both variables are subject to errors. *J. R. Stat. Soc. Ser. B.* **1994**, *56* (1), 167–183.
- Addock, R.J. Note on the method of least squares. *Analyst* **1877**, *4*, 183–184.
- Feldmann, U.; Schneider, H.; Klinkers, H. A multivariate approach for the biometric comparison of analytical methods in clinical chemistry. *J. Clin. Chem. Clin. Biochem.* **1981**, *19*, 121–137.
- Hood, K.; Nix, B.A.J.; Iles, T.C. Asymptotic information and variance-covariance matrices for the linear structural model. *Statistician* **1999**, *48* (4), 477–493.
- Chinchilli, V.M.; Esinhart, J.D.; Miller, W.G. Partial likelihood analysis of within-unit variances in repeated measurement experiments. *Biometrics* **1995**, *51* (1), 205–216.
- Gleser, L.J.; Hwang, J.T. The non-existence of $100(1 - \alpha)\%$ confidence sets of finite expected diameter in errors-in-variables and related models. *Ann. Stat.* **1987**, *15*, 1351–1362.
- Gleser, L.J. Confidence Intervals for the Slope in a Linear Errors-in-Variables Regression Model. In *Advances in Multivariate Statistical Analysis*; Gupta A.K.; Ed.; D. Reidel: Dordrecht, 1987, 85–109.
- Sprent, P. A generalized least squares approach to linear functional relationships. *J. R. Stat. Soc. Ser. B* **1966**, *28*, 278–297.
- Bowden, R.J.; Turkington, D.A. *Instrumental Variables*; Cambridge University Press: Cambridge, 1984.
- Theil, H. A rank-invariant method for linear and polynomial regression analysis. *Indag. Math.* **1950**, *12*, 85–91, 173–177, and 467–482.
- Passing, H.; Bablok, W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. *J. Clin. Chem. Clin. Biochem.* **1983**, *21*, 709–720.
- Kelly, G.E. Use of the structural equations model in assessing the reliability of a new measurement technique. *Appl. Stat.* **1985**, *34* (3), 258–263.
- Hasenabeldy, N. Fuller, W.A.; Ware, J. Indoor air pollution and pulmonary performance: Investigating errors in exposure assessment. *Stat. Med.* **1988**, *8*, 1109–1126.
- Freedman, L.S.; Carroll, R.J.; Wax, Y. Estimating the relation between dietary intake obtained from a food frequency questionnaire and true average intake. *Am. J. Epidemiol.* **1991**, *134* (3), 310–320.
- Rosner, B.; Spiegelman, D.; Willett, W.C. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error. *Am. J. Epidemiol.* **1990**, *132* (4), 734–745.
- Duits, A.A.; Duivenvoorden, H.J.; Boeke, S.; Taams, M.A.; Mochtar, B.; Krauss, X.H.; Passchier, J.; Erdman, R.A.M. A structural modeling analysis of anxiety and depression in patients undergoing coronary-artery bypass graft-surgery—A model generating approach. *J. Psychosom. Res.* **1999**, *46* (2), 187–200.
- Johnson, J.A.; Coons, S.J.; Hays, R.D. The structure of satisfaction with pharmacy services. *Med. Care.* **1998**, *36* (2), 244–250.
- Carrasco, J.L.; Jover, L. Assessing individual bioequivalence using the structural equation model. *Stat. Med.* **2003**, *22*, 901–912.
- Schall, R.; Luus, H.G. On population and individual bioequivalence. *Stat. Med.* **1993**, *12*, 1109–1124.
- Roy, S.N. On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Stat.* **1953**, *24*, 220–238.
- Schuurmann, D.J. A comparison of the two one-sided test procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* **1987**, *15*, 657–680.
- Center for Drug Evaluation and Research (CDER). *U.S. Food and Drug Administration. Guidance for Industry: Statistical Approaches to Establishing Equivalence*; CDER 2001.
- <http://www.bioconductor.org/CRAN/src/contrib/PACKAGES.html#sem> (accessed November, 2003).
- Muthén, K.L.; Muthén, B.O. *Mplus Statistical Analysis With Latent Variables User's Guide*; Muthén and Muthén: Los Angeles, 2003.
- Lee, S.Y. *Structural Equation Modelling: A Bayesian Approach*; John Wiley & Sons: New York, 2007.
- Carrasco, J.L.; Jover, L. The structural error-in-equation model to evaluate individual bioequivalence. *Biomet. J* **2005**, *47*, 623–634.

Stuart–Maxwell Test

Jiang-Ming Johnny Wu

Biostatistics, DOV Pharmaceutical, Inc., Somerset, New Jersey, U.S.A.

Abstract

This entry discusses the Stuart–Maxwell test and its applications, and provides examples of clinical trials that implement the Stuart–Maxwell Test.

Keywords: Categorical data analysis, χ^2 -test; Contingency tables; Matched pairs; SAS macros.

INTRODUCTION

In clinical trials, categorical data analysis plays a crucial rule due to its broad applications that can be applied to different studies. There are many known test procedures that are used in categorical data analysis. For example, in the 2×2 independent samples cases: Pearson χ^2 -test of homogeneity, Fisher's exact test, likelihood ratio test, and Cochran–Mantel–Haenszel (CMH) test may be used. For 2×2 matched pairs case, the McNemar's χ^2 -test^[1,2] and the sign test, etc. can be used. Some of the test procedures mentioned in the 2×2 independent sample cases are still applicable for 3×3 and higher-order contingency tables. However, for 3×3 and higher-order matched pairs cases, only the Stuart–Maxwell test^[3,4] can be applied.

Due to the FDA requirements, it is well known that the SAS® language is the dominant software to be used for generating clinical tables, listings, and graphs in the pharmaceutical industry. However, there is no developed SAS procedure that can perform such a test. This lack of a developed SAS procedure creates a need for some pharmaceutical companies who use this test to perform the inferential analysis for their product in clinical trials. In this article, the author introduces this useful test and attempts to bridge this gap by developing an SAS® macro via SAS/IML®^[5] language to perform this test. The Stuart–Maxwell test will be presented and discussed in the second section. In the third section, a brief introduction regarding the use of the author's SAS macro is presented. A sample SAS main program that applies to an artificial data set is also included. In the fourth section, an application of the Stuart–Maxwell test with an example from the real-world clinical study using the developed SAS macro is presented. Finally, discussions regarding some special cases are included. The developed SAS macro codes that perform the Stuart–Maxwell χ^2 -test can be found in the paper by Wu.^[6]

THE STUART–MAXWELL TEST

This test deals with n matched pairs with r response categories with data configuration as shown in Table 1.

Here, a_{ij} is the number of pairs in the (i, j) th cell, $a_{i.}$ is the i th row sum, and $a_{.j}$ is the j th column sum, $i, j = 1, 2, \dots, r$. Also, the word “Group” is a generic term that may refer to a “Treatment” in clinical trials. Denote $p_i(p_{.j})$ to be the proportion of persons in group 2 who are in response category $i(j)$. Then the null hypothesis for the Stuart–Maxwell test can be formulated as follows:

$$H_0: p_{1.} = p_{.1}, p_{2.} = p_{.2}, \dots, p_{r.} = p_{.r}, r = 2, 3, 4, \dots \quad (1)$$

against the alternative hypothesis

$$H_1: p_{i.} \neq p_{.i} \text{ for at least one } i = 1, 2, \dots, r. \quad (2)$$

For the $r \times r$ table, the test statistic for testing $H_0: (p_{1.}, p_{2.}, \dots, p_{r.}) = (p_{.1}, p_{.2}, \dots, p_{.r})$ is

$$T = f' \mathbf{V}^{-1} f \approx \chi^2(r-1), \quad (3)$$

that is χ^2 distributed with $(r-1)$ degrees of freedom. Here,

$$f' = (a_{1.} - a_{.1}, a_{2.} - a_{.2}, \dots, a_{r.} - a_{.r}), \quad (4)$$

$$\mathbf{V} = [v_{ij}], \quad i, j = 1, 2, \dots, r-1, \quad (5)$$

$$v_{ii} = a_{i.} + a_{.i} - 2a_{ii}, \quad i = 1, 2, \dots, r-1, \quad (6)$$

$$v_{ij} = -(a_{ij} + a_{ji}), \quad \text{for } i \neq j, i, j = 1, 2, \dots, r-1, \quad (7)$$

and $\mathbf{V}^{-1}$ is the inverse matrix of $\mathbf{V}$.

Table 1 Data Configuration for the Stuart–Maxwell Test

		Group 1 Response category					Total
		1	2		r		
Group 2 Response category	1	a_{11}	a_{12}		a_{1r}		$a_{1.}$
	2	a_{21}	a_{22}		a_{2r}		$a_{2.}$
	...	...	...		...		
	r	a_{r1}	a_{r2}		a_{rr}		$a_{r.}$
Total		$a_{.1}$	$a_{.2}$		$a_{.r}$	$a_{..} = n$	

The only assumption for the Stuart–Maxwell test is that the data arise from a matched random sample with proportions $p_{.j}(p_{i.})$ for categories $1, 2, \dots, r$ for group 1(2). When $r = 2$ the Stuart–Maxwell test reduces to the well-known McNemar's test.^[1] Fleiss and Everitt^[7] has also shown that, when $r = 3$, the test statistic T can be reduced to the following format:

$$T = [\hat{a}_{23}(a_{1.} - a_{1.})^2 + \hat{a}_{13}(a_{2.} - a_{2.})^2 + \hat{a}_{12}(a_{3.} - a_{3.})^2] / [2(\hat{a}_{12}\hat{a}_{23} + \hat{a}_{12}\hat{a}_{13} + \hat{a}_{13}\hat{a}_{23})], \quad (8)$$

where

$$\hat{a}_{ij} = (a_{ij} + a_{ji}) / 2, \quad i, j = 1, 2, 3. \quad (9)$$

As usual, the decision rule for a α -level test is: reject H_0 if $T > \chi^2_{1-\alpha}(r-1)$.

In some situations a 3×3 table may be collapsed into a 2×2 table by combining two categories of response into one. In this case the McNemar's χ^2 -test can be used to perform the analysis. However, Fleiss^[8] suggested that under this case one should use a degree of freedom of 2, instead of 1.

The following is an example selected from Woolson's book:^[9]

Example

In a study of the psychological and social effects on the wives of myocardial infarction patients, a semistructured interview was conducted and tape-recorded on two occasions, 2 months after and 1 year after the husband's infarction. One item that was rated was the wife's degree of anxiety or distress. The data are as follows:

		At two months			
	Degrees of anxiety	None or mild	Moderate	Severe	Total
At one year	None or mild	20	30	10	60
	Moderate	5	10	5	20
	Severe	5	5	10	20
	Total	30	45	25	100

Determine whether there is sufficient evidence to conclude that the anxiety levels change from 2 months to 1 year.

Computations

From Eqs. 8 and 9,

$$\hat{a}_{12} = (30 + 5) / 2 = 17.5, \quad \hat{a}_{13} = (10 + 5) / 2 = 7.5, \\ \hat{a}_{23} = (5 + 5) / 2 = 5.$$

$$a_{1.} = 30, \quad a_{2.} = 45, \quad a_{3.} = 25, \quad a_{.1} = 60, \quad a_{.2} = 20, \quad a_{.3} = 20.$$

$$T = [5(60 - 30)^2 + 7.5(20 - 45)^2 + 17.5(20 - 25)^2] / [2(5 \times 17.5 + 17.5 \times 7.5 + 7.5 \times 5)] \\ = 9625 / 512.5 = 18.78.$$

$$H_0 : (p_1, p_2, p_3) = (p_{.1}, p_{.2}, p_{.3}) \text{ vs} \\ H_1 : (p_1, p_2, p_3) \neq (p_{.1}, p_{.2}, p_{.3}).$$

Because $\chi^2_{0.95}(2) = 5.99 < 18.78$, H_0 is rejected and in favor of H_1 . It is concluded that the wife's anxiety levels do change from 2 months to 1 year after husband's infarction.

THE SAS PROGRAM

A SAS[®] macro, written in SAS/IML[®] language, that performs the Stuart–Maxwell test has been developed by the author and can be found in the paper by Wu.^[6] To use that macro, one must do the following things to the data set.

- Rename the response variables as *yvar1* and *yvar2*.
- Sort the data set by visit number with final visit number equals to "99".
- Obtain the frequency distribution of the matched pairs (*yvar1*, *yvar2*) by visit number and store the results in an output data set with name "out".

The following is a sample SAS main program that uses the developed macro, STUART, with an artificial test data set.

Sample SAS Main Program

*** Sample main program to use the macro STUART ***;

```
libname in "/data";
data main;
set in.test;
if trtmnt = 1 then yvar1 = wound1;
else if trtmnt = 2 then yvar1 = wound2;
if control = 1 then yvar2 = wound1;
else if control = 2 then yvar2 = wound2;
if day in ("F") then vnum = 99;
else vnum = input(day,2.);
drop wound1 wound2;
```

```

run;
proc print data = main;
  title "Raw data set";
run;
proc sort data = main;
  by vnum;
run;
proc freq data = main noprint;
  by vnum;
  tables yvar1**yvar2/out = out;
run;
data main;
  set out(keep = vnum yvar1 yvar2 count);
  by vnum;
run;
proc sort data = main;
  by vnum;
run;
%macro useit(visit = );
%if &visit = %then %do; *** Use the macro for all visits
  ***;
  %do i = 1% to 10; *** This number can be changed to fit
    into your own situation ***;
  %stuart(&i,4,prnt = );
  %end;
  %stuart(99,4,prnt = );
%end;
%else %do; *** Use the macro for the selected visit only
  ***;
  %stuart(&visit,4);
%end;
proc print data = final;
  title "data final—summary output for Stuart–Maxwell
   $\chi^2$ -test (all visits)";
run;
%mend useit;
%useit(visit = 5); *** Example 1—Use the macro for visit
  = 5 only ***;
%useit(visit = ); *** Example 2—Use the macro for all
  visits ***;
endsas;

```

THE APPLICATION TO CLINICAL TRIALS

The application of the Stuart–Maxwell test using the developed SAS macro, STUART, applied to a real-world clinical data set will be discussed in this section. The discussion begins with a brief description of the clinical study design and its purpose.

Clinical Study

The W. Company developed a Phase IV protocol to study its product, XXXXXXXX Silver-Coated Dressing that could cure “burn wounds.” The study design was a multicenter (10–15 centers), inpatient, matched-wound, randomized, open-label clinical trial. Patients with burn wounds were enrolled for treatment. Two large wounds of similar size on the body were selected from each patient and then both the study drug and the reference drug were randomly applied to these wounds. The planned sample size was 100 evaluable patients but eventually only 64 patients were obtained. The objective of the study was: To compare the safety, efficacy, cost effectiveness and quality of life benefits of XXXXXXXX silver-coated dressing to a standard burn treatment/dressing (silver sulfadiazine) in the treatment of partial-thickness burn wounds.

For the above clinical study the developed SAS macro, STUART, is applied to the data set to obtain the desired inferential analysis results. Two selected tables are presented below which include:

- **Table 2:** Analysis for reepithelialization for improved patients (ITT)
- **Table 3:** Analysis of physician’s evaluation of therapeutic response (ITT)

Here the reepithelialization includes four levels: light, moderate, marked, and completed. Also, the physician’s satisfaction includes four different levels: excellent, good, fair, and poor. It is seen that both responses are 4×4 matched pair cases and, because the study/reference drugs are

Table 2 Analysis of Reepithelialization for Improved Patients Intent-to-Treat Population

Reepithelialization	Counts (%)		Btwn-trt <i>p</i> -values ^a
	XXXXXXX (<i>N</i> = 61)	SSD (<i>N</i> = 61)	
Final day			
Light/mod. (1%–50%)	13 (21.3)	17 (27.9)	0.034 ^b
Marked (51%–99%)	15 (24.6)	14 (23.0)	
Completed (100%)	9 (14.8)	3 (4.9)	
Total	37 (60.7)	34 (55.7)	

Intent-to-treat population includes all subjects who had baseline and at least one post-baseline scores for primary efficacy assessments.

^a $p \leq 0.01$. $p \leq 0.001$.

^a p -values for categorical data were computed using Maxwell–Stuart χ^2 -test.

^b $p \leq 0.05$.

Table 3 Analysis of Physician's Evaluation of Therapeutic Response Intent-to-Treat Population

Satisfaction	Counts (%)		Btwn-trt p -values ^a
	XXXXXXX ($N = 61$)	SSD ($N = 61$)	
Final day			
Excellent	15 (24.6)	5 (8.2)	0.025 ^b
Good/fair	34 (55.7)	45 (73.8)	
Poor	10 (16.4)	9 (14.8)	
Total	59 (96.7)	59 (96.7)	

Intent-to-treat population includes all subjects who had baseline and at least one post-baseline scores for primary efficacy assessments.

^b $p \leq 0.01$. ^a $p \leq 0.001$.

^a p -Values for categorical data were computed using Maxwell–Stuart χ^2 -test.

^b $p \leq 0.05$.

applied to the same patients, the response variables are not independent. The results of using the Stuart–Maxwell test via the developed SAS macro are reproduced on the next page. Note that only the last page is displayed for Table 2. Also, both examples include two categories collapsed into one combined category, as can be seen in the results.

Both results show that the test drug has a significant effect (p -value = 0.034 and 0.025, respectively) over the reference drug, regarding “cure of the burn wounds” and “physician’s satisfaction.”

DISCUSSION

Because none of the references mentioned in this paper discussed what to do if the matrix $\mathbf{V}$ in Eq. 3 is singular (the matrix $\mathbf{V}^{-1}$ does not exist), it is worthwhile to have a raise this issue here. In the normal case, this test procedure is not applied if $\mathbf{V}^{-1}$ does not exist. However, because the clinical data usually is difficult to obtain and is expensive, we should utilize the clinical data as much as possible. Therefore, in order to utilize the data to the greatest extent, it is suggested to collapse the $r \times r$ table ($r \geq 3$) to be $(r-1)(r-1)$ by combining two categories into one new category, when zero rows and zero columns exist. To illustrate, consider the following example. Suppose $r = 3$ with 3 possible responses being: 1 = mild, 2 = moderate, and 3 = severe, and the matched pairs combination as shown in the following matrix.

$$\mathbf{A} = \begin{bmatrix} 5 & 0 & 10 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad (10)$$

If no collapsing category is done, then we will have:

$$f' = (10, 0) \text{ and } \mathbf{V} = \begin{bmatrix} 10 & 0 \\ 0 & 0 \end{bmatrix} \quad (11)$$

The usual analysis using Eq. 3 will result in $\mathbf{V}^{-1}$ does not exist and, hence, no conclusion will be reached. However, if

we combine category 1 and 2 as a new category, mild/moderate, for example, using the McNemar’s test will result in a test statistic $T = 10$ that will reject the null hypothesis. Examining the $\mathbf{A}$ matrix above, this appears to be a very fair conclusion.

The second suggestion is to always remember to use degrees of freedom ($r - 1$) when dealing with an $r \times r$ table, even if there are two response categories being combined into one category, as suggested by Fleiss^[8] for the $r = 3$ case.

In Wu,^[6] the proposed SAS macro will automatically search for zero rows and zero columns and combine them into one category, as long as at least one zero row and one zero column exist. To do this, a simple algorithm, described below, is used.

Algorithm to Find the Zero Row and Zero Column for the $\mathbf{A}$ Matrix

1. Find the number of zero rows (columns) and reduce the row (column) rank by one each time a zero row (column) is found.
2. Start with the smallest number of zero rows (columns) and remember its index, the i th row (column), $1 \leq i \leq r$, is zero and then try to find out whether the corresponding i th column (row) is zero or not.
3. If the corresponding i th column (row) is also zero, reduce the $\mathbf{A}$ matrix to be of dimension $(r-1)(r-1)$ by deleting both the i th zero row (column) and i th zero column (row).
4. If the corresponding i th column (row) is not zero, reduce the $\mathbf{A}$ matrix to the dimension $(r-1)(r-1)$ by deleting the i th zero row (column) and the first zero column (row) found.
5. When the $\mathbf{A}$ matrix contains only one zero row and one zero column, this row and column will automatically be deleted to form a new reduced matrix, $\mathbf{A}_3$, with a dimension of $(r-1)(r-1)$.

For the above example, the $\mathbf{A}$ matrix has one zero column (column 2) and two zero rows (row 2 and row 3).

Therefore, using the above algorithm, the macro will first identify column 2 to be deleted and then search for row 2 and determine whether it is a zero row. Because it is a zero row in this case, the macro then deletes row 2 to form a new reduced 2×2 matrix **A3** as follows:

$$\mathbf{A3} = \begin{bmatrix} 5 & 10 \\ 0 & 0 \end{bmatrix}. \quad (12)$$

Also, when such a reduction of matrix or combination of two categories into one has been performed, a warning message will automatically be displayed in the output file to warn the user that the interpretation of the table is different.

When the Stuart–Maxwell test is not applied, should the corresponding p -value always be assigned as missing? My personal opinion is no. For example, in the case where the matrix **A** is symmetric, it can be shown that the corresponding **V** matrix will always be singular and, therefore, the Stuart–Maxwell test will not be applied. However, in this case, a p -value of 1 makes much more sense than a missing p -value. A simple example is illustrated below for this purpose.

Let

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (13)$$

then we have

$$f' = (0,0) \text{ and } \mathbf{V} = \begin{bmatrix} 4 & -4 \\ -4 & 4 \end{bmatrix} \quad (14)$$

that is singular. As a consequence, the Stuart–Maxwell test in Eq. 3 does not apply. However, the symmetry of the **A** matrix tells you that there is no difference between those matched pair responses and, therefore, a p -value of 1 is reasonable in this case.

ARTICLES OF FURTHER INTEREST

Analysis of 2K Tables; Analysis of Clustered Binary Data; p -values.

REFERENCES

1. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157.
2. Conover, W.J. *Practical Nonparametric Statistics*; Wiley: New York, 1971.
3. Stuart, A. The comparison of frequencies in matched samples. *Br. J. Statist. Psychol.* **1957**, *110*, 29–32.
4. Maxwell, A.E. Comparing the classification of subjects by two independent judges. *Br. J. Statist. Psychol.* **1970**, *116*, 651–655.
5. SAS Institute Inc. SAS/IML®. Software: Usage and Reference Version 6SAS Institute Inc., Cary, NC, 1989.
6. Wu, J.M. A SAS procedure for the Stuart–Maxwell chi-square test and its application to clinical trials. *ASA Proceedings, Section on Biopharmaceutical Statistics*, ASA: Alexandria, VA, 2004, 926–932.
7. Fleiss, J.L.; Everitt, B.S. Comparing the marginal totals of square contingency tables. *Br. J. Math. Statist. Psychol.* **1971**, *24*, 117–123.
8. Fleiss, J.L. *Statistical Methods for Rates and Proportions*; 2nd Ed.; Wiley: New York, 1981.
9. Woolson, R.F. *Statistical Methods for the Analysis of Biomedical Data*; Wiley: New York, 1987.

Subgroup Analysis

Patricia B. Cerrito

University of Louisville, Louisville, Kentucky, U.S.A.

Abstract

Subgroup analysis can and should be performed when examining the results of both clinical trials and observational studies. This entry discusses the need for subgroup analysis and provides alternative methods of analysis and statistical examples of the benefits of this type of analysis.

INTRODUCTION

Because there are many different segments of the population as a whole and because these different subgroups of the population respond differently to treatment, subject recruitment for randomized, controlled trials is now actively attempting to be fairly representative of the population as a whole. However, this does not always occur.^[1] The statistical methods used need to investigate the outcomes for subgroups, in addition to the need to examine for statistical significance on the entire population. To do this, estimation and exploratory techniques need to be used. Validation by fresh data also needs to be emphasized.

NEED FOR SUBGROUP ANALYSIS

Most statistical hypotheses testing that occur in clinical trials make an assumption that the population is relatively homogeneous. This translates statistically into an assumption that the population has a distribution that resembles a bell-shaped curve (Fig. 1). However, consider the study^[2] that demonstrated that race and ethnicity are important in determining the absorption rates of drugs. Because body weight is also of concern, gender is also a real concern. In terms of trial outcomes, no distinctions are made to determine differences within the populations. This means that it is unlikely that the various symptoms and reactions will present themselves in the form of a bell-shaped curve. Phase I trials are designed to examine dosage levels for all subsequent trials. Typically they use small samples—samples too small to determine all differences related to race, gender, and ethnicity.

Consider a contrast to the assumption of homogeneity when heterogeneity exists in the population (Fig. 2). The assignment of “normal” is usually defined as the amount of the curve that contains 95% (or often 75%) of the area under the bell-shaped curve (Fig. 3). However, such an

assumption will overpresent one subgroup while underpresenting another. It can be seen clearly that half of the second peak falls above the established “norm.” This becomes even more noticeable when the distributions for the two subpopulations are graphed separately (Fig. 4). In fact, the assumption of normality in the presence of bimodality tends to redefine what is normal for one subgroup as an abnormal reaction, because it lies beyond the “norm” of the dominant subgroup (Fig. 5). In Fig. 5, the mean for the second group actually is placed at the upper limit for the first subgroup.

Thus a minority population could be misdiagnosed or treated improperly because of a reliance upon the assumption of homogeneity. Unfortunately, different subpopulations cannot always be clearly distinguished. Although gender, race, and ethnicity are fairly obvious indicators of subpopulations, other characteristics should also be considered. In this case, it is also a matter of finding out what subpopulations exist, as well as which ones should be investigated separately.

The use of subgroups by race or ethnicity is not without controversy. Ref. [3] strongly recommends that race and ethnicity be discarded as population subgroups, to be replaced by genetic markers. It states:

Studies using new technologies for understanding, measuring, and conceptualizing the sources of human variation reveal that approximately 85% of all variation in gene frequencies occurs within populations or races and only 15% variation occurs between such populations; there is no genetic basis for racial classification.

Contrast this statement with one made by Dr. Edward Sondik, Director for the National Center for Health Statistics: “Data showing racial differences in incidence and mortality have important scientific implications. These data should be looked at as offering clues for further research versus providing answers in and of itself.”

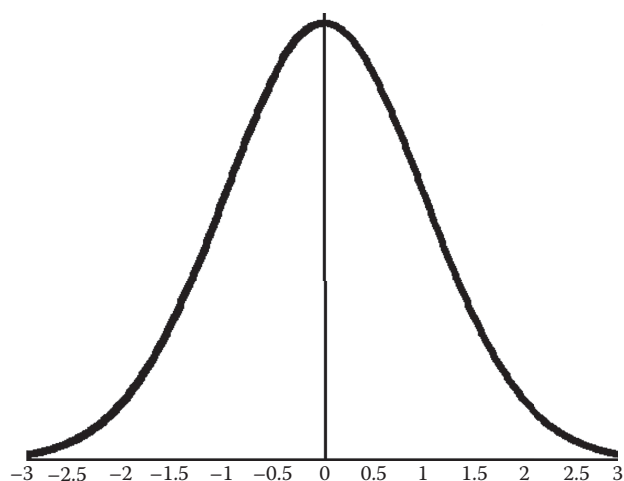


Fig. 1 Generalized representative distribution of a homogeneous population

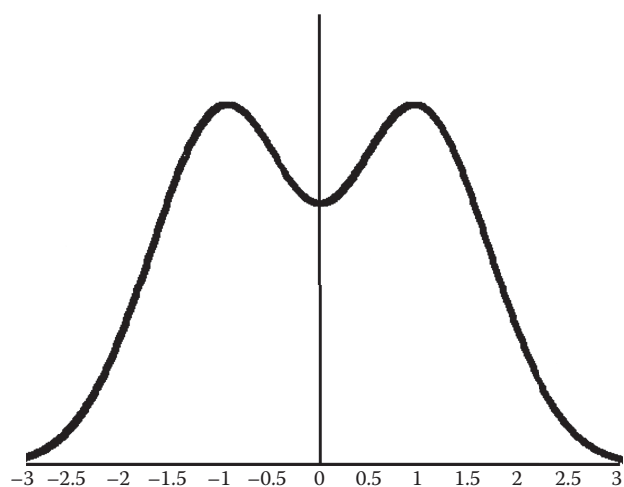


Fig. 2 Generalized representation of a heterogeneous population

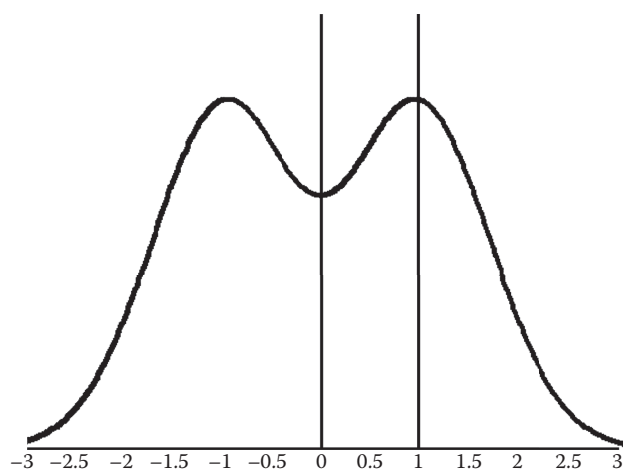


Fig. 3 Generalized representation of a heterogeneous population distribution showing the same 70% upper tail norm established under the assumption of homogeneity

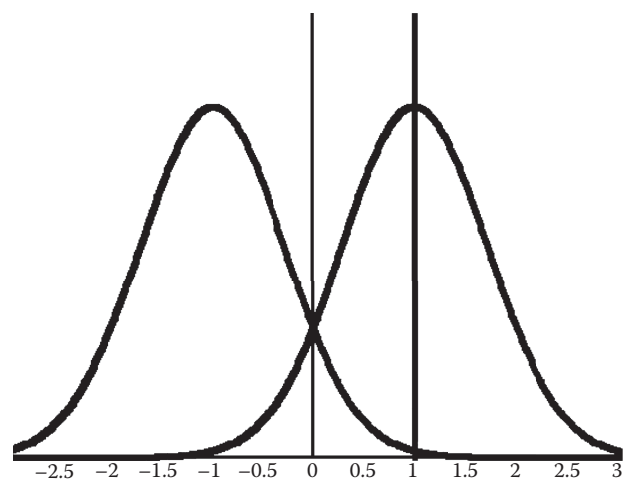


Fig. 4 Generalized representation of probability distributions for two homogeneous subpopulations showing the 70% upper tail norm established under the assumption that the entire population is homogeneous

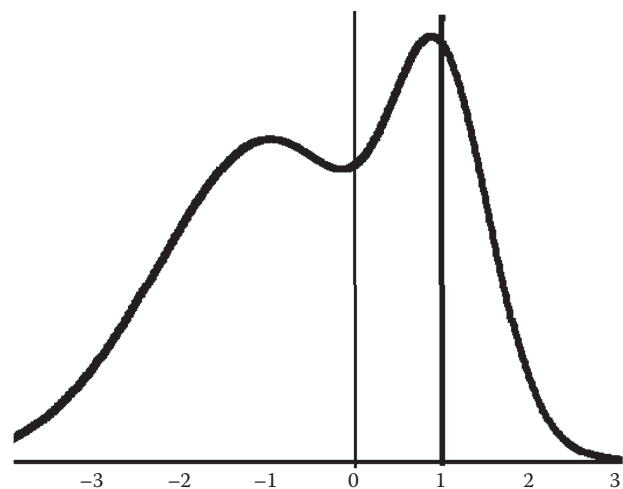


Fig. 5 Generalized representation of two subpopulations demonstrating that one subpopulation can redefine the norms for a second subpopulation under the assumption of homogeneity

The first comment assumes that it is possible to collect data on the genetic makeup of clinical trial subjects; unless the trial itself examines genetic variation, this is not possible. The second statement provides a more realistic attitude in that statistical analysis does not examine why subgroups occur within the population; it only examines their occurrence. Although the claim is being made that race is a social construct and that it is not biological, the fact is that even a social construct determines existence.

ALTERNATIVE METHODS OF ANALYSIS

There are two nonparametric tools that can be used to estimate effects in a multimodal population. Both methods

are characterized by an ability to find subgroups as well as an ability to estimate the effects of treatment on those subgroups.

Mixing Distributions

As shown in Ref. [4], mixture models involve assuming that each observation x_j is taken from a superpopulation G , which is a mixture of a finite number of populations $G_1, \dots, G_g$ that occur in G in some proportions $\pi_1, \dots, \pi_g$, respectively, such that $\pi_1 + \dots + \pi_g = 1$. Then the density function for G is equal to:

$$f(x) = \sum_{i=1}^g \pi_i f_i(x)$$

where f_i is the density function of the population G_i . The function f then can only be found by estimating f_i for $i = 1, \dots, g$ as well as by estimating $\pi_1, \dots, \pi_g$. It is also sometimes necessary to determine the value of g —the number of populations involved. Therefore an attempt must be made to decompose f into its separate parts. The total number of populations g can be either known or unknown.

It is not possible to estimate if no sampling of G_i is done. However, it is also not possible to estimate f accurately. In fact, any attempt to estimate f without sampling from all components will result in severe bias. The largest of the populations tends to dominate the distribution. If the populations are of roughly equal size, both populations are misrepresented.

The advantage of using this method is that it is possible to determine just how many subpopulations should be considered and how many confounding factors should be used in making a diagnosis. Therefore it is possible to subdivide the population and to consider the extremes of symptoms for separate populations.

Kernel Density Estimation

Kernel density estimation uses data to develop an estimate of the population density function. For the collected data $\{X_1, \dots, X_n\}$, the estimate is given by the formula:

$$\hat{f}(x) = \frac{1}{na_n} \sum_{j=1}^n K\left(\frac{x - X_j}{a_n}\right)$$

where K is a known density function and a_n is a parameter that controls the degree of smoothness of f . As the value of a_n decreases, the number of peaks in the estimator increases. It becomes necessary to find an optimal value of a_n to determine the optimal number of peaks in the data. A method for doing this is discussed in Ref. [5]. This technique also gives a nice visual representation of the population distribution and enables the investigator to determine if

minority populations respond differently from the majority. However, this technique will also result in severe bias if the sample comes only from one segment of the population.

The tail probability can be estimated in the same way in which a cumulative distribution function can be estimated. As given in Ref. [5], this function can be defined as:

$$\hat{f}(u) = \frac{1}{na_n} \sum_{i=1}^n K\left(\frac{u - X_i}{a_n}\right)$$

where K is the cumulative distribution function of the kernel:

$$K(u) = \int_{-\infty}^u K(u) du$$

and $X_1, X_2, \dots, X_n$ is the collected sample. Thus the tails of the probability distribution can be determined in detail. The value u can also define a vector of random variables $u = (x, y, z, t)$. Then $f(u)$ is modified to equal (for the discrete variable y):

$$\begin{aligned} \hat{f}(x, y, z, t) &= \frac{1}{na_n^3} \sum_{j=-\infty}^{\infty} W(\hat{S}_n, i, j) \\ &\times \sum_{l \in C_n(j)} k\left(\frac{x - X_l}{a_n}, \frac{z - Z_l}{a_n}, \frac{t - T_l}{a_n}\right) \end{aligned}$$

PROBLEMS WITH SUBGROUP ANALYSIS

Finding Significance Where None Exists

There is a saying among statisticians that if you torture the data, they will confess. This is particularly true of subgroup analysis and retrospective studies. Therefore any results from these types of analysis can be only preliminary, not conclusive. This falls under the general topic of data mining.^[6] Although a data mining procedure can be used to identify a model or a variable of interest, only fresh data can be used to verify the significance of the model. The need for verification is clearly stated:^[7]

The difficulties associated with direct reliance on data in a whole series are exacerbated if one attempts to pick out subclasses marked by strikingly good or bad results. The idea seems reasonable enough, but it ignores a somewhat subtle, inescapable fact: one can always expect to find some good-looking and some bad-looking subsets in any body of data, even when no bias has influenced any part of it. Furthermore, such differences among subsets can easily be large enough to look quite convincing to the unsophisticated analyst.

The pitfalls of subgroup analysis were very vividly presented by the ISIS-2 trial, where a statistically significant subgroup result on the prevention of fatal heart attack by ingesting aspirin was obtained by classifying subjects by astrological sign.^[8]

One way to avoid this problem is to preplan a small number of subgroup analyses by stratifying the data in the protocol. A second way is to perform the primary assessment to demonstrate statistical significance before investigating any subgroups for significance. As stated in Ref. [7]: “[P]ost-hoc subgroup results should be viewed as hypothesis-generating rather than proof.” Unfortunately, published results do not generally indicate how many subgroup analyses were actually performed. The main way to avoid the problem is to regard any results of subgroup analysis as hypothesis-generating, to be validated by the collection of a fresh data set to see if the results still hold.

Two recent papers^[9,10] rely very heavily on subgroup analysis. In Ref. [9], no significant difference was found between treatment groups. Therefore the significant difference found for subjects age 60 years and younger should definitely be regarded as a hypothesis generated, with a need for verification by additional data collection. The paper went on to examine subgroups of the subgroup, all of which were highly significant. As discussed in Ref. [7]: “[I]f n subjects are divided at random into k equal-sized subgroups, then the difference between the largest and the smallest subgroup means has an expected value that is approximately k times the standard error of the mean of the whole group.” This alone demonstrates that such high P values are to be expected. Although the paper does discuss the problem of subgroup analysis and the care taken to ensure that the results were not spurious, the choice of the age of 60 years was made because this was the optimal value to find statistical significance.

In Ref. [10], the difference between treatment and control groups was not significant. As stated, “[A] retrospective survival analysis was performed in patient subsets that showed favorable responses to VMO treatment in the first interim analysis.” It is clear that subgroups were chosen after the fact. Any difference in survival found retrospectively has to be tested with preplanned stratification before the result is considered valid. This paper does not discuss the problems of subgroup analysis. The paper does indicate that the study is not definitive and should be done with sufficient numbers of patients in all subgroups.

The problem of these types of subgroup analyses is clearly examined:^[11]

The controversy began with the retrospective analysis of results by age subgroups, a method of evaluation unanticipated in the design of all but one of the trials; these trials were planned with marginal power that was sufficient only for the whole-group examination of the data. When used to interpret the end results of the trials, retrospective

subgroup analysis reduces the power of the results and thus substantially increases the likelihood that an observed benefit will not achieve statistical significance... [P]ost hoc data-dredging for subgroups of patients or subsets of events, however qualitatively dramatic their results, should only generate hypotheses for the next trial, not conclusions from this one.

Thus subgroup analysis can create controversies within the medical community by stating conclusions that are only generated hypotheses. Further studies, often not performed, are needed to make the conjecture conclusive.

Because few people actually understand the statistical complications in the use of subgroup analysis, these should be recognized in any paper discussing a study where subgroup analysis was performed.

Insufficient Power

Before the data for any test are collected, there are four different parameters that must be considered:^[12]

Type I (or α) error.
Type II (or β) error.
Effect size.
Sample size.

These four variables are closely related, so three are specified by the investigator and the fourth is computed as a function of the other three.

Type I Error

The purpose of any statistical hypothesis test is to make an inference about a population by using a relatively small sample. The possibility of error always exists. Because it cannot be eliminated, it should be controlled. Type I error is defined to be the probability of being wrong if the investigator rejects H_0 . Typically, $\alpha = 0.05$ is considered significant and $\alpha = 0.01$ is considered highly significant. Then the hypothesis test is stated:

H_0 : treatment is not effective.
 H_1 : treatment is effective.

Type II Error

More importantly, type II error should also be considered, although it is usually ignored. Type II error is defined to be the probability of being wrong if the investigator accepts H_0 . Unfortunately, if the α error is low, the β error can be quite high. Therefore consider the question posed earlier:

H_0 : treatment is not effective.
 H_1 : treatment is effective.

If the value of β is not computed, then it is not possible to decide whether to accept H_0 . Therefore the best that can be decided is whether to reject H_0 or to fail to reject H_0 . Failing to reject H_0 is often assumed as equivalent to accepting H_0 , although it is not. In fact, independent of the size of the sample, the maximum possible β error is always $1 - \alpha$. Thus if $\alpha = 0.05$, it is possible to have a β error of 0.95. Only by defining an effect size can α and β be reduced simultaneously.

In statistical terminology, the power of the test is defined to be $1 - \beta$. Typically, in a medical study, α is considered acceptable at 0.05 and β is acceptable at a level of 0.20. In this case, the power of the test is stated as 80%. However, when subgroup analysis is performed, it is performed with only a subset of the data. In this case, the result will be of low power, and it will not be possible to conclude that the subgroup reacts differently than the rest of the population.

Effect Size

If a new treatment increases patient survival by 1 day, does it provide sufficient improvement to be worthwhile? Statistical significance cannot determine whether the size of the difference is important—that question must be answered by the investigator. The effect size is defined to be the point at which the difference between treatments becomes important. Once the effect size is defined, the hypothesis can be rewritten (when limited to two treatments):

$$\begin{aligned} H_0: u_1 &= u_2 \\ H_1: u_1 - u_2 &> u' \text{ or } u_2 - u_1 > u' \end{aligned}$$

where u' is the effect size. Once a specified effect size is given, a sample size can be computed so that both the α and β errors remain small.

Sample Size

The smaller the effect size, the larger the needed sample size to keep both α and β under control. If the effect size is too small, it is possible that the required sample size is too large for an experiment to be conducted, except for prohibitive cost. On the other hand, if the sample size is limited by cost, then the effect size may be so large that H_0 can be rejected only in such extreme situations that it would be virtually obvious without any statistical analysis.

Therefore effect size is usually identified as a percentage of the population variability, with the corresponding sample size computed for small, medium, and large effect sizes. Once this has been done, the null hypothesis can be either accepted or rejected, depending upon the outcome of the collected data.

If the effect size is not considered prior to the start of an experiment, it is often the case that the effect size is so large as to render the experiment meaningless. For example, in a study of 71 randomized trials, it was found that

50 of them had less than 90% power of detecting a 50% improvement. Thus if a surgical procedure were to result in the survival of 50% or fewer of the patients, the hypothesis test would not detect this as statistically significant.^[7]

The main statistical techniques used to avoid the problem of small sample size is the use of meta-analysis. The results from different studies are combined to increase the power so that the conclusions will be definitive. However, when the results from meta-analyses were compared to the results of subsequent large trials, it was found that they differed over 30% of the time. Another method more recently employed has been to set up national depositories.^[13] Data from all parts of the country are collected to be examined. The data lack quality control and randomization and so the results must be viewed as tentative. A test for equivalency remains a serious problem. The best solution might be to make the test on the overall population and then to examine various subpopulations by estimating outcomes rather than using inferential techniques.

EXAMPLES FROM THE MEDICAL LITERATURE

Comparison of Subgroups Within the Population

Ref. [14] examined a subgroup of the general population, sampling 127 African Americans. They were divided into two experimental groups and one control group. Of the 127, 16 were lost to follow-up. Using a 5% level of significance and 80% power, this sample of 111 is sufficient to detect a difference of 20% between groups.^[12]

Identification of Subgroups to Investigate

These studies are generally conducted retrospectively. Subgroups are identified and tested simultaneously. The results of such studies cannot be conclusive; all must be examined by using fresh data. Ref. [15] provides a classic example of subgroup analysis. There were 1395 patients originally recruited into the study. Of the initial sample, 262 patients were identified as having mild to moderate congestive heart failure. This subgroup of patients was followed for 1 year. Using power analysis for log-rank statistics, with a 5% significance level and 80% power, this patient sample is sufficiently large to detect a 15% difference in mortality.^[16]

This is in contrast to Ref. [17], which examined the relationship between aspirin and enalapril by retrospectively adding a variable to the analysis. Although the title indicates that this is a subgroup analysis, the analysis was conducted over the entire patient base.

The hypothesis generation of subgroup analysis is clearly demonstrated in Ref. [18]. As stated in the paper:

After the minimum hematocrit during cardiopulmonary bypass had been determined to be an independent risk

factor for mortality, it was necessary to perform a cutoff point analysis to establish a specific hematocrit level at which there is a statistically significant difference in post-operative mortality. The cutoff point analysis was achieved by dividing the entire patient population into dichotomous groups using different minimum hematocrit levels during cardiopulmonary bypass as a cutoff point. All of these different variables were then entered into a multiple logistic regression model that contained the other statistically significant preoperative risk factors for risk adjustment. This cutoff point analysis is able to pinpoint a specific hematocrit level at which there is a statistically significant difference in mortality after adjusting for other preoperative risk factors.

After a specific hematocrit level was determined to be statistically significant for mortality, a subgroup analysis was then performed to distinguish the difference between high-risk and low-risk populations. This subgroup analysis was performed by dividing the entire patient population into high-risk and low-risk subgroups. The high-risk population consists of patients with at least one statistically significant preoperative risk factor for mortality, and the low-risk population consists of patients without any statistically significant risk factors for mortality. Minimum hematocrit level achieved during cardiopulmonary bypass was again examined as an independent risk factor for mortality in each of the subgroups. Cutoff point analysis was again performed.

Cutpoint analysis as just discussed is an attempt to find the optimal hypothesis. Unfortunately, the study did not yield a conclusive result by testing an additional set of subjects. Therefore any conclusions in the paper must be tentative.

All three of these examples have to deal with the issues of finding significant subgroups where none exists. All should be regarded as conjectures in need of validation.

Mammography Screening

One of the more controversial of the subgroup analyses deals with the mammography screening of women ages 40–49 years. The major studies conducted were not specifically designed to address the question; a cutpoint was

considered at age 50 years. Consider two such studies^[19] (Table 1). Note first the extremely large number of observations in both studies compared to the small number of deaths. A subgroup analysis reduces the number of deaths still further. The HIP study did find a significant difference in mortality for women ages 40–49 years, demonstrating a benefit to mammography; the SNB study did not. As a result of these studies, Ref. [20] made the following conclusion:

Decisions must be made concerning the diversion of resources that would be required to implement mass screening for breast cancer. On the positive side of the balance sheet, there is a marginal reduction in deaths from breast cancer in older women. On the negative side are the financial cost and the induced morbidity. Negative factors also include the false-positive results leading to unnecessary operations, the false-negative results that lead to inappropriate reassurance, the raised level of anxiety in the female population, and the tiny but real risk of radiation-induced cancers.

The calculations presented in this study, with only inappropriate surgical intervention used as the index of harm, demonstrate a harm/benefit ratio of between 21:1 and 62:1. Enough data are now available to suggest that the American Cancer Society recommendations should be ignored in view of the net harm caused by screening for breast cancer.

Contrast this conclusion with the one given in Ref. [21]:

The scientific limitations of retrospective subgroup analysis notwithstanding, results show consistent point estimates of mortality reduction in women who are screened compared with controls for women aged 50 and older (average mortality reduction of approximately 30%; statistically significant mortality reduction in several individual trials), probably because of the much larger number of older women in the trials. For women aged 40 to 49 years, most point estimates show mortality reductions ranging from 22% to 49%, but three trials show no benefit and none of the trials individually shows statistically significant results.

Table 1 Results of Two Major Mammography Studies

Value	HIP 10-year mortality study		SNB 7-year mortality study	
	Number	Percent	Number	Percent
<i>Death from breast cancer</i>				
Number of women screened	31,888	100	78,085	100
Screened	146	0.458	87	0.111
Control	192	0.602	125	0.160
Reduction in death from breast cancer	46 (24% relative)	0.144	38 (30% relative)	0.049
<i>Death from all causes</i>				
Screened	2,235	7.009	4,846	6.206
Control	2,116	6.636	4,787	6.131
Increase in death from all causes	119 (6% relative)	0.373	59 (1% relative)	0.076

HIP = Health Insurance Plan of New York; SNB = Swedish National Board of Health and Welfare.

Table 2 Meta-Analysis of Mammography Studies

Trial	Years of follow-up	Study group deaths	Study group person-years	Control group deaths	Control group person-years	Relative mortality	95% confidence interval
Two-country	15	45	264,059	39	207,725	0.91	0.59–1.39
Malmö	15	15	61,000	23	62,000	0.66	0.34–1.27
Stockholm	12	25	173,866	12	87,826	1.05	0.53–2.09
Gothenburg	10	19	106,000	37	129,000	0.62	0.36–1.08
Edinburgh	10	25	97,206	31	88,766	0.73	0.43–1.25
HIP	18	49	248,454	65	253,085	0.77	0.53–1.11
NBSS	10	73	252,060	66	251,814	1.10	0.78–1.54
Turku	7	3	48,068	9	41,532	0.29	0.07–1.08

Thousands of lives are probably lost each year because women are not being screened. We believe that it is much more prudent to endorse mammographic screening now, risking the unlikely subsequent determination that the effort was ineffective, than to withhold screening until it is determined whether “proof” of efficacy will be obtained, risking the loss of so many lives.

A meta-analysis was performed on the eight major studies of mammography^[22] (Table 2). The study concludes:

The decision whether or not to screen this age group will depend on subjective views of the importance of the benefit and the ability to provide the resources involved. It seems clear, however, that while the size and timing of the mortality reduction require further research, the existence of such a reduction is no longer in question.

This confirms the result of a similar meta-analysis.^[23]

CONCLUSION

Subgroup analysis can and should be performed when examining the results of both clinical trials and observational studies. Some of the subgroups within the population are obvious and should be collected in the course of the study; however, others are more subtle and often are examined because of the trial outcomes. Because subgroup analysis can suffer from lack of statistical power and also from too many subgroup investigations, the results should be considered tentative pending further studies. These studies can take the form of large trials with sufficient power to be conclusive or of meta-analysis. Meta-analysis and large trials disagree in outcome approximately 30% of the time. Risks, benefits, and costs should all be examined when taking the outcomes of subgroup analysis to apply to clinical practice.

REFERENCES

- Ness, R.B.; Nelson, D.B.; Kumanyika, S.K.; Grisso, J.A. Evaluating minority recruitment into clinical studies: How good are the data? *Ann. Epidemiol.* **1997**, *7*, 472–478.
- Caraco, Y.; Lagerstrom, P.O.; Wood, A.J.J. Ethnic and genetic determinants of omeprazole disposition and effect. *Clin. Pharmacol. Ther.* **1996**, *60*, 157–167.
- Freeman, H.P. The meaning of race in science—Considerations for cancer research. *Cancer* **1998**, *82*, 219–225.
- McLachlan, G.J.; Basford, K.E. Mixture Models. In *Inference and Applications to Clustering*; Statistics: Textbooks and Monographs, Marcel Dekker: New York, 1988, Vol. 84.
- Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Chapman and Hall: London, 1986.
- Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. The KDD process for extracting useful knowledge from volumes of data. *Commun. ACM* **1996**, *39*, 27–34.
- Moses, L.E. The Series of Consecutive Cases as a Device for Assessing Outcomes of Intervention. In *Medical Uses of Statistics*; Bailer, J.C., Mosteller, F., Eds.; Massachusetts Medical Society: Waltham, MA, 1986.
- ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomized trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet* **1988**, *2*, 349–360.
- Balch, C.M., et al. Efficacy of an elective regional lymph node dissection of 1 to 4 mm thick melanomas for patients 60 years of age and younger. *Ann. Surg.* **1996**, *224*, 255–266.
- Wallack, M.K., et al. Increased survival of patients treated with a vaccinia melanoma oncolysate vaccine. *Ann. Surg.* **1997**, *226*, 198–206.
- Sickles, E.A.; Kopans, D.B. Mammographic screening for women aged 40 to 49 years: The primary care practitioner's dilemma. *Ann. Intern. Med.* **1995**, *122*, 534–538.
- Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd Ed.; Lawrence Erlbaum Associates: Hillsdale, NJ, 1988.
- Shroyer, A.L.; Edwards, F.H.; Grover, F.L. Updates to the data quality review program: The Society of Thoracic Surgeons' adult cardiac national database. *Ann. Thorac. Surg.* **1998**, *65*, 1494–1497.
- Alexander, C.N.; Schneider, R.H., et al. Trial of stress reduction for hypertension in older African Americans: II. Sex and risk subgroup analysis. *Hypertension* **1996**, *28*, 228–237.
- Herlitz, Z.J.; Waagstein, F., et al. Effect of metoprolol on the prognosis for patients with suspected acute myocardial infarction and indirect signs of congestive heart failure (a subgroup analysis of the Goteborg Metoprolol trial). *Am. J. Cardiol.* **1997**, *80*, 40J–44J.

16. Freedman, L.S. Tables of the number of patients required in clinical trials using the log-rank test. *Stat. Med.* **1982**, *1*, 121–129.
17. Nguyen, K.N.; Aursnes, I.; Kjekshus, J. Interaction between Enalapril and aspirin on mortality after acute myocardial infarction: Subgroup analysis of the cooperative new Scandinavian Enalapril Survival Study II (Consensus II). *Am. J. Cardiol.* **1997**, *79*, 115–119.
18. Fang, W.C.; Helm, R.E.; et al. Impact of minimum hematocrit during cardiopulmonary bypass on mortality in patients undergoing coronary artery surgery. *Circulation* **1997**, *96* (Suppl. II), II-194–II-199.
19. Wright, C.J. Breast cancer screening: A different look at the evidence. *Surgery* **1986**, *100*, 594–598.
20. Tabar, L.; Fagerbert, C.J.; et al. Reduction in mortality from breast cancer after mass screening with mammography. *Lancet* **1985**, *1*, 829–832.
21. Falun Meeting. Breast-cancer screening with mammography in women aged 40–49 years. Report of the organizing committee and collaborators. *Int. J. Cancer* **1996**, *68*, 693–699.
22. Smart, C.R.; Hendrick, R.E.; Rutledge, J.H., III; Smith, R.A. Benefit of mammograph screening in women ages 40 to 49 years. *Cancer* **1995**, *75*, 1619–1626.
23. Feig, S.A. Mammographic screening of women aged 40–49 years. Benefit, risk and cost considerations. *Cancer* **1995**, *76*, 2097–2106. (Suppl.).

Subject-Treatment Interaction

Gary L. Gadbury

Department of Statistics, Kansas State University, Manhattan, Kansas, U.S.A.

Abstract

The focus of this entry is subject-treatment interaction as a variability in the magnitude of true individual effects of a treatment and, when such variability is large, a practical consequence may be that the effect of treatment for some individuals has important differences from the average effect.

Keywords: Causation; Clinical trials; Individual effects; Potential outcomes; Treatment heterogeneity.

INTRODUCTION

That the effect of a treatment will vary among subjects is not surprising, nor is it a recent concept. Roses^[1] provided an 1892 quote by Sir William Osler, “If it were not for the great variability among individuals medicine might as well be a science and not an art.” Subject-treatment (S-T) interaction is, as the term implies, an interaction of specific subjects with applied treatment(s). The result of such interaction is a variability of “individual treatment effects” or “individual treatment heterogeneity” in a population of interest. Effects of treatment may be assessed as a measurement of efficacy or toxicity because variability in either may have important consequences. Although such variability has often been acknowledged as an important consideration, medicine today generally makes use of statistical information gathered about the general population (often about the “average” subject) and then applies it to the individual.^[2] One exception is the interest in individual bioequivalence for drug “switchability” (cf., Ref. [3]) where effects of subject by formulation interaction are investigated, typically in crossover designs (e.g., Refs. [4–6]). S-T interaction in a crossover design is a topic considered in a later section.

When there is a high degree of S-T interaction in a population, there may be a nonnegligible proportion of the population responding differently to a treatment, and possibly in the opposite direction, from the average subject. Even within a carefully designed clinical trial, the validity of results may be compromised if there is a wide variability of individual treatment effects, prompting at least one author to suggest that inference about the mean treatment effect be supplemented with information about suspected variability of effects.^[7] Variability in magnitude has been called a noncrossover interaction as long as the direction of the effect is the same across subjects, and variation in direction of individual effects has been termed a

crossover interaction.^[8] The latter is usually of more concern to researchers, possibly playing a role in adverse drug reactions.^[9]

Advances in the field of pharmacogenomics have raised hopes that genetic information may be used to identify subjects who will “succeed” with a specific treatment.^[10] Recent high-throughput technologies such as microarrays have enhanced the ability to search for such genetic contributors.^[11,12] Gene-treatment interactions are only one component of S-T interaction. Other components may relate to sex, age, environment, other medications, etc.,^[13,14] thus S-T interaction is an upper bound for gene-treatment interactions.^[15] How much S-T interaction is explained by gene-treatment interactions is a nontrivial problem as clinical studies are often not designed to evaluate S-T interaction.^[15]

The degree of S-T interaction in a population can be quantified by a parameter that is related to the variance of individual treatment effects, although this parameter cannot be directly estimated using observable data.^[16] Sometimes a comparison between a proportion of subjects benefiting on a new treatment and the proportion benefiting on a standard treatment is interpreted as being related to the probability that an individual subject will benefit on the new treatment. This is not quite true. Even with crossover trials, identification of S-T interaction depends on how one defines a “true individual treatment effect.” However, S-T interaction can produce effects that are testable using such data. Methods to detect these effects have focused on identifying subsets or covariates to explain perceived treatment heterogeneity.^[8,17–20]

Kent and Hayward^[21] comment that “baseline risk” across subjects in a study may be highly skewed, thus nearly ensuring that the summary results (e.g., estimates of average effects) will be different from the typical patient. They, as well as Kent et al.,^[22] suggest that baseline risk and competing risk heterogeneity may explain treatment

effect heterogeneity, and stratifying on metrics derived from these risks can enhance detection of subgroups that respond differently to treatments. They propose a risk stratified analysis (versus a conventional subgroup analysis) for detecting important differences in treatment effects, and stress the importance of understanding the complex relationship among baseline risk, treatment related harm, and competing risk when making good individual-patient recommendations and decisions.

One method to detect the presence of S-T interaction is to examine the proportion of similar responses between the treatment outcome (say, X) and the control outcome (say, Y). This overlap has been called the proportion of similar responses (PSR).^[23] Given two density functions $f(x)$ and $g(x)$, the PSR is defined as the area under the smaller of the two population density functions and is expressed as

$$PSR(f, g) = \int \min(f(x), g(x)) dx.$$

Stine and Heyse^[24] have developed a non-parametric estimate of the PSR by using kernel density estimators to estimate both $f(x)$ and $g(x)$. Note that the kernel density estimates are made on the marginal densities of the two outcome variables without consideration of the joint distribution. As such, the value of the (non-estimable) correlation parameter, ρ_{xy} , cannot be included in the calculation of the PSR. Whether the marginal distributions and, thus, the degree of overlap between them (i.e., PSR) applies to the population of interest is another matter because patients in clinical trials are often not representative of the target populations.^[25] The issue of sample selection in a clinical study is not directly considered herein, nor are the various issues that arise in clinical trials such as compliance and dropout that pose challenges for even estimating average effects (c.f., Ref. [26]).

OVERVIEW

The focus of this entry is S-T interaction as a variability in the magnitude of true individual effects of a treatment and, when such variability is large, a practical consequence may be that the effect of treatment for some individuals has important differences from the average effect. This focus is consistent with that used in earlier papers.^[16,27-9] The goals of the entry are the following: Define a true individual treatment effect; quantify S-T interaction as a nonestimable population parameter; quantify the probability that an “individual” will experience an “unfavorable” effect of treatment and show that this probability is a function of S-T interaction (hence also nonestimable); show some connections between S-T interaction and causation; demonstrate what can and cannot be learned about S-T interaction using observable data; and finally, suggest design modifications or extensions that may facilitate a more precise evaluation

of the magnitude of S-T interaction (herein, “nonestimable” means that observable data contain insufficient information to allow direct estimation of a parameter). Details will be provided for a two-sample completely randomized design; however, some discussion is included later for matched-pairs and crossover designs. It is hoped that the entry will at least alert the researcher to the assumptions that must be made when equating variability of “perceived” individual treatment effects with variability of true individual effects.

COMPLETELY RANDOMIZED DESIGN COMPARING TWO TREATMENTS

Individual Treatment Effects and Potential Outcomes

Suppose two treatments, T and C , are being compared in a two-sample completely randomized design. Further, suppose that at a particular point in time, the value of an outcome variable, Y , will be measured. This outcome may be a change from baseline, and it may be quantitative or dichotomous (i.e., success or failure). For a subject receiving treatment T , the outcome $Y^{(T)}$ is observable, and for a subject receiving treatment C the outcome $Y^{(C)}$ is observable. The bivariate pair $[Y^{(T)}, Y^{(C)}]$ are potential outcomes^[30,31] in that only one of the two is observable for a specific subject at a particular time.

To avoid the need for the superscripts and thus later simplifying notation, hereafter X is the outcome to treatment T and Y is the outcome to treatment C . The bivariate pair (X, Y) are potential outcomes, and the variable $D = X - Y$ defines a true individual treatment effect. Although D cannot be observed for any subject, its definition helps to conceptualize what is meant by a true individual effect and thus clarify what can and cannot be learned about these effects using observable data. The potential outcomes framework has found use in the area of statistical causality because the ultimate interest in causal inference is effects of causes (i.e., treatments) on specific subjects. (p. 947)^[32] The “fundamental problem of causal inference” is that only one of the two potential outcomes, X or Y , can be observed for an individual subject.^[32] One can imagine a study evaluating k treatments, and the potential outcomes would be a vector containing k outcomes (rather than two); only one of the k outcomes would be observable for a given subject at a particular time.

It is the average treatment effect, $E(D)$, that is usually of interest but it is the variance, $\text{Var}(D)$, that quantifies the degree of variability of individual treatment effects and hence the magnitude of S-T interaction. S-T interaction is present when $\text{Var}(D) > 0$, but this quantity cannot be directly estimated using observed data because of the fundamental problem of causal inference mentioned earlier; however, bounds for it can be derived and these can be estimated.

A Model for Continuous Outcomes and Consequences of S-T Interaction

Consider N subjects in a two-sample design and the variables X and Y represent outcomes to treatments T and C , respectively. The set of N potential outcomes has the form given below (left) that, after treatment assignment produces observed outcomes of the form shown (right), and where the “?” represents an unobservable potential outcome.

$$\begin{pmatrix} X_1 & Y_1 \\ \vdots & \vdots \\ X_N & Y_N \end{pmatrix} \xrightarrow{\text{Treatment Assignment}} \begin{pmatrix} X_1 & ? \\ ? & Y_2 \\ \vdots & \vdots \\ ? & Y_{N-1} \\ X_N & ? \end{pmatrix} \quad (1)$$

An assumption is needed that a subject's outcome will be unaffected by the treatment assignment outcome for other subjects in the study. This was termed no interference between units^[33] and generalized later as a stable unit treatment value assumption (SUTVA).^[34]

One could define a treatment indicator variable and develop an estimate of $E(D)$ and its standard error with respect to a finite population randomization distribution.^[28,35] Here we will suppose an infinite population model, and potential outcomes (X_i, Y_i) , $i = 1, \dots, N$ are a random sample from a bivariate normal distribution with mean vector (μ_x, μ_y) , variances, σ_x^2 and σ_y^2 , and correlation ρ_{xy} . Selection bias associated with a nonprobabilistic recruitment process are beyond the scope of this entry and is discussed elsewhere.^[36] After a random treatment assignment, we have the usual two-group parallel design and a mean treatment effect, $\mu_D = \mu_x - \mu_y$, and standard error are easily estimated using observed data because they are functions of parameters in the marginal distributions of X and of Y . The quantity for S-T interaction, $\sigma_D^2 = \text{Var}(D) = \text{Var}(X - Y)$, cannot be directly estimated because $\sigma_D^2 = \sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y\rho_{xy}$ and there is no available information about ρ_{xy} in observable data; that is, it is nonestimable.

One can write σ_D^2 as follows:

$$\sigma_D^2 = (\sigma_x - \sigma_y)^2 + 2\sigma_x\sigma_y(1 - \rho_{xy}) \quad (2)$$

So there is S-T interaction present unless $\sigma_x = \sigma_y$ and $\rho_{xy} = 1$, and the former condition can be tested using observed data but the latter cannot. Whether or not S-T interaction produces unfavorable consequences on a subset of the population depends upon how large σ_D is with respect to μ_D , a quantity that depends on ρ_{xy} . As an illustration, assume $\sigma_x = \sigma_y = \sigma$, that there are n subjects in each of two treatment groups (so $N = 2n$), and that a positive μ_D above some threshold is a “beneficial average treatment effect” (without loss of generality, assume this threshold is zero). Letting $\tau = \mu_D/\sigma_D$, the power to detect a positive average treatment

effect is $\text{Power} = 1 - T(t_{1-\alpha}, 2n-2, \tau\sqrt{n/2})$, where $T(t_0, k, \lambda)$ is a cumulative distribution function of a noncentral t random variable with k degrees of freedom and noncentrality parameter λ , evaluated at t_0 . The proportion of the population experiencing a negative effect of treatment T with respect to C is $P_- = P(D < 0) = \Phi(-\tau/\sqrt{2(1-\rho_{xy})})$. Fig. 1 shows the relationship between particular ranges of τ (top horizontal axis), P_- (bottom horizontal axis), and Power (vertical axis) when the nonestimable $\rho_{xy} = 0.7$. For example, one may have high power (>80%) to detect a positive average treatment effect when more than 10% of the population experience a negative effect, unless μ_D is 1.3 times larger than σ when $n = 20$. The issue is more pronounced for larger sample sizes and for smaller values of ρ_{xy} .

In practice, one will not know the true value of ρ_{xy} and, so, will be unable to directly evaluate the size of σ_D with respect to μ_D . Little else can be done with observable data except to note that letting $\rho_{xy} = 1$ and -1 produces bounds for σ_D^2 that can be estimated.^[16] In a data example, Gadbury, Iyer, and Allison^[27] assessed the sensitivity of estimated σ_D^2 and its consequence, $P(D < 0)$, to a range of values for ρ_{xy} using maximum likelihood estimation. Sampling variability was assessed using large sample properties of maximum likelihood estimators (MLEs) and also using a bootstrap technique for smaller samples.

A Model for Binary Outcomes and Consequences of S-T Interaction

Gadbury, Iyer, and Albert^[29] showed the following multinomial population model for binary potential outcomes:

$$\begin{array}{ccccc} (x, y) & (0, 0) & (0, 1) & (1, 0) & (1, 1) \\ P(X = x, Y = y) & \pi_1 & \pi_2 & \pi_3 & \pi_4 \end{array} \quad (3)$$

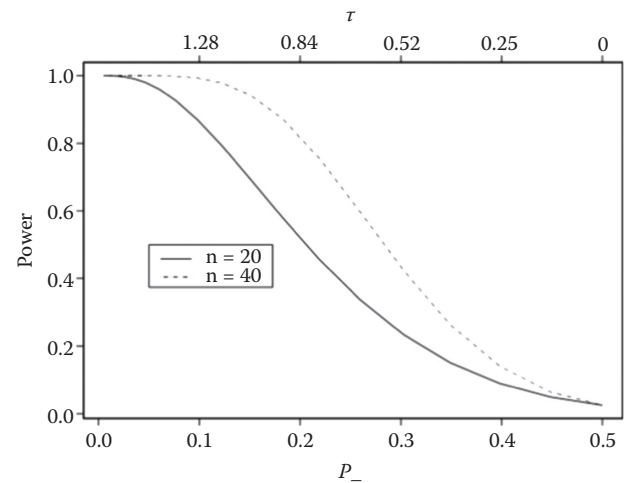


Fig. 1 Consequences of S-T interaction in a bivariate population when $\sigma_x = \sigma_y = \sigma$, $\tau = \mu_D/\sigma$ and $\rho_{xy} = 0.7$. There are n subjects in each of the two treatment groups

where $\sum_i \pi_i = 1$. Suppose that an outcome equal to 1 is a success and 0, a failure (such outcomes may also reflect a collapsing of continuous outcomes into two categories, improvement from baseline or no improvement). At a particular time, treatment T is a success with respect to C for a proportion π_3 , and $D = 1$ for these subjects. However, treatment C is superior to T for a proportion π_2 , and $D = -1$ for this subset. There is no difference, $D = 0$, for a proportion $\pi_1 + \pi_4$. S-T interaction is present in the population unless one of the three quantities is equal to 1: π_3 , π_2 , or $\pi_1 + \pi_4$. However, the individual parameters π_i , $i = 1, \dots, 4$ are nonestimable in observable data and only parameters in the marginal distributions can be estimated. The mean treatment effect is $E(D) = E(X - Y) = \pi_3 - \pi_2$ and it is this quantity that is estimated in the usual comparison of two population proportions. Suppose $E(D) = 0.4$, then it could be that $\pi_3 = 0.4$ and $\pi_2 = 0.0$, or $\pi_3 = 0.7$ and $\pi_2 = 0.3$. In the latter case, 30% of the population would be better off with treatment C although treatment T was superior on average. With no additional information, the most that can be done is to bound the individual parameters with quantities that can be estimated.^[29] For example, estimable bounds for π_2 are given by the following inequalities:^[29]

$$\max(0, \pi_2 - \pi_3) \leq \pi_2 \leq \min\{1 - (\pi_3 + \pi_4), \pi_2 + \pi_4\} \quad (4)$$

The Role of a Covariate

Suppose that a covariate, Z , is observable for all subjects in a study and that Z is unaffected by treatment. Examples are a subject's age, baseline outcome, gender, or a particular genotype. Assume that, as before, two treatments are being compared and either X or Y is observed for a subject depending on treatment assignment outcome. A well-known role for Z is to improve the efficiency in estimating a mean treatment effect. It can also be used to refine the estimable bounds for σ_D^2 .

Suppose that (X, Y, Z) is a trivariate normal random variable and let β_{XZ} and β_{YZ} be the population regression coefficients relating Z to X and to Y , respectively. Let $\sigma_{X|Z}$ and $\sigma_{Y|Z}$ be the standard deviations of the conditional distributions of X and of Y , respectively, given Z , let $\rho_{XY|Z}$ be the partial correlation between X and Y given Z , and σ_Z^2 be the variance of the marginal distribution of Z . Then the following can be shown.

$$\sigma_D^2 = (\sigma_{X|Z} - \sigma_{Y|Z})^2 + (\beta_{XZ} - \beta_{YZ})^2 \sigma_Z^2 + 2\sigma_{X|Z}\sigma_{Y|Z}(1 - \rho_{XY|Z}) \quad (5)$$

Several points can be made regarding Eq. 5. First, all parameters in Eq. 5 can be estimated using observed data except the partial correlation. Second, if Z is a perfect linear predictor of X and of Y , then all S-T interaction is explained by a covariate-treatment interaction as $\sigma_{X|Z} = \sigma_{Y|Z} = 0$. Third, if this is not the case, then $\sigma_D^2 = 0$ if

and only if $\sigma_{X|Z} = \sigma_{Y|Z}$, $\beta_{XZ} = \beta_{YZ}$, and $\rho_{XY|Z} = 1$, and the last equality cannot be tested using observed data. Fourth, letting $\rho_{XY|Z} = -1$ and 1 results in upper and lower bounds for σ_D^2 that can be estimated. Interestingly, these bounds are exactly those that result from the positive definiteness requirement of the three-dimensional correlation matrix for (X, Y, Z) .^[16] Gadbury and Iyer^[16] developed MLEs for these bounds and derived their large sample properties. Gadbury, Iyer, and Allison^[27] showed an example where data suggested a covariate by treatment interaction (i.e., $\beta_{XZ} \neq \beta_{YZ}$) and that subjects with higher baseline blood pressure appeared to have larger decreases in blood pressure after 12 weeks on a calcium supplement vs. the placebo group. They noted that although some S-T interaction is explained by $\beta_{XZ} \neq \beta_{YZ}$, there can remain individual treatment heterogeneity in subpopulations defined by values of the covariate Z . Conditioning on a value of Z , similar results for the two-sample problem (without the covariate) can now be applied to the conditional bivariate distribution of X and Y given $Z = z$, where the mean treatment effect is $\mu_{D|Z=z} = \mu_X - \mu_Y + (\beta_{XZ} - \beta_{YZ})(z - \mu_Z)$ and the conditional variance of D is $\sigma_{D|Z}^2 = \sigma_{X|Z}^2 + \sigma_{Y|Z}^2 - 2\sigma_{X|Z}\sigma_{Y|Z}\rho_{XY|Z}$. Gadbury, Iyer, and Allison^[27] then, for varying $\rho_{XY|Z}$ between -1 and 1, assessed the sensitivity of MLEs for $\sigma_{D|Z}$ and associated MLEs for the probability of a negative treatment effect for a subpopulation of subjects defined by a value of $Z = z$, i.e., $P(D < 0 | Z = z)$. Large sample properties of MLEs were used to compute confidence intervals.

Working out the details of using Z as a predictor for binary outcomes and the associated refinement of bounds for π_2 is a subject for future work. Another use for Z , or any other subjective information regarding the subjects in a study, is to use this information to match subjects into pairs. Some discussion of this situation is given in the next section along with discussion of the two-period crossover design.

INDIVIDUAL TREATMENT EFFECTS IN TWO OTHER DESIGNS

Matched-pairs designs allow for observation of a treatment effect for a pair and computation of the variance of paired differences. Whether the paired differences equal true effects for individuals depends on the quality of the matching criteria. Crossover designs allow for a treatment effect to be observed for each individual because each individual receives each of the two treatments separated by a washout period. The next two subsections define individual effects and S-T interaction for these two types of designs.

Matched-Pairs Design

It is assumed that subjects are matched into pairs a priori of treatment assignment using covariate information

or other subjective information thought to create homogeneity of treatment outcomes within pairs. Subjects are arbitrarily labeled within pairs as Subject 1 and Subject 2. Two treatments, T or C , are assigned within pairs via a random coin toss. There are $N = 2n$ subjects in a study resulting in n matched pairs. The potential outcomes are, again, defined using variables (X, Y) for outcomes to T and C , respectively.

Continuous Outcomes

One way to parameterize the set of $2n$ potential outcomes was shown by Gadbury.^[35] This is shown below along with one pattern of observed outcomes post treatment assignment.

$$\left(\begin{array}{cc} X_1 - \varepsilon_1 & Y_1 - \eta_1 \\ X_1 + \varepsilon_1 & Y_1 + \eta_1 \\ \text{-----} & \text{-----} \\ X_2 - \varepsilon_2 & Y_2 - \eta_2 \\ X_2 + \varepsilon_2 & Y_2 + \eta_2 \\ \text{-----} & \text{-----} \\ \vdots & \vdots \\ \text{-----} & \text{-----} \\ X_n - \varepsilon_n & Y_n - \eta_n \\ X_n + \varepsilon_n & Y_n - \eta_n \end{array} \right) \xrightarrow{\text{Treatment Assignment}} \left(\begin{array}{cc} X_1 - \varepsilon_1 & ? \\ ? & Y_2 + \eta_1 \\ \text{-----} & \text{-----} \\ ? & Y_2 - \eta_2 \\ X_2 + \varepsilon_2 & ? \\ \text{-----} & \text{-----} \\ \vdots & \vdots \\ \text{-----} & \text{-----} \\ ? & Y_n - \eta_n \\ X_n + \varepsilon_n & ? \end{array} \right) \quad (6)$$

A true individual treatment effect for Subject 1 in the i th pair is $D_{i1} = X_i - Y_i - (\varepsilon_i - \eta_i)$ and for Subject 2 it is $D_{i2} = X_i - Y_i + (\varepsilon_i - \eta_i)$. The average treatment effect for the two subjects in the i th pair is $X_i - Y_i$, and the two individual effects are not equal unless $\varepsilon_i = \eta_i$. The parameters, ε_i, η_i , $i = 1, \dots, n$, reflect the quality of the matching criteria. An observed treatment effect for the i th pair can be written as,

$$d_i = \{X_i - Y_i - (\varepsilon_i + \eta_i)\}\delta_i + \{X_i - Y_i + (\varepsilon_i + \eta_i)\}(1 - \delta_i) \quad (7)$$

where $\delta_i = 1$ implies that Subject 1 received treatment T and Subject 2 received treatment C ; if $\delta_i = 0$, the assignment is reversed for that pair and $P(\delta_i = 1) = 1/2$ for each pair. The condition, $\varepsilon_i = \eta_i = 0$, $i = 1, \dots, n$, is sufficient for equating an observed treatment effect for a pair with the true effects for the two individuals in the pair. S-T interaction can then be directly estimated using the variance of the observed paired differences.

Binary Outcomes

Gadbury, Iyer, and Albert^[29] presented the issues involved in estimating π_2 in the population model shown in expression 3 using a matched-pairs design. There are now 16

outcomes in the potential outcomes framework, with four possible outcomes for each subject in a pair. The potential outcomes for Subjects 1 and 2 in a pair are represented by $\{(X_1, Y_1), (X_2, Y_2)\}$. Four (of the 16) outcomes reflect subjects who are perfectly matched on both variables. These outcomes are $\{(x_1, y_1), (x_2, y_2)\} = \{(0,0), (0,0)\}, \{(0,1), (0,1)\}, \{(1,0), (1,0)\},$ and $\{(1,1), (1,1)\}$. If the sum of the probabilities for these four outcomes is equal to 1, then only perfectly matched pairs are available in the matched population.

Suppose subjects have been randomly labeled within each of n pairs and, without loss of generality, Subject 1 in each pair receives treatment T and Subject 2 receives treatment C . So observable variables for a pair are of

the form (X_1, Y_2) and resulting data are of the form for the usual matched 2×2 table for binary outcomes, or can be shown as

X_1	Y_2	Frequency
0	0	s_1
0	1	s_2
1	0	s_3
1	1	s_4

(8)

where $\sum_i s_i = n$. Gadbury, Iyer, and Albert 19 showed that a sufficient condition for s_i/n to be unbiased for π_i is that the population of matched subjects are matched on at least one of the two potential outcome variables; that is, the proportion that are mismatched on both X and Y equals zero. There are four potential outcomes that are a double mismatch, but exchangeability within pairs (as a result of the random labeling) allows for these four to be expressed as two unique double mismatches: $\{(x_1, y_1), (x_2, y_2)\} = \{(0,0), (1,1)\}$ and $\{(x_1, y_1), (x_2, y_2)\} = \{(0,1), (1,0)\}$. A necessary and sufficient condition for s_i/n to be unbiased for π_i is that the probabilities of these two different occurrences of a double-mismatched pair are equal.^[29]

Of course, without additional information, one will not know how well subjects are matched in the population. One can still obtain bounds for π_i similar to what was shown earlier for the two independent sample framework. The quality of the matching criteria cannot be evaluated using observed data unless the design is altered to facilitate this. For example, some pairs could be selected so that both subjects in the pair received treatment T , and other pairs where both subjects received treatment C . Gadbury, Iyer, and Albert^[29] constructed a simulated example where $\pi_2 = 0.125$, and an estimated upper bound for π_2 without this design extension was 0.34. After using the extension to the design that provided information on the matching criteria, the estimated upper bound was tightened to 0.21.

An obvious extension to these results for matched pairs is to using covariates or other information on subjects to group subjects into homogeneous blocks. This design will likely be of more practical use to researchers. Evaluating S-T interaction in this type of design for a binary outcome variable was developed and reported in Albert et al.^[37]

A Two-Period Crossover Design

When feasible to implement, the crossover design allows a direct observation for an individual effect as outcomes for each subject are observed for both treatments. (For more discussion on this, see Ref. [15] and for more details on crossover designs in general, see Ref. [38], or [39] for a review of papers on crossover designs.) As mentioned earlier, these designs have been used in individual bioequivalence studies. Whether or not these observed individual effects are true individual effects depends on how one defines a true individual effect. Suppose that $N = 2n$ subjects are in a trial and that during the first time period, n will be randomly selected to receive treatment T with the other n receiving C (this balance in the design is not necessary but is simply used for convenience). At some point in time (time 1), outcomes are observed. After a washout period, treatment assignments are reversed and at some second point in time (Time 2), outcomes are observed. The two sequences can be abbreviated TC (i.e., treatment T in period one and control C in period 2) and CT (control C in period 1 and treatment T in period 2). It is assumed that the washout period is sufficient so that there are no carryover effects of treatments at Time 1 to Time 2. Potential outcomes for this scenario were shown by Gadbury^[35] and are repeated below.

Subject	Time 1		Time 2	
1	$X_1 - t_1$	$Y_1 - \tau_1$	$X_1 + t_1$	$Y_1 + \tau_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$2n$	$X_{2n} - t_{2n}$	$Y_{2n} - \tau_{2n}$	$X_{2n} + t_{2n}$	$Y_{2n} + \tau_{2n}$

(9)

There is a true individual treatment effect for each subject at each time period. The parameters t, τ reflect time effects from Time 1 to Time 2. It may be reasonable to define a true

individual treatment effect for the i th subject as a combination of the two effects at each time period. The average will be used here so the true treatment effect for the i th subject is $D_i = (1/2)\{(X_i - t_1) - (Y_i - \tau_1) + (X_i + t_2) - (Y_i + \tau_2)\} = X_i - Y_i$. Let $\delta_i = 1$ if the i th subject is assigned the first treatment sequence, T then C , and $\delta_i = 0$ if assigned the reverse sequence, C then T . The observed treatment effect for the i th subject is $d_i = \{(X_i - t_1) - (Y_i + \tau_2)\}\delta_i + \{(X_i + t_2) - (Y_i - \tau_1)\}(1 - \delta_i)$. A sufficient condition for this observed effect to equal the true effect is that $(t_1 + \tau_2) = 0$, meaning that the sequence to which the subject was assigned had no effect on outcomes. Bounds for the variance of individual treatment effects, σ_D^2 , will depend on nonestimable parameters in the distribution of time effects, particularly the variance-covariance matrix of (t, τ) . If a Balaam design is used, where some subjects are allocated to receive the same treatment in both time periods (i.e., include the sequences TT and CC in the above design), then variances of the time effect parameters can be estimated but the correlation cannot.^[40] Still, bounds for σ_D^2 should be estimable by letting the correlation range from -1 to 1 .

CONCLUSION

This entry highlighted the role of potential outcomes for defining individual treatment effects and quantifying S-T interaction. Three designs were discussed and a few details were provided for the two-sample completely randomized design. Evaluating the degree of S-T interaction, and its consequences, is a challenging problem and often only bounds can be derived that can be directly estimated using data. The ability to estimate S-T interaction requires assumptions about the degree to which “observed individual treatment effects” are equal to true individual effects. These assumptions may be perfectly reasonable in many applications where much is known about the treatment and/or the disease that is being treated. Such clinical knowledge and past experience are often outside the realm of what data can show in a single experiment. This entry demonstrated some limitations of using data alone to evaluate individual treatment heterogeneity. It is hoped that this will help illuminate the assumptions that are made when equating observed heterogeneity with true heterogeneity and, thus, help investigators to evaluate the reasonableness of such assumptions in particular contexts. Perhaps new designs or extensions to current designs will be considered for clinical trials with an added goal in mind, as Longford remarked, “... inference about the mean treatment effect be supplemented by inference about the variation of the treatment effects”. (p. 1473)^[7]

ACKNOWLEDGMENTS

The author thanks Hari K. Iyer for reviewing an earlier draft of this entry and Robert Poulson and Roshan Serasinghe for assisting with updates.

REFERENCES

- Roses, A.D. Pharmacogenetics and the practice of medicine. *Science* **2000**, *405*, 857–865.
- Marshall, A. Laying the foundations for personalized medicines. *Nat. Biotechnol.* **1997**, *15*, 954–957.
- Chow, S.-C.; Liu, J.-P. Preface. *Stat. Med.* **2000**, *19*, 2719.
- Enrenyi, L.; Tothfalusi, L. Subject-by-formulation interaction in determinations of individual bioequivalence: bias and prevalence. *Pharm. Res.* **1999**, *16*, 186–190.
- Gould, A.L. A practical approach for evaluating population and individual bioequivalence. *Stat. Med.* **2000**, *19*, 2721–2740.
- Chow, S.-C.; Shao, J.; Wang, H. Individual bioequivalence testing under 2×3 designs. *Stat. Med.* **2002**, *21*, 629–648.
- Longford, N. Selection bias and treatment heterogeneity in clinical trials. *Stat. Med.* **1999**, *18*, 1467–1474.
- Gail, M.; Simon, R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* **1985**, *41*, 361–372.
- Phillips, K.A.; Veenstra, D.L.; Oren, E.; Lee, J.K.; Sadee, W. Potential role of pharmacogenomics in reducing adverse drug reactions. *JAMA* **2001**, *286*, 2270–2279.
- Evans, W.E.; Relling, M.V. Pharmacogenomics: Translating functional genomics into rational therapeutics. *Science* **1999**, *286*, 487–491.
- Gunther, E.; Stone, D.J.; Gerwein, R.W.; Bento, P.; Heyes, M.P. Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 9608–9613.
- Allison, D. Statistical Methods for Microarray Research for Drug Target Identification. *2002 Proceedings of the American Statistical Association [CD ROM]*; American Statistical Association: Alexandria, VA, 2002.
- Evans, W.E.; Johnson, J. Pharmacogenomics: The inherited basis for interindividual differences in drug response. *Annu. Rev. Genomics Hum. Genet.* **2001**, *2*, 9–39.
- Wilson, J.F.; Weale, M.E.; Smith, A.C.; Gratrix, F.; Fletcher, B.; Thomas, M.G.; Bradman, N.; Goldstein, D.B. Population genetic structure of variable drug response. *Nat. Genet.* **2001**, *29*, 265–269.
- Senn, S. Individual therapy: New dawn or false dawn? *Drug Inf. J.* **2001**, *35*, 1479–1494.
- Gadbury, G.L.; Iyer, H.K. Unit-treatment interaction and its practical consequences. *Biometrics* **2000**, *56*, 882–885.
- Good, P.I. Detection of a treatment effect when not all experimental subjects will respond to treatment. *Biometrics* **1979**, *35*, 483–489.
- Simon, R. Patient subsets and variation in therapeutic efficacy. *Br. J. Clin. Pharmacol.* **1982**, *14*, 473–482.
- Conover, W.J.; Salsburg, D.S. Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to “respond” to treatment. *Biometrics* **1988**, *44*, 189–196.
- Horwitz, R.I.; Singer, B.H.; Makuch, R.W.; Viscoli, C.M. Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs of clinical inquiry and drug regulation. *J. Clin. Epidemiol.* **1996**, *49*, 395–400.
- Kent, D.M.; Hayward, R.A. Limitations of applying summary results of clinical trials to individual patients. *JAMA* **2007**, *298*, 1209–1212.
- Kent, D.M.; Ali, A.A.; Hayward, R.A. Competing risk and heterogeneity of treatment effect in clinical trials. *Trials* **2008**, *9*, 30.
- Rom, D.M.; Hwang, E. Testing for individual and population equivalence based on the proportion of similar responses. *Stat. Med.* **1996**, *15*, 1489–1505.
- Stine, R.A.; Heyse, J.F. Non-parametric estimates of overlap. *Stat. Med.* **2001**, *20*, 215–236.
- Senn, S. Addes values: controversies concerning randomization and additivity in clinical trials. *Stat. Med.* **2004**, *23*, 3729–3753.
- McNamee, R. Intention to treat, per protocol, as treated and instrumental variable estimators given non-compliance and effect heterogeneity. *Stat. Med.* **2009**, *28*, 2639–2652.
- Gadbury, G.L.; Iyer, H.K.; Allison, D. Evaluating subject-treatment interaction when comparing two treatments. *J. Biopharm. Stat.* **2001**, *11*, 313–333.
- Gadbury, G.L.; Iyer, H.K. On Estimating Subject-Treatment Interaction. *Advances on Methodological and Applied Aspects of Probability and Statistics*; Taylor and Francis: New York, 2002, 349–364.
- Gadbury, G.L.; Iyer, H.K.; Albert, J. Individual treatment effects in randomized trials with binary outcomes. *J. Stat. Plan.*, **2004**, *121*, 163–174.
- Neyman, J. On the application of probability theory to agricultural experiments. Essay on principles. *Rocz. Nauk Rol. X* (in Polish) **1923**, 1–51.—(English translation of Section 9 by D. M. T. P. Dabrowska, *Speed, Stat. Sci.* **1990**, *5*, 465–480).
- Rubin, D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Br. J. Educ. Psychol.* **1974**, *66*, 688–701.
- Holland, P.W. Statistics and causal inference. *J. Am. Stat. Assoc.* **1986**, *81*, 945–960.
- Cox, D.R. *The Planning of Experiments*; Wiley: New York, 1958; 19.
- Rubin, D.B. Comment on Basu’s randomization analysis of experimental data. *J. Am. Stat. Assoc.* **1980**, *75*, 591–593.
- Gadbury, G. Randomization inference and bias of standard errors. *Am. Stat.* **2001**, *55*, 310–313.
- Longford, N.T.; Nelder, J.A. Statistics versus statistical science in the regulatory process. *Stat. Med.* **1999**, *18*, 2311–2320.
- Albert, J.M.; Gadbury, G.L.; Mascha, E.J. Assessing treatment effect heterogeneity in clinical trials with blocked binary outcomes. *Biom. J.* **2005**, *47*, 662–673.
- Senn, S. *Cross-Over Trials in Clinical Research*; John Wiley and Sons Ltd.: Chichester, 1993.
- Senn, S. Cross-over trials in *Statistics in Medicine: the first ‘25’ years*. *Stat. Med.* **2006**, *25*, 3430–3442.
- Ndum, E. *Individual Treatment Effect Heterogeneity in Multiple Time Point Trials*. PhD Dissertation; Kansas State University: Manhattan, KS, 2009.

Surrogate Endpoint

Shu Zhang

Sepracor, Inc., Marlborough, Massachusetts, U.S.A.

Edward F. C. Pun

Agouron Pharmaceuticals, San Diego, California, U.S.A.

Abstract

Surrogate endpoints play a particularly important role in the early development of a pharmaceutical product. This entry reviews the benefits and the limitations of using surrogate endpoints and provides statistical analysis and regulatory requirements.

INTRODUCTION

In the development of a pharmaceutical product, a crucial step is to conduct clinical trials demonstrating the efficacy and safety of this product. The measurement used to evaluate the effect of the treatment is termed an *endpoint*. Endpoints used to measure clinical benefit, such as prolongation of survival time and reduction of mortality rate, are usually considered “true” endpoints because clinical benefit is the ultimate purpose of a medical treatment. However, the feasibility and practicality of measuring a “true” endpoint may be limited because of various reasons. Therefore it is crucial to have some alternative measurement that would allow early assessment of a potential treatment effect.

OVERVIEW

The most important reason that limits the use of a “true” endpoint is the time and the cost needed to conduct clinical trials. For example, time to death from the time of enrollment in a clinical trial is frequently used as an endpoint to evaluate a treatment for a life-threatening disease such as cancer or AIDS. However, to use survival time as the endpoint may require several hundred patients and several years of follow-up. During the process of developing a new drug, without evidence that the new product may be effective, it is unlikely that a sponsor will be willing to devote a large resource to a clinical trial that will take years to demonstrate a candidate’s therapeutic viability. The need to reduce the duration of clinical trials becomes more urgent in the search for treatment for life-threatening diseases.

Other factors may also limit the use of a “true” endpoint. For example, a new treatment may be developed during the trial period, making the treatment under investigation obsolete. Additionally, competing risk, i.e., death as a result of other reasons during the trial, constitutes another

problem for the evaluation of the treatment effect of the drug. These problems become more serious with longer follow-up times. Finally, the use of a “true” endpoint may also be limited by the current knowledge and technology for measuring the endpoint. One example is pain, which cannot be measured directly and objectively because of the current limited knowledge about pain. Because in most cases objective measurements, which are proven to be sensitive to the “true” degree of pain, are either difficult to obtain or do not exist, most clinical trials rely on patient self-assessment of pain as the endpoint, which can be viewed as the “surrogate” for the “true” severity of pain.^[1]

The limitations of “true” endpoints prompted the use of *surrogate endpoints*. A surrogate endpoint “is an endpoint that is intended to relate to a clinically important outcome but does not in itself measure a clinical benefit.”^[2]

A surrogate endpoint is sometimes referred to as an *intermediate endpoint*^[3,4] or a *surrogate marker*.^[5–7] A *marker* usually refers to some laboratory test result that is associated with a certain disease, usually indicating the progress of the disease. For example, significant elevations of SGOT and SGPT are strong indications for hepatitis or other liver damage, so they are considered markers for hepatitis. Similarly, an increase in HIV viral load usually indicates a worsened condition for AIDS patients, so HIV viral load is a marker for the progression of AIDS. A marker can be used to detect diseases, monitor disease progression, or predict clinical outcome, and because of its strong association with the clinical endpoint, it can also be used as a surrogate for the “true” endpoint. The focus of this discussion will be on the use of a surrogate marker as an endpoint, and the terms surrogate endpoint and surrogate marker may be used interchangeably. Because of the fact that a change in a surrogate endpoint occurs prior to a change in a “true” endpoint—for instance, an increase in HIV viral load occurs before the onset of AIDS—a surrogate endpoint may also be referred to as an intermediate endpoint.

Table 1 Typical Cardiovascular Trials with Surrogate and True Endpoints: Sample Sizes and Follow-Up Periods

Event	True endpoint trial			Surrogate endpoint trial		
	Endpoint	Size	Length	Endpoint	Size	Length
Myocardial infarction	Death	4,000	5 yr	Coronary artery patency	200	90 min
Myocardial infarction	Death	4,000	5 yr	Ejection fraction	30	2–4 wk
Stroke	Stroke	25,000	5 yr	Diastolic blood pressure	200	1–2 yr

Copyright © John Wiley & Sons Limited. (From Ref. [8].)

The use of surrogate endpoints can significantly reduce the follow-up time and sample size required to demonstrate treatment effect in a clinical trial. As an example, [Table 1](#)^[8] illustrates the reduction in time and cost resulting from using surrogate endpoints in clinical trials for the treatment of cardiovascular diseases.

Surrogate endpoints play a particularly important role in the early development of a pharmaceutical product, when surrogate endpoints can be used to demonstrate if the drug has any pharmacological effect, as postulated by the *in vitro* model or in preclinical *in vivo* results. Additionally, surrogate endpoints can provide guidance in selecting a dose range for later confirmatory trials.

It has been suggested that the “true” endpoint usually measures clinical benefit, and the surrogate endpoint measures mostly a specific aspect of disease progression.^[8] For a given disease, there can be different stages and different biological mechanisms affecting the progress of the disease. As a result, for a specific disease, the surrogate endpoint may change depending upon the stage of the disease as well as the specific physiopathological mechanism that the drug is targeting. For instance, both serum cholesterol level and diastolic blood pressure are widely used surrogate endpoints for coronary heart disease. To decide which of these two is appropriate to use will depend on the biological action of the drug under study. [Table 2](#)^[9] presents surrogate endpoints used in developing an antidiabetes drug. It can be seen that the choice of a surrogate endpoint depends upon the stage of diabetes as well as on the stage of the drug development, *i.e.*, whether it is an early- or late-phase clinical trial. [Table 2](#) shows that the sample size and the treatment duration also change according to the stage of drug development. Early-phase trials tend to be shorter and use fewer people as compared with the late-phase trials.

A potential surrogate is identified in several ways.^[8] The core for identifying a surrogate endpoint is a strong biological rationale. For example, any effective treatment for cancer must have some effect on the tumor itself, which suggests that tumor shrinkage might be used as a surrogate endpoint in the studies of cancer treatment. Epidemiological studies can also reveal endpoints that have strong predictive values to the clinical outcome. For instance, in many observational studies for AIDS, CD4 lymphocyte counts were shown to be the most powerful predictor of disease progression and

subsequent survival.^[10–13] This led to the extensive use of CD4 counts as a surrogate endpoint in early AIDS research. Lastly, clinical trials may also indicate a potential surrogate endpoint. In a clinical trial, if a treatment that suppresses a given marker produces more clinical benefit than another treatment that does not suppress the marker, that marker may be used as a potential surrogate endpoint. However, the use of the surrogate endpoints identified through these ways needs to be validated through further studies.

SCIENTIFIC ISSUES—LIMITATIONS OF USING SURROGATE ENDPOINT

Although yielding potentially important benefits, the use of surrogate endpoints can sometimes introduce ambiguous results and thus generate controversy:

There are two principal concerns with the introduction of any proposed surrogate variable. First, it may not be a true predictor of the clinical outcome of interest. For example, it may measure treatment activity associated with one specific pharmacological mechanism, but may not provide full information on the range of actions and ultimate effects of the treatment, whether positive or negative. There have been many instances where treatments showing a highly positive effect on a proposed surrogate have ultimately been shown to be detrimental to the subjects' clinical outcome; conversely, there are cases of treatments conferring clinical benefit without measurable impact on proposed surrogates.^[14]

Some cancer treatments provide an example of a treatment that has a positive effect on the surrogate endpoint but a detrimental effect on a patient's health. In this instance, although a treatment may achieve a certain degree of tumor shrinkage, the toxicity of the drug itself may in fact worsen a patient's general health and thus shorten survival time. Conversely, in AIDS trials, it has been shown that many treatments with a positive effect on survival time have null or even a negative effect on CD4 count, the proposed surrogate.^[15] A second concern relating to the use of surrogate endpoints is that “the proposed surrogate variables may not yield a quantitative measure of clinical benefit that can be weighed directly against adverse effects.”^[14]

Table 2 Clinical Endpoints vs. Surrogate Endpoints in Different Phases of Clinical Trials for Diabetes

Clinical endpoint	Phase	Subject characteristics	Group sample size	Treatment duration	Surrogate endpoints
Acute glucose lowering	I	Normals or diabetics	6–12	≥ Single dose	◆ Plasma glucose ◆ Plasma insulin ◆ Other (e.g., insulin clamp)
Improved glycemic control	I	Diabetics	12–36	2–12 wk	◆ Plasma glucose ⇒fasting ⇒2-hr postprandial ⇒24-hr profile
	III	Diabetics	40–200	6–12 mo	◆ Glycohemoglobin (HBA _{1c}) ◆ Glycated serum proteins ◆ Hypoglycemic episodes
Complications: anatomic/physiologic benefit	II–III	Diabetics with early complications	50–300	6 mo–2 yr	◆ Nerve conduction velocity ◆ Nerve/kidney biopsy ◆ Albuminuria, a GFR ◆ Funduscopy
Complications: actual clinical benefit	IV	Diabetics with early or no complications	200–800	≥ 3–5 yr	◆ Neuropathic pain ◆ Sensorimotor deficit ◆ GFR, renal failure ◆ Visual acuity ◆ Direct and indirect costs

(From Ref. [9].)

REGULATORY REQUIREMENTS

As just mentioned, the use of surrogate endpoints has limitations. However, in the case of life-threatening diseases, where the availability of a new treatment may save many lives, there is a strong need to reduce the duration of a clinical trial and to accelerate drug approval, and, very often, using a surrogate endpoint is necessary to evaluate a treatment more quickly.

The Food and Drug Administration (FDA) issued guidelines for accelerated approval based on the surrogate endpoints. The guideline states that:

FDA may grant marketing approval on the basis of adequate and well-controlled clinical trials establishing that the drug product has an effect on a surrogate endpoint that is reasonably likely, based on epidemiologic, therapeutic, pathophysiologic, or other evidence to predict clinical benefit or on the basis of an effect on a clinical endpoint other than survival or irreversible morbidity. Approval under this section will be subject to the requirement that the applicant study the drug further, to verify and describe its clinical benefit, where there is uncertainty as to the relation of the surrogate endpoint to clinical benefit or of the observed clinical benefit to ultimate outcome.^[15]

STATISTICAL METHODS

In the effort to identify reliable surrogate endpoints, several statistical methods have been proposed. In 1987, Prentice^[16] introduced the following statistical definition of a surrogate endpoint as “a response variable for which a test

of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypotheses based on the true endpoint.” Based on this definition, he proposed the use of the following two conditions as the operational criteria to determine a surrogate:

$$\lambda_T\{t; S(t), x\} = \lambda_T\{t; S(t)\}$$
 (1)

and

$$\lambda_T\{t; S(t)\} \neq \lambda_T(t)$$
 (2)

where t denotes the time from enrollment in a clinical trial, T denotes a true time-to-failure endpoint, $x = (x_1, \dots, x_p)$ consists of indicator variates for $p \geq 1$ of the $p + 1$ treatments to be compared in respect to the corresponding (instantaneous) failure rate, $\lambda_T\{t; x\}$, and $S(t) = \{Z(u); 0 \leq u < t\}$ denotes the history prior to time t of a possibly vector-valued stochastic process $Z(U) = \{Z_1(u), Z_2(u), \dots\}$, which may be used to specify a surrogate for T .

Criterion 1 requires that conditioning on the surrogate marker process, the treatment is independent of the failure rate for T . Thus the surrogate marker captures all the treatment effect of x on T . Criterion 2 requires that the surrogate marker has some predictive value to T . Recall that in “Introduction,” it was noted that a surrogate endpoint must have strong predictive value to the clinical outcome. The Prentice criterion of Eq. 2 is equivalent to this requirement. Between the two criteria, Eq. 1 is considered the principal criterion.^[17]

Although expressed in terms of survival analysis, these criteria can be interpreted in a general way. Feigal,^[18] from the FDA, expressed the following heuristic model for a drug effect:

$$\text{Outcome} = \text{Placebo rate} + \text{Drug effect}$$

In the instance of a surrogate marker, he further decomposed the model as

$$\begin{aligned} \text{Outcome} = & \text{Placebo rate} + \text{SM drug effects} \\ & + \text{Other drug effects} \end{aligned}$$

where SM is the abbreviation of surrogate marker. He pointed out that the Prentice criterion of Eq. (1) requires essentially that conditioning on the surrogate marker, the “other drug effects” reduce to 0.

This Prentice criterion is very strict. In 1992, Freedman and Graubard^[3] proposed another criterion for validating a surrogate, which is less stringent than that of Prentice. The development of Freedman and Graubard’s criterion is described briefly next.

With a slight change in notation, let T be a binary outcome, X indicate the treatment groups, and S denote the surrogate endpoint. Then the Prentice criteria can be rewritten as

$$P(T = 1 | S, X) = P(T = 1 | S) \quad (3)$$

and

$$P(T = 1 | S) \neq P(T = 1) \quad (4)$$

For simplicity, assume that there are two treatment groups ($X = 1$ and $X = 2$) and that S is discrete and takes k values $s_1, \dots, s_k$. Let $p(T = 1 | S = s_i, X = j) = p_{ij}$, then Eq. (3) is equivalent to

$$p_{i1} = p_{i2} \quad (i = 1, \dots, k) \quad (5)$$

Consider the following model:

$$g(p(T = 1 | S, X = j)) = h(S) + \tau_j \quad (6)$$

where S and X have an additive effect on some function (g) of the probability of disease, $h(S)$ is some function of S , and τ_j is the j th treatment effect. If we further choose logit function for g and assume that

$$h(S) = \mu + \sigma_i$$

where μ represents the overall mean, then Eq. (6) becomes the usual linear logistic model with the constraint of $\sigma_1 + \dots + \sigma_k = 0$ and $\tau_1 + \tau_2 = 0$, and it can be further expanded to include other effects. Consequently, the test of the null hypothesis (Eq. [5]) becomes the test of $\tau_1 = \tau_2 = 0$.

If this hypothesis is rejected, S cannot be used as a surrogate endpoint. However, the lack of significance in this test does not warrant S to be a valid surrogate endpoint. To quantify further the effect of S , consider the quantity

$$1 - \frac{\hat{\tau}_{1a}}{\hat{\tau}_1}$$

where $\hat{\tau}_{1a}$ is the adjusted (by S) treatment effect in group 1 and $\hat{\tau}_1$ is the unadjusted treatment effect in group 1. This quantity is the estimate of the proportion of the drug effect that is explained by the surrogate endpoint. Under the criteria of Prentice, this proportion is expected to equal 1. However, we may allow this proportion to be larger than some critical values, such as 0.5 and 0.75, but less than 1. In other words, we may allow the surrogate endpoint to explain a large portion of the drug effect but not all of the effect. Using the notation in the heuristic model, Feigal^[18] presented Freedman and Graubard’s^[3] model as

$$\begin{aligned} & \text{Proportion of drug effect attributed to SM drug effects} \\ &= \frac{\text{SM drug effects}}{\text{Total drug effects}} \end{aligned}$$

RECENT DEVELOPMENTS

The use of surrogate endpoints in clinical trials is an important issue because they can significantly reduce the duration and the cost of clinical trials and thereby accelerate the development of medical treatments. Surrogate endpoints had long been used to assess treatment effects, especially for rare and life-threatening diseases. However, there is some controversy surrounding their use. In 1987, a meeting was held to discuss these issues.^[1,8,16,19,20] In this meeting, Prentice proposed statistical operational criteria to identify surrogate endpoints.^[16] Since then, Prentice’s criteria, as described earlier, have been applied by several researchers and have been used to evaluate several commonly used surrogate markers.^[17,21] In 1992, the FDA issued regulations for accelerated approval based on the surrogate endpoints for new drugs and biological products treating serious or life-threatening illness.^[22] In the same year, Freedman and Graubard^[3] proposed the modification to Prentice’s criteria, making the criteria less stringent. These criteria have also been applied to biomedical research.^[6] In addition, other methods, such as using meta-analysis to identify surrogate endpoints, have also been proposed.^[23]

In recent years, the FDA has issued several guidelines prepared under the auspices of the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use.^[2,14,15] The use of surrogate endpoints was discussed in the “general considerations for clinical trials”^[2] and was specifically

addressed in “guideline on statistical principles for clinical trials.”^[14] The FDA later established a “surrogate marker methodology group” to evaluate each potential surrogate endpoint.^[18] The FDA’s guideline for each disease also includes the use of a specific marker as a surrogate endpoint for that disease.

In addition to the use of a surrogate endpoint in place of the true endpoint for the efficacy analysis, different methods have been proposed in which a surrogate endpoint was used to improve the efficiency of the estimate of the treatment effect on the true endpoint. Flandre and O’Quigley^[24] proposed a two-stage approach. In the first stage, the relationship between the true and the surrogate endpoints, both measured for all subjects, is examined. This relationship is then applied in the second stage to other subjects, where the follow-up will be terminated once the surrogate endpoint is observed. Lefkopoulou and Zelen^[4] proposed a test to examine the impact of an intermediate endpoint on the true endpoint, which is the survival time. Cox^[25] proposed a method using a surrogate endpoint to recover information and predict the survival distribution in the case of right censoring. This idea was further extended by Fleming et al.,^[17] and they referred to the endpoint that provides additional information on the true endpoint as an *auxiliary endpoint* (referred to as a surrogate endpoint in the paper by Cox^[25]) and distinguished it from the surrogate endpoint. Kosorok and Fleming^[26] also proposed a method using surrogate failure-time data to increase efficiency in estimating the treatment effect on the true endpoint.

CONCLUSION

As medical science advances, human beings are revealing the mystery of many complicated diseases, which may take decades to evolve, such as Alzheimer’s disease. In the process of studying these diseases, the use of surrogate endpoints is almost inevitable. With the help of the development of medical technologies, such as the study of genes and protein structure, as well as the development of statistical methods, more reliable surrogate endpoints will be identified to enable humankind to understand an increasing number of diseases. Meanwhile, one should keep in mind the limitation of the surrogate endpoints, and use them with caution.

REFERENCES

1. Ellenberg, S.; Hamilton, J.M. *Stat. Med.* **1989**, *8*, 405–413.
2. Federal Register 62:242, §3.2.2.4, 1997.
3. Freedman, L.S.; Graubard, B.I. *Stat. Med.* **1992**, *11*, 167–178.
4. Lefkopoulou, M.; Zelen, M. *Lifetime Data Anal.* **1995**, *1*, 73–85.
5. Fleming, T.R. *Stat. Med.* **1994**, *13*, 1423–1435.
6. Lin, D.Y.; Fleming, T.R.; DeGruttola, V. *Stat. Med.* **1997**, *16*, 1515–1527.
7. Choi, S.; Lagakos, S.W.; Schooley, R.T.; Volberding, P.A. *Ann. Intern. Med.* **1993**, *118*, 674–680.
8. Wittes, J.; Lakatos, E. *Stat. Med.* **1989**, *8*, 414–425.
9. MacLean, D. *Diabetes Mellitus: Therapeutic Approaches and Issues in Drug Development*. Seminar, Pfizer Central Research-Groton, CT, 1997.
10. Redfield, R.R.; Burke, D.S. *Sci. Am.* **1988**, *259*, 90–98.
11. Polk, B.F.; Fox, R.; Brookmeyer, R.; Kanchanaraksa, S.; Kaslow, R.; Visscher, B.; Rinaldo, C.; Phair, J. N. *Engl. J. Med.* **1987**, *316*, 62–66.
12. Goedert, J.J.; Biggar, R.J.; Melbye, M.; Mann, D.L.; Wilson, S.; Gail, M.H.; Grossman, R.J.; DiGivita, R.A.; Sanchez, W.C.; Weiss, S.H.; Blattner, W.A. *JAMA* **1987**, *257*, 331–334.
13. Safai, B.; Johnson, K.G.; Myskowski, P.L.; Koziner, B.; Yang, S.Y.; Cunningham-Rundles, S.; Godbold, J.H.; Dupont, B. *Ann. Intern. Med.* **1985**, *103*, 744–750.
14. ICH Harmonized Tripartite Guideline, Statistical Principles for Clinical Trials, §2.2.6, 1998.
15. Code of Federal Regulations 21, §314.510, 1997.
16. Prentice, R.L. *Stat. Med.* **1989**, *8*, 431–440.
17. Fleming, T.R.; Prentice, R.L.; Pepe, M.S.; Glidden, D. *Stat. Med.* **1994**, *13*, 955–968.
18. Feigal, D.W., Jr. PK/PD Modeling in Regulatory Decisions for Effectiveness. In *Modeling and Simulation of Clinical Trials in Drug Development and Regulation*, Herndon, VA, 1977.
19. Herson, J. *Stat. Med.* **1989**, *8*, 403–404.
20. Hillis, A.; Seigel, D. *Stat. Med.* **1989**, *8*, 427–430.
21. Lin, D.Y.; Fischl, M.A.; Schoenfeld, D.A. *Stat. Med.* **1993**, *12*, 835–842.
22. Federal Regist. **1992**, *57*, 239, pp. 58942.
23. Daniels, M.J.; Hughes, M.D. *Stat. Med.* **1997**, *16*, 1965–1982.
24. Flandre, P.; O’Quigley, J. *Biometrics* **1995**, *51*, 969–976.
25. Cox, D.R. *J. R. Stat. Soc., B.* **1983**, *45*, 391–393.
26. Kosorok, M.R.; Fleming, T.R. *Biometrika* **1993**, *80*, 823–833.

Survival Analysis

Dirk F. Moore

Department of Biostatistics, University of Medicine and Dentistry of New Jersey School of Public Health, New Jersey, U.S.A.

Abstract

Survival analysis is the study of time-to-event data. The distinguishing characteristics of such data are well-defined starting and ending points. This entry reviews a randomized clinical trial example and several models for comparing survival analyses.

INTRODUCTION

Survival analysis is the study of time-to-event data. The distinguishing characteristics of such data are well-defined starting and ending points, and examples may be found in such diverse fields as demography, actuarial science, engineering, medicine, and epidemiology. Survival data are often subject to *right censoring*, a form of incomplete observation in which some endpoints are only known to be greater than an observed censoring time. In a clinical trial, for example, right-censored observations can occur when subjects drop out of a study, or if a study ends before all subjects have experienced the endpoint of interest. Other forms of incomplete observation include left and interval censoring, and a form of length-biased sampling known as truncation. Correctly accounting for censoring and truncation is essential for obtaining unbiased estimates of the survival distribution and for drawing correct inferences.

A RANDOMIZED CLINICAL TRIAL EXAMPLE

A randomized clinical trial of survival is an experimental method for comparing the survival time distributions of two (or possibly more) treatments for a disease. The survival time is the time from when a patient enters the trial until and the occurrence of an event of interest. This event could be death or recurrence of disease, or even a positive event, such as response to treatment; the key requirement is that the event-of-interest be precisely defined. In the first phase of a trial, called the accrual period, patients are entered into a trial and randomly assigned to one of the treatments. After a desired number of patients are entered, accrual is closed, and the patients who have been entered are followed for a specified period of time known as the follow-up period. A feature of most clinical trials is that at the end of the follow-up period, some patients may not yet have experienced the event of interest. These patients are said to be right-censored—their time-to-event has not

been observed, but is known to exceed the censoring time. For each patient two variables are recorded. The first, the time from entry until the event of interest or, for censored patients, the time from entry until the end of follow-up, is known as the *survival time*. The second variable is a *censoring indicator*, a binary observation that indicates whether the observed survival time is the time to an event or the time to censoring. The goal of the study is to compare the survival times of patients receiving each of the treatments, while correctly accounting for the censoring. A key assumption is that the censoring must be independent of the survival times. This assumption must be carefully assessed, particularly if there is censoring due to patient drop-out.

A simple example, which is typical of randomized clinical trials, is a study comparing the efficacy of maintenance therapy for patients with acute myelogenous leukemia, or AML.^[1] A total of 23 patients in remission following chemotherapy were randomly assigned to either receive or not receive maintenance therapy, and the survival times were defined to be time from randomization until recurrence of disease or censoring. The survival times (remission times in weeks) are as follows:

Maintained group: 91313+182328+313445+48161+

Non – maintained group: 55881216+232730334345

For example, the first two patients in the maintained group were in remission for 9 and 13 weeks, respectively, before recurrence of disease. The third patient in the maintained group was censored at 13 weeks; that is, that patient was still in remission at the end of the follow-up period. The other observations are interpreted in an analogous manner. These results may be presented as *Kaplan-Meier survival curves* (Fig. 1). These survival curves are plots of the estimated probability of being disease-free versus the number of weeks from randomization. Note that both curves start at 1 (the probability of being disease-free at the time of randomization) and decrease as a step function over time,

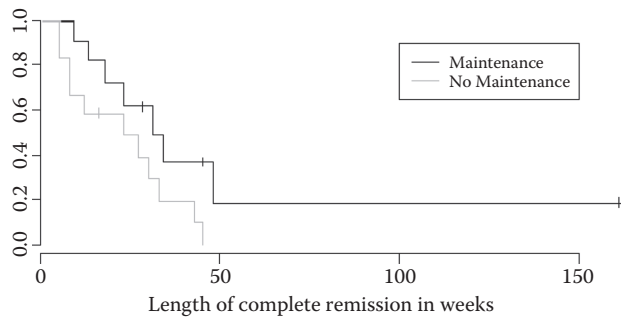


Fig. 1 Kaplan-Meier survival curves for the AML maintenance study

with steps occurring at the event times. Censoring times are indicated by short vertical marks. Visual inspection of the two survival curves shows that, in this trial, patients receiving maintenance therapy remained disease-free longer on average than did patients not receiving the therapy. However, drawing an inference about the difference between two treatments requires a *log rank test*, a specialized procedure for comparing censored survival data discussed later in this article.

SURVIVAL FUNCTIONS AND HAZARD RATES

Let $T > 0$ denote a survival random variable. The survival function is defined to be $S(t) = \text{pr}(T > t)$, and gives, for any time t , the probability that the random variable T exceeds t . Thus, $S(0) = 1$, and the function S is a decreasing (or, more precisely, a monotone non-increasing) function of t . A closely related function is the cumulative distribution function, or (in epidemiology) the cumulative risk of disease, defined by $F(t) = 1 - S(t)$. The probability density function is defined by $f(t) = -\frac{dS(t)}{dt}$. For survival data it is often more convenient to work with the *hazard* function, which may be thought of as the instantaneous risk of failure. Formally, this may be written as

$$h(t) = \lim_{\Delta \rightarrow 0^+} \frac{\text{pr}(t \leq T < t + \Delta \mid t \leq T)}{\Delta},$$

that is, the probability that a subject who has survived up to time t will fail in the next “instant” of time, Δ . In terms of the survival and density functions, this may be written as $h(t) = f(t)/S(t)$. Another useful result is that the survival function may be written in terms of the hazard function as follows:

$$S(t) = \exp \left\{ - \int_0^t h(u) du \right\}.$$

Using these formulas, one may derive the survival function from the hazard function, and vice versa.

PARAMETRIC SURVIVAL MODELS

One approach to modeling survival data is to construct a likelihood function using a parametric family of survival distributions. One may then use standard likelihood theory for estimation and testing. The simplest parametric family for survival analysis is the *exponential* distribution. Its survival and hazard functions are given by $S(t) = e^{-\lambda t}$ and $h(t) = \lambda$, respectively, where λ is called a *rate* parameter. Note that the hazard function is constant over time, indicating that the risk of failure is the same regardless of how long a subject has survived. This is often known as the lack-of-memory property, and limits its usefulness in many applications; human survival times in particular rarely have this property.

The *Weibull* distribution is a more flexible distribution. Its survival and hazard functions are given by $S(t) = \exp(-\lambda t^\alpha)$ and $h(t) = \alpha \lambda t^{\alpha-1}$, respectively, where λ is a scale parameter and α is a *shape* parameter. Three examples of Weibull hazard functions for various choices of the shape parameter are given in Fig. 2. Note that the exponential distribution results as a special case when $\alpha = 1$.

Other distributions useful in survival analysis include the *gamma* and *log-normal* distributions. Details of these and other parametric distributions may be found in, for example, Klein and Moeschberger^[2] and Cox and Oakes.^[3]

Once a parametric model is chosen, it may be fitted to a survival data set by constructing a *likelihood* function. This likelihood consists of a product of factors (if the observations are independent), one for each subject. For data that include right-censored observations, and assuming that the censoring is independent of the failure times, the likelihood may be written as follows:

$$L(\theta) = \prod_{i \in D} f(t_i; \theta) \prod_{i \in C} S(t_i; \theta),$$

where D denotes the set of all subjects who are observed to fail, C denotes the set of subjects who are right-censored, and θ denotes the parameters that define the parametric model. As before, f and S denote the probability density function and survival function, respectively. If the data set

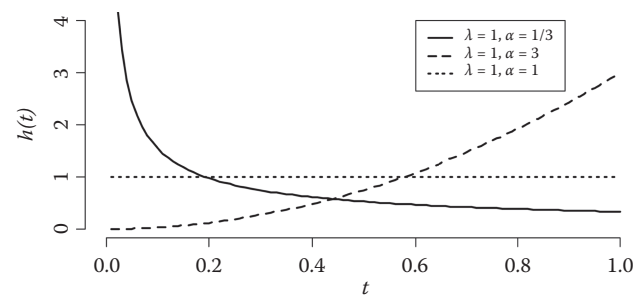


Fig. 2 Weibull hazard functions with $\lambda = 1$ various choices of α . Note that the case when $\alpha = 1$ corresponds to an exponential hazard function

includes subjects who are subject to left or interval censoring, or truncation (due to length-biased sampling), they may be entered into the likelihood using appropriate factors.^[2] Once a likelihood function has been defined, the parameters θ may be estimated using maximum likelihood techniques.

If a parametric family with a sufficiently flexible hazard can be found, estimation and testing may be carried out using standard likelihood theory. The strength of parametric modeling lies in its relative simplicity; complex sampling experiments, as well as clustered survival times, may often be accommodated in a relatively straightforward way. However, a parametric model may not accurately represent the true distribution of survival times, and thus may lead to biased estimates of parameters and incorrect inferences. Additional flexibility in modeling the hazard function may be obtained by using the *piecewise exponential* distribution. In this generalization of the exponential, the hazard function is constant within a small number of specified intervals, but is free to vary from one time interval to the next.^[2] This model may be used in applications where simple parametric models do not provide the needed modeling flexibility, and where fully non-parametric models are not computationally tractable. An example of this may be found in Moore et al.,^[4] who develop a non-standard likelihood for estimating the age-of-onset distribution of breast cancer among carriers of a genetic mutation.

NON-PARAMETRIC SURVIVAL MODELS

Unlike parametric methods, non-parametric survival methods make minimal assumptions about the shape of the hazard function. These methods are derived from actuarial life-table methods, in which time is divided into a number of intervals (often five-year intervals), and the probability of surviving each interval is computed. The survival distribution is computed as a cumulative product of these probabilities. The product-limit estimator, also known as the Kaplan-Meier estimator, of the survival distribution results when the time intervals are sufficiently small that only one event occurs per interval,^[5] although modifications for tied observation times are possible, as discussed by numerous researchers (e.g., Ref. [3]). An asymptotically equivalent method, known as the Nelson-Aalen estimator, models the hazard as a piecewise constant exponential distribution, again with one event per interval.^[2] The flexibility of these models has made them among the most frequently used for modeling human survival data. The Kaplan-Meier and Nelson-Aalen estimators of the survival function are given by, respectively,

$$\hat{S}_{KM}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{r_i}\right)$$

$$\hat{S}_{NA}(t) = \exp \left\{ - \sum_{t_i \leq t} \frac{d_i}{r_i} \right\}$$

where the t_i are the failure times, r_i is the number of subjects at risk at time t_i , and d_i is the number of failures that occur at time t_i . Both of these estimates yield step-function estimates, as in the example shown in Fig. 1.

COMPARING SURVIVAL CURVES

One may compare two survival curves by computing a *log rank statistic*,

$$U = \sum_{t_i \in D} \left(d_{1i} - \frac{d_i r_{1i}}{r_i} \right)$$

It has variance

$$V = \sum_{t_i \in D} d_i \frac{r_{0i} r_{1i} (r_i - d_i)}{r_i^2 (r_i - 1)},$$

where D is the set of all failure times, d_{1i} and d_i are the numbers of deaths at time t_i in group 1 and in both groups 0 and 1, respectively, r_{0i} and r_{1i} are the numbers of subjects at risk for failure at time t_i in groups 0 and 1, respectively, and $r_i = r_{0i} + r_{1i}$. The *log rank test* for a difference of two survival curves is based on the statistic $Z = U/V^{1/2}$, which has approximately a standard normal distribution. A weighted version of U , with weights given by $w_i = n\hat{S}(t_i)$, is known as the Prentice modification of the Gehan statistic. This modified statistic places more weight on early survival differences; details may be found in Cox and Oakes.^[3]

Sample size calculations for a two-sample log-rank test with a specified significance level α and power may be conducted most directly in terms of the number of events d required to detect a log-hazard ratio $\beta = \log[\lambda_1(t)/\lambda_0(t)]$ between two groups with hazard functions $\lambda_1(t)$ and $\lambda_0(t)$, respectively. For equal numbers in the two groups, the total number of events required is given by

$$d = \frac{4(z_\alpha + z_{1-\text{power}})^2}{\beta^2}.$$

Selection of an appropriate value of β may be made in terms of the inverse of the desired ratio of median survival times. For example, if the survival times are exponentially distributed, the required β is given by $\beta = \log_e(m_0/m_1)$, where m_1 and m_0 are the median survival times in groups 1 and 0, respectively. See Therneau and Grambsch^[6] for additional approaches to selecting β for estimating a sample size.

A more difficult problem is to calculate the total number of subjects that must be enrolled into a trial to produce d events. The total number depends on the lengths of the accrual and follow-up periods and on the survival functions in the control and treatment groups. Of course, the larger the amount of censoring that a design calls for, the larger

the total sample size will need to be to produce the required number of events. See Schoenfeld^[7] for details.

A covariate vector $\tilde{z}$ may be accommodated using the *proportional hazards model*,^[8] in which an individual's hazard function is given by

$$\lambda(t; \tilde{z}) = \lambda_0(t) \exp(\tilde{z}\beta).$$

Here $\lambda_0(t)$ is a nonparametric baseline hazard function, and $\exp(\tilde{z}\beta)$ is a multiplicative constant. Estimation of the covariate vector β proceeds via a *partial* likelihood, a semi-parametric version of a likelihood. The *log rank test* discussed earlier may be derived from the proportional hazards model by using a treatment indicator variable as the covariate z and constructing a *score test* of $H_0: \beta = 0$.^[3] The mathematical underpinnings of the non-parametric survival curve methods and the semi-parametric Cox model are based on the theory of counting processes and martingale theory.^[9,10]

Covariate information may also be accommodated using parametric models.^[3] Parametric models lack the flexibility of the Cox proportional hazards model, but may be preferred in applications where computational difficulties render the Cox model impractical.

COMPETING RISKS

When a subject fails due to any of several causes, and one wants to study each of them separately, we have a problem of *competing risks*. We observe, for each individual, a survival time T and a failure type V . For each failure type, we may estimate a cause-specific hazard function

$$h_v(t) = \lim_{\Delta \rightarrow 0^+} \frac{\text{pr}(t \leq T < t + \Delta, V = v | t \leq T)}{\Delta}.$$

Unfortunately, $h_v(t)$ may not be used to estimate the individual hazard functions or survival distributions unless the event time distributions are mutually independent.^[3,11] This assumption may be plausible in some engineering applications, but is problematic in human data, where complex dependencies among competing events are likely. The *cumulative incidence function*, also known as a *subdistribution function*, is the probability of failing from failure type V before time t . It is given by

$$I_v(t) = \int_0^t h_v(u) S(u) du,$$

where $S(u)$ is the survival probability based on failure from any cause and $h_v(u)$ is the cause-specific hazard.

To illustrate these ideas, we consider the disease trajectories of men diagnosed with prostate cancer with regard to death from prostate cancer and death from other

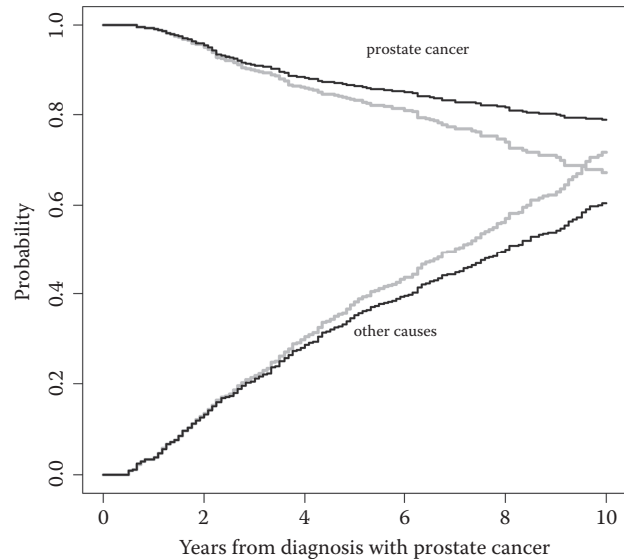


Fig. 3 Survival and cumulative incidence estimates of the probability of death from prostate cancer and from other causes. Competing risks estimates based on cause-specific hazards are in solid black, and naive Kaplan-Meier estimates are in gray

causes.^[12,13] Fig. 3 presents, for elderly men diagnosed with poorly differentiated prostate cancer, the survival distribution of time from diagnosis to death from prostate cancer and the cumulative incidence of death from other causes. Calculation details may be found in Putter, Fiocco, and Geskus.^[14] If each of these two causes is viewed as censoring for the other, one obtains crossing curves (gray), which are not possible, as the probabilities add to more than one. For example, the naïve (gray) survival curve for death from prostate cancer incorrectly treats death from other causes as a censored observation, so that the event potentially could happen in the future. As a result, the survival probabilities are underestimated. When one correctly accounts for the competing risks, the solid black curves are obtained.

OTHER ISSUES IN SURVIVAL ANALYSIS

While right censoring is the most common form of incomplete observation in survival analysis, other kinds of censoring, such as left censoring and interval censoring, are possible. Furthermore, length-biased sampling issues arise in certain applications. For example, in some situations, such as in estimating the latency distribution in studies of AIDS, only cases that occur before a certain time are included.^[15] This is known as *right truncation*. In other applications, the data are subject to left truncation, where subjects become at risk at a known time. A thorough treatment of likelihood construction for various types of censoring and truncation may be found in Klein and Moeschberger.^[16] A general discussion of the non-parametric estimation of the survival distribution under general censoring and truncation patterns may be found in Turnbull.^[17]

Another issue that arises in some applications is clustering of the survival times. For example, one may be interested in the age of onset of a type of cancer among family members, or one may have several time measures on the same individual. These may be accommodated using a marginal model^[18] or by including a common *frailty* in the model for each cluster. Other issues arise when one's interest is in times to multiple occurrences of an event. A thorough treatment of clustering and multiple events in survival data may be found in Therneau and Grambsch.^[6]

REFERENCES

1. Miller, Jr., R.G. *Survival Analysis*; Wiley: New York, 1981.
2. Klein, J.P.; Moeschberger, M.L. *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd Ed.; Springer: New York, 2005.
3. Cox, D.R.; Oakes, D. *Analysis of Survival Data*; Chapman and Hall: London, 1984.
4. Moore, D.F.; Chatterjee, N.; Pee, D.; Gail, M.H. Pseudo-likelihood estimates of the cumulative risk of an autosomal dominant disease from a kin-cohort study. *Gen. Epidemiol.* **2001**, *20*, 210–227.
5. Kaplan, E.L.; Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **1958**, *53*, 457–481.
6. Therneau, T.M.; Grambsch, P.M. *Modeling Survival Data*; Springer: New York, 2000.
7. Schoenfeld, D.A. Sample-size formula for the proportional-hazards regression model. *Biometrics* **1983**, *39*, 499–503.
8. Cox, D.R. Regression models and life-tables (with discussion). *J. R. Stat. Soc., B* **1972**, *34*, 187–220.
9. Fleming, T.R.; Harrington, D.P. *Counting Processes and Survival Analysis*; Wiley: New York, 1991.
10. Andersen, P.K.; Borgan, O.; Gill, R.D.; Keiding, N. *Statistical Models Based on Counting Processes*; Springer-Verlag: New York, 1993.
11. Prentice, R.L.; Kalbfleisch, J.D.; Peterson, A.V.; Flournoy, N.; Farewell, V.T.; Breslow, N.E. The analysis of failure time data in the presence of competing risks. *Biometrics* **1978**, *34*, 541–554.
12. Lu-Yao, G.L.; Albertsen, P.C.; Moore, D.F.; Shih, W.J.; Lin, Y.; DiPaola, R.S.; Barry, M.J.; Zietman, A.; O'Leary, M.; Walker-Corkery, E.; Yao, S.-L. Outcomes of localized prostate cancer following conservative management. *JAMA* **2009**, *302*, 1202–1209.
13. Warren, J.L.; Klabunde, C.N.; Schrag, D.; Bach, P.B.; Riley, G.F. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Med. Care* **2002**, *40*, IV-3–18.
14. Putter, H.; Fiocco, M.; Geskus, R.B. Tutorial in biostatistics: Competing risks and multi-state models. *Stat. Med.* **2007**, *26*, 2389–2430.
15. Lagakos, S.W.; Barraj, L.M.; Degruittola, V. Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika* **1988**, *75*, 515–523.
16. Klein, J.P.; Moeschberger, M.L. *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd Ed.; Springer: New York, 2005.
17. Turnbull, B.W. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. R. Statist. Soc., B* **1976**, *38*, 290–295.
18. Wei, L.J.; Lin, D.Y.; Weissfeld, L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Am. Stat. Assoc.* **1989**, *84*, 1065–1073.

Targeted Clinical Trials

Jen-pei Liu

Division of Biometry, Department of Agronomy, and Institute of Epidemiology, National Taiwan University, Taipei, Taiwan; and Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan

Abstract

This entry illustrates the drawbacks of the current approach to targeted clinical trials based on the continuous data. It also discusses the EM algorithm for binary data and provides formulas for sample size determination.

Keywords: Bootstrap method; EM algorithm; Positive predictive value; Targeted clinical trials.

INTRODUCTION

For the traditional clinical trials, clinical endpoints and clinical-pathological signs or symptoms are often used as the inclusion and exclusion criteria for the intended patient population in which the new treatments are evaluated with a concurrent control. However, considerable variation exists in the responses to the new treatment even for the patients with the same diagnosis. In other words, these endpoints are not well correlated with the clinical benefits of the treatments in the patient population defined by clinical-based inclusion and exclusion criteria. Therefore current paradigm to develop and evaluate a drug or a treatment uses a shotgun approach that may not be beneficial for most of patients. One of the reasons is that the traditional inclusion/exclusion criteria do not utilize the potentially important genetic or genomic variability of the trial participants.

After completion of a human genome project, new breakthrough technology such as microarray, mRNA transcript profiling, single nucleotide polymorphism (SNP), or genome-wide association studies (GWAS) emerges. As a result, the disease targets at molecular level can be identified. Treatments specific for the molecular targets can therefore be developed for the patients who are most likely to benefit. These treatments are referred to as the targeted treatments or drugs. Unlike the shotgun blast approach, targeted therapy employed a precision guided-missile methodology to reach the molecular targets and personalized medicine finally becomes a reality.^[1,2]

Development of targeted treatments is different from the traditional approach. For a targeted therapy, one must have (i) the knowledge of the molecular targets involved in the disease pathogenesis, (ii) a device for detection of the molecular targets, and (iii) a treatment aimed for the molecular targets. Hence, development of targeted therapies involves translation from assessment of the accuracy

of diagnostic devices for the molecular targets to evaluation of the efficacy and safety of the treatment modality for the patient population with the targets. Therefore, evaluation of targeted treatments posts new challenges in clinical research and development. To address these new emergent issues, in April 2005, the U.S. Food and Drug Administration (U.S. FDA) issued the *Drug-Diagnostic Co-development Concept Paper*.^[3] The enrichment design given in Fig. 1^[4] is one of the three designs introduced in the Concept Paper for evaluation of the targeted treatments. The enrichment design consists of two phases. Phase 1 is the enrichment phase during which patients are screened by a diagnostic device for detection of the pre-defined molecular targets. Then only the patients with a positive result for the molecular target are randomized to receive either the targeted treatment or the concurrent control in Phase 2.

However, in practice, no diagnostic test is perfect with 100% positive predicted value (PPV). In addition, measures for diagnostic accuracy such as sensitivity, specificity, PPV, or negative predicted value (NPV) are in fact estimators with variability. Therefore, some of the patients enrolled in targeted clinical trials under the enrichment design might not have the specific targets and hence the treatment effects of the drug for the molecular targets could be under-estimated.^[5] Under the enrichment design, Liu and Lin^[6] and Liu et al.^[7] propose using the EM algorithm^[8–10] in conjunction with the bootstrap technique^[11] to incorporate the uncertainty on the accuracy of the diagnostic device in detection of the molecular targets for the inference of the treatment effects with binary and continuous endpoints. In the next section, we illustrate the drawbacks of the current approach based on the continuous data. The application of EM algorithm to the inference of continuous endpoints is reviewed in section “The EM Procedures.” The EM algorithm for binary data is similar

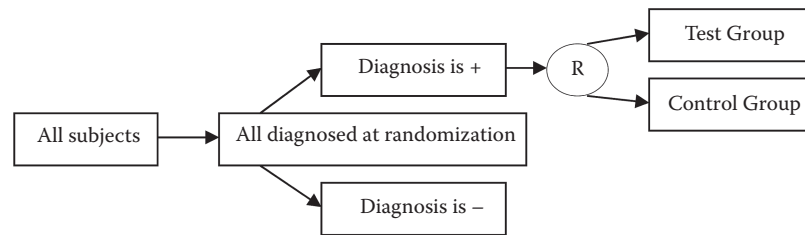


Fig. 1 Enrichment design (From Ref. [3])

and will not be given in this entry. See Liu and Lin^[6] for details. Simulation results of both continuous and binary data are given in section “Simulation Study.”^[6,7] section “Sample Size Determination” provides formulas for sample size determination with consideration of the variability of PPV. Discussion and concluding remarks are given in section “Discussion.”

CURRENT APPROACHES

For the purpose of illustration and simplicity, consider a situation where a particular molecular target involved with the pathway in pathogenesis of the disease has been identified and there is a validated diagnostic device available for detection of the identified molecular target is considered. Also suppose that a test drug for the particular molecular target is currently being developed. Under an enrichment design, one of the objectives of targeted clinical trials is to evaluate the treatment effects of the molecular targeted test treatment in the patient population truly having the molecular target. With respect to Fig. 1, we consider a two-group parallel design in which patients with a positive result by the diagnostic device are randomized in a 1:1 ratio to receive the molecular targeted test treatment (T) or a control treatment (C). We further assume that the primary efficacy endpoint is an approximately normally distributed variable, which is denoted by Y_{ij} , $j=1, \dots, n_i$; $i=T, C$. Table 1 gives the expected values of Y_{ij} by treatment and diagnostic result of the molecular target. In Table 1, μ_{T+} , μ_{C+} (μ_{T-} , μ_{C-}) are the means of test and control groups for the patients truly having (or without) the molecular target. Our parameter of interest is the treatment effects for the patients truly having the molecular target $\theta = \mu_{T+} - \mu_{C+}$. The hypothesis for detection of treatment difference in the patient population truly having the molecular target is the hypothesis of interest:

$$H_0 : \mu_{T+} - \mu_{C+} = 0 \quad \text{vs.} \quad H_a : \mu_{T+} - \mu_{C+} \neq 0 \quad (1)$$

Let $\bar{Y}_T$ and $\bar{Y}_C$ be the sample means of test and control treatments, respectively. Liu and Chow^[5] showed that

$$E(\bar{Y}_T - \bar{Y}_C) = \gamma(\mu_{T+} - \mu_{C+}) + (1 - \gamma)(\mu_{T-} - \mu_{C-}), \quad (2)$$

where γ is the PPV.

Liu and Chow^[5] indicated that the expected value of the difference in sample means consists of two parts. The first part is the treatment effects of the molecular target drug in patients with a positive diagnosis who truly have the molecular target of interest. The second part is the treatment effects of the patients with a positive diagnosis but in fact they do not have the molecular target. It follows that the difference in sample means obtained under the enrichment design for targeted clinical trials actually underestimates the true treatment effects of the molecular targeted drug in the patient population truly having the molecular target of interest if $\mu_{T+} - \mu_{C+} > \mu_{T-} - \mu_{C-}$. As it can be seen from,^[2] the bias of the difference in sample means decreases as the PPV increases. On the other hand, the PPV of a diagnostic test increases as the prevalence of the disease increases.^[12] For a disease which is highly prevalent, say greater than 10%, even with a high diagnostic accuracy of a 95% sensitivity and a 95% specificity for the diagnostic device, the PPV is only about 67.86%. It follows that the downward bias of the traditional difference in sample means could be substantial for estimation of treatment effects of the targeted drug in patients truly having the target of interest. In addition, since $\bar{Y}_T - \bar{Y}_C$ under-estimates $\mu_{T+} - \mu_{C+}$, the planned sample size may not be sufficient for achieving the desired power for detecting the true treatment effects in the patients truly having molecular target of interest. It should be emphasized that the above arguments for a downward bias also holds for the binary endpoints.^[6]

Table 1 Population Means by Treatment and Diagnosis

Positive diagnosis	True target condition	Accuracy of diagnosis	Test group	Control group	Difference
+	+	γ	μ_{T+}	μ_{C+}	$\mu_{T+} - \mu_{C+}$
+	-	$1-\gamma$	μ_{T-}	μ_{C-}	$\mu_{T-} - \mu_{C-}$

γ is the PPV.

(From Liu et al. [7]).

THE EM PROCEDURES

Although all patients randomized under the enrichment design have a positive diagnosis, the true status of the molecular target for individual patients in the targeted clinical trials is in fact unknown. Liu et al.,^[7] proposed to apply the EM algorithm for the inference of the treatment effects for the patients truly having the molecular targets. It follows that under the assumption of homogeneity of variance, Y_{ij} are independently distributed as a mixture of two normal distributions with mean μ_{i+} and μ_{i-} respectively, and a common variance σ^2 .^[12]

$$\varphi(y_{ij} | \mu_{i+}, \sigma^2)^\gamma \varphi(y_{ij} | \mu_{i-}, \sigma^2)^{1-\gamma} \quad i = T, C; j = 1, \dots, n_i, \quad (3)$$

where $\varphi(\cdot)$ denotes the density of a normal variable.

However, γ is an unknown PPV which must be estimated from the data. Therefore, the data obtained from the targeted clinical trials are incomplete because the true status of the molecular target for the patients is missing.

Let X_{ij} is the latent variable indicating the true status of the molecular target of patient j in treatment i ; $i, j = 1, \dots, n_i, i = T, C$. Under the assumption that X_{ij} are i.i.d. Bernoulli random variables with probability γ for the molecular target, the complete-data log-likelihood function for ψ is given by

$$\begin{aligned} \log L_c(\Psi) = & \sum_{j=1}^{n_T} X_{Tj} [\log \gamma + \log \varphi(y_{Tj} | \mu_{T+}, \sigma^2)] \\ & + \sum_{j=1}^{n_T} (1 - X_{Tj}) [\log(1 - \gamma) + \log \varphi(y_{Tj} | \mu_{T-}, \sigma^2)] \\ & + \sum_{j=1}^{n_C} X_{Cj} [\log \gamma + \log \varphi(y_{Cj} | \mu_{C+}, \sigma^2)] \\ & + \sum_{j=1}^{n_C} (1 - X_{Cj}) [\log(1 - \gamma) + \log \varphi(y_{Cj} | \mu_{C-}, \sigma^2)] \end{aligned} \quad (4)$$

where $\Psi = (\gamma, \mu_{T+}, \mu_{T-}, \mu_{C+}, \mu_{C-}, \sigma^2)'$, and $y_{\text{obs}} = (y_{T1}, \dots, y_{Tn_T}, y_{C1}, \dots, y_{Cn_C})$.

Liu et al.^[7] provides procedures for implementation of the EM algorithm in conjunction with the bootstrap procedure for inference of θ in the patient population truly having the molecular target.

Let $\hat{\theta}$ be the estimator for the treatment effects in the patients truly having the molecular target obtained from the EM algorithm and S_B^2 denote the estimator of the variance of $\hat{\theta}$ obtained by the bootstrap procedure. It follows that the null hypothesis is rejected in Eq. 1 at the α level if

$$t = \left| \frac{\hat{\theta}}{\sqrt{S_B^2}} \right| \geq z_{\alpha/2}, \quad (5)$$

where $z_{\alpha/2}$ is the upper $100(\alpha/2)$ percentile of a standard normal distribution.

Liu et al.^[7] showed that the corresponding $(1 - \alpha)100\%$ asymptotic confidence interval for $\theta = \mu_{T+} - \mu_{C+}$ can be constructed as

$$\hat{\theta} \pm z_{1-\alpha/2} \sqrt{S_B^2} \quad (6)$$

They also pointed out that although the assumption that $\mu_{T+} - \mu_{C+} > \mu_{T-} - \mu_{C-}$ is one of the reasons for developing the targeted treatment, this assumption is not used in the EM algorithm for estimation of θ . Hence, the inference for θ by the proposed procedure is not biased in favor of the targeted treatment.

The procedure of application of the EM algorithm to the binary endpoints under the enrichment design can be found in Liu and Lin.^[6]

SIMULATION STUDY

Extensive simulation studies for comparing the performance of the EM procedures and traditional approach in statistical inference in terms of relative bias for estimation and size and power of hypothesis testing were conducted by Liu and Lin^[6] and Liu et al.^[7] for binary and continuous endpoints respectively. Some of their results are presented below.

First, they empirically demonstrated that the standardized $\hat{\theta}$ follows a standard normal distribution. Fig. 2 presents a Q-Q plot of 5000 standardized $\hat{\theta}$ versus the standard normal deviates. This Q-Q plot is a 45-degree straight line with a p-value of 0.2 for the normality by the Kolmogorov and Smirnov test. Therefore, empirical evidence from the simulation supports the normality of $\hat{\theta}$.

For continuous endpoints, Table 2 show that the absolute relative bias of the estimator for θ by the current method ranges from 8.5% to more than 50% and it increases as the PPV decreases. On the other hand, the absolute relative bias of the estimator by the EM algorithm does not exceed 4.6% while most of them are smaller than 2%. The variability has little impact on the bias of both methods. The relative bias of the current method without consideration of the true status of molecular target can be as high as 50% when the PPV is low. Consequently, the empirical coverage probabilities of the corresponding 95% confidence interval can be as low as only 8.6% when the PPV is 50%, mean difference is 20, standard deviation is 20, and n is 100. 14.1% of the 95% confidence intervals by the current method in Table 2 exceed 0.9449. On the contrary, only 9.4% of the 64 coverage probabilities of the 95% confidence intervals by the EM method are below 0.9449. However, 62 of the 64 coverage probabilities of the 95% confidence interval constructed by the EM algorithm are above 0.94. No coverage probability of the EM method is below 0.92. Table 3 provides the absolute relative bias and coverage probabilities

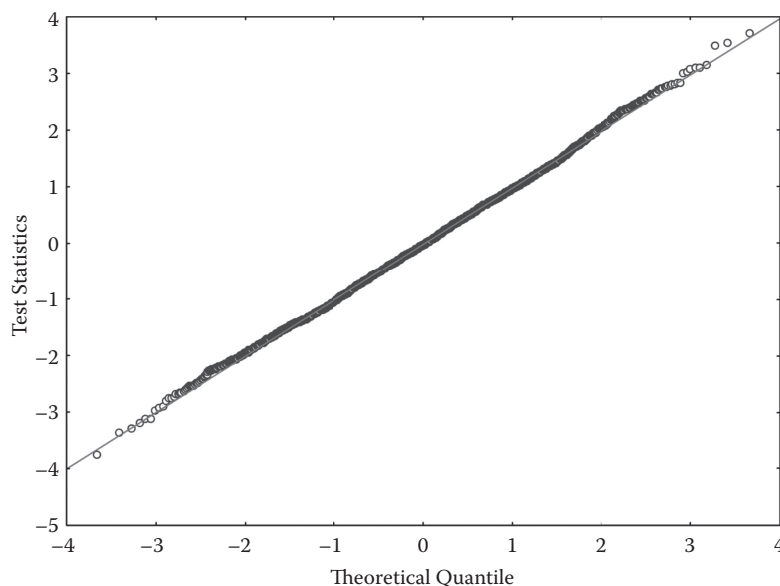


Fig. 2 The Quantile-Quantile plot when the positive predicted value is 0.9, $n = 50$, and $SD = 20$ for the continuous endpoints. (From Liu et al. [7])

for the binary data. From Table 3, similar conclusion can be reached for the binary endpoints.

The empirical powers of the simulation for the continuous and binary endpoints are given respectively in Tables 4 and 5. In addition, Fig. 3 presents the power curves for a two-sided test based on the continuous endpoints when $n = 50$, $\sigma = 20$ and PPV is 0.8 while Fig. 4 provides a comparison of the power curves for the one-sided hypothesis based on binary endpoints for PPV being 0.8. In Table 4, the power of the current method is an increasing function of the PPV. For both methods, the power increases as the sample size increases. However, the simulation results clearly demonstrate that the proposed testing procedure for the treatment effects based on the EM algorithm in the patient population truly having the molecular target is uniformly more powerful than the current method as depicted in Fig. 3. Additional simulation studies also show that the impact of the heterogeneity of variances and the value of $\mu_{T+} - \mu_{C+}$ on the bias, coverage probability, size and power is inconsequential. As shown in Table 5 and Fig. 4, similar results can be found for the binary endpoints.

SAMPLE SIZE DETERMINATION

The required sample size with incorporation of inaccuracy and uncertainty of diagnostic accuracy to achieve $1 - \beta$ power for rejecting the null hypothesis^[1] at the α significance level for a two-tailed alternative is given as

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \{2\sigma^2 + \gamma(1-\gamma)[(\mu_{T+} - \mu_{T-})^2 + (\mu_{C+} - \mu_{C-})^2]\}}{(\mu_{T+} - \mu_{C+})^2} \quad (7)$$

Liu et al.^[7] also conducted a small simulation to empirically investigate whether the proposed sample size can provide sufficient power. The results are reproduced in Table 6. From Table 6, the empirical power provided by the required sample size of the proposed EM method ranges from 0.7898 to 0.8076. On the other hand, the empirical power of the sample size for the unpaired two-sample t -test ranges from 0.4938–0.7174. Although the required sample sizes of the proposed EM procedure and the unpaired two-sample t -test are approximately the same, because the traditional unpaired two-sample t -test does not consider inaccuracy and uncertainty of the accuracy estimate for the diagnostic device, as demonstrated in Table 6, it fails to provide sufficient power. On the contrary, our proposed EM method can provide sufficient power at our targeted value of 80%. The simulation results provided in Table 6 are also consistent with the empirical powers given in Table 4. For example, in Table 6, when $\sigma = 20$, $\mu_{T+} - \mu_{C+} = 8$ and $PPV = 0.7$, the sample sizes of both methods are around 100 with a power 0.4938 and 0.8038, respectively, for the unpaired two-sample t -test and the proposed EM method. From Table 4, the empirical power for $\sigma = 20$, $\mu_{T+} - \mu_{C+} = 8$, $n = 100$ and $PPV = 0.7$ is 0.4866 and 0.7876, respectively, for the unpaired two-sample t -test and the proposed EM method.

DISCUSSION

Under the enrichment design, all patients must have a positive diagnosis for the molecular targets by the diagnostic device to be randomized to receive either the molecular targeted drug or the control treatment in the

Table 2 Relative Bias and the Coverage Probability for Continuous Endpoints

<i>n</i>	SD	Diff	PPV							
			0.5		0.7		0.8		0.9	
			Traditional	EM	Traditional	EM	Traditional	EM	Traditional	EM
100	20	2	−47.3983 ^a	−0.0013	−24.1185	4.2568	−16.6116	2.3478	−8.4813	0.9974
			0.9340 ^b	0.9536	0.9492	0.9516	0.9484	0.9488	0.9522	0.9522
		4	−49.0581	−1.5821	−30.0769	−1.6014	−19.8553	−0.9049	−9.1672	0.3048
			0.8840	0.9478	0.9268	0.9478	0.9414	0.9508	0.9468	0.9494
		6	−49.5276	−2.0853	−30.2413	−1.7987	−19.9750	−0.9682	−10.6642	−1.0877
			0.8252	0.9480	0.9010	0.9494	0.9308	0.9490	0.9458	0.9530
		8	−49.8556	−2.3385	−30.8579	−2.2230	−20.5837	−1.2181	−9.6916	0.4514
			0.7214	0.9480	0.8612	0.9516	0.9086	0.9490	0.9400	0.9496
		10	−50.2429	−2.4300	−30.0024	−1.0020	−19.6690	0.5666	−10.4483	1.2284
			0.5922	0.9508	0.8088	0.9484	0.8966	0.9502	0.9332	0.9472
	60	12	−49.4109	−1.6261	−29.4555	0.1289	−19.5609	1.6223	−9.7678	2.3619
			0.4664	0.9526	0.7700	0.9484	0.8636	0.9464	0.9286	0.9484
		15	−49.5798	−1.2384	−30.1965	0.3199	−19.9585	1.1326	−9.9061	1.8471
			0.2810	0.9532	0.6636	0.9490	0.8222	0.9468	0.9162	0.9478
		20	−50.1847	4.0625	−29.9946	−0.0939	−20.0528	0.0681	−9.9316	0.9709
			0.0886	0.9248	0.4832	0.9436	0.7270	0.9472	0.8872	0.9440
		6	−50.9761	−3.5186	−28.6085	−0.1531	−23.5692	−4.5887	−12.3944	−2.9119
			0.9358	0.9518	0.9500	0.9536	0.9448	0.9500	0.9534	0.9534
		12	−50.7794	−3.3462	−30.0373	−1.6684	−20.8976	−1.8963	−11.3324	−1.8289
			0.8904	0.9486	0.9250	0.9490	0.9424	0.9514	0.9468	0.9502
		18	−50.6359	−3.0353	−29.9622	−1.4980	−20.6520	−1.6591	−10.4509	−0.8401
			0.8208	0.9458	0.9002	0.9480	0.9292	0.9560	0.9454	0.9500
		24	−49.7365	−2.1908	−29.8482	−1.2946	−19.5856	−0.2424	−10.3801	−0.0652
			0.7136	0.9486	0.8706	0.9522	0.9164	0.9504	0.9394	0.9500
		30	−49.4661	−1.9321	−30.0834	−1.2077	−20.1803	0.1411	−9.8132	1.5869
			0.5948	0.9450	0.8208	0.9510	0.8888	0.9454	0.9336	0.9494
		36	−50.0172	−2.1800	−30.2903	−0.6902	−20.2224	0.6928	−10.4163	1.3597
			0.4618	0.9400	0.7550	0.9490	0.8654	0.9504	0.9264	0.9548
		45	−49.8378	−1.2902	−29.6934	0.4980	−19.3310	1.5422	−9.9478	2.1432
			0.2796	0.9494	0.6672	0.9542	0.8356	0.9518	0.9166	0.9426
100	55		−50.4262	1.5676	−29.7992	0.6112	−19.9513	0.8690	−10.1621	1.3317
			0.1212	0.9356	0.5508	0.9504	0.7528	0.9422	0.9062	0.9490

^aRelative bias (%)
^bCoverage probability.
(From Liu et al. [7]).

targeted clinical trials. However, because no diagnostic device is perfect with 100% PPV, some patients with a positive result may in fact not have the molecular target. As a result, the current estimation method may produce a biased estimator for the treatment effects of the molecular targeted test drug in the patients truly having the molecular target. EM algorithm can be applied to incorporate information of the PPV for inference of the treatment effects in the patient population truly having with the

molecular target. Simulation results clearly show not only that the EM estimation method is unbiased and can provide sufficient coverage probability but also that the EM testing procedure can adequately control the type I error rate at the nominal level and is uniformly more powerful than the current method. Liu et al.^[7] has suggested that the traditional sample means, $\bar{Y}_T$ and $\bar{Y}_C$, of the test and control treatments are reasonable initial values for the proposed method.

Table 3 Relative Bias (%) and the Coverage Probability for Binary Endpoints

<i>n</i>	Diff	PPV							
		0.5		0.7		0.8		0.9	
		Traditional	EM	Traditional	EM	Traditional	EM	Traditional	EM
100	0.05	−50.3520	−0.7835	−28.7879	0.7974	−20.9639	−1.0978	−7.0919	2.7808
		0.9908	0.9800	0.9858	0.9760	0.9826	0.9760	0.9808	0.9760
	0.1	−50.6080	−0.9713	−30.8180	−0.9572	−20.8140	−0.8657	−10.4360	−0.4761
		0.9968	0.9798	0.9934	0.9848	0.9860	0.9746	0.9842	0.9758
	0.15	−50.4413	−0.7302	−29.5573	0.2144	−20.0920	−0.2561	−11.1320	−1.1958
		0.8056	0.9640	0.8946	0.9580	0.9136	0.9476	0.9314	0.9462
	0.2	−50.7330	−0.6837	−29.9350	−0.1267	−19.7950	0.0854	−9.8150	0.0737
		0.6012	0.9654	0.8120	0.9570	0.8902	0.9584	0.9368	0.9616

PPV: positive predictive value. *n*: sample size per group.

Upper row: Relative bias; lower row: coverage probability

(From Liu and Lin [6]).

Table 4 Comparison of Empirical Powers for Continuous Endpoints, SD = 20

<i>n</i>	Diff	PPV							
		0.5		0.7		0.8		0.9	
		Traditional	EM	Traditional	EM	Traditional	EM	Traditional	EM
50	4	0.0828	0.1606	0.1064	0.1580	0.1418	0.1764	0.1436	0.1664
	8	0.1724	0.4756	0.2794	0.4912	0.3446	0.4858	0.4266	0.5048
	12	0.3030	0.7962	0.5330	0.8250	0.6496	0.8218	0.7498	0.8204
	20	0.6378	0.9886	0.9050	0.9968	0.9700	0.9964	0.9918	0.9988
100	4	0.1066	0.2812	0.1666	0.2844	0.2020	0.2926	0.2490	0.2942
	8	0.2934	0.7850	0.4866	0.7876	0.6076	0.7910	0.7158	0.7908
	12	0.5584	0.9820	0.8320	0.9838	0.9146	0.9804	0.9672	0.9862
	20	0.9064	1.0000	0.9968	1.0000	0.9996	1.0000	1.0000	1.0000
200	4	0.1684	0.4998	0.2980	0.5248	0.3546	0.5106	0.4338	0.5112
	8	0.5164	0.9728	0.7986	0.9806	0.8886	0.9762	0.9432	0.9730
	12	0.8338	1.0000	0.9888	1.0000	0.9956	0.9998	0.9990	0.9998
	20	0.9974	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

(From Liu et al. [7]).

Table 5 Comparison of Empirical Powers for Binary Data

<i>n</i>	Difference	PPV							
		0.5		0.7		0.8		0.9	
		Traditional	EM	Traditional	EM	Traditional	EM	Traditional	EM
100	0.05	0.0698	0.1270	0.0878	0.1344	0.1028	0.1260	0.1180	0.1346
	0.1	0.1218	0.4054	0.2052	0.3952	0.2662	0.3944	0.3270	0.3944
	0.15	0.2336	0.7828	0.4308	0.7794	0.5382	0.7712	0.6480	0.7634
	0.2	0.3794	0.9834	0.6874	0.9824	0.8202	0.9816	0.9212	0.9778

PPV: positive predictive value. *n*: sample size per group.

(From Liu and Lin [6]).

In this entry, we have only considered continuous and binary endpoints. However, the inferential procedures for the treatment effects based on the censored endpoints for the molecularly targeted test drug in the patients truly having the

molecular target require further research. Bayesian method is another approach to incorporating the uncertainty in accuracy of the diagnostic device for the molecular target into the inference of the treatment effects of the targeted drug.

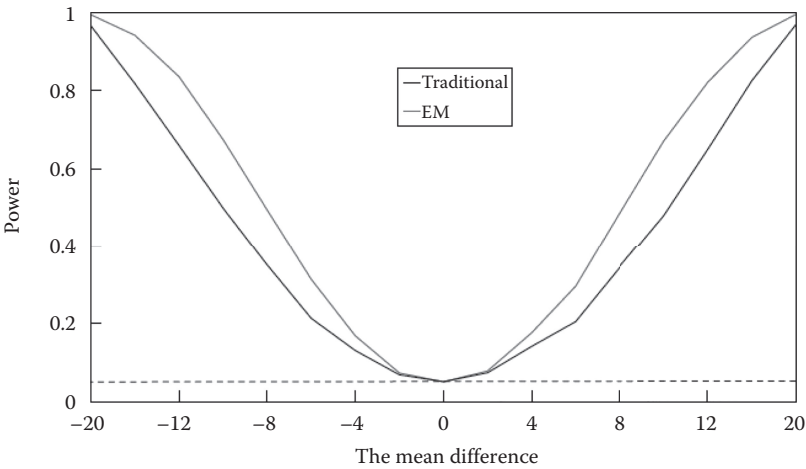


Fig. 3 Empirical power curve when the positive predicted value is 0.8, $n = 50$ and $SD = 20$ for the continuous endpoint. (From Liu et al. [7])

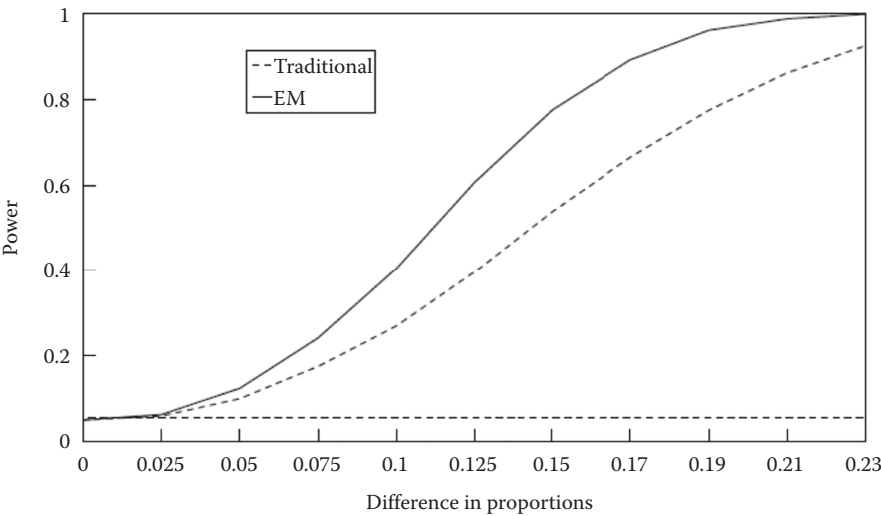


Fig. 4 Empirical Power curve when PPV is 0.8 for binary endpoints. (From Liu et al. [7])

Table 6 Sample Sizes and Empirical Powers for Continuous Endpoints

		PPV					
		0.7		0.8		0.9	
SD	Diff	Traditional ^a (power)	EM ^b (power)	Traditional ^a (power)	EM ^b (power)	Traditional ^a (power)	EM ^b (power)
20	8	99(0.4938)	100(0.8038)	99(0.6108)	100(0.8002)	99(0.6958)	99(0.7956)
	10	63(0.5034)	65(0.7960)	63(0.6050)	65(0.7940)	63(0.7066)	64(0.7916)
	12	44(0.4784)	46(0.7998)	44(0.5880)	45(0.7928)	44(0.7036)	45(0.7986)
40	8	393(0.5008)	395(0.8000)	393(0.6004)	394(0.8076)	393(0.7032)	394(0.8036)
	10	252(0.4916)	253(0.8062)	252(0.6184)	253(0.8018)	252(0.7120)	252(0.8018)
	12	175(0.4964)	177(0.8062)	175(0.5962)	176(0.8042)	175(0.7082)	176(0.8026)
60	8	884(0.5000)	885(0.8052)	884(0.5998)	885(0.7898)	884(0.7092)	884(0.8048)
	10	566(0.5006)	567(0.7964)	566(0.6114)	567(0.8024)	566(0.7082)	566(0.8016)
	12	393(0.5120)	395(0.8048)	393(0.6214)	394(0.7928)	393(0.7174)	394(0.7992)

^aThe sample size required for the traditional unpaired two-sample t -test.

^bThe sample size required for the proposed EM procedure.

(From Liu et al. [7]).

ACKNOWLEDGMENT

This research is partially supported by the Taiwan National Science Council Grants: NSC 95 2118-M-002-007-MY2 and 972118-M-002-002-MY2 to J.P. Liu.

REFERENCES

1. U.S. FDA. *Guidance on Pharmacogenetic Tests and Genetic Tests for Heritable Marker*; The U.S. Food and Drug Administration: Rockville, MD, 2007.
2. U.S. FDA. *Draft Guidance on In Vitro Diagnostic Multivariate Index Assays*; The U.S. Food and Drug Administration: Rockville, MD, 2007.
3. U.S. FDA. *The draft concept paper on Drug-Diagnostic Co-Development*; The U.S. Food and Drug Administration: Rockville, MD, 2005.
4. Chow, S.C.; Liu, J.P. *Design and Analysis of Clinical Trials*, 2nd Ed.; Wiley: New York, 2004.
5. Liu, J.P.; Chow, S.C. Issues on the diagnostic multivariate index assay and targeted clinical trials. *J. Biopharm. Stat.* **2008**, *18*, 167–182.
6. Liu, J.P.; Lin, J.R. Statistical methods for targeted clinical trials under enrichment design. *J. Formo. Med. Assoc.* **2008**, *107*, S35–S42.
7. Liu, J.P.; Lin, J.R.; Chow, S.C. Inference on treatment effects for targeted clinical trials under enrichment design. *Pharm. Stat.* **2009**, *8* (4), 356–370.
8. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Roy. Stat. Soc. B.* **1977**, *39*, 1–38.
9. McLachlan, G.J.; Krishnan, T. *The EM Algorithm and Extensions*; Wiley: New York, 1997.
10. McLachlan, G.J.; Peel, D. *Finite Mixture Models*; Wiley: New York, 2000.
11. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman and Hall: New York, 1993.
12. Fleiss, J.L.; Levin, B.; Paik, M.C. *Statistical Methods for Rates and Proportions*, 3rd Ed.; Wiley: New York, 2003.

Testing for Qualitative Interaction

Xin Yan

Merck Research Laboratories, Blue Bell, Pennsylvania, U.S.A.

Xiaogang Su

Department of Statistics and Actuarial Science, University of Central Florida,
Orlando, Florida, U.S.A.

Abstract

This entry reviews the most commonly used statistical methods designed for testing against qualitative interaction by summarizing the testing procedures and providing tables for illustration of each. It also discusses the advantages and disadvantages of each method in terms of applicability.

INTRODUCTION

In clinical trials involving comparison of two treatment groups, often investigators wish to know whether and how the between-treatment difference varies across patient subsets, which may be defined by prognostic factors such as age, gender, disease severity status, or clinical centers. This leads to the consideration of the treatment-by-patient-subset interaction.

While interaction may exist in numerous ways and with various patterns, Peto^[1] first distinguished *qualitative interaction* from *quantitative interaction*. A quantitative interaction, which is frequently referred to as noncross-over interaction, occurs if the treatment effect varies in magnitude but not in direction across all patient subsets; qualitative interaction, or crossover interaction, refers to the situation where the between-treatment difference changes direction among patient subsets. The existence of quantitative interaction is subject to the measurement scale and is relatively less problematic in terms of its clinical implications, whereas the presence of qualitative interaction cannot be eliminated by transformations and often has therapeutic significance. Therefore, it is important to determine whether apparent crossovers in treatment effect among patient subsets are mainly attributed to chance of variation.

In this article, we review the most commonly used statistical methods designed for testing against qualitative interaction. Our main focus is on the following three methods: the likelihood ratio test (LRT) proposed by Gail and Simon (GS),^[2] the range test considered by Piantadosi and Gail,^[3] and the multiple comparison testing procedure suggested by Pan and Wolfe.^[4] We summarize each testing procedure by outlining the authors' formulation of the test statistics and providing tables of critical values. We also discuss the advantages and pitfalls of each method in terms of power and applicability in practice. In addition,

some follow-up variations and enhancements available in the literature are briefly mentioned as well.

TESTS FOR QUALITATIVE INTERACTION

Assume that two treatments are being compared across I patient subsets. Let μ_{ij} be the expected response of the j th treatment in the i th subset where $j = 1, 2$ and $\delta_i = \mu_{i1} - \mu_{i2}$ be the treatment effect in the i th subset for $i = 1, \dots, I$. As pointed out by Russek-Cohen and Simon,^[6] μ_{ij} can be thought of as the log odds in logistic regression, and hence δ_i is the log odds ratio for the i th subset. When a Cox proportional hazards model is used to model survival times, μ_{ij} can be thought of as the logarithm of hazard, and hence δ_i is the logarithm of relative risk between the two treatment groups for the i th patient subset.

Let D_i be the estimated treatment effect for the i th patient subset. We assume that D_i 's are independent and normally distributed with mean δ_i and known variance $\sigma_i^2 = w_i \sigma^2$, where w_i is a known constant that depends on the sample size and σ^2 is typically the population error variance. Note that this formulation is generally enough because asymptotic normality broadly holds for various estimators. To test for no qualitative interaction, the null hypothesis is that δ_i 's are all non-negative or all nonpositive.

Likelihood Ratio Test

The LRT proposed by Gail and Simon^[2] is the most influential and inspiring contribution in testing for qualitative interaction. To illustrate, let $\delta = (\delta_1, \dots, \delta_I)'$, $O^+ = \{\delta : \delta_i \geq 0, \text{all } i\}$, and $O^- = \{\delta : \delta_i \leq 0, \text{all } i\}$. Therefore, we wish to test

$$H_0 : \delta \in O^+ \cup O^- \text{ vs. } H_1 : \text{not } H_0 \quad (1)$$

The likelihood ratio for this hypothesis is

$$\frac{\max_{\delta \in O^+ \cup O^-} \exp \sum_{i=1}^I \left\{ -(D_i - \delta_i)^2 / (2w_i \sigma^2) \right\}}{\max_{\delta} \exp \sum_{i=1}^I \left\{ -(D_i - \delta_i)^2 / (2w_i \sigma^2) \right\}}$$

The denominator reaches the maximum at $D_i = \delta_i$ and the LRT is thus

$$\max_{\delta \in O^+ \cup O^-} \exp \sum_{i=1}^I \left[-(D_i - \delta_i)^2 / w_i \right] < k \cdot \sigma^2$$

where the constant k is chosen such that the probability of rejecting the null hypothesis does not exceed the significant level α . The preceding maximization problem can be simplified further so that Gail and Simon's test rejects the null hypothesis of no qualitative interaction if both

$$Q^+ = \sum_{i=1}^I \left\{ D_i^2 / w_i \right\} I(D_i > 0) > c \cdot \sigma^2$$

and

$$Q^- = \sum_{i=1}^I \left\{ D_i^2 / w_i \right\} I(D_i < 0) > c \cdot \sigma^2$$

where $I(\cdot)$ is the indicator function and $c = -2\log(k)$. The quantities Q^+ and Q^- are the minimum values of $\sum_{i=1}^I [(D_i - \delta_i)^2 / w_i]$ over O^+ and O^- , respectively. Therefore, the LRT can be simply expressed as

$$\min(Q^+, Q^-) / \sigma^2 > c \quad (2)$$

As shown in GS, the critical value c is determined by the following equation:

$$\sum_{j=1}^{I-1} B(j; n = I - 1, p = 0.5) \{1 - \chi_j^2(c)\} = \alpha \quad (3)$$

where $B(j; n, p)$ is the binomial mass function with parameters n and p ; $\chi_j^2(\cdot)$ is the cumulative distribution function (cdf) of the chi-square distribution with j degrees of freedom. Gail and Simon gave a table, as reproduced in Table 1, for critical values c at significance level 0.001, 0.025, 0.05, 0.10, and 0.20 for the LRT. The above method can also be used to test the hypothesis that all treatment differences are greater (or less) than a prespecified threshold d by simply applying the method to the translated estimate $D_i - d$, which appears to be a useful and practical extension.

The variance σ_i^2 for each patient subset is assumed known in GS. Even though the LRT procedures are valid for large samples if consistent estimators of σ^2 are employed instead, Silvapulle^[7] noticed that the asymptotic critical values given in Table 1 might not be appropriate for small samples, which motivated him to derive the exact finite sample results.

Table 1 Critical Values c for the LRT $\min(Q^+, Q^-) / \sigma^2 > c$

Number of groups	Significance level				
	0.20	0.10	0.05	0.025	0.001
2	0.71	1.64	2.71	3.84	9.35
3	1.73	2.59	4.23	5.54	11.76
4	2.59	4.01	5.43	6.86	13.47
5	3.39	4.96	6.50	8.02	14.95
6	4.14	5.84	7.48	9.09	16.32
7	4.86	6.67	8.41	10.09	17.58
8	5.56	7.48	9.29	11.05	18.78
9	6.75	8.26	10.15	11.98	19.93
10	6.92	9.02	10.99	12.87	21.03
12	8.24	10.50	12.60	14.57	23.13
14	9.53	11.93	14.15	16.25	25.17
16	10.79	13.33	15.66	17.85	27.10
18	12.03	14.70	17.13	19.41	28.96
20	13.26	16.04	18.57	20.94	30.79
25	16.28	19.34	22.09	24.65	35.15
30	19.25	22.55	25.50	28.23	39.33

(From Ref. [2].)

Let S^2 be an estimate of σ^2 such that $\nu S^2 / \sigma^2 \sim \chi_\nu^2$ and S^2 is independent of D_i 's. Typically, S^2 can be chosen as the usual mean square error. Silvapulle^[7] constructed a test statistic

$$F_Q = \frac{\min(Q^+, Q^-)}{S^2} \quad (4)$$

which resembles the usual F -ratio. The exact α -level critical value c for the test with unknown variance is given by the solution of the following equation:

$$\sum_{j=1}^{I-1} B(j; n = I - 1, p = 0.5) \Pr\{F_{j,\nu} > c/j\} = \alpha \quad (5)$$

where $B(j; n = I - 1, p = 0.5)$ is the binomial probability of j successes out of $I - 1$ trials with a probability of success 0.5. $F_{j,\nu}$ is the F distribution with degrees of freedom j and ν . If the sample value of FQ is f_q , then the exact p -value is

$$\sum_{j=1}^{I-1} B(j; n = I - 1, p = 0.5) \Pr\{F_{j,\nu} > f_q/j\} \quad (6)$$

Silvapulle^[7] provided a table (Table 2) that covers a wide range of possible values of ν , the total patient subsets I , and selected sets of critical values that is useful for practical purposes.

The LRT for the qualitative interaction can be extended to the situation where between-treatment difference D_i 's are dependent and have an arbitrary covariance matrix Σ . This typically happens when treatment effect is estimated with respect to some extraneous covariates. The test rejects H_0 if both

Table 2 Exact^a Critical Values for the *F*-Type Test Against Qualitative Interaction at 5% Level

ν^b	Number of groups										
	2	3	4	5	6	7	8	9	10	12	14
4	4.54	—	—	—	—	—	—	—	—	—	—
5	4.06	—	—	—	—	—	—	—	—	—	—
6	3.78	6.45	—	—	—	—	—	—	—	—	—
7	3.59	6.05	—	—	—	—	—	—	—	—	—
8	3.46	5.77	7.72	—	—	—	—	—	—	—	—
9	3.36	5.56	7.41	—	—	—	—	—	—	—	—
10	3.29	5.41	7.17	8.79	—	—	—	—	—	—	—
12	3.18	5.18	6.84	8.34	9.76	—	—	—	—	—	—
14	3.10	5.03	6.61	8.04	9.39	10.68	—	—	—	—	—
16	3.05	4.92	6.44	7.82	9.12	10.36	11.56	—	—	—	—
18	3.01	4.83	6.32	7.66	8.92	10.12	11.28	12.42	—	—	—
20	2.97	4.77	6.22	7.53	8.76	9.93	11.06	12.17	13.25	—	—
25	2.92	4.65	6.05	7.31	8.48	9.60	10.68	11.73	12.76	14.77	—
30	2.88	4.58	5.94	7.17	8.30	9.39	10.43	11.45	12.44	14.38	16.26
40	2.84	4.49	5.81	6.99	8.09	9.13	10.13	11.11	12.06	13.91	15.71
50	2.81	4.43	5.73	6.89	7.96	8.98	9.96	10.91	11.84	13.64	15.38
70	2.78	4.38	5.65	6.77	7.82	8.81	9.76	10.69	11.59	13.33	15.02
90	2.76	4.34	5.60	6.71	7.74	8.72	9.66	10.57	11.45	13.16	14.82
110	2.75	4.32	5.57	6.67	7.69	8.66	9.59	10.49	11.36	13.06	14.70
200	2.73	4.28	5.51	6.59	7.60	8.55	9.46	10.34	11.19	12.85	14.45
∞^c	2.71	4.23	5.43	6.50	7.48	8.41	9.29	10.15	10.99	12.60	14.15

^aThe observations are assumed to be independent and normal with a common variance.

^bThe parameter ν is the degrees of freedom for the estimate of error.

^cThe critical values in the last line corresponding to d.f. = ∞ are the same as those in Ref. [2]. (From Ref. [7].)

$$Q^+ = \min_{\Delta \in O^+} [(D - \delta)' \Sigma^{-1} (D - \delta)] > c$$

and

$$Q^- = \min_{\Delta \in O^-} [(D - \delta)' \Sigma^{-1} (D - \delta)] > c$$

The difficulty involved in the above minimization problem is substantial. It should be solved by quadratic programming,^[8] and the critical value c usually depends on the covariance matrix Σ . This idea has been further explored by Russek-Cohen and Simon^[6] in their qualitative interaction tests for multifactor studies. Various topics concerning the hypothesis to test for normal means based on quadratic form under the constraints of linear inequalities or equalities can be found in Refs. ^[6,9–11]

The statistical properties of the LRT were discussed by Zelterman.^[12] Let $d(\mathbf{D})$, where $\mathbf{D} = (D_1, \dots, D_I)'$, denote a decision rule for testing H_0 vs. H_1 such that one rejects H_0 if $d(\mathbf{D}) = 1$ and cannot reject H_0 if $d(\mathbf{D}) = 0$. Since the labels 1, 2, ..., I can be assigned arbitrarily, the decision rule $d(\mathbf{D})$ is invariant under any permutation of the elements of $\mathbf{D}$. Furthermore, this decision rule is invariant if signs of all D_i 's are changed simultaneously. Zelterman^[12] also pointed out that the vector of true mean difference

$\delta = \mathbf{0}$ poses the most difficult practical situation and is generally where the LRT tends to have the least power and the largest bias.

Silvapulle^[7] modified the LRT based on M -estimators in order to achieve robust power against long-tailed error distributions. In addition, Zelterman and Kershner^[13] and Berger^[11] each introduced a uniformly more powerful test for the qualitative interaction in the case of two patient subsets by augmenting the critical region of the GS test with disjoint regions in the sample space. Nevertheless, Russek-Cohen and Simon^[6] commented that “these tests have some nonintuitive properties and would not be accepted by nonstatisticians.”

The Range Test

Piantadosi and Gail^[3] proposed a test procedure for detecting the qualitative interaction based on the range test. They compared the power of the range test with that of the LRT developed by Gail and Simon.^[2] The range test rejects H_0 at a level α if both

$$\max_{1 \leq i \leq I} \{D_i / \sigma_i\} > C_{2\alpha} \quad \text{and} \quad \min_{1 \leq i \leq I} \{D_i / \sigma_i\} < -C_{2\alpha} \quad (7)$$

where $C_2\alpha$ is the critical value corresponding to the test level α . Table 3 provides the critical values for a two-sided test with respect to various numbers of subsets and test levels.

The above method can be conveniently adopted for one-sided hypotheses, e.g., $H_{01}: \Delta \in O^+$, which intends to test the “better” treatment. For the one-sided test, one rejects the null hypothesis H_{01} if $\min\{D_i/\sigma_i\} < -C_{1\alpha}$. The critical value for situations with I subsets is equal to the entry $I+1$ in Table 3. That is, $-C_{1\alpha}$ for I subsets equals $-C_{2\alpha}$ for $I+1$ subsets.

Piantadosi and Gail^[3] compared the power of the LRT with the power of the range test. To this end, they generated independent normal variates D_i for $i = 1, 2, \dots, I$ with mean δ_i and variance 1 from the uniform distribution using the Box–Muller^[14] method. The McNemar^[15] procedure was used to compare the LRT with the range test for the significance of differences in proportions from the matched samples. It was found through simulations that the range test has more power than the LRT when the new treatment is harmful only in very few subsets in terms of δ_i 's, while the LRT often has more power when the new treatment is harmful in several subsets in terms of δ_i 's. Pan and Wolfe^[5] reached the conclusion that the power of the range test has characteristics similar to those of the LRT.

Test Procedure Based on Simultaneous Inference

In addition to the testing procedures based on the likelihood ratio and the range test, Pan and Wolfe^[4] proposed a procedure to test for qualitative interaction based on simultaneous confidence intervals.

Since a small directional change in treatment differences across patient subsets may not have practical importance, Pan and Wolfe^[4] suggested distinguishing *severe* qualitative interaction from *slight* qualitative interaction. Correspondingly, they introduced a predetermined margin $d > 0$ to divide a clinically insignificant treatment effect from a clinically significant treatment effect. And they modified the hypotheses for testing no qualitative interaction as follows:

$$H_0(d): \{I \text{ of } \delta_i > -d \text{ or } I \text{ of } \delta_i < d \text{ for } 1 \leq i \leq I\}$$

$$H_1(d): \{\text{at least one } \delta_i < -d \text{ and one } \delta_j > d,$$

$$\text{where } 1 < i \neq j \leq I\} \quad (8)$$

Clearly, the margin d depends on the nature and objective of the trial. Often, it has clinical implications and should be predetermined based on the relevant clinical criteria. To draw a statistical inference on the qualitative interaction, they proposed the following test procedure $C(d)$, which is based on simultaneous z -intervals.

Table 3 Critical Values $C_2\alpha$ for the Two-Sided Range Test

Number of groups	Significance level				
	0.20	0.10	0.05	0.025	0.001
2	0.84	1.28	1.64	1.96	3.09
3	1.25	1.63	1.95	2.24	3.29
4	1.46	1.82	2.12	2.39	3.40
5	1.60	1.94	2.23	2.49	3.48
6	1.71	2.04	2.32	2.57	3.54
7	1.79	2.11	2.39	2.63	3.59
8	1.86	2.17	2.44	2.69	3.63
9	1.92	2.22	2.49	2.73	3.66
10	1.97	2.27	2.53	2.77	3.69
12	2.05	2.34	2.60	2.83	3.74
14	2.12	2.41	2.66	2.89	3.78
16	2.18	2.46	2.71	2.93	3.82
18	2.23	2.50	2.75	2.97	3.85
20	2.27	2.54	2.78	3.00	3.88
25	2.36	2.62	2.86	3.07	3.93
30	2.42	2.68	2.92	3.13	3.98

(From Ref. [3].)

1. Compute the joint confidence level $P_E = 2(1 - \alpha)^{1/(I-1)} - 1$
2. Construct a z -confidence interval (L_i, U_i) with confidence level P_E for each patient subset i , $i = 1, \dots, I$, i.e.,

$$(L_i, U_i) = \left[D_i - z \left(\frac{1 - P_E}{2} \right) \sigma_i, D_i + z \left(\frac{1 - P_E}{2} \right) \sigma_i \right], \quad 1 \leq i \leq I \quad (9)$$

where $z(\gamma)$ is the upper γ critical point of a standard normal distribution.

3. Reject null hypothesis $H_0(d)$ if there is a pair $1 \leq i \neq j \leq I$ such that

$$U_i < -d \text{ and } L_j > d \quad (10)$$

Otherwise, do not reject $H_0(d)$.

The testing procedure is easy to implement, and it is essentially a Bonferroni approach. It should be noted that the test procedure could be very conservative, especially when I , the number of patient subsets, is quite large. Pan and Wolfe tabulated simultaneous confidence coefficients for the calculations of the confidence intervals (Table 4). They also showed that the proposed test has a size not greater than α for any given value of d . That is,

$$\sup_{H(d, \delta)} \Pr[C(d) \text{ rejects } H_0] \leq \alpha$$

Table 4 Confidence Coefficients Used to Construct $C(d)$

Number of groups	Significance level				
	0.20	0.10	0.05	0.025	0.001
2	0.60000	0.80000	0.90000	0.95000	0.99800
3	0.78885	0.89736	0.94936	0.97484	0.99900
4	0.85664	0.93098	0.96610	0.98319	0.99933
5	0.89148	0.94801	0.97452	0.98738	0.99950
6	0.91270	0.95830	0.97959	0.98990	0.99960
7	0.92699	0.96519	0.98298	0.99158	0.99967
8	0.93725	0.97012	0.98540	0.99278	0.99971
9	0.94498	0.97383	0.98722	0.99368	0.99975
10	0.95102	0.97672	0.98863	0.99438	0.99978
12	0.95984	0.98094	0.99070	0.99540	0.99982
14	0.96596	0.98386	0.99212	0.99611	0.99985
16	0.97047	0.98600	0.99317	0.99663	0.99987
18	0.97392	0.98764	0.99397	0.99702	0.99988
20	0.97665	0.98894	0.99461	0.99734	0.99990
25	0.98149	0.99124	0.99573	0.99789	0.99992
30	0.98467	0.99275	0.99647	0.99825	0.99993

(From Ref. [4].)

An additional feature is that the power function is explicitly available:

$$\begin{aligned}
& 1 - \prod_{i=1}^I \Phi \left\{ \frac{d - \delta_i}{\sigma_i} - z \left[1 - (1 - \alpha)^{1/(I-1)} \right] \right\} \\
& - \prod_{i=1}^I \Phi \left\{ \frac{d + \delta_i}{\sigma_i} - z \left[1 - (1 - \alpha)^{1/(I-1)} \right] \right\} \\
& + \prod_{i=1}^I \left[\Phi \left\{ \frac{d - \delta_i}{\sigma_i} - z \left[1 - (1 - \alpha)^{1/(I-1)} \right] \right\} \right. \\
& \left. - \Phi \left\{ \frac{-d - \delta_i}{\sigma_i} - z \left[1 - (1 - \alpha)^{1/(I-1)} \right] \right\} \right] \quad (11)
\end{aligned}$$

Note that the conclusion of Pan and Wolfe's^[4] method depends on whether or not there is a pair of confidence intervals lying on both sides of the interval $(-d, d)$, which renders the procedure more suitable for analysis involving a small or a medium number of patient subsets. For example, consider an equivalence trial conducted in a large number of centers. The conclusion regarding the existence of qualitative interaction may be dramatically affected by the inclusion of a couple of centers that happen to yield outlying treatment effect estimates on both sides of the equivalence margins. In order to circumvent this difficulty, Yan^[16] modified the testing procedure by Pan and Wolfe^[4] so that the total number of patient subsets is taken into account in assessing qualitative treatment-by-center interaction.

The method proposed by Yan^[16] introduces a tuning parameter k to quantify the number of outlying centers that is tolerable. The tuning parameter k is imbedded into the hypothesis as follows:

$$H_0(d, k) : \{I - k \text{ of } \delta_i > -d\} \text{ or}$$

$$\{I - k \text{ of } \delta_i < d\}$$

$$H_1(d, k) : \{\text{at least } k + 1 \text{ of } \delta_i < -d$$

$$\text{and another } k + 1 \text{ of } \delta_i > d\} \quad (12)$$

Note that the null hypothesis above is identical to that of Pan and Wolfe's method when $k = 0$. However, the null hypothesis above allows for $k > 0$. This provides the flexibility for selection of a fixed number of centers whose confidence intervals on the between-treatment difference are on both outside of the equivalence margins without indicating qualitative interaction. In fact, k can be considered as a tuning parameter to quantify the tolerance level of the qualitative interaction and should be carefully predetermined based on both the nature of the clinical trial and the total number of investigational centers. This needs to be paid substantial attention in order to prevent the proposed testing procedure from being rendered suspect due to the selection of a value of k . Yan^[16] recommended that k should be less than 5% of the total number of the centers when the design is balanced. Moreover, for an unbalanced design, it is necessary to use a sample size weighting scheme to calculate the simultaneous confidence intervals. Since large centers would be more informative than small

centers, if the confidence intervals on δ_i 's for large centers are completely on both outsides of the equivalence margins, the qualitative interaction may not be negligible.

Based on the hypothesis above, qualitative interaction is confirmed if at least $k + 1$ confidence intervals of the between-treatment difference are entirely above the upper equivalence margin d and at least another $k + 1$ confidence intervals of the between-treatment difference are completely below the lower equivalence margin $-d$. Because of the introduction of the tuning parameter k , the construction of I simultaneous confidence intervals should be based on a confidence level of

$$P_E = 2 \cdot (1 - \alpha)^{1/(I-k-1)} - 1$$

as derived in Ref. [16]. The size of the test is shown to be less than α . Therefore, the introduction of k allows for drawing a valid statistical conclusion without subjectively removing those “extremely bad” centers. Moreover, Yan^[16] also gave the explicit power function:

$$\begin{aligned} & 1 - \prod_{i \in A} \Phi \left[\frac{d + \delta_i}{\sigma_i} + z \left(\frac{1 - P_E}{2} \right) \right] \\ & \times \prod_{j = j_1, j_2, \dots, j_l \in A^c} \Phi^{\omega(j)} \left[\frac{-d - \delta_j}{\sigma_j} - z \left(\frac{1 - P_E}{2} \right) \right] \\ & - \prod_{i \in B} \Phi \left[\frac{d - \delta_i}{\sigma_i} + z \left(\frac{1 - P_E}{2} \right) \right] \\ & \times \prod_{j = j'_1, j'_2, \dots, j'_m \in B^c} \Phi^{\omega(j)} \left[\frac{-d + \delta_j}{\sigma_j} - z \left(\frac{1 - P_E}{2} \right) \right] \\ & + \prod_{i \in C} \left\{ \Phi \left[\frac{d - \delta_i}{\sigma_i} + z \left(\frac{1 - P_E}{2} \right) \right] \right. \\ & \left. - \Phi \left[\frac{-d - \delta_i}{\sigma_i} - z \left(\frac{1 - P_E}{2} \right) \right] \right\} \\ & \times \prod_{j = j_1, j_2, \dots, j_l \in A^c} \Phi^{\omega(j)} \left[\frac{-d - \delta_j}{\sigma_j} - z \left(\frac{1 - P_E}{2} \right) \right] \\ & \times \prod_{j = j'_1, j'_2, \dots, j'_m \in B^c} \Phi^{\omega(j)} \left[\frac{-d + \delta_j}{\sigma_j} - z \left(\frac{1 - P_E}{2} \right) \right] \end{aligned} \quad (13)$$

where the set A is the index set for the confidence intervals (L_i, U_i) whose upper bound $U_i \geq -d$, i.e., $A = \{i \mid U_i \geq -d, \text{ for all } i\}$; the set B is the index set for the confidence intervals (L_i, U_i) whose lower bounds $L_i \leq d$, i.e., $B = \{i \mid L_i \leq d, \text{ for all } i\}$; and the sets A^c and B^c are the complementary sets of A and B , respectively. Also, $\omega(j)$ is an indicator function for the j th confidence interval, i.e., $\omega(j) = 1$ if index j is in set A^c or set B^c , otherwise, $\omega(j) = 0$; $z(\gamma)$ is the

upper $1 - \gamma$ quantile of the standard normal distribution; and $\gamma = (1 - P_E)/2 = 1 - (1 - \alpha)^{1/(I-k-1)}$.

CONCLUSIONS

Evaluation of treatment effect among patient subsets induced by some prognostic factors is of practical importance. In particular, when the prognostic factor is clinical center, the qualitative treatment-by-center interaction is typically of interest. We have reviewed three standard statistical methods designed for detecting the qualitative interaction: the test based on likelihood ratio, the range test, and simultaneous confidence intervals. These methods have received favorable coverage in the literature and have been widely used in practice.

Three sets of test hypotheses are introduced in (1), (8), and (12). For the evidence of no qualitative interaction, the original GS hypothesis (1) defines the situation that all between-treatment differences are of the same sign. Hypothesis (8) focuses on the situation where the between-treatment differences are all above a prespecified clinical insignificance lower margin $-d$ or are all below a prespecified clinical insignificance upper margin d . Hypothesis (12) emphasizes the situation where $I - k$ of the between-treatment differences are all above $-d$ or are all below d , accounting for the total number of patient subsets. Hypotheses (8) and (12) are suitable for a balanced design. For an unbalanced design, a sample size weighting scheme should be used. Besides the test hypotheses (1), (8), and (12), other hypothesis formulations are also arguably more reasonable, and details can be found in Ref. [17].

There are several issues related to the test for the qualitative interaction. In the review, we have focused on two-sided hypotheses concerning the comparison of two treatments among all patient subsets. However, these testing procedures are readily modified to incorporate one-sided hypotheses. In this situation, we test how consistent the advantage of the treatment over the control is among all patient subsets. This is of central interest in noninferiority trials. Secondly, most existing methods concern situations where prespecified patient subsets are disjoint. This is a limitation; however, there are good practical reasons for doing so.^[2] In most applications, one can assume independence among patient subsets to simplify the testing procedure. Nevertheless, it should be noted that the testing procedure could be more complicated when patient subsets are not disjoint. Thirdly, the power for detecting the qualitative interaction is always challenging. This is mainly because, given the general problem setting, only summary information for each patient subset is used. This results in a very limited sample size (i.e., number of subsets), which directly leads to reduction in statistical power for any proposed test.

In addition to the testing procedures discussed, several other methods are noteworthy. Shuster and Van Eys^[18] considered the evaluation of qualitative interaction

between treatment and a continuous covariate, say x , with an effort to divide the values of x into three sets where treatment A is superior, treatment B is superior, and neither is superior. Azzalini and Cox^[19] presented a test for qualitative interaction in the two-way analysis of variance setting under the assumption of no main effects, an assumption that is “unlikely to be true in practice” as indicated by Russek-Cohen and Simon.^[6] In addition, Ciminera et al.^[20] proposed a “pushback” procedure for checking qualitative treatment-by-center interaction. Their procedure first “pushes back” the standardized deviations of D_i ’s from a reference value, e.g., the median, by appropriate amounts and then restores them to the original scale. The appearance of opposite signs among the restored deviations is then taken as substantial evidence of qualitative interaction. Since no joint significance level is justified, their procedure may better serve as a tool for retrospectively detecting qualitative interaction.

REFERENCES

1. Peto, R. Statistical aspects of cancer trials. In *Treatment of Cancer*; Halnan, K.E., Ed.; Chapman and Hall: London, 1982; 868–871.
2. Gail, M.; Simon, R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* **1985**, *41*, 361–372.
3. Piantadosi, S.; Gail, M.H. A comparison of the power of two tests for qualitative interactions. *Stat. Med.* **1993**, *12*, 1239–1248.
4. Pan, G.; Wolfe, D.A. Test for qualitative interaction of clinical significance. *Stat. Med.* **1997**, *16*, 1645–1652.
5. Pan, G.; Wolfe, D.A. Likelihood ratio tests for qualitative interaction when variances are unknown. *ASA Proceedings of the Biopharmaceutical Section*; American Statistical Association: Alexandria, VA, 1996, 151–155.
6. Russek-Cohen, E.; Simon, R.M. Qualitative interactions in multifactor studies. *Biometrics* **1993**, *49*, 467–477.
7. Silvapulle, M.J. Test against qualitative interaction: exact critical values and robust test. *Biometrics* **2001**, *57*, 1157–1165.
8. Shrager, R.I. Quadratic programming for nonlinear regression. *Commun. Assoc. Comput. Mach.* **1972**, *15*, 41–45.
9. Kudo, A. A multivariate analogue of the one-side test. *Biometrika* **1963**, *50*, 403–418.
10. Wolak, F.A. An exact test for multiple inequality and equality constraints in the linear regression model. *J. Am. Stat. Assoc.* **1987**, *82*, 782–793.
11. Berger, R.L. Uniformly more powerful tests for hypothesis concerning linear inequalities and normal means. *J. Am. Stat. Assoc.* **1989**, *84*, 192–199.
12. Zelterman, D. On tests for qualitative interactions. *Stat. Prob. Lett.* **1990**, *10*, 59–63.
13. Zelterman, D.; Kershner, R.P. Tests for qualitative interactions. In *ASA Proceedings of the Biopharmaceutical Section*; American Statistical Association: Alexandria, VA, 1987, 131–133.
14. Box, G.E.P.; Muller, M.E. A note on the generation of random deviates. *Ann. Math. Stat.* **1958**, *29*, 610–611.
15. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157.
16. Yan, X. Test for qualitative interaction in equivalence trials when the number of centers is large. *Stat. Med.* **2004**, *23*, 1711–1722.
17. Chow, S.C.; Shao, J. *Statistics in Drug Research: Methodologies and Recent Developments*; Marcel Dekker Inc: New York, 2002, 263–268.
18. Shuster, J.; Van Eys, J. Interaction between prognostic factors and treatment. *Control. Clin. Trials* **1983**, *4*, 209–214.
19. Azzalini, A.; Cox, D.R. Qualitative interactions in multifactor studies. *J. R. Stat. Soc. Ser. B.* **1984**, *46*, 335–343.
20. Ciminera, J.L.; Heyse, J.F.; Nguyen, H.H.; Tukey, J.W. Test for qualitative treatment-by-center interaction using a ‘pushback’ procedure. *Stat. Med.* **1993**, *12*, 1033–1045.

Therapeutic Equivalence

Jen-pei Liu

Division of Biometry, Department of Agronomy, and Institute of Epidemiology, National Taiwan University, Taipei, Taiwan; and Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan

Abstract

Traditionally, most of the trials conducted by the pharmaceutical industry during phases II and III are aimed at obtaining the necessary evidence for establishing the effectiveness and safety of the new therapeutic entity under development. This entry is focused on the process of equivalence trials and recent developments in hypothesis testing in terms of determining therapeutic equivalence.

Keywords: Active control; Equivalence limit; Non-inferiority; Sample size.

INTRODUCTION

Traditionally, most of the trials conducted by the pharmaceutical industry during phases II and III are aimed at obtaining the necessary evidence for establishing the effectiveness and safety of the new therapeutic entity under development. Consequently, these trials are, in general, the superiority trials in which the primary goal is to demonstrate the better efficacy of the treatment against the concurrent placebo control. However, as innovative drugs become available for the treatment of a number of diseases, attention has focused on searching for new therapeutic modalities to compete with the standard products on the market. These new products may offer some specific advantages over the standard drug, including a better safety profile, improvement on the quality of life, an easy administration route, a short duration of treatment, and, most importantly, reduction of cost. As a result, because of these additional benefits, the new drug is not required to show its superior efficacy over the standard drug but to prove that its effectiveness is no worse than that of the standard. In other words, the sponsors have to provide the evidence that the test product is at least as efficacious as the standard. This concept of demonstration of no worse efficacy but with some specific added values provided by the new treatment is referred to as therapeutic equivalence. As a result, studies with objectives for therapeutic equivalence are then called the non-inferiority trials. Because therapeutic equivalence only requires establishing that efficacy of the new treatment is no worse than that of the standard, it is referred to as one-sided equivalence.

On the other hand, after the patent of an innovative product expires, other pharmaceutical companies can manufacture generic copies under the Drug Price Competition and Patent Term Restoration Act passed in 1984 through the Abbreviated New Drug Application

(ANDA). The generic copies must be demonstrated to be equivalent to the innovative product based on the pharmacokinetic (PK) responses such as the area under plasma concentration-time curve (AUC) or peak concentration ($C_{\max}$) obtained from bioequivalence studies.^[1] According to the current requirement,^[2] a generic copy is claimed to be bioequivalent to the market reference product if the average bioavailability of the generic product is neither too high nor too low as compared to that of the reference product. As a result, bioequivalence emphasizes two-sided equivalence. Bioequivalence includes average bioequivalence, population bioequivalence, and individual bioequivalence that are reviewed and discussed in details in the entry Equivalence Trials.

ONE-SIDED EQUIVALENCE AND NON-INFERIORITY TRIALS

Let μ_T and μ_S be the average efficacy of the test and the standard drugs, respectively, and let a large value in average imply a better efficacy. For a superiority trial, investigators are interested in detecting whether a difference in efficacy exists between the test and the standard drugs. As a result, this question can be answered by the following formulation of hypotheses:

$$H_0 : \mu_T = \mu_S \text{ vs. } H_a : \mu_T \neq \mu_S \quad (1)$$

The alternative hypothesis in Eq. (1) is in the form of a two-sided hypothesis. Others^[3] argue that for a superiority trial, the alternative should be formulated as a one-sided hypothesis:

$$H_a : \mu_T > \mu_S \quad (2)$$

For a detailed discussion and debate of one-sided versus two-sided hypotheses from the perspectives of both the pharmaceutical industry and the regulatory authority, see Koch,^[4] Dubey,^[5] and Chow and Liu.^[6]

If the null hypothesis in Eq. (1) is rejected at the α significance level, then depending upon the alternative in Eqs. (1) or (2), either the existence of a true difference in average efficacy between the test and the standard drugs or a better average efficacy provided by the test drug can be proven at the α significance level. However, when the null hypothesis that $\mu_T = \mu_S$ is not rejected, the conclusion of the trial should never be that the average efficacy of the test drug is the same as that of the standard because the null hypothesis can never be proven in a trial.^[7] In addition, a statistically non-significant, yet clinically meaningful, difference may result from a poorly conducted trial that provides variable data and a large experimental error. On the other hand, well-designed and carefully executed trials generate reliable data to provide estimates of treatment effects with high accuracy and precision. However, less variable data of good quality is able to declare a small difference of no clinical significance to be statistically significant and to conclude that the test and the standard drugs have different average efficacy. The reason for this paradox is that the formulation of hypotheses in Eq. (1) is the wrong hypothesis for evaluation of one-sided equivalence in a non-inferiority trial.

The correct formulation for the hypotheses of interest will be the one-sided hypothesis to test whether the test drug is at least as effective as the standard:^[8–11]

$$H_0 : \mu_T - \mu_S \leq L \text{ vs. } H_a : \mu_T - \mu_S > L \quad (3)$$

where L is the maximum allowable limit of no clinical significance.

The objective of the trial is to prove that the test drug is no worse than the standard treatment—although the average efficacy of the test drug is below that of the standard, but within a range of no clinical significance. It follows that this statement is formulated as the alternative hypothesis in Eq. (3). The equivalence limit L in Eq. (3) is usually a small negative number to define the lower limit for a range of no clinical significance or better efficacy. For the continuous endpoints from a two-group parallel design comparing a test drug with n_T patients to the standard drug with n_S patients, let $\bar{Y}_T$ and $\bar{Y}_S$ be the observed means of the test and the standard drugs, respectively, and let s^2 be the pooled variance. The null hypothesis of Eq. (7) is rejected, and the test drug is then claimed to be at least as efficacious as the standard at the pre-specified α nominal level of significance if

$$T = (\bar{Y}_T - \bar{Y}_S - L) / s \sqrt{1/n_T + 1/n_S} > z(\alpha) \quad (4)$$

where $z(\alpha)$ is the α th upper quantile of the standard normal distribution.

Assessment of equivalence is not only a testing procedure but also an estimation problem. Many have advocated the use of confidence interval for evaluation of both types of equivalence.^[1,7,12–15] For one-sided equivalence, the same conclusion can be reached if the lower $(1 - \alpha)100\%$ confidence limit

$$(\bar{Y}_T - \bar{Y}_S) - z(\alpha)(s\sqrt{1/n_T + 1/n_S}) \quad (5)$$

is larger than the lower equivalence limit L . The confidence interval approach is more appealing than the testing hypothesis procedure. It not only provides the magnitude and range of the average difference between the two treatments but also offers the investigators or readers an opportunity to make their own judgment about whether the test and the standard drugs are equivalent. For a two-group parallel design with equal allocation, the same size per group required to achieve a $(1 - \beta)$ power at α the significance level for non-inferiority trial with respect to hypothesis 3 can be estimated by the following formula:

$$n = 2[\sigma^2/(\delta + L)^2][z(\alpha) + z(\beta)]^2 \quad (6)$$

where $\delta = \mu_T - \mu_S > L$ is the assumed true difference in average efficacy between the test and the standard drugs, σ^2 is the common variance, and $z(\beta)$ is the upper β th quantile of the standard normal distribution.

From Eq. (6), the sample size is a decreasing function of the true unknown difference $\mu_T - \mu_S$. When $\delta = 0$, formula 6 reduces to

$$n = 2[\sigma/L]^2[z(\alpha) + z(\beta)]^2$$

For the sample size for evaluation of two-sided therapeutic equivalence, see Liu;^[16] and for bioequivalence under a standard two-sequence, two-period crossover design, see Liu and Chow.^[17] Blackwelder^[9] has suggested the following procedure for evaluation of one-sided equivalence based on binary endpoints from a two-group parallel trial: The corresponding hypotheses are given below as

$$H_0 : P_T - P_S \leq L \text{ vs. } H_a : P_T - P_S > L \quad (7)$$

where P_T and P_S are the response rates of the test and the standard treatments. Let p_T and p_S be the corresponding observed response rates. The null hypothesis of Eq. 7 is rejected, and the test drug is then claimed to be no worse than the standard at the pre-specified nominal α level of significance if

$$Z = (p_T - p_S - L)/SE > z(\alpha) \quad (8)$$

where $SE^2 = \{[p_T(1 - p_T)/n_T] + [p_S(1 - p_S)/n_S]\}$. The corresponding lower $(1 - \alpha)100\%$ confidence limit and sample-size estimation formula are given respectively as

$$(p_T - p_S) - z(\alpha)SE$$

and

$$n = \left\{ SE^2 / [(P_T - P_S) + L]^2 \right\} \{ z(\alpha) + z(\beta) \}^2$$

Durrleman and Simon^[7] and Jennison and Turnbull^[18] have demonstrated the use of repeated confidence intervals for evaluation of equivalence in a sequential testing manner. Wellek^[19] has proposed a log-rank test for assessment of equivalence between two survival functions.

ACTIVE CONTROL EQUIVALENCE TRIALS

Because non-inferiority trials try to demonstrate that the efficacy of the test drug is no worse than that of the concurrent standard competitor, they are also referred to as active control equivalence trials (ACETs). Issues regarding ACET are thoroughly documented, see, for example, Makuch and Johnson,^[14] Leber,^[20] Senn,^[21] Fleming,^[22] Temple and Ellenburg,^[23] Ellenburg and Temple,^[24] Siegel,^[25] and Fisher et al.^[26] Table 1 provides a set of criteria for evaluating quality of ACETs suggested by Kirshner.^[27] The most critical criterion is the effectiveness of the standard drug. When the test and the standard drugs are therapeutically equivalent, they can be both efficacious or both inefficacious. Leber^[20] provides an excellent illustration of the danger for inference of the results from an active control trial. A series of six trials were conducted to investigate the effectiveness of a new drug for treatment of endogenous depression with an active control imipramine, considered as a standard antidepressant. The results are given in Table 2. If we only look at the comparisons in reduction of the Hamilton depression scale (HAM-D) from baseline between the test drug and the imipramine, both drugs produce a clinically meaningful mean reduction in

HAM-D. However, the null hypothesis of equal reduction in HAM-D is not rejected at the 5% significance level, and the power to detect a 30% difference is quite low. It seems that the test drug and the imipramine have similar efficacy. On the other hand, from Table 2, except for Study V311, the mean reduction in HAM-D for the placebo is also numerically similar to those of the test drug and the imipramine. In other words, except for V311, one cannot reject the null hypothesis that the mean reductions from baseline in HAM-D of both active treatments equal to that of the placebo; hence, both test drugs and imipramine are ineffective as compared to the placebo.

If a concurrent placebo control were not included in these six studies, the test drug would have been claimed effective based on the conjecture that the test drug and the imipramine have similar efficacy. This example illustrates the importance of the inclusion of a concurrent placebo control in an ACET unless the active standard has been proven to be efficacious in adequate well-controlled studies against a concurrent control. As a result, Huque and Dubey,^[28] Pledger and Hall,^[29] ICH,^[10] and Chow and Liu^[6] have recommended that an ACET include three treatments: a test drug, an active standard, and a placebo concurrent control. In addition to evaluation of therapeutic equivalence between the test and the standard drugs, an ACET is required to provide internal validity by verifying effectiveness of both test and standard drugs against the concurrent placebo. As a result, the following hypotheses are formulated to evaluate these two major objectives:

$$\begin{aligned} H_0 : \mu_T - \mu_S \leq L \text{ or } \mu_T - \mu_p \leq 0 \text{ or } \mu_S - \mu_p \leq 0 \\ \text{vs.} \\ H_a : \mu_T - \mu_S > L \text{ and } \mu_T - \mu_p > 0 \text{ and } \mu_S - \mu_p > 0 \end{aligned} \quad (9)$$

where μ_p is the average efficacy of the placebo.

In addition to hypothesis 3 for assessment of one-sided equivalence in efficacy between the test and the standard drugs, there are two more one-sided hypotheses for evaluation of the superiority in effectiveness of the test and the standard drugs as compared to the placebo:

Table 1 Criteria for Evaluation of Active Control Trials

1. Is the standard therapy effective, and does the experimental therapy have an advantage which is generalizable?
2. Can an acceptable minimum effect be defined in terms of the outcome measures?
3. Can this effect be measured precisely?
4. Was the assignment of patients to treatments really randomized?
5. Was the time frame for follow-up sufficient to conclude that the therapies are equivalent?
6. Were factors that which determined participation in the study identical to those for receiving the standard therapy?
7. Is it probable that the groups differ by less than a minimum effect?
8. Were all patients who entered the study accounted for at its conclusion?
9. Were the treatments administered as they would be in practice?

(From Kirshner [27]).

Table 2 Hamilton Depression Scale Endpoint Mean Score

Study	Baseline	Placebo	Test	Imipramine	<i>p</i> value ^a	Power ^b
R301	23.9	14.8 (<i>n</i> = 36)	13.4 (<i>n</i> = 33)	12.8 (<i>n</i> = 33)	0.78	0.40
G305	26.6	13.9 (<i>n</i> = 36)	13.0 (<i>n</i> = 39)	13.4 (<i>n</i> = 30)	0.86	0.45
C311	28.1	18.9 (<i>n</i> = 13)	19.4 (<i>n</i> = 11)	20.3 (<i>n</i> = 11)	0.81	0.18
V311	29.6	23.5 (<i>n</i> = 7)	7.3 ^c (<i>n</i> = 7)	9.5 ^c (<i>n</i> = 8)	0.63	0.09
F313	37.6	22.0 (<i>n</i> = 8)	21.9 (<i>n</i> = 7)	21.9 (<i>n</i> = 8)	1.00	0.26
K317	26.1	10.5 (<i>n</i> = 36)	11.2 (<i>n</i> = 37)	10.8 (<i>n</i> = 32)	0.85	0.33

^aTwo-tailed *p*-values for detection of a difference between the test drug and imipramine.

^bCalculated for a 30% difference between the test drug and the imipramine.

^cStatistically significant difference from placebo at *p* < 0.001, two-tailed. (From Leber [20]).

$$\begin{aligned} H_{0T} : \mu_T - \mu_p \leq 0 \text{ vs. } \mu_T - \mu_p > 0 \\ \text{and} \\ H_{0S} : \mu_S - \mu_p \leq 0 \text{ vs. } \mu_S - \mu_p > 0 \end{aligned} \tag{10}$$

The parameter space of the null hypothesis in Eq. (9) is the union of the spaces of the three null hypotheses in Eqs.(3) and (10). On the other hand, the parameter space of the alternative hypothesis is the intersection of the spaces of the three alternative hypotheses. Hence, if all three one-sided null hypotheses are rejected at the a significant level, then, by intersection-union principle,^[30] the null hypothesis in Eq. (9) is rejected at the α significance level, and it is concluded that the efficacy of both test and reference drugs are superior to the placebo and are equivalent. This procedure may be called the three one-sided test procedure. Although this procedure can control the consumer’s risk under the nominal level of significance, it is undoubtedly very conservative.

EQUIVALENCE LIMIT

For bioequivalence testing and some therapeutic areas, the equivalence limits are usually expressed in terms of the standard response.^[2,28] As a result, equivalence limits are unknown parameters that must be estimated. See the entry “Equivalence Trials” for discussion of this issue. On the other hand, for one-sided therapeutic equivalence, the lower limit *L* may be determined by previous experience about estimated relative efficacy with respect to the placebo and by the maximum allowance below which clinicians consider to be therapeutically acceptable.

With these considerations, it is quite awkward to have an equivalence limit for evaluation of no worse efficacy as large as or even larger than the treatment difference between the standard drug and the placebo. For example, in the treatment of perennial allergic rhinitis, the efficacy is assessed based on an average daily total symptom score over duration of treatment, which is the sum of four individual symptom scores, each with a range from 0–3. It seems reasonable to select an equivalence limit of 3.2 for

evaluation of therapeutic equivalence between a test product and the standard drug because 3.2 represents 25% of the range for the total symptom scores. However, this limit of 3.2 is considered unreasonably liberal as one realizes that improvement in each of the four individual symptom scores, provided by the standard drug over the placebo and estimated from adequate well-controlled studies with sufficient power, is only about 0.5.

Another example is provided in Continuous Infusion versus Double-bolus Administration of Alteplase (COBALT). For fibrinolytic therapy in treatment of suspected acute myocardial infarction, the 30-day mortality rate is around about 12% for the placebo group and is about 8% for tissue plasminogen activator (t-PA, or alteplase^[31]). As a result, the treatment effect of t-PA against placebo is estimated as 4%. To test the hypothesis that double-bolus alteplase is at least as effective as accelerated infusion of alteplase, the COBALT investigators employed an equivalence limit of 0.4%.^[32] This limit is 1/10 of estimated relative treatment effect against placebo. This implies that the upper limit of deaths allowed for the double-bolus alteplase, to be considered therapeutic equivalent to accelerated infusion, is 4 more deaths per 1,000 patients. This number is fewer than 5.6 per 1,000 patients between alteplase and streptokinase.^[32] The estimated difference in the 30-day mortality between double-bolus and accelerated infusion of alteplase is 0.44% with a 95% upper confidence limit of 1.49%. As a result, double-bolus alteplase is not therapeutically equivalent to accelerated infusion of alteplase with respect to an equivalence limit of 0.4%. Therefore, pre-specified equivalence limit for therapeutic equivalence evaluated in a non-inferiority trial should always be selected as a quantity smaller than the difference between the standard drug and the placebo that a superior trial is designed to detect. Even so, the selection of equivalence limit remains the most controversial issue in a non-inferiority trial. The reasons, as illustrated above, are twofold: (i) the ultimate and real goal of any non-inferiority trial is to prove that efficacy of the test treatment is better than that of the placebo although it is no worse than the active control, and (ii) the placebo group is not concurrently included

in non-inferiority trials. Historical information about the relative efficacy of the active control as compared to the placebo group must be utilized in the selection of equivalence limit. For more discussion of the determination of equivalence limit in non-inferiority trials, see Hung,^[33] Snapinn,^[34] Kallen and Larsson,^[35] Tsong et al.,^[36] and Wittes.^[37] For Bayesian approach to incorporating the historical active and the placebo effect in evaluating therapeutic equivalence, see Simon.^[38]

RECENT DEVELOPMENT

In the framework of hypothesis testing and regulatory considerations, Rothmann, et al.^[39] have addressed the design and analysis of non-inferiority trial based on survival for oncology trials. The European Committee for Medical Products for Human Use (CHMP) issued a guideline for selection of non-inferiority margin.^[40] In addition, Chan^[41] has reviewed exact non-inferiority tests based on binary data. Non-inferiority testing has been applied to evaluation of accuracy of diagnostic devices.^[42–46] Liu^[47] has demonstrated the application of non-inferiority to evaluation the safety hazard of drug products. For more discussion on therapeutic equivalence, see Rohmel,^[48] Ebbutt and Frith,^[49] Fleming,^[50] Ng,^[51] and two recently published special issue on non-inferiority trial in *Statistics in Medicine*.^[52,53]

CONCLUSION

For a non-inferiority trial, the objectives of evaluation for equivalence between the test and the standard drugs and for superiority of both drugs over the placebo should be clearly stated in the protocol. In addition, the equivalence limit should also be pre-specified in the protocol with clinical and statistical justification. Because the equivalence limit L is always smaller than the average difference between the active treatments and the placebo, the sample size per group determined by the formulas for equivalence will be large enough to provide sufficient power for verifying effectiveness of both test and standard drugs as formulated in hypotheses 10.

As indicated in ICH,^[10] an adequate active control in a non-inferiority trial should be a widely used standard, whose efficacy with respect to the placebo in the relevant indication has been clearly established and quantified in well-designed and well-documented superiority trial. Furthermore, the standard drug can be expected to exhibit a similar efficacy in the current planned ACET. As a result, the current ACET should have the same critical design features as the previously conducted superiority trials in which the standard has clearly demonstrated clinically relevant efficacy. These critical design features include primary

endpoints, the dose of the standard, inclusion and exclusion criteria, and so on. However, with respect to inference to the target population, there is a distinct difference between superiority trials and ACETs. For superiority trials, the patients randomly assigned to the standard should be comparable with those assigned to the placebo in terms of their risk of developing the outcome. On the other hand, for ACETs, one must assume that patients assigned to the test drugs are nearly identical in terms of how they would respond to the standard drug.^[27]

In general, the intention-to-treat (ITT) analysis will provide an estimate for the treatment effect of a smaller magnitude than the per-protocol (PP) analysis. As a result, for a superiority trial, ITT analysis is conservative in establishing the efficacy. On the other hand, others^[54] claim for a non-inferiority trial; an ITT analysis will not be conservative and is more likely to conclude therapeutic equivalence than a PP analysis. However, patients in an ITT analysis include protocol violators such as failure to satisfy the inclusion/exclusion criteria, poor compliance, loss to follow-up, and missing data. The PP analysis includes the patients who satisfy some criteria such as completion of a minimum pre-specified exposure to the treatment, availability of measurements of the primary endpoints, and absence of any major protocol violation. Therefore, the patients in an ITT analysis are more heterogeneous than those included in a PP analysis. Therefore, the ITT analysis will use a larger variability in evaluation of equivalence than the PP analysis. Hence, the ITT analysis is not necessarily more likely to conclude therapeutic equivalence.

REFERENCES

1. Chow, S.C.; Liu, J.P. *Design and Analysis of Bioavailability and Bioequivalence Studies*. 3rd Ed.; CRC/Chapman & Hall: New York, 2008.
2. The FDA Guidance. *Bioavailability and Bioequivalence Studies for Orally Administered Drug Products—General Considerations*; Food and Drug Administration: Rockville, MD, 2003.
3. Fisher, L.D. The use of one-sided tests in drug trials: An FDA advisory committee member's perspective. *J. Biopharm. Stat.* **1991**, *1*, 151–156.
4. Koch, G.C. One-sided and two-sided tests and p-values. *J. Biopharm. Stat.* **1991**, *1*, 161–170.
5. Dubey, S.D. Some thoughts on the one-sided and two-sided tests. *J. Biopharm. Stat.* **1991**, *1*, 139–150.
6. Chow, S.C.; Liu, J.P. *Design and Analysis of Clinical Trials: Concepts and Methodologies*. 2nd Ed.; John Wiley and Sons: New York, 2004.
7. Durrleman, S.; Simon, R. Planning and monitoring of equivalence studies. *Biometrics* **1990**, *46*, 329–336.
8. Dunnett, C.W.; Gent, M. Significance testing to establish equivalence between treatments with special reference to data in the form of 2x2 table. *Biometrics* **1997**, *33*, 593–602.

9. Blackwelder, W.C. "Prove the null hypothesis" in clinical trials. *Control. Clin. Trials* **1982**, 3, 345–353.
10. ICH International Conference on Harmonisation. In Draft Guidance on Statistical Principles for Clinical Trials, The U.S. Federal Register, 1997.
11. Ware, J.H.; Antman, C.G. Equivalence trials. *N. Engl. J. Med.* **1997**, 337, 1159–1162.
12. Makuch, R.W.; Simon, R. Sample size requirements for evaluating a conservative therapy. *Cancer Treat. Rep.* **1978**, 6, 1037–1040.
13. Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioequivalence. *J. Pharmacokinet. Biopharm.* **1987**, 15, 657–680.
14. Makuch, R.W.; Johnson, M. Issues in planning and interpreting active control equivalence studies. *J. Clin. Epidemiol.* **1989**, 42, 503–511.
15. Makuch, R.W.; Johnson, M. Active Control Equivalence Studies: Planning and Interpretation. In *Statistical Issues in Drug Research and Development*; Peace, K.E., Ed.; Marcel Dekker, Inc: New York, 1990, 238–246.
16. Liu, J.P. Letter to the editor on "sample size for therapeutic equivalence based on confidence interval" by S.C. Lin. *Drug Inf. J.* **1995**, 29, 1063–1064.
17. Liu, J.P.; Chow, S.C. Sample size determination for the two one-sided tests procedure in bioequivalence. *J. Pharmacokinet. Biopharm.* **1992**, 20, 101–104.
18. Jennison, C.; Turnbull, B.W. Sequential equivalence testing and repeated confidence intervals, with applications to normal and binary responses. *Biometrics* **1993**, 49, 31–43.
19. Wellek, S. A log-rank test for equivalence of two survivor functions. *Biometrics* **1993**, 49, 877–881.
20. Leber, P.D. Hazards of inference: The active control interpretation. *Epilepsia* **1989**, 30, S57–S63.
21. Senn, S. Inherent difficulties with active control equivalence studies. *Stat. Med.* **1993**, 12, 2367–2375.
22. Fleming, T.R. Design and interpretation of equivalence trials. *Am. Heart J.* **2000**, 139, S172–S176.
23. Temple, R.; Ellenburg, S.S. Placebo-controlled trials and active-controlled trials in the evaluation of new treatment: Part 1. Ethical and scientific issues. *Ann. Intern. Med.* **2000**, 133, 464–470.
24. Ellenburg, S.S.; Temple, R. Placebo-controlled trials and active-controlled trials in the evaluation of new treatment: Part 2. Practical issue and specific cases. *Ann. Intern. Med.* **2000**, 133, 464–470.
25. Siegel, J.P. Equivalence and noninferiority. *Am. Heart J.* **2000**, 139, S166–S170.
26. Fisher, L.D.; Gent, M.; Buller, H.R. Active-control trials: How would a new agent compare with placebo? A method illustrated with clopidogrel, aspirin, and placebo. *Am. Heart J.* **2001**, 141, 26–32.
27. Kirshner, R. Methodological standards for assessing therapeutic equivalence. *J. Clin. Epidemiol.* **1991**, 44, 839–849.
28. Huque, M.F.; Dubey, S. Design and Analysis of Therapeutic Equivalence Clinical with a Binary Clinical Endpoint. In *The 1990 Proceedings of the Biopharmaceutical Section of the American Statistician Association, Alexandria, VA, USA*; The American Statistical Association: Alexandria VA, 1990, 46–52.
29. Pledger, G.; Hall, D. Active Control Equivalence Studies: Do They Address the Efficacy Issue. In *Statistical Issues in Drug Research and Development*; Peace, K.E., Ed.; Marcel Dekker, Inc: New York, 1990, 226–238.
30. Berger, R.L. Multiparametric hypothesis testing and acceptance sampling. *Technometrics*, **1982**, 24, 295–300.
31. Collins, R.; Peto, R.; Baigent, C.; Sleight, P. Aspirin, heparin, and fibrinolytic therapy in selected acute myocardial infarction. *N.Engl. J. Med.* **1997**, 336, 847–860.
32. The Continuous Infusion versus Double-bolus Administration of Alteplase (COBALT) Investigators. A comparison of continuous infusion of alteplase and double-bolus administration for acute myocardial infarction. *N. Engl. J. Med.* **1997**, 337, 1124–1130.
33. Hung, H.M.J. Noninferiority: A dangerous toy. *ICSA Bull Jan.* **2001**, 27–29.
34. Snapinn, S.M. Alternative for discounting historical data in the analysis of noninferiority trial. *ICSA Bull Jan.* **2001**, 29–33.
35. Kallen, A.; Larson, P. Equivalence studies can not be used to claim equivalence of two active treatments. *ICSA Bull Jan.* **2001**, 33–35.
36. Tsong, Y.; Wang, S.J.; Cui, L.; Hung, J.H.M. Placebo control, historical control, and active control trials. *ICSA Bull Jan.* **2001**, 36–39.
37. Wittes, J. Active-control trials: A linguistic problem. *ICSA Bull Jan.* **2001**, 39–40.
38. Simon, R. Bayesian design and analysis of active control clinical trials. *Biometrics* **1999**, 55, 484–487.
39. Rothmann, M.; Li, N.; Chen, G.; Chi, G.Y.H.; Thou, H.H. Design and analysis of non-inferiority mortality in oncology. *Stat. Med.* **2003**, 22, 239–264.
40. Committee for Medical Products for Human Use (CHMP). Guideline on the choice of the non-inferiority margin; <http://www.emea.eu.int/pdfs/human/ewf/215899en.pdf>.
41. Chan, I.S.F. Proving non-inferiority or equivalence of two treatments with dichotomous endpoints using exact method. *Stat. Meth. Med. Res.* **2003**, 12, 37–58.
42. Hsueh, H.M.; Liu, J.P.; Chen, J.J. Unconditional exact tests for equivalence or non-inferiority for paired binary data. *Biometrics* **2001**, 57 (2), 478–483.
43. Liu, J.P.; Hsueh, H.M.; Hsieh, E.; Chen, J.J. Tests for equivalence or non-inferiority for paired binary data. *Stat. Med.* **2002**, 21, 231–245.
44. Liu, J.P.; Wu, C.Y.; Tai, J.Y.; Ma, M.C. Tests of equivalence and non-inferiority for diagnostic accuracy based on the paired areas under ROC curves. *Stat. Med.* **2006**, 25, 1219–1239.
45. Li, C.R.; Liao, C.T.; Liu, J.P. On the exact interval estimation for the difference in paired areas under the ROC curves. *Stat. Med.* **2008**, 27, 224–242.
46. Liu, C.R.; Liao, C.T.; Liu, J.P. A non-inferiority test for diagnostic accuracy based on the paired partial areas under ROC curves. *Stat. Med.* **2008**, 27, 1762–1778.
47. Liu, J.P. Rethinking statistical approaches to evaluating drug safety. *Yonsei Med. J.* **2007**, 48, 895–900.
48. Rohmel, J. Therapeutic equivalence investigations: Statistical considerations. *Stat. Med.* **1998**, 17, 1703–1714.
49. Ebbutt, A.F.; Frith, L. Practical issue in equivalence trials. *Stat. Med.* **1998**, 17, 1691–1701.

50. Fleming, T.R. Current issues in non-inferiority trials. *Stat. Med.* **2008**, *27*, 317–320.
51. Ng, T.H. Noninferiority hypotheses and choice of noninferiority margin. *Stat. Med.* **2009**, *28*, 1668–1679.
52. D’Agostino, R.B. Noninferiority trials: Advances in concepts and methodology. *Stat. Med. (Spec. Issue)* **2003**, *22* (2), 155–336.
53. D’Agostino, R.B. Noninferiority trials: continued advancements in concepts and methodology (special papers for the 25th anniversary of Statistics in Medicine). *Stat. Med.* **2006**, *25* (2), 181–360.
54. Lewis, J.A.; Machin, D. Intention to treat—Who should use ITT. *Br. J. Cancer* **1993**, *68*, 647–650.

Thorough QT Studies

Joanne Zhang, Dalong (Patrick) Huang, and Yi Tsong

Office of Biostatistics, Office of Translational Sciences, CDER, U.S. Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

After the International Council for Harmonization (ICH) E14 guidance was released in 2005, the thorough QT (TQT) clinical study has been required in early drug development for most compounds. The purpose of the guidance is to provide recommendations to sponsors concerning the design, conduct, analysis, and interpretation of TQT studies. The objective of the study is to exclude a clinically significant increase of drug-induced QTc (QT intervals corrected by the heart rate) intervals. However, the guidance is not clear or specific on a lot of practical details. In this entry, we will address issues including study design, sample size, statistical analysis, assay sensitivity, QT correction methods, how to prespecify time points, and how to collect baseline values for a TQT study. Other than the traditional crossover and parallel studies, a special “nested” crossover design within a parallel study will also be discussed. In addition, using concentration-QTc (C-QTc) modeling to study the relationship between Δ QTc (change from baseline) and $\Delta\Delta$ QTc (placebo adjusted change from baseline) and plasma concentration is also discussed. Suggestions of how to use C-QTc relationship analysis to avoid TQT studies are also provided. The ICH E14 Question and Answer (Q&A) document (2014) emphasized the value of C-QTc modeling, and it enables pharmaceutical companies to use C-QTc modeling as the primary analysis for assessing QTc prolongation risk of new drugs. A new approach by including the time factor into the current C-QTc model is also introduced. Cautions of using C-QTc method in early phase are discussed as well.

Keywords: Thorough QT studies; QT study design; QT correction methods; TQT analysis; Assay sensitivity; Concentration-QTc modeling.

INTRODUCTION

The QT interval is measured from the beginning of Q wave to the end of T wave on an ECG tracing (Fig. 1). It reflects the duration of ventricular depolarization and subsequent repolarization of ventricular myocytes. For some drugs, significant prolongation of the QT interval has been associated with precipitating a potentially fatal cardiac arrhythmia called Torsades de Pointes (TdP), which can degenerate into ventricular fibrillation, leading to sudden cardiac death.^[1,2] Due to the risk of TdP, several drugs have been withdrawn from US market.^[3]

In 2005, the International Council for Harmonization (ICH) issued the E14 guidance, “The Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Nonantiarrhythmic Drugs.” After this guidance was released, regulatory agencies requested that each pharmaceutical company conduct one TQT study when submitting an application for most new nonantiarrhythmic drugs, except for large therapeutic proteins and products that do not enter the bloodstream in any meaningful amount. The purpose of the guidance is to provide recommendations to sponsors concerning the design, conduct, analysis, and interpretation of TQT studies. The guidance suggests that

the TQT study be conducted in early clinical development when some information about the pharmacokinetics of the drug has been obtained.

According to ICH E14 guidance,^[4] a typical TQT study should be a randomized, double blinded, placebo- and active-controlled, crossover, or parallel-arm study conducted with healthy subjects. All subjects are administered either a single dose or multiple doses of the study drug. The objective of a TQT study is to demonstrate the lack of a QTc effect; here QTc is the heart rate corrected QT interval, which is used to remove the effect of heart rate.

If the data collected provide enough evidence to show that the drug does not prolong the QTc interval in a clinically significant manner when compared with placebo, the TQT study is claimed to be a “negative study.” The guidance also recommends using a concurrent positive control arm to establish assay sensitivity. An appropriate positive control will increase the confidence to distinguish a small difference between placebo and the investigational drug.

The ICH E14 analysis is a conservative safety tool and a TQT study is not inexpensive. In 2012, the FDA QT-Interdisciplinary Review Team (QT-IRT) presented the idea of potentially avoiding a TQT study in a public forum.^[5] The presentation reported the coherence result between the

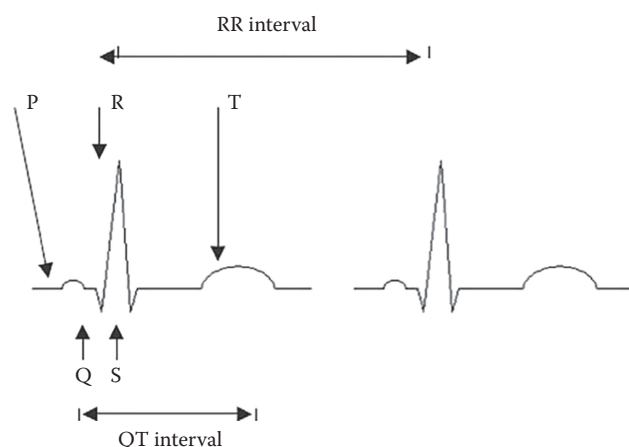


Fig. 1 ECG Waves. The P wave represents depolarization occurring within the atria of the heart. The QRS complex represents depolarization occurring within the ventricles of the heart. The T wave represents repolarization occurring within the ventricles of the heart. The QT wave is the period of time it takes to complete both depolarization and repolarization within the ventricles and is measured from the beginning of the Q wave to the end of the T wave

relationship of QTc to drug concentration (this relationship is measured through the slope) and ICH E14 findings (positive or negative TQT studies) based on hundreds of QT studies submitted to the FDA. The conclusion was that if the slope was not positive, in most cases (86%) the corresponding QT studies were also negative. If the slope was positive, 92% of the corresponding QT studies were also positive. The FDA QT-IRT proposed that by studying the relationship between concentration and QTc, one might use early phase I concentration and QTc data to eliminate the need for TQT studies. This presentation was later published in 2015.^[6] After this presentation, more research results have been published.^[7,8] The ICH E14 Question and Answer (Q & A) document was revised in December 2015. This document allows concentration-QTc (C-QTc) modeling to be the primary method for assessing the QTc interval. However, it is not clear whether a TQT study is still needed. The authors believe that for some compounds, C-QTc modeling might still have its limitations and a TQT type of study might still play an important role.

In this entry, the statistical approaches and issues in the design and analysis of a TQT study as well as C-QTc modeling will be discussed. The terms of negative study and negative compound are interchangeable (meaning that the QT liability of the new drug can be excluded), as with positive study and positive compound (meaning that the QT liability of the new drug cannot be excluded).

QT INTERVAL CORRECTED BY ITS HEART RATE

The ECG records the electrical activities of a human heart. Figure 1 displays an ECG of two heartbeats. The QRS

complex represents depolarization occurring within the ventricles of the heart. The T wave represents repolarization occurring within the ventricles of the heart. The period from the beginning of the Q wave to the end of the T wave defines the QT interval and represents the total time required for both ventricular depolarization and repolarization. The time between two consecutive R waves, the RR interval, is inversely proportional to the heart rate, or QT interval increases with RR interval. In order to eliminate unwanted heart rate effects, the QT interval is generally corrected for heart rate, denoted as the QTc interval.

Commonly used correction formulas include fixed correction methods, population correction methods, and individual correction methods.

A fixed correction method uses a fixed correction formula that is not derived from the current trial data. The most commonly used fixed correction methods are Bazett's^[9] and Fridericia's^[10] formulae (1920). Based on his empirical experience, Bazett assumed the relationship $QT = \alpha \times RR^{1/2}$, which yields $QTc = QT/RR^{1/2}$. Fridericia's formula is $QTc = QT/RR^{1/3}$. This formula is often considered an improvement over Bazett's correction, because it improves the QTc consistency among different subjects.

In addition, Schlamowitz^[11] proposed to use $QTc = QT/RR^{0.205}$ for correction and the Framingham^[12] study (1992) used $QTc = QT/RR^{0.154}$. The fixed correction methods do not account for the variation in the QT versus RR relationship from population to population. This led to the population-based correction method.

A population-based correction method assumes that subjects under the same treatment (on or off drug) share the same but unknown correction factors which need to be estimated through the data. This method usually fits a regression model to the pooled drug-free QT-RR data from all subjects in the current trial.

However, some researchers pointed out the variation in the QT versus RR relationship from subject to subject.^[13–15] They presented evidence that each individual has a different profile of QT against RR. This has led to individual correction methods, an idea first introduced by Dr. Marek Malik.^[13,16] This method often fits a regression model to QT-RR data from each individual's QT and RR measurements at baseline.

Different correction methods have been compared under different models in numerous studies using clinical data.^[1,13,17–20] The goal of a good correction method is to remove the dependency of QTc on RR intervals within subjects. The conclusions of the comparisons were dependent on subject composition and the observed range of RR intervals. Most of the comparisons were focused on the model error which is the intra-subject mean difference. With these constraints, the "error" term would be partially attributed to the data grouping and the outcome of the "error" term may not entirely correspond to the actual biases of the correction methods.

For drugs without a substantial effect on the heart rate, the difference among different correction methods might be small. Based on our experience, Fridericia's correction method works well for a well-conducted QT study. However, if the study drug changes heart rate or has other autonomic effects, the individual correction method with a wide range of RR intervals might work better.^[21] Like any data-driven correction method, the individual correction method will introduce uncertainty for the correction factors. The variability of correction factors is usually ignored and the issue needs further investigation.

There is a concern about the discrepancy of the results by using different correction methods. If this occurs, careful evaluation of the data quality is needed.

ASSESSMENT OF TREATMENT-INDUCED PROLONGATION OF QTc INTERVAL

The ICH E14 guidance^[4] sets the threshold level of regulatory concern for the test drug at a mean effect on QTc of around 5 ms but requires as evidence that a one-sided 95% (or two-sided 90%) upper confidence bound for the largest baseline adjusted mean difference between the study drug and placebo be less than 10 ms. This actually does not make any statistical sense. An upper bound less than 10 ms only indicates that the mean effect is no greater than 10 ms. In practice, instead of testing 5 ms, the mean difference of 10 ms between the study drug and placebo has been tested. The non-inferiority margin of 10 ms was chosen based on clinical experience of the ICH E14 committee. The authors of this entry think the margin may be changed when more experience has been gained.

Let $\mu_T(j)$ and $\mu_P(j)$ denote the mean QTc (or baseline-adjusted QTc) interval of subjects receiving test and placebo treatment, respectively, at time point j , and let $\delta_{TP}(j) = \mu_T(j) - \mu_P(j)$, $j = 1$ to J , where J is the total number of time points. Then, the following statistical hypotheses can be considered for the ICH E14 analysis:

$$H_0 : \text{Max}_{j=1 \text{ to } J} [\delta_{TP}(j)] \geq 10 \text{ ms}$$

versus

$$H_a : \text{Max}_{j=1 \text{ to } J} [\delta_{TP}(j)] < 10 \text{ ms} \quad (1)$$

In practice, it is represented by testing the following null hypotheses by time points:

$$\begin{aligned} H_0(j) : \delta_{TP}(j) &\geq 10 \text{ ms} \text{ versus} \\ H_a(j) : \delta_{TP}(j) &< 10 \text{ ms at all } j = 1 \text{ to } J \end{aligned} \quad (2)$$

The trial is concluded negative when all J null hypotheses in (Eq. 2) are rejected. TQT trials are usually carried out in a single clinical center with healthy subjects. The null hypotheses are rejected by showing that

$$T_j = \frac{\hat{\delta}_{TP}(j) - 10}{e(\hat{\delta}_{TP}(j))} < t(n^*; 0.05)$$

at each time point, where $\hat{\delta}_{TP}(j)$ is the unbiased estimate of $\delta_{TP}(j)$, $e(\hat{\delta}_{TP}(j))$ is the standard error of $\hat{\delta}_{TP}(j)$, and $t(n^*, 0.05)$ is the fifth percentile of the t distribution with n^* degrees of freedom (determined by the sample sizes of the test and placebo groups as well as the design model).

Equivalently, one may reject each of the null hypotheses of (Eq. 2) by showing that one-sided 95% upper confidence interval of $\delta_{TP}(j)$ is less than 10 ms at each time point,

$$U = \hat{\delta}_{TP}(j) + t(n^*; 0.95)e(\hat{\delta}_{TP}(j)) < 10 \text{ ms}, j = 1 \text{ to } J$$

It has been shown that this intersection-union test is very conservative and the type I error rate α is well controlled.^[22] Furthermore, this approach was shown by Berger^[23] to be uniformly most powerful among all monotone, α -level tests for the linear hypotheses (Eq. 2). However, when J is large, this approach usually has low power.

Boos et al.^[24] showed that $\text{Max}_{j=1 \text{ to } J} \hat{\delta}_{TP}(j)$ is a biased estimate of $\text{Max}_{j=1 \text{ to } J} \delta_{TP}(j)$ and $E[\text{Max}_{j=1 \text{ to } J} \hat{\delta}_{TP}(j)] \geq \text{Max}_{j=1 \text{ to } J} \delta_{TP}(j)$. An appropriate bias correction method has not been developed yet.

Cheng et al.^[25] proposed an asymptotic test based on distribution of the maximum of correlated random variable to study drug induced QTc prolongation. At present, there is not much experience with this method.

VALIDATION TEST

For the purpose of clinical trial validation (i.e., assay sensitivity), a positive control arm is usually considered in a TQT study. Such a validation process will enhance confidence in the ability of the study to detect a small change in QTc intervals, especially when the outcome of the trial is negative. ICH E14 guidance^[4] states that “the positive control should have an effect on the mean QT/QTc interval of about 5 ms (i.e., an effect that is close to the QT/QTc effect that represents the threshold of regulatory concern, around 5 ms). This statement can be formed as a statistical hypothesis.^[26]

Let $\mu_C(j)$ and $\mu_P(j)$ denote the mean QTc interval at the j -th time point after the baseline adjustment for the positive control and placebo, respectively. Let $\delta_{CP}(j) = \mu_C(j) - \mu_P(j)$. The hypotheses of interest for assay sensitivity are

$$H_0 : \text{Max}_{j=1 \text{ to } J} \delta_{CP}(j) \leq 5 \text{ ms}$$

versus

$$H_a : \text{Max}_{j=1 \text{ to } J} \delta_{CP}(j) > 5 \text{ ms} \quad (3)$$

Similar to the non-inferiority test of treatment-induced QTc prolongation, the conventional approach is to test the J sets of hypotheses,

$$\begin{aligned} H_0(j) : \delta_{CP}(j) &\leq 5 \text{ ms against} \\ H_a(j) : \delta_{CP}(j) &> 5 \text{ ms for } j = 1 \text{ to } J \end{aligned} \quad (4)$$

the null hypothesis can be rejected as long as the lower limit of at least one confidence interval of $\delta_{CP}(j)$ is greater than 5 ms, i.e.

$$L = \hat{\delta}_{CP}(j) + t(n^*; \alpha^*)e(\hat{\delta}_{CP}(j)) > 5 \text{ ms} \quad (5)$$

for at least one j , $j = 1$ to J with a properly adjusted type I error rate α^* .

As the statistic introduced in Section II, $\text{Max}_{j=1 \text{ to } J} \hat{\delta}_{CP}(j)$ is also a biased estimate of $\text{Max}_{j=1 \text{ to } J} \delta_{CP}(j)$. Unlike testing (Eq. 1) or (Eq. 2), testing (Eq. 3) or (Eq. 4), using $\alpha = 0.05$ for each individual test would inflate family type I error rate because of multiple comparisons. A simple conservative approach would be using the Bonferroni adjustment with $\alpha^* = 0.05/J$ for each individual test in the conventional intersection-union test or estimating each individual confidence interval in (Eq. 5). Power of the validation test can be improved using many Bonferroni-modified (stepwise) procedures. For example, let $p_{(1)} \geq p_{(2)} \geq \dots \geq p_{(J)}$ be the ordered p -values of the J tests, and $H_{0(1)}, H_{0(2)}, \dots, H_{0(J)}$ be the corresponding ordered null hypotheses. Using Holm's procedure (1979),^[27] the study is validated if $p_{(J)}$, the smallest p -value is less than $0.05/J$. Validation fails if $p_{(J)} \geq 0.05$. On the other hand, using Hochberg's procedure (1988),^[28] the study is validated if $p_{(j)} < 0.05/j$ for any $j = 1, \dots, J$. Furthermore, if the covariance structures of the measurements across all the time points are known, the power can further be improved by using Holm's procedure.^[29]

In practice, it is not efficient to perform statistical tests over the entire time course if the peaking time of the positive control arm can be anticipated based on the pharmacokinetic (PK) information. For instance, if we know that the observed maximum QTc effect of the positive control occurs around times $t_1, t_2, \dots, t_K$, then it is sufficient to focus our statistical test just on those K points, where $K < T$.

With a clear signal of QT prolongation and well-behaved PK profile, moxifloxacin, a fluoroquinolone antibiotic, is usually considered for a positive control in a TQT study.^[30] Its maximum QTc effect is around 10 ms or larger and usually occurs between hour 1 and 4 hours after a single dose of 400 mg tablet of administration; therefore, it might be sufficient to evaluate the QTc effect of moxifloxacin only between hours 1 and 4.^[26] Therefore, the multiple comparison adjustment can be performed only on 3 or 4 time points.

Alternatively, in order to test if moxifloxacin has a 5 ms effect, we can look at the average baseline adjusted mean differences of the positive control and placebo around $T_{\max}$.

The parameter of interest is $m_M - \mu_P = \frac{1}{K} \sum_{i=1}^K (\mu_M(i) - \mu_P(i))$,

where $\mu_M(i)$ and $\mu_P(i)$ are the baseline adjusted mean QTc interval at the i -th time point for moxifloxacin and placebo, respectively, and K is usually 3 or 4. If the statistical hypotheses is set up based on this parameter $m_M - \mu_P$, then there is no need to adjust for multiple time points.^[26]

To complete the validation test, usually the QTc profile of the baseline adjusted mean difference between moxifloxacin and placebo over the entire time course should be consistent with the expected QTc moxifloxacin time course profile. A typical moxifloxacin profile is displayed in Fig. 2 with a rising phase, the peaking time, and a declining phase.

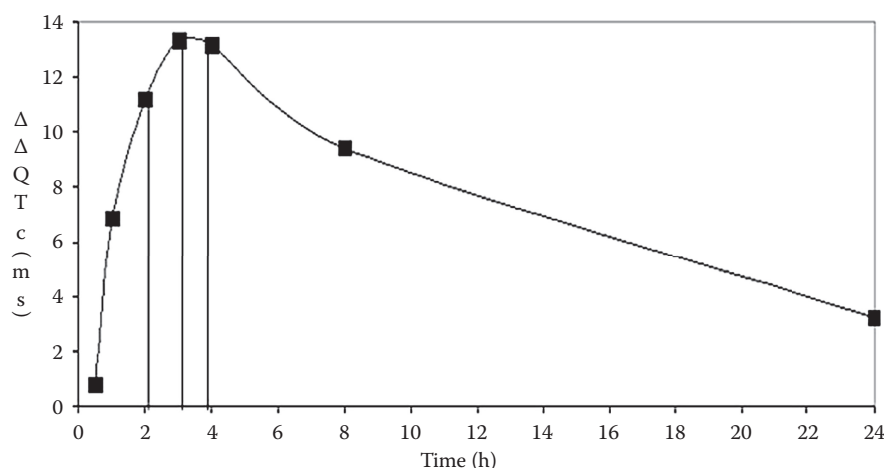


Fig. 2 General time course of Moxifloxacin effect of a double-blind randomized study (Mean baseline adjusted Moxifloxacin – placebo difference and its 90% confidence interval)

SAMPLE SIZE AND DESIGN ISSUES OF TQT STUDIES

Like any clinical trial, the sample size of a TQT study can be calculated based on the hypotheses, statistical models, the desired power for a chosen alternative hypothesis and the variability of the study.^[22,31,32] For different study designs, the sample size required for the study will be different. For a typical crossover TQT study, approximately 45 subjects is a reasonable estimate for any number of treatment arms. For a parallel study though, each arm should have about 45 subjects. Therefore, a parallel study needs a much larger sample size, especially when treatment arms are large.

A TQT study is a randomized, double-blinded, placebo- and positive-controlled, crossover or parallel study with single or multiple doses of the study drug. TQT studies are usually conducted in populations of healthy subjects. However, there are some drugs or the high doses of certain drugs that healthy volunteers cannot tolerate due to safety concerns. In such cases, a TQT study can be conducted in patient populations.

The ICH E14 guidance^[4] suggests that a crossover TQT study is preferable since it needs significantly fewer subjects than a parallel study. In addition, since the same subject acts as her/his own control, the precision of the estimated treatment differences is greater than that for a parallel design with the same number of subjects. One of the major concerns for a crossover study is the potential carryover effect. Hence, it is important to have a sufficient washout period between any two treatments.

For a crossover design, a sequence is a prespecified ordering of the treatments to the periods. The randomization of the clinical trial is achieved by randomly assigning the subjects to one of the sequences for the treatments. For example, the following sequences may be considered for a three-arm trial:

Latin Square 1: Sequence #1: T/P/C;
Sequence #2: C/T/P;
Sequence #3: P/C/T;
Latin Square 2: Sequence #4: T/C/P;
Sequence #5: C/P/T;
Sequence #6: P/T/C;

Here, T, C, and P are the test, positive control, and placebo treatments, respectively.

A sequence-balanced complete crossover trial design for more than two treatments is often represented by a group of Latin squares. The first three sequences and the second three sequences given above form two different Latin squares. For a study with 48 subjects, a complete balanced crossover design is to have eight subjects randomly assigned to each of the six sequences above. The most frequently used crossover design in a TQT trial is the Williams square design.^[33] A Williams square is a specific Latin square with two treatments appearing in two

consecutive periods and in reversed order exactly once. This design is balanced for the first order carryover effects. The Williams square is a variance-balanced design, which means that the variance between any two estimated direct treatment effects is the same. A variance-balanced design has the property that each treatment will be equally precisely compared with every other treatment.^[34] The direct-by-carryover interaction is significant if the carryover effect of a treatment depends on the treatment applied in the immediately subsequent period. If the study model includes period, sequence, treatment, first-order carryover, and direct-by-carryover interaction as fixed effects, the design based on one or two Williams squares might not have sufficient degrees of freedom to assess direct-by-carryover interaction. If this is the case, one might consider using a design with repeated multiple Latin squares, or complete orthogonal sets of Latin squares. For the three treatments, with 48 subjects, the Williams design is also a completely orthogonal set of Latin squares.^[34,35]

For a trial with an even number of treatment arms, usually one Williams square can achieve the goal of balancing all treatment arms. For a TQT study, if one Williams square is implemented, it is better to conduct a double-blinded trial because if the code of which Williams square being used is broken, the whole study may become open-labeled. The FDA does not require double-blinding the positive control arm, moxifloxacin, in TQT studies; however, both the study drug and placebo arms still need to be double-blinded. In a clinical trial, when one arm is open-labeled, it creates more opportunities to break the blinding code for other treatment arms. Fortunately, it is not difficult to overcome this problem. If moxifloxacin is open-labeled, one can apply at least two Williams squares as shown below.

Williams square 1

	Period 1	Period 2	Period 3	Period 4
Sequence 1	B	C	A	MOXI
Sequence 2	C	MOXI	B	A
Sequence 3	A	B	MOXI	C
Sequence 4	MOXI	A	C	B

Williams square 2

	A	B	C	MOXI
Sequence 1	A	B	C	MOXI
Sequence 2	B	MOXI	A	C
Sequence 3	C	A	MOXI	B
Sequence 4	MOXI	C	B	A

Williams square 3

	C	A	B	MOXI
Sequence 1	C	A	B	MOXI
Sequence 2	A	MOXI	C	B
Sequence 3	B	C	MOXI	A
Sequence 4	MOXI	B	A	C

Here A, B, C denote the two doses of the study drug and placebo arm, respectively. MOXI is for moxifloxacin. Even though patients or investigators know which two Williams squares have been used in a TQT study with an open-labeled moxifloxacin administration, they do not know exactly which other treatments they are receiving. For instance, if moxifloxacin was administered in Period 1, subjects may receive either treatment A or treatment C in period 2, even though they know that they will not receive treatment B. If a third Williams square is added, after moxifloxacin is given, there is an equal opportunity for each other treatment being given after moxifloxacin.

The potential danger one needs to consider is that the sample size might be too small to allow for enough degrees of freedom.

A parallel group design may have to be used when the study drug has a long elimination half-life or long-lived active metabolites for which a long time interval would be required to achieve steady-state, or when the study drug has a large number of dose groups to be evaluated. In a situation where the anticipated dropout rate is high, a parallel design is worth consideration as well since the duration of a parallel study is shorter than a crossover study. The biggest disadvantage of a parallel design is the large number of subjects required. To reduce the sample size, researchers proposed a “nested-crossover” alternative design,^[36,37] in which a combined placebo/positive control group can be considered. If moxifloxacin is used for the positive control, a 2×2 crossover between moxifloxacin and placebo can be “nested” within the parallel study. For demonstration purpose, assume that we have three treatment arms. The study drug needs multiple days (m days) of exposure and a parallel comparison with the placebo is determined for the study drug. By using the conventional three-arm parallel design, approximately 135 subjects are needed (about 45 subjects in each treatment arm). Alternatively, one can consider the following design (Table 1). Forty-five subjects can be assigned to Group 1 (the study drug). Twenty-three subjects can be assigned to Group 2 and Group 3, respectively. By considering this design, only 91 subjects are needed.

Even though the sample size is reduced, the price to pay is to collect more ECG data. In addition to Day 0 and Day m , ECGs at Days 1 and $m + 1$ are also required for the statistical analysis.

The ICH E14 guidance^[4] suggests that the concentration-response relationship for QTc prolongation be

characterized. To achieve this, the duration of dosing or a dosing regimen, along with the timing of the collection of ECGs, should be sufficient to capture the PK and PD properties of the drug and its metabolites. Zhang and Stockbridge^[38] described how to prespecify the time points for a TQT study. In a TQT study, a time-matched ECG between each treatment group including the positive control arm and placebo should be collected. It is recommended^[39] the replicate ECGs at each time point be considered. Usually three to five replicates at each pre-specified time point should be obtained within a 2–5 minute period. It is shown that a triplicate ECG at each time point might be sufficient.^[39]

Due to large inter-subject variability, a time-matched baseline adjustment is usually recommended to correct for diurnal variability. The baseline ECG should be collected at least at those time points where post-dose ECG is measured. However, for a crossover design, if the study drug does not change the heart rate, a time-matched baseline might not be necessary. This is because in a crossover study, the diurnal variability of the QTc is accounted for by the design itself since each subject receives each of the treatments in the post-dose periods at exactly the same time points. For a parallel study though, whole day time-matched baseline measurements are needed in order to adjust for within-subject diurnal variability.^[40] When a wide range of RR intervals are needed, in addition to those time points pre-specified for the treatment groups, more time points at baseline might help for a better QT correction.

LESSONS LEARNED FROM HUNDREDS OF THOROUGH QT STUDIES

Since the release of the ICH E14 guidance^[4] and beginning the use of TQT studies as part of the drug approval process, no drugs have been withdrawn from the market due to TdP risk. However, based on many years of experience, it has been realized that the TQT study and the ICH E14 standard are overly conservative. If a TQT study can be replaced by something less complicated, if not for every new drug, but at least for most new applications, that will benefit the community. This was the motivation behind the 2012 comprehensive review of all the existing QT studies submitted to the FDA QT-IRT.^[5] The focus of this post hoc analysis was to examine the relationship between Δ QTc (change from baseline) or $\Delta\Delta$ QTc (placebo adjusted change from baseline) and plasma concentration. If this relationship could not be detected, in most cases, the corresponding QT studies were also negative. Based on the submissions, most of the QT studies are negative. If those negative studies can be identified without a TQT study, it will be more cost effective without compromising safety.

By the end of the year 2012, FDA had reviewed 277 unique QT studies from submissions by pharmaceutical

Table 1 Alternative design for a conventional parallel study

	Day 1 to Day m	Day $m + 1$
Group 1:	study drug	Placebo
Group 2:	Day 1: moxifloxacin	Placebo
	Day 2 to Day m : placebo	
Group 3:	Day 1 to Day m : placebo	Moxifloxacin

sponsors to the FDA, including 229 TQT studies and 48 non-TQT studies conducted among patients without a positive or a negative control arm. Among 277 TQT and non-TQT studies, 53 studies were missing pharmacokinetics (PK) information and one study had an obvious hysteresis between PK and QT. After excluding those 54 studies, 223 QT studies were analyzed, including 141 TQT crossover, 63 TQT parallel, and 19 non-TQT studies.

For each of the 223 selected QT studies, a basic mixed model of ΔQ_{Tc} or $\Delta\Delta Q_{Tc}$ versus concentration or log of concentration was considered:

$$Y_{ij} = \alpha_i + \beta_i X_{ij} + \epsilon_{ij}, \quad i = 1, 2, \dots, n$$

where $Y_{ij} = \Delta Q_{Tc}$ or $\Delta\Delta Q_{Tc}$ for the i th subject at the time point j ; n is the number of subjects; X_{ij} = concentration or log(concentration) for the i th subject at the j th time point; and β_i is the slope for the i th subject, which is a random variable with the expectation $E(\beta_i) = \beta$. Here, β is the population fixed effect, a parameter of interest. The statistical hypotheses considered were as follows:

$$H_0 : \beta \leq 0 \text{ vs. } H_a : \beta > 0 \tag{6}$$

ΔQ_{Tc} was used for parallel or single-arm studies and $\Delta\Delta Q_{Tc}$ was applied for crossover studies. If there was no significant effect or a negative concentration effect on ΔQ_{Tc} or $\Delta\Delta Q_{Tc}$, no relationship between the ΔQ_{Tc} , or $\Delta\Delta Q_{Tc}$ and the plasma concentration would be concluded.

The baseline values for most parallel studies were collected the day prior to the dosing day, which was a time-matched whole-day baseline. For most crossover studies, the baseline values were based on a few pre-dose time points on the day of dosing. After concentration-QTc relationship (C-QTc relationship) analysis, the results were compared with the findings (positive or negative QT studies) based on the ICH E14 analysis. The two approaches were considered to be coherent if (a) H_0 could not be rejected in Equation (V.1) and H_0 in Eqs. 1 or 2 was rejected or (b) H_0 in Eq. 5 was rejected and H_0 in Eqs. 1 or 2 could not be rejected. The results for the sensitivity and specificity of C-QTc relationship analysis were also reported. Sensitivity was defined to be the proportion of negative studies based on the ICH E14 approach that were correctly identified as having no relationship through hypotheses in Eq. 7. Specificity was defined to be the proportion of positive studies based on ICH E14 analysis that were correctly identified as having a relationship by rejecting H_0 in Eq. 7.

The overall coherence rate was about 87%. The overall sensitivity and specificity were about 86% and 92%, respectively. One can find more detailed results in Zhang et al., 2014. Based on the high sensitivity and specificity, Zhang^[6] recommended studying the relationship between ΔQ_{Tc} or $\Delta\Delta Q_{Tc}$ and the drug plasma concentration when proper single ascending dose (SAD) and multiple ascending dose (MAD) studies are available. If the relationship

could not be detected based on hypotheses in Eq. 6, and the two-sided 90% upper confidence interval at a fixed concentration level (50th or 75th percentile, or mean peak plasma concentration C_{max}) was below a certain threshold level (e.g., 10 ms), then a TQT study might be unnecessary. The reason to calculate a 90% confidence interval is to avoid making a wrong decision when there is no relationship, but the ΔQ_{Tc} or $\Delta\Delta Q_{Tc}$ effect is still high due to either a narrow dose range or missing the rising phase of the PK profile. Theoretically, it does not matter at what concentration level ΔQ_{Tc} or $\Delta\Delta Q_{Tc}$ is evaluated at if the slope is flat. There is no statistical test involved in this evaluation. If the relationship can be established and the 90% lower confidence bound at a fixed concentration level (e.g., mean C_{max}) is greater than 10 ms, further investigation is needed. If the signal is real, one might choose intensive safety monitoring during later drug development for a potentially effective compound. However, there are some gray areas in which this analysis alone cannot determine the potential QT liability of the drug, and a TQT type of study is still worth considering.

A NEW APPROACH TO C-QTC MODELING

Huang et al.^[41] discussed some statistical issues of the traditional C-QTc modeling and introduced a modified C-QTc method including the mean circadian rhythm effect into the model. Based on the characteristics of the study setting, they simulated a SAD type of study (the schema of the study is shown in Table 2). A total of nine different doses (ranging from 5 mg to 800 mg) were considered with three cohorts. Each cohort is a 3 × 3 crossover design with two different doses plus placebo in three periods for each sequence. Twelve subjects were equally distributed among three sequences per cohort. The drug concentrations and ECGs were measured at the baseline, 0.5, 1, 2, 3, 4, 8, 12, 24 hr after the first administration of the drug/placebo.

Table 2 Phase 1 crossover single ascending dose study design schema

Cohort	Number of			
	subjects	Period 1	Period 2	Period 3
1	N = 4	5 mg	50 mg	Placebo
	N = 4	5 mg	Placebo	400 mg
	N = 4	Placebo	50 mg	400 mg
2	N = 4	20 mg	200 mg	Placebo
	N = 4	20 mg	Placebo	600 mg
	N = 4	Placebo	200mg	600 mg
3	N = 4	10 mg	100 mg	Placebo
	N = 4	10 mg	Placebo	800 mg
	N = 4	Placebo	100 mg	800 mg

In the simulation study, the QTc circadian rhythm was introduced using the same multioscillator function^[42] for all the subjects. The QTc interval was assumed to be linearly dependent on the drug concentration. The QTc values among different subjects were accomplished through the subject random effect. The QTc intervals within the same subject but at different periods were generated through the subject-by-period random effect. The random sampling error and the two random effects were assumed to be independent of each other. The modified C-QTc model was set up as follows:

$$QTc_{kjs} = QTc_0 + \text{Effect}(t_s) + \beta * C_{kjs} + b_k + b_{k(j)} + e_{kjs} \quad (7)$$

where QTc_{kjs} is the QTc measurement for the subject k in the j th period at time point t_s and QTc_0 is the grand mean QTc value at the baseline and set to be 410 ms. The mean circadian rhythm effect at time t_s , $\text{Effect}(t_s) = QTc_0(-0.03*\cos(2\pi(t_s-6)/24)-0.016*\cos(2\pi t_s/12))$. C_{kjs} is the observed drug concentration for the subject k in the j th period at time t_s . β is the slope which measures the relationship between QTc prolongation and the drug concentration. b_k , $b_{k(j)}$, and e_{kjs} denote the subject random effect, subject by period random effect and the random error, respectively. Suppose $b_k \sim N(0, \sigma_s^2)$ and $b_{k(j)} \sim N(0, \sigma_p^2)$ for the subject k at the j th period. The random errors have autoregressive order one (AR(1)) covariance structure with variance and correlation σ_e^2 and ρ , respectively. The reason AR(1) instead of the commonly used double-compound symmetry structure was chosen is that the correlation between the two QTc measurements may reduce as the time points are apart.^[43,44] With the total variance fixed at 300, eight scenarios of (σ_s^2 , σ_p^2 , σ_e^2) and 10 different correlations were considered (Table 3).

Each period had one predose and eight postdose QTc measurements at 0, 0.5, 1, 2, 3, 4, 8, 10, and 24 hours. The mean QTc value within each period was generated using the above mentioned function $\text{Effect}(t_s)$. The drug effect slope β was set to be $\Delta\Delta QTc/C_{\max}$. True $\Delta\Delta QTc = 10\text{ms}$ was considered.

Further, the following one-compartmental model^[45] was used to simulate the drug concentration,

$$C_{kjs} = \frac{\text{Dose} * K_{ak}}{V_k * (K_{ak} - K_{ek})} * (\exp(-K_{ek} * t_s) - \exp(-K_{ak} * t_s))$$

Table 3 Simulation scenarios and values of covariance parameters

Scenario	σ_s^2	σ_p^2	σ_e^2	ρ
1	225	30	45	0, 0.1, 0.2, ..., 0.8, 0.9
2	195	60	45	
3	180	45	75	
4	180	75	45	
5	165	90	45	
6	120	90	90	
7	90	120	90	
8	210	45	45	

where $K_{ak} = Ka * \exp(\eta_1)$, $Ka = 0.8 \text{ hr}^{-1}$, the absorption rate, $K_{ek} = CL_k/V_k$, $CL_k = CL * \exp(\eta_2)$,

$CL = 12,000 \text{ mL}$, the mean clearance, $V_k = V * \exp(\eta_3)$, $V = 40,000 \text{ mL}$ volume of distribution. Here, η_1 , η_2 , and η_3 are between-subject variability of all the population pharmacokinetics parameters, and they are assumed to be distributed as $N(0, 0.1^2)$. For demonstration purposes, the 600 mg dose was used to evaluate the QTc prolonging effect. The mean $C_{\max}$ of the 600 mg dose from the one-compartmental PK model was $8.327 \mu\text{g/mL}$. Each simulation was replicated 2000 times.

The most widely used C-QTc modeling is the linear mixed model (LMM) with a random slope and/or random intercept using concentration as the covariate. In this simulation, models with and without time were examined. Models without time effect were indicated as base LMM. Models by adding time effect alone and by adding time effect together with other covariates, such as baseline or treatment or both to the base LMM, were referred to as covariate LMM. Both ΔQTc and $\Delta\Delta QTc$ were evaluated. Both compound symmetry (CS) and unstructured (UN) variance-covariance were inspected (Table 4).

Through the simulation, they evaluated the false negative rate (FNR) (percentage of the upper bounds of 90% CIs that were below 10 ms when the true $\Delta\Delta QTc = 10 \text{ ms}$), the coverage probability (CP) (percentage of the 90% CIs that contained the true $\Delta\Delta QTc = 10\text{ms}$), and the relative bias (percentage bias between predicted $\Delta\Delta QTc$ and true $\Delta\Delta QTc$).

Table 4 Concentration-QTc analysis models in the simulation study

Model type	Model formulation	Covariance matrix structure
Base models (Base LMM)	$\Delta QTc_{kjs} = \alpha + \alpha_k + \beta * C_{kjs} + \beta_k * C_{kjs} + e_{kjs}$ $\Delta\Delta QTc_{kjs} = \alpha + \alpha_k + \beta * C_{kjs} + \beta_k * C_{kjs} + e_{kjs}$	CS, UN (α_k and β_k are random intercept and slope respectively)
Covariate Models (Covariate LMM)	$\Delta QTc_{kjs} = \alpha + \alpha_k + \beta * C_{kjs} + \beta_k * C_{kjs} + (\lambda_1 * T_s) + (\lambda_2 * \text{Baseline}_j) + (\lambda_3 * \text{Trt}) + e_{kjs}$ $\Delta\Delta QTc_{kjs} = \alpha + \alpha_k + \beta * C_{kjs} + \beta_k * C_{kjs} + (\lambda_1 * T_s) + (\lambda_2 * \text{Baseline}_j) + (\lambda_3 * \text{Trt}) + e_{kjs}$	CS, UN (α_k and β_k are random intercept and slope respectively)

The results of the simulation study were presented for the scenario ($\sigma_s^2=225$, $\sigma_p^2=30$, $\sigma_e^2=45$, and $\rho=0.3$), which is close to the variances and correlation observed in real TQT studies. The results for other scenarios in the simulations had a similar trend.

The performance of FNR with different models at three different sample sizes ($n=12, 24$, and 36) is displayed in Fig. 3. As can be seen from the picture, the time factor plays an important role in the C-QTc modeling. When the time factor was not included in the C-QTc analysis, the FNR was inflated in most models, except in the models with $\Delta\Delta\text{QTc}$ being the dependent variable. When the time effect was considered, the FNR was below 0.05 in most cases when $n=24$ or 36 .

Covariance structure has also played a significant role in controlling FNR. Based on the simulation results, the overall FNR under CS was lower than that for UN structure. For CS, the FNR can be controlled as long as the time factor was included into the model. When UN was used, the FNR can only be controlled when sample size was relatively large and time effect was either included into the model alone or included together with other variables if ΔQTc being the dependent variable (Fig. 3).

Based on this simulation study, the time effect should be incorporated into the model. Most models without the time effect would underestimate the true mean effect by more than 12%, but models with the time effect included

overestimate the true mean effect by no more than 2% and the small overestimation attenuated as sample size increased (Fig. 4).

Since FNR was not controlled by most base models, CP results were only presented for the covariate LMM models. As can be seen from Fig. 5, models that controlled FNR had CP at least 90%.

CONCLUSIONS AND DISCUSSION

The ICH E14 guidance^[4] provides recommendations to pharmaceutical companies in terms of how to conduct a TQT study. In practice, there is still a spectrum of encountered problems with details of design, conduct, analysis, and interpretation of such studies. In this entry, we have touched topics including correction methods for QT intervals by its heart rate, sample size, study design, assay sensitivity, time point selection, and baseline ECG collections. It is not required to double-blind the positive control arm, moxifloxacin, in TQT studies. There is no published research reported on data quality between blinded and open-labeled moxifloxacin data. If the study variability is smaller for double-blind moxifloxacin study, open-labeling moxifloxacin probably needs to be reconsidered. Before this issue becomes clear, we need to maintain the quality of the overall study design by double-blinding other

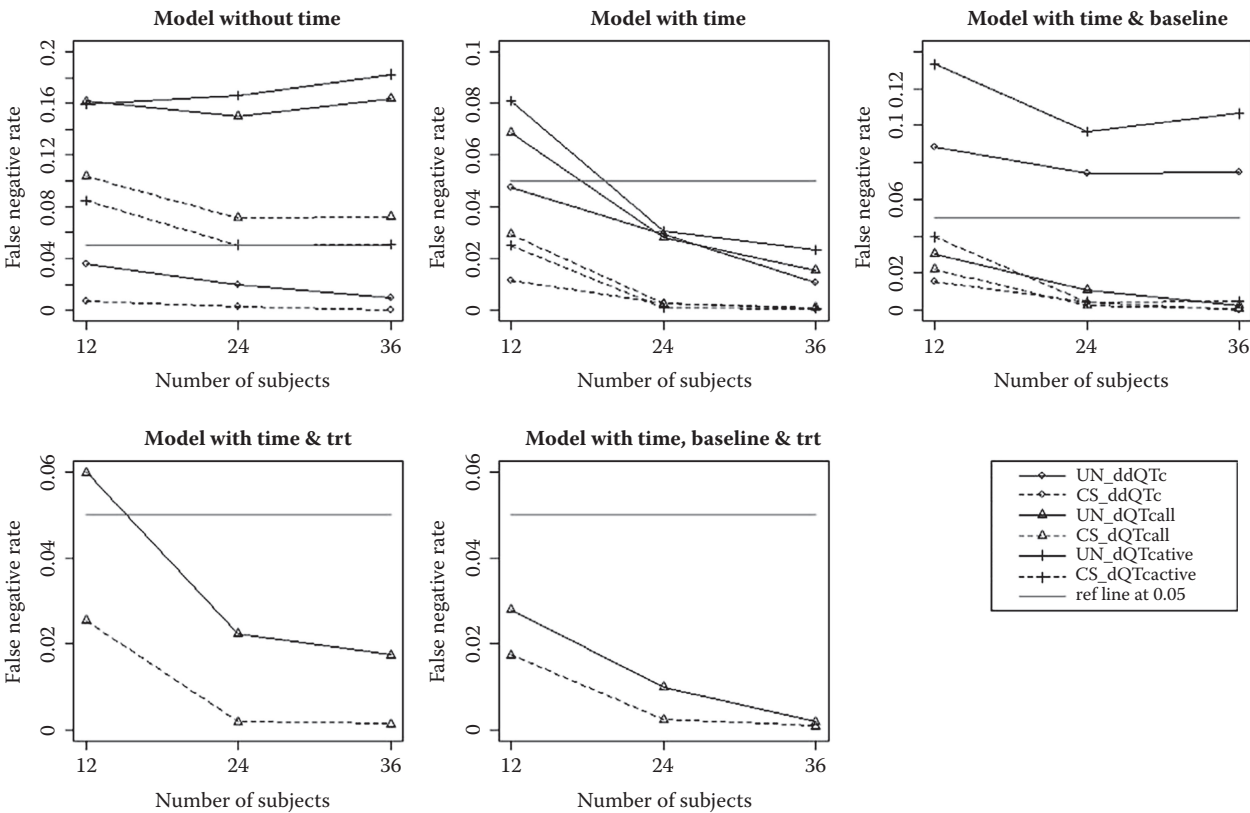


Fig. 3 False negative rate

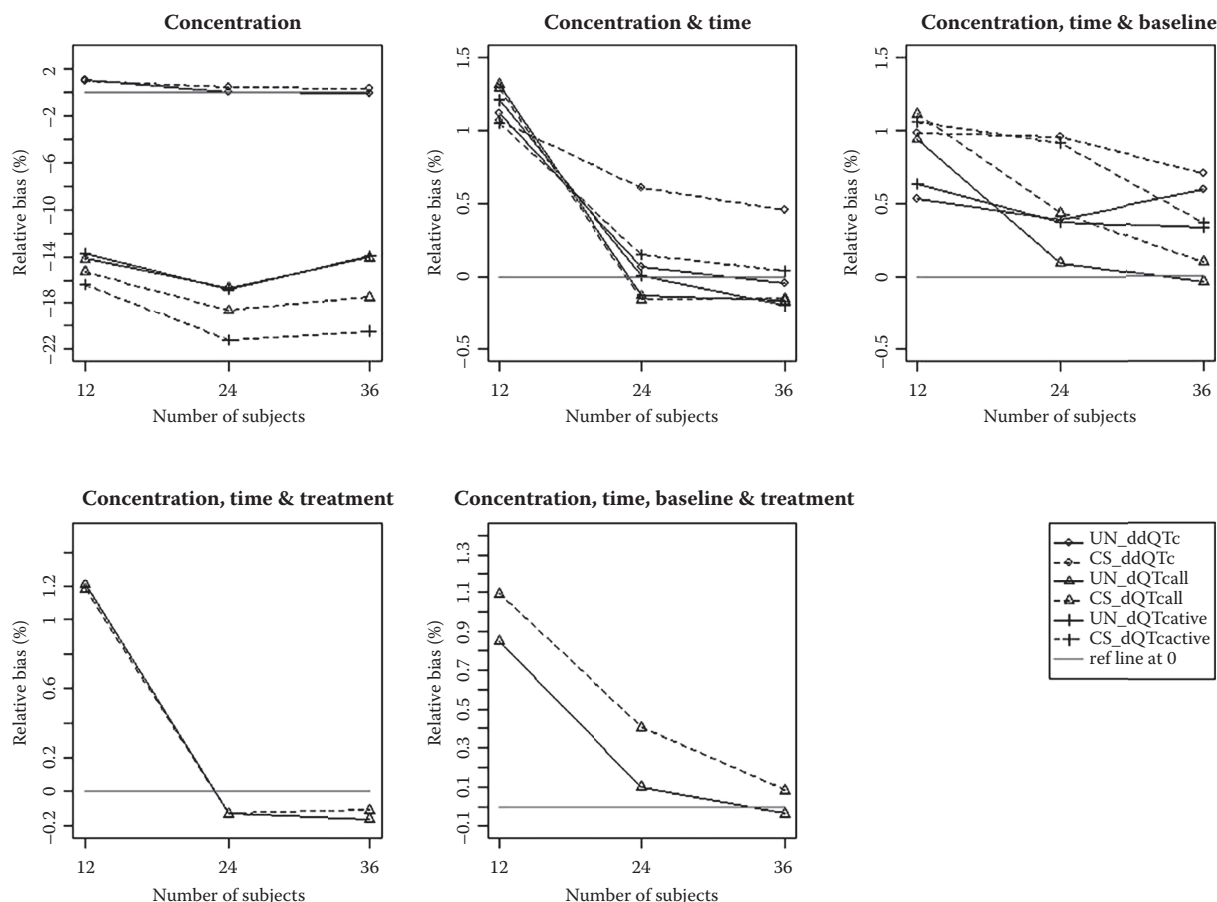


Fig. 4 Relative bias when true $\Delta\Delta\text{QTc} = 10$ ms

treatment arms, which might require implementation of some special designs. For instance, for an even number of treatment arms, instead of using only one Williams square, we might need to use two or three Williams squares. A special “nested” crossover study within a parallel design is also discussed. The “nested” design can help reduce the larger sample size required for a parallel study, but the price to pay is more ECG collection.

It is well understood that the ICH E14 analysis is a conservative tool. If a negative conclusion is made for a compound, the chance of making a mistake for the QT risk is much smaller than 0.05. In recent years, researchers have been actively searching for other alternatives to replace TQT studies.

Garnett et al.^[46] proposed to assess treatment-induced QT prolongation by comparing the predicted one-sided 95% upper confidence interval of $\Delta\Delta\text{QTc}$ at mean $C_{\max}$ with 10 ms, which is done through a linear C-QTc model. Tsong et al.^[47] raised model fitness and other questions. The ICH E14 (Q&A) document^[48] emphasized the value of C-QTc modeling and indicated that C-QTc modeling is “an important component of a totality of evidence assessment of the risk of QT prolongation.” It further suggested that C-QTc modeling can be built into early phase studies as part of the conventional QT study. The ICH E14 (Q&A)

was revised in December 2015, enabling pharmaceutical companies to use a C-QTc modeling as the primary analysis for assessing QTc prolongation risk of new drugs.

The Consortium of Innovation and Quality in Pharmaceutical Development (IQ) and the Cardiac Safety Research Consortium (CSRC) collaborated to evaluate the C-QTc modeling in a small dose escalation clinical study.^[7,8] The joint IQ-CSRC study evaluated five QT-prolonging drugs and one negative drug in 20 healthy subjects. The results showed that C-QTc modeling made correct conclusions for all six drugs. However, as some researchers have pointed out,^[49] the IQ-CSRC study was not designed to assess false negative and false positive rates. The authors of this book entry also believe that IQ-CSRC did not provide enough evidence to show that C-QTc modeling is ready for the primary analysis based on five positive drugs and one negative drug. It makes sense to use the upper bound to exclude a positive compound under certain conditions; however, it is questionable to use the upper bound to judge if a compound is positive. The investigators in the IQ-CSRC study could make an even more “positive” claim by using a much smaller sample size or a badly or poorly designed study with a larger variability. From a safety point of view, it is acceptable to use the upper bound to exclude drugs with QT liability; however, we should also realize the risk of making

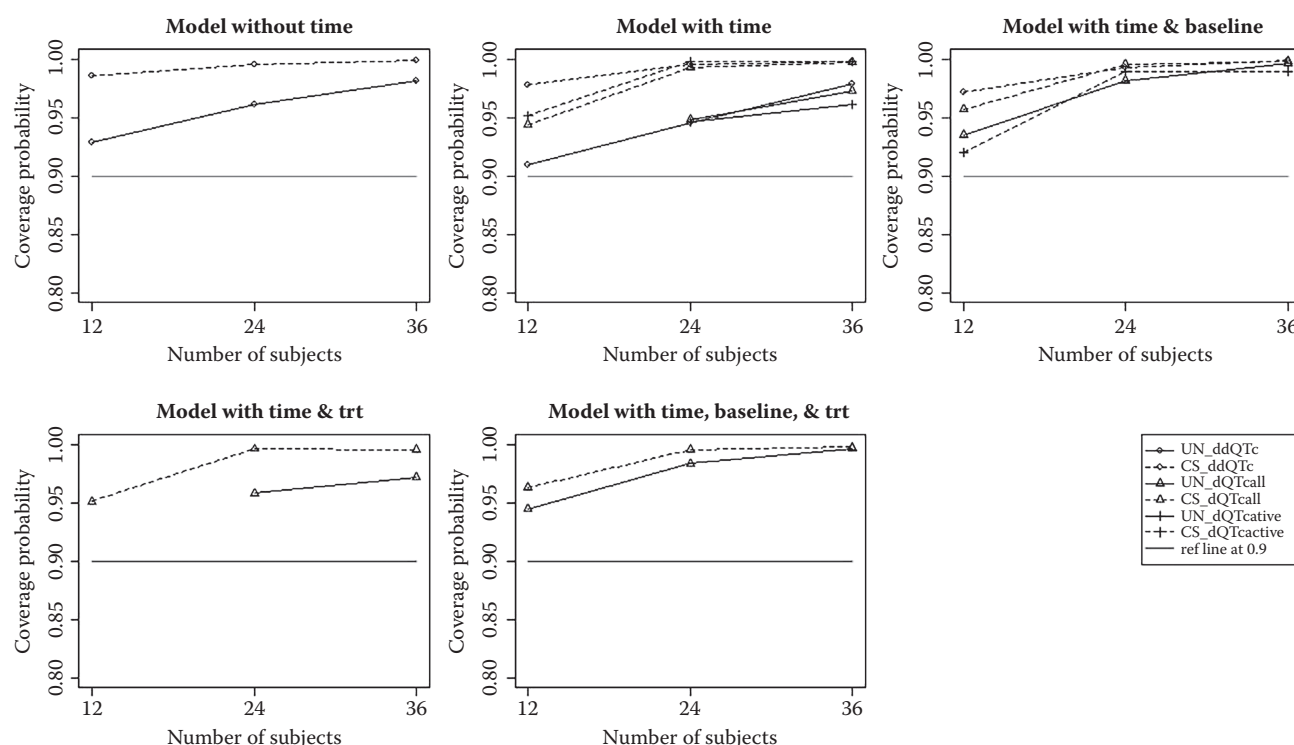


Fig. 5 Coverage probability for models with false negative rate controlled

a false-positive claim by using only the upper bound as the criterion, especially when the sample size is small. This topic is certainly worth further evaluation. However, it is not our intention to conclude that C-QTc modeling cannot be used. The authors strongly support the idea to use the C-QTc method in early drug development to help avoid unnecessary TQT studies for most negative compounds.

It is worthwhile to mention the difference between the C-QTc relationship approach discussed in Section V and the traditional C-QTc modeling that people are familiar with using. For the traditional approach, the conclusion is only based on one realization at mean $C_{\max}$, which by itself has a large variability.^[6] In addition, there are no statistical hypotheses that can be set up properly; therefore, no statistical power or optimal sample size can be obtained. It is not appropriate to compare the “power” between the traditional C-QTc modeling and ICH E14 analysis for a TQT study. The “power” used in the traditional C-QTc modeling does not have the same meaning as defined in a statistical analysis. Even though nothing is wrong with calculating upper or lower bounds at any concentration level, one needs to realize the limitations of the conclusions that may be drawn. Also, the results between the traditional C-QTc approach and the ICH E14 findings do not always agree. Florian et al.^[50] published analysis results using moxifloxacin data based on 20 TQT studies submitted to the FDA. Using C-QTc modeling, the predicted one-sided 95% upper confidence interval for moxifloxacin at $C_{\max}$ was less than 10 ms in two studies, which was not consistent with

the results by ICH E14 analysis. The false negative rate was 10%. The largest mean difference between ICH E14 and C-QTc in one study was as large as 4.4 ms.

In this entry, we present a modified C-QTc model. The performance of this new model was evaluated through a simulation study. The authors believe that the time factor is critical to the modeling and should be included in a C-QTc model in order to control false negative rate. In light of the recent update of the ICH E14 (Q&A), careful consideration needs to be given to a study where C-QTc is used as primary analysis in order to control false negative rates.

There is still room to improve the current C-QTc approach.

REFERENCES

1. Moss, A. Measurement of the QT interval and the risk associated with QTc interval prolongation: a review. *Am. J. Cardiol.* **1993**, 72, 23B–25B.
2. Wysowski, D.; Corken, A.; Gallo-Torres, H.; Talarico, L.; Rodriguez, E. Postmarketing reports of QT prolongation and ventricular arrhythmia in association with cisapride and Food and Drug Administration regulatory actions. *Am. J. Gastroenterol.* **2001**, 96 (6), 1698–1703.
3. Stockbridge, N.; Zhang, J.; Garnett, C.; Malik, M. Practice and challenges of thorough QT studies. *J. Electrocardiol.* **2012**, 45, 582–587.
4. Zhang, J. *FDA IRT Experience with Novel QT Study Designs, Part I: Can a TQT be Avoided? – What Have We*

- Learnt So Far?* DIA Cardiovascular Safety and State-of-art Development Issues, Washington, DC, Apr 17–18, 2012.
5. Zhang, J.; Chen, H.; Tsong, Y.; Stockbridge, N. Lessons learned from hundreds of thorough QT studies. *Ther. Innovation Regul. Sci.* **2015**, *49* (3), 392–397.
 6. Darpo, B.; Garnett, C.; Benson, C.; Keirns, J.; Leishman, D.; Malik, M.; Mehrotra, N.; Prasingle, K.; Riley, S.; Rodriguez, I.; Sager, P.; Sarapa, N.; Wallis, R. The IQ-CSRC prospective clinical phase 1 study: Can early QT assessment using exposure response analysis replace the thorough QT study. *Ann. Noninvasive Electrocardiol.* **2014**, *19* (1), 70–81.
 7. Darpo, B.; Benson, C.; Dota, C.; Ferber, G.; Garnett, C.; Green, C.; Jarugula, V.; Johannesen, L.; Keirns, J.; Krudys, K.; Liu, J.; Ortemann-Renon, C.; Riley, S.; Sarapa, N.; Smith, B.; Stoltz, R.; Zhou, M.; Stockbridge, N. Results from the IQ-CSRC prospective study support replacement of the thorough QT study by QT assessment in the early clinical phase. *Clin. Pharmacol. Ther.* **2015**, *97* (4), 326–335.
 8. Bazett, J.C. An analysis of time relations of electrocardiograms. *Heart* **1920**, *7*, 353–367.
 9. Fridericia, L.S. Die Systolendauer im Elektrokardiogramm bei normalen Menschen und bei Herzkranken. *Acta Medica Scandinavica* **1920**, *53*, 469–486.
 10. Schlamowitz, I. An analysis of the time relationship within the cardiac cycle in electrocardiograms of normal man. I. The duration of the QT interval and its relationship to the cycle length (R-R interval). *Am. Heart J.* **1946**, *31*, 329.
 11. Sagie, A.; Larson, M.G.; Goldberg, R.J.; Bengtson, J.R.; Levy, D. An improved method for adjusting the QT interval for heart rate (the Framingham Heart Study). *Am. J. Cardiol.* **1992**, *70*, 797–801.
 12. Malik, M. Problems of heart rate correction in assessment of drug-induced QT interval prolongation. *J. Cardiovasc. Electrophysiol.* **2001**, *12*, 411–420.
 13. Desai, M.; Li, L.; Desta, Z.; Malik, M.; Flockhart, D. Variability of heart rate correction methods for the QT interval: potential for QTc overestimation with a low dose of haloperidol. *Br. J. Clin. Pharmacol.* **2003**, *55* (6), 511–517.
 14. Batchvarov, V.; Ghuran, A.; Smetana, P.; Hnatkova, K.; Harries, M.; Dilaveris, P.; Camm, A.; Malik, M. QT-RR relationship in healthy subjects exhibits substantial inter-subject variability and high intrasubject stability. *Am. J. Physiol. Heart Circ. Physiol.* **2002**, *282*, H2356–H2363.
 15. Malik, M.; Farbom, P.; Batchvarov, V.; Hnatkova, K.; Camm, A. Relation between QT and RR intervals is highly individual among healthy subjects: implications for heart rate correction of the QT interval. *Heart* **2002**, *87*, 220–228.
 16. Ahnve, S. Correction of the QT interval for heart rate: review of different formulae and the use of Bazett's formula in myocardial infarction. *Am. Heart J.* **1985**, *109*, 568–574.
 17. Funk-Brentano, C.; Jaillon, P. Rate-corrected QT intervals: Techniques and limitations. *Am. J. Cardiol.* **1993**, *72*, 17B–23B.
 18. Hnatkova, K.; Malik, M. "Optimum" formulae for heart rate correction of the QT interval. *Pacing. Clin. Electrophysiol.* **2002**, *22*, 1683–1687.
 19. Malik, M.; Hnatkova, K.; Batchvarov, V. Differences between study-specific and subject-specific heart rate corrections of the QT interval in investigations of drug induced QTc prolongation. *Pace* **2004**, *27*, 791–800.
 20. Garnett, C.E.; Zhu, H.; Malik, M.; Fossa, A.; Zhang, J.; Badilini, F.; Li, J.; Darpo, B.; Sager, P.; Rodriguez, I. Methodologies to characterize the QT/corrected QT interval in the presence of drug-induced heart rate changes or other autonomic effects. *Am. Heart J.* **2012**, *163* (6), 912–993.
 21. Zhang, J.; Machado, S. Statistical issues including design and sample size calculation in thorough QT/QTc studies. *J. Biopharm. Statist.* **2008a**, *18* (3), 451–467.
 22. Berger, R.L. Uniformly most powerful tests for hypotheses concerning linear inequalities and normal means. *J. Am. Statist. Assoc.* **1989**, *84*, 192–199.
 23. Boos, D.; Hoffman, D.; Kringle, R.; Zhang, J. New confidence bounds for QT studies. *Stat. Med.* **2007**, *26*, 3801–3817.
 24. Cheng, B.; Chow, S.C.; Burt, D.; Cosmatos, D. Statistical assessment of QT/QTc prolongation based on maximum of correlated normal random variables. *J. Biopharm. Statist.* **2008**, *18*, 494–501.
 25. Zhang, J. Testing for positive control activity in a thorough QTc study. *J. Biopharm. Statist.* **2008**, *18* (3), 517–528.
 26. Holm, S. A simple sequentially rejective multiple test procedure. *Scan. J. Statist.* **1979**, *6*, 65–70.
 27. Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **1988**, *75*, 800–802.
 28. Tsong, Y.; Yan, L.; Zhong, J.; Nei, L.; Zhang, J. Multiple comparisons of repeatedly measured response: Issues of validation testing of thorough QT trials. *J. Biopharm. Statist.* **2010**, *20* (3), 654–664.
 29. Yan, L.K.; Zhang, J.; Ng, M.J.; Dang, Q. Statistical characteristics of moxifloxacin-induced QTc effect. *J. Biopharm. Stat.* **2010**, *20* (3), 497–507.
 30. Zhang, J. Optimal sample size allocation in a thorough QTc study. *Drug Inform. J.* **2011**, *45*, 455–468.
 31. Tsong, Y.; Sun, A.; Kang, S.-H. Sample size of thorough QTc clinical trial adjusted for multiple comparisons. *J. Biopharm. Stat.* **2013**, *23* (1), 57–72.
 32. Williams, E. Experimental designs balanced for the estimation of residual effects of treatments. *Aust. J. Sci. Res.* **1949**, *2*, 149–168.
 33. Chen, L.; Tsong, Y. Design and analysis for drug abuse potential studies: Issues and strategies for implementing a crossover design. *Drug Inform. J.* **2007**, *41*, 481–489.
 34. Jones, B.; Kenward, M.G. *Design and Analysis of Crossover Trials*. 2nd Ed.; Chapman and Hall: London, 2003.
 35. Zhang, J. *Moxifloxacin and Placebo Can Be Given in a Crossover Fashion During a Parallel-Designed Thorough QT Study*. DIA Cardiovascular safety QT, and arrhythmia in drug development conference, Bethesda, MD, 2009.
 36. Tsong, Y. On the designs of thorough QT/QTc clinical trials. *J. Biopharm. Stat* **2013**, *23* (1), 43–56.
 37. Zhang, J.; Stockbridge, N. Selection of the time points for a thorough QTc Study. *Drug Inform. J.* **2011**, *45*, 713–715.
 38. Patterson, S.; Agin, M.; Anziano, R.; Chuang-Stein, C.; Dmitrienko, A.; Ferber, G.; Francom, S.; Gerald, M.; Ghosh, K.; Mills, T.; Menton, R.; Natarajan, J.; Offen, W.; Saoud, J.; Smith, B.; Suresh, R.; Zariffa, N. Investigating drug induced QT and QTc prolongation in the clinic: Statistical design and analysis considerations. *Drug Inform. J.* **2005**, *39*, 243–266.
 39. Zhang, J.; Dang, Q.; Malik, M. Baseline correction in thorough parallel QT studies. *Drug Safety* **2013**, *36*, 441–453.

40. Huang, D.; Xiao, S.; Dang, Q.; Tsong, Y. *Some Statistical Issues in Concentration-QTc Modeling*, ASA Biopharmaceutical Section Regulatory-Industry Statistics Workshop, Washington, DC, Sept. 29, 2016.
41. Piotrovsky, V. Pharmacokinetic-pharmacodynamic modeling in the data analysis and interpretation of drug-induced QT/QTc prolongation. *AAPS J.* **2005**, *7* (3), E609–E624.
42. Stylianou, A.; Roger, J.; Stephens, K. A statistical assessment of QT data following placebo and moxifloxacin dosing in thorough QT studies. *J. Biopharmaceut. Stat.* **2008**, *18* (3), 502–516.
43. Meng, Z. Simple direct QTc-exposure modeling in thorough QT studies. In *FDA/Industry Workshop*; Rockville, MD, 2008.
44. Bonate, P. Effect of assay measurement error on parameter estimation in concentration–QTc interval modeling. *Pharmaceut. Stat.* **2013**, *12*, 156–164.
45. Garnett, C.; Beasley, N.; Bhattaram, V.; Jadhav, P.; Madabushi, R.; Stockbridge, N.; Tornøe, C.; Wang, Y.; Zhu, H.; Gobburu, J. Concentration-QT relationships play a key role in the evaluation of proarrhythmic risk during regulatory review. *J. Clin. Pharmacol.* **2008**, *48*, 13–18.
46. Tsong, Y.; Shen, M.Y.; Zhong, J.; Zhang, J. Statistical issues of QT prolongation assessment based on linear concentration modeling. *J. Biopharm. Stat.* **2008**, *18* (3), 564–584.
47. ICH E14 Implementation Working Group. *ICH E14 Guideline: The clinical evaluation of QT/QTc interval prolongation and proarrhythmic potential for non-antiarrhythmic drugs questions & answers (R2) international conference on harmonisation*, Geneva, March 2014.
48. Mendzelevski, B. *A Proposal to review the TQT study*, CSRC/FDA Workshop: The Proarrhythmic Assessment of New Chemical Entities, Washington, DC, April 2016.
49. Florian, J.; Tornøe, C.; Brundage, R.; Parekh, A.; Garnett, C. Population pharmacokinetic and concentration-QTc models for moxifloxacin: pooled analysis of 20 thorough QT studies. *J. Clin. Pharmacol.* **2011**, *51*, 1152–1162.
50. International Conference on Harmonisation. *ICH (E14) Guidance: The Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs*, International Conference on Harmonisation, Geneva, Switzerland, May 2005.

Tiered Approach of Analytical Similarity Assessment: Statistical Methods

Yi Tsong

Office of Biostatistics, Office of Translational Science, CDER, FDA, Silver Spring, Maryland, U.S.A.

Abstract

For approval of a biosimilar drug product, the U.S. Food and Drug Administration recommends a stepwise approach for obtaining the totality of the evidence between the proposed biosimilar product and its corresponding innovative biological product be considered. The stepwise approach starts with the assessment of similarity in identified critical quality attributes, which are classified into three tiers depending upon their corresponding criticality or risk ranking relevant to clinical outcomes. Appropriate statistical tests (e.g., equivalence test, quality range approach, and raw data graphical presentation) are then applied to these critical quality attributes in different tiers (e.g., Tier 1, Tier 2, and Tier 3, respectively). This article intends to provide an overview of these statistical tests for analytical similarity assessment in biosimilar studies.

Keywords: Equivalence test; Raw data graphical presentation; Tiered approach; Totality of the evidence; Quality range approach.

BACKGROUND

As defined in 351(i) of the Public Health Service (PHS) Act, biosimilarity is “that the biological product is highly similar to the reference product notwithstanding minor differences in clinically inactive components” and that “there are no clinically meaningful differences between the biological product and the reference product in terms of the safety, purity and potency of the product.” For the assessment of biosimilarity for biosimilar products, the U.S. Food and Drug Administration (FDA) published a few guidances. As indicated in its recent guidance, the FDA recommends a stepwise approach for obtaining the totality of the evidence for demonstrating biosimilarity between a proposed biosimilar product and an innovative (reference) biological product.^[1–3] The stepwise approach starts with analytical studies for functional and structural characterization of critical quality attributes (CQAs) that are relevant to clinical outcomes at various stages of the manufacturing process followed by animal studies for the assessment of toxicity, clinical pharmacology pharmacokinetic (PK) or pharmacodynamic (PD) studies, and clinical studies for the assessment of immunogenicity, safety/tolerability, and efficacy (see Fig. 1).

Analytical similarity serves as the foundation of biosimilarity assessment. CQAs in the manufacturing process of drug products are referred to as chemical, physical, biological, and microbiological attributes that can be defined, measured, and continually monitored to ensure the finished products remain within acceptable quality limits. For the analytical studies, CQAs should be identified and classified into three tiers according to their criticality or risk ranking based on the mechanism of action or PK. A few CQAs that are most relevant to clinical outcomes will be classified as Tier 1, those that are least relevant to clinical outcomes or non-numerical ones will be classified as Tier 2, and the rest will be classified as Tier 3, respectively.^[3,4] Accordingly, we proposed different statistical approaches for the different tiers of the CQAs.

STATISTICAL APPROACHES USED FOR ASSESSING SIMILARITY

In analytical similarity assessments, an extensive structural and functional characterization of both the proposed biosimilar product and the reference product is conducted. Considering practicality, a three-tiered statistical approach for utilizing various statistical methods was proposed to evaluate analytical similarity data. The quality attributes (QAs) should be ranked depending on the risk of the attributes based on their potential impacts on clinical activity, PK/PD, safety, and other factors. In general,

Note: This article reflects the views of the author and should not be construed to represent the FDA's views or policies.

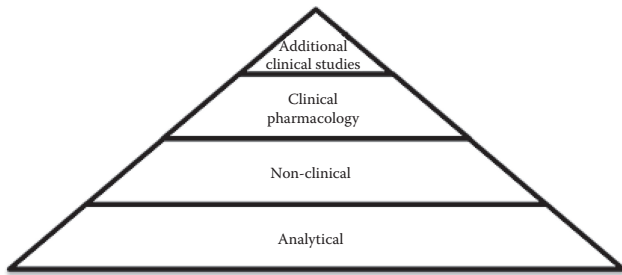


Fig. 1 A stepwise approach for demonstration of biosimilarity

statistical equivalence test is proposed to apply to some of the selected highest ranked QAs; quality range statistical approach is proposed to assess the analytical similarity of other selected high-ranked quantitative QAs; side-by-side comparison approaches are proposed for the rest QAs. In the following subsections, we will present more details of the statistical approaches, especially for Tier 1 statistical approach. Overall, the equivalence test approach proposed for Tier 1 CQAs is more rigorous than the quality range approach proposed for Tier 2 CQAs, which is statistically more rigorous than the side-by-side graphic comparison approach for Tier 3 CQAs, in terms of the stringency of the approaches for its study objective. It should not be interpreted as if a CQA that satisfied the equivalence test proposed for Tier 1 CQA would also satisfy the criteria of the quality range approach proposed for Tier 2 CQAs.

STATISTICAL EQUIVALENCE TEST FOR TIER 1 CQAS

Similar to the bioequivalence testing for generic drugs,^[5] for Tier 1 CQAs, equivalence in means between the proposed biosimilar and reference products can be measured either by the ratio of means or by the difference of means. Here, for simplicity to illustrate, let us focus only on the normally distributed data and difference of product means. The two one-sided hypotheses can be written as follows:

$$\begin{aligned} H_{01} : \mu_T - \mu_R \leq -\delta & \quad \text{vs.} \quad H_{a1} : \mu_T - \mu_R > -\delta \\ H_{02} : \mu_T - \mu_R \geq \delta & \quad \text{vs.} \quad H_{a2} : \mu_T - \mu_R < \delta \end{aligned} \quad (1)$$

where $\delta > 0$ is the equivalence margin. More details of the choice of δ values will be presented in the later section. Based on the properties of two one-sided tests (TOSTs), the familywise type I error rate is controlled at $\alpha = 0.05$ if each of the two one-sided hypotheses H_{01} and H_{02} is tested at $\alpha = 0.05$ separately.

With a proper design, for each CQA to be analyzed, there are J replicates of the testing results from each of the n_T biosimilar lots and from each of the n_R reference lots. Let σ_{Ti}^2 and σ_{Ri}^2 be the variances of the i th biosimilar lot and i' -th reference lot, respectively. Let X_{Ti} $X_{Ri'}$ be the

lot sample mean value of replicates within each lot for test and reference, respectively. For the following discussions, X_{Ti} and $X_{Ri'}$ are assumed to follow $N(\mu_T, \sigma_T^2)$ and $N(\mu_R, \sigma_R^2)$ distributions, respectively.

Note that when there are J replicates in each of reference and biosimilar lots, X_{Ti} and $X_{Ri'}$ represent the sample means of J replicates and they have the variance $\sigma_k^{*2} = \frac{1}{J} \sigma_{kiJ}^2$ where $k = T$, and R and σ_k^{*2} represents the true value between lot variances. For simplicity, we assume that $\sigma_k^2 = \sigma_k^{*2} + \frac{1}{J} \sigma_{kiJ}^2$ with equal σ_{kiJ}^2 across all lots.

For testing against H_{01} and H_{02} , the following test statistics are used:

$$T_1 = \frac{(\bar{X}_T - \bar{X}_R) + \delta}{\sqrt{\frac{S_T^2}{n_T} + \frac{S_R^2}{n_R}}}, T_2 = \frac{(\bar{X}_T - \bar{X}_R) - \delta}{\sqrt{\frac{S_T^2}{n_T} + \frac{S_R^2}{n_R}}}, \quad (2)$$

where $\bar{X}_l = \sum_{i=1}^{n_l} X_{li} / n_l$, $S_l^2 = \sum_{j=1}^{n_l} (X_{lj} - \bar{X}_l)^2 / (n_l - 1)$; X_{li} is the lot attribute value, estimated by the average of the replicates within the j th lot for product l , $l = T, R$; and n_T and n_R are the number of lots from the proposed biosimilar and reference products, respectively. Please note that the expectation of S_l^2 is a linear combination of both within- and between-lot variances, and it decreases toward σ_i^2 when the number of replicates increases within each lot.

Under the null hypotheses, when the margin is a known constant, T_1 and T_2 follow an approximate t -distribution with v degrees of freedom, where v is approximated by Satterthwaite^[5] as follows:

$$v = \left(S_T^2 / n_T + S_R^2 / n_R \right)^2 / \left(\left(S_T^2 / n_T \right)^2 / (n_T - 1) + \left(S_R^2 / n_R \right)^2 / (n_R - 1) \right)$$

On the boundary of H_{01} , $T_1 \sim t_v$, where t_v is the t -distribution with a degree of freedom v . On the boundary of H_{02} , T_2 follows the same distribution as T_1 . Thus, H_{01} and H_{02} of Eq. 1 are rejected with the conclusion of statistical equivalence in means if $T_1 > t_{v,0.95}$ and $T_2 < t_{v,0.05}$.

The above hypothesis testing approach is consistent with the decision rule of using the $(1 - 2\alpha)$ 100% two-sided confidence interval of the mean difference under normality, and σ_R in the equivalence margin is assumed to be known or determined using exterior information. For normally distributed attributes, the two one-sided hypotheses (Eq. 1) are used to test for the following equivalence hypothesis:

$$\begin{aligned} H_0 : \mu_T - \mu_R \leq -\delta \text{ or } \mu_T - \mu_R \geq \delta \\ H_a : -\delta < \mu_T - \mu_R < \delta \end{aligned} \quad (3)$$

Analytical similarity is supported if the null hypothesis H_0 of Eq. 3) is rejected. In other words, statistical equivalence in means is established if we can show that

$-\delta < \mu_T - \mu_R < \delta$ by using the test statistics in Eq. 2 or by showing that $(\bar{X}_T - \bar{X}_R) \pm t_{1-\alpha, \nu} \sqrt{S_T^2/n_T + S_R^2/n_R}$ is completely contained within the interval $(-\delta, \delta)$.^[6]

Note that when the attribute value Y follows a lognormal distributions, we may use $X = \log(Y)$ which has a normal distribution for the equivalence test.

QUALITY RANGE APPROACH FOR TIER 2 CQAS

For the selected QAs (CQAs) of Tier 2 category, the similarity assessment is focused on the data spray of biosimilar product against the reference product. For this reason, instead of evaluating the equivalence of means of the lot attribute values, the objective is to show that the biosimilar data overlap with the reference product well within a reasonably defined range of its qualities. The basic concept of this statistical approach is that it requires a large proportion, say 90%, of the test results from the biosimilar lots to be within an acceptance range determined by the reference product. Assume the test results of the reference product follow the normal distribution as defined earlier, e.g., $X_R \sim N(\mu_R, \sigma_R^2)$, then a certain proportion of the observations from the reference product can be found within the interval of $\mu_R \pm k \times \sigma_R$. Let us name this interval of $\mu_R \pm k \times \sigma_R$ as the quality range. For example, there is about 68% coverage of all lot values of the reference product within the interval of $\mu_R \pm \sigma_R$; there is about 95% coverage of the reference distribution within $\mu_R \pm 2\sigma_R$; there is about 99.7% coverage within $\mu_R \pm 3\sigma_R$; and there is about 99.99% coverage of the reference distribution within $\mu_R \pm 4\sigma_R$. The problems of precise estimation of the range with small sample size are well known. The mean $\pm k$ SD approach uses the point estimate of the target interval with the intended coverage. This approach has its own statistical issues and limitations, especially when the data are clearly not normally distributed. However, as it is often used in specification determination in the early development stage and is also used in an appropriate quality control chart, chemists and biologists can easily understand the mean $\pm k$ SD approach better than any other alternative range approach. The determination of using an alternative multiplier k would mainly depend on scientific judgment.

GRAPHIC AND RAW DATA COMPARISON FOR TIER 3 CQAS

Graphic and raw data comparison approaches are often used for Tier 3 CQAs. The approaches are often chosen for information presentations in a multivariate nature. Due to the lack of rigidness of the approaches in statistical properties, the potential of being subjective and biased is usually large. These approaches are chosen because they have less impact on CQAs with less impact on clinical outcomes in

the sense that a notable dissimilarity will not affect clinical outcomes. The results of these approaches are more suitable to characterize the similarity or dissimilarity between biosimilar and reference products that is not attainable using the parameters to represent the information.

DETERMINATION OF EQUIVALENCE MARGIN

As in any equivalence test, determination of a proper equivalence margin is crucial and yet challenging. Preferably, a biologically or clinically meaningful equivalence margin may be derived from scientific knowledge or past experience. However, such a predefined margin is not available for biosimilar development programs most of the time. Furthermore, a one-size-fits-all margin may not be suitable to serve as an acceptance criteria for QAs due to the following reasons: (1) a biosimilar application may have multiple selected CQAs with difference variability to be analyzed using the equivalence test. Using a fixed margin, it can lead to a requirement of different sample sizes (number of lots) for each QA. Sometimes, the required sample size may be much larger than what is practically possible. (2) Different biosimilar development programs with a common reference product may have different QAs selected for using the equivalence test or may apply different analytical methods for the same QAs. Thus, a unified but not fixed margin is needed. (3) The sample size (number of lots) available in the analytical similarity studies is usually very limited and pre-given. Thus, analytical similarity studies with small sample sizes would have a low power to pass the equivalence test using a one-size-fits-all margin unless the margin is overly wide. Therefore, the development of a statistical equivalence margin is helpful when a biologically or clinically meaningful margin is not available.

Due to the properties of the biologics and the unique challenges in analytical similarity assessment, the statistical equivalence margin was derived with the following considerations: first, a unified representation of the margin is needed for all selected CQAs with a different variability, and second, a margin needs to be determined by achieving a sufficient power of passing the equivalence test with the practically limited sample sizes. With these two considerations in mind, it led to an equivalence margin as a function of the reference product variability.

Under the equal sample size and equal variance assumption, Limentani et al.^[7] recommends a margin determination for analytical chemistry. That is,

$$\delta = \delta_0 + s^* \left[t_{(1-\alpha, 2n-2)} + t_{(1-\beta/2, 2n-2)} \right] \sqrt{2/n}, \quad (4)$$

where δ_0 is a prespecified tolerable shift, s^* is the upper 95% confidence limit of standard deviation of the reference product, and $t_{(a,b)}$ is the 100 a th percentile of t -distribution

with b degrees of freedom. Limentani et al.^[7] recommended using $1 - \beta/2 = 0.99$ for any sample size at the alternative hypothesis $\mu_T - \mu_R = \delta_0$. Therefore, the Limentani approach defines the margin as a function of s^* to have 99% power to pass equivalence test for any sample size if the true mean difference is no larger than a prespecified difference of δ_0 . In general, this margin is considered to be too wide, especially for small sample sizes. The Limentani approach does not reward large sample sizes, thus providing little motivation to increase the sample size.

To overcome the caveat of the Limentani approach, an equivalence margin δ is considered as a function of the variability of the reference product with a form of $\delta = f \times \sigma_R$. The multiplier f is adjusted by the sample size and the power. In our approach for determination of f , the first step is to preassign the power as a monotonic increasing function of the sample size under a given true mean difference, the target “similarity.” The second step is to obtain the margin to achieve the preassigned power at the first step with the given sample size and true mean difference. In practice, σ_R is unavailable to both the biosimilar sponsor and the regulatory agency. Conventionally, the sample standard deviation is used as the estimate of the variability in margin determination. The related statistical issues will be discussed in later sections.

As mentioned earlier, for most of biosimilar development programs, the sample size is usually limited. Ten lots per product is generally considered feasible. With such a practical issue, more than 60% power is allocated to pass the equivalence test at the assumed true mean difference when the number of lot is less than 10. The possible power values assigned to different sample sizes with biologists were derived through internal discussion in the FDA with various disciplines involved; the initial power values assigned to the given sample sizes that may cover extremely small to moderate sample sizes in biosimilar development are provided in Table 1.

To generalize the power to any number of lots, a monotonic exponential function was fitted through the five values in Table 1. The fitted curve is

$$\text{Power}(n) = 1 - \exp(-0.53948 - 0.14694n + 0.00205n^2). \quad (5)$$

With this monotonic function, the fitted power curve of the power is shown in Fig. 2.

Table 1 Initial power assignment for the given number of lots assuming the true mean difference meets the “targeted” similarity criteria

Number of lots per product	Assigned power (%)
6	72
10	85
15	90
20	92
25	95

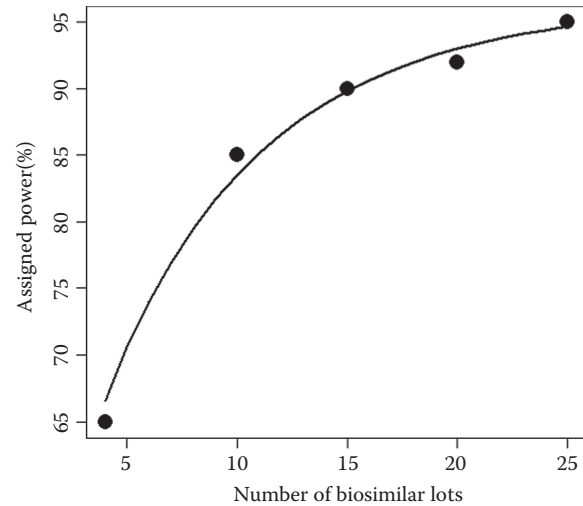


Fig. 2 Fitted curve of assigned power as a function of the number of lots per product assuming the true mean difference meets the “targeted” similarity criteria

Next, a target similarity criterion Δ_0 to determine the multiplier f for the margin $\delta = f \times \sigma_R$ is derived using the power calculation. The two one-sided equivalence test hypotheses become

$$\begin{aligned} H_{01} : \mu_T - \mu_R \leq -f \times \sigma_R & \text{ vs. } H_{a1} : \mu_T - \mu_R > -f \sigma_R \\ H_{01} : \mu_T - \mu_R \geq f \times \sigma_R & \text{ vs. } H_{a1} : \mu_T - \mu_R < f \sigma_R \end{aligned} \quad (6)$$

The corresponding two tests are

$$T_1 = \frac{(\bar{X}_T - \bar{X}_R) + f \times \sigma_R}{\sqrt{S_T^2/n_T + S_R^2/n_R}}, T_2 = \frac{(\bar{X}_T - \bar{X}_R) - f \times \sigma_R}{\sqrt{S_T^2/n_T + S_R^2/n_R}}.$$

For the objectives as described, a target similarity criterion Δ_0 is identified in order to determine the multiplier f for the margin $\delta = f \times \sigma_R$ using the power calculation. The sample size–power function for TOST with equal sample sizes and equal variance assumption when is derived for the target mean difference Δ_0 is $\eta \times \sigma_R$ and the equivalence margin is $f \times \sigma_R$, where σ_R is assumed to be the known standard deviation of the reference product:

$$\begin{aligned} \text{Power}(n) &= \Pr(T_1 > t_{1-\alpha, 2(n-1)}, T_2 < t_{\alpha, 2(n-1)}) \\ &= E_{U^2} \left[\Phi \left(\frac{f - \eta}{\sqrt{2/n}} - t_{1-\alpha, 2(n-1)} \sqrt{\frac{U^2}{2(n-1)}} \right) \right. \\ &\quad \left. - \Phi \left(\frac{-f - \eta}{\sqrt{2/n}} + t_{1-\alpha, 2(n-1)} \sqrt{\frac{U^2}{2(n-1)}} \right) \right] \end{aligned} \quad (7)$$

where U^2 follows the chi-square distribution with $2(n-1)$ degrees of freedom, Φ is the cumulative probability function of the standard normal distribution, and E is the expectation with respect to U^2 . The power curves are generated for $\eta = 1/16, 1/8, 3/16, 1/4, 5/16$, and $1/2$. After

Table 2 Proposed multiplier f for the margin $\delta = f \times \sigma_R$ to achieve the assigned power with a true mean difference of $1/8 \sigma_R$ and $\alpha = 0.05$ with the given number of lots per product n

	$n = 6$	$n = 7$	$n = 8$	$n = 9$	$n = 10$	$n = 11$	$n = 12$	$n = 13$	$n = 14$	$n = 15$
Assigned power (%)	74	77	79	82	84	85	87	88	89%	90%
f	1.74	1.64	1.55	1.50	1.45	1.39	1.36	1.33	1.30	1.27
	$n = 16$	$n = 17$	$n = 18$	$n = 19$	$n = 20$	$n = 21$	$n = 22$	$n = 23$	$n = 24$	$n = 25$
Assigned power (%)	91	91	92	93	93	93	94	94	94	95
F	1.25	1.21	1.20	1.19	1.16	1.13	1.13	1.11	1.09	1.09

carefully studying the results with representatives of all related disciplines in the FDA, the target mean difference Δ_0 is selected as $1/8 \sigma_R$. Then, under $\Delta_0 = 1/8 \sigma_R$, the corresponding multiplier f values in the equivalence margin of $\delta = f \times \sigma_R$ to achieve the assigned power calculated using Eq. 7 are listed in Table 2. The table shows that the multiplier f decreases with an increase in the sample size. When the sample size is small, the equivalence margin is wider compared to that for the larger sample size. However, this does not mean that the acceptance criteria of analytical similarity are easier for small sample sizes. Smaller sample sizes are adequately justifiable if the product is available for small patient population.

Keep in mind that the power of passing the equivalence test is controlled at lower levels for small sample sizes although the margin appears to be wider. In order to make the implementation of the equivalence test easier, the same value of the multiplier f for all sample sizes may be needed but adjusted for the value of α instead. For this objective, $n = 10$ is selected as the reference point and $f = 1.5$ is assigned instead of 1.45 for the margin. This increases the power from 84% to 87% when using $\alpha = 0.05$. The power values assigned to other sample sizes also slightly change.

Finally, the equivalence margin as $\delta = 1.5 \times \sigma_R$ is chosen by an FDA committee of various disciplines. Correspondingly, the two one-sided equivalence hypotheses become

$$\begin{aligned} H_{01} : \mu_T - \mu_R &\leq -1.5\sigma_R \text{ vs. } H_{a1} : \mu_T - \mu_R > -1.5\sigma_R \\ H_{02} : \mu_T - \mu_R &\geq 1.5\sigma_R \text{ vs. } H_{a2} : \mu_T - \mu_R < 1.5\sigma_R \end{aligned} \quad (8)$$

As mentioned earlier, the equivalence test can also be performed by comparing the $(1 - 2\alpha)$ 100% two-sided confidence interval of the mean difference against the equivalence margin, where $(1 - 2\alpha)$ 100% is the confidence level. When applying the power function (Eq. 7) with a margin of $1.5\sigma_R$, the confidence levels required are derived for each equal sample size from 6 to 15 and listed in Table 3.

ISSUES OF THE PROPOSED STATISTICAL APPROACHES FOR TIER 1 CQAS

There are potential issues regarding our proposed statistical equivalence test for Tier 1. First, let us examine the

Table 3 Confidence level with $\delta = 1.50\sigma_R$ for various sample sizes

Number of lots per product	Margin	Assigned power at mean difference of $1/8\sigma_R$ (%)	Confidence level (%)
6	$1.50\sigma_R$	78	78.4
7		82	81.4
8		84	85.0
9		86	87.4
10		87	90.0
15		90	90.0

underlying assumptions to ensure proper statistical properties of our proposed equivalence testing. The first assumption is that the samples are randomly selected and are representative of the product. In practice, the reference lots are purchased from the market with unknown manufacturing dates. If the reference lots are selected from a wide period of manufacturing dates, then it is likely to have proper estimate of reference variability of the whole population of reference lots. In contrast, if the reference lots are chosen from a correlated sample, e.g., successive lots may use some materials in common, then the test results may not be representative of the product or the variability can be underestimated and the mean estimate may also be biased. Using a random-effects model to adjust for the variance based on the small sample of lots may be biased. On the other hand, the biosimilar lots are usually freshly produced with known manufacturing dates and known manufacturing process. Thus, more knowledge of the biosimilar variability will be attained in the future. But in the same way as the reference sample, the sample mean and variance may not represent the population of biosimilar lots to be manufactured in the future. For details of the discussion, please refer to the works of Burdick et al.,^[8] Tsong et al.,^[9] and Shen et al.^[10]

The second assumption of the proposed equivalence testing is that the standard deviation σ_R for the margin determination is a known parameter. In practice, σ_R is usually unknown and is estimated from samples. For this reason, using estimate as constant may have the impact on type I error rate and power. This issue is also studied

and reported separately by Burdick et al.,^[8] Dong and Weng et al.^[11]

In addition, the lot value X is often the average of k replicates within the same lot. Then, X would have a mean value of μ_i and a variance of $\sigma_a^2 + \sigma_e^2/k$, where σ_a^2 and σ_e^2 are the between- and within-lot variances, respectively. As discussed earlier, the variance of $\bar{X}$ (i.e., the sample mean value across different lots) is a linear combination of those two variance components; then its value reduces with an increase number of replicates within the lot. The expression of variance of $\bar{X}$ could be more complicated if the within-lot variance is different from lot to lot. For some detailed discussion, please refer to the work of Tsong et al.^[13]

The third unstated assumption is that the data are considered blind if they are analyzed only once. In reality, more samples may be added in a later stage before Biologics License Application submission. In addition, the number of reference lots can be much larger than the number of biosimilar lots, because purchasing sufficient reference lots from the market is the first step in the biosimilar development. Thus, the sample size imbalance may cause some concerns in analytical similarity evaluation for Tier 1. Dong et al.^[14] describe those concerns and the potential statistical approaches to addressing the issues.

As described earlier, the multiplier 1.5 in the proposed equivalence margin is determined based on an assumed mean difference of $\mu_T - \mu_R = \sigma_R/8$. Both the assumed mean difference and the multiplier, 1.5, were selected based on power simulation results. These values may need to be revised when the statistical approaches are updated in the future.

REFERENCES

1. FDA. Scientific Considerations in Demonstrating Biosimilarity to a Reference Product. U.S. Food and Drug Administration: Silver Spring, MD, 2015.
2. FDA. Quality Considerations in Demonstrating Biosimilarity of a Therapeutic Protein Product to a Reference Product. U.S. Food and Drug Administration: Silver Spring, MD, 2015.
3. FDA. Clinical Pharmacology Data to Support a Demonstration of Biosimilarity to a Reference Product. U.S. Food and Drug Administration: Silver Spring, MD, 2015.
4. Christl, L. Overview of Regulatory Pathway and FDA's Guidance for Development and Approval of Biosimilar Products in US. Presented at FDA ODAC meeting, January 7, 2015, Silver Spring, MD, 2015. www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeeting-Materials/Drugs/OncologicDrugsAdvisoryCommittee/UCM436387.pdf.
5. Satterthwaite, F.E. An approximate distribution of estimates of variance components. *Biomet. Bull.* **1946**, 2, 110–114.
6. Chow, S.-C., Shao, J. A note on statistical methods for assessing therapeutic equivalence. *Cont. Clin. Trials*. 2002, 23, 515–520.
7. Limentani, G.B.; Ringo, M.C.; Ye, F.; Bergquist, M.L., McSorley, E.O. Beyond the t-test: Statistical equivalence testing. *Anal. Chem.* **2005**, 77, 221A–226A.
8. Burdick, R., Coeffey, T., Gutka, H., Gratzl, G., Conlon, H.D., Huang, C.-T., Boyne, M., Kuehne, H. Statistical approaches to assess biosimilarity from analytical data. *AAPS J.* **2016**. doi:10.1208/s12248-016-9968-0.
9. Tsong, Y., Xia, Q., Weng, Y.-T. Commentary on “Statistical approaches to assess biosimilarity from analytical data” by Burdick et al. *AAPS J.* **2016**. doi: 10.1208/s12248-016-9987-x.
10. Shen, M., Wang, T., Tsong, Y. Statistical consideration regarding correlated lots in analytical biosimilar equivalence test. *J. Biopharm. Stat.* **2016**. doi: 10.1080/10543406.2016.1265541.
11. Dong, X., Bian, Y., Tsong, Y., Wang, T. Exact test-based approach for equivalence test with parameter margin. *J. Biopharm. Stat.* **2016**. doi: 10.1080/10543406.2016.1265546.
12. Weng Y.-T., Tsong Y., Shen M., Wang C. Improved Wald test for equivalence assessment of analytical biosimilarity. *Inter. J. Clin. Biostat. Biometrics*. DOI:10.23937/2469-5831/1510016.
13. Tsong, Y. Development of statistical methods for analytical biosimilarity assessment. *J. Biopharm. Stat.* **2016**. doi: 10.1080/10543406.2016.1272606.
14. Dong, X., Weng, Y.-T., Tsong, Y. Adjustment for unbalanced sample size for analytical biosimilar equivalence assessment. *J. Biopharm. Stat.* **2016a**. doi: 10.1080/10543406.2016.1265544.

Titration Design

Marilyn A. Agin

Pfizer Global Research and Development, Ann Arbor, Michigan, U.S.A.

Edward F. C. Pun

Agouron Pharmaceuticals, San Diego, California, U.S.A.

Abstract

In the typical titration study, the response is defined according to some predetermined criterion, the dose is increased over time, and a subject may receive more than one dose level. This entry provides reasons and rationale for utilizing titration design in clinical trials as well as a sample data analysis.

INTRODUCTION

One of the most critical and challenging tasks in developing a new drug is to determine a dose range that provides an optimal therapeutic risk/benefit ratio in the targeted population. Clinical trials carried out to explore this optimal dose range can be grouped into two types of study design.

The *parallel-groups, fixed-dose* study is the first type of design. It is the simplest and most commonly used design to explore the dose range. In the parallel study, subjects are randomly assigned to receive one of a set of predetermined doses for the entire study duration. No dose adjustments are allowed, so a subject receives only one dose level.

The *dose titration* study is the second type of design. In this study, all subjects are started on the lowest of a pre-specified set of dose levels. Subjects who respond on the lowest dose stay on that dose; those who do not respond on the lowest dose after a fixed treatment period receive the next highest dose. This procedure is repeated until all subjects either drop out, respond, or reach the highest dose level.

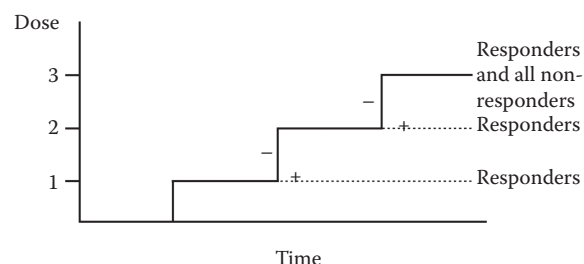
OVERVIEW

In the typical titration study, the response is defined according to some predetermined criterion, the dose is increased over time, and a subject may receive more than one dose level. Variations on the typical titration design include the use of a concurrent placebo group, the titration of the dose level up or down, and a forced titration, in which all subjects receive a higher dose regardless of their response at a lower dose, assuming they do not drop out for safety or other reasons.

The layout of the parallel design is:



The layout of the titration design is:



The fundamental difference between a titration study and a parallel study is that in a titration study, a subject receives a higher dose only if the subject fails to respond at a lower dose. At each titration step, the target populations are different. For example, in a parallel study, the target population for subjects who receive dose level 2 is all subjects who have the disease. In a titration study, the target population for subjects who receive dose level 2 is all subjects who have the disease but failed to respond on dose level 1. Thus the target populations at dose level 2 are different in each study. This is why the proportion of responders at a particular dose level in a parallel study should not be compared directly with the proportion of responders at that same level in a titration study. It is also the main reason why the analysis and the interpretation of a titration study are more complicated than that of a parallel study.^[1,2]

In general, if the purpose of the study is to compare the drug as a whole to placebo, then titration can provide a clear evidence of effectiveness and valuable information on population average and individual dose–response relationships.^[3] If the purpose of the study is to compare several doses to placebo, then the use of titration makes it more difficult to evaluate the effect of the drug by dose, and a careful analysis is needed.

USE OF TITRATION DESIGNS

Clinical Rationale

In certain circumstances, a titration design is desirable and sometimes even necessary for clinical reasons. In a titration study, subjects are exposed only to the amount of drug necessary to control their disease. This is good clinical practice and is the main reason to titrate the dose. Titration studies also provide information on the sequential and cumulative effects of the drug, so they are often used in clinical trials with drugs intended for long-term use at varying dose levels, such as antihypertensive drugs. Titration designs are also used in clinical trials in evaluating formulations of insulin for diabetic subjects, because it is usually necessary that the dose for each individual subject be titrated up or down until the desired response is reached. Titration may also be necessary because a subject may build up tolerance to a drug. For example, when taking an alpha blocker, a subject's dose level may need to be increased gradually to avoid orthostatic hypotension.

Even if titration appears to be appropriate for clinical reasons, the following requirements are necessary for the appropriate use of a titration study:

1. The response to the titration must be well-defined and easily measurable.
2. The number of dropouts cannot be excessive.^[3]
3. The cumulative effect of the drug must be minimal.^[3]
4. The time until the response develops must be relatively short. For example, in a hypertension study, the response might be a decrease in supine diastolic blood pressure to 90 mm Hg or less, which is sustained over a 2-week period.^[1,3]
5. If the response to a titration is irreversible, such as cure of a bacterial infection, then the use of a titration design is not appropriate.^[3]
6. Titration designs may not be suitable in diseases such as osteoporosis, Alzheimer's disease, or depression, because a desired response may develop only after a long period of time, and even if the treatment period is extended, the response may be difficult to measure.

Disadvantages of Titration

In a typical titration study, the dose is confounded with the duration of treatment because a lower dose is always followed by a higher dose. Over the course of treatment, the disease may progress, the subjects may exhibit spontaneous improvement, or the drug may have some cumulative effects. As a result, dose and time effects are hard to separate. One of the key issues in a titration study therefore is deciding when to titrate the dose. If the dose is titrated upward before its full effect is realized, the subject will be exposed to a level of drug that is unnecessarily high, and thus the recommended dose will be overestimated. This can happen in a confirmatory study where the goal is to compare the efficacy of several doses, or when a high dose is considered appropriate because of the urgent need for an effective treatment. In cases where the dose-related toxicity is relatively small and there is no safety incentive to minimize the dose, a likely tendency is to choose the highest dose that is reasonably well tolerated. Freston^[4] presented several examples of this tendency in hypertension studies. In another example, given by Temple,^[5] one-twelfth of the highest dose level used in a titration study was eventually considered to be the maximum effective dose. This problem can be minimized somewhat by extending the treatment time on each drug level.

With some drugs, a carryover effect may exist from one dosing period to another. In a titration study, because the doses are given in increasing order, these carryover effects may be harder to estimate than in other types of designs. For example, in a crossover study, subjects are given several different levels of drug, but in a preset random order rather than in a strictly increasing order. In general, however, crossover studies and titration studies are used to answer different questions.

Advantages of Titration

Titration designs provide practical as well as statistical advantages. The main practical advantages of titration designs are that they are in accordance with good clinical practice, they allow subjects to be exposed only to the amount of drug necessary to control their disease, and they mimic the way the drug will be prescribed. One statistical advantage is that because each subject may receive more than one dose, titration designs can provide information on the distribution of individual dose–response curves as well as on the population-average, dose–response curve. Another statistical advantage is that titration studies can be used to examine the response over a wide range of doses, usually with a smaller number of subjects, than a parallel study with the same statistical power.^[1,2,3] Sheiner et al.^[6] compared parallel, crossover, and titration designs in dose-ranging studies. They concluded that if a

patient-specific parametric dose–response model is used, then a titration study can be more advantageous than either a parallel study or a crossover study.

Titration designs are useful in phase I tolerance trials, in which the goal is to assess the safety of the drug. They are also valuable in early phase I/II exploratory studies, when the drug is assumed to have an effect but little is known about the dose–response relationship.^[1,7] In these studies, the goal is to examine the effectiveness of the drug over a wide range of doses, and to define the shape of the dose–response curve. Thus the titration design may be helpful in choosing doses for a later, confirmatory parallel study, in which hypotheses about the drug’s efficacy are tested.^[3]

Titration designs are also appropriate in phase III randomized trials if the intended way for the drug to be prescribed is in a titration scheme. An additional benefit in this case is that subjects used in phase III studies more closely resemble the target population of subjects with the disease. With the inclusion of a placebo group and appropriate analysis, these later-stage titration studies can also provide pieces of evidence of the safety and overall efficacy of the drug as well as evidence of the effectiveness of various doses.^[8]

ANALYSIS OF A TITRATION DESIGN

Types of Analysis

Life Table Approach

In the discrete life table approach to the analysis of a titration study,^[1] three general assumptions are made. First, the period during which one dose is administered is long enough for the physiological systems to stabilize. Second, a subject who responds at one dose will respond at all higher doses. Third, given the number of subjects who receive a particular dose, the number of responders at that dose follows a binomial distribution. The probability of response in that binomial distribution is a conditional probability, given that the subject failed to respond at all lower doses. If all three assumptions are satisfied, then a relationship exists between the unconditional probability of response at a particular dose level, as commonly defined in a parallel study, and the conditional probability of the titration study. This method of analysis resembles a life table approach in that the end event is defined by the response to the drug and the follow-up periods by the dose level.

In Chuang’s life table approach, a logistic linear model is used to define the dose–response relationship. The parameters in the model are estimated using iterated reweighted least squares, and the procedure is applied to the study of an antihypertensive compound.

Life table estimates can be imprecise, however, if subjects in the titration study skip doses, receive too high a dose, or receive a low dose after a high dose. Shih et al.^[8] address these problems in the analysis of a phase III study.

Incomplete Design Approach

The incomplete design approach to the analysis of a titration study^[9] generalizes Chuang’s life table approach. In a complete design, all subjects are exposed to all dose levels. The titration design can be viewed as an incomplete design because all subjects are not exposed to all doses. Wong^[9] also provides estimates of lower and upper bounds for the unconditional probabilities of response.

The assumption that a subject who responds at one dose will respond at all higher doses is not made in the incomplete design approach. If this assumption is included, however, the estimates from the life table and incomplete design approaches will be the same. As in the life table approach, dropouts are assumed to occur at random.

Contingency Table Approach

Once a dose–response relationship has been established in phase I/II clinical trials, efficacy and safety are confirmed in phase III. A phase III titration study can be used to determine if the overall response rate is satisfactory, the starting dose is therapeutically useful, and a maximally effective dose is reached whether or not one titrates up to the high dose.

Shih et al.^[8] analyze the data in a phase III antihypertensive study as a one-dimensional contingency table including incomplete (grouped) information on the number of dropouts at each dose level, obtaining maximum likelihood parameter estimates through the EM algorithm. They do not require knowledge of the number of subjects who receive a particular dose level as in the life table approach, but only of the lowest dose level to which each subject responded before withdrawal. This is because in large phase III titration studies, even if the design is well defined, nonresponders may skip steps in the titration scheme, or responders may be escalated to an unnecessarily high dose or even have their dose reduced. The EM algorithm is not affected by these irregularities, but the life table approach can be affected.

As in the life table approach, it is assumed that the dropouts are withdrawals between two successive doses and that their censoring mechanism is independent of efficacy. A method for incorporating dropouts into the efficacy evaluation is given by Chuang-Stein and Shih.^[10]

Titration with Placebo Controls

If a placebo effect is expected to be high, the magnitude of the effect as well as the efficacy of the drug as a whole can be estimated by the inclusion of a placebo control group. If placebo controls are included concurrently with each dose group, the effect of the dose can be measured relative to the placebo controls for that dosing group.

Chuang-Stein^[7] extended her life table approach to include a placebo control. She used logistic linear models and iterated reweighted least squares in two approaches to compare the responses of the test drug groups to those of the placebo controls. In the first approach, similar models were assumed for both the drug and the control groups. In the second approach, it was assumed that the response rates of the test drug groups were consistently higher than those of the placebo groups, and the response rates were evaluated relative to those of the corresponding placebo groups.

Pun^[2] compared and contrasted parallel and titration dose-ranging studies. He recommended the inclusion of a concurrent placebo group in a titration design in order to investigate the risk/benefit ratio for subjects who do not respond to a lower dose and must be titrated up to a higher dose. He showed that for a fixed sample size, the titration design has more accuracy and power in determining the dose range than does the parallel design.

Analysis of Safety Data

Even the response is easily measurable and occurs in a short period of time, the analysis of safety data in a titration study presents special problems, and the estimation of the risk/benefit ratio is more challenging. The incidence of an adverse event is usually defined as the proportion of subjects who experience the event at a given dose. Because the effect of the drug is confounded with time, the pattern of occurrence of an adverse event or a laboratory abnormality might be mistaken for a drug effect. For example, suppose a particular adverse event tends to occur early in drug treatment. As the dose is titrated upward, the incidence of the adverse event gets smaller, mistakenly indicating a negative dose effect.

In addition, the assumptions in analyzing the safety data in a titration study are not the same as those in analyzing the efficacy data. A necessary assumption in most analyses of efficacy data in a titration study is that a subject who responds at one dose level will respond at all higher dose levels. If a subject experiences an adverse event at one dose level, however, it does not imply that the subject would have experienced the same adverse event at all higher dose levels. It also does not imply that increasing the dose necessarily makes the adverse event worse, although sometimes this is true. For example, suppose a subject experiences a headache on the lowest dose level and dizziness on the next higher dose level.

The dizziness may be due only to the time on drug and not to the higher dose level. That is, dizziness would have been experienced in the second treatment period even if the subject had not been titrated up to the higher dose. On the other hand, a side effect that appears on one dose level but disappears at a higher dose level may be due to the subject's having adjusted to the pharmacological effect of the drug and not to the higher dose level. Thus the analysis of safety data in a titration study presents special problems.

Hsu and Laddu^[11] state that because of the pattern of occurrence of adverse events in a titration study, no dose-specific evaluation of adverse events can be made. In order to evaluate the time trend, they recommend the inclusion of a placebo control group, the dose level, the time of occurrence of the adverse event, and the speed of titration. They also recommend the comparison of the results to those from a parallel study.

CONCLUSION

Parallel and titration designs are used extensively in dose-finding and comparative studies. The basic difference between them is that they are used to answer different questions.^[2,8,12] Parallel studies are useful in determining the effectiveness of various doses through the use of independent dose groups having the same duration of treatment, in anticipation of fixed-dose, short-term prescriptions. Titration studies provide information on the sequential and cumulative effects of the drug at various dose levels, and are useful when the drug will be prescribed for long-term use at varying dose levels.^[8]

When titration designs are used and analyzed appropriately, they can provide valuable information on the safety and the efficacy of the drug. By including a concurrent placebo group and taking into account the dose, time, and speed of titration, the titration design can provide information on the overall safety and efficacy of the drug as well as on the effect by dose, but the results for each individual subject remain somewhat confounded by time.^[3,11,13,14]

REFERENCES

1. Chuang, C. The analysis of a titration study. *Stat. Med.* **1987**, *6*, 583–590.
2. Pun, E.F.C. A parallel dose–titration design to determine dose range. *ASA Pro Biophys.* **1990**; 159–164.
3. ICH Harmonized Tripartite Guideline, E-4. In *Dose–Response Information to Support Drug Registration*; 1994.
4. Freston, J.W. Dose-ranging in clinical trials: Rationale and proposed use with placebo or positive controls. *Am. J. Gastroenterol.* **1986**, *81*, 307–311.
5. Temple, R. Government viewpoint of clinical trials of cardiovascular drugs. *Cardiovasc. Pharmacother.* **1989**, *III*, 495–509.

6. Sheiner, L.B.; Beal, S.L.; Sambol, N.C. Study designs for dose-ranging. *Clin. Pharmacol. Ther.* **1989**, *46* (1), 63–77.
7. Chuang-Stein, C. Two approaches for the analysis of a titration study with a placebo control. *Commun. Stat., Theory Methods* **1988**, *17*, 821–832.
8. Shih, W.J.; Gould, A.L.; Hwang, I.K. The analysis of titration studies in phase III clinical trials. *Stat. Med.* **1989**, *8*, 583–591.
9. Wong, R. Estimating the probability of a response from titration studies. *ASA Pro Biophys.* **1992**; 137–139.
10. Chuang-Stein, C.; Shih, W.J. A note on the analysis of titration studies. *Stat. Med.* **1991**, *10*, 323–328.
11. Hsu, P.; Laddu, A.R. Analysis of adverse events in a dose titration study. *J. Clin. Pharmacol.* **1994**, *34*, 136–141.
12. Turri, M.; Stein, G. The determination of practically useful doses of new drugs: Some methodological considerations. *Stat. Med.* **1986**, *5*, 449–457.
13. Ruberg, S.J. Dose response studies: I. Some design considerations. *J. Biopharm. Stat.* **1995**, *5*, 1–14.
14. Ruberg, S.J. Dose response studies: II. Analysis and interpretation. *J. Biopharm. Stat.* **1995**, *5*, 15–42.

Toxicological Studies

Guenther Baus

Aventis Pharma Deutschland GmbH, Frankfurt, Germany

Thomas Hoffman

Aventis Pharma Deutschland GmbH, Hattersheim, Germany

Abstract

Toxicology is defined as the knowledge of harmful effects of chemical substances on living organisms. The focus of this entry is on detailing several testing methods for determining toxicology and also stresses the importance of statistical analysis being performed before the study itself.

INTRODUCTION

Toxicology is defined as “the knowledge of harmful effects of chemical substances on living organisms.” The methods of toxicology are used to assess the potential risk of substances to man and animals, in order to avoid adverse effects on health. Quite simply spoken, a substance is called a poison when it can produce harmful effects in an organism. However, it should be always kept in mind that solely the dose of a substance determines whether damage occurs or not. The physician Paracelsus recognized already in the 16th century that all substances are poisons and there is no nonpoison. It is only the dose that distinguishes a poison from a remedy.

A variety of methods to assess toxic effects method. Acute toxicity testing measures morbidity and mortality after a single administration and is probably the simplest method for assessing toxic effects. On the other hand, subchronic and chronic toxicity tests investigate effects caused by a long-term administration of a compound. A further classification of toxic effects can be drawn between local and systemic effects. As an example, damage by acid burns at the site of exposure is a local effect. In contrast, systemic effects involve the affection of target organs irrespective of the exposure site. Chronic exposure to benzene, as an example, can lead to damages to the bone marrow and to leukemia after dermal exposure or inhalation. Effects of compounds on the genetic material are investigated using genotoxicity tests. Reproductive toxicology studies are carried out in order to detect effects on fertility and early embryonic development, on prenatal and postnatal toxicity, and on embryogenesis and fetogenesis. The capability of chemical compounds to induce tumors is usually examined in long-term carcinogenesis bioassays in rats and mice. Furthermore, a variety of study types to examine other end points (e.g., skin and eye irritation, skin sensitization, local injection tolerance, immunotoxicity and neurotoxicity) exist.

Although the main study types including those without statistical evaluation are summarized below in order to cover the whole spectrum, the description of the statistical analysis is mainly focused on subchronic and chronic toxicity studies. However, these methods can also be applied to other studies, if data requirements are met.

REGULATORY REQUIREMENTS

Drugs, agricultural products, and industrial chemicals must be characterized regarding safety before getting marketing authorization. With respect to this, toxicological studies are routinely required by regulatory agencies and must be conducted according to standard study protocols, which are published as so-called “guidelines” by the respective regulatory agencies. As an example, the outlines of toxicological studies needed for new drugs were evaluated by the International Conference on Harmonization (ICH). These ICH guidelines represent the European Union, Japan, and the United States, and are implemented at the national level by adopting or publishing. Regulatory requirements before an investigational new drug can be applied to humans usually include: acute toxicity testing in two species (usually rats and mice) via two administration routes, of which one should represent the intended clinical route; safety pharmacology studies covering central nervous system, cardiovascular, and respiratory end points; a minimum of three genotoxicity tests (usually Ames test, chromosomal aberration study in vitro, micronucleus test); and repeated dose studies (at least 14 days) in one rodent species and nonrodent species. The duration of the latter studies should be at least equal to the duration of the clinical trial. Generally, the route of administration should mimic the intended clinical route. Additional testing for local injection tolerance is necessary in case of parenteral administration. For a New Drug Application (NDA)

submission, additional studies investigating the effects of the drug on reproduction, on the developing fetus (usually rats and rabbits), and on prenatal and postnatal development are required, as well as chronic toxicity (6-month rodent study and 6- to 12-month nonrodent study) and carcinogenicity studies in two species (usually rats and mice). In case of industrial chemicals and agricultural products, toxicity testing requirements are similar to those applying to new drugs and are covered by the guidelines published by the U.S. Environmental Protection Agency.

STUDY DESIGN AND AIM OF STUDIES

A large number of study types were published for a variety of end points. A comprehensive description of all these study types is beyond the scope of the present paper. Therefore the following sections focus mainly on studies that are routinely applied during the safety assessment of new drugs, agrochemicals, and industrial chemicals.

Acute Toxicity

Single-Dose Toxicity

One of the first steps in studying the toxic effects of an unknown substance is the determination of its toxicity after a single administration. This provides a first indication of the kind of toxic effects and gives information about the lethal dose. In this study, the test compound is administered usually by oral, dermal, inhalational, intravenous, or other routes of exposure. The route of administration is dependent on the intended clinical use in case of drugs, or the possible route of exposure in case of agrochemicals and industrial chemicals. After administration, the animals are observed at least once daily for mortality and clinical signs of intoxication, including—but not limited to—changes in skin and fur, mucous membranes, behavior, sensory reflexes, motor activity, breathing, circulatory effects, and central nervous effects. Body weight is recorded at the start of the study and then weekly thereafter. After death or at the end of the study (usually 14 days after administration), the animals are necropsied and checked for macroscopically visible changes. In the classic test design, different dosages are selected in order to determine the LD_{50} (median lethal dose). This value describes the dose that would kill 50% of the test animals and is calculated by Probit analysis.^[1] A prerequisite for this calculation is that at least two mortality ratios are between 0% and 100%. For agricultural products and industrial chemicals, the LD_{50} value is used for the classification of compounds. However, this classic test can be replaced by the determination of the approximate LD_{50} in many cases, which saves a substantial number of test animals. Nowadays, the determination of an exact LD_{50} value is no longer needed for the approval of new drugs,^[2] agrochemicals, and industrial chemicals.^[3–6]

Testing for Acute Dermal Irritation

These tests are conducted in order to evaluate the irritant and/or corrosive effects of a compound to the skin of mammals. Whereas dermal irritation is a reversible inflammation of the skin, dermal corrosion is an irreversible damage resulting in scar formation. Usually, this test is conducted for industrial chemicals, agrochemicals, and pharmaceuticals used for dermal application.

In the skin irritation test, 0.5 g or 0.5 mL of the test compound is applied at three sites to the shaved skin on the back of one albino rabbit and covered with a semioclusive dressing. The patches are removed successively after 3 min, 1 hr, and 4 hr, and remnants of the test compound are gently removed. If no corrosive effect occurs after 4 hr, then two animals are added with one patch and an exposure period of 4 hr. If corrosive effects are observed after 3 min or 1 hr of exposure, no further testing is required. The treated skin surface is evaluated 30–60 min, 24 hr, 48 hr, and 72 hr after the removal of the patch for erythema and eschar formation as well as for edema formation. Dermal irritation is scored according to gradings of 0 (*no irritation*) to 4 (*severe irritation*). The observation period is prolonged in order to evaluate reversibility if changes are still present after 72 hr.^[7] Depending on the irritation scores, the compound is classified as nonirritant, irritant, or corrosive, if the skin is destroyed through its whole thickness.

Testing for Acute Eye Irritation

Noncorrosive substances are tested for irritant effects on the eye or mucous membranes. In this study, 100 mg (solids) or 0.1 mL (liquids) of the test substance is applied to the conjunctival sac of albino rabbits. A single animal can be used initially if a severe irritant effect is expected. In other cases, at least three animals are required. The compound is removed by rinsing the eyes 24 hr later with physiological saline. The eyes are observed 1, 24, 48, and 72 hr after application, and corneal opacity, inflammation of the iris, conjunctival irritation, and conjunctival swelling are scored according to defined gradings. The compound is then classified according to these scores. If irritation occurs after 24 hr of exposure, it may be appropriate to test additional animals with a washout 30 sec after instillation.^[8]

A number of alternative test systems have been developed to replace test systems for dermal and eye irritation due to animal welfare reasons. They can be used as a screening test for different formulations, or may replace the standard test if corrosive effects are shown.

Test for Skin Sensitization

Skin sensitization studies are carried out in guinea pigs for the determination of an allergic potential (contact

dermatitis). Several study protocols have been developed for this purpose, which in common cover the induction phase, the resting period, and the challenge period. Some of these tests include the use of Freund's complete adjuvant (FCA) for the enhancement of the immunologic process. The guinea pig maximization test of Magnusson and Kligman,^[9] which uses adjuvants, and the Buehler test,^[10] which does not use adjuvants, are considered to be preferable. During the induction period, a slightly irritant concentration of the test compound is applied by dermal or intradermal route for a possible stimulation of immune-competent cells. During the subsequent 2-week resting period, a potential antibody formation reaches a maximum. During the challenge period, the test compound is then applied as a patch to the skin in a concentration that has been shown to be a nonirritant in preliminary tests. In case of sensitization, the immunological reaction leads to erythema and/or edema formation, which should be absent in a control group induced with the vehicle only but challenged identical to the test group. The skin is examined 24 and 48 hr after removal of the patch, and erythema and edema formation are scored using a grading system. In case of the Buehler test, usually 20 animals in the treatment group and 10 animals in the control group are used. In case of the maximization test, usually 10 animals in the treatment group and 5 animals in the control group are used. However, in case of questionable results, further animals are added in this test, yielding 20 treatment animals and 10 control animals.^[11]

Depending on the animal model used, the compound is regarded as a skin sensitizer if at least 15% (tests without the use of FCA) or 30% (tests with the use of FCA) of the animals show a positive skin reaction.

Screening tests have been developed for the detection of skin sensitizers, the local lymph node assay,^[12] and the mouse ear swelling test.^[13] In case of positive results in these tests, a compound can be classified as a sensitizer without further testing.

Safety Pharmacology Studies

Safety pharmacology studies are conducted usually preclinically with new drugs in order to examine the possible adverse effects of test substances on physiological functions at therapeutic dosages and above, which could impact human safety. In vivo, ex vivo, and in vitro studies can be used. In vivo or ex vivo studies are usually carried out after a single administration of the test compound. The route of administration should be the expected clinical route. Usually three dosages are tested, of which the dose levels should equal the pharmacologically active dose and the high dose should exceed the therapeutic range by a factor of 30–100 or be the maximum tolerated dose. The central nervous system, the cardiovascular system, and the respiratory system are examined in a first step. If safety concerns arise from these studies or during later clinical

development, then supplemental safety pharmacology studies are needed.^[14]

Testing for effects on the central nervous system include behavioral tests, motor activity, coordination, and body temperature sensory and motor reflex responses, and can be measured in rats or mice according to modified Irwin's test^[15] or by using a functional observation battery.^[16]

Effects on the respiratory system should be assessed quantitatively and include measurements of respiratory rate and other respiratory parameters (e.g., tidal volume, minute volume, inspiratory time, expiratory time), which can, for example, be done in rats or guinea pigs using whole body plethysmography.

Cardiovascular function is examined by measuring blood pressure, heart rate, and the quantitative evaluation of electrocardiograms (e.g., PR, QT, QTc). Usually these tests are carried out with unanesthetized dogs instrumented for telemetric measurements. Effects of test compounds on resting potential, repolarization, and conduction abnormalities can be examined electrophysiologically in vitro using Purkinje fibers of isolated hearts, for example.

Subchronic and Chronic Toxicity Studies

These study designs are used to investigate the toxic effects of the test substance after repeated daily administration. The duration of the study ranges from 14–28 days up to 12 months. In case of drugs, the study duration depends on the intended period of therapeutic use in humans. The route of administration should reflect on the expected exposure or clinical use in humans. Usually three different dosage groups and one control group, which is handled identical to the dosage groups and receives the vehicle only, are used. The lowest dose should not cause any adverse effects, whereas the high dose should produce toxicity but no mortality or severe suffering of the animals and usually does not need to exceed 1000 mg/kg body weight (limit dose). Besides the main groups sacrificed after the end of the treatment period, additional animal groups that are allowed a treatment-free recovery period postdosing may be added in order to determine reversibility or nonreversibility of effects seen at the end of the treatment period. The animals are observed daily for health condition and at least twice daily for morbidity and mortality. Ophthalmological examinations are performed before the start and toward the end of the study. Body weight gain and food consumption are recorded at regular intervals. In case of industrial chemicals and agrochemicals, a neurotoxicological assessment is performed including sensory reactivity, grip strength, and motor activity. At the end of the treatment and recovery period, hematological and clinical chemistry examinations are carried out (e.g., erythrocyte counts, leukocyte counts, thrombocyte counts, hemoglobin, hematocrit, blood coagulation, differential blood count, electrolytes, marker enzymes, urea, creatinine, glucose, cholesterol, lipids, electrophoresis, urinalysis).

At necropsy, all animals are examined for macroscopically visible changes, the main organs are weighed, and many organs are processed for histopathological examination. In studies with a 13-week or longer administration period, additional time points or animal groups may be added in order to examine the time course and the reversibility of observed effects. The required number of animals depends on the study duration and ranges from 5 to 20 animals per group and sex in rodents. In case of subchronic and chronic studies in nonrodents (usually dogs), which are required for the registration of new drugs and agrochemicals, usually four animals per sex and group are used. Further details on subchronic studies are given in the respective guidelines (e.g., Refs. [17–20]).

The aim of these studies is to evaluate the toxicological profile of the test compound [i.e., determination of target organs, dose–response relationships, and the ‘No Observed Effect Level’ (NOEL) at which no adverse effects are encountered after repeated administration of the compound]. The NOEL is used for the extrapolation of the effects seen in animals to humans. The classic procedure uses an uncertainty factor of 10 to extrapolate from the NOEL of the most sensitive animal species to humans, and a further uncertain factor of 10 for the variability between humans. Nowadays, extrapolation is based on the corresponding plasma levels of the test compound in this species at the NOEL, and the expected plasma levels in humans, if known.

Carcinogenicity Studies

These studies are conducted in order to evaluate a possible tumorigenic action of the test compound and are described in detail in the article *Carcinogenicity Studies of Pharmaceuticals*.

Genotoxicity Studies

Heritable changes of the genetic information stored in the DNA are called mutations. On one hand, mutations are the prerequisite for the evolution process and its mechanisms such as recombination, genetic drift, or selection. On the other hand, mutations are generally random and result in deleterious effects in most cases. Therefore the aim of mutagenicity tests is to detect interactions of the test compound with the genetic material. Three principle types of mutations can be distinguished: gene mutations, chromosome mutations, and genome mutations. Gene mutations are not visible with the aid of light microscopy. One example is a change in a biochemical function, which can be induced either by a substitution of a base pair in the DNA (substitution mutation), or by a deletion/insertion of a base pair in the DNA (frameshift mutation). Changes in the structure of the chromosomes are defined as chromosome mutations and can be detected by light microscopic

techniques. They cover gaps (parts of the chromosomes that are not stained), breaks (dislocated parts of the chromatids), missing parts (deletions), or exchanges between different chromosomes. Changes in the number of chromosomes are defined as genome mutations. Like chromosome mutations, they can be detected using light microscopy. Many different test protocols for the detection of mutagens have been published, but only few are used for regulatory purposes. In principle, these tests can be divided into tests for detecting gene mutations, methods for detecting chromosome aberrations, and DNA indicator tests for identifying nonspecific interactions of the test compound with the genetic material. Selected genotoxicity assays representing these three categories are described below.

Methods for Detection of Gene Mutations

Ames assay The Ames assay^[21] uses defect mutants of the bacteria strain *Salmonella typhimurium*. In contrast to the wild type, these strains are no longer able to synthesize the amino acid histidine and therefore cannot grow into colonies in agar medium without histidine. However, this defect mutation can be reverted by mutagenic agents, and growth without histidine is enabled. The *Salmonella* strains most commonly used are TA 98, TA 100, TA 1535, TA 1537, and TA 102, which carry either substitution or frameshift mutation markers. A compound, which induces reverse mutations in this system, is detected by a concentration-dependent increase in the number of colonies. Because bacteria have only a limited capacity to metabolize xenobiotics, a fraction of rat liver homogenate (S9mix) is added in order to enable metabolism processes present in mammalian organisms. It should be kept in mind that the results of this bacterial test may have only limited relevance for mammalian systems. However, the Ames test is an easy and quick test for the detection of mutagens, and the results have a high predictability.^[22] A detailed information on the conduction of the Ames test is given in the respective guidelines (e.g., Ref. [23]).

Other test systems Test systems using mammalian cells in vitro are closer to the target species than the Ames test. Assays commonly used are the HGPRT test^[24] and the mouse lymphoma test.^[25] An important difference in bacterial systems is that, in general, two copies of each gene are present in a normal mammalian cell. Therefore mutations are studied using hemizygous or heterozygous cells. An example for a hemizygous gene (X chromosomal) is the gene coding for the enzyme hypoxanthine-guanine phosphoribosyltransferase (HGPRT or HPRT). The function of this enzyme is to salvage the degradation products of purine metabolism. However, this pathway is not essential for survival because the synthesis of purine bases can also take place de novo. Toxic purine analogues such as thioguanine and azaguanine are incorporated into DNA or RNA

in cells containing HGPRT, resulting in the death of these cells. Mutant cells with altered or absent HGPRT function are able to survive and to grow into colonies because purine bases are synthesized de novo and toxic analogues are not incorporated. Thus the principle of the HGPRT test is to detect mutagens by an increase in the number of thioguanine-resistant or azaguanine-resistant colonies after an exposure of the cells to the test compound. Because cultured cells possess only a limited metabolic capability, an exogenous metabolizing system is added. This may be done either by the addition of S9 mix, or by a coculture with metabolically competent cells (e.g., hepatocytes).

In the mouse lymphoma test cells, heterozygous cells with one active allele of the gene coding for the enzyme thymidine kinase (TK + / -) are used. This enzyme incorporates thymidine into DNA. Toxic pyrimidine base analogues such as trifluorothymidine or bromodesoxyuridine are incorporated into DNA or RNA in cells containing at least one active allele of TK, resulting in the death of these cells. Like in the HGPRT test, mutagens are detected by an increase in the number of trifluoromethylthymidine-resistant or bromodesoxyuridine-resistant colonies after the incubation of the cells with the test substance. A detailed information on the conduction of these assays is given in the respective guidelines (e.g., Ref. [26]).

Tests for Detection of Chromosome Aberrations

Chromosomal aberration test in vitro Cytogenetic tests in vitro using mammalian cell cultures are usually performed with CHO cells (ovary cells), V79 cells (fibroblasts), or human lymphocytes. The cells are incubated for a period of 3 or 24 hr with different concentrations of the test compound. After incubation, mitosis is arrested in the metaphase with Colcemid® or Colchicine. Analogous to other in vitro systems, possible metabolism can be simulated by the addition of S9 mix to the test compound. From each concentration including the control cells, at least 100 metaphase plates are evaluated microscopically for structural chromosome changes. The data are usually presented as aberrant metaphases with and without gaps and metaphases with exchanges. A compound is considered clastogenic (i.e., inducing chromosome aberrations) if a concentration-dependent or a statistically significant increase in the number of metaphases with aberrations (excluding metaphases showing only gaps) is observed. A detailed information on the conduction of this assay is given in the respective guidelines (e.g., Ref. [27]).

Chromosomal aberration test in vivo Cytogenetic tests in vivo were performed in Chinese hamsters in the past because these species has a set of only 22 chromosomes, which can be distinguished easily. However, nowadays, this species is no longer used and cytogenetic examinations in vivo are usually performed in rats or mice. The test

compound is administered to the animals twice at an interval of 24 hr at least in three doses, of which the highest one should be the maximum tolerated one. Twenty hours after the last dose, mitosis is arrested in the metaphase and the animals are killed 4 hr later. Metaphase plates of the bone marrow are examined microscopically for structural abnormalities. A compound is considered clastogenic if a dose-dependent or a statistically significant increase in the number of metaphases with aberrations (excluding metaphases showing only gaps) is observed. A detailed information on the conduction of this assay is given in the respective guidelines (e.g., Ref. [28]).

Micronucleus test The micronucleus test is used for the detection of chromosome aberrations or the interference of the test compound with the mitotic spindle apparatus. Micronuclei are DNA fragments or whole chromosomes that are not separated correctly into the daughter cells during mitosis. They can be easily detected in erythrocytes because the nucleus of these cells is expelled during maturation. Originally, this test was developed using mice, but it can also performed in rats. Groups of five male and female animals receive the test compound twice within an interval of 24 hr in usually three different dosages. After 24 hr, the animals are necropsied and bone marrow smears are taken of the femora. The incidence of micronuclei in polychromatic erythrocytes is determined in the control and dose groups. A compound is considered clastogenic in this test system if a dose-dependent or statistically significant increase in the incidence of micronuclei is observed. A detailed information on the conduction of this assay is given in the respective guidelines (e.g., Ref. [29]).

DNA Indicator Tests

Besides the test systems described above covering specific end points of mutagenesis, a number of test systems have been established for which such a clear assignment cannot be given. The results of these studies give solely a hint of a general interaction with the genetic material. One test of this class is the test for induction of unscheduled DNA synthesis (UDS test).

UDS test The UDS test measures the repair of DNA damage induced by the test compound. This test is usually performed in hepatocytes either in vitro or ex vivo. In the former case, hepatocytes are incubated with the test compound, whereas hepatocytes from treated animals are used in the latter case. DNA synthesis (incorporation of [³H]thymidine) is measured in cells, which are arrested in the G phase of the cell cycle either by autoradiography or by liquid scintillation counting of the DNA. This (unscheduled) DNA synthesis is considered to reflect repair synthesis due to previous DNA damage. A detailed information

on the conduction of this assay is given in the respective guidelines (e.g., Ref. [30]).

Reproductive Toxicity Studies

Reproductive toxicity studies are conducted in order to evaluate effects to any stage of the reproductive cycle and are described in detail in the article *Reproductive/Development Studies*.

AIMS AND ISSUES IN STATISTICAL ANALYSIS

The following sections mainly deal with the statistical analysis of continuous parameters in subchronic and chronic toxicity studies. Nevertheless, some of the statistical techniques and tests presented below could easily be adapted to other study types, depending on the kind of data and study design to be analyzed. Only a few hints and additional literature for the analysis of some study types mentioned above will be given now. The statistical analysis of the Ames assay, including its specific statistical problems (i.e., the high variability of the data and the overlapping toxic and mutagenic effects at higher dose levels), has been described, for example, by Schumacher and Schmoor^[31] or by Hoffmann.^[32] A proposal for a statistical analysis for the HGPRT assay was given by, for example, Liu and Chow.^[33] To analyze the outcome of a micronucleus test statistically, a pairwise one-sided exact Wilcoxon test (see also “Statistical Approaches”) could be used in order to compare the incidence of micronuclei of the treated animals with the control. Another approach based on the Fisher–Pitman permutation test has been established by Leimer et al.^[34]

Aim of the Analysis

The main scope of statistical analysis for a toxicological study is to support the interpretation of the study outcome. In a typical subchronic or chronic toxicity study, around 50 parameters of hematology, clinical chemistry, urinalysis, and organ weights are under investigation, sometimes measured twice or even more during the course of the study. In addition, body weight and food consumption often are measured weekly. Due to this high amount of parameters and measurement times in most types of studies, it is hardly feasible to get one statistical decision on a predefined significance level for a complete study, which reflects the aim of the studies in most cases better, to make several statistical decisions on a studywise multiple significance level. In order to support the interpretation of the study results according to the aim of the studies (i.e., identification of the NOEL), one has to take into account that, on one hand, enough power is available to find toxicological relevant effects and, on the other hand, the studywise significance level should still be under control, if feasible. Using tests on differences, which is still the

commonly accepted method in toxicology, the aspect of an adequate power is most important as also discussed in following sections. Therefore it has been widely agreed to implement a multiple significance level of 5% for each parameter, measurement time, and sex, and not try to achieve a more comprising multiple significance level. Nevertheless, sometimes, still more than 100 statistical decisions have to be made, each of them on a 5% significance level. Discussions around this issue can be found in Hothorn.^[35]

How could statistical results gathered in that way be interpreted adequately? In general, the study director himself makes the final decision about toxicological effects. Either the study director accepts some minor toxicological effects—if any has been found at all—or the study director concludes to reject the substance. Depending on the study director’s knowledge about statistics, and attitude toward statistics, the statistical test results are used in a more or less descriptive way in most cases, especially when using routine analyzing systems without the involvement of statisticians. In very simple terms, the number of significances weighted by the importance of the parameters under investigation, combined with some general impressions based on the study data, leads to a general judgment about the substance under investigation. In addition, individual test outcomes on each parameter help to identify possible toxic end points. Only in some rare cases, especially when critical decisions have to be made, do statisticians have to be involved for a deeper analysis to support the interpretation of the results. In all other cases, the final decision is still made based on a high amount of experience of the toxicologist. Therefore it must be the aim of statistical analysis to influence this decision by adequate, objective inferential statistical methods. This should lead to more objective conclusions on the outcome of studies.

Equivalence Tests—The Better Choice?

Because it is the aim of toxicological studies to find a dose level where no relevant effect of the substance under investigation is present (NOEL), the use of statistical tests on differences between a treatment group and a control group seems to be questionable. Only if the results of these tests show no significant difference for a dose group under investigation is it concluded that no relevant effect occurred. However, it has to be emphasized that the probability for a wrong decision—the so-called β error, or error of the second kind—in this case is unknown. Therefore the power of the tests, defined as $1 - \beta$, is also not under control. However, because these tests are still in common use, the following sections rely on them and describe how to use them in a proper manner, which is definitely not optimal due to the above described property. Only a short introduction in equivalence tests, which would fit much better to make safety assessments, restricted to the two-sided case is given now.

The null hypothesis H_0 of an equivalence test is formulated as a two-sided difference and the alternative H_A is formulated as an equivalence hypothesis:

$$\begin{aligned} H_0: x_T &\geq x_0 + \delta \text{ or } x_T \leq x_0 - \delta \\ H_A: x_0 - \delta &< x_T < x_0 + \delta \end{aligned}$$

where x_0 is the mean of the control group; x_T is the mean of the treatment group; and δ is the limit of relevance region.

The main issue of an equivalence test is the a priori determination of the relevance region by a definition of the relevance limit δ . Although the a posteriori definition of δ has been described in Hauschke and Hothorn,^[36] an a priori definition should be preferred because it leads, in general, to more objective test results not affected by the outcome of the study itself. Once a suitable relevance limit has been found, it can be seen that there are two advantages of equivalence tests compared to tests on differences. First, as already described, the probability of erroneously deciding for equivalence—and therefore no toxicological effect—is fixed now. It is the error of the first kind or significance limit α that has to be a priori defined in the study protocol. Second, it is possible now, by defining δ correspondingly, to include a relevance of the statistical decision instead of only relying on the absence of a statistical significance while testing on differences.

In toxicological studies, a general issue for the a priori definition of the relevance limit δ is the high amount of parameters. Individual definitions of δ for each parameter based on the experience or knowledge of the study director about an adequate level of relevance would be highly influenced by a subjective opinion of toxicologists. A more objective approach for the determination of an adequate relevance limit δ could be based on a rule, which fits for all parameters and takes into account the estimated standard deviation of control animals based on historical data. Wellek^[37] describes an approach to use 75% of the standard deviation for normally distributed data, if no a priori definition of the relevance limit could be achieved, but not in the context of toxicological studies. Therefore some more experience has to be gained before using this method. It has to be discussed in detail if, on one hand, a general approach for all parameters is feasible and if, on the other hand, 75% of the estimated standard deviation based on historical controls could be accepted by toxicologists as a reasonable value for the definition of toxicological equivalence.

Some aspects of the practical use of equivalence tests (e.g., the description of the one-sided test problem and two-sided test problem, the determination of the NOEL under the assumption of a monotone dose–response relationship, and the use of parametric and nonparametric tests) have been described in Hauschke and Hothorn.^[36] However, due to the still existing issue in determining δ , further discussions within this paper only refer to the

assumption that tests on differences are used to analyze toxicological data.

Statistical Significance Vs. Toxicological Relevance

When using tests on differences, a general question arises by the discussion about the relevance of statistical significances. The power of a statistical test to find an existing difference between treatment groups and control is the essential criterion for the quality of a test on differences. It depends on the actual difference, the sample size, the variance of the underlying distribution, and the predefined α level. Most toxicological studies are planned based on recommendations of guidelines especially as far as the sample size is concerned. Also, a significance level of 5% is in common use. Therefore the power to find an existing difference may vary for each parameter within a study, as the variance of the underlying distribution could be different. This may lead to situations where even small differences in means between treatment and control could result in significant findings, which may not be relevant from a toxicological point of view. On the contrary, relevant effects might not be significant due to a low test power. As long as tests on differences are in use, this major disadvantage has to be taken into account while interpreting significances. It has to be pointed out that the power of a test cannot be determined a posteriori based on study data because the true difference in treatment and control means and the true variances of the underlying distribution of the data are unknown.

Issue of Heterogeneous Variances

The variance homogeneity assumption for all groups under investigation, which is the prerequisite for parametric statistical methods, does not hold in general. Unfortunately, variances could be affected by treatment (e.g., increasing dose levels could not only affect the mean values but could also lead to increasing variances). Other variance structures would also be conceivable but seem to occur rather exceptionally in toxicology. Some parametric statistical tests have been described in the literature, which are able to adjust for heterogeneous variances. The Welch *t*-test as a pairwise comparison seems to be well approved (a description could be found in a section below). In the field of trend tests, two examples should be mentioned here: the Welch trend test and the robust contrast-based trend test proposed by Meng et al.^[38] Both tests are described in Bailey.^[39] Some further investigations on the behavior of the robust contrast-based trend test performed by the University of Applied Sciences in Darmstadt^[40] showed two major disadvantages that lead us to the conclusion not to use or to recommend it in toxicological studies. First, for given mean values depending on the variance structure of the control and treatment groups, the power of the test could range from about 20% up to 90%, as presented in the original paper. This extreme variation in

the power is hardly acceptable in toxicology because it is an important quality criterion for a test on differences. Second, depending on the number of groups under investigation and on the sample size balance of the groups, the influence of groups is reduced or even eliminated. For example, in the case of three groups with equal sample sizes, only groups 1 and 3 are taken into account for the calculation of the test statistic. Therefore essentially a pairwise comparison is performed and no trend alternative is tested. This varying behavior of the test, depending on the variance structure and even the number of the groups under investigation, is also not desirable.

The Welch trend test is also described in Bailey,^[39] who points out that the amalgamation procedure for the adaptation of group means to bring them in a nondecreasing order used in this test could lead to misleading test results. Therefore this test seems also not suitable for the routine use in toxicology.

For nonparametric methods in general, although the assumption of same distributions, which includes equal variances, should be kept, these methods are robust against violations of variance homogeneity. The reason seems to be the transformation of the original data into ranks, which are obviously only affected by changes in variances between groups if these changes lead to different rank orders of the original values.

In summary, one can say that the issue of variance heterogeneity is well examined and could be taken into account in the case of parametric pairwise comparison procedures and also for nonparametric procedures. Some investigative work still has to be done if parametric trend tests should be established.

STATISTICAL APPROACHES

As already described, a 5% significance level per parameter, sex, and measurement time is recommended for the analysis of toxicological studies and, as long as tests on differences are used, the power of the methods is essential. Therefore it is important for the analysis to increase the power as much as possible in order to decrease the false-negative error rate (i.e., ensuring that existing toxicological effects lead to statistical significances with a high probability, and therefore minimizing the consumer's risk). First of all, several statistical techniques are described in the following sections, which are useful to achieve this aim, given the standard toxicological study designs described above. A discussion of some useful tests on differences follows. It is not intended to present a complete list of possible tests but a useful subset of established methods for toxicology. Other tests could also be used in combination with the presented statistical techniques, if adequate. A general step-down procedure is described; pooling control groups, the use of one-sided tests, parametric vs. nonparametric testing, the use of trend tests vs. pairwise testing,

and baseline corrected values are discussed as possible statistical techniques to increase the power. All of them may be used for the parameters under investigation either in combination with each other or a subset of them as long as the assumptions for an individual parameter are fulfilled and also accepted by the responsible study directors. However, for the correctness of the methods described below, it is required that a suitable randomization procedure has been performed to assign animals to study groups.

Other approaches will not be mentioned here. The decision tree approach especially, often used in former times, is not recommended for toxicology. Bailey^[39] discussed the disadvantages of this method.

Techniques to Improve Power of the Analysis

Multiple Comparison Procedures

Based on the above described study designs, the statistical inference could be viewed as an analysis of variance (ANOVA) problem. Due to the described aim of the studies, many-to-one comparisons of treatment groups with the control within the ANOVA model, either based on pairwise tests or trend tests, is adequate. In study types, where increasing doses of a substance are under investigation, usually the dose-response relationship is monotone, although there are exceptions from this rule. If this assumption was not verified or could not be accepted a priori for a toxicological study, it would not be possible to identify a NOEL as described in the sections above. For general situations, where this assumption does not hold, pairwise testing of the treatment groups vs. the control is recommended with an additional correction of the p values per parameter, sex, and measurement time based on the procedure of Holm^[41] in order to keep to a multiple significance level of 5%. One special situation, which could, for example, occur in the analysis of mutagenicity testing, is the so-called rainbow alternatives. At low doses, there is an increase in the response, but at high doses, there is a decrease. Schmoor and Schumacher^[42] have discussed some adaptive procedures to handle these situations. In a first step, a range is estimated, where the response seems to be increasing. The highest dose groups not belonging to this range are eliminated and then, for example, a trend test could be performed based on the remaining groups.

If a monotone dose-response relationship is confirmed, the following statistical method for identification of a NOEL, at least restricted to the parameter under investigation, is reasonable. It is a step-down procedure for general use, which is very easy to implement and could be used for all pairwise and trend tests described below. It has been established by Hothorn and Lehmacher:^[43]

1. Test for the highest dose group vs. control, or trend test for all groups on a significance level of α .
2. If significance is found, eliminate the highest dose group and go back to step 1 with the remaining

groups. If no dose groups are left, stop the procedure; no NOEL could be identified. If no significance is found, go on with step 3.

3. The highest dose group of the remaining groups is identified as the NOEL. Lower dose groups are expected to show no relevant difference to control.

Hothorn and Lehmacher showed that this procedure is based on the closed testing principle and keeps to a multiple significance level of α . In the case of a monotone dose–response relationship, the power of this procedure is much better than the more general Holm procedure due to the use of an uncorrected α for each individual test.

Trend Test Vs. Pairwise Comparisons

Also based on the assumption of a monotone dose–response relationship, the use of trend tests is justified. In comparison to corresponding pairwise comparisons of each treatment group with the control, trend tests are in general more powerful due to the use of more information, which is given by the groups between the control and the highest dose group. Nevertheless, one has to take into account that limitations are given in the parametric case, if variance homogeneity cannot be assumed as already pointed out. Either parametric trend tests not adjusting for heterogeneity could be used, relying on their robustness, or pairwise comparisons (e.g., Welch *t*-test) accepting the lower power could be used.

Pooling Control Groups

In order to increase the power of a statistical test for some parameters, it is feasible to pool male and female animals. The decision about pooling could either be based on the experience of toxicologists, that no difference between both sexes exists, or based on investigations of historical control data. However, pooling is only adequate for control groups because it cannot be foreseen if a substance under investigation affects both sexes in the same way or not. Therefore a separate analysis should be done for males and females, but the same common control group could be used for both comparisons.

One-Sided Testing vs. Two-Sided Testing

A toxicological effect on a parameter either leads to an increase or a decrease of the values of treatment groups in comparison to a control. In most cases, the direction a priori is not known. Therefore two-sided testing is adequate in general. If a step-down testing procedure is used (see the section “Multiple Comparison Procedures”), the test direction could be determined on the test of the highest dose group. Therefore due to the assumption of a monotone dose–response relationship, one could easily change to one-sided tests based on the predetermined test direction

for all following steps (i.e., for the tests of the lower dose groups). Obviously, the power of these tests is increased.

For some parameters, only either an increase or a decrease of the treatment group mean in comparison to the control is relevant in order to support toxicological questions. In these situations, it is recommended to use one-sided tests instead of the general use of two-sided tests.

Parametric Method Vs. Nonparametric Method

Parametric methods rely on the assumption of an underlying normal distribution. Although the robustness of parametric methods concerning violations of this assumption has been widely accepted, nonparametric methods are available, where the assumption of a normal distribution is not required. Only the assumptions of equal distributions for all groups and variance homogeneity have to be fulfilled—as for a lot of parametric methods also. Therefore from a theoretical point of view, nonparametric methods are a well-proofed alternative and are able to handle situations where the underlying distribution is not known or known to be not normal. Many authors have described the behavior and have recommended the use of these methods in case of uncertainty about the distribution. For example, Hollander and Wolfe^[44] emphasize that in most cases, nonparametric procedures show only a slightly less efficiency than parametric procedures in case of normality. Bailey^[39] even points out that nonparametric methods are more powerful than the equivalent parametric tests when the assumptions are not fulfilled for a parametric analysis.

Also from a practical point of view, nowadays it is adequate to use nonparametric methods. Some time ago, one had to rely on asymptotic tests due to computational issues of the exact distribution. Especially in toxicology, the use of asymptotic tests should be avoided due to the small sample sizes. During the recent decade, powerful computer algorithms have been developed and established in statistical software to enable exact and fast calculations of *p* values for nonparametric methods. Two algorithms are mentioned here because both have been in use by the authors and seem to work with a high efficiency. First, the shift algorithm developed by Streitberg and Röhmel^[45] could be implemented for most types of programming language. In addition, the network algorithm introduced by Metha and Patel^[46] could be found in the software packages StatXact and SAS (version 8.0 and higher). Both software tools provide exact *p* value calculations, for example, for the Wilcoxon and Jonckheere–Terpstra tests (see following sections).

For the presentation of results from nonparametric methods, it is recommended to show the median in addition to the mean, standard deviation, and sample size as descriptive statistics. Especially when only small sample sizes are in use, the mean value could be affected enormously by outlying values while the median is not.

Baseline Correction

Depending on the study type for some parameters, baseline values, taken right before the treatment of the animals starts, are available. In a descriptive way, sometimes these values are used to check if no undesired randomization for the assignment from animals to study groups occurred. For example, if all heavy animals have been assigned randomly to the control group, significant effects in the treatment groups (e.g., a reduced absolute food consumption) may not have occurred due to treatment. Another way to use these baseline values could be the correction of values taken under treatment or during recovery (i.e., calculation of changes to baseline). Under certain circumstances, these differences could lead to a reduction of the variance of the corrected values in comparison to the absolute values. In these cases, the power of the statistical analysis would be improved.

Body weight changes seem to be a well-established example for a baseline correction in toxicology. Obviously, only the change of the body weight is reflected without looking at the initial weight. At least under the assumption that treatment could affect animals independently of the initial body weight, all randomization issues concerning this parameter are eliminated.

For some study types, baseline values are also available for other parameters, such as hematology or clinical chemistry. Because the changes in these values are not in common use, it is not recommended to perform baseline corrections in these cases.

Statistical Tests

The scope of the next sections is the presentation of some univariate parametric and nonparametric trend and pairwise tests, which could be used in the analysis of toxicological studies. It is not intended to describe these tests in detail, but some hints on further literature are given. In addition, SAS procedures (version 8.0 and higher) to perform these tests are mentioned, if available. One assumption has to be kept for all of the following tests: The parameter values under investigation have to be stochastically independent. Therefore only one value per animal is allowed to be included in the test. Multivariate procedures would allow analyzing more than one value per animal (e.g., some time points of one parameter or some parameters in one analysis). Because these procedures are not investigated well enough in the framework of toxicological studies, only univariate procedures will be described.

Student's *t*-Test

The *t*-test is a pairwise test to identify a difference in means between two groups of parameter values. The populations of the two groups have to be normally distributed with equal variances (variance homogeneity) with regard to the parameter under investigation. In the two-sided

test, the null hypothesis “no difference in means” (i.e., the populations of both groups are equal) is tested against the alternative “existing difference in means” (i.e., populations are different with regard to the mean).

The null hypotheses of one-sided tests are constructed as follows: either the mean of group 1 is smaller than or equal to group 2, or group 1 is bigger than or equal to group 2. The alternatives of one-sided tests are formulated correspondingly: either the mean of group 1 is bigger than group 2, or vice versa.

In toxicological studies, because variances have to be equal between control and treatment groups, it would be reasonable to calculate a common variance estimation for all groups, which is a better estimation due to higher sample size, than to calculate variances for each of the two groups under investigation individually. This common variance estimation can be used in all individual tests.

A detailed description of Student's *t*-test was, for example, given by Yamane.^[47] Calculations could be done with the SAS procedure TTEST. A common variance estimation could be performed using the procedure GLM.

Welch *t*-Test

As the *t*-test described above, the Welch *t*-test is also a pairwise test to identify a difference in means between two groups of parameter values. Normal distribution is required, whereas variances are allowed to differ between the groups. In case of same variances in both groups, the Welch *t*-test is less powerful than the *t*-test. The hypotheses to be tested in the one-sided and two-sided cases are the same as described for the *t*-test.

A detailed description of the Welch *t*-test could be found, for example, in Lehmann.^[48] Calculations could be done with the SAS procedure TTEST.

Williams Test

The Williams test is a parametric trend test based on the *t*-test described above. The assumptions of the *t*-test have to be given (normal distribution and variance homogeneity) and, in addition, a monotone dose-response relationship has to be assumed. If the sample means are not in increasing or decreasing order, an amalgamation procedure is performed first, recalculating the means to correct the order. In some rare cases, this procedure leads to the same results for a data set using one-sided tests in both directions. This has to be taken into account while interpreting the test results.

In the two-sided case, the null hypothesis “no difference in means” is tested against the following alternative:

Either monotone-increasing means with the mean of the highest dose group is bigger than the mean of the control, or monotone-decreasing means with the mean of the highest dose group is smaller than mean of the control.

One-sided alternatives are constructed straightforwardly using only the corresponding part of the two-sided alternative.

The Williams test including the amalgamation procedure is well described in Bailey.^[39] Critical values for this test, which are important for implementation, are given in the original publications by Williams.^[49,50]

Linear Contrasts Within an ANOVA

Linear contrasts within an ANOVA can be used for a test on a linear trend. The populations of the groups under investigation have to be normally distributed with equal variances. Only one-sided linear trend alternatives can be tested with linear contrasts. Some useful contrasts for testing other alternatives have been presented by Bailey.^[39] Computation considerations could be found, for example, in Neter et al.^[51] The SAS procedure GLM can be used to calculate tests based on linear contrasts.

Unfortunately, sometimes it is recommended to perform a standard ANOVA first—testing the null hypothesis of no differences between any group under investigation against the alternative of at least one group being different from the others without taking into account the monotone dose–response relationship. Only when a significance is found in this overall test are additional contrasts calculated to identify the NOEL. It has to be taken into account that the overall test is less powerful than a linear contrast for all groups under investigation, because it does not consider the assumption of a monotone dose–response relationship. Instead, a very general alternative is tested. Therefore it is more reasonable to avoid this overall test.

ANOVA with Rank Transformations

An easy method to perform nonparametric tests using the same linear contrasts in an ANOVA model as described above could be implemented by rank-transforming the data first (i.e., sorting data values in increasing order, assigning ranks to all data values, and using these ranks for the following analysis), which can be performed as described above. This procedure has been presented by Conover and Iman.^[52] The rank transformation could easily be done by the SAS procedure RANK in the following analysis as described in “Linear Contrasts Within an ANOVA.”

Wilcoxon Test

The Wilcoxon test is a nonparametric test to perform pairwise comparisons between two groups. The calculation of the test statistic is based on the ranks of the parameter values. The same continuous distributions in both groups are required; differences in the means of the underlying distributions can be identified. Nevertheless, also discrete data can be analyzed if the number of ties is not too big.

The hypotheses in the one-sided and two-sided cases are the same as for the *t*-test.

It is recommended to calculate exact *p*-values (see the section “Parametric Method vs. Nonparametric Method”) when using this test in toxicological studies due to the small sample sizes.

Hollander and Wolfe^[44] described the Wilcoxon test in detail. The exact *p* values could be calculated with the SAS procedure NPAR1WAY.

Jonckheere–Terpstra Test

The Jonckheere–Terpstra test is a nonparametric trend test to identify monotone-increasing or monotone-decreasing trends of the parameter under investigation with regard to the increasing dose levels. The distributional assumptions are the same as for the Wilcoxon test. The hypotheses to be tested in the one-sided and two-sided cases can be found in “Williams Test.”

As with the Wilcoxon test, exact *p*-values should be calculated.

A detailed description of the Jonckheere–Terpstra test could be found, for example, in Hollander and Wolfe.^[44] The SAS procedure FREQ delivers exact *p* values.

Normal Range Comparisons

Normal ranges could be used as an additional statistical instrument to discuss the relevance of test decisions, if tests on differences are used. Any statistically proofed differences between dose group and control might still be interpreted as not relevant if both groups are identified as normal according to the calculated normal ranges. Otherwise, if the dose group is proofed to be abnormal and the control is not, the relevance of the test result seems to be verified. Two major problems should be mentioned when analyzing historical data. First, in some study types, multiple measurements per animal may be taken. Therefore not all values of the historical data pool are stochastically independent. In this case, a randomized selection of one value per animal is recommended. Second, it should be taken into account that age-dependent parameters could exist. This is not relevant for short-term studies, if only animals with similar age had been used. If the age of the animals in the historical data pool varies strongly, one should think about an adequate correction of the values with regard to age.

Only one method to calculate normal ranges, able to classify a complete group of animals as normal or abnormal, should be mentioned here. Passing et al.^[53] described a nonparametric prognose region, which could be used to calculate a normal region containing, on average, 95% of the population. A group of animals of a current study is classified as normal when at least 75% of the values lie within the corresponding range.

The basis of normal range calculations is professional data management, which has to ensure that historical data

are stored and updated correctly over a long period of time, including age, species, strain information, and other available parameters to be taken into account for the calculations.

Two Concepts for Analyzing Chronic or Subchronic Toxicity Studies

Two possible concepts for the routine analysis of hematology, clinical chemistry, and urinalysis parameters of a chronic or subchronic toxicity study will be presented here, both using some of the techniques and statistical tests presented above. Two hypothetical companies are investigated for demonstration. In addition, some recommendations are given for the analysis of body weights, organ weights, and food consumption, valid for both companies. One general rule should always be taken into account: Keep the analysis as simple as possible.

First Concept

No knowledge about the distribution of the parameters is available in the first company, where toxicological studies have never been done before. Therefore it has been decided to use a nonparametric test instead of relying on uncertain assumptions of normally distributed parameters. Because only increasing doses of one substance will be analyzed routinely and the toxicologists agreed on the assumption of a monotone dose–response relationship, a nonparametric Jonckheere–Terpstra trend test is selected for the analysis of all parameters. For special studies, where this assumption seems obviously be not valid, a statistician should be involved. Because the toxicologists in this company are inexperienced, also no knowledge is available about differences between sexes for individual parameters and relevant directions of possible toxic effects could be gathered; the decision is made by analyzing sexes separately and performing two-sided tests. A step-down procedure according to Hothorn and Lehmacher (see the section “Multiple Comparison Procedures”) is adequate, with the detection of the test direction based on all dose groups followed by one-sided test for further steps. Because no historical data are available, normal range calculation is not possible.

Second Concept

In a second company, historical data of control animals have been gathered and well managed and maintained since more than 10 years ago. In addition, toxicologists have decided to go on with the same animal strains that have been used since the beginning of toxicological studies in this company, for the future. Now, in order to define a new statistical concept, the decision is made to analyze the historical data pool. A basic question is the normal distribution assumption for the individual parameters. Even if it is expected that about 80% of the parameters could not be identified as normally distributed (according to Bailey^[39]),

toxicologists are interested in improving the power by using parametric methods for about 20% of the parameters found to be normally distributed. It has been decided to use the Williams test for normally distributed parameters and the Jonckheere–Terpstra test for others.

The toxicologists have agreed on a list of parameters, where no differences between untreated male and female animals are expected, and also a list of parameters, where one toxicologically relevant test direction has been identified. Therefore pooled control groups and one-sided tests as described above are implemented for the corresponding parameters. Normal ranges are calculated as described in the “Normal Range Comparisons” and will be used for additional information because many of the historical data are available.

Analysis of Body Weights, Organ Weights, and Food Consumption

In both companies, toxicologists confirm that body weights, absolute organ weights, and food consumption are normally distributed parameters. They also insist on the analysis of body weight changes (i.e., baseline corrected values). Because it cannot be assumed that variances of control and treatment groups are equal, and no well-working trend tests have been found in the literature, the decision is made to use a pairwise two-sided Welch-*t* test for routine analysis. In some situations, organ weights will have to be analyzed relative to final body weight or to brain weight, which is even better due to a lower intra-animal variation in contrast to body weight. In addition, food consumption should be analyzed relative to corresponding body weights. It is a fact that normally distributed parameters divided by each other are not normally distributed anymore. Therefore nonparametric methods as described in the first concept will be used in these cases.

Comparison of Concepts

The advantage of the first concept is its easy implementation and straightforward interpretation of the test results. The second concept is a more advanced combination of the statistical techniques presented above. Without a doubt, the implementation is expensive and takes some time in planning to discuss details very carefully. Once established, toxicologists have to be aware of the methods in use in order to prevent a misinterpretation of the statistical results. Nevertheless, on the average, some more significances will be found due to a higher power in some parameters.

CONCLUSION

Before starting a toxicological study, a concept for the statistical analysis has to be determined and described in a study plan. On one hand, adequate statistical tests

for toxicological parameters have to be selected; on the other hand, the general statistical techniques described above should be discussed very intensively. As long as assumptions for a statistical test are fulfilled, it is a second-rate question as to which test should be preferred for parameters, if more than one is available. A selection of possible tests has been described in the sections above to provide an overview of well-accepted methods. In order to increase the power of the statistical analysis, it is even more important to apply the above described statistical techniques, if possible. Most of them are independent of the use of a special test. However, they are powerful tools, which are able to improve the quality of a statistical analysis of toxicological parameters, if used adequately. Nevertheless, the concept for the analysis of toxicological studies has to be well discussed and accepted by statisticians and toxicologists.

Although there are still statistical areas where an improvement of the methods is necessary in order to provide a most efficient support of toxicological decisions—especially variance heterogeneity situations and multivariate analysis should be mentioned here—many achievements have already been made. The broad acceptance of nonparametric methods and multiple comparison procedures as the step-down procedure described above is one important example.

Some statistical areas are well discussed in the literature but are still not in common use. Here, it should be emphasized that especially the practical use of equivalence tests needs more investigative work. Not only the way of providing equivalence limits for a huge amount of toxicological parameters is still not well discussed and generally accepted, also the interpretation of the results is different from the common methods in use. It has to be well discussed by toxicologists and statisticians in order to get a common understanding and a broad acceptance—not only by people involved in the conduction of studies but also by the administrative persons responsible.

REFERENCES

1. Finney, D.J. The median lethal dose and its estimation. *Arch. Toxicol.* **1985**, *56*, 215–218.
2. Arcy, P.F.; Harron, D.W.G. International Conference of Harmonisation. *Proceedings of the First International Conference on Harmonization Brussels*, 1991; Queen's University of Belfast: Belfast, 1992, 183–184.
3. U.S. Environmental Protection Agency. *Health Effects Guidelines* OPPTS 870.1100, *Acute Oral Toxicity*. No. 712-C-98-190; EPA: Washington, DC, 1998.
4. Organization for Economic Cooperation and Development. *Guidelines for the Testing of Chemicals: Guideline 420*. In *Acute Oral Toxicity—Fixed Dose Method*, OECD; OECD Publications: 2 Rue André-Pascal, 75775 Paris Cedex 16, France, 1992.
5. Organization for Economic Cooperation and Development. *Guidelines for the Testing of Chemicals: Guideline 423*. In *Acute Oral Toxicity—Acute Toxic Class Method*, OECD; OECD Publications: 2 Rue André-Pascal, 75775 Paris Cedex 16, France, 1996.
6. Organization for Economic Cooperation and Development. *OECD Guidelines for the Testing of Chemicals: Guideline 425*. In *Acute Oral Toxicity—Up and Down Procedure*, OECD; OECD Publications: 2 Rue André-Pascal, 75775 Paris Cedex 16, France, 1998.
7. U.S. Environmental Protection Agency. *Acute Dermal Irritation*. No. 712-C-98-196; EPA *Health Effects Guidelines* OPPTS 870.2500; Washington, DC, 1998.
8. U.S. Environmental Protection Agency. *Acute Eye Irritation*. No. 712-C-98-195; EPA *Health Effects Guidelines* OPPTS 870.2400; Washington, DC, 1998.
9. Magnusson, B.; Kligman, A.M. The identification of contact allergens by animal assay. The guinea pig maximization test. *J. Invest. Dermatol.* **1969**, *52*, 268–276.
10. Buehler, E.V. Occlusive patch method for skin sensitization in guinea pigs: The Buehler method. *Food Chem. Toxicol.* **1994**, *32*, 97–101.
11. U.S. Environmental Protection Agency. *Skin Sensitization*. No. 712-C-98-197; EPA *Health Effects Guidelines* OPPTS 870.2600; Washington, DC, 1998.
12. Kimber, I.; Hilton, J.; Weidenberger, C. The murine local lymph node assay for identification of contact allergens: A preliminary evaluation of in situ measurement of lymphocyte proliferation. *Contact Dermatitis* **1989**, *21*, 215–220.
13. Gad, S.C.; Dunn, B.J.; Dobbs, D.W.; Reilly, C.; Walsh, R.D. Development and validation of an alternative dermal sensitization test: The mouse ear swelling test (MEST). *Toxicol. Appl. Pharmacol.* **1986**, *84*, 93–114.
14. International Conference of Harmonisation. *Safety Pharmacology Studies for Human Pharmaceuticals*, 2000; Adopted by CPMP/ICH/539/00.
15. Irwin, S. Comprehensive observational assessment: Ia. A systematic, quantitative procedure for assessing the behavioural and physiologic state of the mouse. *Psychopharmacologia (Berlin)* **1969**, *13*, 222–257.
16. Mattsson, J.L.; Spencer, P.; Albee, R.R. A performance standard for clinical and functional observation battery examinations of rats. *J. Am. Coll. Toxicol.* **1996**, *15*, 1–239.
17. Committee for Proprietary Medical Products (CPMP). *Note for Guidance on Repeated Dose Toxicity*. CPMP/SWP/1042/99, 2000.
18. U.S. Environmental Protection Agency. *Repeated Dose 28-Day Oral Toxicity Study in Rodents*. No. 712-C-00-366; EPA *Health Effects Guidelines* OPPTS 870.3050; Washington, DC, 2000.
19. U.S. Environmental Protection Agency. *90-Day Oral Toxicity Study in Rodents*. No. 712-C-98-199; EPA *Health Effects Guidelines* OPPTS 870.3100; Washington, DC, 1998.
20. U.S. Environmental Protection Agency. *90-Day Oral Toxicity Study in Nonrodents*. No. 712-C-98-200; EPA *Health Effects Guidelines* OPPTS 870.3150; Washington, DC, 1998.
21. Ames, B.N.; McCann, J.; Yamasaki, E. Methods for detecting carcinogens and mutagens with the *Salmonella*/mammalian-microsome mutagenicity test. *Mutat. Res.* **1975**, *31*, 347–364.
22. Zeiger, E. Identification of rodent carcinogens and noncarcinogens using genetic toxicity tests: Premises, promises,

- and performance. Regul. Toxicol. Pharmacol. **1998**, 28, 85–95.
23. U.S. Environmental Protection Agency. *Bacterial Reverse Mutation Test*. No. 712-C-98-247; EPA *Health Effects Guidelines OPPTS 870.5100*; Washington, DC, 1998.
 24. Li, A.P.; Carver, J.H.; Choy, W.N.; Hsie, A.W.; Gupta, R.S.; Loveday, K.S.; O'Neill, J.P.; Riddle, J.C.; Stankowski, Jr., L.F.; Yang, L.L. A guide for the performance of the Chinese hamster ovary cell/hypoxanthine-guanine phosphoribosyl transferase gene. Mutation assay. Mutat. Res. **1987**, 189, 135–141.
 25. Clive, D.; Johnson, K.O.; Spector, J.F.S.; Batson, A.G.; Brown, M.M. Validation and characterization of the L5178Y/TK + /– mouse lymphoma mutagen assay system. Mutat. Res. **1979**, 59, 61–108.
 26. U.S. Environmental Protection Agency. In Vitro Mammalian Cell Gene Mutation Test. No. 712-C-98-221; EPA *Health Effects Guidelines OPPTS 870.5300*; Washington, DC, 1998.
 27. U.S. Environmental Protection Agency. In Vitro Mammalian Chromosome Aberration Test. No. 712-C-98-223; EPA *Health Effects Guidelines OPPTS 870.5375*; Washington, DC, 1998.
 28. U.S. Environmental Protection Agency. Mammalian Bone Marrow Chromosome Aberration Test. No. 712-C-98-225; EPA *Health Effects Guidelines OPPTS 870.5385*; Washington, DC, 1998.
 29. U.S. Environmental Protection Agency. Mammalian Erythrocyte Micronucleus Test. No. 712-C-98-226; EPA *Health Effects Guidelines OPPTS 870.5395*; Washington, DC, 1998.
 30. U.S. Environmental Protection Agency. Unscheduled DNA Synthesis in Mammalian Cells in Culture. No. 712-C-98-230; EPA *Health Effects Guidelines OPPTS 870.5550*; Washington, DC, 1998.
 31. Schumacher, M.; Schmoor, C. Statistical Analysis of the Ames Assay. In *Lecture Notes in Medical Informatics, Statistical Methods in Toxicology*; Hothorn, L., Ed.; Springer-Verlag: New York, 1991.
 32. Hoffmann, W.P. The In Vitro Ames Test. In *Design and Analysis of Animal Studies in Pharmaceutical Development*; Chow, S.C.; Liu, J.P., Eds.; Marcel Dekker, Inc: New York, 1998.
 33. Liu, J.; Chow, S.C. The In Vitro Chinese Hamster Ovary Cell Mutagenesis Studies. In *Design and Analysis of Animal Studies in Pharmaceutical Development*; Chow, S.C.; Liu, J.P., Eds.; Marcel Dekker, Inc: New York, 1998.
 34. Leimer, I.; Peil, H.; Ellenberger, J. Statistical Analysis of the Micronucleus Test with the Fisher–Pitman Permutation Test. In *Lecture Notes in Medical Informatics, Statistical Methods in Toxicology*; Hothorn, L., Ed.; Springer-Verlag: New York, 1991.
 35. Hothorn, L. General Statistical Principles in Testing of Toxicological Studies. In *Lecture Notes in Medical Informatics, Statistical Methods in Toxicology*; Hothorn, L., Ed.; Springer-Verlag: New York, 1991.
 36. Hauschke, D.; Hothorn, L.A. Safety Assessment in Toxicological Studies: Proof of Safety Versus Proof of Hazard. In *Design and Analysis of Animal Studies in Pharmaceutical Development*; Chow, S.C.; Liu, J.P., Eds.; Marcel Dekker, Inc: New York, 1998.
 37. Wellek, S. *Statistische Methoden zum Nachweis von Äquivalenz*; Gustav Fischer Verlag: Stuttgart, 1994.
 38. Meng, C.; Davis, S.; Roth, A. Robust Contrast Based Trend Tests. *1993 Proceedings of the Biopharmaceutical Section*; American Statistical Association, 1993, 127–132.
 39. Bailey, S. Subchronic Toxicity Studies. In *Design and Analysis of Animal Studies in Pharmaceutical Development*; Chow, S.C.; Liu, J.P., Eds.; Marcel Dekker, Inc: New York, 1998.
 40. Overbeck-Larisch, M.; Sanns, W. Final Report on the MDR Project. *Checking the Suitability of the Meng Davis Roth Test with Regard to Its Use in Toxicology*; University of Applied Sciences, 2000. property of Aventis Pharma.
 41. Holm, S. A simple sequentially rejective multiple test procedure. Scand. J. Statist. **1979**, 6, 65–70.
 42. Schmoor, C.; Schumacher, M. Adaptive statistical procedures for the analysis of nonmonotone dose–response relationships. Biom. Inform. Med. Biol. **1992**, 23 (3), 113–126.
 43. Hothorn, L.; Lehmacher, W. A simple testing procedure “control versus k treatments” for one-sided ordered alternatives, with application in toxicology. Biom. J. **1991**, 33, 179–189.
 44. Hollander, M.; Wolfe, D.A. *Nonparametric Statistical Methods*; Wiley and Sons, Inc.: New York, 1973.
 45. Streitberg, B.; Röhm, J. Exakte Verteilung für Rang- und Randomisierungstests im allgemeinen Stichprobenproblem. In *EDV in Medizin und Biologie*; Verlag Eugen Ulmer GmbH and Co./Gustav Fischer Verlag KG: Stuttgart, 1987, Vol. 18, 12–19.
 46. Metha, C.R.; Patel, N.R. A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. J. Am. Stat. Assoc. **1983**, 78, 427–434.
 47. Yamane, T. *Statistics, An Introductory Analysis*, 3rd Ed.; Harper International Edition; Harper International, 1973.
 48. Lehmann, E.L. *Testing Statistical Hypotheses*, 2nd Ed.; John Wiley and Sons, Inc.: New York, 1986.
 49. Williams, D.A. A test for differences between treatment means when several dose levels are compared with a zero dose control. Biometrics **1971**, 27, 103–117.
 50. Williams, D.A. The comparison of several dose levels with a zero dose control. Biometrics **1972**, 28, 519–531.
 51. Neter, J.; Wasserman, W.; Kutner, M. *Applied Linear Statistical Models*, 2nd Ed.; Richard Irwin: Homewood, IL, 1985.
 52. Conover, W.; Iman, R. Rank transformations as a bridge between parametric and nonparametric statistics. Am. Stat. **1981**, 35, 124–129.
 53. Passing, H.; Bambynek, G.; Deyssenroth, H.; Grieve, A.P.; Helmstädter, G.; Knappen, F.; Lüdin, E.; Peil, H.; Unkelbach, H.D. Normalbereiche im Tierversuch. In *Biometrie in der Chemischpharmazeutischen Industrie*; Vollmar, J., Ed.; G. Fischer Verlag: Stuttgart, 1983, Vol. 1, 57–73.

Traditional Chinese Medicine (TCM): Clinical Development

Fuyu Song

Peking University Clinical Research Institute, Peking University Health Science Center,
Beijing, China

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

In recent years, the use of complementary and alternative medicine including botanical drug products and traditional Chinese (herbal) medicine (TCM) in humans for treating critical and/or life-threatening diseases has received much attention. In pharmaceutical/clinical development of a given TCM, one of the major criticisms is lack of objectively scientific evidence (documents) of clinical safety and efficacy. Unlike the Western medicines (WMs), TCM often consists of multiple components (active ingredients) whose pharmacological activities are often unknown or are not fully characterized or understood. Thus, standard methods for WM clinical trials may not be appropriately applied directly to TCM clinical trials. In this article, some statistical considerations including selection of study design, preparation of matching placebo, development of study endpoint, validation of an instrument, calibration of the validated instrument, and power calculation for sample size estimation are discussed. These considerations have an impact on effectively and scientifically evaluation of clinical safety and efficacy of TCM in clinical trials. In addition, some practical issues regarding test for consistency in raw materials, stability of drug substance, and animal studies are also discussed.

Keywords: Botanical drug products; Calibration; Herbal medicine; Matching placebo; QOL-like instrument.

INTRODUCTION

In recent years, the search for complementary and alternative medicine such as botanical drug products and traditional Chinese (herbal) medicine (TCM) for treating critical and/or life-threatening diseases has received much attention. This has led to the study of the potential use of promising TCMs. As indicated by Chow et al.^[1] and Chow,^[2] TCM originated in ancient China has evolved over several thousands of years, which usually refers to a broad range of Chinese medicine practice including various forms of herbal medicine, acupuncture, massage (*Tui Na*), exercise (*Qi Gong*), and dietary therapy. In pharmaceutical/clinical development of a test treatment, one of the major criticisms for the development of TCM is lack of objectively scientific evidence (documents) of clinical safety and efficacy. Unlike the Western medicines (WMs), TCM often consists of multiple components (active ingredients) whose pharmacological activities are often unknown or are not fully characterized or understood. Thus, standard methods for evaluation of WM clinical trials (see, e.g., the works of Chow et al.^[1–3]) may not be appropriately applied directly to TCM clinical trials.

In TCM clinical development, it is a concern whether a TCM can be scientifically evaluated in the Western way due to some fundamental differences between a WM and a TCM. These fundamental differences include differences in formulation and drug administration, medical theory/practice, diagnostic procedure, and criteria for evaluation.^[1] As an example, the Chinese diagnostic procedure for patients with certain diseases consists of four major techniques: inspection, auscultation and olfaction, interrogation, and pulse taking and palpation (see also the work of Chow^[2]). Under these differences, it is of interest to the investigators regarding how to design and conduct a scientifically valid (i.e., an adequate and well-controlled) clinical trial for the evaluation of the clinical safety and efficacy of the TCM under investigation. In addition, it is also of particular interest to the investigators as to how to translate an observed significant difference detected by the Chinese diagnostic procedure to a clinically meaningful difference based on some well-established clinical study endpoints. The purpose of this article is not only to describe some perspectives regarding TCM development but also to provide some basic considerations regarding practical issues that are commonly encountered during the

conduct of clinical trials in the development of TCMs in the Western way. These statistical considerations include selection of study design, preparation of matching placebo, development of study endpoint, validation of an instrument, calibration of a validated instrument, and power calculation for sample size estimation.

In the next section, some perspectives regarding TCM clinical development are described. The “Basic Considerations” section provides some basic (statistical) considerations for TCM clinical trials. Some practical issues that are commonly encountered during the process of development of a TCM are reviewed in the “Practical Issues” section. Some concluding remarks are given in the last section.

PERSPECTIVES OF TCM CLINICAL DEVELOPMENT

For clinical development of TCMs, before a TCM clinical trial is conducted, the following questions are necessarily asked.

- 1. Will the TCM clinical trial be conducted by Chinese doctors alone, Western clinicians alone, Western clinicians who have some background and experience of Chinese herbal medicine alone, or both Chinese doctors and Western clinicians?
- 2. Will traditional Chinese diagnostic procedure and/or trial procedures be used throughout the TCM clinical trial?
- 3. Upon approval, is the TCM intended for use by Chinese doctors, Western clinicians, or both?

For the first two questions, if the TCM clinical trial is to be conducted by Chinese doctors alone, the following questions arise: first, should the Chinese diagnostic procedure be validated in terms of its sensitivity, specificity, and consistency (reproducibility) in order to provide an accurate and reliable assessment of the TCM? In addition, it is of interest to determine how an observed difference obtained from the Chinese diagnostic procedure can be translated to the clinical endpoint commonly used in similar WM clinical trials with the same indication. These two questions can be addressed statistically by the calibration and validation of the Chinese diagnostic procedure with respect to some well-established clinical endpoints for evaluation of WMs. If the TCM clinical trial is to be conducted by Western clinicians or Western clinicians who have some background of Chinese herbal medicine, the standards and consistency of clinical results compared to those WM clinical trials are ensured. However, the good characteristics of TCM may be lost during the process of the conduct of the TCM clinical trials. On the other hand, if the TCM clinical trial is to be conducted by both Chinese doctors and Western clinicians, differences in medical practice

and/or possible disagreement regarding the diagnosis, treatment, and evaluation are the major concerns.

For the third question, if the TCM is intended for use of Chinese doctors but it is conducted by Western clinicians, difference in perception regarding how to prescribe the TCM is of great concern. The preparation of a package insert based on the clinical data could be a major issue, not only to the sponsor but also to regulatory authorities. Similar comments apply to the situation where the TCM is intended for use of Western clinicians, but the trial is conducted by Chinese doctors.

As a result, it is suggested that the intention of use (i.e., labeling for the indication) be clearly evaluated when planning a TCM clinical trial. In other words, the sponsor needs to determine whether the TCM is intended for use of Western clinician only, Chinese doctors only, or both Western clinicians and Chinese doctors at the planning stage of a TCM clinical trial, for an adequate package insert of the target diseases under study.

BASIC CONSIDERATIONS

To effectively and scientifically evaluate the clinical safety and efficacy of a TCM under investigation, as indicated by Chow and Song^[4], a valid study design with some basic considerations for an intended clinical trial is necessarily considered for achieving study objectives with a desired power under the perspectives described in the previous section (see also the FDA guideline^[5]).

Selection of Study Design

Let WW and CW denote conducting a clinical trial the Western way (i.e., based on Western well-established clinical endpoints) and the Chinese way (i.e., based on traditional Chinese diagnostic procedures), respectively. Depending upon the study endpoints (or diagnostic procedures), clinical trials can be classified into the following four different types (see Table 1):

Type I clinical trials, denoted by WM₁₁ are classic clinical trials for WMs that utilize well-established Western clinical endpoints before and after treatment.

Type II (TCM₁₂) clinical trials usually refer to those TCM clinical trials utilizing well-established Western

Table 1 Types of WM and TCM Clinical Trials

Pretreatment (screening)	Posttreatment	
	1 (WW)	2 (CW)
1 (WW)	Type I = WM ₁₁	Type II = TCM ₁₂
2 (CW)	Type III = WM ₂₁	Type IV = TCM ₂₂

WW, Western way; CW, Chinese way; WM₁₁, utilizing WW in both pre- and post-treatment; WM₂₁, utilizing CW in the pre-treatment and WW in the post-treatment.

clinical endpoints at screening (pretreatment) and evaluating the test treatment the Chinese way (i.e., based on Chinese diagnostic/evaluating procedures) posttreatment.

Type III (WM₂₁) clinical trials usually refer to those WM clinical trials utilizing Chinese four diagnostic procedures at screening (pretreatment) and evaluating the test treatment the Western way posttreatment.

Type IV clinical trials, denoted by TCM₂₂, are typical TCM clinical trials adopting Chinese diagnostic and evaluating procedures pre- and posttreatment.

As can be seen from Table 1, for evaluation of a TCM under investigation, one may consider conducting a clinical trial that can be either type II (TCM₁₂) or type IV (TCM₂₂). In practice, however, WM type of clinical trials (type I and type III) and TCM type of clinical trials (type II and type IV) are very likely arriving at different conclusions on the test treatment under investigation due to some fundamental differences between WMs and TCMs. In the interest of testing for consistency between the Western way and the Chinese way for the evaluation of the TCM under investigation, the following study designs are useful:

Design A—Subjects will first be screened using the well-established Western study endpoints. Qualified subjects will then be randomly assigned to receive either a test treatment (T) or a control (C). In each treatment group, subjects will further be randomly split into two subgroups: one subgroup will be evaluated the Western way (WW) and the other subgroup will be evaluated the Chinese way (CW). Design A is illustrated in Fig. 1. Under this study design, not only the treatment effect can be evaluated by means of either the Western way or the Chinese way but also the consistency between WW and CW can be evaluated.

Design B—Subjects will first be screened by using the Chinese diagnostic procedure. Qualified subjects will then be randomly assigned to receive either a test treatment (T) or a control (C). In each treatment group, subjects will further be randomly split into

two subgroups: one subgroup will be evaluated the Western way (WW) and the other subgroup will be evaluated the Chinese way (CW). Design B is illustrated in Fig. 2. Under this study design, similarly, not only the treatment effect can be evaluated by means of either the Western way or the Chinese way, but also the consistency between WW and CW can be evaluated.

Design C—This design is a combination of Design A and Design B. Subjects will be first randomly split into two subgroups. Subjects in one subgroup will be screened using the well-established Western study endpoints, whereas subjects in the other subgroup will be screened by means of the Chinese diagnostic procedure. In each subgroup, qualified subjects will then be randomly assigned to receive either a test treatment (T) or a control (C). In each treatment group, subjects are further randomly split into two subgroups: one subgroup will be evaluated the Western way (WW) and the other subgroup will be evaluated the Chinese way (CW). Design C is illustrated in Fig. 3. Under this study design, the treatment effect in each subgroup can be evaluated by means of either the Western way or the Chinese way. In addition, the consistency between WW and CW across subgroups can be evaluated.

Alternative design—In practice, it may not be ethical if the disease under study is critical and/or life-threatening provided that a WM is available. Thus, alternatively, in 2009, Hsiao et al. proposed a randomized, placebo-control, crossover clinical trial or a parallel-group design consisting of three arms (i.e., the TCM under study, a WM as an active control, and a placebo).^[6] The three-arm, parallel-group design allows the establishment of non-inferiority/equivalence of the TCM compared to the active control (WM) and the demonstration of the superiority of the TCM with respect to the placebo. One of the advantages of a crossover clinical trial is that a comparison within each individual can be made, although it will take a longer time to complete the study. Although a crossover design requires a smaller sample size compared to a parallel-group design, there are some limitations for the use of crossover design. First, baselines

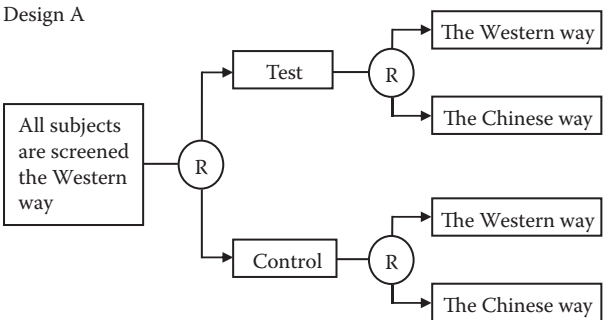


Fig. 1 Design for WM clinical trials comparing the Western way and the Chinese way

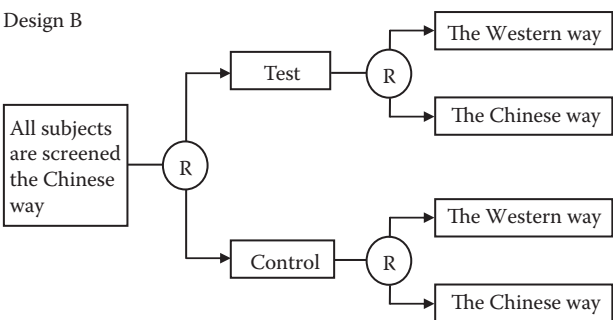


Fig. 2 Design for TCM clinical trials comparing the Western way and the Chinese way

Design C

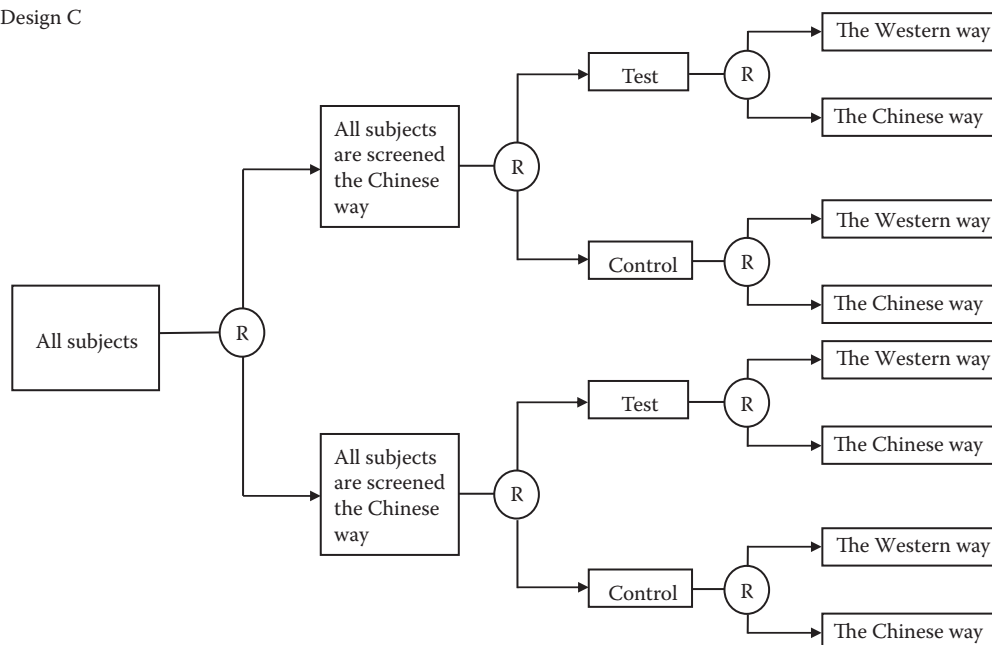


Fig. 3 Design for combined WM and TCM clinical trials comparing the Western way and the Chinese way

prior to dosing may not be the same. Second, when a significant sequence effect is observed, we would not be able to isolate the effects of period effect, carry-over effect, and subject-by-treatment effect that are confounded to one another.

Preparation of Matching Placebo

In clinical development, double-blind, placebo-control, randomized clinical trials are often conducted for the evaluation of the safety and effectiveness of a test treatment under investigation. To maintain blindness, a matching placebo should be identical to the active drug in all aspects of size, color, coating, taste, texture, shape, and order except that it contains no active ingredient. In clinical trials, as an advanced technique available for formulation, a matching placebo is not difficult to make because most WMs contain a single active ingredient. Unlike WMs, TCMs usually consist of a number of components, which often have different tastes. In TCM clinical trials, the TCM under investigation is often encapsulated. However, the test treatment will be easily unblinded if either the patient or the Chinese doctor breaks the capsule. As a result, the preparation of a matching placebo in TCM clinical trials plays an important role for the success of the TCM clinical trials.

In 2011, Fai et al. pointed out that in many TCM clinical trials, it is very difficult to make a quality matching placebo to achieve the purpose of blinding.^[7] Ideally, the characteristics of the test drug and matching placebo should be identical in color, appearance, smell, and taste. The quality matching placebo should be identical to the test drug in physical form, sensory perception, packaging,

and labeling, and it should have no pharmaceutical activity. For this purpose, Fai et al. developed a placebo capsule to match an herbal medicine in terms of its physical form, chemical nature, appearance, packaging, and labeling. Based on the assessment results, the developed placebo capsule assessment results suggested that the placebo was found satisfactory in these aspects.^[7] Thus, Fai et al. concluded that a matching placebo could be created for a randomized controlled trial involving herbal medicine.^[7] In addition, Fai et al. also discussed the means to acquire patent for a developed matching placebo.^[7]

It should be noted that the preparation of matching placebo is extremely important to maintain the integrity of blinding to avoid any possible operational biases that may be introduced due to the knowledge of the treatment assignment. The oral dosage form of capsule is often considered for preparation of matching placebo for clinical trials involving Chinese herbal medicines as it may remove the strong smell and taste of the herbal medicines. However, one of the major challenges is that patients or clinicians will reveal the treatment assignments if they break the capsules. Thus, standard operating procedures (SOPs) for preventing patients and clinicians from breaking the capsules are necessarily developed.

Development of Clinical Endpoint

Unlike WMs, the primary study endpoints for the assessment of safety and effectiveness of a TCM are usually assessed subjectively by a quantitative instrument by experienced Chinese doctors. Although the quantitative instrument is developed by the community of Chinese

doctors and is considered a gold standard for the assessment of safety and effectiveness of the TCM under investigation, it may not be accepted by the Western clinicians due to fundamental differences in medical theory, perception, and practice. In practice, it is very difficult for a Western clinician to conceptually understand the clinical meaning of the difference detected by the subjective Chinese quantitative instrument. Consequently, whether the subjective quantitative instrument can accurately and reliably assess the safety and effectiveness of the TCM is always a concern to Western clinicians.

As an example, for the assessment of safety and efficacy of a drug product for the treatment of ischemic stroke, a commonly considered primary clinical endpoint is the functional status assessed by the so-called Barthel index. The Barthel index is an ordinal scale used to measure the performance of activities of daily living, which was introduced by Mahoney and Barthel.^[8] The Barthel index is a weighted functional assessment scoring technique composed of 10 items with a minimum total score of 0 (functional incompetence) and a maximum total score of 100 (functional competence). The Barthel index is a weighted scale measuring performance in self-care and mobility, which is widely accepted in ischemic stroke clinical trials. A patient may be considered a responder if his/her Barthel index is ≥ 60 . On the other hand, Chinese doctors usually consider a quantitative instrument developed by the Chinese medical community as the standard diagnostic procedure for assessment of ischemic stroke. The standard quantitative instrument is composed of six domains, which capture different information regarding patient's performance, functional activities, and signs and symptoms and status of the disease.

In practice, it is of interest to both Western clinicians and Chinese doctors how an observed clinically meaningful difference by the Chinese quantitative instrument can be translated to that of the primary study endpoint assessed by the Barthel index. To reduce the fundamental differences in medical theory/perception and practice, it is suggested that the subjective Chinese quantitative instrument be calibrated and validated with respect to that of the clinical endpoint assessed by the Barthel index before it can be used in TCM ischemic stroke clinical trials.

Validation of a Quality-of-Life-Like Instrument

In TCM medical practice, a Chinese doctor usually collects information from the patient with a certain disease through the four subjective diagnostic procedures as described earlier. The purpose of these subjective approaches is to collect information on various aspects of the disease under study such as signs, symptoms, patient's performance, and functional activities. In practice, a quality-of-life-like (QOL-like) instrument with a number of questions/items is often considered to quantitatively assess the treatment effect. These items are usually grouped to form subscales,

composite scores (domains), or an overall score for a simple and easy interpretation of the treatment effect. Since the items (subscales) in each subscale (composite score) are correlated, the structure of responses to a quantitative instrument is usually multidimensional, complex, and correlated.

In 1954, Guilford discussed several methods such as Cronbach's α for measuring the reliability of internal consistency of a quantitative instrument.^[9] In 1989, Guyatt et al. indicated that a quantitative instrument should be validated in terms of its validity, reproducibility, and responsiveness.^[10] In 1991, Hollenberg et al. discussed several methods for validation of a quantitative instrument, such as consensual validation, construct validation, and criterion-related validation.^[11] There is, however, no gold standard as to how a quantitative instrument should be validated. As indicated in the work of Chow and Liu^[3] in 2013, the validity of a quantitative instrument is the extent to which the instrument measures what is designed to measure. It is a measure of biasedness of the instrument.^[3] The biasedness of a quantitative instrument reflects the accuracy of the instrument. The reliability of a quantitative instrument measures the variability of the instrument, which directly relates to the precision of the instrument. On the other hand, the responsiveness of a quantitative instrument is usually referred to as the ability of the instrument to detect a difference of clinical significance within a treatment.

In 2009, Hsiao et al. considered a specific design for calibration/validation of the Chinese diagnostic procedure.^[6] In their proposed study design, qualified subjects are randomly assigned to receive either a TCM or a WM. Each patient will be evaluated by a Chinese doctor and a Western clinician independently, regardless of which treatment group he/she is in. As a result, there are four groups of data: (i) patients who receive TCM and are evaluated by a Chinese doctor, (ii) patients who receive TCM but are evaluated by a Western clinician, (iii) patients who receive WM but are evaluated by a Chinese doctor, and (iv) patients who receive WM and are evaluated by a Western clinician. Groups (iii) and (iv) are used to establish a standard curve for calibration between the TCM and the WM. Groups (i) and (ii) are then used to validate the Chinese diagnostic procedure based on the established standard curve.

Calibration of a Validated Instrument

Unlike WMs, the primary study endpoints for the assessment of safety and effectiveness of a TCM are usually assessed by a quantitative instrument or the four diagnostic procedures by experienced Chinese doctors. The assessment by a quantitative instrument has been criticized in many ways. First, it may not capture the true health status of the patients with diseases under study (e.g., by asking wrong questions). Second, it may not detect the effect of the test treatment under investigation. As an example,

consider a quantitative instrument with possible scores from 0 (perfect health) to 100 (worst possible disease status). Suppose the scores can be classified into the following categories of health status: *Health* (0–25), *Mild* (26–50), *Moderate* (51–75), and *Severe* (76–100). In this case, there is a significant difference between a patient with a score of 25 (*Health*) and a patient with a score of 26 (*Mild*) although they only differ by one point. On the other hand, a patient with a score of 26 and a patient with a score of 50 are both considered to have *Mild* disease status although they differ by 24 points. Thus, the assessment based on a quantitative instrument by experienced Chinese doctors is not only subjective but also lacks validity. Consequently, the reliability of the assessment is a concern, especially when there is evidence of large rater-to-rater variability.

Thus, although the quantitative instrument is developed by the community of Chinese doctors and is considered a gold standard for the assessment of safety and effectiveness of the TCM under investigation, it may not be accepted by the Western clinicians due to not only the lack of validity and reliability but also the interpretation of the assessment (or translation of assessment to well-established and widely accepted clinical endpoints). In practice, it is very difficult for a Western clinician to conceptually understand the clinical meaning of the difference detected by the subjective Chinese quantitative instrument due to fundamental differences in medical theory, perception, and practice.

Thus, for modernization or Westernization of TCMs, whether the subjective quantitative instrument can accurately and reliably assess the safety and effectiveness of the TCM is a concern for development of TCM. In practice, it is then suggested that a clinical trial be conducted to calibrate the subjective quantitative assessment against either life events or well-established clinical endpoints that are commonly used in the assessment of WMs. The clinical trials should consist of two arms: one arm will include subjects with diseases under study diagnosed by the subjective quantitative instrument and the other arm will include subjects diagnosed by Western diagnostic or testing procedures. Each subject posttreatment will be assessed by both Chinese doctors using the quantitative instrument and Western clinicians based on the well-established and widely accepted study endpoints.^[6]

Power Calculation for Sample Size Estimation

In clinical trials, sample size is usually selected to achieve a desired power for detecting a clinically meaningful difference in one of the primary study endpoints for the intended indication of the treatment under investigation.^[12] As a result, sample size calculation depends upon the primary study endpoint and the clinically meaningful difference that one would like to detect. Different primary study endpoints may result in very different sample sizes.

For illustration purposes, consider the example concerning a TCM for treatment of ischemic stroke, which

was developed with more than 30 years of clinical experience with humans. Suppose a sponsor would like to conduct a clinical trial to scientifically evaluate the safety and efficacy of the TCM the Western way compared to an active control (e.g., aspirin). Thus, the intended clinical trial is a double-blind, parallel-group, placebo-control, randomized trial. The primary clinical endpoint is the response rate (a patient is considered a responder if his/her Barthel index is ≥ 60) based on the functional status assessed by the Barthel index. Sample size calculation is performed based on the response rate after 4 weeks of treatment under the hypotheses of testing for superiority. As a result, a sample size of 150 patients per treatment group is required for achieving an 80% power for establishment of superiority of the TCM over the active control agent. Alternatively, we may consider the quantitative instrument developed by experienced Chinese doctors as the primary study endpoint for sample size calculation. Based on a pilot study, about 80% (79 out of 122) of ischemic stroke patients were diagnosed by one domain of the quantitative instrument. A patient is considered a responder if his/her domain score is ≥ 7 . Based on this primary study endpoint, a sample size of 90 per treatment group is required to achieve an 80% power for establishment of superiority.

The difference in sample size leads to the question of whether the use of the primary endpoint of response rate based on one domain of the Chinese quantitative instrument could provide substantial evidence of safety and effectiveness of the TCM under investigation.

PRACTICAL ISSUES

Before the test treatment under investigation can be used in humans, some practical issues regarding sufficient information of chemistry, manufacturing, and control (CMC); clinical pharmacology; and toxicology are necessarily considered.^[13] However, since most TCMs consist of multiple components with unknown pharmacological activities, information regarding CMC, clinical pharmacology, and toxicology is often difficult to obtain. In what follows, these difficulties are briefly described.

Test for Consistency

Unlike most WMs, TCMs usually consist of a number of components, which are extracted from herbal samples. The herbal samples are normally dried at 60°C to a constant weight, followed by grinding in a mortar and storing in a desiccator. For water-soluble substances, an appropriate amount of water is first added to the dried material and boiled for about 1 h. For alcohol-soluble substances, 60% ethanol is added and the mixture is extracted at 60°C for 1 h in an ultrasonic bath. After cooling to room temperature, the extract can be cleared by filtration through net or centrifugation at 12,000g for 10 min at 20°C, and the

supernatant is used for further applications. The pharmacological activities, interactions, and relative proportions of these components are usually unknown. In practice, TCM is usually prescribed subjectively by an experienced Chinese doctor. As a result, the actual dose received by each individual varies depending upon the signs and symptoms as perceived by the Chinese doctor. Although the purpose of this medical practice is to reduce the within-subject (or intra-subject) variability, it could also introduce nonnegligible variability such as variations from a component to a component and from a rater to a rater (a Chinese doctor to another). Consequently, reproducibility or consistency of clinical results is questionable. Thus, how to ensure the reproducibility or consistency of the observed clinical results has become a great concern to regulatory agencies in the review and approval process. It is also a great concern to the sponsor of the manufacturing process. To address the question of reproducibility or consistency, a valid statistical quality control process on the raw materials and final product is essential (see, e.g., the works of Tse^[14] and Lu et al.^[15]).

Stability Analysis

Most regulatory agencies require that the expiration dating period (or shelf life) of a drug product be indicated in the immediate container label before it can be released for use. To fulfill this requirement, stability studies are usually conducted in order to characterize the degradation of the drug product. For drug products with a single active ingredient, statistical methods for determination of drug shelf life are well established (see, e.g., the guidelines of FDA^[16] and ICH^[17]). However, for drug products with multiple components, regulatory requirements for estimation of drug shelf life are not available.

Following the concept of estimating the shelf life for drug products with a single active ingredient, two approaches are worth considering. First, we may (conservatively) consider the *minimum* of the shelf lives obtained from each component of the drug product. This approach is conservative, and yet may not be feasible due to the fact that (i) not all of the components of a TCM can be accurately and reliably quantitated, and (ii) the resultant shelf life may be too short to be useful. Alternatively, we may consider a two-stage approach for determination of drug shelf life. At the first stage, an attempt should be made to identify the most active component(s) whenever possible. A shelf life can then be obtained based on the method suggested in the FDA and ICH guidelines.^[16,17] At the second stage, the obtained shelf life is adjusted based on the relationship and/or interactions of the most active ingredient(s) and other components. As an alternative, Chow and Shao in 2007 proposed a statistical method for determining the shelf life of a TCM following a similar idea suggested by the FDA, assuming that the components are linear combinations of some factors.^[18]

Animal Studies

The purpose of animal studies is not only to study the possible toxicity in animals but also to suggest an appropriate dose for use in humans, assuming that the established animal model is predictive of the human model. For a newly developed drug product, animal studies are necessary. However, for some well-known TCMs, which have been used in humans for years and have a very mild toxicity profile, it is questionable whether animal studies are necessary. It is suggested that all components of TCMs as described in Chinese Pharmacopoeia (CP) be classified into several categories depending upon their potential toxicities and/or safety profiles as a basis for regulatory requirements for animal studies. In other words, for some well-known TCM components such as Ginseng, animal studies for testing toxicity may be waived depending upon the past experience of human use, although health risks or side effects following the proper administration of designated therapeutic dosages were not recorded in human use. Note that the German regulatory authority's herbal watchdog agency, commonly called Commission E, has conducted an intensive assessment of the peer-reviewed literature on some 300 common botanicals with respect to the quality of the clinical evidence and the uses for which the herb can be reasonably considered effective.^[19]

CONCLUDING REMARKS

Although TCM has a long history of being used in humans, little or no scientifically valid documentations are available for the demonstration of clinical safety and efficacy of the TCM. In the interest of modernization or Westernization of TCM development, as indicated by the FDA, substantial evidence regarding safety and effectiveness of the test treatment under investigation can only be obtained by conducting adequate and well-controlled clinical trials.^[5] In TCM clinical trials, however, it is a concern whether a TCM can be scientifically evaluated the Western way due to some fundamental differences between a WM and a TCM.

The purpose of this article is to provide some basic considerations for providing substantial evidence of clinical safety and efficacy of a TCM under investigation during the conduct of TCM clinical trials in the development of TCMs the Western way. These statistical considerations include selection of study design, preparation of matching placebo, development of study endpoint, validation of an instrument, calibration of a validated instrument, and power calculation for sample size estimation.

In addition, as discussed earlier, before a TCM under investigation can be used in humans, sufficient information regarding CMC, clinical pharmacology, and toxicology is necessarily provided. In practice, such information, which has the impact on the scientific validity for

the assessment of the TCM under investigation, is difficult, if not impossible, to obtain. Thus, it is suggested that some practical issues such as test for consistency, stability analysis for shelf-life estimation, and animal studies for toxicity be evaluated before the conduct of the intended TCM clinical trials.

REFERENCES

1. Chow, S.C.; Pong, A.; Chang, Y.W. On traditional Chinese medicine clinical trials. *Drug Inform. J.* **2006**, *40*, 395–406.
2. Chow, S.C. *Quantitative Methods for Traditional Chinese Medicine Development*; Chapman and Hall/CRC Press, Taylor & Francis: New York, 2015.
3. Chow, S.C.; Liu, J.P. *Design and Analysis of Clinical Trials—Revised and Expanded*, 3rd Ed.; John Wiley & Sons: New York, 2013.
4. Chow, S.C.; Song, F.Y. Statistical considerations for traditional Chinese medicine clinical trials. *J. Case Stud.* **2015**, *4* (5), 134–142.
5. FDA. *Guideline for the Format and Content of the Clinical and Statistical Sections of New Drug Applications*, The United States Food and Drug Administration, Rockville, MD, 1988.
6. Hsiao, C.F.; Tsou, H.H.; Pong, A.; Liu, J.P.; Lin, C.H.; Chang, Y.J.; Chow, S.C. Statistical validation of traditional Chinese diagnostic procedure. *Drug Inform. J.* **2009**, *43*, 83–95.
7. Fai, C.K.; Qi, G.D.; Wei, D.A.; Chung, L.P. Placebo preparation for the proper clinical trial of herbal medicine—requirements, verification and quality control. *Recent Pat. Inflamm. Allergy Drug Discov.* **2011**, *5*, 169–174.
8. Mahoney, F.; Barthel, D. Functional evaluation: The Barthel Index. *Md. Med. J.* **1965**, *14*, 61–65.
9. Guilford, J.P. *Psychometric Methods*, 2nd Ed.; McGraw-Hill: New York, 1954.
10. Guyatt, G.H.; Van Zanten, S.V.; Feeny, D.H.; Patric, D.L. Measuring quality of life in clinical trials: A taxonomy and review. *Can. Med. Assoc. J.* **1989**, *140*, 1441–1448.
11. Hollenberg; N.K., Testa, M., Williams, G.H. Quality of life as a therapeutic endpoint—An analysis of therapeutic trials in hypertension. *Drug Saf.* **1991**, *6*, 83–93.
12. Chow, S.C.; Shao, J.; Wang, H. *Sample Size Calculation in Clinical Research*. 2nd Ed., Taylor & Francis: New York, 2007.
13. FDA *Guidance for Industry—Botanical Drug Products*. The United States Food and Drug Administration, Rockville, MD, 2004.
14. Tse, S.K.; Chang, J.Y.; Su, W.L.; Chow, S.C.; Hsiung, C.; Lu, Q. Statistical quality control process for traditional Chinese medicine. *J. Biopharm. Stat.* **2006**, *16*, 861–874.
15. Lu, Q.; Chow, S.C.; Tse, S.K. Statistical quality control process for traditional Chinese medicine with multiple correlative components. *J. Biopharm. Stat.* **2007**, *17*, 791–808.
16. FDA. *Guideline for Submitting Documentation for the Stability of Human Drugs and Biologics*. Center for Drugs and Biologics, Office of Drug Research and Review, Food and Drug Administration, Rockville, MD, 1987.
17. ICH. *Stability Testing of New Drug Substances and Products. Tripartite International Conference on Harmonization Guideline Q1A*. Geneva, Switzerland, 1993.
18. Chow, S.C.; Shao, J. Stability analysis for drugs with multiple ingredients. *Stat. Med.* **2007**, *26*, 1512–1517.
19. PDR. *Physicians' Desk Reference for Herbal Medicines*. Medical Economics Company, Montvale, NJ, 1998.

Traditional Chinese Medicine (TCM): General Considerations

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Annpey Pong

Merck Research Laboratories, Summit, New Jersey, U.S.A.

Yu-Wei Chang

Temple University, Philadelphia, Pennsylvania, U.S.A.

Abstract

The purpose of this entry is not only to introduce Consortium for Globalization of Chinese medicine, but also to provide some basic considerations regarding practical issues that are commonly encountered when conducting a traditional Chinese medicine clinical trial the Western way.

Keywords: Calibration; Flexible dose; Global balance/harmony; Multiple components; Test for consistency; Validation.

INTRODUCTION

In recent years, as more and more innovator drug products are going off patent, the search for new medicines that treat critical and/or life-threatening diseases has become the center of attention of many pharmaceutical companies. This leads to the study of the potential use of promising traditional Chinese medicines (TCM), especially for those intended for treating critical and/or life-threatening diseases. A TCM is defined as a Chinese herbal medicine developed for treating patients with certain diseases as diagnosed by the four Chinese major techniques of inspection, auscultation and olfaction, interrogation, and pulse taking and palpation based on traditional Chinese medical theory of global dynamic balance among the functions/activities of all organs of the body. Unlike evidence-based clinical research and development of a Western medicine (WM), clinical research and development of a TCM is usually experience-based with anticipated variability due to subjective evaluation of the disease under study. The use of TCM in humans for treating various diseases has a history of more than five thousand years, although no scientific documentation is available regarding clinical evidence of safety and efficacy of these TCMs.

In the past several decades, regulatory agencies of both China and Taiwan have debated which direction the TCM should take—Westernization or modernization. The Westernization of TCM is referred to the adoption of the typical (Western) process of pharmaceutical research and development for scientific evaluation of the safety and effectiveness of the TCM products under investigation,

whereas the modernization of TCM is to evaluate the safety and effectiveness of TCM the Chinese way (i.e., different sets of regulatory requirements and evaluation criteria) scientifically. Although both China and Taiwan do attempt to build up an environment for the modernization of TCM, they seem to adopt the Westernization approach. In pursuit of advancing the TCM through the modernization way or the Westernization approach to benefit patients who suffer from critical and/or life-threatening diseases, a Consortium for Globalization of Chinese medicine (CGCM) was formed in 2003.

In practice, it is a concern whether a TCM can be scientifically evaluated the Western way due to some fundamental differences between a WM and a TCM. These include differences in formulation, medical practice, drug administration, diagnostic procedure and criteria for evaluation, and flexibility. Based on these differences, it is then of interest to researchers explore how to conduct a scientifically valid (i.e., an adequate and well-controlled) clinical trial for evaluation of the clinical safety and efficacy of the TCM under investigation. In addition, it is also of particular interest to the investigators how to translate an observed significant difference detected by the Chinese diagnostic procedure to a clinically meaningful difference based on some well-established clinical study endpoint. The purpose of this entry is not only to introduce CGCM, but also to provide some basic considerations regarding practical issues that are commonly encountered when conducting a TCM clinical trial the Western way.

The next section describes some fundamental differences between WM and TCM that have an impact on the Westernization of TCM. These fundamental differences

Table 1 Fundamental Differences between a WM and a TCM

Description	Western medicine (WM)	Traditional Chinese medicine (TCM)
Active ingredient	Single	Multiple
Dose	Fixed	Flexible
Diagnostic procedure	Objective; validated	Subjective; not validated
Therapeutic index	Well established	Not well established
Medical mechanism	Specific organs	Global dynamic balance/harmony among organs
Medical perception	Evidence-based	Experience-based
Statistics	Population	Individual

include the concept of global dynamic balance/harmony among organs of the body (TCM) versus local site action (WM); subjective diagnostic techniques of inspection, auscultation and olfaction, interrogation, pulse taking and palpation (TCM) versus objectively clinical evaluation (WM); and personalized flexible dose with multiple components (TCM) versus fixed dose of single active ingredient (WM). The section “Basic Considerations of TCM Clinical Trials” provides some basic considerations of TCM clinical trials. These basic considerations include study design, the validation of a quantitative instrument developed for the four major TCM diagnostic techniques, the use/preparation of matching placebo, and sample size calculation. Some practical issues commonly encountered when conducting a TCM clinical trial are discussed in the section “Other Issues in TCM Research and Development,” such as test for consistency in statistical quality control of raw material and/or final product, stability analysis, animal studies, regulatory requirements, and indication, label and packaging insert. The section “Consortium for Globalization of Chinese Medicine” introduces the organization of the Consortium for Globalization of Chinese Medicine and its recent development. Some concluding remarks, including future strategy and recommendations in TCM research and development, are given in the last section.

FUNDAMENTAL DIFFERENCES

As indicated earlier, the process for pharmaceutical research and development of Western medicines (WM) is well established, and yet it is a lengthy and costly process. This lengthy and costly process is necessary to ensure the efficacy, safety, quality, stability, and reproducibility of the drug product under investigation. For pharmaceutical research and development of a TCM, one may consider directly applying this well-established process to the TCM under investigation. However, this process may not be feasible due to some fundamental differences between a TCM and a WM, which are summarized in Table 1.

These fundamental differences are briefly described below.

Medical Theory/Mechanism and Practice

Medical Theory and Mechanism

TCM is a 3,000-year-old holistic medical system encircling the entire scope of human experience. It combines the use of Chinese herbal medicines, acupuncture, massage, and therapeutic exercise such as Qigong (the practice of internal *air*) and Taigie for both treatment and prevention of disease. With its unique theories of etiology, diagnostic systems, and abundant historical literature, TCM itself consists of Chinese culture and philosophy, clinical practice experience, and the use of many medical herbs.

Chinese doctors believe TCM functions in the body based on the eight principles (Table 2), five-element theory (Table 3), five *Zang* and six *Fu* (Table 4), and knowledge about channels and collaterals. The eight principles consist of *Yin* and *Yang* (i.e., negative and positive) (see also, Table 5), cold and hot, external and internal, and *Shi* and *Xu* (i.e., weak and strong). These eight principles help Chinese doctors to differentiate syndrome patterns. For instance, people with *Yin* will develop disease in a negative, passive, and cool way (e.g., diarrhea and back pain), whereas people with *Yang* will develop disease in an aggressive, active, progressive, and warm way (e.g., dry eyes, tinnitus, and night sweats). The five elements (earth, metal, water, wood, and fire) correspond to particular organs in the human body. Each element operates in

Table 2 Eight Principles

English	Chinese
Eight Principles	八綱
Yin (Negative)	陰
Yang (Positive)	陽
Cold	寒
Hot	熱
External	表
Internal	裏
Weak	虛
Strong	實

Table 3 Five Elements

English	Chinese
Five Elements	五行
Metal	金
Wood	木
Water	水
Fire	火
Earth	土

Table 4 Five Zang and Six Fu

English	Chinese
Five Zang	五臟
Heart	心
Lung	肺
Spleen	脾
Liver	肝
Kidney	腎
Six Fu	六腑
Gall bladder	膽
Stomach	胃
Large intestine	大腸
Small intestine	小腸
Urinary bladder	膀胱
Three cavities (chest, epiastrum, hypogastrum)	三焦 (上焦: 橫膈以上 中焦: 橫膈到肚臍 下焦: 肚臍以下)

Table 5 Examples of Yang and Yin

English	Chinese
Yang	陽
Sun	日
Sky	天
Day Time	晝
Hot	熱
Left	左
Up	上
Spring	春
Summer	夏
Six Fu	六腑
Strong	強壯
Yin	陰
Moon	月
Ground	地
Night	夜
Cold	寒
Right	右
Down	下
Autumn	秋
Winter	冬
Five Zang	五臟
Weak	虛弱

harmony with the others. Fig. 1 illustrates the operation relationships between the five elements.

The five Zang (or Yin organs) include heart (including the pericardium), lung, spleen, liver, and kidney, whereas the six Fu (or Yang organs) include gall bladder, stomach, large intestine, small intestine, urinary bladder, and three

cavities (i.e., chest, epigastrium, and hypogastrum). Zang organs can manufacture and store fundamental substances. These substances are then transformed and transported by Fu organs. TCM treatments involve a thorough understanding of the clinical manifestations of Zang-Fu organ imbalance as well as knowledge of appropriate acupuncture points and herbal therapy to rebalance or maintain the balance of the organs. The channels and collaterals are the representation of the organs of the body and are responsible for conducting the flow of energy and blood through the entire body.

The elements of TCM can also help to describe the etiology of disease including six exogenous factors (i.e., wind, cold, summer, dampness, dryness, and fire), seven emotional factors (i.e., anger, joy, worry, grief, anxiety, fear, and fright), and other pathogenic factors (Table 6). Once all of the information is collected and processed into a logical and workable diagnosis, the TCM doctor can determine the treatment approach.

Under the medical theory and mechanism described above, Chinese doctors believe that all of the organs within a healthy subject should reach the so-called *global dynamic balance or harmony* among organs. Once the global balance is broken at certain sites such as heart, liver, or kidney, some signs and symptoms then appear to reflect the imbalance at these sites. An experienced Chinese doctor usually assesses the causes of global imbalance before a TCM with flexible doses is prescribed to fix the problem. This approach is sometimes referred to as a personalized (or individualized) medicine approach.

Medical Practice

Different medical perceptions regarding signs and symptoms of certain diseases could lead to a different diagnosis

Table 6 Six Exogenous Factors and Seven Emotional Factors

English	Chinese
Six Exogenous Factors	外在因素(六淫)
Wind	風
Cold	寒
Summer	暑
Dampness	濕
Dryness	燥
Fire	火
Seven Emotional Factors	內在因素(七情)
Angry	怒
Joy	喜
Worry	憂
Grief	悲
Anxiety	思
Fear	恐
Fright	驚

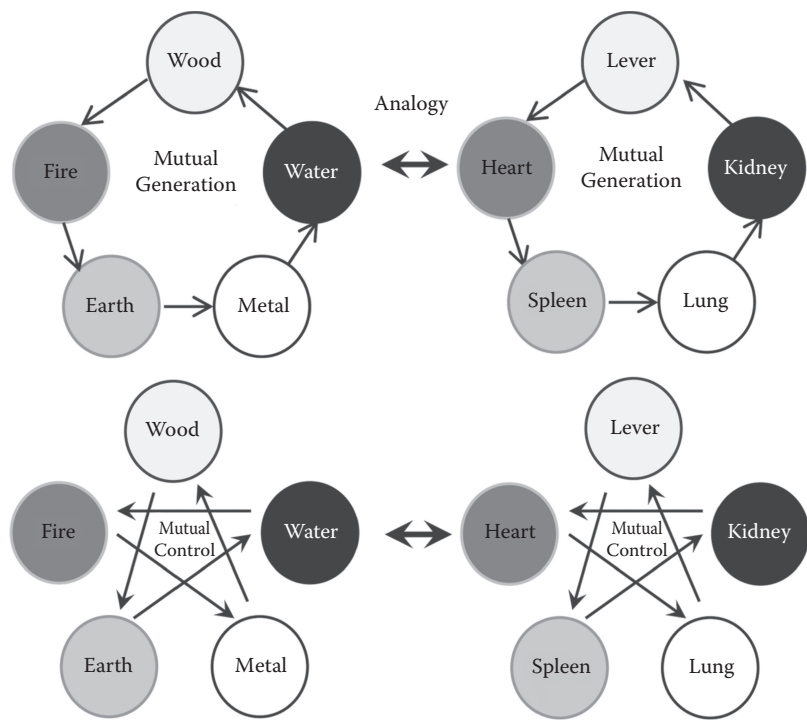


Fig. 1 The relationships between the five elements

and treatment for the diseases under study. For example, the signs and symptoms of Type 2 diabetic subjects could be classified as the disease of *thirsty reduction* by Chinese doctors. The disease of Type 2 diabetes is not recognized by Chinese medical literature, although the disease has the same signs and symptoms as the well-known disease of thirsty reduction. This difference in medical perception and practice has an impact on the diagnosis and treatment of the disease.

In addition, we tend to see therapeutic effect of WM sooner than that of TCM. TCM is often considered for patients who have chronic diseases or non-life-threatening diseases. For critical and/or life-threatening diseases such as cancer or stroke, TCM is often used as the second-or third-line treatment with no other alternative treatments. In many cases such as patients with later phase of cancer, TCM is often used in conjunction with WM without the knowledge of the primary care physicians.

Techniques of Diagnosis

The Chinese diagnostic procedure for patients with certain diseases consists of four major techniques, namely, inspection, auscultation and olfaction, interrogation, and pulse taking and palpation (Table 7). All these diagnostic techniques aim mainly at providing an objective basis for differentiation of syndromes by collecting symptoms and signs from the patient. Inspection involves observing the patient's general appearance (strong or weak, fat or thin), mind, complexion (skin color), five sense organs (eye, ear, nose, lip, and tongue), secretions, and excretions.

Table 7 Four Major Techniques

English	Chinese
Four Major Techniques	四診
Inspection	望
Auscultation and olfaction	聞
Interrogation	問
Pulse taking and palpation	切

Auscultation involves listening to the voice, expression, respiration, vomit, and cough. Olfaction involves smelling the breath and body odor. Interrogation involves asking questions about specific symptoms and the general condition including history of the present disease, past history, personal life history, and family history. Pulse taking and palpation can help to judge the location and nature of a disease according to the changes of the pulse.

The Chinese diagnostic procedure of inspection, auscultation and olfaction, interrogation, and pulse taking and palpation is subjective, with large between rater variability (i.e., variability from one Chinese doctor to another). This subjectivity and variability will have an impact not only on the patient's evaluability but also the prescribability of TCM, which will be further discussed below.

Objective Versus Subjective Criteria for Evaluability

For evaluation of a WM, objective criteria based on some well-established clinical study endpoints are usually considered. For example, response rate (i.e., complete response plus

partial response based on tumor size) is considered a valid clinical endpoint for evaluating clinical efficacy of oncology drug products. Unlike WMs, Chinese diagnostic procedure for evaluation of a TCM is very subjective. The use of a subjective Chinese diagnostic procedure has raised the following issues. First, it is a concern whether the subjective Chinese diagnostic procedure can accurately and reliably evaluate clinical efficacy and safety of the TCM under investigation. Thus, it is suggested that the subjective Chinese diagnostic procedure should be validated in terms of its accuracy, precision, and ruggedness before it can be used in TCM clinical trials. A validated Chinese diagnostic procedure should be able to detect a clinically significant difference if the difference truly exists. On the other hand, it is not desirable to wrongly detect a difference when there is no difference.

In clinical trials, evaluation is usually based on some validated tools (instruments) such as laboratory tests. Test results are then evaluated against some normal ranges for abnormality. Thus, it is suggested that the Chinese diagnostic procedure must be validated in terms of validity and reliability, and its false positive and false negative rates, before it can be used for evaluation of clinical efficacy and safety of the TCM under investigation.

Treatment

TCM prescriptions typically consist of a combination of several components. The combination is usually determined based on the medical theory of global dynamic balance (or harmony) among organs, and the observations from the Chinese diagnostic procedure. The use of Chinese diagnostic procedure is to find out what caused the imbalance among these organs. The treatment is to re-install the balance among these organs. Thus, the dose and treatment duration are flexible in order to achieve the balance. This concept leads to the concept of so-called personalized (or individualized) medicine, which minimizes intra-subject variability.

Single Active Ingredient Versus Multiple Components

Most Western medicines (WMs) contain a single active ingredient. After drug discovery, an appropriate formulation (or dosage form) is necessarily developed so that the drug can be delivered to the site action in an efficient way. At the same time, an assay is necessarily developed to quantitate the potency of the drug. The drug is then tested on animals for toxicity, and humans (healthy volunteers) for pharmacological activities. Unlike WMs, TCMs usually consist of multiple components with certain relative proportions among the components. As a result, the typical approach for evaluation of single active ingredient for WMs is not applicable.

In practice, one may suggest evaluating the TCM component by component. However, this is not feasible due to the following difficulties. First, in practice, analytical

methods for quantitation of individual components are often not tractable. Thus, the pharmacological activities of these components are not known. It should be noted that the component that comprises the major proportion of the TCM may not be the most active component. On the other hand, the component that has the least proportion of the TCM may be the most active component of the TCM. In practice, it is not known which relative proportions among these components can lead to the optimal therapeutic effect of the TCM. In addition, the relative component-to-component and/or component by food interactions are usually unknown, which may have an impact on the evaluation of clinical efficacy and safety of the TCM.

Fixed Dose Versus Flexible Dose

Most WMs are usually administered in a fixed dose (say a 10mg tablets or capsule). On the other hand, because a TCM consists of multiple components with possible varied relative proportions among the components, a Chinese doctor usually prescribes the TCM with different relative proportions of the multiple components based on the signs and symptoms of the patient according to his or her best judgment following a subjective evaluation based on the Chinese diagnostic procedure. Thus, unlike a WM that is prescribed as a fixed dose, a TCM is often prescribed as an individualized flexible dose.

The approach of WM with a fixed dose is a population approach to minimize the between subject (or intersubject) variability, whereas the approach to TCM with an individualized flexible dose is to minimize the variability within each individual. In practice, it is a concern whether an individual flexible dose is compatible with a Western evaluation of the TCM. An individualized flexible dose depends heavily upon the Chinese doctor's subjective judgment, which may vary from one Chinese doctor to another. As a result, although an individualized flexible dose does minimize intra-subject variability, the variability from one Chinese doctor to another (i.e., the doctor-to-doctor or rater-to-rater variability) could be huge, and hence non-negligible.

Remarks

For the research and development of a TCM, before a TCM clinical trial is conducted, the following questions are necessarily asked.

- i. Will the TCM clinical trial be conducted by Chinese doctors alone, Western clinicians alone, Western clinicians who have some background of Chinese herbal medicine alone, or both Chinese doctors and Western clinicians?
- ii. Will traditional Chinese diagnostic and/or trial procedures be used throughout the TCM clinical trial?
- iii. Upon approval, is the TCM intended for use by Chinese doctors or Western clinicians?

With respect to the first two questions, if the TCM clinical trial is to be conducted by Chinese doctors alone, the following questions arise. First, should the Chinese diagnostic procedure be validated in order to provide an accurate and reliable assessment of the TCM? In addition, it is of interest to determine how an observed difference obtained from the Chinese diagnostic procedure can be translated to the clinical endpoint commonly used in similar WM clinical trials with the same indication. These two questions can be addressed statistically by the calibration and validation of the Chinese diagnostic procedure with respect to some well-established clinical endpoints for evaluation of Western medicines. If the TCM clinical trial is to be conducted by Western clinicians or Western clinicians who have some background of Chinese herbal medicine, the standards and consistency of clinical results as compared to those WM clinical trials are ensured. However, the good characteristics of TCM may be lost during the process of the conduct of the TCM clinical trials. On the other hand, if the TCM clinical trial is to be conducted by both Chinese doctors and Western clinicians, difference in medical practice and/or possible disagreement regarding the diagnosis, treatment, and evaluation are major concerns.

For the third question, if the TCM is intended for use of Chinese doctors but is conducted by Western clinicians, the difference in perception regarding how to prescribe the TCM is of great concern. The preparation of a package insert based on the clinical data could be a major issue, not only to the sponsor but also to regulatory authorities. Similar comments apply to the situation where the TCM is intended for use of Western clinicians but the trial is conducted by Chinese doctors.

As a result, it is suggested that the intention of use (i.e., labeling for the indication) be clearly evaluated when planning a TCM clinical trial. In other words, the sponsor needs to determine at the planning stage of a TCM clinical trial whether the TCM is intended for use of Western clinicians only, Chinese doctors only, or both Western clinicians and Chinese doctors in order to plan, for an adequate package insert of the target diseases under study.

BASIC CONSIDERATIONS OF TCM CLINICAL TRIALS

In this section, we describe some basic considerations that are necessary in order to ensure the success of a TCM clinical trial.

Study Design

To demonstrate clinical efficacy and safety of TCM under investigation, it is suggested that a randomized parallel-group, placebo-controlled clinical trial be conducted, as is done in WM. However, doing so may not be ethical if the disease under study is critical and/or life-threatening

provided that a WM is available. Alternatively, a randomized placebo-control crossover clinical trial or a parallel-group design consisting of three arms (i.e., the TCM under study, a WM as an active control, and a placebo) is recommended. The three-arm, parallel-group design allows the establishment of non-inferiority/equivalence of the TCM as compared to the active control (WM) and the demonstration of the superiority of the TCM with respect to the placebo. One of the advantages of a crossover clinical trial is that a comparison within each individual can be made, although it will take a longer time to complete the study. Although a crossover design requires a smaller sample size as compared to a parallel-group design, there are some limitations for the use of crossover design. First, baselines prior to dosing may not be the same. Second, when a significant sequence effect is observed, we would not be able to isolate the effects of period effect, carry over effects, and subject-by-treatment effect which are confounded to one another.

In many cases, factorial designs are used to evaluate the impact of specific components (with respect to the therapeutic effect) by fixing some of the components. For example, we may consider a parallel-group design comparing two treatment groups (one group is treated with the TCM with a specific component, and the other group is treated with the TCM without the specific component). Design of this kind may be useful to identify the most active component of the TCM with respect to the diseases under study. However, it does not address the possible drug-to-drug interactions among the components.

Validation of Quantitative Instrument

In TCM practice, a doctor usually collects information from the patient with a certain disease through the four subjective approaches described in the previous section. The purpose of these subjective approaches is to collect information on various aspects of the disease under study, such as signs, symptoms, patient's performance, and functional activities, so a quantitative instrument with a large number of questions/items is necessary and helpful. For simple analysis and easy interpretation, these questions are usually grouped to form subscales, composite scores (domains), or overall score. The items (or subscales) in each subscale (or composite score) are correlated. As a result, the structure of responses to a quantitative instrument is multidimensional, complex, and correlated. As mentioned above, a standardized quantitative tool (instrument) is necessary to reduce variability from one Chinese doctor to another (prior to the conduct of a clinical trial).

Guilford^[1] have discussed several methods such as Cronbach's α for measuring the reliability of internal consistency of a quantitative instrument. Guyatt et al.^[2] have indicated that a quantitative instrument should be validated in terms of its validity, reproducibility, and responsiveness. Hollenberg et al.^[3] have discussed several methods for

validation of a quantitative instrument, such as consensual validation, construct validation, and criterion-related validation. There is, however, no gold standard as to how a quantitative instrument should be validated. In this paper, we will focus on the validation of a quantitative instrument in terms of validity, reliability (or reproducibility) and responsiveness.^[4,5] As indicated in Chow and Shao,^[6] the validity of a quantitative instrument is the extent to which the instrument measures what is designed to measure. It is a measure of biasedness of the instrument. The biasedness of a quantitative instrument reflects the accuracy of the instrument. The reliability of a quantitative instrument measures the variability of the instrument, which directly relates to the precision of the instrument. On the other hand, the responsiveness of a quantitative instrument is usually referred to as the ability of the instrument to detect a difference of clinical significance within a treatment.

Hsiao et al.^[7] have considered a specific design for calibration/validation of the Chinese diagnostic procedure. In the proposed study design, qualified subjects are randomly assigned to receive either a TCM or a WM. Each patient is evaluated by a Chinese doctor and a Western clinician independently, regardless which treatment group he or she is in. As a result, there are four groups of data, namely (i) patients who receive TCM and evaluated by a Chinese doctor, (ii) patients who receive TCM but are evaluated by a Western clinician, (iii) patients who receive WM but are evaluated by a Chinese doctor, and (iv) patients who receive WM and are evaluated by a Western clinician. Groups (iii) and (iv) are used to establish a standard curve for calibration between the TCM and the WM. Groups (i) and (ii) are then used to validate the Chinese diagnostic procedure based on the established standard curve.

Clinical Endpoint

Unlike WMs, the primary study endpoints for assessment of safety and effectiveness of a TCM are usually assessed subjectively by a quantitative instrument by experienced Chinese doctors. Although the quantitative instrument is developed by the community of Chinese doctors and is considered a gold standard for assessment of safety and effectiveness of the TCM under investigation, it may not be accepted by the Western clinicians due to fundamental differences in medical theory, perception, and practice. In practice, it is very difficult for a Western clinician to conceptually understand the clinical meaning of the difference detected by the subjective Chinese quantitative instrument. Consequently, whether the subjective quantitative instrument can accurately and reliably assess the safety and effectiveness of the TCM is always a concern to Western clinicians.

Take for example the assessment of safety and efficacy of a drug product for treatment of ischemic stroke, a commonly considered primary clinical endpoint is the functional status assessed by the so-called Barthel index.

The Barthel index is a weighted functional assessment scoring technique composed of 10 items with a minimum score of 0 (functional incompetence) and a maximum score of 100 (functional competence). The Barthel index is a weighted scale measuring performance in self-care and mobility, which is widely accepted in ischemic stroke clinical trials. A patient may be considered a responder if his or her Barthel index is greater than or equal to 60. On the other hand, Chinese doctors usually consider a quantitative instrument developed by the Chinese medical community as the standard diagnostic procedure for assessment of ischemic stroke. The standard quantitative instrument is composed of six domains, which capture different information regarding patient's performance, functional activities, and signs and symptoms and status of the disease.

In practice, it is of interest to both Western clinicians and Chinese doctors how an observed clinically meaningful difference by the Chinese quantitative instrument can be translated to that of the primary study endpoint assessed by the Barthel index. To reduce the fundamental differences in medical theory/perception and practice, it is suggested that the subjective Chinese quantitative instrument be calibrated and validated with respect to that of the clinical endpoint assessed by the Barthel index before it can be used in TCM ischemic stroke clinical trials.

Matching Placebo

In clinical development, double-blind, placebo-control randomized clinical trials are often conducted for evaluation of the safety and effectiveness of a test treatment under investigation. To maintain blindness, a matching placebo should be identical to the active drug in all aspects of, size, color, coating, taste, texture, shape, and order except that it contains no active ingredient. In clinical trials, as advanced technique available for formulation, a matching placebo is not difficult to make because most WMs contain a single active ingredient. Unlike WMs, TCMs usually consist of a number of components, which often have different taste. In TCM clinical trials, the TCM under investigation is often encapsulated. However, the test treatment will be easily unblinded if either the patient or Chinese doctor breaks the capsule. As a result, the preparation of matching placebo in TCM clinical trials plays an important role for the success of the TCM clinical trials.

Sample Size Calculation

In clinical trials, sample size is usually selected to achieve a desired power for detecting a clinically meaningful difference in one of the primary study endpoints for the intended indication of the treatment under investigation.^[8] As a result, sample size calculation depends upon the primary study endpoint and the clinically meaningful difference that one would like to detect. Different primary study endpoints may result in very different sample sizes.

Consider the example concerning a TCM for treatment of ischemic stroke that was developed with more than 30 years clinical experience with humans. Suppose a sponsor would like to conduct a clinical trial to scientifically evaluate the safety and efficacy of the TCM the Western way as compared to an active control (e.g., aspirin). Thus, the intended clinical trial is a double-blind, parallel-group, placebo-control, randomized trial. The primary clinical endpoint is the response rate (a patient is considered a responder if his/her Barthel index is greater than or equal to 60) based on the functional status assessed by the Barthel index. Sample size calculation is performed based on the response rate after four weeks of treatment under the hypotheses of testing for superiority. As a result, a sample size of 150 patients per treatment group is required for achieving an 80% power for establishment of superiority of the TCM over the active control agent. Alternatively, we may consider the quantitative instrument developed by experienced Chinese doctors as the primary study endpoint for sample size calculation. Based on a pilot study, about 80% (79 out of 122) of ischemic stroke patients were diagnosed by one domain of the quantitative instrument. A patient is considered a responder if his or her domain score is greater than or equal to 7. Based on this primary study endpoint, a sample size of 90 per treatment group is required to achieve an 80% power for establishment of superiority.

The difference in sample size leads to the question of whether the use of the primary endpoint of response rate based on one domain of the Chinese quantitative instrument could provide substantial evidence of safety and effectiveness of the TCM under investigation.

OTHER ISSUES IN TCM RESEARCH AND DEVELOPMENT

Although TCM has a long history of use in humans, no scientifically valid documentations of its safety or effectiveness are available. As indicated by the U.S. Food and Drug Administration (FDA), substantial evidence regarding safety and effectiveness of the test treatment under investigation can only be obtained by conducting adequate and well-controlled clinical trials. However, before the test treatment under investigation can be used in human, sufficient information regarding chemistry, manufacturing, and control (CMC); clinical pharmacology; and toxicology must be provided.^[9] Because most TCMs consist of multiple components with unknown pharmacological activities, valid information regarding CMC, clinical pharmacology, and toxicology are difficult to obtain. In what follows, we briefly describe these difficulties.

Test for Consistency

As mentioned above, unlike most WMs, TCMs usually consist of a number of components that are extracted

from herbal samples. The herbal samples are normally dried at 60°C to a constant weight, followed by grinding in a mortar and storage in a desiccator. For water-soluble substances, an appropriate amount of water is first added to the dried material and boil for about one hour. For alcohol-soluble substances, 60% ethanol is added and the mixture is extracted at 60°C for one hour in an ultrasonic bath. After cooling to room temperature, the extract can be cleared by filtration through net or centrifugation at 12,000 g for 10 minutes at 20°C, and the supernatant is used for further applications. The pharmacological activities, interactions, and relative proportions of these components are usually unknown. In practice, TCM is usually prescribed subjectively by an experienced Chinese doctor. As a result, the actual dose received by each individual varies depending upon the signs and symptoms as perceived by the Chinese doctor. Although the purpose of this medical practice is to reduce the within-subject (or intra-subject) variability, it could also introduce non-negligible variability such as variations from component-to-component and from rater-to-rater (a Chinese doctor to another). Consequently, reproducibility or consistency of clinical results is questionable. Thus, ensuring the reproducibility or consistency of the observed clinical results has become a great concern to regulatory agencies in the review and approval process. It is also a great concern to the sponsor of the manufacturing process. To address the question of reproducibility or consistency, a valid statistical quality control process on the raw materials and final product is suggested.

Chow et al.^[10] and Tse et al.^[13] have proposed a statistical quality control (QC) method to assess a proposed consistency index of raw materials, which are from different resources and/or final product, which may be manufactured at different sites. The consistency index is defined as the probability that the ratio of the characteristics (e.g., extract) of the most active component among the multiple components of a TCM from two different sites (locations) is within a limit of consistency. The consistency index closest to 1 indicates that the components from the two sites or locations are almost identical. The idea for testing consistency is to construct a 95% confidence interval for the proposed consistency index under a sampling plan. If the constructed 95% confidence lower limit is greater than a pre-specified QC lower limit, then we claim that the raw materials or final product has passed the QC and hence can be released for further process or use. Otherwise, the raw materials and/or final product should be rejected.

Let U and W be the characteristics of the most active component among the multiple components of a TCM from two different sites, where $X = \log U$ and $Y = \log W$ follows normal distributions with means μ_X , μ_Y and variances V_X , V_Y , respectively. Similar to the idea of using $P(X < Y)$ to assess reliability in statistical quality control,^[11,12] Tse et al.^[10] have proposed the following probability as an

index to assess the consistency of raw materials and/or final product from two different sites

$$p = P\left(1 - \delta < \frac{U}{W} < \frac{1}{1 - \delta}\right), \quad (1)$$

where $0 < \delta < 1$ and is defined as a limit that allows for consistency. Tse et al.^[10] refer p as the consistency index. Thus p tends to 1 as δ tends to 1. For a given δ , if p is close to 1, materials U and W are considered to be identical. It should be noted that a small δ implies the requirement of high degree of consistency between material U and materials W . In practice, it may be difficult to meet this narrow specification for consistency. Under the normality assumption of $X = \log U$ and $Y = \log W$, p is given by

$$p = \Phi\left(\frac{-\log(1 - \delta) - (\mu_X - \mu_Y)}{\sqrt{V_X + V_Y}}\right) - \Phi\left(\frac{\log(1 - \delta) - (\mu_X - \mu_Y)}{\sqrt{V_X + V_Y}}\right),$$

where $\Phi(z_0) = P(Z < z_0)$ with Z being a standard normal random variable. Therefore, the consistency index p is a function of the parameters

$$\theta = (\mu_X, \mu_Y, V_X, V_Y),$$

Suppose that observations $X_i = \log U_i, i = 1, \dots, n_X$, $Y_i = \log W_i, i = 1, \dots, n_Y$ are collected in an assay study. Then, using the invariance principle, the maximum likelihood estimator (MLE) of p can be obtained as

$$\hat{p} = \Phi\left(\frac{-\log(1 - \delta) - (\bar{X} - \bar{Y})}{\sqrt{\hat{V}_X + \hat{V}_Y}}\right) - \Phi\left(\frac{\log(1 - \delta) - (\bar{X} - \bar{Y})}{\sqrt{\hat{V}_X + \hat{V}_Y}}\right), \quad (2)$$

where

$$\bar{X} = \frac{1}{n_X} \sum_{i=1}^{n_X} X_i, \bar{Y} = \frac{1}{n_Y} \sum_{i=1}^{n_Y} Y_i, \hat{V}_X = \frac{1}{n_X} \sum_{i=1}^{n_X} (X_i - \bar{X})^2,$$

$$\text{and } \hat{V}_Y = \frac{1}{n_Y} \sum_{i=1}^{n_Y} (Y_i - \bar{Y})^2.$$

In other words, $\hat{p} = h(\theta) = h(\bar{X}, \bar{Y}, \hat{V}_X, \hat{V}_Y)$. Furthermore, it can be easily verified that the following asymptotic result holds. It can be verified that $\hat{p}$ is asymptotically normal with mean $E(\hat{p})$ and variance $Var(\hat{p})$. That is,

$$\frac{\hat{p} - E(\hat{p})}{\sqrt{Var(\hat{p})}} \rightarrow N(0, 1), \quad (3)$$

where $E(\hat{p}) = p + B(p) + o(\frac{1}{n})$ and $Var(\hat{p}) = C(p) + o(\frac{1}{n})$, where $B(p)$ and $C(p)$ are some functions of p .^[13] As a result, an approximate $(1 - \alpha) \times 100\%$ confidence interval for p , i.e., $(LL(\hat{p}), UL(\hat{p}))$, can be obtained as follows

$$LL(\hat{p}) = \hat{p} - B(\hat{p}) - z_{\alpha/2} \sqrt{C(\hat{p})}, \text{ and} \quad (4)$$

$$UL(\hat{p}) = \hat{p} - B(\hat{p}) + z_{\alpha/2} \sqrt{C(\hat{p})}.$$

where $z\alpha$ is the upper α -percentile of a standard normal distribution. In terms of consistency, Tse et al.^[10] proposed the following quality control (QC) criterion. If the probability that the lower limit $LL(\hat{p})$ of the constructed $(1 - \alpha) \times 100\%$ confidence interval of p is greater than or equal to a pre-specified quality control lower limit, say, QC_L , exceeds a pre-specified number β (say $\beta = 80\%$), then we claim that U and W are consistent or similar. In other words, U and W are consistent or similar if $P(QC_L \leq LL(\hat{p})) \geq \beta$, where β is a pre-specified constant. In practice, it is necessary to select a sample size to ensure that there is a high probability, say β , of consistency between U and W when in fact U and W are consistent. It is suggested that the sample size is chosen such that there is more than 80% chance that the lower confidence limit of p is greater than or equal to the QC lower limit, i.e. $\beta = 0.8$. In other words, the sample size is determined such that

$$P\{QC_L \leq LL(\hat{p})\} \geq \beta. \quad (5)$$

Using Eq. (5), this leads to

$$P\{QC_L \leq \hat{p} - B(\hat{p}) - z_{\alpha/2} \sqrt{Var(\hat{p})}\} \geq \beta.$$

Thus,

$$P\{QC_L + z_{\alpha/2} \sqrt{Var(\hat{p})} - p \leq \hat{p} - p - B(p)\} \geq \beta.$$

This gives

$$P\left\{\frac{QC_L - p}{\sqrt{Var(\hat{p})}} + z_{\alpha/2} \leq \frac{\hat{p} - p - B(p)}{\sqrt{Var(\hat{p})}}\right\} \geq \beta.$$

Therefore, the sample size required for achieving a probability higher than β can be obtained by solving the following equation:

$$\frac{QC_L - p}{\sqrt{Var(\hat{p})}} + z_{\alpha/2} \leq -z_{1-\beta}. \quad (6)$$

Assuming that $n_X = n_Y = n$, then the common sample size is given by

$$n \geq \frac{(z_{1-\beta} + z_{\alpha/2})^2}{(p - QC_L)^2} \left\{ \left(\frac{\partial \hat{p}}{\partial \mu_X} \right)^2 V_X + \left(\frac{\partial \hat{p}}{\partial \mu_Y} \right)^2 V_Y + \left(\frac{\partial \hat{p}}{\partial V_X} \right)^2 (2V_X^2) + \left(\frac{\partial \hat{p}}{\partial V_Y} \right)^2 (2V_Y^2) \right\}.$$

The above result suggests that the required sample size will depend on the choices of α , β , V_x , V_y , $\mu_x - \mu_y$ and $p - QC_L$.

It should be noted that Tse et al.^[10] only focus on single (i.e., the most active) component assuming that the most active component can be quantitatively identified among the multiple active components. Lu, Chow, and Tse^[14] extend the results to the case of two correlative components following similar idea. Various sampling plans are obtained for various combinations of study parameters to reflect real practice in quality control of traditional Chinese medicine.

Stability Analysis

Most regulatory agencies require that the expiration dating period (or shelf life) of a drug product must be indicated in the immediate container label before it can be released for use. To fulfill this requirement, stability studies are usually conducted in order to characterize the degradation of the drug product. For drug products with a single active ingredient, statistical methods for determination of drug shelf life are well established.^[15,16] However, regulatory requirements for estimation of drug shelf life for drug products with multiple components are not available.

Following the concept of estimating shelf life for drug products with single active ingredient, two approaches are worthy considering. First, we may (conservatively) consider the minimum of the shelf-lives obtained from each component of the drug product. This approach is conservative, and yet may not be feasible due to the fact that (i) not all of the components of a TCM can be accurately and reliably quantitated, and (ii) the resultant shelf life may be too short to be useful.^[17] Alternatively, we may consider a two-stage approach for determination of drug shelf life. At the first stage, an attempt should be made to identify the most active component(s) whenever possible. A shelf life can then be obtained based on the method suggested in the FDA and ICH guidelines. At the second stage, the obtained shelf life is adjusted based on the relationship and/or interactions of the most active ingredient(s) and other components.

Alternatively, Chow and Shao^[18] have proposed a statistical method for determining the shelf life of a TCM following a similar idea suggested by the FDA, assuming that the components are linear combinations of some factors. Let $y(t)$ denote the p dimensional vector whose k th component is the potency of the k th ingredient at time t after the manufacture of a given TCM, $k = 1, \dots, p$. Assume that, for any t ,

$$y(t) - E[y(t)] = LF_t + \varepsilon_t, \quad (7)$$

where L is a $p \times q$ nonrandom unknown matrix of full rank, F_t and ε_t are unobserved independent random vectors of dimensions q and p , respectively, $E(F_t) = 0$, $Var(F_t) = I_q$ (the identity matrix of order q), $E(\varepsilon_t) = 0$, $Var(\varepsilon_t) = \Psi$, and Ψ

is an unknown diagonal matrix of order p . Let $z(t) = (L'L)^{-1}L'y(t)$. It follows from Eq. (7) that

$$z(t) - E[z(t)] = F_t + (L'L)^{-1}L'\varepsilon_t. \quad (8)$$

Now, let $x(t)$ be an s dimensional covariate vector associated with $y(t)$ at time t . Chow and Shao^[18,19] assume the following model at any time t :

$$E[y(t)] = Bx(t), Var[y(t)] = \Sigma, \quad i = 1, \dots, m, j = 1, \dots, n, \quad (9)$$

where B is a $p \times s$ matrix of unknown parameters and $\Sigma > 0$ is an unknown $p \times p$ positive definite covariate matrix. Since $z(t) = (L'L)^{-1}L'y(t)$, it follows from Eq. (9) that

$$E[z(t)] = \gamma'x(t) = (L'L)^{-1}L'Bx(t), \quad i = 1, \dots, m, j = 1, \dots, n, \quad (10)$$

where $\gamma = B'L(L'L)^{-1}$. Suppose that we independently observe data y_{ij} , $i = 1, \dots, m, j = 1, \dots, n$, where y_{ij} is the j th replicate of $y(t_i)$ and $t_1, \dots, t_m$ are designed time points for the stability analysis. Define

$$x_i = x(t_i), z_{ij} = (L'L)^{-1}L'y_{ij}, \quad i = 1, \dots, m, j = 1, \dots, n \quad (11)$$

Consider the case where $q = 1$. If z_{ij} 's are observed, according to FDA,^[15] the expiration dating period (or shelf life) is given by

$$\tau = \inf\{t : l(t) \leq \eta\}, \quad (12)$$

where η is the lower product specification limit as specified in the USP/NF. However, z_{ij} 's are not observed but y_{ij} 's. Thus, the lower confidence bound $l(t)$ in (12) needs to be modified as follows

$$l(t) = \hat{\gamma}'x(t) - t_{0.95, mn-s} \sqrt{x(t)'Vx(t)}, \quad (13)$$

where

$$V = \frac{mn-1}{mn} \sum_{i=1}^m \sum_{j=1}^n (\hat{\gamma}_{i,j} - \hat{\gamma})(\hat{\gamma}_{i,j} - \hat{\gamma})',$$

in which $\hat{\gamma}_{i,j}$ is the estimator of γ with the (i,j) th data point deleted. For the case where $1 < q \leq p$, let $\hat{B}$ be least squares estimator of B , λ_k be the k th largest eigenvalue of the sample covariance matrix based on $y_{ij} - \hat{B}x_i$, $i = 1, \dots, m, j = 1, \dots, n$, and e_k be the normalized eigenvector corresponding to λ_k . Then, estimator $\hat{L}$ of L is the $p \times q$ matrix whose k th column is $\lambda_k e_k$, $k = 1, \dots, q$. Let $\hat{\gamma}_k$ be the k th column of $\hat{\gamma} = \hat{B}'\hat{L}(\hat{L}'\hat{L})^{-1}$, which is an $s \times q$ matrix. Thus we have

$$l_k(t) = \hat{\gamma}_k'x(t) - t_{0.95/q, mn-s} \sqrt{x(t)'V_kx(t)}, \quad (14)$$

where

$$V_k = \frac{mn-1}{mn} \sum_{i=1}^m \sum_{j=1}^n (\hat{\gamma}_{k,i,j} - \hat{\gamma}_k)(\hat{\gamma}_{k,i,j} - \hat{\gamma}_k)',$$

in which $\hat{\gamma}_{k,i,j}$ is the same as $\hat{\gamma}_k$ but calculated with the (i,j) th data point deleted. Then, $l_k(t)$, $k = 1, \dots, q$ are approximate 95% simultaneous lower confidence bounds for $\zeta(t)$, $k = 1, \dots, q$, where $\zeta_k(t)$ is the k th component of $E[z(t)] = \gamma'x(t)$. An estimated shelf life is then given by

$$\tau = \min_{k=1, \dots, q} \tau_k,$$

where τ_k is defined by the right-hand side of Eq. (12) with $l(t)$ replaced by $l_k(t)$ and is in fact an estimated shelf life for the k th component of z with confidence level $(1 - 0.05/q)\%$.

Animal Studies

The purpose of animal studies is not only to study possible toxicity in animals, but also to suggest an appropriate dose for use in humans, assuming that the established animal model is predictive of the human model. For a newly developed drug product, animal studies are necessary. However, for some well-known TCMs, which have been used in humans for years and have a very mild toxicity profile, it is questionable whether animal studies are necessary. It is suggested that all components of TCMs as described in Chinese Pharmacopoeia (CP) be classified into several categories depending upon their potential toxicities and/or safety profiles as a basis for regulatory requirements for animal studies. In other words, for some well-known TCM components such as Ginseng, animal studies for testing toxicity may be waived depending upon past experience of human use, although health risks or side effects following the proper administration of designated therapeutic dosages were not recorded in human use. Note that the German regulatory authority's herbal watchdog agency, commonly called Commission E, has conducted an intensive assessment of the peer-reviewed literature on some three hundred common botanicals with respect to the quality of the clinical evidence and the uses for which the herb can be reasonably considered effective.^[20]

Regulatory Requirements

Although the use of TCMs in humans has a long history, there have been no regulatory requirements regarding the assessment of safety and effectiveness of the TCMs until recently. For example, the regulatory authorities of both China and Taiwan have published guidelines/guidances for clinical development of TCMs.^[21–23] In addition, the U.S. FDA has also published a guidance for botanical drug products.^[24] These regulatory requirements for TCM research and development, especially for clinical development, are very similar to well-established guidelines/guidances for pharmaceutical research and development for WMs. It is a concern whether these regulatory requirements and the corresponding statistical methods are feasible for research and development of TCM, based on the fact that there are

so many fundamental differences in medical practice, drug administration, and diagnostic procedure. As a result, it is suggested that current regulatory requirements and the corresponding statistical methods should be modified in order to reflect these fundamental differences.

It is strongly recommended that regulatory requirements for development, review and approval process for Premarin (conjugated estrogens tablets, USP) be consulted because Premarin is a WM consisting of multiple components and is thus similar to a TCM.^[25–30] Premarin, which contains multiple components of estrone, equilin, 17α -dihydroequilin, 17α -estradiol, and 17β -dihydroequilin, is intended for treatment of moderate to severe vasomotor symptoms associated with the menopause. The experience with Premarin is helpful in developing appropriate guidelines/guidances for TCM drug products with multiple components.

Indication/Label

As indicated earlier, it is very important to clarify the intention for the use of a TCM (by Chinese doctors alone, Western clinicians alone, or both Chinese doctors and Western clinicians) once it is approved by the regulatory agencies. If a TCM is intended for use by Chinese doctors alone, the clinical trials conducted for obtaining substantial evidence should reflect medical theory of TCM and medical practice of Chinese doctors. The label should provide sufficient information as to how to prescribe the TCM the Chinese way. On the other hand, if the TCM under investigation is intended for use by Western clinicians alone, patients under study should be evaluated based on clinical study endpoints for safety and efficacy the Western way. Consequently, the label should provide sufficient information for prescribing the TCM the Western way. If the TCM is intended for both Western clinicians and Chinese doctors, patients are necessarily evaluated by both Western clinical study endpoints and Chinese diagnostic procedures (e.g., some standardized quantitative instrument) provided that the Chinese diagnostic procedure has been calibrated and validated against the well-established Western clinical endpoint. In this case, there is a clear understanding how an observed difference by Chinese diagnostic procedure can be translated to a clinical effect which is familiar to Western clinicians, and vice versa.

CONSORTIUM FOR GLOBALIZATION OF CHINESE MEDICINE

As indicated earlier, in pursuit of advancing the TCM to benefit patients who suffer from critical and/or life-threatening diseases, a Consortium for Globalization of Chinese Medicine (CGCM) was formed at the University of Hong Kong in 2003. The goals of the consortium are multifold. First, the CGCM is to provide an environment for development of platform technologies required

for advancing Chinese herbal medicine by joint efforts. Second, it is to facilitate the interaction and collaboration among different institutes in advancing Chinese herbal medicine by information sharing. Third, it is to promote high-quality research and development of Chinese medicine internationally. In addition, it is to assist industry and regulatory agencies worldwide in (i) methodology development in TCM research and development and (ii) guidances/guidelines development for evaluation and approval of TCM regulatory submissions. More details can be found at <http://www.tcmmedicine.org>.

The CGCM aims at to establish a network among researchers from academia, industry, and regulatory agencies under an academic collaboration arrangement focusing on two main areas. The first area is for the development of a centralized database of Chinese medicine that is completed with formulations, botanical ingredients, and phytochemical substances along with associated biological information on traditional and modern therapeutic indication, clinical information, animal pharmacology, biochemical activities, toxicities, and manufacturing information. The other area of focus is for the standardization and adoption of a modern quality-control platform for chemical and biological fingerprinting of botanical drugs for scientific and regulatory purpose. The Consortium had 13 institutional members when it was founded in 2003. As of August 2005, it has grown to 32 institutional members worldwide, including Columbia University and Yale University in the United States.

To achieve the goal of globalization of Chinese medicine, the CGCM has established six working groups including (i) quality control, (ii) informatics, (iii) herbal resources, (iv) clinical, (v) external affairs and industrial liaison, and (vi) intellectual property. The Quality Control Working Group develops strategy and appropriate methodology including acceptance criteria, sampling plan, and testing procedure for assuring consistency of raw materials, in-process materials, and final products in post-approval manufacturing process. The Informatics Working Group investigates possible biomarkers such as PK/PD markers or genomic markers for efficacy and safety in traditional Chinese medicine clinical trials. The Working Group for Herbal Resources not only identifies resources of Chinese herbs worldwide, but also determines potential differences in the same species of herbs. The External Affairs and Industrial Liaison Working Group seeks collaboration in TCM research development with the pharmaceutical industry and/or research organizations such as the National Institutes of Health (NIH) of the United States of America. The Working Group for Intellectual Property is developing strategies for patent protection. The efforts of the six working groups will lead advance in TCM research and development in the next century. However, it is also strongly suggested that a working group regarding regulatory requirements for review and approval of TCM be established to assist the sponsors to bring their

promising (efficacious and safe) TCM products to the marketplace. The Working Group for Regulatory Affairs should be a collaborative team that includes regulatory agencies from Taiwan, China, the United States, Europe (such as Germany), and Japan for a harmonized regulatory requirements.

CONCLUDING REMARKS

As indicated earlier, a TCM is defined as a Chinese herbal medicine developed for treating patients with certain diseases as diagnosed by the four major techniques of inspection, auscultation and olfaction, interrogation, and pulse taking and palpation based on traditional Chinese medical theory of global balance among the functions/activities of all organs of the body. When conducting a TCM clinical trial, it is suggested that the fundamental differences between a WM and a TCM, as described in the section "Fundamental Differences," should be evaluated carefully for a valid and unbiased assessment of the safety and effectiveness of the TCM under investigation.

One of the key issues in TCM research and development is the clarification of differences between Westernization of TCM and modernization of TCM. For Westernization of TCM, we follow regulatory requirements at critical stages of the process for pharmaceutical development including drug discovery, formulation, laboratory development, animal studies, clinical development, manufacturing process validation and quality control, regulatory submission, review, and process, despite the fundamental differences between WM and TCM. For modernization of TCM, it is suggested that regulatory requirements should be modified in order to account for the fundamental differences between WM and TCM. In other words, we still ought to be able to see if TCM is really working with modified regulatory requirements using Western clinical trials as a standard for comparison.

In practice, it is recognized that WMs tend to achieve the therapeutic effect sooner than that of TCMs for critical and/or life-threatening diseases. TCMs are found to be useful for patients with chronic diseases or non-life-threatening diseases. In many cases, TCMs have shown to be effective in reducing toxicities or improving safety profile for patients with critical and/or life-threatening diseases. As a strategy for TCM research and development, it is suggested that (i) TCM be used in conjunction with a well-established WM as a supplement to improve its safety profile and/or enhance therapeutic effect whenever possible, and that (ii) TCM should be considered as the second- or third-line treatment for patients who fail to respond to the available treatments. However, some sponsors are interested in focusing on the development of TCM as a dietary supplement due to (i) the lack or ambiguity of regulatory requirements, (ii) the lack of understanding of the medical theory/mechanism of TCM, (iii) the confidentiality

of non-disclosure of the multiple components, and (iv) the lack of understanding of pharmacological activities of the multiple components of TCM.

Because TCM consists of multiple components that may be manufactured from different sites or locations, the post-approval consistency in quality of the final product is both a challenge to the sponsor and a concern to the regulatory authority. As a result, some post-approval tests, such as tests for content uniformity, weight variation, and/or dissolution and (manufacturing) process validation, must be performed for quality assurance before the approved TCM can be released for use.

REFERENCES

- Guilford, J.P. *Psychometric Methods*, 2nd Ed.; McGraw-Hill: New York, 1954.
- Guyatt, G.H.; Veldhuyzen Van Zanten, S.J.O.; Feeny, D.H.; Patric, D.L. Measuring quality of life in clinical trials: A taxonomy and review. *CMAJ* **1989**, *140*, 1441–1448.
- Hollenberg, N.K.; Testa, M.; Williams, G.H. Quality of life as a therapeutic end-point—an analysis of therapeutic trials in hypertension. *Drug Saf.* **1991**, *6*, 83–93.
- Chow, S.C.; Ki, F. On statistical characteristics of quality of life assessment. *J Biopharm stat.* **1994**, *4*, 1–17.
- Chow, S.C.; Ki, F. Statistical issues in quality of life assessment. *J Biopharm stat.* **1996**, *6*, 37–48.
- Chow, S.C.; Shao, J. *Statistics in Drug Research – Methodologies and Recent Development*; Marcel Dekker, Inc.: New York, 2002.
- Hsiao, C.F.; Tsou, H.H.; Pong, A.; Liu, J.P.; Lin, C.H.; Chang, Y.J.; Chow, S.C. Statistical validation of traditional Chinese medicine diagnostic procedure. *Drug Inf. J.* **2009**, *43*, 83–95.
- Chow, S.C.; Shao, J.; Wang, H. *Sample Size Calculation in Clinical Research*; Marcel Dekker, Inc.: New York, 2002.
- Chow, S.C.; Liu, J.P. *Statistical Design and Analysis in Pharmaceutical Science*; Marcel Dekker, Inc.: New York, 1995.
- Chow, S.C.; Tse, S.K.; Lin, M. Statistical methods in translational medicine. *J. Formosa. Med. Assoc.* **2008**, *107*, S60–S72.
- Enis, P.; Geisser, S. Estimation of the probability that $Y < X$. *J. Am. Stat. Assoc.* **1971**, *66*, 162–168.
- Church, J.D.; Harris, B. The estimation of reliability from stress-strength relationships. *Technometrics* **1970**, *12*, 49–54.
- Tse, S.K.; Chang, J.Y.; Su, W.L.; Chow, S.C.; Hsiung, C.; Lu, Q. Statistical quality control process for traditional Chinese medicine. *J. Biopharm. Stat.* **2006**, *16*, 861–874.
- Lu, Q.; Chow, S.C.; Tse, S.K. Statistical quality control process for traditional Chinese medicine with multiple correlative components. *J. Biopharm. Stat.* **2007**, Vol. 17, 791–808.
- FDA. *Guideline for Submitting Documentation for the Stability of Human Drugs and Biologics*; Center for Drugs and Biologics, Office of Drug Research and Review, Food and Drug Administration: Rockville, MD, 1987.
- ICH Q1A. *Stability Testing of New Drug Substances and Products*. Tripartite International Conference on Harmonization Guideline Q1A, Geneva, 1993.
- Pong, A.; Raghavarao, D. Comparing distributions of drug shelf lives for two components in stability analysis for different designs. *J. Biopharm Stat.* **2002**, *12*, 277–293.
- Chow, S.C.; Shao, J. Stability analysis for drugs with multiple active ingredients. *Stat. Med.* **2007**, *26*, 1512–1517.
- Chow, S.C.; Shao, J. Estimating drug shelf-life with random batches. *Biometrics.* **1991**, *47*, 1071–1079.
- PDR. *Physicians' Desk Reference for Herbal Medicines*; Medical Economics Company: Montvale, NJ, 1998.
- MOPH. *Guidance for Drug Registration*; Ministry of Public Health: Beijing, China, 2002.
- DOH. *Draft Guidance for IND of Traditional Chinese Medicine*; The Department of Health: Taipei, Taiwan, 2004.
- DOH. *Draft Guidance for NDA of Traditional Chinese Medicine*; The Department of Health: Taipei, Taiwan, 2004.
- FDA. *Guidance for Industry – Botanical Drug Products*; The United States Food and Drug Administration: Rockville, MD, 2004.
- FDA. *Guidance for In Vivo Bioequivalence and In Vitro Drug Release*; Center for Drug Evaluation and Research, Food and Drug Administration: Rockville, MD, 1991.
- Liu, J.P.; Chow, S.C. Statistical issues on FDA conjugated estrogen tablets guideline. *Drug Inf. J.* **1996**, *30*, 881–889.
- Chow, S.C.; Liu, J.P. *Design and Analysis of Clinical Trials*; John Wiley & Sons: New York, 2000.
- Chow, S.C.; Pong, A.; Chang, Y.W. On traditional Chinese medicine clinical trials. *Drug Inf. J.* **2006**, *40*, 395–406.
- ICH. *Validation of Analytical Procedures: Methodology*. Tripartite International Conference on Harmonization Guideline, 1996.
- USP/NF. *The United States Pharmacopeia XXIV and the National Formulary XIX*; The United States Pharmacopoeial Convention, Inc.: Rockville, MD, 2000.

Traditional Chinese Medicine (TCM): Unified Approach for Evaluation

Bin Cheng

Department of Biostatistics, Columbia University, New York, New York, U.S.A.

Shein-Chung Chow

Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, U.S.A.

INTRODUCTION

In recent years, as more and more innovative drug products are going off patent protection, the search for new medicines that treat critical and/or life-threatening diseases such as cardiovascular diseases and cancer has become the center of attention of many pharmaceutical companies and research organizations such as the National Institutes of Health (NIH). This leads to the study of the potential use of promising traditional Chinese medicines (TCMs), especially for critical and/or life-threatening diseases. Randomized clinical trial (RCT) was used to assess the effect of Chinese herb medicine in treating the irritable bowel syndrome.^[1] However, RCT is not in common use when studying TCM. There are fundamental differences between Western medicines and TCMs in terms of diagnostic procedures, therapeutic indices, medical mechanism, medical theory, and method practice.^[2–4] Besides, TCM often consists of multiple components with flexible dose.

Chinese doctors believe that all of the organs within a healthy subject should reach the so-called *global dynamic balance and harmony* among organs. Once the global balance is broken at certain sites such as the heart, liver, or kidney, some signs and symptoms will appear to reflect the imbalance at these sites. The collective signs and symptoms are then used to determine what disease the individual has contracted. An experienced Chinese doctor usually assesses the causes of global imbalance before a TCM with flexible doses is prescribed to fix the problem. This approach is sometimes referred to as a personalized (or individualized) medicine approach. In practice, TCMs consider inspection, auscultation and olfaction, interrogation, and pulse taking and palpation as the primary diagnostic procedures. The scientific validity of these subjective and experience-based diagnostic procedures has been criticized due to lack of reference standards and anticipated large evaluator-to-evaluator (i.e., Chinese doctor-to-Chinese doctor) variability. For a systematic discussion of the statistical issues of TCM, see the work of Chow.^[4]

In this article, we attempt to propose a unified approach to developing a composite dissimilarity index based on a

number of indices collected from a treated subject and a healthy control, respectively, under the concept of global dynamic balance among organs. Dynamic balance among organs can be defined as follows: Following the concept of testing bioequivalence or biosimilarity, if the 95% confidence upper bound is less than some health limit, then the conclusion is that the treatment achieves dynamic balance among the organs of the subject; hence, it is considered as efficacious. If we fail to reject the null hypothesis, then the conclusion is that the treatment is not efficacious since there is still a signal of illness, i.e., some of signs and symptoms are still out of the health limit.

METHOD

Dissimilarity Index as an Efficacy Measure

Let $X_T = (X_{T1}, \dots, X_{Tk})$ and $X_H = (X_{H1}, \dots, X_{Hk})$ be the k -dimensional health profiles of a subject who is under treatment of TCM and a subject who is a healthy control, respectively. The components X_{Ti} of a health profile could be either continuous or ordinal, and even if it is continuous, its distribution may not be normal. The dimension of the health profile k is potentially high. Under this formulation, a TCM treatment is considered as efficacious if the health profile of subjects who receive treatment is not significantly different from the one of a healthy subject, possibly age and gender matched if needed. To this end, we define a dissimilarity index θ as a kind of pseudo distance between the profile of a treated subject and the profile of a healthy subject. Specifically, let μ_T and μ_H be the mean health profiles of a treated subject and a healthy subject, respectively, and let Σ_T and Σ_H be the covariance matrices of a treated subject and a healthy subject, respectively. Let $\lambda_1(\Sigma)$ be the largest eigenvalue of a symmetric matrix Σ . Define the dissimilarity index as

$$\theta = \frac{(\mu_T - \mu_H)^T (\mu_T - \mu_H) + \lambda_1(\Sigma_T) - \lambda_1(\Sigma_H)}{\max\{\sigma_0^2, \lambda_1(\Sigma_H)\}}, \quad (1)$$

where σ_0^2 is a known constant.

The above definition of dissimilarity index has an interesting connection with principal component analysis. To see this, imagine that X_T and X_H are multivariate normal. Then, $(\mu_T - \mu_H)^T (\mu_T - \mu_H)$ is the squared mean of the first principal component of $X_T - X_H$, $\lambda_1(\Sigma_T)$ is the variance of the first principal component of X_T , and $\lambda_1(\Sigma_H)$ is the variance of the first principal component of X_H . A small θ implies two things. First, it implies a small difference between the mean health profiles μ_T and μ_H , relative to the maximal variance of a healthy profile X_H . Second, it implies that the maximal variance of the diseased profile X_T is at least not much greater than the maximal variance of the healthy profile X_H . Therefore, a small dissimilarity index can be viewed as evidence for efficacy of a treatment.

The efficacy of a treatment could be formulated as a test for equivalence with the health profile of a healthy control. Let X_T denote the health profile of a diseased subject at the completion of the treatment in Eq. 1. The efficacy is claimed if a test rejects the following H_0 :

$$H_0: \theta \geq \varepsilon \text{ versus } H_1: \theta < \varepsilon, \quad (2)$$

where ε is a known positive threshold.

Assessing Treatment Effect of TCM through Dissimilarity Index

Let

$$\gamma = (\mu_T - \mu_H)^T (\mu_T - \mu_H) + \lambda_1(\Sigma_T) - \lambda_1(\Sigma_H) - \in \max\{\sigma_0^2, \lambda_1(\Sigma_H)\}. \quad (3)$$

Then, testing Eq. 2 is equivalent to testing

$$H_0: \gamma \leq 0 \text{ versus } H_1: \gamma < 0. \quad (4)$$

We construct an asymptotic test via the duality between hypothesis test and confidence interval. Specifically, we construct a 95% approximate confidence upper bound for γ based on two independent random samples $X_{Ti} = (X_{Ti1}, \dots, X_{Tki})^T$, $i = 1, \dots, n_T$, $X_{Hj} = (X_{Hj1}, \dots, X_{Hkj})^T$, $j = 1, \dots, n_H$, and reject the H_0 if this 95% confidence upper bound is smaller than 0.

Let B be a $k \times k$ orthogonal matrix such that

$$B^T(q\Sigma_T + \Sigma_H)B = \text{diag}\{\eta_1, \dots, \eta_k\},$$

where $q = n_H/n_T$. Define

$$v = (v_1, \dots, v_k)^T = B(\mu_T - \mu_H),$$

and similarly define

$$\hat{v} = (\hat{v}_1, \dots, \hat{v}_k) = \hat{B}(\hat{\mu}_T - \hat{\mu}_H),$$

where

$$\hat{B}^T(q\hat{\Sigma}_T + \hat{\Sigma}_H)\hat{B} = \text{diag}\{\hat{\eta}_1, \dots, \hat{\eta}_k\},$$

and $\hat{\mu}_i$ and $\hat{\Sigma}_i$ are the sample mean and the sample covariance, respectively, for $i = T, H$.

Then, parameter γ can be rewritten as

$$\gamma = \sum_{i=1}^k v_i^2 + \lambda_1(\Sigma_T) - \lambda_1(\Sigma_H) - \in \max\{\sigma_0^2, \lambda_1(\Sigma_H)\}.$$

It is easily seen that $\hat{v}_i, i = 1, \dots, k$, are asymptotically independent and normally distributed as $N(\nu_i, \eta_i)$. Let $\hat{\eta}_i$ be the i th diagonal element of $\hat{B}^T(q\hat{\Sigma}_T + \hat{\Sigma}_H)\hat{B}$. Then, an asymptotic 95% confidence upper bound for v_i^2 is

$$\sum_{i=1}^k \hat{v}_i^2 + 2z_{0.05} \sqrt{\frac{\sum_{i=1}^k \hat{v}_i^2 \hat{\eta}_i}{n_H}},$$

where $z_{0.05}$ is the 95%th percentile of the standard normal distribution.

Let $l_{1,T}$ be the largest eigenvalue of $\hat{\Sigma}_T$. Then, an asymptotic 95% confidence upper bound for $\lambda_1(\Sigma_T)$ is^[5]

$$\frac{l_{1,T}}{1 - z_{0.05} \sqrt{2/n_T}}.$$

Similarly, an asymptotic 95% confidence lower bound for $\lambda_1(\Sigma_H)$ is

$$\frac{l_{1,H}}{1 + z_{0.05} \sqrt{2/n_H}}.$$

If $\lambda_1(\Sigma_H) \geq \sigma^2$, then γ in Eq. 3 reduces to

$$\gamma = \sum_{i=1}^k v_i^2 + \lambda_1(\Sigma_T) - (1 + \in)\lambda_1(\Sigma_H). \quad (5)$$

Since $\hat{v}_i$'s, $l_{1,T}$ and $l_{1,H}$ are independent, then using the ideas of Howe^[6] and Graybill and Wang^[7], we construct an approximate 95% confidence upper bound for γ as

$$\hat{\gamma}_{U,1} = \sum_{i=1}^k \hat{v}_i^2 + l_{1,T} - (1 + \in)l_{1,H} + \sqrt{\Delta_1}, \quad (6)$$

where

$$\Delta_1 = \frac{4z_{0.05}^2 \sum_{i=1}^k \hat{v}_i^2 \hat{\eta}_i}{n_H} + \left(\frac{l_{1,T}}{1 - z_{0.05} \sqrt{2/n_T}} - l_{1,T} \right)^2 + (1 + \in)^2 \left(\frac{l_{1,H}}{1 - z_{0.05} \sqrt{2/n_H}} - l_{1,H} \right)^2.$$

If $\lambda_1(\Sigma_H) < \sigma_0^2$, then γ in Eq. 3 reduces to

$$\gamma = \delta + \lambda_1(\Sigma_T) - \lambda_1(\Sigma_H) - \epsilon \sigma_0^2. \quad (7)$$

An approximate 95% confidence upper bound for γ is

$$\hat{\gamma}_{U,2} = \sum_{i=1}^k \hat{v}_i^2 + l_{1,T} - l_{1,H} - \epsilon \sigma_0^2 + \sqrt{\Delta_2}, \quad (8)$$

where

$$\Delta_2 = \frac{4z_{0.05}^2 \sum_{i=1}^k v_i^2 \hat{\eta}_i}{n_H} + \left(\frac{l_{1,T}}{1 - z_{0.05} \sqrt{2/n_T}} - l_{1,T} \right)^2 + \left(\frac{l_{1,H}}{1 - z_{0.05} \sqrt{2/n_H}} - l_{1,H} \right)^2.$$

For testing Eq. 2, it is equivalent to test

$$\gamma \leq 0, \text{ versus } H_1: \gamma < 0. \quad (9)$$

We propose the following testing rule: Reject H_0 if $l_{1,H} \geq \sigma_0^2$ and $\hat{\gamma}_{U,1} < 0$, or $l_{1,H} < \sigma_0^2$ and $\hat{\gamma}_{U,2} < 0$.

Specifically, we claim that the testing rule

$$\phi = I(\hat{\gamma}_{U,1} < 0, l_{1,H} \geq \sigma_0^2) + I(\hat{\gamma}_{U,2} < 0, l_{1,H} < \sigma_0^2) \quad (10)$$

has an approximate level of 0.05 for testing H_0 .

When the sample size is small, this test ϕ tends to be aggressive in rejecting the null hypothesis. We recommend replacing $z_{0.05}$ by $z_{0.025}$ in practice unless the sample size is very large.

Sample Size Determination

From the derivation of the test, we can show that if $\lambda_1(\Sigma_H) \geq \sigma_0^2$, as $n_H \rightarrow \infty$,

$$\sqrt{n_H}(\hat{\gamma} - \gamma) \rightarrow_d N\left(0, 4 \sum_{i=1}^k v_i^2 \eta_i^2 + 2q\lambda_1^2(\Sigma_T) + 2(1+\epsilon)^2 \lambda_1^2(\Sigma_H)\right),$$

as $n_H \rightarrow \infty$, and if $\lambda_1(\Sigma_H) < \sigma_0^2$,

$$\sqrt{n_H}(\hat{\gamma} - \gamma) \rightarrow_d N\left(0, 4 \sum_{i=1}^k v_i^2 \eta_i^2 + 2q\lambda_1^2(\Sigma_T) + 2\lambda_1^2(\Sigma_H)\right),$$

as $n_H \rightarrow \infty$.

Therefore, for any fixed alternative value $\gamma_* < 0$ in Eq. 9, to achieve a power of $1 - \beta$ while controlling the type one error rate as α , the sample size needed for the control group is

$$n_H = \left(z_\alpha + z_\beta\right)^2 \sigma^2 / \gamma_*^2,$$

where

$$\sigma^2 = 4 \sum_{i=1}^k v_i^2 \eta_i^2 + 2q\lambda_1^2(\Sigma_T) + 2[1 + \epsilon I(\lambda_1(\Sigma_H) \geq \sigma_0^2)] \lambda_1^2(\Sigma_H).$$

SIMULATION

In our simulation, we set $k=10$, and since γ depends on μ_T and μ_H only through their difference $\mu_T - \mu_H$, we assume

$$\mu_T = (a, ar, ar^2, \dots, ar^9), \quad \mu_H = (0, 0, 0, \dots, 0),$$

where $a \neq 0$, $r \geq 0$. For the covariance matrix, we set

$$\Sigma_T = \begin{pmatrix} \sigma_1 I_5 & 0 \\ 0 & \sigma_2 I_5 \end{pmatrix} \cdot R_T \cdot \begin{pmatrix} \sigma_1 I_5 & 0 \\ 0 & \sigma_2 I_5 \end{pmatrix},$$

where I_5 is a 5×5 identity matrix, σ_1 is the standard deviation of the first five components of X_T , σ_2 is the standard deviation of the last five components of X_T , and R_T is the 10×10 correlation matrix. Similarly, we set

$$\Sigma_H = \begin{pmatrix} \tau_1 I_5 & 0 \\ 0 & \tau_2 I_5 \end{pmatrix} \cdot R_H \cdot \begin{pmatrix} \tau_1 I_5 & 0 \\ 0 & \tau_2 I_5 \end{pmatrix},$$

where τ_1 is the standard deviation of the first five components of X_H , τ_2 is the standard deviation of the last five components of X_H , and R_H is the 10×10 correlation matrix.

For illustration, we choose $\epsilon=2.0$, $\sigma_0=0.5$. Two types of correlation patterns are considered: component symmetric (CS) correlation and the autoregressive of order 1 (AR(1)) correlation, where the parameters for R_T and R_H are ρ_T and ρ_H , respectively. All estimations are based on 10,000 simulation runs and $z_{0.025}$ is used in the test.

The simulated sizes of the test under CS and AR(1) correlations are summarized in Tables 1 and 2, respectively. The size is controlled in all the CS scenarios but slightly inflated in some AR(1) scenarios.

One possible reason is that the estimations of a 10×10 covariance matrix and its largest eigenvalue λ_1 are not accurate enough in these cases when the sample size is as small as 120.

The simulated powers of the test under CS and AR(1) correlations are summarized in Tables 3 and 4, respectively. A balance design, i.e., $n_T = n_H$, typically yields a higher power than the case where more healthy controls are used, i.e., $n_H > n_T$. The simulation indicates that for an effect size of $\gamma = -2.00$, a sample size of 120 in each group would be adequate to achieve 80% power.

Table 1 Estimated sizes based on 10,000 simulation runs, assuming $\varepsilon=2.0$, $\sigma_0=0.5$, and compound symmetric correlation

(a, r)	$(\sigma_1, \sigma_2, \rho_1)$	(τ_1, τ_2, ρ_2)	γ	(n_T, n_H)	Size
(1.450, 0.245)	(1.10, 1.00, 0.10)	(1.00, 1.00, 0.05)	0.00	(50, 100)	0.0091
(1.450, 0.245)	(1.10, 1.00, 0.10)	(1.00, 1.00, 0.05)	0.00	(75, 75)	0.0422
(1.450, 0.245)	(1.10, 1.00, 0.10)	(1.00, 1.00, 0.05)	0.00	(80, 160)	0.0168
(1.450, 0.245)	(1.10, 1.00, 0.10)	(1.00, 1.00, 0.05)	0.00	(120, 120)	0.0435
(0.000, 0.000)	(1.71, 1.00, 0.10)	(1.00, 1.00, 0.05)	0.00	(50, 100)	0.0005
(0.000, 0.000)	(1.71, 1.00, 0.10)	(1.00, 1.00, 0.05)	0.00	(75, 75)	0.0057
(0.000, 0.000)	(1.71, 1.00, 0.10)	(1.00, 1.00, 0.05)	0.00	(80, 160)	0.0015
(0.000, 0.000)	(1.71, 1.00, 0.10)	(1.00, 1.00, 0.05)	0.00	(120, 120)	0.0125

Table 2 Estimated sizes based on 10,000 simulation runs, assuming $\varepsilon=2.0$, $\sigma_0=0.5$, and AR(1) correlation

(a, r)	$(\sigma_1, \sigma_2, \rho_1)$	(τ_1, τ_2, ρ_2)	γ	(n_T, n_H)	Size
(1.325, 0.240)	(1.10, 1.00, 0.100)	(1.00, 1.00, 0.050)	0.00	(50, 100)	0.0162
(1.325, 0.240)	(1.10, 1.00, 0.100)	(1.00, 1.00, 0.050)	0.00	(75, 75)	0.1186
(1.325, 0.240)	(1.10, 1.00, 0.100)	(1.00, 1.00, 0.050)	0.00	(80, 160)	0.0343
(1.325, 0.240)	(1.10, 1.00, 0.100)	(1.00, 1.00, 0.050)	0.00	(120, 120)	0.1517
(0.000, 0.000)	(1.67, 1.00, 0.100)	(1.00, 1.00, 0.050)	0.00	(50, 100)	0.0000
(0.000, 0.000)	(1.67, 1.00, 0.100)	(1.00, 1.00, 0.050)	0.00	(75, 75)	0.0063
(0.000, 0.000)	(1.67, 1.00, 0.100)	(1.00, 1.00, 0.050)	0.01	(80, 160)	0.0002
(0.000, 0.000)	(1.67, 1.00, 0.100)	(1.00, 1.00, 0.050)	0.01	(120, 120)	0.0173

Table 3 Estimated powers based on 10,000 simulation runs, assuming $\varepsilon=2.0$, $\sigma_0=0.5$, and compound symmetric correlation

(a, r)	$(\sigma_1, \sigma_2, \rho_1)$	(τ_1, τ_2, ρ_2)	γ	(n_T, n_H)	Power
(0.300, 0.080)	(1.10, 1.00, 0.100)	(1.00, 1.00, 0.050)	-1.99	(50, 100)	0.5331
(0.300, 0.080)	(1.10, 1.00, 0.10)	(1.00, 1.00, 0.050)	-1.99	(75, 75)	0.8500
(0.300, 0.080)	(1.10, 1.00, 0.100)	(1.00, 1.00, 0.050)	-1.99	(80, 160)	0.8445
(0.300, 0.080)	(1.10, 1.00, 0.100)	(1.00, 1.00, 0.050)	-1.99	(120, 120)	0.9650
(0.000, 0.000)	(1.19, 1.00, 0.100)	(1.00, 1.00, 0.050)	-2.00	(50, 100)	0.4591
(0.000, 0.000)	(1.19, 1.00, 0.100)	(1.00, 1.00, 0.050)	-2.00	(75, 75)	0.8300
(0.000, 0.000)	(1.19, 1.00, 0.100)	(1.00, 1.00, 0.050)	-2.00	(80, 160)	0.8125
(0.000, 0.000)	(1.19, 1.00, 0.100)	(1.00, 1.00, 0.050)	-2.00	(120, 120)	0.9613

Table 4 Estimated powers based on 10,000 simulation runs, assuming $\varepsilon=2.0$, $\sigma_0=0.5$, and AR(1) correlation

(a, r)	$(\sigma_1, \sigma_2, \rho_1)$	(τ_1, τ_2, ρ_2)	γ	(n_T, n_H)	Power
(0.000, 0.080)	(1.10, 1.00, 0.100)	(1.00, 1.00, 0.080)	-1.99	(50, 100)	0.8781
(0.000, 0.080)	(1.10, 1.00, 0.100)	(1.00, 1.00, 0.080)	-1.99	(75, 75)	0.9972
(0.000, 0.080)	(1.10, 1.00, 0.100)	(1.00, 1.00, 0.080)	-1.99	(80, 160)	0.9979
(0.000, 0.080)	(1.10, 1.00, 0.100)	(1.00, 1.00, 0.080)	-1.99	(120, 120)	0.9999
(0.000, 0.000)	(1.00, 1.00, 0.100)	(1.00, 1.00, 0.035)	-2.00	(50, 100)	0.9714
(0.000, 0.000)	(1.00, 1.00, 0.100)	(1.00, 1.00, 0.035)	-2.00	(75, 75)	0.9998
(0.000, 0.000)	(1.00, 1.00, 0.100)	(1.00, 1.00, 0.035)	-2.00	(80, 160)	0.9999
(0.000, 0.000)	(1.00, 1.00, 0.100)	(1.00, 1.00, 0.035)	-2.00	(120, 120)	0.9999

DISCUSSION

In this article, we assume that the health profile has a multivariate normal distribution. When part or all the components of the health profile are not normally distributed, the proposed method will be valid asymptotically when $\min\{n_T, n_H\} \rightarrow \infty$.

In this article, we consider the case that k is moderately large. When k is extremely large, i.e., when $k \gg \max\{n_T, n_H\}$, the methods for sparse principal component analysis, such as those of Zou et al.^[8] or Johnstone and Lu,^[9] should be used to estimate the covariance matrices, and the corresponding 95% confidence upper bound could be constructed using the bootstrap method.

The dissimilarity index defined in this article measures the closeness between a profile of a treated subject and a file of a healthy control using both mean and variance information. However, we did not model the causal relationship among the k components $X_1, \dots, X_k$. Per TCM theory, an alternative model incorporating causal information could be considered. Future research is thus warranted.

The approach we adopt is a population bioequivalence approach. It is therefore most relevant when the main purpose is to establish the treatment effect of a TCM at the population level. It is of interest to develop methods based on individual bioequivalence. Specifically, let X_H and X_T be the health profiles from the *same* subject where the former is the profile when the subject is known in a healthy condition, whereas the latter is the one after the subject received TCM for his sickness. Let μ_i and Σ_i be the mean and covariance of the profile when the subject is at status $i=H,T$, respectively, and let Σ_{TH} be the covariance between X_T and X_H due to the same subject. Then, we could define an illness index as

$$\vartheta = \frac{(\mu_T - \mu_H)^T (\mu_T - \mu_H) + \lambda_1(\Sigma_T) - \lambda_1(\Sigma_H) - 2\lambda_1(\Sigma_{TH})}{\max\{\sigma_0^2, \lambda_1(\Sigma_H)\}}$$

and consider testing

$$H_0 : \vartheta \geq \varepsilon \text{ versus } H_1 : \vartheta < \varepsilon,$$

where $\varepsilon > 0$ is a known health limit. Here, rejecting H_1 implies the efficacy of the TCM. The derivation of the test is more involved and hence will be constructed in a separate paper.

REFERENCES

1. Bensoussan, A.; Talley, N.J.; Hing, M.; Menzies, R.; Guo, A.; Ngu, M. Treatment of irritable bowel syndrome with Chinese herb medicine: A randomized controlled trial. *JAMA* **1998**, *280*, 1585–1589.
2. Chow, S.C.; Pong, A.; Chang, Y.W. On traditional Chinese medicine clinical trials. *Drug Inf. J.* **2006**, *40*, 395–406.
3. Zhou, X.H.; Li, S.L.; Tian, F.; Xie, Y.M.; Pei, Y.; Kang, S.; Fan, M.; Li, J.P. Building a disease risk model of osteoporosis based on traditional Chinese medicine symptoms and western medicine risk factors. *Stat. Med.* **2012**, *31*, 643–652.
4. Chow, S.C. *Quantitative Methods for Traditional Chinese Medicine*; Chapman and Hall/CRC Press, Taylor & Francis: New York, 2015.
5. Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*, 3rd Ed.; Wiley: New York, 2003.
6. Howe, W.G. Approximate confidence limits on the mean of $X+Y$ where X and Y are two tabled independent random variables. *J. Am. Stat. Assoc.* **1974**, *69*, 789–794.
7. Graybill, F.; Wang, C.M. Confidence intervals on nonnegative linear combinations of variances. *J. Am. Stat. Assoc.* **1980**, *75*, 869–873.
8. Zou, H.; Hastie, T.; Tibshirani, R. Sparse principal component analysis. *J. Comput. Graph. Stat.* **2006**, *15*, 265–286.
9. Johnstone, I.M.; Lu, A.Y. On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Stat. Assoc.* **2009**, *104*, 682–693.

Translational Medicine: Concepts, Statistical Methods, and Related Issues

Siu-Keung Tse

City University of Hong Kong, Hong Kong, China

Shein-Chung Chow

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Abstract

Translational Medicine: Concepts, Statistical Methods, and Related Issues This entry focuses on statistical methods commonly employed in translational medicine, which is a multi-disciplinary entity that bridges basic scientific research with clinical development.

INTRODUCTION

In early 2000, the U.S. Food and Drug Administration (FDA) kicked off *Critical Path Initiative* to assist the sponsors to identify possible causes of the scientific challenges underlying the medical product pipeline problems. The Critical Path Opportunities List released by the FDA on March 16, 2006, identified (i) better evaluation tools and (ii) streamlining clinical trials as the top two topic areas to bridge the gap between the quick pace of new biomedical discoveries and the slower pace at which those discoveries are currently developed into therapies. This has led to the consideration of the use of adaptive design methods in clinical development and the focus of translational medicine (or science/research), which attempt to not only to identify best clinical benefit of a drug product under investigation but also to increase the probability of success. Statistical methods for the use of adaptive trial designs in clinical development can be found in Chow and Chang^[1] and Chang.^[2] In this article, however, we will focus on statistical methods that are commonly employed in translational medicine.

Chow^[3] classified translational medicine into three areas, namely, translation in language, translation in information, and translation in (medical) technology. Translation in language is referred to possible lost in translation of informed consent form and/or case report forms in multinational clinical trials. Lost in translation is commonly encountered due not only to differences in language, but also differences in perception, culture, medical practices, and so forth. A typical approach for assessment of the possible lost in translation is to first translate the informed consent form and/or the case report forms by an experienced expert and then translate back by a different experienced but independent expert. The back-translated version is then compared with the original version for consistency. If the

back-translated version passes the test for consistency, then the back-translated version is validated through a small-scale pilot study before it is applied to the intended multinational clinical trial. Translation in information is referred to as bench-to-bedside in translational science/research, which is also known as translational medicine. Translation in technology includes biomarker development and translation in diagnostic procedures between traditional Chinese medicine and Western medicine. In this article, we will focus on statistical methods for translation in information and translation in technology. Note that in practice, translational medicine is often divided into two areas, namely, discovery translational medicine and clinical translational medicine. Discovery translational medicine is referred to biomarker development, bench-to-bedside, and animal model versus human model, while clinical translational medicine includes translation among study endpoints, translation in technology, and the generalization from a target patient population to another.

In the next section, statistical methods for biomarker development, especially for optimal variable screening in microarray analysis, are outlined. Also included in this section is a cross-validation method for model selection and validation. The section “Bench-to-Bedside” discusses the ideas of a one-way and two-way translation process in pharmaceutical/clinical development. Whether or not an established animal model is predictive of a human model is examined in the section “Animal Model Versus Human Model.” The section “Translation in Study Endpoints” summarizes translation among different study endpoints. Issues that are commonly encountered in translation in technology are described in the section “Bridging Studies.” Also included in this section is the generalization of results obtained from a target patient population to another similar but different target patient population. Some concluding remarks are provided in the last section of this article.

BIOMARKER DEVELOPMENT

A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic process, or pharmacologic responses to a therapeutic intervention. Biomarkers can be classified as classifier markers, prognostic markers, and predictive markers. A classifier marker usually does not change over the course of study and can be used to identify patient population who would benefit from the treatment from those do not. A typical example is a DNA marker for population selection in the enrichment process of clinical trials. A prognostic marker informs the clinical outcomes, which is independent of treatment. A predictive marker informs the treatment effect on the clinical endpoint, which could be population-specific. That is, a predictive marker could be predictive for population A but not population B. It should be noted that the correlation between biomarker and true endpoint constitutes a prognostic marker. However, correlation between biomarker and true endpoint does not constitute a predictive biomarker.

In clinical development, a biomarker could be used to select the right population, to identify the natural course of disease, for early detection of disease, and to help to develop personalized medicine. The utilization of biomarkers could lead to better target populations, detection of a large effect size with smaller sample size, and timely decision-making. As indicated in the FDA Critical Path Initiative Opportunity List, better evaluation tools call for biomarker qualification and standards. Statistical methods for early stage biomarker qualification include, but are not limited to, (i) distance-dependent K-nearest neighbors (DD-KNN), (ii) K means clustering, (iii) single/average/complete linkage clustering, and (iv) distance-dependent Jarvis-Patrick clustering. More information can be found at http://www.ncicrf.gov/human_studies.shtml.

In what follows, we will review statistical methods that are commonly used in biomarker development for optimal variable screening. The selected variables will then be used to establish a predictive model through a model selection/validation process.

Optimal Variable Screening

DNA microarrays have been used extensively in medicinal practice. Microarrays identify a set of candidate genes that are possibly related to a clinical outcome of a disease (in disease diagnoses) or a medical treatment. However, there are many more candidate genes than the number of available samples (the sample size) in almost all studies, which leads to an irregular statistical problem in disease diagnoses or treatment outcome prediction. Some available statistical methods deal with a single gene at a time,^[4] which clearly do not provide the best solution for polygenic diseases. In practice, a meta-analysis and/or combining several similar studies is often considered to increase sample

size. These approaches, however, may not be appropriate due to the fact that (i) the combined dataset may still be much too small, and (ii) there may be heteroscedasticity among the data from different studies. Alternatively, Shao and Chow^[5] propose an optimal variable screening approach for dealing with the situation where the number of variables (genes) is much larger than the sample size.

Let y be a clinical outcome of interest and x be a vector of p candidate genes that are possibly related to y . Shao and Chow^[5] simply considered inference on the population of y conditional on x and noted that their proposed method can be applied to the unconditional analysis (i.e., both y and x are random). Consider the following model:

$$y = \beta'x + \varepsilon, \quad (1)$$

where β is a p -dimensional vector and the distribution of ε is independent of x with $E(\varepsilon) = 0$ and $E(\varepsilon^2) = \sigma^2$. Under the model (1), assume that there is a positive integer p_0 (which does not depend on n) such that only p_0 components of β are nonzero. Furthermore, β is in the linear space generated by the rows of $X'X$ for sufficiently large n , where x is the $n \times p_n$ matrix whose i th row is x'_i . In addition, assume that there is a sequence $\{\xi_n\}$ of positive numbers such that $\xi_n \rightarrow \infty$ and $\lambda_{in} = b_i \xi_n$, where λ_{in} is the i th nonzero eigenvalue of $X'X$, $i = 1, \dots, n$ and $\{b_i\}$ is a sequence of bounded positive numbers. Note that in many problems, $\xi_n = n$. Furthermore, there exists a constant $c > 0$ such that $p_n/\xi_n^c \rightarrow 0$. For the estimation of β , Shao and Chow^[5] considered the following ridge regression estimator.

$$\hat{\beta} = (X'X + h_n I_{p_n})^{-1} X'Y, \quad (2)$$

where $Y = (y_1, \dots, y_n)'$, I_{p_n} is the identity matrix of order p_n and $h_n > 0$ is the ridge parameter. The bias and variance of $\hat{\beta}$ are given by

$$\text{bias}(\hat{\beta}) = E(\hat{\beta}) - \beta = -(h_n^{-1} X'X + I_{p_n})^{-1} \beta$$

and

$$\text{var}(\hat{\beta}) = \sigma^2 (X'X + h_n I_{p_n})^{-1} X'X (X'X + h_n I_{p_n})^{-1}$$

Let β_i and $\hat{\beta}_i$ be the i th component of β and $\hat{\beta}$, respectively. Under the assumptions as described earlier, we have $E(\hat{\beta}_i - \beta_i)^2 \rightarrow 0$ (i.e., $\hat{\beta}_i$ is consistent for β_i in mean squared error) if h_n is suitably chosen. Thus, we have

$$\text{var}(\hat{\beta}) = \frac{\sigma^2}{h_n} \left(\frac{X'X}{h_n} + I_{p_n} \right)^{-1} \frac{X'X}{h_n} \left(\frac{X'X}{h_n} + I_{p_n} \right)^{-1}. \quad (3)$$

Hence, $\text{var}(\hat{\beta}_i) \rightarrow 0$ for all i as long as $h_n \rightarrow \infty$. Note that the analysis of the bias of $\hat{\beta}_i$ is more complicated. Let Γ be an orthogonal matrix such that

$$\Gamma'X'X\Gamma = \begin{pmatrix} \Lambda_n & 0_{n \times (n-p_n)} \\ 0_{(p_n-n) \times n} & 0_{(p_n-n) \times (p_n-n)} \end{pmatrix},$$

where Λ_n is a diagonal matrix whose i th diagonal element is λ_{in} and $0_{l \times k}$ is the $l \times k$ matrix of 0's. Then, it follows that

$$\text{bias}(\hat{\beta}) = - \left[\Gamma \left(\frac{\Gamma'X'X\Gamma}{h_n} + I_{p_n} \right) \Gamma' \right]^{-1} \beta = -\Gamma A \Gamma' \beta, \quad (4)$$

where A is a $p_n \times p_n$ diagonal matrix whose first n diagonal elements are $h_n/h_n + \lambda_{in}, i = 1, \dots, n$, and the last diagonal elements are all equal to 1. Under the above mentioned assumptions, combining the results for variance and bias of $\hat{\beta}_i$, i.e., Eqs. (3) and (4), it can be shown that for all i

$$E(\hat{\beta}_i - \beta_i)^2 = \text{var}(\hat{\beta}_i) + [\text{bias}(\hat{\beta}_i)]^2 \rightarrow 0$$

if h_n is chosen so that $h_n \rightarrow \infty$ at a rate slower than ξ_n (e.g., $h_n = \xi_n^{2/3}$). Based on this result, Shao and Chow^[5] proposed the following optimal variable screening procedure:^[5]

Let $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \rightarrow 0$. For each fixed n , we screen out the i th variable if and only if $|\hat{\beta}_i| \leq a_n$.

Note that, after screening, only variables associated with $|\hat{\beta}_i| > a_n$ are retained in the model as predictors. The idea behind this variable screening procedure is similar to that in the Lasso method.^[6] Under certain conditions, Shao and Chow^[5] showed that their proposed optimal variable screen method is consistent in the sense that the probability that all variables (genes) unrelated to y , which will be screened out, and all variables (genes) related to y , which will be retained is 1 as n tends to infinity.

Model Selection and Validation

Suppose that n data points are available for selecting a model from a class of models. Several methods for model selection are available in the literature. These methods include, but are not limited to, Akaike information criterion (AIC),^[7,8] the C_p ,^[9] and the jackknife and the bootstrap.^[10] These methods, however, are not asymptotically consistent in the sense that the probability of selecting the model with the best predictive ability does not converge to 1 as the total number of observations $n \rightarrow \infty$. Alternatively, Shao^[11] proposed a method for model selection and validation using the method of cross-validation. The idea of cross-validation is to split the dataset into two parts. The first part contains n_c data points which will be used for fitting a model (model construction), whereas the second part contains $n_v = n - n_c$ data points that are reserved for assessing the predictive ability of the model (model validation). It

should be noted that all of the $n = n_c + n_v$ data, not just n_v are used for model validation. Shao^[11] showed that all of the methods of AIC, C_p , Jackknife and bootstrap are asymptotically equivalent to the cross-validation with $n_v = 1$, denoted by CV(1), although they share the same deficiency of inconsistency. Shao^[11] indicated that the inconsistency of the leave-one-out cross-validation can be rectified by using leave- n_v -out cross-validation with n_v satisfying $n_v/n \rightarrow 1$ as $n \rightarrow \infty$.

In addition to the cross-validation with $n_v = 1$, denoted by CV(1), Shao^[11] also considered the other two cross-validation methods, namely, a Monte Carlo cross-validation with $n_v (n_v \neq 1)$, denoted by MCCV(n_v), and an analytic approximate CV(n_v), denoted by APCV(n_v). MCCV(n_v) is a simple and easy method utilizing the method of Monte Carlo by randomly drawing (with or without replacement) a collection $\mathfrak{R}$ of b subsets of $\{1, \dots, n\}$ that have size n_v and select a model by minimizing

$$\hat{\Gamma}_{\alpha, n} = \frac{1}{n_v b} \sum_{s \in \mathfrak{R}} \|y_s - \hat{y}_{\alpha, s^c}\|^2$$

On the other hand, APCV(n_v) selects the optimal model based on the asymptotic leading term of balance incomplete CV(n_v), which treats each subset as a block and each i as a treatment. Shao^[11] compared these three cross-validation methods through a simulation study under the following model with five variables with $n = 40$:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + e_i,$$

where are i.i.d. from $N(0,1)$, x_{ki} is the i th value of the k th prediction variable x_k , $x_{1i} = 1$, and the values of x_{ki} , $k = 2, \dots, 5$, $i = 1, \dots, 40$, are taken from an example in Gunst and Mason^[12]. Note that there are 31 possible models, and each model was denoted by a subset of $\{1, \dots, 5\}$ that contains the indices of the variable x_k in the model. Shao^[11] indicated that MCCV(n_v) has the best performance among the three methods under study except for the case where the largest model is the optimal model. APCV(n_v) is slightly better than the CV(1) in all cases. CV(1) tends to select unnecessarily large models. The probability of selecting the optimal model by using the CV(1) could be very low (e.g., less than 0.5).

Remarks

In practice, it is suggested that the optimal variable screening method proposed by Shao and Chow^[5] be applied to select a few relevant variables say 5–10 variables. Then, apply the cross-validation method to select the optimal model based on linear model selection (Shao, 1993) or non-linear model selection.^[13] The selected model can then be validated based on the cross-validation methods as described in the previous subsection.

BENCH-TO-BEDSIDE

Pizzo^[14] defines translational medicine as *bench-to-bedside* research wherein a basic laboratory discovery becomes applicable to the diagnosis, treatment, or prevention of a specific disease and is brought forth by either a physician-scientist who works at the interface between the research laboratory and patient care or by a team of basic and clinical science investigators. Thus, translational medicine is referred to the translation of basic research discoveries into clinical applications. More specifically, translational medicine is to take the discoveries from basic research to a patient and measures an endpoint in a patient. Most recently, scientists are increasingly aware that this bench-to-bedside approach to translational research is a two-way street. Basic scientists provide clinicians with new tools for use in patients and for assessment of their impact, and clinical researchers make novel observations about the nature and progression of disease that often stimulate basic investigations. As indicated by Pizzo,^[14] translational medicine can also have a much broader definition, referring to the development and application of new technologies, biomedical devices, and therapies in patient-driven environments, such as clinical trials, where the emphasis is on early patient testing and evaluation. Thus, translational medicine also includes epidemiological and health-outcomes research and behavioral studies that can be brought to the bedside or ambulatory setting.

Mankoff et al.^[15] pointed out that there are three major obstacles to effective translational medicine in practice. The first is the challenge of translating basic science discoveries into clinical studies. The second hurdle is the translation of clinical studies into medical practice and health care policy. A third obstacle to effective translational medicine is philosophical. It may be a mistake to think that basic science (without observations from the clinic and without epidemiological findings of possible associations between different diseases) will efficiently produce the novel therapies for human testing. Pilot studies such as non-human and non-clinical studies are often used to transition therapies developed using animal models to a clinical setting. Statistical process plays an important role in translational medicine. In particular, we define a statistical process of translational medicine as a translational process for (i) determining association between some independent parameters observed in basic research discoveries and dependent variable observed from clinical application, (ii) establishing a predictive model between the independent parameters and the dependent response variable, and (iii) validating the established predictive model. For example, in animal studies, the independent variables may include *invitro* assay results, pharmacological activities such as pharmacokinetics and pharmacodynamics, and dose toxicities and the dependent variable could be a clinical outcome (e.g., a safety parameter). Statistical methods related

to (i) and (ii) are well documented in statistical literature. However, research effort is needed to develop appropriate criteria to assess the accuracy of a translation process.

If the information observed in basic research discoveries is translated to clinical practice, it is usually referred to as a one-way translation process. However, as indicated by Pizzo,^[14] the translational process should be a two-way translation, where information obtained in clinical study could also be translated back to basic research discoveries. Tse^[16] described the statistical process of how to validate a translational process, either one-way or two-way, based on some probability-based criteria. Furthermore, a measure to assess the degree of lost in translation was also proposed.

The following sections describe basic statistical concepts and statistical tests for one-way and two-way translation between basic research discoveries and clinical outcome under a well-established and validated predictive model. In particular, two different criteria and the corresponding statistical tests for one-way translation are discussed. Furthermore, a procedure for the evaluation of two-way translation and the statistical assessment of the lost in translation is briefly discussed.

One-Way Translation

Let x and y be the observed values from basic research discoveries and clinical application, respectively. In practice, it is important to ensure that the translational process is accurate and reliable with some statistical assurance. One of the statistical criteria is to examine the closeness between the observed response y and the predicted response $\hat{y}$ via a translational process. To study this, we will first study the association between x and y and build up a model. Then, we will validate the model based on some criteria. For simplicity, we assume that x and y can be described by the following linear model

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad (5)$$

where ε follows a normal distribution with mean 0 and variance σ_e^2 .

Suppose that n pairs of observations $(x_1, y_1), \dots, (x_n, y_n)$ are observed in a translation process. To define notation, let $X^T = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix}$ and $Y^T = (y_1 \ y_2 \ \dots \ y_n)$.

Then, under model (1), the maximum likelihood estimates (MLE) of the parameters β_0 and β_1 are given as

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X^T X)^{-1} X^T Y \text{ with } \text{var} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X^T X)^{-1} \sigma_e^2. \text{ Thus, the}$$

following relationship

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (6)$$

can be established between x and y , the observed values from basic research discoveries and clinical application respectively.

Given x_i from Eq. (6), the corresponding fitted value $\hat{y}_i$ is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Furthermore, the corresponding MLE of σ_e^2 is $\hat{\sigma}_e^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{n-2}{n}$ MSE, where MSE is the mean squared errors of the fitted model.

For a given $x = x_0$, suppose that the corresponding observed value is given by y ; using (6), the corresponding fitted value is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_0$. Note that $E(\hat{y}) = \beta_0 + \beta_1 x_0 = \mu_0$ and

$$\text{Var}(\hat{y}) = (1 \quad x_0) (X^T X)^{-1} \begin{pmatrix} 1 \\ x_0 \end{pmatrix} \sigma_e^2 = c \sigma_e^2,$$

where $c = (1 \quad x_0) (X^T X)^{-1} \begin{pmatrix} 1 \\ x_0 \end{pmatrix}$. Furthermore, $\hat{y}$ is normally distributed with mean μ_0 and variance $c \sigma_e^2$, i.e., $\hat{y} \sim N(\mu_0, c \sigma_e^2)$.

Measures of Closeness

We may validate the translation model by considering how close an observed y and its predicted value obtained based on the fitted regression model given by (6). To assess the closeness, we propose the following two measures, which are based either on the absolute difference or the relative difference between y and $\hat{y}$:

$$\text{Criterion I.} \quad p_1 = P\{|y - \hat{y}| < \delta\},$$

$$\text{Criterion II.} \quad p_2 = P\left\{\left|\frac{y - \hat{y}}{y}\right| < \delta\right\}.$$

In other words, it is desirable to have a high probability that the difference or the relative difference between y and $\hat{y}$, given by p_1 and p_2 respectively, is less than a clinically or scientifically meaningful difference δ . Then, for either $i = 1$ or 2 , it is of interest to test the following hypotheses

$$H_0: p_i \leq p_0 \quad \text{vs.} \quad H_a: p_i > p_0, \quad (7)$$

where p_0 is some pre-specified constant. If the conclusion is to reject H_0 in favor of H_a , this would imply that the established model is considered validated. The steps in carrying out the test of hypothesis corresponding to the two criteria are outlined in the following section.

Measure of closeness based on absolute difference It is easy to show that $p_1 = \Phi(\delta/\sqrt{(1+c)\sigma_e^2}) - \Phi(-\delta/\sqrt{(1+c)\sigma_e^2})$, where Φ is the cumulative density function of a standard normal distribution. Thus, the MLE of p_1 is given by $\hat{p}_1 = \Phi(\delta/\sqrt{(1+c)\hat{\sigma}_e^2}) - \Phi(-\delta/\sqrt{(1+c)\hat{\sigma}_e^2})$. Define $V_1 = (2\delta^2/(1+c)n\sigma_e^2) \phi^2(\delta/\sqrt{(1+c)\hat{\sigma}_e^2})$, where $\phi(z)$ is the probability density function of a standard normal distribution. For a sufficiently large sample size n , using Slutsky Theorem, $\hat{p}_1 - p_0/\sqrt{V_1}$ can be approximated by a standard

normal distribution. Based on these results, for the testing of the hypotheses,

$$H_0: p_1 \leq p_0 \quad \text{vs.} \quad H_a: p_1 > p_0,$$

H_0 is rejected if $\hat{p}_1 - p_0/\sqrt{V_1} > z_{1-\alpha}$, where $z_{1-\alpha}$ is the $100(1-\alpha)^{\text{th}}$ percentile of a standard normal distribution.

Measure of closeness based on the absolute relative difference Recall that the measure of closeness based on the absolute relative difference is given by $p_2 = P\{|y - \hat{y}|/y < \delta\}$, where p_2 is determined by a non-central F distribution and can be approximated by a central F distribution. In particular, it can be shown that $p_2 \approx P\{((1-\delta)^2/c)(1+\lambda_2/1+\lambda_1 < F_{v,v'} < (1+\lambda_2/1+\lambda_1)(1+\delta)^2/c\}$, where $\lambda_1 = \mu_0^2/c\sigma_e^2$ and $\lambda_2 = \mu_0^2/c\sigma_e^2$ and $f_{v,v'}$ is a central F distribution with degrees of freedom $v = (v_1 + \lambda_1)2/v_1 + 2\lambda_1 = (1 + \mu_0^2/c\sigma_e^2)^2 / (1 + 2\mu_0^2/c\sigma_e^2)$ and $v' = (v_1 + \lambda_1)^2/v_2 + 2\lambda_2 = (1 + \mu_0^2/c\sigma_e^2)^2 / (1 + 2\mu_0^2/c\sigma_e^2)$. Thus, p_2 can be estimated by

$$\begin{aligned} \hat{p}_2 &= P\left\{\frac{(1-\delta)^2}{c} \frac{1+\hat{\lambda}_2}{1+\hat{\lambda}_1} < F_{\hat{v},\hat{v}'} < \frac{1+\hat{\lambda}_2}{1+\hat{\lambda}_1} \frac{(1+\delta)^2}{c}\right\} \\ &= P\{u_1 < F_{\hat{v},\hat{v}'} < u_2\}; \end{aligned}$$

where $u_1 = (1 + \hat{\lambda}_2/c(1 + \hat{\lambda}_1))(1 - \delta)^2$, $u_2 = (1 + \hat{\lambda}_2/c(1 + \hat{\lambda}_1))x(1 + \delta)^2$ and $(\hat{\lambda}_1, \hat{\lambda}_2, \hat{v}, \hat{v}')$ are the corresponding MLE of $(\lambda_1, \lambda_2, v, v')$. Furthermore, denote $\Delta = \left(\frac{\partial \hat{p}_2}{\partial \hat{\beta}_0}, \frac{\partial \hat{p}_2}{\partial \hat{\beta}_1}, \frac{\partial \hat{p}_2}{\partial \hat{\sigma}_e^2}\right)$, where

$$\begin{aligned} \frac{\partial \hat{p}_2}{\partial \hat{\beta}_0} &= \frac{2(c-1)\hat{\mu}_0}{c^2\hat{\sigma}_e^2(1+\hat{\lambda}_1)^2} \left[(1+\delta)^2 f_{\hat{v},\hat{v}'}(u_2) - (1-\delta)^2 f_{\hat{v},\hat{v}'}(u_1) \right] \\ &\quad + \frac{4\hat{\lambda}_1^2(1+\hat{\lambda}_1)}{\hat{\mu}_0(1+2\hat{\lambda}_1)^2} \int_{u_1}^{u_2} \frac{\partial f_{\hat{v},\hat{v}'}(x)}{\partial \hat{v}} dx \\ &\quad + \frac{4\hat{\lambda}_2^2(1+\hat{\lambda}_2)}{\hat{\mu}_0(1+2\hat{\lambda}_2)^2} \int_{u_1}^{u_2} \frac{\partial f_{\hat{v},\hat{v}'}(x)}{\partial \hat{v}'} dx, \\ \frac{\partial \hat{p}_2}{\partial \hat{\beta}_1} &= x \frac{\partial \hat{p}_2}{\partial \hat{\beta}_0}; \end{aligned}$$

and

$$\begin{aligned} \frac{\partial \hat{p}_2}{\partial \hat{\sigma}_e^2} &= \frac{2(c-1)\hat{\mu}_0}{c^2\hat{\sigma}_e^2(1+\hat{\lambda}_1)^2} \left[(1+\delta)^2 f_{\hat{v},\hat{v}'}(u_2) - (1-\delta)^2 f_{\hat{v},\hat{v}'}(u_1) \right] \\ &\quad - \frac{2\hat{\lambda}_1^2(1+\hat{\lambda}_1)}{\hat{\sigma}_e^2(1+2\hat{\lambda}_1)^2} \int_{u_1}^{u_2} \frac{\partial f_{\hat{v},\hat{v}'}(x)}{\partial \hat{v}} dx \\ &\quad - \frac{2\hat{\lambda}_2^2(1+\hat{\lambda}_2)}{\hat{\sigma}_e^2(1+2\hat{\lambda}_2)^2} \int_{u_1}^{u_2} \frac{\partial f_{\hat{v},\hat{v}'}(x)}{\partial \hat{v}'} dx; \end{aligned}$$

with

$$\partial f_{\hat{v}, \hat{v}'}(x) / \partial \hat{v} = 1/2 \left(f_{\hat{v}, \hat{v}'}(x) \left[(\log \Gamma(\hat{v} + \hat{v}'))^{(1)} - (\log \Gamma(\hat{v}))^{(1)} + \log(\hat{v}x / \hat{v}x + \hat{v}') + \hat{v}'(1-x) / \hat{v}x + \hat{v}' \right] \right),$$

$$\partial f_{\hat{v}, \hat{v}'}(x) / \partial \hat{v}' = 1/2 \left(f_{\hat{v}, \hat{v}'}(x) \left[(\log \Gamma(\hat{v} + \hat{v}'))^{(1)} - (\log \Gamma(\hat{v}'))^{(1)} + \log(\hat{v}' / \hat{v}x + \hat{v}') + \hat{v}(1-x) / \hat{v}x + \hat{v}' \right] \right)$$

and $(\log \Gamma(s))^{(1)}$ is the first order derivative of the natural logarithm of the gamma function with respect to s .

For a sufficiently large sample size, $\hat{p}_2$ can be approximated by a normal distribution with mean p_2 and variance V_2 , where

$$V_2 = \Delta \begin{pmatrix} (X^T X)^{-1} \hat{\sigma}_e^2 & 0 \\ 0' & \frac{2\hat{\sigma}_e^4}{n-2} \end{pmatrix} \Delta^T.$$

Thus, the hypotheses given in (7) for one-way translation based on the probability of relative difference can be tested. In particular, H_0 is rejected if

$$Z = \frac{\hat{p}_2 - p_0}{\sqrt{\hat{V}_2}} > z_{1-\alpha},$$

where $z_{1-\alpha}$ is the $100(1-\alpha)^{\text{th}}$ percentile of a standard normal distribution. Note that $\hat{V}_2$ is an estimate of $\text{var}(\hat{p}_2)$ and can be evaluated simply by replacing the parameters with their corresponding estimates of the parameters.

An Example

For the two measures proposed in the above section, p_1 is based on the absolute difference between y and $\hat{y}$. Given α , p_0 and the selected observation (x_0, y_0) , the hypothesis $H_0: p_1 > p_0$ is rejected in favor of $H_a: p_1 > p_0$ when $Z = \hat{p}_1 - p_0 / \sqrt{\hat{V}_1} > z_{1-\alpha}$. Equivalently, H_0 is rejected when $(\hat{p}_1 - p_0 - z_{1-\alpha} \sqrt{\hat{V}_1}) > 0$. Note that the value of $\hat{p}_1$ depends on the value of δ and it can be shown that $(\hat{p}_1 - p_0 - z_{1-\alpha} \sqrt{\hat{V}_1})$ is an increasing function of δ over $(0, \infty)$. Thus, $(\hat{p}_1 - p_0 - z_{1-\alpha} \sqrt{\hat{V}_1}) > 0$ if and only if $\delta > \delta_0$. Thus, the hypothesis H_0 can be rejected based on δ_0 instead of $\hat{p}_1$ as long as we can find the value of δ_0 for the given x_0 . On the other hand, from a practical point of view, p_2 is more intuitive to understand because it is based on the relative difference, which is equivalent to measure the percentage difference relative to the observed y and δ can be viewed as the upper bound of the percentage error.

For illustration purpose, suppose that the following data are observed in a translational study, where x is a given dose level and y is the associated toxicity measure:

x	0.9	1.1	1.3	1.5	2.2	2.0	3.1	4.0	4.9	5.6
y	0.9	0.8	1.9	2.1	2.3	4.1	5.6	6.5	8.8	9.2

When this set of data is fitted to model (1), the estimates of the model parameters are given by $\hat{\beta}_0 = -0.704$, $\hat{\beta}_1 = 1.851$, $\hat{\sigma}^2 = 0.431$. Thus, based on the fitted results, given $x = x_0$, the proposed translation model is given by $\hat{y} = -0.704 + 1.851x_0$.

In this study, choose $\alpha = 0.05$ and $p_0 = 0.8$. In particular, two dose levels $x_0 = 1.0$ and 5.2 are considered. Based on the study, the corresponding toxicity measures y_0 are 1.2 and 9.0 respectively. However, based on the translation model, the predicted toxicity measures are 1.147 and 8.921 , respectively. In the following, the validity of the translation model is assessed by the two proposed closeness measures p_1 and p_2 , respectively. Without loss of generality, choose $\alpha = 0.05$ and $p_0 = 0.8$.

Case 1: Testing of $H_0: p_1 \leq p_0$ vs. $H_a: p_1 > p_0$

Using the above results, for $x_0 = 1.0$, δ is 1.112 , since $|y_0 - \hat{y}| = |1.2 - 1.147| = 0.053$, which is less than $\delta = 1.112$, therefore H_0 is rejected. Similarly, for $x_0 = 5.2$, the corresponding δ is 1.178 , then $|y_0 - \hat{y}| = |9.0 - 8.921| = 0.079$, which is again smaller than $\delta = 1.178$, thus H_0 is rejected.

Case 2: Testing of $H_0: p_2 \leq p_0$ vs. $H_a: p_2 > p_0$

Suppose that $\delta = 1$, for the given two values of x , estimates of p_2 and the corresponding values of the test statistic Z are given in the following table.

x_0	y_0	$\hat{y}$	$\hat{p}_2$	Z	
1.0	1.2	1.147	0.870	1.183	Do not reject H_0
5.2	9.0	8.921	0.809	1.164	Do not reject H_0

Two-Way Translation

Validation of a Two-Way Translational Process

The translational process described in the section “One-Way Translation” is usually referred to as a *one-way translation* in translational medicine. That is, the information observed at basic research discoveries is translated to clinic. In other words, we can exchange x and y in Eq. (5)

$$x = \gamma_0 + \gamma_1 y + \varepsilon \quad (8)$$

and come up with another predictive model $\hat{x} = \hat{\gamma}_0 + \hat{\gamma}_1 y$.

Following the similar ideas, using either one of the measures p_i , the validation of a two-way translational process can be summarized by the following steps:

Step 1: For a given set of data (x, y) , established a predictive model, say, $y = f(x)$.

- Step 2: Select the bound δ_{yi} for the difference between y and $\hat{y}$. Evaluate $\hat{p}_{yi} = P\{|y - \hat{y}| < \delta_{yi}\}$. Assess the one-way closeness between y and $\hat{y}$ by testing the hypotheses (7). Proceed to the next step if the one-way translation process is validated.
- Step 3: Consider x as the dependent variable and y as the independent variable. Set up the regression model. Predict x at the selected observation y_0 , denoted by $\hat{x}$, based on the established model between x and y (i.e., $x = g(y)$), i.e., $\hat{x} = g(y) = \hat{\gamma}_0 + \hat{\gamma}_1 y$.
- Step 4: Select the bound δ_{xi} for the difference between x and $\hat{x}$. Evaluate the closeness between x and $\hat{x}$ based on a test for the following hypotheses

$$H_0: p_i \leq p_0 \quad \text{vs.} \quad H_1: p_i > p_0$$

where

$$p_i = P\left\{\left|\frac{y - \hat{y}}{y}\right| < \delta_{yi} \quad \text{and} \quad \left|\frac{x - \hat{x}}{x}\right| < \delta_{xi}\right\}$$

The above test can be referred to as a test for two-way translation. If, in Step 4, H_0 is rejected in favor of H_1 , this would imply that there is a two-way translation between x and y (i.e., the established predictive model is validated). However, the evaluation of p involved the joint distribution of $x - \hat{x}/x$ and $y - \hat{y}/y$. An exact expression is not readily available. Thus, an alternative approach is to modify Step 4 of the above procedure and proceed with a conditional approach instead. In particular,

Step 4 (modified): Select the bound δ_{xi} for the difference between x and $\hat{x}$. Evaluate the closeness between x and $\hat{x}$ based on a test for the following hypotheses,

$$H_0: p_{xi} \leq p_0 \quad \text{vs.} \quad H_a: p_{xi} > p_0 \quad (9)$$

where $p_{xi} = \{ |x - \hat{x}| < \delta_{xi} \}$.

Note that the evaluation of p_{xi} is much easier and can be computed in similarly by interchanging the role of x and y for the results given in section “Measures of Closeness.”

An Example (Continued)

Using the data set given in section “An Example,” we set up the regression model $x = \gamma_0 + \gamma_1 y + \varepsilon$ with y as the independent variable and x as the dependent variable. The estimates of the model parameters are $\hat{\gamma}_0 = 0.468$, $\hat{\gamma}_1 = 0.519$, and $\hat{\sigma}^2 = 0.121$. Based on this model, for the same α and p_0 , given $(x_0, y_0) = (1.0, 1.2)$ and $(5.2, 9.0)$, the fitted values are given by $\hat{x} = 0.468 + 0.519y_0$.

Case 1: Testing of $H_0: p_{x1} \leq p_0$ vs. $H_a: p_{x1} > p_0$

Using the above results, for $y_0 = 1.2$, δ is 0.587, since $|x_0 - \hat{x}| = |1.0 - 1.09| = 0.09$, which is less than $\delta_x = 0.587$, therefore H_0 is rejected. Similarly, for $y_0 = 9.0$, the corresponding δ is 0.624, then $|x_0 - \hat{x}| = |5.2 - 5.139| = 0.061$, which is again smaller than $\delta = 0.624$, thus H_0 is rejected.

Case 2: Testing of $H_0: p_{x2} \leq p_0$ vs. $H_a: p_{x2} > p_0$

Suppose that $\delta = 1$, for the given two values of y , estimates of p_{x2} and the corresponding values of the test statistic Z are given in the following table.

x_0	y_0	$\hat{x}_0$	$\hat{p}_{x2}$	Z	
1.0	1.2	1.090	0.809	1.300	Do not reject H_0
5.2	9.0	5.139	0.845	16.53	Do not reject H_0

Lost in Translation

It can be noted that δ_y and δ_x can be viewed as the maximum bias (or possible lost in translation) from the one-way translation (e.g., from basic research discovery to clinic) and from the other way of translation (e.g., from clinic to basic research discovery), respectively. If δ_y and δ_x given in steps 2 and 4 of section “One-Way Translation” are close to 0 with a relatively high probability, then we conclude that the information from the basic research discoveries (clinic) is fully translated to the clinic (basic research discoveries). Thus, one may consider the following parameter to measure the degree of lost in translation

$$\zeta = 1 - p_{xy} p_{yx},$$

where p_{xy} is the measure of closeness from x to y and p_{yx} is the measure of closeness from y to x . When $\zeta \approx 0$, we consider that there is no lost in translation. Overall lost in translation could be significant even lost in translation from the one-way translation is negligible. To illustrate, if there is a 10% lost in translation in one-way translation and 20% lost in translation in the other way, there would be an up to 28% lost in translation. In practice, an estimate of ζ can be obtained for a given set of data (x, y) . In particular, $\hat{\zeta} = 1 - p_{xy} p_{yx}$.

As an illustration, consider the example discussed in the section “An Example (Continued).” Suppose that the measure of closeness based on relative difference is used, given $(x_0, y_0) = (1.0, 1.2)$ and $(5.2, 9.0)$, the corresponding lost in translation for the two-way translation with $\delta = 1$ is tabulated in the following table:

x_0	y_0	$\hat{y}$	$\hat{p}_{xy}$	$\hat{x}$	$\hat{p}_{yx}$	$\hat{\zeta}$
1.0	1.2	1.147	0.870	1.090	0.809	0.296
5.2	9.0	8.921	0.809	5.139	0.845	0.316

ANIMAL MODEL VERSUS HUMAN MODEL

In translational medicine, a commonly asked question is that whether an animal model is predictive of a human model. To address this question, we may assess the similarity between an animal model (population) and a human model (population). For this purpose, we first establish an animal model to bridge the basic research discovery (x) and clinic (y). For example, consider a one-way translation process where $y = \beta_0 + \beta_1 x + \varepsilon$. Let $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ be the predictive model obtained from the one-way translation based on data from an animal population. Thus, for a given x_0 , $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ follows a distribution with mean μ_y and σ_y^2 . Under the predictive model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, denote by (μ_y, σ_y) the target population. Assume that the predictive model works for the target population. Thus, for an animal population, $\mu_y = \mu_{\text{animal}}$ and $\sigma_y = \sigma_{\text{animal}}$, while for a human population, $\mu_y = \mu_{\text{human}}$ and $\sigma_y = \sigma_{\text{human}}$. Assuming that the linear predictive model can be applied to both animal population and human population, we can link the animal and human model by the following

$$\mu_{\text{human}} = \mu_{\text{animal}} + \varepsilon,$$

and

$$\sigma_{\text{human}} = C\sigma_{\text{animal}}.$$

In other words, we expect there are differences in population mean and population standard deviation under the predictive model due to possible differences in response between animal and human. As a result, the effect size adjusted for standard deviation under the human population can be obtained as follows:

$$\left| \frac{\mu_{\text{human}}}{\sigma_{\text{human}}} \right| = \left| \frac{\mu_{\text{animal}} + \varepsilon}{C\sigma_{\text{animal}}} \right| = |\Delta| \left| \frac{\mu_{\text{animal}}}{\sigma_{\text{animal}}} \right|$$

where $\Delta = (1 + \varepsilon/\mu_{\text{animal}})/C$. Chow et al.^[17] refer to Δ as a sensitivity index when changing from a target population to another. As can be seen, the effect size under the human population is inflated (or reduced) by the factor of Δ . If $\varepsilon = 0$ and $C = 1$, we then claim that there is no difference between the animal population and the human population. Thus, the animal model is predictive of the human model. Note that the shift and scale parameters (i.e., ε and C) can be estimated by

$$\hat{\varepsilon} = \hat{\mu}_{\text{human}} - \hat{\mu}_{\text{animal}}$$

and

$$\hat{C} = \frac{\hat{\sigma}_{\text{human}}}{\hat{\sigma}_{\text{animal}}},$$

respectively, in which $(\hat{\mu}_{\text{animal}}, \hat{\sigma}_{\text{animal}})$ and $(\hat{\mu}_{\text{human}}, \hat{\sigma}_{\text{human}})$ are estimates of $(\mu_{\text{animal}}, \sigma_{\text{animal}})$ and $(\mu_{\text{human}}, \sigma_{\text{human}})$, respectively. Thus, the sensitivity index can be assessed as follows:

$$\hat{\Delta} = (1 + \hat{\varepsilon}/\mu_{\text{human}})/\hat{C}.$$

In practice, there may be a shift in population mean (i.e., ε) and/or in population standard deviation (i.e., C), Chow et al.^[18] indicated that shifts in population mean and population standard deviation can be classified into the following four cases where (i) both ε and C are fixed, (ii) ε is random and C is fixed, (iii) ε is fixed and C is random, and (iv) both ε and C are random. For the case where both ε and C are fixed, (v) can be used for estimation of.^[18] Chow et al.^[18] derived statistical inference of Δ for the case where ε is random and C is fixed by assuming that y conditional on μ follows a normal distribution $N(\mu, \sigma^2)$. That is,

$$y|_{\mu=\mu_{\text{human}}} \sim N(\mu, \sigma^2),$$

where μ is distributed as $N(\mu_{\mu}, \sigma_{\mu}^2)$ and σ, μ_{μ} , and σ_{μ} , are some unknown constants. It can be verified that y follows a mixed normal distribution with mean μ_{μ} and variance $\sigma^2 + \sigma_{\mu}^2$. That is, $y \sim N(\mu_{\mu}, \sigma^2 + \sigma_{\mu}^2)$. As a result, the sensitivity index can be assessed based on data collected from both animal and human population under the predictive model.

Note that for other cases where C is random, the above method can also be derived similarly. The assessment of sensitivity index can be used to adjust the treatment effect to be detected under a human model when applying an animal model to a human model, especially when there is a significant or major shift between an animal population and the human population. In practice, it is of interest to assess the impact of the sensitivity index on both lost in translation and the probability of success. This, however, requires further research.

TRANSLATION IN STUDY ENDPOINTS

In clinical trials, it is not uncommon that a study is powered based on expected absolute change from baseline of a primary study endpoint, but the collected data are analyzed based on relative change from baseline (e.g., percent change from baseline) of the primary study endpoint. In many cases, the collected data are analyzed based on the percentage of patients who show some improvement (i.e., responder analysis). The definition of a responder could be based on either absolute change from baseline or relative change from baseline of the primary study endpoint. It is very controversial in terms of the interpretation of the analysis results, especially when a significant result is observed based on a study endpoint (e.g., absolute change from baseline, relative change from baseline, or the responder analysis) but not on the other study endpoint (e.g., absolute change from baseline, relative change from baseline, or responder analysis). In practice, it is then of interest to explore how an observed significant difference of a study endpoint (e.g., absolute change from baseline,

relative change from baseline, or responder's analysis) can be translated to that of the other study endpoint (e.g., absolute change from baseline, relative change from baseline, or responder's analysis).

Power Analysis and Sample Size Calculation

An immediate impact on the assessment of treatment effect based on different study endpoints is the power analysis for sample size calculation. For example, sample size required in order to achieve the desired power based on the absolute change could be very different from that obtained based on the percent change, or the percentage of patients who show an improvement based on the absolute change or relative change at α level of significance. Let w_{1i} and w_{2i} be the measurements of the i -th subject before and after the treatment, respectively. For example, assume that w_{1i} are lognormal distributed, i.e., $\log w_{1i} \sim N(\mu_1, \sigma^2)$. Let $w_{2i} \equiv w_{1i}(\tilde{w}_{1i} + 1)$, where $\log \tilde{w}_{1i} \sim N(\mu_2, \sigma^2)$ and w_{1i} , $\tilde{w}_{1i}$ are independent for $1 \leq i, j \leq n$. It easily follows that $\log(w_{2i} - w_{1i}) \sim N(\mu_1 + \mu_2, 2\sigma^2)$ and $\log(w_{2i} - w_{1i}/w_{1i}) \sim N(\mu_2, \sigma^2)$. Define X_i and Y_i as $X_i = \log(w_{2i} - w_{1i})$, $Y_i = \log(w_{2i} - w_{1i}/w_{1i})$. Then X_i and Y_i represent the logarithm of the absolute change and relative change of the measurements before and after the treatment. It can be shown that both X_i and Y_i are normally distributed. More details on the derivation of these results can be found in the Appendix.

Let μ_{AC} and μ_{RC} be the population means of the logarithm of the absolute change and relative change of a primary study endpoint of a given clinical trial, respectively. Thus, the hypotheses of interest based on the absolute change are given by

$$H_{01} : \mu_{AC} \leq \delta_0 \quad \text{vs.} \quad H_{a1} : \mu_{AC} > \delta_0,$$

where δ_0 is the difference of clinically importance. For the relative change, the hypotheses of interest are given by

$$H_{02} : \mu_{RC} \leq \Delta_0 \quad \text{vs.} \quad H_{a2} : \mu_{RC} > \Delta_0,$$

where Δ_0 is the difference of clinically importance. In practice, for a specific value of δ , where $\delta > \delta_0$, would be equivalent to what value of Δ with $\Delta > \Delta_0$ clinically.

In addition, if we consider that a patient is a responder if the logarithm of his/her absolute change of the primary study endpoint is greater than δ , then, it is of interest to test the following hypotheses:

$$H_{03} : P_{AC} = \eta \quad \text{vs.} \quad H_{a3} : P_{AC} > \eta,$$

where P_{AC} is the proportion of patients whose logarithm of the absolute change of the primary study endpoint is greater than δ . In practice, we may claim superiority (clinically) of the test treatment if we reject the null hypothesis at $\eta = 50\%$ and in favor of the alternative hypothesis that P_{AC} is greater than 50%. However, this lacks statistical

justification. For a non-inferiority (or superiority) trial, how can a non-inferiority margin of μ_{AC} be translated to the non-inferiority margin of P_{AC} ? Similarly, for the relative change, the hypotheses of interest are given by

$$H_{04} : P_{RC} = \eta \quad \text{vs.} \quad H_{a4} : P_{RC} > \eta,$$

where P_{RC} is the proportion of patients whose logarithm of the relative change of the primary study endpoint is greater than δ .

An Example

For example, consider a clinical trial for evaluation of possible weight reduction of a test treatment in female patients. Weight data from 10 subjects are given in Table 1. As it can be seen from Table 1, mean absolute change and mean percent change from pre-treatment are 5.3 lb. and 4.8%, respectively. If a subject is considered a responder if there is weight reduction by more than 5 lb. (absolute change) or by more than 5% (relative change), the response rates based on absolute change and relative change are given by 40% and 30%, respectively. For illustration purpose, Table 2 summarizes sample sizes required for achieving a desired power for detecting a clinically meaningful difference, say,

Table 1 Weight Data from Ten Female Subjects

Pre-treatment	Post-treatment	Absolute change	Relative change (%)
110	106	4	3.6
90	80	10	11.1
105	100	5	4.8
95	93	2	2.1
170	163	7	4.1
90	84	6	6.7
150	145	5	3.3
135	131	4	3.0
160	159	1	0.6
100	91	9	9.0
120.5 (30.5)	115.2 (31.5)	5.3	4.8

Table 2 Sample Size Calculation

Study endpoint	Clinical meaningful difference	Sample size required
Absolute Change	5 lb.	190
Relative Change	5 %	95
Responder 1 ^a	50% improvement	54
Responder 2 ^b	50% improvement	52

^aResponder is defined based on absolute change greater than or equal to 5.5 lb.

^bResponder is defined based on relative change greater than or equal to 5.5%.

by an absolute change of 5.5 lb. and a relative change of 5.5%, for the two study endpoints respectively.

As it can be seen from Table 2, a sample size of 190 is required for achieving an 80% power for detecting a difference of 5.5 lb. (post-treatment absolute change from pre-treatment) at the 5% level of significance, while a much larger sample size of 95 is required in order to have an 80% power for detecting a difference of 5.5% (relative change between post-treatment and pre-treatment) at the 5% level of significance. Based on responder's analysis, the results are different. Based on the analysis of responders, defined as subjects who have greater than or equal to 5.5 lb. of weight reduction, a total sample size of 54 subjects are needed for detecting a 50% improvement at the 5% level of significance. On the other hand, if we define a responder as a subject who has a greater than or equal to 5.5% relative weight reduction, then a total sample size of 52 subjects is required for achieving a 50% improvement at the 5% level of significance.

Remarks

As discussed above, sample sizes required for achieving a desired power for detecting a clinically meaningful difference at the 5% level of significance may be very different depending upon (i) the choice of study endpoint and (ii) clinically meaningful difference. In practice, it will be more complicated if the intended trial is to establish non-inferiority. In this case, sample size calculation will also depend on the selection of non-inferiority margin. To ensure the success of the intended clinical trial, the sponsor will usually carefully evaluate several clinical strategies for selecting the type of study endpoint, clinically meaningful difference, and non-inferiority margin during the stage of protocol development. Commonly considered study endpoints are

- measure based on absolute change;
- measure based on relative change;
- proportion of responders defined based on absolute change; and
- proportion of responders defined based on relative change.

In some cases, investigators may consider composite endpoint based on both absolute change and relative change. For example, in clinical trials for evaluation of the efficacy and safety of a compound for treating patients with active ulcerative colitis, a study endpoint utilizing the so-called Mayo score is often considered. The investigator may define a subject as a responder if he/she has a decrease from baseline in the total score of at least 3 points and at least 30% with an accompanying decrease in the subscore for rectal bleeding of at least one point or absolute subscore for rectal bleeding of 0 or 1 at Day 57. Note that the Mayo scoring system for assessment of ulcerative colitis activity

consists of three domains of (i) Mayo score, (ii) partial Mayo score, and (iii) mucosal healing.^[19]

For the four types of study endpoints derived from the clinical data collected from the sample patient population, clinically meaningful difference or non-inferiority margin that we would like to detect or establish could be based on either absolute change or relative change. For example, based on responder's analysis, we may want to detect a 30% difference in response rate or to detect a 50% relative improvement in response rate. As a result, there are a total of eight clinical strategies for assessment of the treatment effect. In practice, some strategies may lead to the success of the intended clinical trial (i.e., achieve the study objectives with the desired power), while other strategies may not. A common practice is for the sponsors to choose the strategy to their best interest. However, regulatory agencies may challenge the sponsor about the inconsistent results. This has raised the following questions: (i) how is the clinical information translated among different study endpoints (since they are obtained based on the same data collected from the same patient population)? and (ii), which study endpoint is telling the truth? These questions, however, remain unanswered. More research is required to address these questions. The current regulatory position is to require the sponsor to pre-specify which study endpoint will be used for assessment of the treatment effect in the study protocol without any scientific justification.

BRIDGING STUDIES

In recent years, the influence of ethnic factors on clinical outcomes for evaluation of efficacy and safety of study medications under investigation has attracted much attention from regulatory authorities, especially when the sponsor is interested in bringing an approved drug product from the original region (e.g., the United States or the European Union) to a new region (e.g., the Asian Pacific region). To determine whether clinical data generated from the original region are acceptable in the new region, the International Conference on Harmonization (ICH) issued a guideline on Ethnic Factors in the Acceptability of Foreign Clinical Data. The purpose of this guideline is not only to permit adequate evaluation of the influence of ethnic factors, but also to minimize duplication of clinical studies in the new region.^[20] This guideline is known as ICH E5 guideline.

As indicated in the ICH E5 guideline, a bridging study is defined as a study performed in the new region to provide pharmacokinetic (PK), pharmacodynamic (PD), or clinical data on efficacy, safety, dosage, and dose regimen in the new region that will allow extrapolation of the foreign clinical data to the population in the new region. The ICH E5 guideline suggests the regulatory authority of the new region to assess the ability to extrapolate foreign data based on the bridging data package, which

consists of (i) information including PK data and any preliminary PD and dose-response data from the complete clinical data package (CCDP) that is relevant to the population of the new region, and if needed, (ii) bridging study to extrapolate the foreign efficacy data and/or safety data to the new region. The ICH E5 guideline indicates that bridging studies may not be necessary if the study medicines are insensitive to ethnic factors. For medicines characterized as insensitive to ethnic factors, the type of bridging studies (if needed) will depend upon experience with the drug class and upon the likelihood that extrinsic ethnic factors could affect the medicine's safety, efficacy, and dose-response. On the other hand, for medicines that are ethnically sensitive, bridging study is usually needed since the populations in two regions are different. In the ICH E5 guideline, however, no criteria for assessment of the sensitivity to ethnic factors for determining whether a bridging study is needed are provided. Moreover, when a bridging study is conducted, the ICH guideline indicates that the study is readily interpreted as capable of bridging the foreign data if it shows that dose-response, safety, and efficacy in the new region are similar to those in the original region. However, the ICH does not clearly define the similarity.

Shih^[21] interpreted the ICH guideline as consistency among study centers by treating the new region as a new center of multicenter clinical trials. Under this definition, Shih^[21] proposed a method for assessment of consistency to determine whether the study is capable of bridging the foreign data to the new region. Alternatively, Shao and Chow^[22] proposed the concepts of reproducibility and generalizability probabilities for assessment of bridging studies. If the influence of the ethnic factors is negligible, then we may consider the reproducibility probability to determine whether the clinical results observed in the original region are reproducible in the new region. If there is a notable ethnic difference, the generalizability probability can be assessed to determine whether the clinical results in the original region can be generalized in a similar but slightly different patient population due to the difference in ethnic factors. In addition, Chow et al.^[17] proposed to assess bridging studies based on the concept of population (or individual) bioequivalence. Along this line, Hung^[23] and Hung et al.^[24] considered the assessment of similarity based on testing for non-inferiority between a bridging study conducted in the new region as compared to the previous conducted in the original region. This leads to the argument regarding the selection of non-inferiority margin.^[25] Note that other methods such as the use of Bayesian approach have also been proposed in literature.^[26]

Test for Consistency

For assessment of similarity between a bridging study conducted in a new region and studies conducted in the original region, Shih^[21] considered all of the studies conducted

in the original region as a multicenter trial and proposed to test consistency among study centers by treating the new region as a new center of a multicenter trial.

Suppose there are K reference studies in the CCDP. Let T_i denotes the standardized treatment group difference, i.e.,

$$T_i = \frac{\bar{x}_{Ti} - \bar{x}_{Ci}}{s_i \sqrt{(1/m_{Ti} + 1/m_{Ci})}},$$

where $\bar{x}_{Ti}(\bar{x}_{Ci})$ is the sample mean of $m_{Ti}(m_{Ci})$ observations in the treatment (control) group, and s_i is the pooled sample standard deviation. Shih^[21] considered the following predictive probability for testing consistency

$$p(T|T_i, i = 1, \dots, K) = \left(\frac{2\pi(K+1)}{K} \right)^{-K/2} \times \exp \left[-K(T - \bar{T})^2 / 2(K+1) \right].$$

Test for Reproducibility and Generalizability

On the other hand, when the ethnic difference is negligible, Shao and Chow^[22] suggest assessing reproducibility probability for similarity between clinical results from a bridging study and studies conducted in the CCPD. Let x be a clinical response of interest in the original region. Let y be similar to x but is a response in a clinical bridging study conducted in the new region. Suppose the hypotheses of interest are

$$H_0: \mu_1 = \mu_0 \quad \text{vs.} \quad H_a: \mu_1 \neq \mu_0.$$

We reject H_0 at the 5% level of significance if and only if $|T| > t_{n-2}$, where t_{n-2} is the $(1 - \alpha/2)$ th percentile of the t distribution with $n - 2$ degrees of freedom, $n = n_1 + n_2$,

$$T = \frac{\bar{y} - \bar{x}}{\sqrt{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2 / n - 2\sqrt{(1/n_1 + 1/n_2)}}},$$

and $\bar{x}, \bar{y}, s_0^2$, and s_1^2 are sample means and variances for the original region and the new region, respectively. Thus, the power of T is given by

$$p(\theta) = P(|T| > t_{n-2}) = 1 - \mathfrak{Z}_{n-2}(t_{n-2}|\theta) + \mathfrak{Z}_{n-2}(-t_{n-2}|\theta),$$

where

$$\theta = \frac{\mu_1 - \mu_0}{\sigma \sqrt{(1/n_1 + 1/n_0)}},$$

and $\mathfrak{Z}_{n-2}(\bullet|\theta)$ denotes the cumulative distribution function of the non-central t distribution with $n - 2$ degrees of freedom and the non-centrality parameter θ . Replacing θ in the power function with its estimate $T(x)$, the estimated power

$$\hat{p} = P(T(x)) = 1 - \mathfrak{Z}_{n-2}(t_{n-2}|T(x)) + \mathfrak{Z}_{n-2}(-t_{n-2}|T(x))$$

is defined as a reproducibility probability for a future clinical trial with the same patient population. When the ethnic difference is notable, Shao and Chow^[22] recommend assessing so-called generalizability probability for similarity between clinical results from a bridging study and studies conducted in the CCPD.

Test for Similarity

Using the criterion for assessment of population (individual) bioequivalence, Chow et al.^[17] proposed the following measure of similarity between x and y :

$$\theta = \frac{E(x-y)^2 - E(x-x')^2}{E(x-x')^2/2},$$

where x' is an independent replicate of x and y , x , and x' are assumed to be independent. Since a small value of θ indicates that the difference between x and y is small (relative to the difference between x and x'), similarity between the new region and the original region can be claimed if and only if $\theta < \theta_U$, where θ_U is a similarity limit. Thus, the problem of assessing similarity become a problem of testing the following hypotheses

$$H_0: \theta \geq \theta_U \quad \text{vs.} \quad H_a: \theta < \theta_U.$$

Let $k=0$ indicate the original region and $k=1$ indicates the new region. Suppose that there are m_k study centers and n_k responses in each center for a given variable of interest. For simplicity, we only consider the balanced case where centers in a given region have the same number of observations. Let z_{ijk} be the i th observation from the j th center of region k , b_{jk} be the between-center random effect, and e_{ijk} be the within center measurement error. Assume that

$$z_{ijk} = \mu_k + b_{jk} + e_{ijk}, \quad i = 1, \dots, n_k, \quad j = 1, \dots, m_k, \quad k = 0, 1,$$

where μ_k is the population mean in region k , $b_{jk} \sim N(0, \sigma_{bk}^2)$, $e_{ijk} \sim N(0, \sigma_{wk}^2)$, and $\{b_{jk}\}$ and $\{e_{ijk}\}$ are independent. Under the above model, the criterion for similarity becomes

$$\theta = \frac{(\mu_0 - \mu_1)^2 + \sigma_{r1}^2 - \sigma_{r0}^2}{\sigma_{T0}^2},$$

where $\sigma_{Tk}^2 = \sigma_{bk}^2 + \sigma_{wk}^2$ is the total variance (between center variance plus within center variance) in region k . The above hypotheses are equivalent to

$$H_0: \varsigma \geq 0 \quad \text{vs.} \quad H_a: \varsigma < 0,$$

where $\varsigma = (\mu_0 - \mu_1)^2 + \sigma_{T1}^2 - (1 + \theta_U)\sigma_{T0}^2$.

CONCLUDING REMARKS

Translational medicine is a multi-disciplinary entity that bridges basic scientific research with clinical development. As the expense in developing therapeutic pharmaceutical compounds continues to increase and the success rates for getting such compounds approved for marketing and to the patients needing these treatments continues to decrease, a focused effort has emerged to improve the communication and coordination of planning between basic and clinical science. This will likely lead to more therapeutic insights being derived from new scientific ideas, and more feedback being provided back to research so that approaches are better targeted. Translational medicine spans all the disciplines and activities that lead to making key scientific decisions as a compound traverses across the difficult preclinical-clinical divide. Many argue that improvement in decision-making about what doses and regimens should be pursued in the clinic—based on the likely human safety risks of a compound, the likely drug interactions, and the pharmacologic behavior of the compound—is the most critical step in the entire development process. Many of these decisions and the path for uncovering this information within later development are defined at this specific time within the drug development process. Improving these decisions will likely lead to a substantial increase in the number of safe and effective compounds available to combat human diseases.

REFERENCES

1. Chow, S.C.; Chang, M. *Adaptive Design Methods in Clinical Trials*; Chapman Hall/CRC Press, Taylor and Francis: New York, 2006.
2. Chang, M. *Adaptive Design Theory and Implementation Using SAS and R*; Chapman and Hall/CRC Press, Taylor and Francis: New York, 2007.
3. Chow, S.C. Statistics in translational medicine. Presented at Current Advances in Evaluation of Research & Development of Translational Medicine. National Health Research Institutes, Taipei, Taiwan, October 19, 2007.
4. Chen, J.; Chen, C. Microarray gene expression. In *Encyclopedia of Biopharmaceutical Statistics*; Chow, S.C., Ed.; Marcel Dekker, Inc: New York, 2003, 599–613.
5. Shao, J.; Chow, S.C. Variable screening in predicting clinical outcome with high-dimensional microarrays. *J. Multivariate Anal.* **2007**, *98*, 1529–1538.
6. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. Royal Stat. Soc.* **1996**, *58*, 267–288.
7. Akaike, H.A. A new look at statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723.
8. Shibata, R. An optimal selection of regression variables. *Biometrika* **1981**, *68*, 45–54.
9. Mallows, C.L. Some comments on Cp. *Technometrics* **1973**, *15*, 661–675.

10. Efron, B. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Stat. Assoc.* **1983**, 78, 316–331.
11. Shao, J. Linear model selection by cross-validation. *J. Am. Stat. Assoc.* **1993**, 88, 486–494.
12. Gunst, G.F.; Mason, R.L. *Regression Analysis and Its Application*; Marcel Dekker, Inc.: New York, 1980.
13. Li, L.; Chow, S.C.; Smith, W. Cross-validation for linear model with unequal variances in genomic analysis. *J. Biopharm. Stat.* **2004**, 14, 723–739.
14. Pizzo, P.A. *The Dean's Newsletter*; Stanford University School of Medicine: Stanford, CA, 2006.
15. Mankoff, S.P.; Brander, C.; Ferrone, S.; Marincola, F.M. Lost in translation: obstacles to translational medicine. *J. Transl. Med.* **2004**, 2, 14.
16. Tse, S.K. Statistical tests for one-way/two-way translation in translational medicine. Presented at Current Advances in Evaluation of Research & Development of Translational Medicine. National Health Research Institutes, Taipei, Taiwan, October 19; 2007.
17. Chow, S.C.; Shao, J.; Hu, O.Y.P. Assessing sensitivity and similarity in bridging studies. *J. Biopharm. Stat.* **2002**, 12, 385–400.
18. Chow, S.C.; Chang, M.; Pong, A. Statistical consideration of adaptive methods in clinical development. *J. Biopharm. Stat.* **2005**, 15, 575–591.
19. Rutgeerts, P.; Sandborn, W.J.; Feagan, B.G.; Reinisch, W.; Olson, A.; Johanns, J. Infliximab for induction and maintenance therapy for ulcerative colitis. *N. Engl. J. Med.* **2005**, 353, 2462–2476.
20. ICH E5. International Conference on Harmonization Tripartite Guideline on Ethnic Factors in the Acceptability of Foreign Data. *The U.S. Federal Register* **1997**, 83, 31790–31796.
21. Shih, W.J. Clinical trials for drug registrations in Asian Pacific countries: proposal for a new paradigm from a statistical perspective. *Control. Clin. Trials* **2001**, 22, 357–366.
22. Shao, J.; Chow, S.C. Reproducibility probability in clinical trials. *Stat. Med.* **2002**, 21, 1727–1742.
23. Hung, H.M.J. Statistical Issues with Design and Analysis of Bridging Clinical Trial. Presented at the 2003 Symposium on Statistical Methodology for Evaluation of Bridging Evidence, Taipei, Taiwan.
24. Hung, H.M.J.; Wang, S.J.; Tsong, Y.; Lawrence, J.; O'Neil, R.T. Some fundamental issues with non-inferiority testing in active controlled trials. *Stat. Med.* **2003**, 22, 213–225.
25. Chow, S.C.; Shao, J. On non-inferiority margin and statistical tests in active control trials. *Stat. Med.* **2006**, 25, 1101–1113.
26. Liu, J.P.; Hsueh, H.M.; Hsiao, C.F. Bayesian approach to evaluation of the bridging studies. *J. Biopharm. Stat.* **2002**, 12, 401–408.

Trend Estimation

Jian Qing Shi

School of Mathematics and Statistics, Newcastle University, U.K.

Man Lai Tang

Department of Mathematics, Hong Kong Baptist University, Hong Kong

Abstract

This entry discusses hypothesis testing and confidence intervals for determining trend estimation, and provides several models and methods for analysis and a sample analysis for further illustration and application of trend estimations.

OVERVIEW

In dose-effect data analysis, it is of interest to examine the effect (e.g., desirable effect or adverse effect) of different dosages of a drug or chemical in a system (e.g., an organ or a cell). Knowledge of the relationship between dosage and clinical response (effectiveness and adverse effects) is important for the safe and effective use of drugs in individual patients. For example, benzyl chloride, an occupational toxicant, is a colorless liquid with a pungent smell. It causes marked irritation of mucous membranes and conjunctiva. In various laboratory studies, the substance was shown to be carcinogenic after subcutaneous, epicutaneous, and intragastric administration to rats or mice; both local and systematic tumors were induced after intestinal uptake of benzyl chloride. In an animal carcinogenesis experiment,^[1] groups of 52 male mice were given benzyl chloride doses of 0, 50, or 100 mg/kg body weight in corn oil by gavage, three times weekly for two years. The proportions of mice with forestomach carcinomas and papillomas at the end of the study are reported in Table 1. In general, it is the purpose of dose-response trend analysis to examine if, for instance, the cure rate (risk) of certain disease (side effects) increases with increased intake (exposure) to a certain drug (toxicant). In this entry, the response of the subject will be assumed to be binary (e.g., absence or presence of a certain outcome), and the logistic model will be used to describe the dose-response relationship.

Assume that K doses of a substance are tested in an experiment, and that the k -th dose (denoted as x_k) is administered to n_k subjects, of whom r_k respond. In a case-control epidemiological study, r_k is the number of cases, and $n_k - r_k$ is the number of controls for the k th dose level. Let π_k be the probability that a subject shows response or the probability of there being a case at the k th dose level.

We assume independence among individuals within doses and among doses, and assume that the response

probability π_k satisfies a logit model. The model is defined as:

$$r_k \sim \text{Bin}(n_k, \pi_k), \quad \pi_k = \frac{1}{1 + \exp(\alpha + \beta x_k)}, \quad (1)$$

for $k = 1 \dots K$. Parameter β , which is called the trend parameter, measures the dose-response relationship or the association between the log-odds ratio (of being a case vs. being a control) and dose level. This is the parameter of interest. Parameter α is usually the nuisance parameter, i.e., parameter of no interest. The likelihood function of the observed counts $(r_1, \dots, r_k)$ is given by

$$L(\alpha, \beta) = \prod_{k=1}^K \binom{n_k}{r_k} \pi_k^{r_k} (1 - \pi_k)^{n_k - r_k} \quad (2)$$

$$\propto \frac{\exp\left[\alpha \sum_k r_k + \beta \sum_k x_k r_k\right]}{\prod_k \left[1 + \exp(\alpha + \beta x_k)\right]^{n_k}}. \quad (3)$$

Maximum likelihood estimates of α and β can be calculated by maximizing the above likelihood.

HYPOTHESIS TESTING AND CONFIDENCE INTERVAL

Asymptotic Approach

For the hypothesis testing method, it suffices to test the following hypotheses:

$$H_0: \beta = 0 \quad \text{vs} \quad H_A: \beta \neq 0. \quad (4)$$

In this case, a rejection of the null hypothesis indicates a dose-response trend. For the confidence interval method,

one can construct a $(1 - \gamma) \times 100\%$ confidence interval for β . If the value 0 is lying outside the constructed confidence interval, one may conclude the existence of a dose-response trend. In particular, if the lower (upper) confidence limit is greater (less) than zero, a positive (negative) linear trend is anticipated. In practice, one can perform either of the methods via (unconditional) asymptotic and (conditional) exact approaches.

For the asymptotic approach, the unconditional likelihood inference is based on maximizing the likelihood function in Eq. (3). Usually, asymptotic statistics (e.g., likelihood ratio, score, and Wald) can be used to test the null hypothesis in Eq. (4). Take the Wald statistic as an example. Let $\hat{\beta}$ be the maximum likelihood estimate of β in Eq. (3) and $s\beta$ the corresponding standard error. The Wald statistic calculates a Z statistic, which is $z = \hat{\beta} / s\beta$. According to the hypothesis testing method, a dose-response trend is guaranteed at γ level if $|z| \geq z_{\gamma/2}$, where $z_{\gamma/2}$ is the $(1 - \gamma/2) \times 100\%$ percentile of the standard normal distribution. Equivalently, one may construct a $(1 - \gamma) \times 100\%$ Wald confidence interval for β , i.e., $\hat{\beta} \pm z_{\gamma/2} s\beta$. According to the confidence interval method, a dose-response trend exists if 0 is not in $(\hat{\beta} - z_{\gamma/2} s\beta, \hat{\beta} + z_{\gamma/2} s\beta)$. For the carcinogenicity study with benzyl chloride in Table 1, we have $z = 5.50$ with p -value < 0.0001 , and we have reason to believe that there exists a dose-response trend between the carcinoma rate and the dosage of benzyl chloride. In addition, the 95% confidence interval for β is given by (1.98, 4.16), whose lower confidence limit is greater than 0. That further indicates a positive trend between the carcinoma rate and the dosage of benzyl chloride. It is noteworthy that the asymptotic approach is reliable only when the sample size is large enough.

Conditional Exact Approach

When the total sample size (i.e., $n = \sum_k n_k$) is small or the data structure (i.e., $r = (r_1, \dots, r_k)$) is sparse, the asymptotic approach is generally believed to be unreliable. Under these situations, exact inference, based on enumerating the exact distribution of sufficient statistics of the parameter of interest, is widely recommended.^[2-4] To perform conditional inference, we first note that the sufficient statistics for the parameters α and β can be shown to be $s = \sum_k r_k$ and $t = \sum_k x_k r_k$, respectively. To eliminate the nuisance parameter α from the analysis,^[5,6] one can consider the distribution of $r = (r_1, \dots, r_k)'$ conditioned on s . That is,

$$\Pr[r | s] = \frac{\prod_k \left[\binom{n_k}{r_k} \exp(\beta x_k r_k) \right]}{\sum_{u \in \Omega(s)} \prod_k \left[\binom{n_k}{u_k} \exp(\beta x_k u_k) \right]},$$

where $\Omega(s) = \{u = (u_1, \dots, u_K) : 0 \leq u_k \leq n_k, k = 1, \dots, K \text{ s.t. } \sum_k u_k = s\}$. Particularly, the conditional distribution of t given s is given by

$$\Pr[t | s; \beta] = \frac{c(s, t) \exp(\beta t)}{\sum_{v \in \Gamma} c(s, v) \exp(\beta v)} \text{ with}$$

$$c(s, t) = \sum_{u \in \Omega(t|s)} \prod_k \binom{n_k}{u_k},$$

where $\Omega(t | s) = \{u : u \in \Omega(s) \text{ and } \sum_k x_k u_k = t\}$ and $\Gamma_s = \{v : v = \sum_k x_k u_k \text{ and } u \in \Omega(s)\}$.

To illustrate the implementation of the hypothesis testing method based on the conditional exact approach, we consider the following famous chi-square Cochran-Armitage statistic for linear trend^[7,8] $X^2(s, t) = t^2 / V_t$, where $V_t = [s(n - s) / (n(n - 1))] \sum_k n_k (x_k - \bar{x})^2$ and $\bar{x} = \sum_k x_k / K$. To determine the p -value of $X^2(s, t)$ under H_0 , we note that the conditional distribution of t given s is

$$f_0(t | s) = \frac{c(s, t)}{\sum_{v \in \Gamma} c(s, v)}.$$

Hence, an exact conditional p -value based on $X^2(s, t)$ can be computed as

$$p_e = \sum_{X^2(s, t) \leq X^2(s, v)} f_0(v | s).$$

We may then conclude the existence of a dose-response trend at γ level if p_e is less than or equal to γ . It is noteworthy that a lot of existing statistics (e.g., Helmert, reverse-Helmert, maximin, pairwise, linear, analogon, and reverse-analogon statistics) have been derived for detecting dose-response trend for dichotomous endpoints,^[9] and the above conditional exact p -value calculation method can be readily extended to these statistics. For the confidence interval method based on the exact approach, the exact $(1 - \gamma) \times 100\%$ confidence interval for β is given by (β_L, β_U) , where β_L and β_U satisfy the following equations

$$G(\beta_L; t) = 1 - \Pr[T < t | s; \beta_L] = \gamma / 2 \text{ and}$$

$$H(\beta_U; t) = \Pr[T \leq t | s; \beta_U] = \gamma / 2.$$

Again, a dose-response trend exists if 0 is not in (β_L, β_U) . For the carcinogenicity study with benzyl chloride in Table 1, we have $p_e < 0.0001$, and there exists a dose-response trend between the carcinoma rate and dosage of benzyl chloride.

Table 1 Proportions of Mice That Developed Forestomach Carcinomas and Papillomas in the Carcinogenicity Study with Benzyl Chloride

Forestomach carcinomas and papillomas	Dosage (mg/kg body weight)		
	0	50	100
Yes	0	4	32
No	51	48	20

Moreover, the 95% exact confidence interval for β is given by (1.96, 4.46), whose lower confidence limit is greater than 0, and a positive trend between the carcinoma rate and dosage of benzyl chloride can be concluded.

The methods and approaches discussed in this section have been available from the LOGISTIC procedure in Release 8.1, SAS/STAT software. Readers should notice that the linear logistic regression model has been assumed throughout this section. Unfortunately, methods (especially the Cochran-Armitage statistic) discussed in this section are sometimes wrongly applied and erroneously interpreted as a linear trend analysis without scrutiny of the fit of the underlying linear logistic model.^[10] Hence, readers are strongly recommended to test the linearity assumption before a linear trend is claimed.

In theory, exact unconditional inference can be similarly developed for logistic regression analysis.^[11] Since the exact unconditional method involves tedious computations (e.g., maximizing the nuisance parameters over their support and exhausting all possible data configurations), it seems to be applicable to simple situations such as 2×2 contingency table analyses.

ESTIMATION

Trend Estimation by Using an Empirical Log-Odds Ratio

In practice, we can use an empirical log-odds ratio to estimate the trend parameter β . Without loss of generality, let the first dose class be the baseline group with zero dosage and let y_k be the empirical log-odds ratio or adjusted log-odds ratio at the k th dose level, compared with the baseline group for $k = 2, \dots, K$. If we define $y = (y_2, \dots, y_K)'$, then an approximation approach for trend estimation is to fit a linear regression through the origin, i.e.,

$$y_k = \beta x_k + s_k \varepsilon_k, \quad \text{for } k = 2, \dots, K, \quad (5)$$

where ε_k 's are assumed to be standard normal errors, and s_k is the standard error of y_k . For crude log-odds ratio, s_k can be approximated by $1/r_1 + 1/(n_1 - r_1) + 1/r_k + 1/(n_k - r_k)$.^[12] In this model, we do not consider group correlation for log-odds ratios. Applying least squares to Eq. 5, we obtain the estimate of the dose-response regression slope $\hat{\beta}$ and its

$$\text{variance as } \hat{\beta} = \frac{\sum_{i=1}^m \hat{\beta}_i / s_{\beta i}^2}{\sum_{i=1}^m 1 / s_{\beta i}^2}, \text{ and } \text{var}(\hat{\beta}) = \frac{1}{\sum_{i=1}^m 1 / s_{\beta i}^2}.$$

Since the log-odds ratios are calculated with the same baseline group, they are correlated. To allow for within-study correlation between log-odds ratios, we use the following model^[13]

$$y_k = \beta x_k + \varepsilon_k, \text{ with} \quad (6)$$

$$\varepsilon = (\varepsilon_2, \dots, \varepsilon_K) \sim N(0, \Sigma),$$

where the $(j-1, k-1)$ -th off-diagonal element of Σ is given by $r_{jk} s_j s_k$ with $r_{jk} \approx v_j v_k / (v_j v_k)^{1/2}$, $v_1 = 1/r_1 + 1/(n_1 - r_1)$, and $v_k = 1/r_k + 1/(n_k - r_k)$ for $j, k = 2, \dots, K$. Here, v_k is the approximated standard error of the crude log-odds ratio. The estimate of β and its standard error are then

$$\text{given by } \hat{\beta} = \frac{\sum_{k=2}^K x_k y_k / s_k^2}{\sum_{k=2}^K x_k^2 / s_k^2}; \quad s_{\beta}^2 = \frac{1}{\sum_{k=2}^K x_k^2 / s_k^2} \text{ where } x = (x_2, \dots, x_K)'$$

Trend Estimation with Grouped Data

In trend estimation, the dose (exposure) levels for each subject are often not recorded exactly, but are grouped into class intervals. We suppose that exposure (dose) is an underlying continuous covariate belonging to observable intervals J_k , so that x_k in J_k . For the example of breast cancer and alcohol use, alcohol consumption can only be recorded as grouped data, because of the nature of the questionnaire, which are shown as the first column in Table 2. The customary method of analysis is to assign a single value x_k as a representative of J_k . In Table 2, the assigned values for x_k are listed in the second column. If the class intervals cover a wide range of levels of exposure, the assigned-value method is clearly unsatisfactory. Different choices of assigned values will give different estimates. Thus, assigning a single value to each J_k leads to inaccurate estimates and the underestimation of variance.^[14]

Suppose the exposure levels of all individuals in a particular study are sampled from some distribution with a probability density function $f(x)$. Then, if the probability of being a case given dose x is π_x , the probability that an individual in class interval J is a case is

$$\pi_J = \frac{\int_J \pi_x f(x) dx}{\int_J f(x) dx}. \quad (7)$$

Table 2 Case-control data on alcohol use and breast cancer

Alcohol (g/day)	Assigned dose x	No. of cases	No. of controls	Crude OR	Adjusted OR (CI)
0	0	165	172	1.00	1.00
<2.5	2	74	93	0.83	0.80 (0.51–1.27)
2.5–9.3	6	90	96	0.98	1.16 (0.73–1.85)
>9.3	11	122	90	1.41	1.57 (0.99–2.51)

The observed number of individuals in this class interval who are cases therefore follows a binomial distribution with Eq. (7) as the probability of success. If we continue to adopt the logit model with $\pi_x = 1 / (1 + \exp(\alpha + \beta x))$, the likelihood for (α, β) is still given by Eq. 2 with π_k being replaced by π_{j_k} in Eq. 7. Following the derivation given in Shi and Copas,¹⁴ β can be estimated by maximizing the following log-likelihood function:

$$\hat{\beta} = s_{\beta}^2 x' \Sigma^{-1} y, \quad s_{\beta}^2 = [x' \Sigma^{-1} x]^{-1}$$

where $A(\beta) = (A_2(\beta), \dots, A_k(\beta))$, $A_k(\beta) = \log[p_k(\beta) / (1 - p_k(\beta))]$, and $-\frac{1}{2} \log |\Sigma| - \frac{1}{2} \{y - A(\beta)\}' \Sigma^{-1} \{y - A(\beta)\}$, with $q_k = \int_{j_k} f(x) dx$ and $\hat{\alpha} = \log[r_1 / (n_1 - r_1)]$. This approach is also approximately valid for an adjusted log-odds ratio, provided the covariate adjustments are not too large.

To implement this procedure, we need to know the dose distribution $f(x)$. We may use a parametric model fitted to the observed frequencies of the dose intervals. For the example of breast cancer and alcohol use, a log-normal distribution is used. In principle, any parametric or non-parametric method could be used to estimate the density function $f(x)$. Prior knowledge from historical data can also be incorporated.

COMBINING DOSE-RESPONSE DATA

In epidemiology and other areas, different studies on a similar topic may fail to give meaningful or statistically significant conclusions addressing a particular issue. Meta-analysis is used to summarize and combine results from these individual studies and provide an overall conclusion.^[15]

Meta-Analysis for Trend Estimation

Suppose there are m studies in a meta-analysis, with the log-odds ratios (or their adjusted values) from individual studies satisfying model Eq. (6). Let $\hat{\beta}_i$ and s_{β_i} be the estimate of β_i and its standard error calculated from Eq. (6) for $i = 1, \dots, m$. A fixed-effect meta-analysis model assumes that the true value of β is constant across the studies, i.e.,

$$\hat{\beta}_i \sim N(\beta, s_{\beta_i}^2). \quad (8)$$

This gives the pooled estimate of the trend parameter and its variance as

$$\sum_{i=1}^m \left\{ -\frac{1}{2} \log(\tau^2 + s_{\beta_i}^2) - \frac{1}{2} (\hat{\beta}_i - \beta)^2 / (\tau^2 + s_{\beta_i}^2) \right\}.$$

Since the design quality and sample selection could vary extremely among different studies,^[16] this may lead to the

problem of heterogeneity in combining dose-response data. To allow for heterogeneity among different studies, we assume that

$$\hat{\beta}_i \sim N(\beta, s_{\beta_i}^2 + \tau^2). \quad (9)$$

Here, the coefficient β measures the overall association between log-odds ratio and dose level, and τ measures the magnitude of the variation between studies. The maximum likelihood estimate of (β, τ) is calculated by maximizing

$$p_k(\beta) = \frac{1}{q_k} \int \frac{\exp(\hat{\alpha} + \beta x)}{1 + \exp(\hat{\alpha} + \beta x)} f(x) dx$$

The standard error of β is calculated by $1 / \sum (\tau^2 + s_{\beta_i}^2)^{-1}$, where τ is evaluated at its maximum likelihood estimate. We can use a hypothesis testing method to test $H_0: \beta = 0$. The p -value can be calculated by using, for example, the likelihood ratio test. The confidence interval method can also be used here. This is an easy approach, which can be readily extended to grouped values of x_k discussed in the previous section. A comprehensive discussion for meta-analysis and trend estimation is given in Shi and Copas.^[14]

The method discussed in this section is based on the normal approximation of $\hat{\beta}_i$ in Eq. 8 or Eq. (9). However, if the total sample size in some studies is small, the normal approximation may be not valid. Shi^[17] developed an exact approach based on binomial distributions. The method also allows for the heterogeneity among different studies and addresses the problem of publication bias.

EXAMPLE: ASSOCIATION BETWEEN BREAST CANCER AND ALCOHOL USE

Inference in a Single Study

Table 2,^[18] reports data from a case-control study investigating the association between the consumption of alcohol and the risk of breast cancer. Here, for each of the 451 cases and 415 controls, we have an estimate of alcohol consumption grouped into one of four exposure bands, including the baseline group with zero dose. Each row gives an estimate of the odds of being a case versus being a control, and the estimates of the odds ratios reported in the fifth column. In practice, as in this study, we usually adjust these estimates according to other covariates known to be related to risk, such as age. The sixth column in Table 2 gives the adjusted odds ratios and their confidence limits.

We first take the x_k 's to be the assigned dose levels listed in the second column of Table 2, and use model Eq. (6) to allow for within-group correlation for log-odds ratios. Using the formulas given after model Eq. 6 yields the maximum likelihood estimate $\hat{\beta} = 0.0479$ with standard error 0.205.

However, due to the nature of the questionnaire, we can only record a class interval for a dose level, which is reported in the first column of Table 2. By using all the observed frequencies of the dose intervals for similar studies (they are the data collected and used in meta-analysis as discussed below), the estimated density function $f(x)$ for dose level has a log-normal distribution. By using the method for grouped dose level, we have the estimates $\hat{\beta} = 0.019$ and the standard error 0.0094.

Meta-Analysis and Trend Estimation

Longnecker et al.^[19] and Greenland and Longnecker^[20] conducted a system review for studying the overall association between alcohol consumption and risk of breast cancer. We consider 13 studies here to compare different methods for meta-analysis. We first use the same assigned values for x_k as used by those authors of these individual studies. Those assigned values are given based on the authors' experiences or some historical data, and therefore may not be very accurate. This method gives the overall estimate of the trend parameter as 0.0088, with standard error 0.0014, based on the fixed-effect model Eq. 8. On the other hand, the random-effect model Eq. 9 produces the random-effect trend estimate of 0.0112, with standard error 0.0026 and $\hat{\tau} = 0.0065$. The corresponding p -value for $H_0: \beta = 0$ is less than 0.001, suggesting that the association between breast cancer and the consumption of alcohol is highly significant.

Suppose that alcohol consumption has a log-normal distribution over all the subjects in these studies. Taking a log-transformation for alcohol consumption and fitting all the observed frequencies of the dose intervals to a normal distribution gives the estimated normal distribution, with mean 1.533 and standard error 1.380. We next take $f(x)$ to be this fitted log-normal distribution and maximize the log-likelihood after model Eq. 7. This gives the estimate of β_i in each of the individual studies. The results are reported in the columns named "Grouped values approach" in Table 3.

For comparison purpose, we set the x_k 's by their expected log dose calculated from the fitted dose distribution (i.e., the log-normal distribution). This provides the column named "Mean approach" in Table 3. Comparing the results reported in the first paragraph of this section with the results of "Mean approach" reported in Table 3, we confirm that the estimates are very sensitive to the choices of the assigned values of dose levels.

When the studies are pooled together using the fixed effects model, the assigned values approaches suggest a much stronger dose trend than the grouped values approach. Based on the random effects model, the estimates from the assigned values approaches are even higher. The value of $\hat{\tau}$ for the grouped values approach is much smaller than that for the assigned values approaches. In fact, the fixed and random effects models are almost indistinguishable when

Table 3 Meta Analysis Estimates for Alcohol and Breast Cancer Studies

Study number	Mean approach		Grouped values approach	
	β	SE	β	SE
1	0.0054	0.0028	0.0027	0.0016
2	0.0111	0.0044	0.0070	0.0029
3	0.0149	0.0030	0.0067	0.0010
4	0.0328	0.0164	0.0157	0.0069
5	0.0107	0.0033	0.0087	0.0030
6	0.0256	0.0080	0.0204	0.0076
7	0.0030	0.0033	0.0024	0.0027
8	0.0000	0.0072	0.0000	0.0043
9	0.0026	0.0030	0.0020	0.0024
10	0.0211	0.0089	0.0190	0.0094
11	0.0081	0.0041	0.0053	0.0035
12	0.0112	0.0054	0.0080	0.0045
13	0.0092	0.0067	0.0066	0.0054
Pooled estimate by fixed effect model				
	0.0084	0.0012	0.0055	0.0007
Pooled estimate by random effect model				
	0.0090	0.0016	0.0054	0.0009
$\hat{\tau}$	0.0034		0.0013	
p -value of $\tau = 0$	0.1164		0.3814	

the grouped values approach is used. It seems that the extra uncertainty through the variability of dose levels within classes has almost entirely explained the heterogeneity. The consequence of having a large value of τ is because more weight is assigned to those smaller studies.

Based on the assigned-value approach, the random effects trend estimate is 0.0112, implying that one extra drink daily (about 13g of alcohol, according to Longnecker^[21]) increases the risk by about 16%. By the grouped values approach, the risk increases only about 7% (from an estimated trend of 0.0054), which is less than half of the percentage reported by the assigned-value approach. The latter seems to be a more reasonable estimate according to a latest extensive meta-analysis.²¹

CONCLUSION

When the total sample size is not too small, we can usually use the asymptotic approach to calculate p -value for hypothesis testing and confidence interval construction for trend parameter β . In practice, we can simply use an empirical log-odds ratio to estimate the value of β . This method can be easily extended to meta-analysis, which combines many small dose-response data to give an overall result.

However, when the total sample size is small, or the data structure is sparse, the exact conditional approach is usually recommended. The exact approaches for hypothesis testing and confidence interval method are discussed in this entry. For meta-analysis based on exact approach, refer to Shi.^[17]

Due to the nature of the experiments in epidemiological studies and other areas, grouped-dose data are quite common. As shown in this entry, the estimate of trend parameters can be very sensitive to the choice of the assigned values of dose levels. We therefore recommend an approach that considers the grouped-dose data discussed in this entry.

The computation involved in the models discussed in this entry is quite straightforward and can be implemented in SAS or R. Program codes for meta-analysis and trend estimation are available at <http://www.staff.ncl.ac.uk/j.q.shi>.

ARTICLES OF FURTHER INTEREST

Analysis of $2 \times K$ Tables; Combination Drug Clinical Trial; Dose Proportionality; Dose-Response Analysis in Clinical Trials; Dose-Response Study Design; Hypothesis Testing; Logistic Regression; Meta-Analysis of Therapeutic Trials; Odds Ratio.

REFERENCES

1. Henschler, D. *Occupational Toxicants: Critical Data Evaluation for MAK Values and Classification of Carcinogens/Commission for the Investigation of Health Hazards of Chemical Compounds in Work Area*; VCH: Weinheim, New York, 1991.
2. Agresti, A. A survey of exact inference for contingency tables. *Stat. Sci.* **1992**, *7*, 131–177.
3. Hirji, K.F.; Tang, M.L. A comparison of tests for trend. *Commun. Stat. Theory Methods* **1998**, *27*, 943–963.
4. Corcoran, C.; Mehta, C.; Patel, N.; Senchaudhuri, P. Computational tools for exact conditional logistic regression. *Stat. Med.* **2001**, *20*, 2723–2739.
5. Tarone, R.E.; Gart, J.J. On the robustness of combined tests for trends in proportions. *J. Am. Stat. Assoc.* **1980**, *75*, 110–116.
6. Cox, D.R. The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B* **1958**, *20*, 215–242.
7. Cochran, W.G. Some methods for strengthening the common tests. *Biometrics* **1954**, *10*, 417–451.
8. Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics* **1955**, *11*, 375–385.
9. Neuhauser, M. Trend tests for dichotomous endpoints with application to carcinogenicity studies. *Drug Info. J* **1997**, *31*, 463–469.
10. Tang, M.L.; Hirji, K.F. Simple polynomial multiplication algorithms for exact conditional tests of linearity in a logistic model. *Comput. Methods Programs Biomed.* **2002**, *69*, 13–23.
11. Tang, M.L. On tests of linearity for dose-response data: asymptotic, exact conditional and exact unconditional tests. *J. Appl. Stat.* **2000**, *27*, 871–880.
12. Cox, D.R.; Snell, E.J. *Analysis of Binary Data*; Chapman and Hall: New York, 1989.
13. Greenland, S.; Longnecker, M.P. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *Am. J. Epidemiol.* **1992**, *135*, 1301–1309.
14. Shi, J.Q.; Copas, J.B. Meta-analysis for trend estimation. *Stat. Med.* **2004**, *23*, 3–19.
15. Greenland, S. Quantitative methods in the review of epidemiologic literature. *Epidemiol. Rev.* **1987**, *9*, 1–30.
16. Thompson, S.G.; Sharp, S.J. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat. Med.* **1999**, *18*, 2693–2708.
17. Shi, J.Q. Meta-Analysis and Latent Variable Models for Binary Data. In *Handbook of Latent Variable and Related Models*; Lee, S.Y.; Kontoghiorghe, E.J., Eds.; Elsevier: London, 2007; Chapter 12.
18. Rohan, T.E.; McMichael, A.J. Alcohol consumption and risk of breast cancer. *Int. J. Cancer* **1988**, *41*, 695–699.
19. Longnecker, M.P.; Berlin, J.A.; Orza, M.J.; Chalmers, T.C. A meta-analysis of alcohol consumption in relation to risk of breast cancer. *JAMA* **1988**, *260*, 652–656.
20. Greenland, S.; Longnecker, M.P. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *Am. J. Epidemiol.* **1992**, *135*, 1301–1309.
21. Longnecker, M.P. Alcoholic beverage consumption in relation to risk of breast cancer: meta-analysis and review. *Cancer Causes Control.* **1994**, *5*, 73–82.

Two-Stage Adaptive Design: Analysis

Shein-Chung Chow, PhD

Duke University School of Medicine, Durham, North Carolina, U.S.A.

Min Lin, PhD

Food and Drug Administration, Silver Spring, Maryland, U.S.A.

Abstract

The purpose of an adaptive design is not only to efficiently identify clinical benefits of the test treatment under investigation, but also to increase the probability of success in the drug development process. One of the commonly considered adaptive designs is a two-stage seamless adaptive design, which addresses study objectives within a single trial that are normally achieved through separate trials. In practice, the two-stage seamless adaptive designs can be classified into four categories depending on study objectives and study endpoints at different stages. In this article, a comparison between the two-stage seamless adaptive designs and the traditional approach is briefly outlined. An overview of the characteristics and analysis methods for four categories of two-stage seamless adaptive designs is provided. Issues and approaches regarding the sample size calculation for a two-stage seamless adaptive design are also discussed.

Keywords: Adaptive designs; Sample size calculation; Seamless phase II/III; Two-stage.

INTRODUCTION

Adaptive design clinical trials have gained much attention over decades. The purpose of an adaptive design is not only to efficiently identify clinical benefits of the test treatment under investigation, but also to increase the probability of success in the drug development process. With an adaptive design, the investigators have a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on accumulating data from subjects in the ongoing study.^[1,2] The use of adaptive design methods in clinical trials may allow the researchers to react earlier to surprises at the interim looks and to correct assumptions used at the planning stage. Regardless of positive or negative interim results of an adaptive design study, its flexible features can be ethical with respect to both efficacy and safety of the treatment under investigation and may better reflect real-world medical practice. However, adaptations applied after the review of interim (unblinded) data may raise some critical concerns to the accurate and reliable evaluation of the test treatment. One major concern is how to control the overall type I error rate at a prespecified level of significance. It is also of concern that the scientific integrity and validity of a trial may not be warranted due to operational biases.

One of the commonly considered adaptive designs is a two-stage seamless adaptive design, which addresses study objectives within a single trial that are normally achieved through separate trials. In practice, a typical two-stage seamless adaptive design consists of two stages (phases),

namely, a learning (or exploratory) phase (Stage 1) and a confirmatory phase (Stage 2).^[3,4] The learning phase provides the investigator the opportunities to obtain information regarding the uncertainty of the test treatment under investigation and to stop the trial early due to safety and/or futility/efficacy based on accrued data. The objective of the confirmatory phase is to confirm the findings observed from the learning phase. At the end of the study, one would use data collected from both stages in the final analysis. The idea behind a two-stage seamless adaptive design trial is to reduce lead time between two traditional studies.

In practice, the commonly employed two-stage seamless adaptive designs include a two-stage seamless phase I/II adaptive design and a two-stage seamless phase II/III adaptive design. For a two-stage seamless phase I/II adaptive design, the objective of the first stage may be biomarker development, whereas the study objective of the second stage is to establish early efficacy. For a two-stage seamless phase II/III adaptive design, the study objective is for treatment selection (or dose finding) in the first stage, and the study objective of the second stage is efficacy confirmation.

In what follows, a comparison between the two-stage seamless adaptive designs and the traditional approach is briefly outlined. An overview of the characteristics and statistical methods for different types of two-stage seamless adaptive designs is provided. Issues and approaches regarding the sample size calculation for a two-stage seamless adaptive design are discussed. Some concluding remarks are then given.

TWO-STAGE SEAMLESS ADAPTIVE DESIGN VS. TRADITIONAL APPROACH

Possible advantages and disadvantages of a two-stage seamless adaptive design have been noted.^[3–5] As mentioned earlier, a two-stage seamless adaptive design may help in reducing lead time between two studies in the traditional approach. For example, it usually takes 6–12 months to kick off the phase III study after the completion of the phase II trial. The lead time between phase II and phase III trials are generally used (i) to clean and lock database, perform statistical analysis, and write a final clinical study report of the phase II trial; and (ii) to finalize the study protocol, initiate participating study site/centers, and obtain institutional review board (IRB) review/approval of the phase III study. As a comparison, the study protocol of a two-stage seamless phase II/III adaptive design covers both phase II (Stage 1) and phase III (Stage 2) studies. Thus, there is no need to resubmit the protocol for IRB approval during the lead time between Stage 1 and Stage 2. In addition, compared to the traditional approach, fewer subjects may be required for a two-stage seamless phase II/III adaptive design because it allows investigators to fully utilize data collected from both stages for a combined analysis. However, the use of a two-stage seamless phase II/III adaptive design may not save the overall time of a product development as it may require extra upfront planning time to provide design justification, possibly perform extensive simulation studies, and go through multiple regulatory review cycles. The start of a two-stage seamless adaptive design study can be significantly delayed due to those processes.

ANALYSIS OF TWO-STAGE SEAMLESS ADAPTIVE DESIGN

Types of Two-Stage Seamless Adaptive Design

Depending upon study objectives and study endpoints at different stages, two-stage seamless adaptive designs can be classified into the four categories^[3,4]: Category I designs—the study objectives and the study endpoints are the same; Category II designs—the study objectives are the same but the study endpoints are different; Category III designs—the study objectives are different but the study endpoints are the same; Category IV designs—the study objectives and the study endpoints are different. Note that different study objectives could be treatment selection or dose finding for Stage 1 and efficacy confirmation for Stage 2, whereas different study endpoints are usually referred to biomarker (surrogate) endpoints, or clinical endpoints with a shorter duration for Stage 1 versus clinical endpoints for Stage 2. Category I design is often viewed as a similar design to a traditional group sequential design with one interim analysis. In this article,

our emphasis is on Category II designs. The results would be similarly obtained to address the issues in Category III and Category IV designs.

Analysis of Category I Designs

As mentioned previously, Category I design with the same study objectives and study endpoints at both stages is similar to a typical group sequential design with one planned interim analysis. However, standard statistical methods for group sequential design may not be appropriate due to various adaptations applied in Category I designs. Available analysis approaches include (i) Fisher's methods for combining independent p -values,^[6–8] (ii) weighted test statistics,^[9] (iii) the conditional error function method,^[10,11] (iv) conditional power approaches,^[12] and (v) Chang's method based on the approach of sum of p -values (MSP).^[13] In this section, for illustration purposes, we apply the general framework of the MSP to the K -stage seamless adaptive designs as outlined in the works of Chang^[13] and Chow and Chang.^[14] The results can be simplified to fit a two-stage seamless adaptive design.

Let us consider a K -stage seamless adaptive design with a hypothesis test at each of K stages. One may want to treat it as a clinical trial with K preplanned interim analyses, whereas the final analysis is the K th interim analysis. The objective of the trial can be written as the intersection of the individual hypothesis tests from the interim analyses:

$$H_0 : H_{01} \cap \dots \cap H_{0K}, \quad (1)$$

where $H_{0i}, i = 1, \dots, K$, is the null hypothesis to be tested based on a subsample at the i th stage. Without loss of generality, the hypothesis test can be given as

$$H_{0i} : \eta_{i1} \geq \eta_{i2} \quad \text{vs.} \quad H_{ai} : \eta_{i1} < \eta_{i2}, \quad (2)$$

for comparing two treatment groups at the i th stage, where η_{i1} and η_{i2} are the responses of the two treatment arms, and we assume bigger values are better. One constraint is that rejection of any $H_{0i} : i = 1, \dots, K$ will result in the same clinical implication (e.g., new treatment is efficacious). Thus, all $H_{0i} : i = 1, \dots, K$ are constructed for testing the same endpoint within a trial. Note that the p -value p_i for the subsample at the i th stage is often uniformly distributed on $[0, 1]$ under H_0 when $\eta_{i1} = \eta_{i2}$. Following the idea of Fisher's combination approach,^[6–8] Chang^[13] considered a linear combination of the p -values to construct a test statistic for multiple-stage adaptive designs up to the k th stage:

$$T_k = \sum_{i=1}^k w_{ki} p_i, \quad i = 1, \dots, k, \quad (3)$$

where $w_{ki} > 0$. T_k is viewed as cumulative evidence against H_0 . As a result, the smaller the T_k is, the stronger the

evidence is. One can also consider $T_k = \sum_{i=1}^k p_i / K$, which can be treated as an average of the evidence against H_0 . Intuitively, stopping rules may be set up as

$$\begin{cases} \text{Stop for efficacy} & \text{if } T_k \leq \alpha_k \\ \text{Stop for futility} & \text{if } T_k \geq \beta_k, \\ \text{Continue} & \text{otherwise} \end{cases} \quad (4)$$

where T_k , α_k , and β_k are monotonic increasing functions of k , $\alpha_k < \beta_k$, $k = 1, \dots, K-1$, and $\alpha_K = \beta_K$. Note that α_k and β_k are referred to as the efficacy and futility boundaries, respectively. For a K -stage seamless adaptive design, it has to pass 1 to $(k-1)$ th stages to get to the k th stage. Thus, Chang^[13] defined a so-called proceeding probability as unconditional probability:

$$\begin{aligned} \psi_k(t) &= P(T_k < t, \alpha_1 < T_1 < \beta_1, \dots, \alpha_{k-1} < T_{k-1} < \beta_{k-1}) \\ &= \int_{\alpha_1}^{\beta_1} \dots \int_{\alpha_{k-1}}^{\beta_{k-1}} \int_{-\infty}^t f_{T_1 \dots T_k}(t_1, \dots, t_k) dt_k dt_{k-1} \dots dt_1, \end{aligned} \quad (5)$$

where $t \geq 0$, t_i , $i = 1, \dots, k$, is the test statistic at the i th stage, and $f_{T_1 \dots T_k}$ is the joint probability density function. Note that $\psi_k(t)$ is an unconditional probability. Therefore, one can calculate the error rate at the k th stage using

$$\pi_k = \psi_k(\alpha_k). \quad (6)$$

Therefore, the experiment-wise type I error rate is the sum of π_k , $k = 1, \dots, K$, because the type I error rates at different stages are mutually exclusive. That is,

$$\alpha = \sum_{k=1}^K \pi_k. \quad (7)$$

Hence, stopping boundaries can be determined with appropriate choices of α_k . The adjusted p -value can be calculated using the same approach in a classic group sequential design.^[15] Note that the p -value is equal to alpha spent $\sum_{i=1}^k \pi_i$ when the test statistic at the k th stage $T_k = t = \alpha_k$ (i.e., just on the efficacy stopping boundary). This is true regardless of which error spending function is used and consistent with the p -value definition of the traditional design. Chang^[13] defined the adjusted p -value corresponding to an observed test statistic $T_k = t$ at the k th stage as

$$p(t; k) = \sum_{i=1}^{k-1} \pi_i + \psi_k(t), \quad k = 1, \dots, K. \quad (8)$$

As compared to p_i in Eq. (3), the stage-wise (unadjusted) p -values from a subsample at the i th stage, $p(t; k)$, are adjusted p -values calculated from the test statistic, which are based on the cumulative sample up to the k th stage

where the trial stops. Note that Eqs. (7) and (8) are valid regardless of how p_i are calculated.

Analysis of Category II Designs

In this section, we consider a Category II two-stage seamless phase II/III adaptive design that has the same study objectives but different study endpoints. Without loss of generality, we further consider that the study endpoint for the Stage 1 (phase II) is a biomarker (or a surrogate endpoint) with continuous measurements, whereas the study endpoint for the Stage 2 (phase III) is a continuous clinical endpoint. Chow et al.^[16] proposed to obtain the predicted values of the clinical endpoint using data collected from the biomarker based on an established relationship between the biomarker and the clinical endpoint. In the final analysis, these predicted values are combined with the data collected at the confirmatory phase (Stage 2) to derive a statistical inference on the treatment effect of the test product under investigation.

Let x_i be the observed value of a biomarker (or a surrogate endpoint) from the i th subject in phase II (Stage 1), $i = 1, \dots, n$, and y_j be the observed value of the primary clinical endpoint from the j th subject in phase III (Stage 2), $j = 1, \dots, m$. We assume that x_i 's and y_j 's are independently and identically distributed with $E(x_i) = \nu$ and $\text{Var}(x_i) = \tau^2$, and $E(y_j) = \mu$ and $\text{Var}(y_j) = \sigma^2$, respectively. For simplicity, assume that x and y can be correlated with the following straight-line relationship

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad (9)$$

where ε is the random error with zero mean and variance ε^2 . ε is assumed to be independent of x . In practice, if we consider that this relationship is well established, then the parameters β_0 and β_1 are assumed to be known. Hence, the observations x_i obtained in Stage 1 can be transformed into $\beta_0 + \beta_1 x_i$ (denoted by $\hat{y}_i$) using Eq. (9). $\hat{y}_i$ are then combined with those observations y_i collected in the Stage 2 to estimate the treatment mean μ . The following weighted mean estimator was proposed by Chow, Lu, and Tse^[16] as

$$\hat{\mu} = \omega \bar{\hat{y}} + (1 - \omega) \bar{y}, \quad (10)$$

where $\bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^n \hat{y}_i$, $\bar{y} = \frac{1}{m} \sum_{j=1}^m y_j$, and $0 \leq \omega \leq 1$. In practice, ω is usually estimated by

$$\hat{\omega} = \frac{n/s_1^2}{n/s_1^2 + m/s_2^2}, \quad (11)$$

where s_1^2 and s_2^2 are the sample variances of $\hat{y}_i$'s and y_j 's, respectively. The corresponding estimator of μ is then denoted by

$$\hat{\mu}_{GD} = \hat{\omega} \bar{\hat{y}} + (1 - \hat{\omega}) \bar{y}, \quad (12)$$

which is referred to as the Graybill-Deal (GD) estimator of μ .^[17]

Based on the GD estimator, the comparison of the equality of the two treatment groups can be made by testing the following hypotheses:

$$H_0: \mu_A = \mu_B \quad \text{vs.} \quad H_1: \mu_A \neq \mu_B. \quad (13)$$

Let $\hat{y}_{it}$ be the predicted value (based on $\beta_0 + \beta_1 x_{it}$) of the clinical endpoint for the i th subject in phase II (Stage 1) and y_{ij} be the observed value of the clinical endpoint from the j th subject in phase III (Stage 2) under treatment, where $t=A$ or B . From Eq. (12), the GD estimator of μ_t is given by

$$\hat{\mu}_{GDt} = \hat{\omega}_t \bar{\hat{y}}_t + (1 - \hat{\omega}_t) \bar{y}_t, \quad (14)$$

where $\bar{\hat{y}}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} \hat{y}_{it}$, $\bar{y}_t = \frac{1}{m_t} \sum_{j=1}^{m_t} y_{tj}$, and $\hat{\omega}_t = \frac{n_t/s_{1t}^2}{n_t/s_{1t}^2 + m_t/s_{2t}^2}$, with s_{1t}^2 and s_{2t}^2 being the sample variances of $(\hat{y}_{t1}, \dots, \hat{y}_{tn_t})$ and $(y_{t1}, \dots, y_{tm_t})$, respectively. The following test statistic can be used to test hypotheses (Eq. 12):

$$\tilde{T} = \frac{\hat{\mu}_{GDA} - \hat{\mu}_{GDB}}{\sqrt{\widehat{Var}(\hat{\mu}_{GDA}) + \widehat{Var}(\hat{\mu}_{GDB})}} \quad (15)$$

where

$$\widehat{Var}(\hat{\mu}_{GDt}) = \frac{1}{n_t/s_{1t}^2 + m_t/s_{2t}^2} \left[1 + 4\hat{\omega}_t(1 - \hat{\omega}_t) \left(\frac{1}{n_t - 1} + \frac{1}{m_t - 1} \right) \right]$$

is an exact expression of the variance of $\widehat{Var}(\hat{\mu}_{GDt})$ in the form of an infinite series.^[18] Thus, an approximate 100(1 - α)% confidence interval of $\mu_1 - \mu_2$ is given as

$$\left(\hat{\mu}_{GDA} - \hat{\mu}_{GDB} - z_{\alpha/2} \sqrt{\widehat{V}_T}, \hat{\mu}_{GDA} - \hat{\mu}_{GDB} + z_{\alpha/2} \sqrt{\widehat{V}_T} \right), \quad (16)$$

where $\widehat{V}_T = \widehat{Var}(\hat{\mu}_{GDA}) + \widehat{Var}(\hat{\mu}_{GDB})$. Therefore, the null hypothesis H_0 will be rejected if the above confidence interval does not contain 0.

Analysis for Category III and IV Designs

In this section, we will discuss the statistical inference for Category III and IV designs. For a Category III design, the study objectives at different stages are different (e.g., dose selection at Stage 1 vs. efficacy confirmation at Stage 2), whereas the study endpoints are the same at different stages. For a Category IV design, both study objectives and endpoints at different stages are different (e.g., dose selection using surrogate endpoint at Stage 1 vs. efficacy confirmation with clinical endpoint at Stage 2).

Let us consider a seamless phase II/III adaptive design. We suppose a surrogate (biomarker) endpoint and a well-established clinical endpoint are available for the treatment effect assessment. For comparing k treatments groups, $E_1, \dots, E_k$, with a control group C , let θ_i and ψ_i , $i = 1, \dots, k$, be the treatment effect assessed by the surrogate (biomarker) endpoint and the clinical endpoint for comparing E_i with C , respectively. Hence, the following hypothesis can be adopted to test the treatment effect using the surrogate (biomarker) endpoint:

$$H_{0,1}: \theta_1 = \dots = \theta_k, \quad (17)$$

whereas the hypothesis

$$H_{0,2}: \psi_1 = \dots = \psi_k \quad (18)$$

is used for testing the treatment effect with the clinical endpoint.

In practice, the investigators may have one or more interim analyses at each stage under a Category III or IV design. Cheng and Chow^[19] considered a seamless phase II/III adaptive design with two interim analyses at Stage 2 (phase III) and assumed that ψ_i is a monotone increasing function of the corresponding θ_i . The hypotheses (17) and (18) will be tested at four critical milestone points (i.e., end of Stage 1 analysis, Stage 2a analysis, Stage 2b analysis, and final analysis) based on accrued data. Their proposed tests are briefly described next. For simplicity, the variances of the surrogate (biomarker) endpoint and the clinical outcome are assumed to be known and denoted by σ^2 and τ^2 .

End of Stage 1 analysis—At Stage 1, $(k + 1)n_1$ subjects are randomized to either one of the k treatment arms or the control arm at a 1:1 ratio with n_1 subjects in each group. At the end of Stage 1 analysis, the most effective treatment will be selected compared to the control based on the surrogate endpoint. The selected treatment group along with the control group will then proceed to the subsequent stage. For pairwise comparison, Cheng and Chow^[19] considered test statistics $\hat{\theta}_{i,1}$, $i = 1, \dots, k$ and $S = \arg \max_{1 \leq i \leq k} \hat{\theta}_{i,1}$ (i.e., the arguments of the maxima for the test statistics $\hat{\theta}_{i,1}$, $i = 1, \dots, k$). In this case, if $\hat{\theta}_{S,1} \leq c_1$ for some prespecified critical value c_1 , then the trial is stopped for futility and we are in favor of $H_{0,1}$; and if $\hat{\theta}_{S,1} > c_{1,1}$ for a prespecified critical value $c_{1,1}$, the treatment E_S is selected to be the most promising treatment among the k treatments. Only those subjects who receive either the promising treatment or the control will be followed for the clinical endpoint. All other subjects will undergo necessary safety monitoring without treatment assessment.

Stage 2a analysis—At Stage 2a, $2n_2$ additional subjects are equally assigned to receive either the treatment E_S or the control C . The Stage 2a analysis is planned when the surrogate endpoint measures from $2n_2$ Stage 2 subjects and the clinical endpoint measures from $2n_1$ Stage 1

subjects who receive either the treatment E_s or the control C become available. Let us denote $T_{1,1} = \hat{\theta}_{s,1}$ and $T_{1,2} = \hat{\psi}_{s,1}$ as the test statistics from Stage 1 based on the surrogate endpoint and the primary endpoint, respectively. Let $\hat{\theta}_{s,2}$ be the test statistic based on the surrogate endpoint from Stage 2. We define

$$T_{2,1} = \sqrt{\frac{n_1}{n_1 + n_2}} \hat{\theta}_{s,1} + \sqrt{\frac{n_2}{n_1 + n_2}} \hat{\theta}_{s,2} \quad (19)$$

If $T_{2,1} \leq c_{2,1}$ for a prespecified critical value $c_{2,1}$, then we accept $H_{0,1}$ and stop the trial for futility; if $T_{2,1} > c_{2,1}$ and $T_{1,2} > c_{1,2}$, then we reject both $H_{0,1}$ and $H_{0,2}$, and stop the trial for efficacy; otherwise, if $T_{2,1} > c_{2,1}$ but $T_{1,2} \leq c_{1,2}$, then we will move on to Stage 2b.

Stage 2b analysis—At Stage 2b, no additional subjects will be recruited. The Stage 2b analysis will be performed when $2n_2$ Stage 2a subjects receive their clinical endpoint assessment. Let

$$T_{2,2} = \sqrt{\frac{n_1}{n_1 + n_2}} \hat{\psi}_{s,1} + \sqrt{\frac{n_2}{n_1 + n_2}} \hat{\psi}_{s,2}, \quad (20)$$

where $\hat{\psi}_{s,2}$ is the test statistic from Stage 2b. We will reject $H_{0,2}$ and stop the trial for efficacy if $T_{2,2} > c_{2,2}$ for a prespecified critical value $c_{2,2}$. Otherwise, we continue the enrollment to reach the prespecified number of subjects.

Final analysis—We will recruit $2n_3$ additional subjects and follow them till their primary clinical endpoints; then the final analysis will be performed. Let

$$T_3 = \sqrt{\frac{n_1}{n_1 + n_2 + n_3}} \hat{\psi}_{s,1} + \sqrt{\frac{n_2}{n_1 + n_2 + n_3}} \hat{\psi}_{s,2} + \sqrt{\frac{n_3}{n_1 + n_2 + n_3}} \hat{\psi}_{s,3}, \quad (21)$$

where $\hat{\psi}_{s,3}$ is the test statistic based on $2n_3$ additional subjects. If $T_3 > c_3$ for a prespecified critical value c_3 , then we will reject $H_{0,2}$ and declare efficacy; otherwise, accept $H_{0,2}$.

In the above design, all the parameters, $n_1, n_2, n_3, c_1, c_{1,1}, c_{2,1}, c_{2,2}$, and c_3 are determined such that an overall type I error rate of at a prespecified level of significance α will be controlled with a target power of $1 - \beta$. Note that the most promising treatment is selected based on the surrogate data rather than assessing $H_{0,2}$ at Stage 1. In other words, a dose does not need to be significant upon completion of Stage 1 in order to move forward.

The above approach is essentially a group sequential procedure with treatment selection at interim. No additional adaptations are involved. Cheng and Chow^[19] discussed a sample size reestimation procedure within the above design framework. Assume that ψ and θ are correlated with a local linear relationship. This assumption is reasonable if we focus only on the values at a neighborhood of the most promising treatment E_s . Therefore, at

the end of Stage 2a, the treatment effect of the primary endpoint can be reestimated using

$$\hat{\delta}_s = \frac{\hat{\psi}_{s,1}}{\hat{\theta}_{s,1}} T_{2,1}.$$

Thus, the sample size for final analysis can be reestimated based on a modified treatment effect of the primary endpoint $\delta = \max\{\hat{\delta}_s, \delta_0\}$, where δ_0 is a minimal clinically relevant treatment effect. Let m be the reassessed sample size for final analysis based on δ . If $m \leq n_3$ (originally planned), then there is no modification for the procedure; if $m > n_3$, then m subjects (instead of n_3) per arm will be recruited. Such an adaptation allows to adjust the treatment effect of the clinical outcome with a validated relationship between the surrogate endpoint and the clinical outcome.

SAMPLE SIZE CALCULATION FOR TWO-STAGE SEAMLESS ADAPTIVE DESIGN

When we apply a two-stage seamless adaptive design clinical trial, one of the questions commonly asked is how to perform sample size calculation. For Category I designs, Chow and Chang^[14] described the methods based on individual p values. These methods, however, are not appropriate for Category II–IV designs. Chow and Tu^[3] derived formulas and/or procedures for sample size calculation for Category II designs with various data types including continuous, discrete (binary response), and time-to-event endpoints assuming a well-established relationship between the Stage 1 endpoint and the Stage 2 endpoint.

Let us consider the same settings in Section “Analysis of Category II Designs” and suppose we will test the hypotheses in Eq. (13). Thus, under the local alternative hypothesis $H_1: \mu_A - \mu_B = \delta \neq 0$, the required sample size for achieving a $1 - \beta$ power at the significance level of α satisfies

$$-z_{\alpha/2} + |\delta| / \sqrt{\text{Var}(\hat{\mu}_{GDA}) + \text{Var}(\hat{\mu}_{GDB})} = z_\beta.$$

If we let $m_t = \rho n_t$, where $t = A$ or B , and $n_B = \gamma n_A$. Therefore, the total sample size, denoted by N_T , is given by $N_T = (1 + \rho)(1 + \gamma)n_A$. Then, we solve n_A as

$$n_A = \frac{1}{2} UV \left(1 + \sqrt{1 + 8(1 + \rho) \frac{W}{U}} \right), \quad (22)$$

where $U = \frac{(z_{\alpha/2} + z_\beta)^2}{\delta^2}$, $V = \frac{\sigma_A^2}{\rho + r_A^{-1}} + \frac{\sigma_B^2}{\gamma(\rho + r_B^{-1})}$, and

$W = V^{-2} \left[\frac{\sigma_A^2}{r_A(\rho + r_A^{-1})^3} + \frac{\sigma_B^2}{\gamma^2 r_B(\rho + r_B^{-1})^3} \right]$, with $r_t = \beta_1^2 \tau_t^2 / \sigma_t^2$, $t = A$ or B .

For testing a superiority hypothesis $H_1: \mu_A - \mu_B = \delta_1 > \delta$, the required sample size to achieve a $1 - \beta$ power at the significance level of α satisfies

$$-z_\alpha + (\delta_1 - \delta) / \sqrt{\text{Var}(\hat{\mu}_{GD_A}) + \text{Var}(\hat{\mu}_{GD_B})} = z_\beta.$$

This gives

$$n_A = \frac{1}{2} V X \left(1 + \sqrt{1 + 8(1 + \rho) \frac{W}{X}} \right), \quad (23)$$

$$\text{where } X = \frac{(z_\alpha + z_\beta)^2}{(\delta_1 - \delta)^2}.$$

If one wants to test for equivalence with the local alternative hypothesis that $H_1: \mu_A - \mu_B = \delta_1$, where $|\delta_1| < \delta$. The required sample size to achieve $1 - \beta$ power with a significance level α satisfies

$$-z_\alpha + (\delta - \delta_1) / \sqrt{\text{Var}(\hat{\mu}_{GD_A}) + \text{Var}(\hat{\mu}_{GD_B})} = z_\beta.$$

Thus, n_A is given by

$$n_A = \frac{1}{2} V Y \left(1 + \sqrt{1 + 8(1 + \rho) \frac{W}{Y}} \right), \quad (24)$$

$$\text{where } Y = \frac{(z_\alpha + z_{\beta/2})^2}{(\delta - |\delta_1|)^2}.$$

Under Category II designs, formulas for sample size calculation for binary response and time-to-event endpoints can be similarly derived for testing equality, non-inferiority, superiority, and equivalence. For Category III and IV designs, however, the above derived formulas for sample size calculation should be modified for controlling the overall type I error at a prespecified level for achieving a target power.

CONCLUDING REMARKS

In summary, a two-stage seamless adaptive design can be attractive to investigators due to its potential efficiency, including reducing the lead time between two traditional studies (e.g., a phase II trial and a phase III study), saving subjects by combining data collected from both stages for a final analysis, and possibly making an early decision (e.g., stop the trial early for futility). Nonetheless, as mentioned earlier, to achieve the above advantages for implementing a two-stage seamless adaptive design, the investigators usually need to spend extra time on the planning stage. Hence, the overall time of a product development may be actually longer than adopting a traditional approach (i.e., two separated studies).

In practice, as discussed earlier, two-stage seamless adaptive designs can be classified into four categories based on the study objectives and the study endpoints used at different stages. In the case when the study objectives and/or the study endpoints are different at different stages, it is important to determine how to combine the data collected from both stages to perform the final analysis. It is also critical to know how the sample size calculation should be done for achieving a desired overall type I error rate with a target power to meet the study objectives originally set for the two stages (separate studies).

Chow and Chang^[14] expressed concerns that the standard statistical methods derived for a group sequential design trial with one planned interim analysis are often directly applied to a two-stage seamless adaptive design regardless of whether the study objectives and/or the study endpoints are the same at different stages. In many cases, especially when the study objectives and/or study endpoints at different stages are different, the obtained p -value and confidence interval for assessment of the treatment effect may not be correct or reliable with the direct application of standard statistical methods. Moreover, sample size obtained under a standard group sequential trial design may not be sufficient for achieving the study objectives under a two-stage seamless adaptive design. In practice, the major challenge for utilizing a two-stage seamless adaptive design is not only the development of valid statistical methods but also the development of a set of criteria for choosing the most appropriate design for valid and reliable assessment of test treatment under investigation. Sometimes, a traditional approach (i.e., two separated studies) is more promising.

DISCLAIMER

The views presented in this article have not been formally disseminated by the US Food and Drug Administration and should not be construed to represent any agency determination or policy.

REFERENCES

1. FDA. Draft Guidance for Industry—Adaptive Design Clinical Trials for Drugs and Biologics, 2010; <https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf>.
2. FDA. Adaptive Designs for Medical Device Clinical Studies—Guidance for Industry and Food and Drug Administration Staff, 2016; <https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm446729.pdf>.
3. Chow, S.C.; Tu, Y.H. On two-stage seamless adaptive design in clinical trials. *J. Formos. Med. Assoc.* **2008**, *107* (12), S52–S60.

4. Chow, S.C.; Lin, M. Analysis of Two-Stage Adaptive Seamless Trial Design, 2015; <http://dx.doi.org/10.4172/2153-2435.1000341>.
5. Lin, M.; Lee, S.; Zhen, B.G.; Scott, J.; Horne, A.; Solomon, G.; Russek-Cohen, E. CBER's experience with adaptive design clinical trials. *-Ther. Innov. Regul. Sci.* **2016**, *50* (2), 195–203.
6. Bauer, P.; Kohne, K. Evaluation of experiments with adaptive interim analyses. *Biometrics* **1994**, *50* (4), 1029–1041.
7. Bauer, P.; Rohmel, J. An adaptive method for establishing a dose-response relationship. *Stat. Med.* **1995**, *14* (14), 1595–1607.
8. Posch, M.; Bauer, P. Interim analysis and sample size reassessment. *Biometrics* **2000**, *56* (4), 1170–1176.
9. Cui, L.; Hung, H.M.J.; Wang, S.J. Modification of sample size in group sequential clinical trials. *Biometrics* **1999**, *55* (3), 853–857.
10. Liu, Q.; Chi, G.Y.H. On sample size and inference for two-stage adaptive designs. *Biometrics* **2001**, *57* (1), 172–177.
11. Proschan, M.A.; Hunsberger, S.A. Designed extension of studies based on conditional power. *Biometrics* **1995**, *51* (4), 1315–1324.
12. Li, G.; Shih, W.C.J.; Wang, Y.N. Two-stage adaptive design for clinical trials with survival data. *J. Biopharm. Stat.* **2005**, *15* (4), 707–718.
13. Chang, M. Adaptive design method based on sum of p -values. *Stat. Med.* **2007**, *26* (14), 2772–2784.
14. Chow, S.C.; Chang, M. *Adaptive Design Methods in Clinical Trials*, 2nd Ed.; Chapman & Hall/CRC, Taylor & Francis: New York, 2011.
15. Jennison, C.; Turnbull, B.W. *Group Sequential Methods with Applications to Clinical Trials*; Chapman & Hall/CRC: London and Boca Raton, FL, 2000.
16. Chow, S.C.; Lu, Q.S.; Tse, S.K. Statistical analysis for two-stage seamless design with different study endpoints. *J. Biopharm. Stat.* **2007**, *17* (6), 1163–1176.
17. Graybill, F.A.; Deal, R.B. Combining unbiased estimators. *Biometrics* **1959**, *15*, 7.
18. Khatri, C.G.; Shah, K.R. Estimation of location of parameters from two linear models under normality. *Commun. Stat. Theory Methods* **1974**, *3*, 647–663.
19. Cheng, B.; Chow, S.C. Statistical inference for a multiple-stage transitional seamless trials designs with different study objectives and endpoints. Submitted. **2018**.

Two-Stage Design: Phase II Cancer Clinical Trials

Sin-Ho Jung

Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, U.S.A.

Abstract

Phase II trials have been very widely conducted and published every year for cancer clinical research. Because of their small sample sizes, statistical methods based on large sample approximation are not appropriate for design and analysis of phase II trials. As a prospective clinical research, the analysis method of a phase II trial is predetermined at the design stage, and it is analyzed during and at the end of the trial as planned by the design. The analysis method of a trial should be matched with the design method. For two-stage, single-arm phase II trials, Simon's method has been the standard for choosing an optimal design, but the resulting data have been analyzed and published ignoring the two-stage design aspect with small sample sizes. In this article, we review the analysis methods that exactly get along with the exact two-stage design method. We also discuss some statistical methods to improve the existing design and analysis methods for single-arm, two-stage phase II trials.

Keywords: Accrual overrun; Admissible design; Confidence interval; P -value; Two-stage design; Uniformly minimum unbiased estimator.

INTRODUCTION

Cancer clinical trials are to investigate the efficacy and toxicity of experimental cancer therapies. If an appropriate dose level of an experimental drug is determined from a phase I trial, the drug's anticancer activity is assessed through phase II clinical trials. Phase II clinical trials are to screen out inefficacious experimental therapies before they proceed to further investigation through large-scale phase III trials. In order to expedite this process, phase II trials traditionally use a single-arm design to treat patients by experimental therapies only. So, the efficacy of an experimental therapy is compared with that of a standard therapy that is called a historical control. The most popular primary endpoint of phase II cancer clinical trials is tumor response that is measured by the change in tumor size before and during treatment. For a solid tumor, if the size, defined as the sum of the largest diameters of the target tumors, decreases by at least 30% compared to that at the baseline, we call it a partial response by RECIST 1.1.^[1] If the target tumor disappears during treatment, we call it a complete response (CR). Overall response is defined as partial response or CR.

Phase II trials generally require shorter study periods than phase III trials. Consequently, phase II trials have small sample sizes, so that exact statistical methods are preferable to the asymptotic methods for their design and analysis. Various exact methods have been published for phase II trials with binary outcomes such as tumor response. For ethical reasons, two-stage designs

are commonly used for phase II cancer clinical trials. A typical single-arm, two-stage trial with a futility stopping value (a_1), a superiority stopping value (b_1), and a stage 2 rejection value (a) is conducted as follows:

- Stage 1: Treat n_1 patients and count the number of responders X_1 .
 - If $X_1 \leq a_1$, then reject the experimental therapy and stop the trial.
 - If $X_1 \geq b_1$, then accept the experimental therapy and stop the trial.
 - If $a_1 < X_1 < b_1$, then proceed to the second stage.
- Stage 2: Treat an additional n_2 patients and count the number of responders X_2 .
 - If $X_1 + X_2 \leq a$, then reject the experimental therapy.
 - Otherwise, accept the experimental therapy for further investigation.

We usually do not stop the trial early for superiority since there is no ethical issue to continue treating patients with an efficacious therapy, and we want to collect as many data as possible to use when designing a subsequent phase III trial. In this case, we choose $b_1 = n_1 + 1$.

The design and analysis of phase II trials require exact statistical methods accounting for two-stage design and small sample sizes. The most popular for phase II cancer clinical trials has been a two-stage design with a futility stopping only based on Simon's^[2] minimax or optimality criterion.^[3–5] But most publications reporting the results of phase II trials fail to appropriately address these issues.

When the aforementioned two-stage phase II trial is completed, we should be able to accept or reject the experimental therapy based on the critical values and sample sizes (a_1, b_1, a, n_1, n_2) . In a usual clinical trial, we would recruit slightly more patients than required to make up for possible attrition due to ineligibility and dropout. As a result, the observed sample size tends to be different from that specified by the design. In a typical two-stage phase II trial, the sample size at the stopping stage is different from (usually larger, but rarely smaller, than) the planned one. In this case, the critical values at the stopping stage become meaningless. This may be the major reason why investigators do not make clear conclusions from their phase II trials. Addressing this issue, we propose to extend the two-stage testing rule by calculating a p -value or a confidence interval of the true response rate (RR). These methods exactly coincide with the two-stage testing rule if the observed sample sizes are identical to the prespecified (n_1, n_2) . In this article, we discuss further design and analysis methods for single-arm phase II trials with tumor response as the primary endpoint. We further extend Simon's criteria to search for an alternative two-stage design.

OPTIMAL TWO-STAGE DESIGNS

In this section, we focus on the popular two-stage designs with a futility interim test only, i.e., $b_1 = n_1 + 1$. These designs are defined by the number of patients to be treated, n_1 and n_2 , during stages 1 and 2, and rejection values, a_1 and a ($a_1 < a$), so that we specify them by $(a_1/n_1, a/n)$, where $n = n_1 + n_2$ is called the maximal sample size. The values of $(a_1/n_1, a/n)$ are determined based on some prespecified design parameters as described next. Let p_0 denote the maximum unacceptable probability of response which is usually chosen by the RR of a historical control, and p_1 the minimum acceptable probability of response with $p_0 < p_1$. For the true RR p of the experimental therapy, we want to test $H_0: p \leq p_0$ against $H_1: p > p_0$. In this statistical testing, rejecting H_0 means accepting the experimental therapy. Given (p_0, p_1) , we can calculate the type I error probability α and power $1 - \beta$ of a two-stage design $(a_1/n_1, a/n)$ based on the fact that the number of responders from the two stages, X_1 and X_2 , are independent binomial random variables.

Suppose that we want to find a two-stage design with a type I error probability no larger than α^* and a power no smaller than $1 - \beta^*$ for given (p_0, p_1) values. We call $(p_0, p_1, \alpha^*, \beta^*)$ design parameters. Given (p_0, p_1) , there are infinitely many two-stage designs satisfying the $(\alpha^*, 1 - \beta^*)$ constraint. Noting that the sample size of a two-stage trial is a random variable taking n_1 or n , we can calculate the expected sample size of the trial given a true RR of the experimental therapy. Simon^[2] proposes two criteria to select a good two-stage design among these candidate designs. The minimax design minimizes the maximal sample size, n , among the designs satisfying the $(\alpha^*, 1 - \beta^*)$ constraint. On the other

hand, the so-called optimal design minimizes the expected sample size under $H_0: p = p_0$, denoted as EN.

Although the minimax and optimal designs have both been widely used for long, other two-stage designs have been largely ignored in the past. Oftentimes, the two Simon's criteria result in very different two-stage designs, i.e., the minimax design may have an excessively large EN compared to the optimal design and the optimal design may have an excessively large maximal sample size n compared to the minimax design. This results from the discrete nature of the exact binomial method.

Example 1

For the design parameters $(p_0, p_1, \alpha^*, 1 - \beta^*) = (0.1, 0.3, 0.05, 0.85)$, the minimax design is given by $(a_1/n_1, a/n) = (2/18, 5/27)$ and the optimal design by $(1/11, 6/35)$. The maximal sample size n for the minimax design is less than that for the optimal design by $8(=35-27)$. However, the expected sample size EN under H_0 for the optimal design is 18.3, which is only slightly smaller than $EN = 20.4$ for the minimax design. We may consider choosing the minimax design since, compared to the optimal design, it largely saves the maximal sample size while sacrificing the expected sample size by only about $2(=20.4-18.3)$.

In this case, however, there exists a practical compromise without changing the statistical operating characteristics appreciably.

Example 1 (revisited)

For the same design parameters $(p_0, p_1, \alpha^*, 1 - \beta^*) = (0.1, 0.3, 0.05, 0.85)$, the design given by $(a_1/n_1, a/n) = (1/13, 5/28)$ requires only one more patient in the maximal sample size n than the minimax design, but its expected sample size EN under H_0 is very comparable to that of the optimal design (18.7 vs 18.3).

This design is a good compromise between the minimax design and the optimal design.^[6] There can be multiple compromising designs. Jung et al.^[7] show that they are admissible designs in terms of the loss function combining the maximal sample size and the expected sample size under H_0 . Simon's minimax and optimal designs belong to the family of admissible designs too. An interactive computer program to find admissible designs is available in www.dukecancerinstitute.org/research/shared-resources/biostatistics/clinical-trial-design-systems/.

Figure 1 is a snapshot of this program for Example 1. It displays the plot of EN against n for various designs with $n \leq 37$. Simon's minimax design given by $(a_1/n_1, a/n) = (2/18, 5/27)$ is shown at the leftmost in the figure and the optimal design given by $(1/11, 6/35)$ is shown at the very bottom. The program provides the specification of a design $(a_1/n_1, a/n)$ along with $(\alpha, 1 - \beta)$, EN, and the probabilities of early termination for $p = p_0$ and p_1 when the circle representing a candidate design, actually (n, EN) , is clicked with a pointer.

In this section, we have considered two-stage designs with a futility stopping value only. Chang et al.^[8] propose optimal multistage designs with both futility and superiority stopping boundaries by minimizing the average of expected sample sizes under H_0 and H_1 . Mander et al.^[9] extends the admissible design concept to multistage designs with both futility and superiority stopping

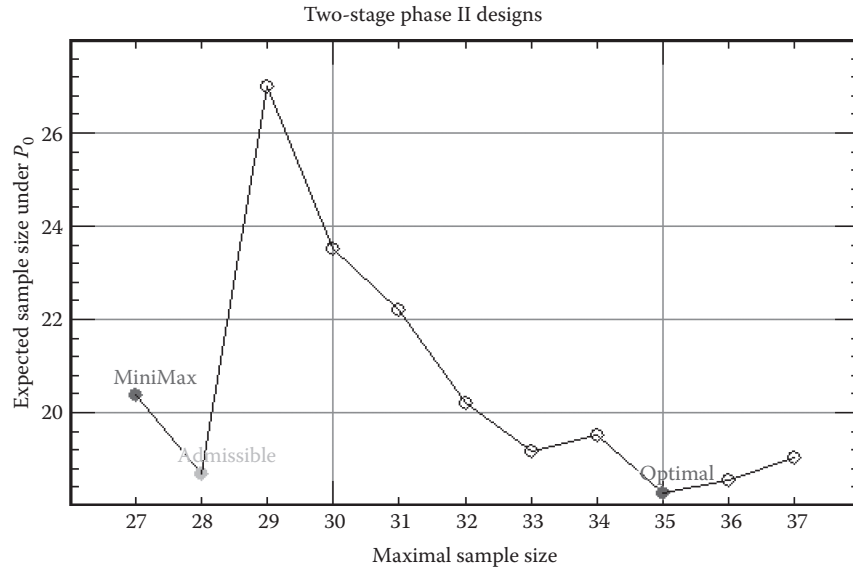


Fig. 1 Two-stage designs for $(p_0, p_1, \alpha^*, \beta^*) = (0.1, 0.3, 0.05, 0.15)$ with $N = 37$

boundaries by using a loss function combining the maximal sample size, the expected sample size under H_0 , and the expected sample size under H_1 . Englert and Kieser^[10] propose optimal adaptive two-stage designs whose stage 2 sample sizes are determined based on the number of responders from stage 1. They do not consider early stopping. Accounting for uncertainty of the true treatment effect, Kieser^[11] proposes to specify an interval of possible true RR in designing a two-stage design.

INFERENCE ON RR

Estimation of RR

In a two-stage phase II trial, let M ($=1$ or 2) denote the stopping stage, and S the cumulative number of responders in the stopping stage, i.e., $S = X_1$ if $M = 1$ and $S = X_1 + X_2$ if $M = 2$. The most popular estimator of RR p for $(M, S) = (m, s)$ is the sample proportion:

$$\hat{p} = \begin{cases} s/n_1 & \text{if } m = 1 \\ s/(n_1 + n_2) & \text{if } m = 2 \end{cases}$$

By not reflecting the two-stage design aspect of the trial, this estimator, called the maximum likelihood estimator (MLE), is always negatively biased for standard two-stage trials with

futility stopping only.^[12] As a result, if p_0 of a future trial is chosen by the sample proportion of a historical control taken from a previous two-stage phase II trial, then it will underestimate the true RR of the historical control and result in a higher chance of false positivity for the future trial.

For $(M, S) = (m, s)$, the uniformly minimum-variance unbiased estimator (UMVUE) of p for two-stage phase II trials is given by

$$\tilde{p} = \begin{cases} s/n_1 & \text{if } m = 1 \\ \frac{\sum_{x_1=(a_1+1) \vee (s-n_2)}^{s \wedge (b_1-1)} \binom{n_1-1}{x_1-1} \binom{n_2}{s-x_1}}{\sum_{x_1=(a_1+1) \vee (s-n_2)}^{s \wedge (b_1-1)} \binom{n_1}{x_1} \binom{n_2}{s-x_1}} & \text{if } m = 2 \end{cases} \quad (1)$$

where $a \wedge b = \min(a, b)$, $a \vee b = \max(a, b)$, $\binom{m}{k} = m! / \{k!(m-k)!\}$, and $x! = x(x-1) \times \cdots \times 2 \times 1$ [12]. Note that the UMVUE and the MLE are identical if the trial stops after stage 1, i.e., $m = 1$.

For a true RR of p , the probability mass function (PMF) of (M, S) , $f(m, s|p) = \Pr(M = m, S = s)$, is given as^[12]

$$f(m, s|p) = \begin{cases} p^s (1-p)^{n_1-s} \binom{n_1}{s} & \text{if } m = 1, 0 \leq s \leq a_1 \text{ or } b_1 \leq s \leq n_1 \\ p^s (1-p)^{n_1+n_2-s} \sum_{x_1=a_1+1}^{(b_1-1) \wedge s} \binom{n_1}{x_1} \binom{n_2}{s-x_1} & \text{if } m = 2, a_1+1 \leq s \leq b_1-1+n_2 \end{cases} \quad (2)$$

Since the UMVUE of p is a function of (M, S) , its PMF is derived from $f(m, s|p)$.

Example 2

Under $(p_0, p_1, \alpha^*, 1 - \beta^*) = (0.2, 0.4, 0.05, 0.8)$, the Simon's optimal two-stage design with a futility stopping value only is given as $(a_1/n_1, a/n) = (3/13, 12/43)$. Table 1 gives the UMVUE and MLE for observations (m, s) and their PMF for various RR values in the two-stage design. When $m = 1$, two estimates are exactly the same as noted earlier. When $m = 2$, the MLE is much smaller than the UMVUE for small s values. The UMVUE is shown to have a comparable variance compared to the MLE overall.^[12]

In order to account for ineligible or unevaluable patients, we often accrue slightly more (or possibly fewer) patients than the

planned sample size at the stopping stage, especially in multi-center trials.^[13,14] But this does not become any issue in the calculation of the UMVUE, since it does not require the specification of the critical values (a_m, b_m) at the stopping stage m , where $a_2 = a$ and $b_2 = a + 1$. If the study is terminated after stage 1 (i.e., $m = 1$), then the UMVUE is calculated by considering the observed number of stage 1 patients as n_1 . If the study is proceeded to the second stage (i.e., $m = 2$), then the UMVUE in Eq. 1 depends only on the rejection values for stage 1 (a_1, b_1) and the number of patients for the two stages (n_1, n_2) . When $m = 2$, the interim test is always conducted using (a_1, b_1) when exactly n_1 patients have response data as planned. However, the number of stage 2 patients may be slightly different from n_2 . In this case, we calculate the UMVUE by considering the observed number of stage 2 patients as n_2 with (n_1, a_1, b_1) fixed by the design.

Table 1 UMVUE, MLE, and PMF for true p for each observation in a two-stage design with $(a_1/n_1, a/n) = (3/13, 12/43)$

m	s	UMVUE	MLE	$f(m, s p)$ for p				
				0.1	0.2	0.3	0.4	0.5
1	0	0.000	0.000	0.254	0.055	0.010	0.001	0.000
1	1	0.077	0.077	0.367	0.179	0.054	0.011	0.002
1	2	0.154	0.154	0.245	0.268	0.139	0.045	0.010
1	3	0.231	0.231	0.100	0.246	0.218	0.111	0.035
2	4	0.308	0.093	0.001	0.000	0.000	0.000	0.000
2	5	0.312	0.116	0.004	0.002	0.000	0.000	0.000
2	6	0.317	0.140	0.007	0.006	0.001	0.000	0.000
	7	0.322	0.163	0.008	0.015	0.002	0.000	0.000
2	8	0.328	0.186	0.006	0.027	0.006	0.000	0.000
2	9	0.335	0.209	0.004	0.038	0.015	0.001	0.000
2	10	0.343	0.233	0.002	0.043	0.030	0.003	0.000
2	11	0.351	0.256	0.001	0.041	0.049	0.008	0.000
2	12	0.360	0.279	0.000	0.033	0.068	0.018	0.001
2	13	0.371	0.302	0.000	0.023	0.081	0.033	0.003
2	14	0.382	0.326	0.000	0.014	0.084	0.054	0.006
2	15	0.395	0.349	0.000	0.007	0.076	0.076	0.013
2	16	0.409	0.372	0.000	0.003	0.062	0.096	0.025
2	17	0.424	0.395	0.000	0.001	0.044	0.107	0.042
2	18	0.440	0.419	0.000	0.001	0.029	0.108	0.063
2	19	0.458	0.442	0.000	0.000	0.017	0.098	0.085
2	20	0.477	0.465	0.000	0.000	0.009	0.080	0.105
2	21	0.496	0.488	0.000	0.000	0.004	0.059	0.116
2	22	0.517	0.512	0.000	0.000	0.002	0.040	0.118
2	23	0.538	0.535	0.000	0.000	0.001	0.025	0.108
2	24	0.560	0.558	0.000	0.000	0.000	0.014	0.091
2	25	0.582	0.581	0.000	0.000	0.000	0.007	0.069
2	26	0.605	0.605	0.000	0.000	0.000	0.003	0.048
2	27	0.628	0.628	0.000	0.000	0.000	0.001	0.030
2	28	0.651	0.651	0.000	0.000	0.000	0.001	0.017
2	29	0.674	0.674	0.000	0.000	0.000	0.000	0.009

(Continued)

Table 1 UMVUE, MLE, and PMF for true p for each observation in a two-stage design with $(a_1/n_1, a/n) = (3/13, 12/43)$ (Continued)

m	s	UMVUE	MLE	$f(m, s p)$ for p				
				0.1	0.2	0.3	0.4	0.5
2	30	0.698	0.698	0.000	0.000	0.000	0.000	0.004
2	31	0.721	0.721	0.000	0.000	0.000	0.000	0.002
2	32	0.744	0.744	0.000	0.000	0.000	0.000	0.001
2	33	0.767	0.767	0.000	0.000	0.000	0.000	0.000
2	34	0.791	0.791	0.000	0.000	0.000	0.000	0.000
2	35	0.814	0.814	0.000	0.000	0.000	0.000	0.000
2	36	0.837	0.837	0.000	0.000	0.000	0.000	0.000
2	37	0.861	0.861	0.000	0.000	0.000	0.000	0.000
2	38	0.884	0.884	0.000	0.000	0.000	0.000	0.000
2	39	0.907	0.907	0.000	0.000	0.000	0.000	0.000
2	40	0.930	0.930	0.000	0.000	0.000	0.000	0.000
2	41	0.954	0.954	0.000	0.000	0.000	0.000	0.000
2	42	0.977	0.977	0.000	0.000	0.000	0.000	0.000
2	43	1.000	1.000	0.000	0.000	0.000	0.000	0.000

Example 2 (revisited)

Suppose that a phase II trial was designed for the Simon's optimal two-stage design with a futility stopping value only given as $(a_1/n_1, a/n) = (3/13, 12/43)$ under $(p_0, p_1, \alpha^*, 1 - \beta^*) = (0.2, 0.4, 0.05, 0.8)$, but it was completed with $s=7$ responders from a total of 45 patients after stage 2. Then, by using $(a_1, n_1) = (3, 13)$ and conditioning n_2 at $32 (= 45 - 13)$, we have $(a_1 + 1) \vee (s - n_2) = (3 + 1) \vee (7 - 32) = 4$ and $s \wedge (b_1 - 1) = 7 \wedge (14 - 1) = 13$. Hence, from Eq. 1, the UMVUE is calculated by

$$\tilde{p} = \frac{\sum_{x_1=4}^7 \binom{12}{x_1-1} \binom{32}{7-x_1}}{\sum_{x_1=4}^7 \binom{13}{x_1} \binom{32}{7-x_1}} = \frac{1,362,988}{4,241,380} = 0.321$$

whereas the MLE is only $\hat{p} = 7/45 = 0.156$.

Bowden and Wason^[15] investigate various estimators of RR including the UMVUE and compare their performance.

Confidence Interval

Jung and Kim^[12] prove that the ordering of the sample space for (M, S) in terms of the magnitude of the UMVUE is identical to that proved by Jennison and Turnbull^[16] (see also Armitage^[17]; Tsiatis et al.^[18]). In other words, we have

$$\begin{aligned} &\tilde{p}(1,0) < \tilde{p}(1,1) < \cdots < \tilde{p}(1,a_1) \\ &< \tilde{p}(2,a_1+1) < \cdots < \tilde{p}(2,b_1-1+n_2) \\ &< \tilde{p}(1,b_1) < \cdots < \tilde{p}(1,n_1) \end{aligned} \quad (3)$$

where $\tilde{p}(m,s)$ denotes the UMVUE for $(M, S) = (m, s)$. Hence, the confidence intervals calculated by Clopper and Pearson^[19] and the stochastic ordering based on the magnitude of the UMVUE are identical to those by Jennison and Turnbull.^[16]

With the UMVUE $\tilde{p}(m,s)$, an exact $100(1 - \alpha)\%$ equal tail confidence interval (p_L, p_U) for p is given by

$$\Pr(\tilde{p}(M,S) \geq \tilde{p}(m,s) | p = p_L) = \alpha/2$$

and

$$\Pr(\tilde{p}(M,S) \leq \tilde{p}(m,s) | p = p_U) = \alpha/2,$$

where the probabilities are calculated using the PMF of (M, S) as defined in Eq. 2. Confidence limits p_L and p_U can be obtained by solving the equations using a numerical method such as the bisection method, which can be described to solve an equation $g(x) = 0$ as follows:

- Find an interval $[x_1, x_2]$ such that $g(x_1)g(x_2) < 0$.
- For $x_3 = (x_1 + x_2)/2$,
 - If $g(x_1)g(x_3) < 0$, then $x_3 \rightarrow x_2$.
 - Otherwise, $x_3 \rightarrow x_1$.
- Continue (B) until $|x_1 - x_2|$ is small enough and choose x_3 as the solution to the equation.

Example 3

Suppose that we observed $(m,s)=(2,7)$ from a two-stage study with lower boundaries only, $(n_1, n_2, a_1, a) = (13, 30, 3, 12)$. In this case, we have $b_1 = n_1 + 1 = 14$, $(a_1 + 1) \vee (s - n_2) = (3 + 1) \vee (7 - 30) = 4$, and $s \wedge (b_1 - 1) = 7 \wedge (14 - 1) = 7$. From Eq. 1, the UMVUE is

$$\tilde{p}(2, 7) = \frac{\sum_{x_1=4}^7 \binom{13-1}{x_1-1} \binom{30}{7-x_1}}{\sum_{x_1=4}^7 \binom{13}{x_1} \binom{30}{7-x_1}} = .322.$$

Using Eq. 2, we have

$$\Pr(\tilde{p}(M, S) \geq .322 | p = .103) = .025$$

and

$$\Pr(\tilde{p}(M, S) \leq .322 | p = .538) = .025$$

so that a 95% confidence interval on p is given as (0.103, 0.538), which is the same as the one according to Jennison and Turnbull.^[16] In contrast, a naive exact 95% confidence interval calculated by Clopper and Pearson^[19] ignoring the two-stage aspect of the study design is given as (0.068, 0.307). Note that the latter is narrower than the former by ignoring the group sequential feature of the study. Furthermore, the former is slightly shifted to the right from the latter to reflect the fact that the study has been continued to stage 2 after observing more responders than $a_1 = 3$ in stage 1. The popularly used asymptotic confidence interval $\hat{p} \pm 1.96 \sqrt{\hat{p}(1-\hat{p})/n} = (.052, .273)$ with 95% significance level is even narrower and further shifted to the left than the naive exact confidence interval.

The Jennison–Turnbull confidence interval based on the stochastic ordering (Eq. 3) has a desirable property: given (p_0, α^*) , the testing result based on the two-stage design $(a_1/n_1, a/n)$ exactly coincides with that based on the one-sided confidence interval with a significance level of $100(1 - \alpha^*)\%$, i.e., we reject $H_0 : p = p_0$ by the two-stage design if and only if the lower confidence limit is larger than p_0 . This property is very valuable especially when the observed sample size for the stopping stage is different from the sample size specified by the study design. In this case, the two-stage design does not provide us a proper testing rule any more since the rejection values, (a_1, b_1) or a , are meaningful only when the number of patients are identical to the prespecified sample size. But we can calculate a Jennison–Turnbull confidence interval of RR by treating the observed sample size at the stopping stage like the planned sample size. This is possible since Eqs. 2 and 3 do not require specification of the rejection values at the stopping stage. Various confidence interval methods have been proposed for multistage designs,^[20,21] but no other confidence intervals have this property.

Example 4

Kim et al.^[22] published a phase II trial of concurrent chemoradiotherapy for newly diagnosed patients with limited-stage extranodal natural killer/T-cell lymphoma, nasal type. The primary endpoint is the CR rate. The study had the Simon's optimal two-stage design $(a_1/n_1, a/n) = (4/6, 22/27)$ under $(p_0, p_1, \alpha^*, 1 - \beta^*) = (0.7, 0.9, 0.05, 0.8)$. The trial proceeded to the second stage to treat a total of 30 eligible and evaluable patients of which 27 achieved CR. Since the observed maximal sample size is different from the planned one, the rejection value $a = 22$ was of no use to determine the positivity of the study. The authors

concluded the study to be positive without statistical ground. Noting this, Shimada and Suzuki^[23] claimed that this trial was a negative study by calculating an asymptotic confidence interval with two-sided 95% significance level and showing that it covers $p_0 = 0.7$. This claim is misleading since their interval (i) has a two-sided $100(1 - \alpha^*) = 95\%$ confidence level while the two-stage design has 1-sided $\alpha^* = 5\%$, (ii) ignores the two stage feature of the trial, and (iii) is using a large-sample approximation. Kim et al.^[24] defended their conclusion by showing that the 1-sided 95% Jennison–Turnbull confidence interval (0.762, 1] is above $p_0 = 0.7$.

P-VALUE CALCULATION

Through a phase II trial, we intend to conduct a statistical testing to reject or accept its experimental therapy. If we reject or fail to reject the null hypothesis, we should be able to provide a p -value as a measure of how strong evidence the decision is based on against the null hypothesis. However, publications from phase II trials rarely report p -values to support their conclusions. In most of the publications, it is not even clear if the investigators conclude to accept their therapies or not.

Using the stochastic ordering of UMVUE (Eq. 3), Jung et al.^[25] propose a p -value method for two-stage phase II clinical trials. Noting that a p -value is defined as the probability to observe an extreme test statistic value toward the direction of H_1 when H_0 is true, they propose to calculate the probability to observe a UMVUE value larger than that obtained from the study under H_0 . Let $\tilde{p}$ denote the UMVUE of the RR observed from a two-stage phase II trial specified by (a_1, b_1, a, n_1, n_2) . Given $(M, S) = (m, s)$, the p -value = $\Pr\{\tilde{p}(M, S) \geq \tilde{p}(m, s) | p_0\}$ based on UMVUE can be calculated as

$$p\text{-value} = \begin{cases} \sum_{j=8}^{n_1} f(1, j | p_0) & \text{if } m = 1, s \geq b_1 \\ 1 - \sum_{j=0}^{s-1} f(1, j | p_0) & \text{if } m = 1, s \leq a_1 \\ \sum_{j=b_1}^{n_1} f(1, j | p_0) + \sum_{j=8}^{b_1-1+n_2} f(2, j | p_0) & \text{if } m = 2 \end{cases} \quad (4)$$

Jung et al.^[25] show that, if the observed sample size at each stage is identical to that specified by the design, the decision by the two-stage design exactly matches with that by the p -value compared with the α^* value. As discussed in the previous section, the confidence interval calculated by Jennison and Turnbull^[16] has this property, too.

Example 5

For $H_0 : p_0 = 0.4$ vs. $H_1 : p_1 = 0.6$, the two-stage design $(n_1, n_2, a_1, b_1, a) = (34, 5, 17, 35, 20)$ is the Simon's minimax design among those with $(\alpha^*, 1 - \beta^*) = (0.05, 0.8)$ and a futility

Table 2 UMVUE, MLE, and p -values using these estimators for a two-stage design of $(n_1, n_2, a_1, b_1, a) = (34, 5, 17, 35, 20)$ with $p_0 = 0.4$

m	s	$f(m, s p)$	Estimate		p -value		
			UMVUE	MLE	UMVUE	MLE	Naive
1	0	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000
1	1	0.0000	0.0294	0.0294	1.0000	1.0000	1.0000
1	2	0.0000	0.0588	0.0588	1.0000	1.0000	1.0000
1	3	0.0001	0.0882	0.0882	1.0000	1.0000	1.0000
1	4	0.0003	0.1176	0.1176	0.9999	0.9999	0.9999
1	5	0.0010	0.1471	0.1471	0.9997	0.9997	0.9997
1	6	0.0034	0.1765	0.1765	0.9986	0.9986	0.9986
1	7	0.0090	0.2059	0.2059	0.9952	0.9952	0.9952
1	8	0.0203	0.2353	0.2353	0.9862	0.9862	0.9862
1	9	0.0391	0.2647	0.2647	0.9659	0.9659	0.9659
1	10	0.0652	0.2941	0.2941	0.9268	0.9268	0.9268
1	11	0.0948	0.3235	0.3235	0.8617	0.8617	0.8617
1	12	0.1211	0.3529	0.3529	0.7669	0.7669	0.7669
1	13	0.1366	0.3824	0.3824	0.6458	0.6458	0.6458
1	14	0.1366	0.4118	0.4118	0.5092	0.5092	0.5092
1	15	0.1214	0.4412	0.4412	0.3726	0.3726	0.3726
1	16	0.0961	0.4706	0.4706	0.2512	0.2478	0.2512
1	17	0.0679	0.5000	0.5000	0.1550	0.1388	0.1550
2	18	0.0033	0.5294	0.4615	0.0872	0.2512	0.2653
2	19	0.0129	0.5337	0.4872	0.0838	0.1517	0.1713
2	20	0.0219	0.5403	0.5128	0.0709	0.0709	0.1021
2	21	0.0217	0.5508	0.5385	0.0490	0.0490	0.0559
2	22	0.0145	0.5672	0.5641	0.0273	0.0273	0.0280
2	23	0.0075	0.5897	0.5897	0.0128	0.0128	0.0128
2	24	0.0033	0.6154	0.6154	0.0053	0.0053	0.0053
2	25	0.0013	0.6410	0.6410	0.0020	0.0020	0.0020
2	26	0.0005	0.6667	0.6667	0.0007	0.0007	0.0007
2	27	0.0002	0.6923	0.6923	0.0002	0.0002	0.0002
2	28	0.0000	0.7179	0.7179	0.0001	0.0001	0.0001
2	29	0.0000	0.7436	0.7436	0.0000	0.0000	0.0000
2	30	0.0000	0.7692	0.7692	0.0000	0.0000	0.0000
2	31	0.0000	0.7949	0.7949	0.0000	0.0000	0.0000
2	32	0.0000	0.8205	0.8205	0.0000	0.0000	0.0000
2	33	0.0000	0.8462	0.8462	0.0000	0.0000	0.0000
2	34	0.0000	0.8718	0.8718	0.0000	0.0000	0.0000
2	35	0.0000	0.8974	0.8974	0.0000	0.0000	0.0000
2	36	0.0000	0.9231	0.9231	0.0000	0.0000	0.0000
2	37	0.0000	0.9487	0.9487	0.0000	0.0000	0.0000
2	38	0.0000	0.9744	0.9744	0.0000	0.0000	0.0000
2	39	0.0000	1.0000	1.0000	0.0000	0.0000	0.0000

Traditional
(cont.)-Z-Score

stopping boundary only. Table 2 displays the p -values and the PMF at $p_0 = .4$ for each sample point in the descending order for UMVUE. For example, for $(m, s) = (2, 19)$, we calculate the UMVUE by

$$\begin{aligned}\tilde{p} &= \frac{\sum_{x_1=(17+1)\vee(19-5)}^{19\wedge(35-1)} \begin{pmatrix} 34-1 \\ x_1-1 \end{pmatrix} \begin{pmatrix} 5 \\ 19-x_1 \end{pmatrix}}{\sum_{x_1=(17+1)\vee(19-5)}^{19\wedge(35-1)} \begin{pmatrix} 34 \\ x_1 \end{pmatrix} \begin{pmatrix} 5 \\ 19-x_1 \end{pmatrix}} \\ &= \frac{\begin{pmatrix} 33 \\ 17 \end{pmatrix} \begin{pmatrix} 5 \\ 1 \end{pmatrix} + \begin{pmatrix} 33 \\ 18 \end{pmatrix} \begin{pmatrix} 5 \\ 0 \end{pmatrix}}{\begin{pmatrix} 34 \\ 18 \end{pmatrix} \begin{pmatrix} 5 \\ 1 \end{pmatrix} + \begin{pmatrix} 34 \\ 19 \end{pmatrix} \begin{pmatrix} 5 \\ 0 \end{pmatrix}} = 0.5337.\end{aligned}$$

Hence, when $(m, s) = (2, 19)$ is observed, we calculate the p -value based on UMVUE as

$$\begin{aligned}\Pr(\tilde{p}(M, S) \geq .5337 | p_0) &= \sum_{j=19}^{39} f(2, j | p_0) \\ &= .0129 + .0219 + .0217 + .0145 + .0075 + .0033 + .0013 \\ &\quad + .0005 + .0002 + .0000 + \dots + .0000 \\ &= 0.838.\end{aligned}$$

Since $p\text{-value} > \alpha^* (=0.05)$, we fail to reject H_0 . Since $s = 19$ after stage 2 ($=m$) is smaller than $a = 20$, we fail to reject H_0 by the two stage testing rule too. Table 2 reports the p -values based on the ordering by MLE values and the p -values calculated by ignoring the fact that the two-stage design aspect, denoted as “naive.” Note that the p -values based on UMVUE and MLE orderings are different for large s values with $m = 1$ and for small s values with $m = 2$. This occurs because of the difference between the UMVUE ordering and the MLE ordering for the small s values with $m = 2$. The naive p -values are quite different from the p -values based on the UMVUE ordering too.

The strength of the above p -value method is that it can be extended to the cases where the observed sample size at the stopping stage is different from that specified by the design. This becomes possible because the PMF $f(m, s | p)$ and the UMVUE depend on the stopping boundaries only up to stage $m - 1$, i.e., $\{(a_k, b_k), 1 \leq k \leq m - 1\}$, where $a_2 = b_2 - 1$ denotes a , the rejection value at the second stage. Suppose that a study stopped after stage $m = 1$ with $S = s$ responders from n_1 patients that may be different from the stage 1 sample size specified by the design. Then, Eq. 4 can be modified to calculate a p -value:

$$p\text{-value} = \begin{cases} \sum_{j=s}^{n_1} p^s (1-p)^{n_1-s} \begin{pmatrix} n_1 \\ s \end{pmatrix} & m=1 \text{ and the study} \\ & \text{is stopped for superiority} \\ 1 - \sum_{j=0}^{s-1} p^s (1-p)^{n_1-s} \begin{pmatrix} n_1 \\ s \end{pmatrix} & m=1 \text{ and the study} \\ & \text{is stopped for futility} \end{cases}.$$

Note that the p -value with $m = 1$ is calculated as if the study had a single-stage design with sample size n_1 , so that it does not require specification of stopping values (a_1, b_1) .

On the other hand, suppose that the study proceeded to the second stage, i.e., $m = 2$. In this case, the stage 1 decision will be made based on (a_1, b_1, n_1) as specified by the design, but the sample size for the second stage may be slightly different from n_2 that is specified by the design. In this case, based on the observed values of (n_2, s) , we calculate

$$\begin{aligned}p\text{-value} &= \sum_{j=b_1}^{n_1} p^s (1-p)^{n_1-s} \begin{pmatrix} n_1 \\ s \end{pmatrix} \\ &\quad + \sum_{j=s}^{b_1-1+n_2} p^s (1-p)^{n_1+n_2-s} \sum_{x_1=a_1+1}^{(b_1-1)\wedge s} \begin{pmatrix} n_1 \\ x_1 \end{pmatrix} \begin{pmatrix} n_2 \\ s-x_1 \end{pmatrix}\end{aligned}$$

when $m = 2$. This p -value does not require specification of a either. We can reject H_0 and accept the study therapy when $p\text{-value} < \alpha^*$ regardless of the observed sample size at the stopping stage.

It is easy to show that the ordering of MLE depends on the critical values at the stopping stage. Hence, we cannot calculate an MLE-based p -value if the sample size at the stopping stage is different from that specified by the design. Furthermore, the decision rule based on the MLE-based p -value may not match that based on the two-stage testing rule irrespective of whether the observed sample is identical to the planned one.

Example 4 (revisited)

In Example 4, with 27 CRs out of $n = 30$ patients after the second stage, the p -value is 0.0086 by the above method for $p_0 = 0.7$ and $(a_1, n_1) = (4, 6)$. Since $p\text{-value} < \alpha^*$, this is another justification that the study by Kim et al.^[22] is a positive study. For $\alpha^* = 0.05$, this study will be positive with 26 CRs ($p\text{-value} = 0.0259$), but not with 25 CRs ($p\text{-value} = 0.0605$).

As mentioned previously, the critical values of a two-stage design cannot be used for testing if the realized sample size at the stopping stage is different from that specified by the design. In this case, we can conduct a statistical testing by calculating the p -value with the sample size at the stopping stage conditioned on the observed value and checking if it is smaller than the prespecified α level. In the previous section, we show that the Jennison–Turnbull confidence interval can be used for this purpose too. This property makes the p -value, together with the Jennison–Turnbull confidence interval, very useful for the analysis of two-stage phase II trials. Other p -value methods for multistage design^[26,27] do not have these desirable properties.

Green and Dahlberg^[13] and Herndon^[14] have also considered under- or over-accrual of patients in each stage and proposed some ad hoc approaches to the selection of rejection values depending on the realized sample sizes. Chen and Ng^[28] considered a set of combinations for possible sample sizes for stages 1 and 2 (n_1, n_2) and provided the rejection values (a_1, a) to minimize the maximal sample size or the average expected sample, assuming that each combination in the set has the same probability to be observed. The latter approach seems to have some issues: (i) If the combination of realized sample sizes does not belong to the prespecified set, then it does not provide valid rejection values; and (ii) for each combination of (n_1, n_2) , rejection values (a_1, a) are given. For each n_1 value, there are multiple n_2 values with their own a_1 values. So, it is not clear which rejection value a_1

should be used for a realized n_1 after stage 1. Furthermore, suppose that we observed a combination (n_1, n_2) through two stages. Then, we will use a that is assigned to this combination, but a_1 that has been used after stage 1 may be different from the stage 1 rejection value assigned to this combination of sample sizes.

For $N_1 = n_1$ and $N_2 = n$, Chang and O'Brien^[29] use the likelihood ratio

$$\frac{f(m, s | \hat{p})}{f(m, s | p_0)} = \left(\frac{\hat{p}}{p_0} \right)^s \left(\frac{1 - \hat{p}}{1 - p_0} \right)^{N_m - s}$$

to measure how far the MLE $\hat{p} = \hat{p}(m, s)$ is from p_0 . Chang et al.^[30] and Cook^[31] use the likelihood ratio ordering to derive p -values from sequential tests. Note that the ordering of the sample space based on the likelihood ratio is two sided in nature, so that it is inappropriate for p -values for phase II trials which have one-sided hypotheses. It is obvious from Eq. 3 that the UMVUE gives p -values satisfying the properties: (i) the p -values in the acceptance region of H_0 are larger than those in the rejection region, and (ii) the p -value for the critical value matches with the type I error probability of the sequential testing. These properties are not satisfied by p -values derived from other orderings.

REFERENCES

- Eisenhauer, E.A.; Therasse, P.; Bogaerts, J.; et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur. J. Cancer* **2009**, *45*, 228247.
- Simon, R. Optimal two-stage designs for phase II clinical trials. *Controlled Clin. Trials* **1989**, *10*, 110.
- Ludwig, H.; Viterbo, L.; Greil, R.; et al. Randomized phase II study of bortezomib, thalidomide, and dexamethasone with or without cyclophosphamide as induction therapy in previously untreated multiple myeloma. *J. Clin. Oncol.* **2013**, *31*, 247–255.
- Friedberg, J.W.; Mahadevan, D.; Cebula, E.; et al. Phase II study of alisertib, a selective aurora A kinase inhibitor, in relapsed and refractory aggressive B- and T-cell non-Hodgkin lymphomas. *J. Clin. Oncol.* **2014**, *32*, 44–50.
- Koon, H.B.; Krown, S.E.; Lee, J.Y.; et al. Phase II trial of imatinib in AIDS-associated Kaposi sarcoma: AIDS malignancy Consortium Protocol 042. *J. Clin. Oncol.* **2014**, *32*, 402–408.
- Jung, S.H.; Carey, M.; Kim, K.M. Graphical search for two-stage phase II clinical trials. *Controlled Clin. Trials* **2001**, *22*, 367372.
- Jung, S.H.; Lee, T.Y.; Kim, K.M.; George, S.L. Admissible two-stage designs for phase II cancer clinical trials. *Stat. Med.* **2004**, *23*, 561–569.
- Chang, M.N.; Therneau, T.M.; Wieand, H.S.; et al. Designs for group sequential phase II clinical trials. *Biometrics* **1987**, *43*, 865874.
- Mander, A.P.; Wason, J.M.S.; Sweeting, M.J.; Thompson, S.G. Admissible two-stage designs for phase II cancer clinical trials that incorporate the expected sample size under the alternative hypothesis. *Pharm. Stat.* **2012**, *11*, 91–96.
- Englert, S.; Kieser, M. Optimal adaptive two-stage designs for phase II cancer clinical trials. *Biom. J.* **2013**, *55*, 955–968.
- Kieser, M.; Englert, S. Performance of adaptive designs for single-armed phase II oncology trials. *J. Biopharm. Stat.* **2015**, *25* (3), 602–615.
- Jung, S.H.; Kim, K.M. On the estimation of the binomial probability in multistage clinical trials. *Stat. Med.* **2004**, *23*, 881896.
- Green, S.J.; Dahlberg, S. Planned vs attained design in phase II clinical trials. *Stat. Med.* **1992**, *11*, 853–862.
- Herndon, J. A design alternative for two-stage, phase II, multicenter cancer clinical trials. *Controlled Clin. Trials* **1998**, *19*, 440–450.
- Bowden, J.; Wason, J. Identifying combined design and analysis procedures in two-stage trials with a binary end point. *Stat. Med.* **2012**, *31* (29), 3874–3884.
- Jennison, C.; Turnbull, B.W. Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. *Technometrics* **1983**, *25*, 4958.
- Armitage, P. Numerical studies in the sequential estimation of a binomial parameter. *Biometrika* **1958**, *45*, 115.
- Tsiatis, A.A.; Rosner, G.L.; Mehta, C.R. Exact confidence intervals following a group sequential test. *Biometrics* **1984**, *40*, 797803.
- Clopper, C.J.; Pearson, E.S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **1934**, *26*, 404413.
- Duffy, D.E.; Santner, T.J. Confidence intervals for a binomial parameter based on multistage tests. *Biometrics* **1987**, *43*, 8193.
- Tsai, W.Y.; Chi, Y.; Chen, C.M. Interval estimation of binomial proportion in clinical trials with a two-stage design. *Stat. Med.* **2008**, *27*, 1535.
- Kim, S.J.; Kim, K.; Kim, B.S. et al. Phase II trial of concurrent radiation and weekly cisplatin followed by VIPD chemotherapy in newly diagnosed, stage IE to IIE, nasal, extranodal NK/T-cell lymphoma: Consortium for Improving Survival of Lymphoma study. *J. Clin. Oncol.* **2009**, *27*, 60276032.
- Shimada, K.; Suzuki, R. Concurrent chemoradiotherapy for limited-stage extranodal natural killer/T-cell lymphoma, nasal type. *J. Clin. Oncol.* **2010**, *28* (14), e229.
- Kim, S.J.; Jung, S.H.; Kim, W.S. Reply to K Shimada et al. *J. Clin. Oncol.* **2010**, *28* (14), e230.
- Jung, S.H.; Owzar, K.; George, S.L. et al. P-value calculation for multistage phase II cancer clinical trials (with discussion). *J. Biopharm. Stat.* **2006**, *16*, 765783.
- Cook, T.D. P-value adjustment in sequential clinical trials. *Biometrics* **2002**, *58*, 10051011.
- Koyama, T.; Chen, H. Proper inferences from Simon's two-stage designs. *Stat. Med.* **2008**, *27*, 31453154.
- Chen, T.T.; Ng, T.H. Optimal flexible designs in phase II clinical trials. *Stat. Med.* **1998**, *17*, 2301–2312.
- Chang, M.N.; O'Brien, P.C. Confidence intervals following group sequential trials. *Controlled Clin. Trials* **1985**, *7*, 18–26.
- Chang, M.N.; Gould, A.L.; Snapinn, S.M. P-values for group sequential testing. *Biometrika* **1995**, *82*, 650–654.
- Cook, T.D. P-value adjustment in sequential clinical trials. *Biometrics* **2002**, *58*, 1005–1011.

US Food and Drug Administration

Stacy R. Lindborg

Eli Lilly and Company, Indianapolis, Indiana, U.S.A.

Susan S. Ellenberg

U.S. Food and Drug Administration, Rockville, Maryland, U.S.A.

Abstract

The Food and Drug Administration (FDA) is a U.S. government agency under the Department of Health and Human Services that is responsible for ensuring that the food we eat is safe, the cosmetics we use are pure and safe, and medicines and medical products are safe and effective for uses indicated in the product labels. This entry is dedicated to giving an overview of the FDA's history and role in American life.

INTRODUCTION

The Food and Drug Administration (FDA) is a U.S. government agency under the Department of Health and Human Services (DHHS) that plays a complex and important role in the lives of all Americans, as well as in the lives of consumers of many products grown, processed, and/or manufactured in the United States. As mandated by federal law, the FDA helps ensure that the food we eat is safe and wholesome, the cosmetics we use are pure and safe, and medicines and medical products are safe and effective for uses indicated in the product labels. The FDA is also responsible for overseeing veterinary medical and food products.

OVERVIEW

The FDA mission, formalized in the FDA Modernization Act of 1997, is as follows:

- To promote public health by promptly and efficiently reviewing clinical research and by taking appropriate action on the marketing of regulated products in a timely manner.
- With respect to such products, to protect public health by ensuring that foods are safe, wholesome, sanitary, and properly labeled; human and veterinary drugs are safe and effective; there is reasonable assurance of the safety and the effectiveness of devices intended for human use; cosmetics are safe and properly labeled; and public health and safety are protected from electronic product radiation.
- To participate through appropriate processes with representatives of other countries to reduce the burden of regulation, to harmonize regulatory requirements, and to achieve appropriate reciprocal arrangements.

- As determined to be appropriate by the Secretary [of DHHS], to carry out items (1) through (3) in consultation with experts in science, medicine, and public health, and in cooperation with consumers, users, manufacturers, importers, packers, distributors, and retailers of regulated products.

To accomplish these goals, the FDA is divided into six centers, along with a network of regional field offices (see [Fig. 1](#)). Three centers focus on human medical products: the Center for Biologics Evaluation and Research (CBER), the Center for Devices and Radiological Health (CDRH), and the Center for Drug Evaluation and Research (CDER). The Center for Food Safety and Applied Nutrition (CFSAN) is responsible for overseeing food and food products; the Center for Veterinary Medicine (CVM) enables the marketing of effective and safe veterinary medical products; and the National Center for Toxicological Research (NCTR) is a basic research facility that primarily studies problems related to the biological mechanisms underlying the toxicity of products faced by all of the centers with regulatory responsibilities.

The goal of this entry is to provide an overview of the FDA while discussing the role of statistics in drug regulation. While this discussion will be far from comprehensive, we will attempt to provide practical commentary regarding the interaction between the FDA and its customers. We will begin by discussing the origin of the FDA and the important historical events that shaped its current structure and purpose.

HISTORICAL DEVELOPMENT

The organization we know today has evolved via a long and interesting series of events, beginning in 1785 when

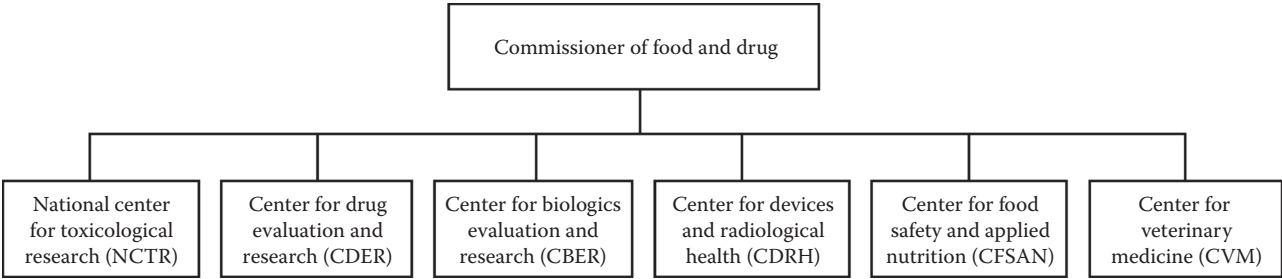


Fig. 1 Organizational structure of the FDA

the state of Massachusetts enacted the first food adulteration law in the United States. The FDA itself had its origins in the establishment of the Bureau of Chemistry in the Department of Agriculture in 1862 by President Abraham Lincoln. A national food and drug law was first recommended in 1880, but one was not enacted until 1906, following a flurry of congressional activity between 1879 and 1906 resulting in over 100 bills relating to the safety of food and drug products.

Three landmarks of food and drug law are the Biologics Control Act of 1902, and the Food and Drugs Act and the Meat Inspection Act, both of 1906. The first established the basis for the regulation of sera, vaccines, and related products, and represents the origin of biologics regulation in the United States. The second prohibited interstate commerce in misbranded and adulterated food and drugs, and provided the basis for regulation in these areas. The third dealt with sanitation in meat-packing plants. These laws were all motivated by tragic events and/or disclosures of clearly dangerous conditions and procedures in the preparation of food and drugs marketed to the U.S. population.

In 1938, the Federal Food, Drug, and Cosmetic (FDC) Act was passed. This act substantially extended the responsibilities of the executive branch with regard to the regulation of these products. Cosmetics and therapeutic devices became subject to safety requirements; drugs were required to be shown safe prior to marketing; inspections of manufacturing facilities were authorized; and a variety of standards were established for foods. Shortly after this legislation was enacted, the regulatory authority for food and drugs was moved from the Department of Agriculture to the Federal Security Agency, and the first Commissioner of Food and Drugs, Walter G. Campbell, was appointed in 1940. (In 1953, the Federal Security Agency became the Department of Health, Education, and Welfare, the predecessor of the current DHHS.)

In 1962, the FDC Act was amended to require for the first time that drugs be demonstrated effective as well as safe prior to being approved for marketing. Provisions strengthening drug safety requirements were also included. These legislative actions were motivated largely by the discovery that thalidomide, a sleeping pill marketed widely outside the United States, was responsible for an epidemic of birth defects.

More recent milestones in food and drug regulation include the transfer of the Bureau of Radiological Health to the FDA in 1971; the transfer of responsibility for the regulation of biologics from the NIH to the FDA in 1972 and the initiation of over-the-counter drug review that same year; the passage of medical device amendments in 1976, formalizing and strengthening regulation in this growing area; the Orphan Drug Act of 1983, promoting the development of drugs needed to treat rare diseases; and the Prescription Drug User Fee Act (PDUFA) of 1992, requiring manufacturers of drugs and biologics to pay fees related to the submission of marketing applications. As a result of these acts, the FDA is authorized by Congress to monitor activity surrounding U.S.\$1 trillion worth of goods annually, approximately one-quarter of the gross national product. Each of the six centers of the FDA plays an important role in helping to accomplish this enormous task.

FOOD AND DRUG ADMINISTRATION CENTERS AND THE PRODUCT DEVELOPMENT PROCESS

Four of the FDA’s six centers—CBER, CDER, CDRH, and CVM—focus on biopharmaceutical/biomedical products. We will focus most of our discussions on these four centers. The primary tasks of these centers relate to the oversight of product development and the evaluation of licensing applications for new products (or new uses of old products). Regulatory requirements for drugs and biologics can be found in the Code of Federal Regulations, Title 21 (Food and Drugs), Chapter 1 (The Food and Drug Administration), Subchapter D (Parts 300–499); requirements specific to veterinary products in Subchapter E (Parts 500–599); requirements specific to biologics in Subchapter F (Parts 600–699); and requirements for medical devices and radiological health in Subchapters H–J (Parts 800–1050).

For drugs and biologics intended for human use, the regulatory cycle begins when the product developer, or sponsor, intends to begin the first studies in humans. At that time, an investigational new drug (IND) application is submitted to the agency. Sufficient preclinical data must be provided in this application to support the safety and the rationale for the use of the product. For medical devices,

such applications are required only for specific classes of devices that pose “substantial risk.” In these cases, an investigational device exemption (IDE) application will be submitted. Reviewers for the FDA (representing a variety of disciplines depending on the product) will consider the application and will determine whether clinical investigations may proceed.

Plans for each clinical study to be carried out must be submitted to the agency prior to initiation if the studies are to be carried out in the United States. The review of these studies by the FDA serves several functions. First, this review provides an additional check on safety assurance that the study is unlikely to result in harm to participants. Second, the FDA reviewers will assess whether the study design is likely to achieve its goals, and they may make suggestions to the sponsor regarding the approach taken. An interaction between the sponsor and the FDA at the early stages of drug development can be extremely helpful in ensuring that the development strategy laid out by the sponsor will ultimately provide the persuasive data regarding safety and efficacy that will be necessary to gain marketing approval.

The culmination of the sponsor’s evaluation of the product is the submission of a dossier to the appropriate agency: for pharmaceuticals, a new drug application (NDA) to the CDER; for biologics, a biologics licensing application (BLA) to the CBER; and for medical devices, a premarket assessment (PMA) to the CDRH. This comprehensive package must summarize the efficacy and the safety of the investigational drug, biologic, or device as demonstrated by clinical and other studies. It must also contain adequate evidence that the product will have the effect that the sponsor claims under the conditions of use (505(d) FDA Act). The necessary components of a submission package have been clearly defined by the appropriate center of the FDA and can be found easily on the Internet (<http://www.fda.gov>). In the case of an NDA, the dossier must contain an application summary, copies of study protocol(s), study reports, integrated summaries of efficacy and safety, and proposed product labeling (package insert).

In addition to reviewing data from studies evaluating product, the FDA engages in a wide variety of oversight activities. Product manufacturing processes are held to established standards, and manufacturing facilities are inspected by the FDA staff at regular intervals to ensure compliance with the regulations. The FDA inspectors also visit clinical sites where studies are carried out to document that correct and appropriate data were recorded, that all patients treated with investigational product were accounted for in study reports, and that study procedures were followed according to the study protocol. The records of Institutional Review Boards (IRBs) are also subject to scrutiny by the FDA inspectors.

The FDA’s responsibility for ensuring product safety does not end with the decision that a product may be marketed. Manufacturers are required to report adverse

experiences associated with their products to the FDA within timelines dependent on the type of experience. These reports, together with others received directly from health care providers and consumers, are reviewed by the FDA to determine if changes are needed in the way the product is labeled, marketed, or prescribed.

STATISTICS AT THE FOOD AND DRUG ADMINISTRATION

Each of the FDA’s six centers has its own statistical unit. In 1998, the number of statisticians employed by the FDA was approximately 150. FDA statisticians participate in all aspects of regulatory review, assessing laboratory and pre-clinical data as well as data from pharmacokinetic studies, early clinical assessments, and large-scale clinical trials. They are also involved in setting and evaluating standards for testing and quality control; assessing data from post-marketing safety surveillance programs; collaborating with other FDA scientists on research projects; evaluating newly proposed statistical methods potentially applicable to regulatory settings; conducting methodological research; providing training to other scientific staff; and contributing to policy development.

The CDER has the largest cadre of statisticians, divided among three divisions, and a staff devoted to special projects. Each of the three divisions includes about 20–30 statisticians and works with a particular subset of medical review divisions. For example, the Division of Biometrics I evaluates drugs for cancer, heart, kidney, and neurological/mental disorders. All other centers have a single group of statisticians, ranging in approximate size from 5 to 20. Most FDA statisticians have doctoral degrees in statistics or biostatistics.

Probably the largest fraction of statistical effort at the FDA goes into the review of licensing applications for new products. The data from studies that the sponsor believes support the efficacy and the safety of the product are generally provided in electronic format, facilitating the FDA statistical analysis. The FDA statistical reviewers will assess the analyses presented by the sponsor and may perform supplementary analyses to assist in interpreting the data presented. Some FDA statisticians attend the advisory committee meetings wherein the application is discussed and may make a presentation in cases where statistical issues need to be highlighted. The statistical assessment of the data in the application is key to the decision to approve, or not to approve, a product for marketing.

A major ongoing challenge for FDA statisticians is the evaluation of the applicability of new statistical approaches to the regulatory setting. For example, there has been increased interest in applying Bayesian methodology to clinical trials of investigational products.^[1] Statisticians at the CDRH are actively evaluating the incorporation of Bayesian methodology in the clinical development process.

Because so many new devices differ in only minimal ways from earlier versions, the formal inclusion of information on the effectiveness and the safety of these earlier versions in the evaluation of new devices is appealing. As reported by the February 1998 International Medical Device Regulatory Monitor, "The FDA is preparing to issue a guidance on an advanced statistical methodology, known as the Bayesian model, that incorporates historical information into medical device clinical trial data."^[2] For a discussion of the use of Bayesian methodology in medical device development, see Berry.^[3]

FOOD AND DRUG ADMINISTRATION RELATIONSHIP TO SPONSORS

The relationship formed between the FDA and a sponsor is extremely important. Each drug, device, or biologic that a sponsor decides to research on and submit for regulatory approval represents an enormous and risky investment. Only one in approximately 10,000 compounds tested ever makes it to the market. Additionally, of those that do eventually make it to market, the cost of testing and satisfying the regulators worldwide has been estimated as U.S.\$528–792 million by the Association of British Pharmaceutical Industry (ABPI). Due to the costly nature of the registration process, the importance of good communication between the sponsor and the FDA cannot be overemphasized.

Each center of the FDA has an office dedicated to communication with sponsors and other interested parties. These offices work to ensure that all of the FDA's customers (including representatives of the pharmaceutical industry, health care professionals, government officials in other agencies, the media, and consumers) have easy and open access to information and are fully informed about the drug regulation process.

The FDA encourages communication with the sponsor throughout the review process and specifically advocates meetings prior to the development of an IND application and immediately preceding the submission of a marketing application. In planning a meeting with the FDA, the sponsor should take great care in designing the agenda. As noted by Waymack and Rutan,^[4] sponsors should carefully consider the attendees and the content of this meeting to ensure that the relevant issues can be fully and efficiently addressed.

One important component of the review process is the presentation to one of the FDA's advisory committees by the sponsor and the FDA review team of their respective assessments of the data contained in a marketing application. These committees consist of a panel of experts—external to the FDA and independent of the sponsor—that generally includes physicians, statisticians, other scientists and health professionals, and a lay consumer representative. The committees are governed by the Federal Advisory Committee Act of 1972 and later amendments.

The primary purposes of FDA advisory committees are to review available data (efficacy and safety) for medical products under investigation for human use and to address issues as requested by the FDA to assist in labeling decisions. Advisory committees may also be asked to comment on methodological issues germane to many applications. The FDA is not required to follow the recommendations of advisory committees but does act in accordance with their recommendations in most instances. While the size of the committee can range between 10 and 15 members, the panel can be supplemented with additional experts on an ad hoc basis when necessary for particular applications. A thorough discussion of the FDA's use of advisory committees can be found in an Institute of Medicine report that was issued in 1992^[5] and that formed the basis for many recent changes in operating procedures. For a practical article related to FDA advisory committees, see, for example, Farley.^[6]

FOOD AND DRUG ADMINISTRATION GUIDANCE DOCUMENTS

A list of guidelines that have been prepared by the FDA to aid sponsors in their pursuit of regulatory product approval is maintained by each FDA center. It is important to note that guidance documents differ from regulations, in that sponsors are not required to abide by practices described in guidance documents; rather, these documents represent what the FDA considers acceptable approaches to regulated research, although there may be other approaches that are also acceptable. While guidance is provided by the FDA on many aspects of design and analysis, the responsibility for adequate studies and credible analyses ultimately lies with the sponsor. A list of all guidelines that apply to the FDA's regulations across all centers, as well as the corresponding documents, can be found at the FDA's website (<http://www.fda.gov/opacom/morechoices/industry/guidedc.htm>).

An important ongoing initiative of the FDA is participation in the International Conference on Harmonization (ICH). The ICH comprises a joint effort of regulatory authorities and the pharmaceutical industry in the United States, Europe, and Japan to standardize regulatory requirements and to establish standards for pharmaceutical research to the extent possible among these three regions, thereby reducing inefficiencies and ensuring a consistently high quality of research for manufacturers in an increasingly global industry. The ICH has developed documents in the areas of quality, safety, and efficacy, and for some crosscutting topics as well. A description of the ICH, a listing and brief description of all documents issues, and access to the full documents themselves can be found at the ICH website (<http://www.ifpma.org.ich1.html>).

An ICH guideline entitled "Statistical Principles for Clinical Trials" has recently been issued by the FDA as

a guidance document (<http://www.fda.gov/ohrms/dockets/98fr/091698c.txt>). This document discusses approaches to trial design and analysis in terms of the prevention of bias and the preservation of the desired error probabilities. This document should help ensure that adequate attention is given to statistical considerations for studies of new pharmaceutical products performed in the three ICH regions, and should facilitate the acceptability of registration data collected for marketing authorization in one region by regulatory authorities in the others.

COMMON PROBLEMS AND CURRENT CONTROVERSIES

As methods for clinical research become more sophisticated, the approaches to evaluating investigational medical products also evolve. In many cases, FDA regulations and guidelines may be updated to account for new information. For example, guidance documents were updated in the late 1980s to account for methods of interim analysis that were becoming widely used in trials with mortality or major morbidity endpoints. In addition, patient groups in recent years have raised concerns about FDA regulations being so focused on the protection of study subjects that many individuals were being denied access to clinical trials. These policies had the effect (perhaps inappropriately) of restricting access to investigational therapy of some patient groups, as well as reducing the generalizability of the data generated. As a result, guidance documents have been made available to stress the importance that all relevant subgroup populations should be included in clinical trials.

Despite the ready availability of FDA regulations and guidelines, as described earlier, and the continual reassessment of optimal research approaches, certain problems regularly arise. These issues include the size of studies, the appropriate control group, the appropriate population(s) to study, and the primary and secondary study endpoints that will define the product efficacy. It is important to demonstrate efficacy and safety in a way that not only meets the requirements of regulatory agencies, but also documents and motivates use from both a clinical and a commercial perspective.

Study Size

Study size is determined by the desired error probabilities and the estimated effects of the investigational and control treatments. Companies wish to perform the smallest possible studies that will still be adequate to demonstrate efficacy and safety. In some cases, FDA reviewers believe that the company is overly optimistic about the expected effects of the new treatment and suggest that a study be enlarged to be able to detect more modest, but still clinically important, effects. In some instances, the study size proposed may be adequate to assess the primary endpoint

but not important secondary endpoints or specific safety concerns.

Control Group

There are a number of aspects to the selection of a control group: chronology (historical vs. concurrent controls); treatment assignment method (randomized vs. systematic); type of control treatment (active vs. inactive); and approach to blinding (double-blind, single-blind, open label). The randomized, double-blind, placebo-controlled trial is generally considered the optimal approach but is not feasible in all instances. For example, active controls rather than placebo controls would be used when available active treatment is known to prevent or delay death or major morbidity. Random allocation of patients between treatment and control is considered essential, in most instances, to a reliable interpretation of observed treatment effects. However, historical controls have been accepted as a basis for approval in some instances when the target patient population is extremely small and/or the effect of the new product is so dramatic as to make selection bias an unlikely explanation for the outcome.

Study Population

The definition of the study population is extremely important. Clearly, one wishes to study a new treatment in a population in which it is most likely to prove effective. Thus individuals whose disease is so minimal that improvements would be difficult to document and individuals whose disease is so advanced that no treatment is likely to be effective are generally not considered for studies of new products. On the other hand, it is important to study the treatment in the population for which it is likely to be prescribed, should the study be positive and should the product be marketed. Attention is given at the design stage to ensure the inclusion of the elderly, women, and children in studies of treatments likely to be used in these populations.

Primary and Secondary Endpoints

The *primary* endpoint for the trial must provide an adequate basis for assessing the clinical value of the treatment. Most of the discussions between the FDA and the sponsors center around the choice of clinical trial endpoints. Endpoints may be measured as continuous, binary, or multinomial variables, as time to an event of interest, as counts of events over a specified time interval, and as the slope of a regression line, just to give a few examples. In some cases, a *composite* endpoint—counting any of two or more outcomes as a primary endpoint—may be used. Such endpoints may be particularly relevant when treating conditions that have multiple manifestations, any of which might be improved by an effective treatment. The choice of endpoint will depend on the particular circumstances of

the trial and must be clearly specified in the study protocol prior to initiating the study.

Secondary endpoints are those that are important indicators of disease status and thus may provide a supportive evidence of treatment efficacy or safety. The number of secondary endpoints specified should be limited to those that are truly important to patients, and to physicians managing patients with the disease in question and likely to reflect the effects of the treatment being evaluated. Secondary endpoints generally cannot provide the basis for marketing approval in the absence of effects on the primary endpoint (see Fisher^[7] and Moye^[8] for recent extensive discussion of this issue) but can provide useful supportive data as well as further understanding of the product's mechanisms of activity.

In trials of therapies that are potentially lifesaving or have other potentially major benefits, "surrogate" endpoints—laboratory or early clinical measures that are believed to be predictive of the clinical outcome of interest—have been accepted as the basis for product approvals, with the condition that adequate studies documenting effects on the primary clinical outcome are ongoing. Because rapid approval based on surrogate endpoints implies shorter-term studies, there is a risk that longer-term safety issues might not be identified in such studies and that surrogates might not accurately predict the true risk-benefit ratio of a product.^[9,10] For this reason, the action based on surrogate endpoints is limited to settings of serious disease where agents with improved efficacy are urgently needed.

Many other topics that are debated on frequently in regulatory meetings include washout periods in crossover designs, the need for and/or adequacy of blinding, the treatment of dropouts and noncompliers in study analyses, the need for statistical adjustment for multiple comparisons, procedures for carrying out interim analyses, a definition of an equivalence margin in active control trials, and the necessary extent of follow-up for safety and efficacy assessments. Some of these issues are discussed by Lynch and Lachenbruch^[11] in the context of trials for biologics, but most of the discussion is relevant to clinical trials more generally. An excellent source of insight on current clinical trial controversies can be seen in Senn.^[12]

CHALLENGES FOR THE FUTURE

In a time of rapid pharmaceutical development and the opening of innovative research made possible by the explosion in biotechnology, the FDA is facing a number of difficult but exciting challenges. Perhaps the most obvious of these is the need to review an increasing number of new products rapidly so that new products that meet standards of safety and effectiveness can be made available to the public as quickly as possible. The PDUFA of 1992, which has been extended through September 2002, commits the FDA to faster review times for most new drugs and biologics applications. It was anticipated that this act would

enable the new product application review to be cut from 15.5 to 10 months. As shown in a progress report to Congress in 1998, the FDA has been successful in hiring and training new reviewers and is ahead of schedule in meeting the escalating PDUFA goals.

Other challenges relate to the application of new biological knowledge to the development of new products. Gene therapy is in its infancy, as are therapies involving the manipulation of a patient's own cells and then replacing them. Both of these approaches are showing the potential for improving clinical status in disease settings with currently little or no available satisfactory treatment. Gene therapy is one of a number of promising approaches for which theoretical safety concerns will need to be carefully studied. Other products, such as thalidomide (now approved for treatment of leprosy and AIDS) and xenotransplants (transplants of animal tissues to humans; not yet approved but under study), raise even stronger safety concerns. The task of balancing the risks of dramatically new treatment approaches carrying major risks with the potential benefits of such treatments for extremely ill patients will be confronted more frequently in the coming years.

ACKNOWLEDGEMENTS

Minor updates have been made by the Editor-in-Chief in order to reflect developments since publication of the *Encyclopedia of Biopharmaceutical Statistics, Third Edition*.

REFERENCES

1. Spiegelhalter, D.J.; Freedman, L.S.; Parmar, M.K.B. *JRSS A* **1994**, *157*, 357–416.
2. FDA Officials Endorse Statistical Model That Uses Historical Data in Clinical Trials. In *International Medical Device Regulatory Monitor*; Newsletter Services Inc: Lanham, MD, 1998, Vol. 6 (2), 1–2.
3. Berry, D. *Using a Bayesian Approach in Medical Device Development*; Duke University: Durham, NC, 1997, <http://www.stat.duke.edu/people/personal/db.html>.
4. Waymack, J.P.; Rutan, R. *Appl. Clin. Trials* **1996**, 28–34.
5. *Institute of Medicine, Food and Drug Administration Advisory Committees*; National Academy Press: Washington, DC, 1992.
6. Farley, D. *FDA Consumer*, 2nd Ed.; 1995, 31–35.
7. Fisher, L.D. *Contr. Clin. Trials* **1999**, *20*, 16–39.
8. Moye, L.A. *Contr. Clin. Trials* **1999**, *20*, 40–49.
9. Temple, R.; Nimmo, W.S.; Tucker, G.T. A Regulatory Authority's Opinion About Surrogate Endpoints. In *Clinical Measurement in Drug Evaluation*; Nimmo, W.S., Tucker, G.T., Eds.; Wiley: New York, 1995, 3–22.
10. Fleming, T.R.; DeMets, D.L. *Ann. Intern. Med.* **1996**, *125*, 605–613.
11. Lynch, C.J.; Lachenbruch, P.A. *Drug Inf. J.* **1996**, *30*, 921–932.

USP Tests

John R. Murphy

Eli Lilly and Company, Indianapolis, Indiana, U.S.A.

Abstract

This brief entry provides an overview of United States Pharmacopeia tests which are methods and procedures by which the acceptability of drug substances and drug products are determined and adjudicated.

INTRODUCTION

United States Pharmacopeia (USP) tests are methods and procedures by which the acceptability of drug substances and drug products are determined and adjudicated. In the United States, the standards published in the United States Pharmacopeia and the National Formulary (USP/NF) are official and binding by law under the Food, Drug, and Cosmetic Act, and they apply at any time in the life of a drug, from production to consumption.^[1] Some tests for drug substances or drug products provided in the USP, such as assays for identity, strength, and purity, are specific to the particular drug, and these tests and procedures are described in detail for each drug in a section of the USP called “Official Monographs.” Other tests, such as uniformity of dosage units, dissolution, disintegration, deliverable volume, and particulate matter, are common to many drugs and are provided in what are called “General Chapters.”

HISTORY

The USP is a publication of the USP Convention, which is an organization of both individuals and institutions having interest and expertise in drug standards. The first meeting of the USP Convention was in 1820; the first USP was published later that year, consisting of 272 pages and listing 217 drugs worthy of recognition, for guiding pharmacists and physicians in making simple mixtures and preparations. In its present-day form, the USP contains standards of identity, quality, strength, and purity for over 3000 drugs and substances in “Official Monographs.” In addition to the compilations in the “Official Monographs,” the sections called “General Tests and Assays” and “General Information” cover such subjects as analytical methods, biological assays, how to clean glassware, the laws and regulations governing drug manufacturing and distribution, stability issues, validation of compendial methods, sterility, etc. Similar compendia, such as the *British Pharmacopeia*, the *European Pharmacopeia*, and the *Japanese*

Pharmacopeia, apply to drug products marketed in those areas of the world.

A great deal of confusion exists concerning the application of USP tests. For example, it is widely believed that manufacturers must perform the USP tests to release products to the market and that they must perform them exactly as given in the USP. Neither statement is true, as the USP clearly states in its “Preamble.” On the other hand, as the USP tests and procedures are the standards by which drug products will be evaluated after they are released to the market, they are sometimes used as lot release criteria. This practice carries a degree of risk because there is little assurance that the product will pass future USP tests, given only that it has passed the USP test once at release. As the cost associated with the failure of the product in the marketplace to meet compendial requirements is great, manufacturers often adopt “in-house” product release procedures that differ from the actual USP tests but that still provides assurance that the product will meet USP standards throughout its shelf life.

Many USP tests have the look and feel of statistical sampling plans, but the USP discourages this interpretation. In its “Preface,” the USP explains that the standards apply to specimens of compendial articles and that although some tests apply to specimens consisting of multiple-dosage units, they still comprise one determination of the particular attributes of the specimen. Thus, for example, although the USP Uniformity of Dosage Forms calls for up to 30 individual dosage units, the result of the test is the determination of uniformity or the lack thereof of one single specimen. The confusion of USP tests with sampling and acceptance plans also contributes to a misunderstanding about what it means when a USP test is not met. The USP position is that inferences about the quality of the batch or the process is not necessarily warranted from a single test on a single specimen. Nevertheless, a prudent manufacturer will want to be satisfied that there is a high probability that any specimen in commercial distribution has a high probability of meeting USP requirements and will gain assurance that this is the case through internal sampling

and acceptance procedures. Although the USP tests should not be interpreted as sampling and acceptance plans, it is still possible to evaluate their hypothetical performance by means of operating characteristic (OC) curves, which display the probability that a particular test will be passed as a function of one or more measures of conformance.

CONCLUSION

In summary, USP tests are the official and statutory means by which identity, quality, strength, and purity are judged for drugs marketed in the United States. The various tests and procedures for performing them are given in the USP/NF in “Official Monographs” and in “General Chapters.” Any specimen of any article in commercial distribution

must comply with USP standards at any time throughout shelf life, and manufacturers are expected to perform the necessary sampling and testing to have a high level of assurance that pharmaceutical products conform to USP standards.

ARTICLES OF FURTHER INTEREST

Content Uniformity; Release Targets; Specifications.

REFERENCE

1. USP/NF. *United States Pharmacopeia XXIV and the National Formulary XIX*; The United States Pharmacopoeial Convention, Inc: Rockville, MD, 2000.

Validation of Quantitative and Qualitative Assays

Bob Zhong

Janssen Pharmaceutical Companies, Johnson and Johnson, Malvern, Pennsylvania, U.S.A.

Chunfu Qiu

Sanofi-Aventis Pharmaceuticals, Bridgewater, New Jersey, U.S.A.

Dejun Tang

Novartis Pharmaceuticals, East Hanover, New Jersey, U.S.A.

Abstract

This entry introduces typical studies conducted in the validation process of quantitative and qualitative assays and explains for each study its purpose, statistical model, parameters, and the hypothesis being tested.

Keywords: Assay validation; Quantitative; Qualitative assays.

INTRODUCTION

After an assay has been developed in a laboratory and before it can be used in clinical application (such as the diagnosis of a patient's disease), a rigid clinical validation is carried out to demonstrate that the procedure is fit for use. The clinical validation process can be divided into two categories: one for quantitative assay; the other for qualitative one. (Whether an assay is quantitative or qualitative is decided during the developing stage of an assay.) Within each category, a full, partial, or cross validation might be needed.^[1]

This entry will introduce typical studies conducted in the validation process of quantitative and qualitative assays. The key difference and similarity between quantitative and qualitative assays are discussed in the section "The Key Difference and Similarity Between Quantitative and Qualitative Assays." Typical studies conducted to validate quantitative and qualitative assays are addressed in sections "Evaluation of Quantitative Assays" and "Evaluation of Qualitative Assays." Some concluding remarks are given in the last section.

This entry is written based on clinical and laboratory standards established by the Clinical and Laboratory Standards Institute (CLSI), which was renamed in 2005 (it was originally called the National Committee on Clinical Laboratory Standards (NCCLS). Occasionally, the other standards-setting body is also mentioned in this entry: the U.S. Pharmacopeia (USP). USP is an official public standards-setting authority for all prescription and over-the-counter medicines and other health-care products manufactured or sold in the United States. USP also sets widely recognized standards for food ingredients and dietary supplements and for the quality, purity, strength, and consistency of these products.

THE KEY DIFFERENCE AND SIMILARITY BETWEEN QUANTITATIVE AND QUALITATIVE ASSAYS

The key difference and similarity between quantitative and qualitative assays are discussed below.

An obvious key difference between quantitative and qualitative assays can be seen from "quantitative" and "qualitative." A quantitative assay usually measures the concentration of an analyte. The reported concentration level is distributed within a continuous range of values. A qualitative assay, on the other hand, produces dichotomous results only, such as "positive/negative," "abnormal/normal" statements. In practice, both quantitative and qualitative assays have broad applications. For example, Pharmacokinetics analysis, based on values measured by quantitative assays, gives indications of how much drug is in the blood stream and how long it lasts; hence, reasonable treatment regimens and frequency can be determined. An example for the use of qualitative assays is to diagnose if a patient is HIV-positive or -negative.

A key similarity between quantitative and qualitative assays is that usually a monotone increasing relationship exists between the response and concentration level, where the response of a qualitative assay is its raw response (before it is categorized into "positive/negative"). The monotone relationship of a quantitative assay between the response and concentration level can be seen through linearity study (See section "Accuracy"). The monotone relationship of a qualitative assay between the raw response and concentration level can be seen from the precision study of a qualitative assay. The monotone relationship of a qualitative assay had not been incorporated into clinical studies, as pointed out in.^[2]

The key difference constitutes the different validation studies of quantitative and qualitative assays; and the key similarity constitutes the similar study, as discussed in the following sections.

EVALUATION OF QUANTITATIVE ASSAYS

Introduction

The objective of validation of a quantitative assay procedure is to demonstrate that the assay is suitable for its intended purpose.^[3] Typical performance characteristics for consideration during the validation of the analytical procedures are accuracy, precision, selectivity (specificity), limit of detection (LOD), limit of quantitation (LOQ), linearity, and range. Which characteristic needs to be validated for a particular assay is described in.³ The validation of accuracy, precision, selectivity (section “Matrix Effects”), linearity and range, and LOD and LOQ are included in the following subsections.

Accuracy

The accuracy (trueness) of an analytical procedure expresses the closeness of agreement between the “true” value and the value found, where the “true” value is accepted either as a conventional true value or an accepted reference value.^[3] Using the concentration of an analyte in a sample as an example, the true concentration is the targeted true value. This true concentration can be “known” or unknown as explained below.

A sample prepared by adding a certain amount of the analyte to a tube of excipient is assumed to have a “known” concentration. The “known” concentration of the sample is calculated using appropriate (theoretical) methods, though the calculated concentration might be different from the real concentration due to possible interaction between the added analyte and excipient. In this case, a validation called recovery study is usually conducted.

The other case is that when patient samples are used to validate the accuracy of an assay. In this case, the true concentration is unknown. The validation of the accuracy of an assay is usually conducted by comparing the proposed assay’s results to the results of an established assay. This validation is called “method comparison.”

It should be pointed out that the recovery study and method comparison cannot be viewed as a replacement for each other. The recovery study is conducted in the development stage of an assay and confirmed at validation stage. On the other hand, even an assay is validated through a recovery study; it might not be ready for clinical applications, since clinical normal range, etc., might not have been established yet.

Hence, the validation of accuracy is divided into two parts according to the true value is known (a conventional

true value exists) or unknown (a conventional true value does not exist but an accepted reference value exists). The validation for these two cases will be discussed below, respectively.

Recovery Study

As discussed above, the accuracy of an assay method is usually assessed using samples prepared by adding known amounts of ingredient to excipient in a recovery study. In this case, the concentrations of the samples are assumed fixed. However, the reported concentrations by the proposed assay are random variables. The validation can be based on the percent recovery, the absolute bias, and the percent bias.^[4] An alternative approach for the assessment of accuracy is linear regression analysis, as discussed further below.

Let y be the result by the proposed assay when measuring a sample at concentration x and y_{ij} be the result by the proposed assay at x_i at the j th repeated measurement, $i = 1, \dots, n, j = 1, \dots, k$. The first step is to assess if a linear relationship between y and x exists, using the statistical methods in section “Linearity and Range.” Once the linear relationship between y and x is established, the next step is to check if, for example, the 95% confidence intervals of a and b are within their acceptable ranges, $(-0.5, 0.5)$ and $(0.9, 1.1)$, respectively, with 95% assurance, where a and b are parameters in the following model

$$y_{ij} = a + bx_i + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2), i = 1, \dots, n; \quad j = 1, \dots, k.$$

If these confidence intervals are within their acceptable ranges, then the assay is accepted as accurate; otherwise, it is rejected. When the measurement errors are not constant across all concentration levels, the weighted regression approach might be used.^[5]

Method Comparison

The purpose of method comparison is primarily to determine whether the method being validated is a suitable replacement for a current method. During a quantitative assay validation process, a typical situation is that other existing assays are available in the market. These existing assays are perhaps more expensive, less accurate, harder to operate, slower in speed, etc. The new assay being validated is expected to overcome some or all of these deficiencies while still maintaining values similar to the existing ones for clinical diagnoses. The similarity of values makes a full scale of clinical establishment of normal range, etc., unnecessary. Therefore, a typical design is to select one established assay as a reference, and compare the new assay to this reference assay. If the results reported by both assays are comparable, then the new assay is acceptable. The following part of section reviews the model setup, parameters to be validated, and statistical methods for method comparison.

To determine whether the new assay being validated (y) is a suitable replacement of the selected reference assay (x) or if the two methods yield equivalent results, the conditional means of both methods are usually compared. Suppose that z is an analyte concentration level of a specimen (a tube of body fluid, such as blood or urinary sample). This concentration z varies from person to person. Hence, z is a random variable. In addition, the true concentration z is usually unknown and might be unobservable. Suppose that x and y are measurements of the same specimen of true concentration z using the reference and the new methods, respectively. Therefore, a general model is

$$\begin{cases} x|z = E(x|z) + \varepsilon_{x|z}, E(\varepsilon_{x|z}) = 0, \\ y|z = E(y|z) + \varepsilon_{y|z}, E(\varepsilon_{y|z}) = 0, \end{cases}$$

where $\varepsilon_{x|z}$ is the measurement error for a given z , $E(x|z)$ is the conditional expectation for a given z , etc. As commonly seen from the precision study (section “Precision”), $Var(\varepsilon_{x|z})$ and $Var(\varepsilon_{y|z})$ are approximately proportional to z^2 . The task of method comparison is usually to show that $E(x|z)$ and $E(y|z)$ are close enough to ascertain the “agreement” of two assays. The meaning of “ $E(x|z)$ and $E(y|z)$ are close enough” will be explored further below.

Traditionally, “ $E(x|z)$ and $E(y|z)$ are close enough” is evaluated under linear assumption; that is,

$$E(y|z) = a + b, E(x|z),$$

where a and b are unknown constants. If $a = 0$ and $b = 1$, then $E(x|z) = E(y|z)$ for all z . This certainly means that $E(x|z)$ and $E(y|z)$ are close enough. Hence, the validation is to check if a is close to 0 and b is close to 1. Traditionally, the hypotheses are set up as: $H_0 : a = 0$ and $b = 1$ versus $H_1 : a \neq 0$ or $b \neq 1$.

Three classical methods to test H_0 are: classical linear regression, Deming’s regression,^[6] and Passing-Bablok method.^[7] (A complete review of these methods can be found in.^[8]) A brief description of classical regression only is given here. Classical regression assumes:

$$y = a + bx + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

where σ^2 is an unknown positive parameter. The check of linearity between y and x is done through visual inspection of a scatter plot (Fig. 1) and a bias plot (Fig. 2). The rest of the validation process follows classical linear regression procedures.

There are a few serious drawbacks of classical linear regression. The regression assumes: (i) there is no measurement error in x , the reference method; (ii) the measurement error of the assay being validated is not related to the concentration level (in other words, the measurement error is a constant); and (iii) a linear relationship between y and x . All of these assumptions might be invalid. The

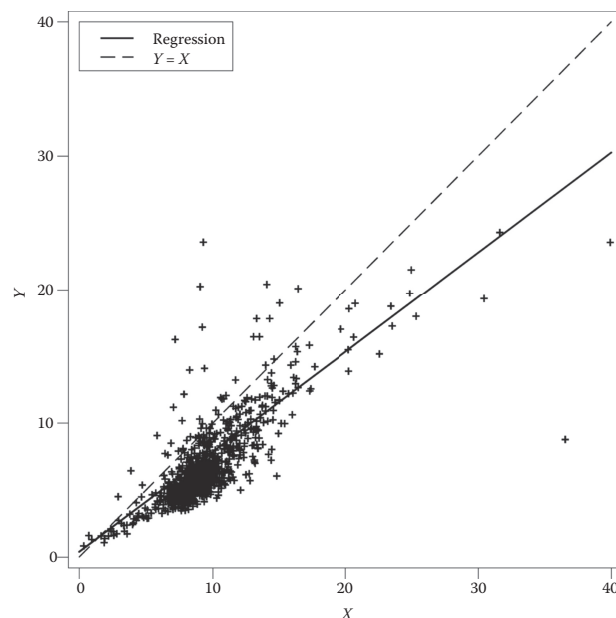


Fig. 1 Scatter plot of Y vs. X . The data were from a real clinical trial, with 1258 specimens

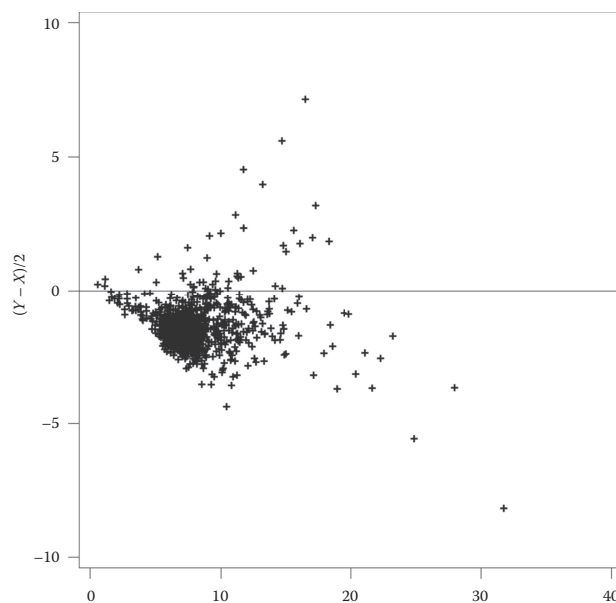


Fig. 2 Bias plot of data of Fig. 1

assumption (iii) is essential to all regression-based methods. In practice, this assumption is often violated (see Fig. 1). When repeated measures of each specimen by each assay are available, Zhong and Shao⁹ recently proposed a few new methods, which require few assumptions. As a result, all three assumptions imposed by linear regression are not needed anymore. These methods are based on a different measure of the agreement between two assays,

$$\delta = E \left\{ \left[E(x - y | z) \right]^2 / \left[Var(x - y | z) \right] \right\}$$

δ has several nice properties. For example, $\delta = 0$ when $E(x|z) = E(y|z)$. Another example, a larger δ usually indicates greater conditional mean differences (i.e., if δ is large, $E(x|z) - E(y|z)$ is greater). The hypotheses are $H_0: \delta \geq \delta_0$ vs. $H_0: \delta \geq \delta_0$, for pre-selected δ_0 . The detailed testing methods can be found in.^[9]

Another measure of the agreement between two assays is

$$\delta_{\downarrow} 1 = E[E(x - y | z)]^2 / (E[V(x - y | z)])$$

Note that generally, $\delta_1 > \delta$. The detailed testing methods based on δ_1 can be found in Ref. [28].

Other methods for evaluating agreement of two assays are related to the concordant correlation coefficient.^[10] A serious drawback of these methods is that the reference method (x) is assumed to be a gold standard. This assumption is impractical.

Precision

The precision of an assay expresses the closeness of agreement (degree of scatter) between a series of measurements obtained from multiple sampling of the same homogenous sample under the prescribed conditions.^[3] Precision may be considered at three levels: repeatability, intermediate precision, and reproducibility. Repeatability (intra-assay precision) expresses the precision under the same operating conditions over a short interval of time. Intermediate precision expresses within-laboratories variation: different days, different analysts, different equipment, etc. Reproducibility expresses the precision between laboratories. The intermediate precision and reproducibility are also called ruggedness in the USP/NF.^[11]

Therefore, the design of validation of precision varies according to the intended use of an assay. The intended use of an assay also decides how many factors should be included in the validation. For example, if all measurements of an assay is performed by a central laboratory using only one instrument and one reagent lot, then laboratory, instrument, and reagent lot should not be included as factors in the validation. On the other hand, if more than one laboratory, reagent lot, and instrument will be used, then these factors should be included in the validation study.

A commonly used parameter to describe precision is the total coefficient of variation (CV), which is defined as the total standard deviation divided by the grand mean, in which the total variance is the sum of all variance components (factors) being considered, such as instrument, reagent lot, day and run; and the grand mean is the average of all results. The total CV, or the total %CV, instead of the total standard deviation, is used as the precision parameter since the total standard deviation changes with the concentration level of the sample (hence the grand mean). The commonly seen industrial standards of precision in %CV for assays related to hematology are below 15%. This requirement should be met within a clinically meaningful

range. A few concentration levels are usually chosen within the clinically meaningful range and the validation is then to focus on each concentration level.

To simplify our discussion, we assume that these concentration levels have been chosen and only two factors (day and run) are included in the validation. The study designs and statistical inferences in this particular situation will be discussed in the following subsections.

Study Designs and Statistical Inferences

A design of two factors (day and run) to verify the precision of an assay needs to specify how many replicates are required within a run, how many runs are required within a day, and how many days are required for the study. There are a few NCCLS guidelines which provide a few designs with different numbers of days, runs and replicates.^[12,13] However, there is no statistical justification regarding when these designs should be used and which one is optimal. A commonly used design is to run an assay in 20 days, with two runs within a day and 2 replicates within a run. A general statistical model for these designs follows.

A model for a design which operates within at least two days, with at least two replicates within each run, and with at least two runs within each day is

$$y_{ijk} = \mu + d_i + r_{j(i)} + \varepsilon_{k(ij)},$$

where y_{ijk} is the result reported by the assay, d_i is the random factor representing between-day variations, $r_{j(i)}$ is the random factor for between-run variations nested in d_i , and $\varepsilon_{k(ij)}$ is the random factor for within-run variations nested in $r_{j(i)}$; $d_i \sim N(0, \sigma_d^2)$, $r_{j(i)} \sim N(0, \sigma_r^2)$, $\varepsilon_{k(ij)} \sim N(0, \sigma_\varepsilon^2)$, d_i , $r_{j(i)}$, and $\varepsilon_{k(ij)}$ are mutually independent; μ , σ_d , σ_r , and σ_ε are unknown constants; $i = 1, \dots, D$, $j = 1, \dots, R$, $k = 1, \dots, N$; and D is the number of days, R is the number of runs, and N is the number of replicates. The total variance is defined as

$$\sigma_{total}^2 = \sigma_\varepsilon^2 + \sigma_r^2 + \sigma_d^2$$

The total CV is defined as σ_{total}/μ . In this model, the unbiased point estimate of each variance component and the total variance, and confidence intervals of σ_{total} and σ_{total}/μ are calculated as follows.

where $\bar{y}_{i..} = \frac{1}{RN} \sum_{j=1}^R \sum_{k=1}^N y_{ijk}$, $\bar{y}_{ij.} = \frac{1}{N} \sum_{k=1}^N y_{ijk}$, and $\bar{y}_{...} = \frac{1}{DRN} \sum_{i=1}^D \sum_{j=1}^R \sum_{k=1}^N y_{ijk}$. Denote $MS_\varepsilon = \frac{1}{(N-1)RD} \sum_{i=1}^D \sum_{j=1}^R \sum_{k=1}^N (y_{ijk} - \bar{y}_{ij.})^2$, $MS_r = \frac{1}{(R-1)D} \sum_{i=1}^D \sum_{j=1}^R (\bar{y}_{ij.} - \bar{y}_{i..})^2$, and $MS_d = \frac{1}{D-1} \sum_{i=1}^D (\bar{y}_{i..} - \bar{y}_{...})^2$. Then, the unbiased estimates for the variance components using the method of moments are (negative estimates should be set to 0):

$$\hat{\sigma}_\varepsilon^2 = MS_\varepsilon, \hat{\sigma}_r^2 = \frac{1}{N} (MS_r - MS_\varepsilon), \hat{\sigma}_d^2 = \frac{1}{NR} (MS_d - MS_r),$$

$$\hat{\sigma}_{total}^2 = \hat{\sigma}_\varepsilon^2 + \hat{\sigma}_r^2 + \hat{\sigma}_d^2 = \frac{N-1}{N} MS_\varepsilon + \frac{R-1}{NR} MS_r + \frac{1}{NR} MS_d.$$

The sums of squares and their expectations are:

Source	DF	SS (Sum of Squares)	MS (Mean Square)	E(MS)
Day	$D - 1$	$NR \sum_{i=1}^D (\bar{y}_{i..} - \bar{y}_{...})^2$	$\frac{NR}{D-1} \sum_{i=1}^D (\bar{y}_{i..} - \bar{y}_{...})^2$	$\sigma_\varepsilon^2 + N\sigma_r^2 + NR\sigma_d^2$
Run	$(R - 1)D$	$N \sum_{i=1}^D \sum_{j=1}^R (\bar{y}_{ij.} - \bar{y}_{i..})^2$	$\frac{N}{(R-1)D} \sum_{i=1}^D \sum_{j=1}^R (\bar{y}_{ij.} - \bar{y}_{i..})^2$	$\sigma_\varepsilon^2 + N\sigma_r^2$
Rep	$(N - 1)RD$	$\sum_{i=1}^D \sum_{j=1}^R \sum_{k=1}^N (y_{ijk} - \bar{y}_{ij.})^2$	$\frac{1}{(N-1)RD} \sum_{i=1}^D \sum_{j=1}^R \sum_{k=1}^N (y_{ijk} - \bar{y}_{ij.})^2$	σ_ε^2
Total	$NRD - 1$	$\sum_{i=1}^D \sum_{j=1}^R \sum_{k=1}^N (y_{ijk} - \bar{y}_{...})^2$		

An approximately $100*(1 - \alpha)\%$ confidence interval for σ_{total} is:

$$\left(\frac{\sqrt{T} \hat{\sigma}_{total}}{\sqrt{\chi^2_{1-\frac{\alpha}{2}}(T)}}, \frac{\sqrt{T} \hat{\sigma}_{total}}{\sqrt{\chi^2_{\frac{\alpha}{2}}(T)}} \right),$$

$$T = \frac{D[(N-1)R \cdot MS_\varepsilon + (R-1)MS_{rr} + MS_{dd}]^2}{(N-1)R \cdot MS_\varepsilon^2 + (R-1)MS_{rr}^2 + \frac{D}{D-1} MS_{dd}^2},$$

where $\chi_p^2(T)$ is the p th quantile of the χ^2 -distribution with the degrees of freedom $T(p = 1 - \alpha/2$ or $\alpha/2)$.^[14] A commonly acceptable approximate $100*(1 - \alpha)\%$ confidence interval for the total CV is

$$\left(\frac{\sqrt{T} \hat{\sigma}_{total}}{\bar{y}_{...} \sqrt{\chi^2_{1-\frac{\alpha}{2}}(T)}}, \frac{\sqrt{T} \hat{\sigma}_{total}}{\bar{y}_{...} \sqrt{\chi^2_{\frac{\alpha}{2}}(T)}} \right),$$

where $\bar{y}_{...}$ is treated as a constant in constructing the confidence interval.

Matrix Effects

The results of quantitative assays, as well as qualitative assays, are affected by many factors in practice. These factors are called “matrix effects.” For instance, the result of a hematology-related assay may be affected by interfering substances, and chemical and physical properties that are different from the ideal specimen collection and measurement conditions (such as drugs, alcohol, anticoagulants, storage conditions, etc.). (When the validation is conducted to assess unequivocally the analyte in the presence of components which may be expected to be present, it is called selectivity (specificity).^[3,11]) Using a blood-related test which measures the ferritin concentration as a specific example, the question is, does a person with excessive lipids

in his/her blood stream after a fatty meal have a positive or negative impact on the result? Another example is that specimens are usually frozen and shipped to a central laboratory for testing. If the testing results are quite different from the results tested when samples are fresh, this kind of procedure is certainly not appropriate. Hence, the objective of matrix effect validation is to confirm which matrix effect has an impact on the assay's result and which does not.

As proposed by NCCLS EP-7A,^[15] a matrix effect should be evaluated at least at two medical concentrations of the analyte. We will use the ferritin concentration in blood testing where lipid is a possible interfering substance as an example. First, at each concentration, two pools of specimens of equal concentrations of ferritin are prepared. Second, two tubes of equal amounts of solvent are prepared, one without lipids, one with “excessive” lipids. The “excessive” means that once the tube of solvent with lipids is added to a pool, it simulates the “worst case” conditions when lipids are interfering with ferritin. Finally, the two tubes of solvent are added to the two pools of specimens. One pool without added lipids is called a control pool. The other is called a test pool. These two pools should have quantity enough to allow for repeated measurements, as mandated by the power requirement to declare that there is or there isn't an interfering effect.

Let $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_m$ denote the results of measurements from the control pool and test pool, respectively. Assume $x_i \sim N(\mu_0, \sigma^2)$, $y_j \sim N(\mu_1, \sigma^2)$, $i = 1, \dots, n$; $j = 1, \dots, m$; and μ_0, μ_1 , and σ are unknown constants. The matrix effect analysis is to test the following statistical hypotheses

$$H_0: |\mu_1 - \mu_0| \geq \delta \quad \text{vs.} \quad H_a: |\mu_1 - \mu_0| < \delta,$$

where δ is a pre-specified clinical acceptance limit (in difference). Sometimes, the clinical acceptance limit is given by the relative difference. Then the above hypotheses can be rewritten as

$$H_0: \left| \frac{\mu_1 - \mu_0}{\mu_0} \right| \geq \delta_r \quad \text{vs.} \quad H_a: \left| \frac{\mu_1 - \mu_0}{\mu_0} \right| < \delta_r$$

accordingly. The tests of the hypotheses above are equivalent to whether the corresponding confidence intervals are within $(-\delta, \delta)$ or $(-\delta_p, \delta_p)$, respectively, where the $(1 - \alpha)$ confidence interval for $\mu_1 - \mu_0$ is

$$(\bar{y} - \bar{x}) \pm t\left(1 - \frac{\alpha}{2}, m + n - 2\right) \cdot s_p \sqrt{\frac{1}{m} + \frac{1}{n}},$$

and the approximate $(1 - \alpha)$ confidence interval for $\frac{\mu_1 - \mu_0}{\mu_0}$ is

$$\frac{\bar{y} - \bar{x}}{\bar{x}} \pm t\left(1 - \frac{\alpha}{2}, m + n - 2\right) \cdot \frac{s_p}{\bar{x}} \sqrt{\frac{1}{m} + \frac{1}{n}}$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \bar{y} = \frac{\sum_{i=1}^m y_i}{m}, s_p = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}},$$

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, s_y^2 = \frac{1}{m-1} \sum_{i=1}^m (y_i - \bar{y})^2,$$

and $t(1 - \frac{\alpha}{2}, m + n - 2)$ is the $(1 - \alpha/2)$ percentile of t distribution with the degrees of freedom $m + n - 2$.

Another typical design is the one that does not pool specimens. Under this design, two tubes of specimens are collected from each subject. Then a certain amount of solvent is added to each tube, one with the interfering substance, one without the interfering substance. Each tube of specimens is then tested repeatedly by the assay. The appropriate analyses for this design are the statistical methods presented in the “Method Comparison” section.

Linearity and Range

For a quantitative assay, one essential requirement is that the results reported by the assay are linear to the true concentrations in the reportable range, where the results are the final analytical results the assay reports. The reportable range is the interval for which it has been demonstrated that the assay has a suitable level of precision, accuracy, and linearity. The reason for requiring this linear relationship is based on clinical grounds. For example, assuming that a scale of weight is an “assay,” we expect that a person weighing 200 lbs. is truly twice as heavy as a person weighing 100 lbs.

Hence, the purpose of the validation is to determine the linearity of an assay. The statistical process for this validation is to examine if fitting a model with the linear relationship is adequate or if using the quadratic or some higher-order polynomials is more appropriate.

Study Designs and Statistical Analysis

NCCLS Guideline EP6-P2^[6] recommends several study designs. One of them is the high-low pool method of preparing samples at different concentrations. The details of

this method follow: (i) Choose two samples with known concentrations, one sample with high concentration and one sample with low concentration. The concentrations of the high and low samples should be close to the upper and lower limits of the reportable range, respectively. (ii) Create a few samples with different concentration levels between the high and low concentrations by pooling various proportions of the high and low concentration samples. The concentration levels of these created samples should be evenly distributed between the concentrations of the high and low samples. For instance, if we want to create three samples with concentrations evenly distributed between the high and low concentrations, we can prepare these samples using 75%L + 25%H (75% of one unit of high concentration sample and 25% of one unit of low concentration sample), 50%L + 50%H, and 25%L + 75%H respectively. The concentrations of these created samples can be calculated from the known concentrations of the high and low samples and the proportions used to create them.

Once these samples are created, a couple of measurements of each sample are usually conducted. Suppose a third-order polynomial model is postulated to validate the linearity of the assay, i.e.,

$$y_{ij} = a + bx_i + cx_i^2 + dx_i^3 + \varepsilon_{ij},$$

where x_i is the known concentration level; y_{ij} is the measurement of the sample with concentration x_i ; a, b, c and d are unknown constants; $\varepsilon_{ij} \sim N(0, \sigma^2)$; and $i = 1, 2, \dots, n, j = 1, 2, \dots, k_i$. If the relationship between y_{ij} and x_i is linear, then c and d should be 0. Therefore, the statistical validation of linearity focuses on whether c and d are 0. A confidence interval approach based on traditional linear regression is usually utilized to reach a conclusion about the linearity of the assay. For example, if the 95% confidence interval of c is within $(-\delta_c, +\delta_c)$, then c is considered to be equivalent to 0 (δ_c is a pre-specified acceptance limit).

Further Considerations

There are a few points of caution about the statistical validation procedures presented above. First, note that a constant measurement error across all concentrations is assumed. However, this assumption is usually not true as can be seen from precision studies. If the measurement errors are proportional to the concentration level, which is true for many assays, then the weighted regression approach is preferred.^[5] Second, a third-order polynomial model is assumed first, and then we establish that the second- and third-order terms are not significant. However, confirming that the second- and third-order terms are not significant does not automatically exclude the significance of even higher-order terms. One possible solution follows:

Establish the best-fit polynomial (using a model-selection approach), $p(x)$, and the linear-fit polynomial, $l(x) = a + bx$, between y and x . The search for the best-fit

polynomial $p(x)$, for example, starts with the highest possible order of the polynomial, the number of concentration levels, and uses a backward model-selection approach. Then the biases at each concentration x , defined by

$$B(x) = p(x) - l(x),$$

are calculated. If all biases are within the pre-specified medical acceptance limits, then the linearity of the assay is accepted; otherwise, it is not accepted. However, a statistically rigid testing procedure, based on $B(x)$, needs to be studied further.

Limits of Detection and Quantitation

Limit of detection (LOD) of an assay is the lowest amount of analyte in a sample which can be detected but not necessarily quantitated as an exact value. Limit of quantitation (LOQ) of an assay is the lowest amount of analyte in a sample which can be quantitatively determined with suitable precision and accuracy.^[3] Hence, LOD is a parameter which merely substantiates that analyte concentration is above or below a certain level;^[4] LOQ is a parameter of a quantitative assay for low levels of compounds in sample metrics, and is used particularly for the determination of impurities and/or degradation products. However, there are no accepted standard interpretations of these terminologies. Although each therapeutic area might have its own conventional interpretations, it is important to understand whether the estimates are useful for the intended use.

Traditionally, the LOD of an assay has been calculated from an intra-assay measure reflecting imprecision of the zero calibrator.^[17] However, this method is generally considered clinically irrelevant and usually giving an overly optimistic estimate of LOD.^[18] A more clinically relevant measure, LOQ, or functional sensitivity, is generally more preferable. One commonly used interpretation of LOQ is the lowest concentration level at which an assay achieves a CV of 20%, which represents the lowest amount of an analyte in a sample that can be quantitatively determined with suitable precision and accuracy. This is the definition used in the estimation below.

Some ways of estimating LOD and LOQ follow. (ICH Guideline, Q2B, "Validation of analytical procedures: methodology" gives some standard and simple ways of estimating LOD and LOQ.)

The estimation of LOD is sometimes connected with analytical sensitivity. Analytical sensitivity is a threshold (cutoff) such that when an assay's result of a sample is greater than the threshold, it is considered that the sample contains some positive amount of the analyte (with a high probability). In that sense, the analyte is detected in the sample. Analytical sensitivity is obviously of interest in testing for impurities, where the presence or absence of certain undesirable substance is the critical information desired from the assay.

Analytical sensitivity is usually determined as the $(1 - \alpha)$ 100 percentile of the distribution of assay results from blank samples (with zero amount of the analyte). LOD is then estimated as the concentration of the analyte in a sample at which the assay's result is greater than the analytical sensitivity with $(1 - \alpha)$ probability.

Suppose that y is the response (after calibration) at concentration x , and

$$y \sim N(x, (a + bx)^2),$$

where a and b are unknown positive constants. Then, an example of study design and estimation procedure of estimating LOD is given below ($\alpha = 0.05$):

1. Prepare a panel with 5 members such that the target concentrations of the panel members sandwiching the underline LOD and LOQ (the concentration of the i th member is denoted as x_i).
2. Assay each member 30 times, repeatedly. (Let y_{ij} be the j th repeated measurement at x_i , $j = 1, \dots, 30$; $i = 1, \dots, 5$.)
3. Calculate $SD(x_i)$, $i = 1, \dots, 5$, using formula

$$SD(x_i) = \frac{1}{29} \sum_{j=1}^{30} (y_{ij} - \bar{y}_i)^2, \text{ where } \bar{y}_i = \frac{1}{30} \sum_{j=1}^{30} y_{ij}.$$

4. Calculate linear regression estimators $\hat{a}$ and $\hat{b}$ of a and b , using $SD(x_i)$ and x_i , $i = 1, \dots, 5$.
5. Estimate LOD using the formulas

$$LOD = 3.290\hat{a}/(1 - 1.645\hat{b}), LOQ = \hat{a}/(0.2 - \hat{b}).$$

The estimates of LOD and LOQ are "snapshots" or the point estimates of the LOD and LOQ. A better way is to present 95% confidence intervals of the LOD and LOQ. This can be achieved through classical linear regression techniques.

EVALUATION OF QUALITATIVE ASSAYS

Introduction

Sensitivity and specificity are two indices commonly used to evaluate the performance of a qualitative assay. As mentioned earlier, qualitative assays usually are designed to result only in positive/negative responses. Naturally, if patients are diseased, we want to know what the positive rate is by the assay; if patients are non-diseased, we want to know what the negative rate is by the assay. These two rates (positive and negative) are called sensitivity and specificity, respectively. However, using sensitivity and specificity as the performance measures of a qualitative assay is controversial.^[19] The question is: do these two parameters vary from population to population?

The answer to this question is “yes.” This can be seen in the following example. Suppose that the concentration of the analyte a qualitative assay measures is continuous (this is a commonly seen practical situation). As discussed in the following section, qualitative assays are established first by selecting a cutoff point in a receiver operating characteristic (ROC) analysis, and then a qualitative assay’s “raw” response is dichotomized further into “positive” or “negative,” i.e., if a raw response is greater than or equal to the cutoff, then the specimen is classified as positive; otherwise, it is negative. All assays have variations in their responses to a given concentration level; hence, a population with more diseased patients whose concentration levels are around 0 will have a greater chance of being misdiagnosed as negative, even if the two populations have equal proportions of diseased patients. This point is demonstrated further in Fig. 3.

Though sensitivity and specificity have this drawback, there are no other commonly accepted substitute measures. This limitation, however, can be resolved if each assay is validated in the population for which it is intended and labeled accordingly. This is commonly required by most regulatory agencies.

Most validations of qualitative assays are centered on sensitivity and specificity. The precision study of a qualitative assay is an exception. The precision study of a qualitative assay is almost an exact copy of the precision study of a quantitative assay. Refer to section “Precision” for further details.

The ROC analysis, the method used to estimate sensitivity and specificity, and matrix validation will be presented in the following subsections, respectively.

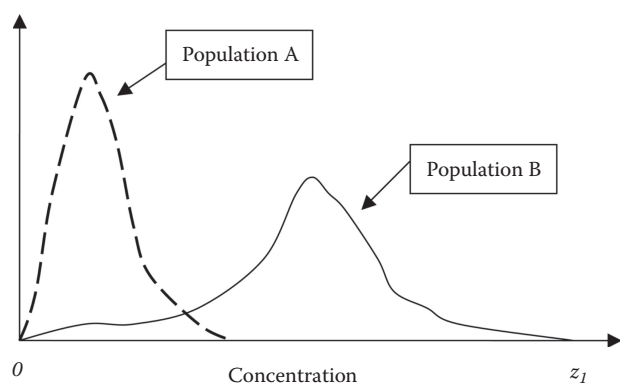


Fig. 3 Distributions of two populations with equal proportions of diseased patients. *Note:* For example, suppose an assay is used to diagnose HIV. In this case, though Population A and Population B have equal disease-prevalence rates, Population A has more patients with HIV analyte concentration levels close to 0 than Population B. That is, more patients from Population A were infected with HIV recently. Hence, the same HIV assay will have lower sensitivity in Population A than in Population B

ROC Analysis

As discussed in the sections “The Key Difference and Similarity Between Quantitative and Qualitative Assays” and “Introduction,” typically, a qualitative assay’s raw response (before it is dichotomized into positive/negative) is distributed within a continuous range of values and the dichotomization is based on a decision threshold, or cutoff. That is, for example, if the raw response of the assay is less than the cutoff, then the assay’s final diagnostic result is negative; otherwise, the final result is positive. The patient will be diagnosed as non-diseased (negative) or diseased (positive), accordingly. Hence, the determination of the cutoff is critical to this assay.

Since assays are rarely perfect, there are always overlaps between the raw responses of non-diseased and diseased patients. Therefore, by varying the cutoff value, any sensitivity from 0 to 1 (specificity from 1 to 0) can be obtained, and each one will have a corresponding specificity.^[20] A ROC curve is a plot of sensitivity versus 1-specificity when the cutoff value varies over the complete range of the assay’s measurements.

A ROC curve is a fundamental evaluation tool in clinical medicine. It provides a global assessment of accuracy by demonstrating the trade-off between sensitivity and specificity, which captures an assay’s ability to distinguish between alternative states of health, over the complete spectrum of operating conditions. Once the curve is generated, a user can readily go on to many activities, such as assessing the value of an assay as a diagnostic tool, selecting an optimal cutoff, and performing comparisons of assays. When the consequence of a diagnosis from an assay is associated with costs, a ROC curve can be incorporated into a clinical strategy by using decision analysis. A brief history of the applications of ROC curves in the area of medicine was described in.^[21]

In a typical process of a qualitative assay’s development, a ROC curve is used to determine an optimal cutoff. This is the focus of this section.

ROC Curve and Its Application

A ROC curve can be constructed using either a parametric or nonparametric method. A parametric method is based on the assumption that the distribution of an assay’s measurements for a diseased population and a non-diseased population are known except a few parameters. Details of parametric methods can be found in.^[22] A nonparametric method is based on the true disease-states (diseased or non-diseased) and an assay’s measurements of all samples used in the validation, which is specified further below:

- Calculate sample sensitivity and specificity when the cutoff value varies over the complete range of the assay’s measurements.

- Plot the sensitivity versus 1-specificity, using sensitivity as the y-axis and 1-specificity as the x-axis, and connect the data points in ascending order of sensitivity.

A ROC curve constructed in the manner described above is an increasing-step function and connects the two points of (0, 0) and (1, 1). An assay with perfect accuracy achieves 100% sensitivity and 100% specificity at one or more cutoff points. The ROC curve of this perfect assay goes through the point (0,1). An assay that does not distinguish between truly positive and truly negative subgroups has a ROC curve as the diagonal line connecting the points of (0, 0) and (1, 1). These are two extreme cases. An assay with good clinical performance achieves both high sensitivity and specificity, which corresponds to high sensitivity (true positive rate) and low 1-specificity (false positive rate). Assays with high diagnostic accuracy (high clinical sensitivity and specificity) have a ROC curve close to the upper left corner of the square. The closer the curve is to the upper left corner, the better the accuracy of the assay.

Based on the ROC curve, an optimal cutoff value can be determined, depending on the optimization criteria. For instance, if the goal is to maximize the sum of sensitivity and specificity, the optimal cutoff value corresponds to the point on the curve which has the longest distance to the line which connects (0, 0) and (1, 1).

An obvious drawback of a standard ROC curve is that its intrinsic connection to the cutoff values has been lost. This connection can be re-established by adding an additional axis, as shown in Fig. 4. Note that the axis labeled “cutoff” in Fig. 4 is not of linear scale. Fig. 4 was produced using SAS (PROC Gplot). The optimal cutoff value is based on the maximization of the sum of sensitivity and specificity.

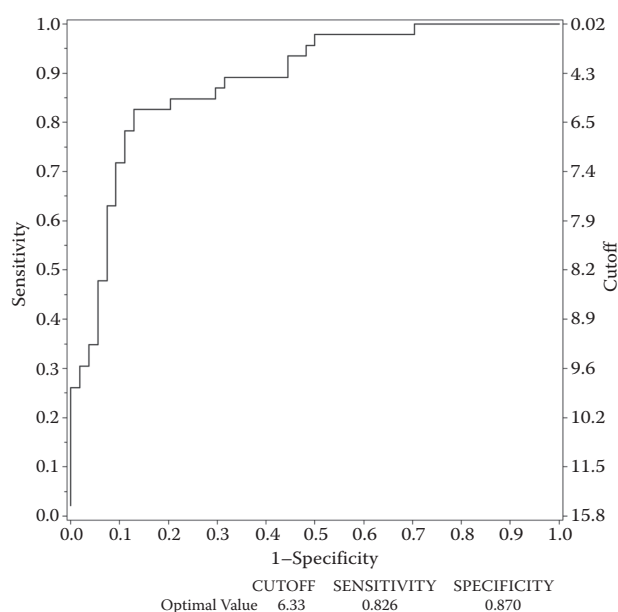


Fig. 4 An example of a ROC curve with cutoff values

Other Applications of the ROC Curves

There are two other important applications of the ROC curves. One is to quantify the diagnostic accuracy of an assay by a number. The most common measure is the Area Under the ROC Curve (AUC). By convention, this area is always at least 0.5. If not, the assay can be improved by simply switching the definitions of “Positive” and “Negative.” The AUC values range between 0.5 (an assay with no test value) and 1.0 (a perfect assay).

The AUC is equivalent to the Wilcoxon rank-sum statistic or Mann-Whitney nonparametric two-sample statistic, with the positive subgroup as one sample and the negative as the other (20). It represents the probability that $X > Y$ (assuming a larger value is more likely to associate with a positive), where X is the assay’s result for a randomly selected person from the positive group and Y for a randomly selected person from the negative subgroup (in case of no tie).

The other important application is to make global comparisons of multiple assays. For a visual comparison, if the ROC curve of Assay A is always above the ROC curve of Assay B, then Assay A is uniformly better than Assay B. If two ROC curves cross at some point, then one assay is better on one side than the other, and vice versa. The AUCs of assays also can be used for statistical comparisons, and this comparison can be non-parametrical or parametrical.^[22]

Sensitivity and Specificity

By definition, sensitivity and specificity are conditional probabilities. If a gold standard testing method exists and can be applied to every specimen, then clinical trial design and analysis are quite simple. As recommended by NCCLS EP12-P,^[23] only specimens identified as “diseased” by the gold standard will be used in an estimate of sensitivity, and only specimens identified as “non-diseased” by the gold standard will be used in an estimate of specificity. In either case, binomial distribution is used in design and statistical inferences. For example, the design and statistical inferences for sensitivity are based on the total number of specimens identified as “positive” by the gold standard and the distribution of the frequency of these specimens tested as “positive” by the assay being validated is used for references.

In practice, however, it is much more difficult to get accurate estimates of sensitivity and specificity. First, this gold standard never exists in practice. Several clinically proven methods are usually used together sequentially to establish this “gold” standard method by some criterion, such as “if tested positive by any of these methods, then the specimen is assumed positive by the ‘gold’ standard; otherwise, it is negative.” Other methods to establish this gold standard can be found in.²⁴ The error introduced by this “fake” gold standard is usually ignored. The same assumption will be assumed here, that is, a gold standard

is established some way, and any error introduced by using this “fake” gold standard will be ignored.

The other, more challenging, problem is that this “fake” gold standard may not be applied to all participating patients due to unacceptable expenses or for the sake of the subjects’ welfare. For example, some diseases require taking a small piece from inside a subject’s body in order to know if the subject is diseased. To do such a test on all participating subjects, even when they are perfectly normal, is certainly unethical. Hence, any design needs to decrease the frequency of using the gold standard while still obtaining estimates of sensitivity and specificity as accurately as possible.

A design and estimation method that has long been used is called “discordant resolution.” Discordant resolution uses two steps for testing. First, it employs both a customarily used assay and the new assay being validated to evaluate all subjects. Second, the gold standard is applied to selected subjects with discordant testing results by the customarily used and the new assays. The true disease states of all subjects with concordant results are estimated as the concordant results, positive or negative, accordingly. This design, together with its estimation method, has been criticized by many.^[25,27] Other methods^[26,27] have been proposed recently. The drawback of Becker’s method^[26] is discussed in.^[2] The drawback of^[27] is that the asymptotical variances of their estimators are asymptotically equivalent to the inverse of the second phase sample size. Hence, its application is limited. One method proposed by^[2] follows.

Suppose that sensitivity is the parameter being validated (specificity is similar to sensitivity). Also, suppose that y is the assay being validated; x is the customarily used assay; z is the gold standard; and the values of x , y , and z are dichotomized from x_1 , y_1 , and z_1 respectively, where x_1 and y_1 are the raw responses of assay x and y , respectively, and z_1 is the true concentration level.

The design is to borrow strength from x_1 and y_1 for the selection of specimens to be retested by z and for the estimation of the true disease state of a specimen. The rationale is that x_1 and y_1 have some quantitative characteristics such as the fact that a larger raw response will be observed when the concentration level of the analyte in a specimen is higher. Keeping that fact in mind, and suppose at 1 (the cutoff), the standard deviations of $SD(x_1)$ and $SD(y_1)$ of x_1 and y_1 are known, then with almost 100% certainty that if a specimen has both x_1 -value less than $1 - 3 * SD(x_1)$ and y_1 -value less than $1 - 3 * SD(y_1)$, then we can conclude that the specimen is negative. Hence, retesting with z needs only to be applied to specimens of x_1 -value above $1 - 3 * SD(x_1)$ or y_1 -value above $1 - 3 * SD(y_1)$. The detailed retesting schema with z and the estimation method follow:

First, test all specimens using x and y , with their x_1 -values and y_1 -values recorded. Then, test all specimens with x_1 -values above $1 - 3 * SD(x_1)$ or y_1 -values above $1 - 3 * SD(y_1)$ using gold standard z . The true disease state of a specimen is the test result of z if the specimen is tested by

z ; otherwise, it is estimated as negative (Zhong’s method). All results then are summarized into the following table:

		Zhong’s method	
		1	0
Y	1	f_{11}	f_{10}
	0	f_{01}	f_{00}

The estimation of $SE(y)$ is:

$$SE(y) = \frac{f_{11}}{f_{01} + f_{11}}.$$

The simulations in² showed that this method is quite accurate. However, the theoretical property of this method needs to be explored further.

Matrix Effects

Matrix effects validation of qualitative assays needs to deal with similar tasks as those of quantitative assays. The design of the experiments and acceptance criteria, however, are quite different. To show the commonly used validation process, the following example uses a study which determines if excessive lipids are interfering with the assay’s ability to detect HIV.

First, approximately 30 healthy volunteers are selected. Next, each volunteer’s blood is collected in two tubes. (These specimens are then tested to confirm that they are HIV negative; HIV positives are removed.) Third, HIV positive material is added to one tube of each donor’s blood. Hence, for each donor, there is one tube of HIV negative blood, one tube of HIV positive blood. Last, a certain amount of lipids is added to each negative and positive blood tube to simulate practical situations of blood samples with excessive lipids. The acceptance criteria are that all negative specimens still tested negatively, and all positive specimens still tested positively by the assay being validated.

There are several problems with the above design and validation procedures related to sensitivity. One problem is that it is hard to decide how much HIV positive material should be added to each tube with HIV negative blood. As discussed in section “Sensitivity and Specificity,” if too much HIV analyte is added, the concentration of the HIV analyte is too high, and it is unlikely to test the blood samples negatively with the added lipids. Though clinicians usually add less than 5% of HIV analyte into each blood tube to create HIV positive specimens, it is difficult to justify too much or too little HIV analyte was added. This is, indeed, an arguable point between regulatory bodies and pharmaceutical companies. Therefore, the question is, how much HIV analyte should be added to simulate practical situations where a patient is HIV positive and the patient’s

Validation of Quantitative and Qualitative Assays

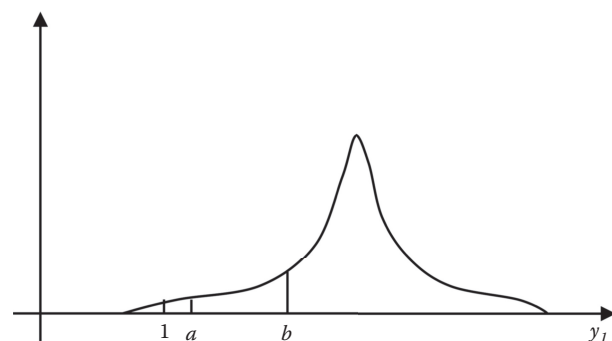


Fig. 5 The distribution of responses of the assay being investigated and its quantiles (diseased specimens)

Note: $P(y_1 < a) = 0.03$, $P(y_1 < b) = 0.06$

blood sample also contain excessive lipids. Another problem is that the sample size is decided subjectively without any statistical justification. Still, another problem is the question what the rationale of the acceptance criteria of “the positives remain positive” and “the negatives remain negative”, or “no qualitative change” is?

Zhong^[2] created a new method which can solve all the above problems for evaluation related to sensitivity. The suggested acceptance criterion for sensitivity is that the sensitivity under the testing condition is not less than the design requirement of the sensitivity of the assay under standard testing conditions, i.e.

$$SE(y | a \text{ testing condition}) \geq SE_0,$$

where SE_0 is the design requirement of sensitivity. The amount of the analyte to be added to the blood collection tube, or in other words, the allowable range $[a, b]$ measured by the assay, should be assay specific and sample-size related. Under the assumption that any specimens with the raw response values above the selected b remain positive under a testing condition, then the design and validation process is illustrated as follows.

1. Calculate the total sample size, n , for the design requirement of sensitivity under standard conditions.
2. Select two positive numbers $0 < q_0 < q_1 < 1$ and estimate the corresponding quantiles of the assay under standard conditions, where $1 - q_0 > SE_0$. For example, suppose that $SE_0 = 0.95$ and two positive numbers are selected as $q_0 = 0.03$, $q_1 = 0.06$. The corresponding quantiles to 0.03 and 0.06 of the raw response of y under standard conditions are estimated as a and b , respectively. See Fig. 5.
3. Calculate the number of positive specimens needed, using $n_1 = (q_1 - q_0) * n$.
4. Create n_1 positive specimens with a raw response distribution in $[a, b]$ the same as the raw response distribution in $[a, b]$ of all specimens under standard clinical testing conditions.

5. Add lipids to these n_1 specimens.
6. Determine if the assay is acceptable. If all n_1 specimens are positive, then the testing condition is acceptable; otherwise, it is unacceptable.

The above schema, however, is not workable for evaluating specificity. The difficulty comes from creating negative specimens with raw responses close to the cutoff. Our current designs might be invalid when evaluating negative populations. This may explain why so many false positive cases have been reported.

CONCLUDING REMARKS

In this entry, we introduced typical studies conducted in the validation of quantitative and qualitative assays. The purpose of each study, the statistical model used, the parameters estimated, and hypotheses tested were explained in each subsection. However, we did not try to cover all studies conducted in the validation process.

Plenty of statistical technologies have been used in the validation process. At the same time, there are many statistical problems unsolved. Some of these unsolved problems are mentioned throughout this entry.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to Ms. Kathryn Nordhaus for her assistance in preparing this manuscript.

REFERENCES

1. Shah, V.; Midha, K.K.; Findlay, J.W.; Hill, H.M.; Hulse, J.D.; McGilveray, I.J.; McKay, G.; Miller, K.J.; Patnaik, R.N.; Powell, M.L.; Tonelli, A.; Viswanathan, C.T.; Yacobi, A. Bioanalytical method validation—A revisit with a decade of progress. *Pharm. Res.* **2000**, *17*, 1551–1557.
2. Zhong, B. Evaluating qualitative assays using sensitivity and specificity. *J. Biopharm. Stat.* **2002**, *12*, 409–424.
3. ICH Guidelines. Text on validation of analytical procedures; Q2A, 1994.
4. Chow, S.C.; Liu, J.P. *Statistical Design and Analysis in Pharmaceutical Science*; Marcel Dekker, Inc.: New York, 1995.
5. Dunn, O.; Clark, V. *Applied Statistics: Analysis of Variance and Regression*; John Wiley & Sons: New York, 1987.
6. Deming, W. *Statistical Adjustment of Data*; Wiley: New York, 1943.
7. Passing, H.; Bablok, W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. *J. Clin. Chem. Clin. Biochem.* **1983**, *21*, 709–720.
8. Zhong, B.; Shao, J. Testing the agreement of two quantitative assays in individual means. *Commun. Stat. Theory Meth.* **2002**, *31*, 1283–1299.

9. Zhong, B.; Shao, J. Evaluating the agreement of two quantitative assays with repeated measurements. *J. Biopharm. Stat.* **2003**, *13*, 75–86.
10. Lin, L.; Hedayat, H.; Sinha, S.; Yang, M. Statistical methods in assessing agreement: models, issues, and tools. *JASA* **1997**, *97*, 257–270.
11. USP/NF. United States pharmacopeia XXII and the national formulary XVII; The United States Pharmacopoeial Convention: Rockville, Md.
12. NCCLS. Evaluation of Precision Performance of Clinical Chemistry Devices, EP5-A; 2001, Wayne, Pennsylvania.
13. NCCLS. Preliminary Evaluation of Quantitative Clinical Laboratory Methods, EP6-A; 1998, Wayne, Pennsylvania.
14. Satterthwaite, F.E. An approximate distribution of estimates of variance components. *Biometrics* **1946**, *2*, 110–114.
15. NCCLS. Interference Testing in Clinical Chemistry, EP7-A; 2002, Wayne, Pennsylvania.
16. NCCLS. Evaluation of the Linearity of Quantitative Analytical Methods, EP6-P2; 2001, Wayne, Pennsylvania.
17. Spencer, C.A.; Takeuchi, M.; Kazarosyan, M.; MacKenzie, F.; Beckett, G.; Wilkinson, E. Interlaboratory/intermethod differences in functional sensitivity of immunometric assays of thyrotropin (TSH) and impact on reliability of measurement of subnormal concentrations of TSH. *Clin. Chem.* **1995**, *41*, 367–374.
18. Hay, I.D.; Bayer, M.F.; Kaplan, M.M.; Klee, G.G.; Larsen, P.R.; Spencer, C.A. American Thyrotropin measurements and guidelines for future clinical assays. *Clin. Chem.* **1991**, *37*, 2002–2008.
19. Brenner, H.; Gefeller, O. Variation of sensitivity, specificity, likelihood rates and predictive values with disease prevalence. *Stat. Med.* **1997**, *16*, 981–991.
20. NCCLS. Assessment of the Clinical Accuracy of Laboratory Tests Using Operating Characteristic (ROC) Plots, GP10-A; **1995**, Wayne, Pennsylvania.
21. Zweig, M.H.; Campbell, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* **1993**, *39*, 561–577.
22. Zhou, X.H.; McClish, D.K.; Obuchowski, N.A. *Statistical Methods in Diagnostic Medicine*; Wiley: New York, 2002.
23. NCCLS. User protocol for evaluation of qualitative test performance, EP12-A; 2002, Wayne, PA.
24. Alonzo, T.; Pepe, M. Using a combination of reference test to access the accuracy of a new diagnostic test. *Stat. Med.* **1999**, *18*, 2987–3003.
25. Hadgu, A. The discrepancy in discrepancy analysis. *Lancet* **1996**, *348*, 592–593.
26. Becker, S. Evaluation a new test using a reference test with estimated sensitivity and specificity. *Commun. Stat.* **1991**, *20*, 2739–2752.
27. Hawkins, D.; Garrett, J.; Stephen, B. Some issues in resolution of diagnostic tests using an imperfect gold standard. *Stat. Med.* **2001**, *20*, 1987–2001.
28. Shao, J.; Zhong, B. Assessing the agreement between two quantitative assays with repeated measurements. *J. Biopharm. Stat.* **2004**, *14*, 201–212.

Weibull Model: Group Sequential Design

Jianrong Wu

Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee, U.S.A.

Abstract

A multistage group sequential procedure is presented under the Weibull model by applying sequential conditional probability ratio test methodology.

Keywords: Brownian motion; Group sequential trial; Randomized clinical trial; Sample size; Time to event; Weibull distribution.

INTRODUCTION

In cancer clinical trials, time to event is often a primary end point for the study design. The primary interest is to compare the survival distributions between treatment groups. For ethical reasons, randomized phase III clinical trials are often monitored for early stopping if a sufficiently large treatment difference is observed during an interim analysis. Various group sequential monitoring methods have been developed in the past few decades; e.g., the type I error spending function approach by Lan and DeMets,^[1] the triangular test by Whitehead and Stratton,^[2] and the sequential conditional probability ratio test (SCPRT) by Xiong.^[3] Comprehensive reviews of these methods are provided by Jennison and Turnbull.^[4]

For survival data, the Weibull distribution is the most frequently used parametric model. In general, a survival trial under the Weibull model can also be designed under the proportional hazards model using the non-parametric log-rank test. However, a parametric test derived under the Weibull model has better small sample properties than the non-parametric log-rank test. The maximum sample size is often large for a phase III group sequential trial. However, the available data could be small in the early stages of interim monitoring; therefore, a group sequential design using a parametric test may perform better than the non-parametric log-rank test. Tsiatis, Boucher, and Kim^[5] derived asymptotic sequential distributions for score and Wald tests in general parametric survival models. Applying results derived by Tsiatis, Boucher, and Kim, Wu and Xiong^[6] proposed a multistage group sequential design under the Weibull model.

SEQUENTIAL TEST STATISTICS

A parametric sequential test statistic is discussed in this section to provide group sequential design for randomized

two-arm survival trials under the Weibull model. Assume that time-to-event variable T_j of a subject from the j^{th} group follows the Weibull distribution with a common shape parameter κ and a scale parameter ρ_j , where $j = 1, 2$. That is, T_j has a survival distribution function,

$$S_j(t) = e^{-(\rho_j t)^\kappa} \quad (1)$$

and a hazard function,

$$h_j(t) = \kappa \rho_j^\kappa t^{\kappa-1} \quad (2)$$

The shape parameter κ indicates the degree of acceleration ($\kappa > 1$) or deceleration ($\kappa < 1$) of the hazard over time. In a cancer trial, the median survival time is an intuitive end point for clinicians. The median survival time of the j^{th} group for the Weibull distribution can be calculated as $m_j = \rho_j^{-1} \{\log(2)\}^{1/\kappa}$. Therefore, the Weibull survival distribution can be expressed as:

$$S_j(t) = e^{-\log(2) \left(\frac{t}{m_j}\right)^\kappa}, \quad j = 1, 2 \quad (3)$$

The one-sided hypothesis of a randomized two-arm trial defined by median survival times can be expressed as:

$$H_0 : m_1 \geq m_2 \quad \text{vs.} \quad H_1 : m_1 < m_2 \quad (4)$$

For notational convenience, we convert the scale parameter ρ_j to a hazard parameter $\lambda_j = \rho_j^\kappa = \log(2)/m_j^\kappa$. Then, the above hypothesis on median survival times is equivalent to the following:

$$H_0^* : \delta \leq 1 \quad \text{vs.} \quad H_1^* : \delta > 1 \quad (5)$$

where $\delta = \lambda_2/\lambda_1 = (m_1/m_2)^\kappa$ is the hazard ratio. The survival distribution is $S_j(t) = e^{-\lambda_j t^\kappa}$ and hazard function is $h_j(t) = \kappa \lambda_j t^{\kappa-1}$ for j^{th} group, in which κ is taken as a

known constant. This indicates that the testing problem can also fit into a proportional hazard model and the log-rank test is applicable for the intended testing.

Now, suppose that during the accrual phase of the trial, n_j subjects of the j^{th} group are enrolled in the study, and let T_{ij} and C_{ij} denote, respectively, the event time and censoring time of the i^{th} subject of the j^{th} group, with both being measured from the time of study entry, Y_{ij} . We assume that the event time T_{ij} is independent of the censoring time C_{ij} and entry time Y_{ij} , and $\{(Y_{ij}, T_{ij}, C_{ij}); i = 1, \dots, n_j, j = 1, 2\}$ are independent and identically distributed within each group. When the data are examined at calendar time $t \leq \tau$, where τ is the study duration, we observe the time-to-event $X_{ij}(t) = T_{ij} \wedge C_{ij} \wedge (t - Y_{ij})^+$ and failure (or event) indicator $\Delta_{ij}(t) = I(T_{ij} \leq C_{ij} \wedge (t - Y_{ij})^+)$, where $i = 1, \dots, n_j$ and $j = 1, 2$. Based on the observed data $L(\lambda_1, \lambda_2; t) = \lambda_1^{d_1(t)} \lambda_2^{d_2(t)} e^{-\lambda_1 U_1(t) - \lambda_2 U_2(t)}$, the observed likelihood function at time t is proportional to:

$$L(\lambda_1, \lambda_2; t) = \lambda_1^{d_1(t)} \lambda_2^{d_2(t)} e^{-\lambda_1 U_1(t) - \lambda_2 U_2(t)} \quad (6)$$

where $d_j(t) = \sum_{i=1}^{n_j} \Delta_{ij}(t)$ is the total number of events observed in the j^{th} group by time t , and $U_j(t) = \sum_{i=1}^{n_j} X_{ij}(t)$ is the cumulative follow-up time by time t penalized by the Weibull shape parameter κ .^[7] The maximum likelihood estimate of $\lambda_j(t)$ can be derived as:

$$\hat{\lambda}_j(t) = d_j(t) / U_j(t) \quad (7)$$

and its variance is approximately $\hat{\lambda}_j^2(t) / d_j(t)$. Therefore, under the null hypothesis, the Wald statistic of the log-hazard ratio $\gamma = \log(\delta)$ at calendar time t is given by:

$$Z(t) = \log\{U_2(t)d_1(t)/U_1(t)d_2(t)\}(d_1^{-1}(t) + d_2^{-1}(t))^{-1/2} \quad (8)$$

and has approximately a standard normal distribution.^[6] To derive the group sequential design, let

$$U(t) = \log\{U_2(t)d_1(t)/U_1(t)d_2(t)\}(d_1^{-1}(t) + d_2^{-1}(t))^{-1} \quad (9)$$

then under the alternative, the statistic $U(t)$ is approximately normal with mean $\log(\delta)V(t)$ and variance $V(t)$ and has an independent increment structure, where $V(t) = (d_1^{-1}(t) + d_2^{-1}(t))^{-1}$. The above results can be obtained from Tsiatis, Boucher, and Kim^[5] who proved the results for general survival parametric models. Since,

$$n_1^{-1}V(t) \rightarrow D(t) = (p_1^{-1}(t) + \pi^{-1}p_2^{-1}(t))^{-1} \quad (10)$$

where $p_j(t) = P(\Delta_{ij}(t) = 1)$ and $\pi = n_2/n_1$ is the treatment allocation ratio. Thus, $B(t^*) = n_1^{1/2}U(t)/D^{1/2}(t) \sim N(\theta t^*, t^*)$ is approximately a Brownian motion with drift parameter $\theta = n_1^{1/2} \log(\delta)D^{1/2}(\tau)$ and information time $t^* = D(t)/D(\tau)$, where $D(\tau)$ is the value of $D(t)$ at $t = \tau$.

SAMPLE SIZE FOR FIXED SAMPLE TEST

The sample size for a fixed sample test is calculated for the situation at the end of the study. Based on test statistic $Z(t)$ at $t = \tau$, under the null hypothesis,

$$Z(\tau) = \log\{U_2(\tau)d_1(\tau)/U_1(\tau)d_2(\tau)\}(d_1^{-1}(\tau) + d_2^{-1}(\tau))^{-1/2} \quad (11)$$

has an approximate standard normal distribution. To calculate the power, let $p_j(\tau)$ be the probability of a subject from the j^{th} group having an event during the study. Then under the alternative $\delta = \lambda_1/\lambda_2 > 1$, $Z(\tau)$ is approximately normal distributed with mean $n_1^{1/2} \log(\delta)D^{1/2}(\tau)$ and unit variance. Therefore, given a significance level α , the power $(1 - \beta)$ of the $Z(\tau)$ test under the alternative is given by:

$$1 - \beta = \Phi\left\{n_1^{1/2} \log(\delta)(p_1^{-1}(\tau) + \pi^{-1}p_2^{-1}(\tau))^{-1/2} - z_{1-\alpha}\right\} \quad (12)$$

where $\Phi(\cdot)$ is the standard normal distribution function and $z_{1-\alpha} = \Phi^{-1}(1 - \alpha)$. Thus, based on the $Z(\tau)$ test, the sample size of the first group can be calculated as:

$$n_1 = \frac{(z_{1-\alpha} + z_{1-\beta})^2(p_1^{-1}(\tau) + \pi^{-1}p_2^{-1}(\tau))}{[\log(\delta)]^2} \quad (13)$$

where $\delta = R^\kappa$. Therefore, the total sample size for the two groups is given by:

$$n = \frac{(\pi + 1)(z_{1-\alpha} + z_{1-\beta})^2(p_1^{-1}(\tau) + \pi^{-1}p_2^{-1}(\tau))}{[\log(\delta)]^2} \quad (14)$$

To calculate the number of subjects required for the study, we need to calculate $p_j(\tau)$, the probability of a subject in the j^{th} group having an event during the study. Typically, we assume that subjects are accrued over an accrual period of length t_a with an additional follow-up period of length t_r . A subject enters the study at time u , the entry time is uniformly distributed on $[0, t_a]$, and no subject is lost to follow-up during the study. Then the probability of a subject having an event during the study under the Weibull model can be calculated by:

$$p_j(\tau) = 1 - \frac{1}{t_a} \int_{t_r}^{t_a+t_r} e^{-\log(2)\left(\frac{u}{m_j}\right)^\kappa} du \quad (15)$$

Therefore, given the design parameters $\kappa, m_1, m_2, \alpha, \beta, \pi, t_r$ and t_a , the number of subjects n required for the study can be calculated using equation 14.

When designing an actual trial with accrual time t_a , calculating the sample size is often impractical, because we may not be able to enroll the planned number of subjects within the given accrual duration. It is more practical to design the study by starting with the given accrual rate r and then calculate the required accrual time t_a . This can

be accomplished under the Weibull model assumption. We can define a root function of the accrual time t_a ,

$$\text{root}(t_a) = \text{rt}_a - \frac{(\pi + 1)(z_{1-\alpha} + z_{1-\beta})^2 (p_1^{-1}(t_a + t_f) + \pi^{-1}p_2^{-1}(t_a + t_f))}{[\log(\delta)]^2} \quad (16)$$

The accrual time t_a can now be obtained by solving the root equation $\text{root}(t_a) = 0$ numerically in R using the “uni-root” function. The total sample size required for the study is approximately $n = [\text{rt}_a]^+$, where $[x]^+$ denotes the smallest integer greater than x .

GROUP SEQUENTIAL PROCEDURE

In this section, we will apply an SCPRT procedure to the test statistic $Z(t)$. The SCPRT has two unique features: 1) the maximum sample size of the sequential test is not greater than the size of the reference fixed sample test and 2) the probability of discordance, or the probability that the conclusion of the sequential test would be reversed if the experiment was not stopped according to the stopping rule but continued to the planned end, can be controlled to an arbitrarily small level.^[8] Furthermore, the power function of the SCPRT is virtually the same as that of the fixed sample test.^[3] The SCPRT boundaries derived in this entry have analytical solutions. All these features make the SCPRT attractive and simple to use.

Let $\{B_{t^*} : 0 < t^* \leq 1\}$ be the Brownian motion $B_{t^*} \sim N(\theta t^*, t^*)$ and B_1 be the B_{t^*} at the final stage with full information $t^* = 1$. Then the joint distribution of (B_{t^*}, B_1) has a bivariate distribution with mean $\mu = (\theta t^*, \theta)$ and a variance matrix $\Sigma = (\sigma_{ij})_{2 \times 2}$ with $\sigma_{11} = \sigma_{12} = \sigma_{21} = t^*$ and $\sigma_{22} = 1$. Therefore, according to multivariate conditional distribution theory,^[9] the conditional density $f(B_{t^*} | B_1)$ is the normal density of $N(B_1 t^*, (1 - t^*) t^*)$. Let $s_0 = z_{1-\alpha}$ be the critical value of B_1 to reject the null for the fixed sample test. Then the conditional maximum likelihood ratio for the stochastic process on information time t^* is:^[3,10]

$$L(t^*, B_{t^*} | z_{1-\alpha}) = \frac{\max_{\{s > s_0\}} f(B_{t^*} | B_1 = s)}{\max_{\{s \leq s_0\}} f(B_{t^*} | B_1 = s)} \quad (17)$$

Taking the logarithm, the log-likelihood ratio can be simplified as:

$$\log\{L(t^*, B_{t^*} | z_{1-\alpha})\} = \pm \frac{(B_{t^*} - z_{1-\alpha} t^*)^2}{2(1 - t^*) t^*} \quad (18)$$

which has a positive sign if $B_{t^*} > z_{1-\alpha} t^*$ and a negative sign if $B_{t^*} < z_{1-\alpha} t^*$. This equation leads to lower and upper boundaries for $B_{t_k^*}$ as:

$$\begin{aligned} a_k &= z_{1-\alpha} t_k^* - \{2at_k^*(1 - t_k^*)\}^{1/2}; \\ b_k &= z_{1-\alpha} t_k^* + \{2at_k^*(1 - t_k^*)\}^{1/2}, \end{aligned} \quad (19)$$

for $k = 1, \dots, K$, where K is the total number of looks, and $t_1^*, t_2^*, \dots, t_K^* (= 1)$ are the information times of the interim looks and the final look. The a in Eq. 19 is the boundary coefficient, and it is crucial to choose an appropriate a for the design such that the probability of conclusion by the sequential test being reversed by the test at the planned end is small but not unnecessarily small. The larger a is, the smaller is the discordance probability, and the wider apart the upper and lower boundaries are, making it harder for the sample path to reach the boundaries and stop early and resulting in larger expected sample sizes. Thus, an appropriate a can be determined by choosing an appropriate discordance probability.^[3,10]

We now apply the SCPRT to the test statistic $B(t^*) = \eta_1^{-1/2} U(t)/D^{1/2}(\tau) \sim N(\theta t^*, t^*)$, which is a Brownian motion in information time $t^* = D(t)/D(\tau)$ on $[0, 1]$, and the drift parameter $\theta = \eta_1^{1/2} \log(\delta) D^{1/2}(\tau)$, where $D(t)$ is defined by Eq. 10 in the section “Sequential Test Statistics.” Suppose k^{th} interim looks are planned at calendar time $t_k^*, k = 1, \dots, K$. Then based on the SCPRT procedure presented above, the lower and upper boundaries for $B(t_k^*)$ at the k^{th} look are given by:

$$\begin{aligned} a_k &= z_{1-\alpha} t_k^* - \{2at_k^*(1 - t_k^*)\}^{1/2}; \\ b_k &= z_{1-\alpha} t_k^* + \{2at_k^*(1 - t_k^*)\}^{1/2} \end{aligned} \quad (20)$$

for $k = 1, \dots, K$, where $t_k^* = D(t_k)/D(\tau)$ is the information time at the k^{th} look at calendar time t_k . The nominal critical p values for testing H_0 are:

$$P_{a_k} = 1 - \Phi(a_k / \sqrt{t_k^*}); \quad P_{b_k} = 1 - \Phi(b_k / \sqrt{t_k^*}) \quad (21)$$

The observed p value at the k^{th} look is:

$$P_{Z(t_k)} = 1 - \Phi(Z(t_k)) \quad (22)$$

The stopping rule for monitoring the trial can be executed by stopping the trial when, for the first time, $P_{Z(t_k)} \geq P_{a_k}$ (accept H_0 and stop for futility) or $P_{Z(t_k)} \leq P_{b_k}$ (reject H_0 and stop for efficacy).

EXAMPLE

Rhabdoid tumors are aggressive pediatric malignancies with a poor prognosis. Over the past years, St. Jude Children’s Research Hospital accrued 14 pediatric patients with recurrent or refractory non-central nervous system (CNS) rhabdoid tumors treated with conventional chemotherapy. The median event-free survival is only about 1 year, where the event is defined as disease relapse or death. All 14 patients had events within about 3 years. The Weibull model was fitted in R to the data, resulting in an estimate (standard error) of the shape parameter $\kappa = 1.37(0.28)$ and median event-free survival time of $m_1 = 0.936$ years, which provides a more satisfactory

model than the exponential model.^[11] Now, suppose that we would like to design a multicenter randomized two-arm trial to assess the effectiveness of the small molecule inhibitor alisertib (treatment 2) versus conventional chemotherapy (treatment 1) for this group of patients. Patients will be randomized with equal allocation to each treatment group and hence $\pi = n_2/n_1 = 1$. The hypotheses of the planned study are $H_0: m_2 \leq m_1$ vs. $H_1: m_2 > m_1$. The investigators would like to detect a half-year increase in the median event-free survival of the alisertib treatment group over that of the conventional chemotherapy group, with 90% power, 5% type I error, and 2 years of follow-up ($t_f = 2$) after the last patient is enrolled in the study. Then, for the alternative hypothesis, $m_2 = m_1 + 0.5 = 1.436$, $\delta = \lambda_1/\lambda_2 = (m_2/m_1)^x = 1.797$, and $\gamma = \log(\delta) = 0.586$. Assume this multicenter trial has the capacity to enroll and treat 20 patients per year. Then, under the assumption of the Weibull model, with uniform entry and no loss to follow-up, the required total accrual time is $t_a = 5.3$ years, calculated by Eq. 16, in which $p_1(\tau)$ and $p_2(\tau)$ are found by Eq. 15 with $\tau = t_a + t_f$. Then, the study duration is $\tau = t_a + t_f = 7.3$ years, and the total sample size is 106 patients (53 per group). Now assuming interim and final looks are planned at calendar times $t_1 = 4$, $t_2 = 5$, and $t_3 = \tau = 7.3$ years, the corresponding information time at each planned interim look t_k can be calculated by $t_k^* = D(t_k)/D(\tau)$, where $D(t) = (p_1^{-1}(t) + \pi^{-1}p_2^{-1}(t))^{-1}$ with:

$$p_j(t) = \frac{1}{t_a} \int_0^{t \wedge t_a} \{1 - S_j(t-u)\} du \quad (23)$$

$S_j(t) = e^{-\log(2)\left(\frac{t}{m_j}\right)^x}$, and $\pi = 1$. By calculation, the corresponding information times are $t_1^* = 0.511$, $t_2^* = 0.706$, and $t_3^* = 1$. Assuming the maximum conditional probability of discordance $\rho = 0.02$, then the boundary coefficient $a = 2.593$ for $K = 3$,^[3] and the maximum probability of discordance is $\rho_{\max} = 0.0043$. That is, under the most unfavorable setting of mean parameter, i.e., the underlying true drift θt^* of Brownian motion $B(t^*)$ is pointing to the cut-off point $z_{1-\alpha}$ of the test at the final stage, on average, for every 1,000 such sequential tests, there would be only 4.3 cases of later reversal of conclusion (significant or futility) at early stopping if the sequential test was not stopped as it should be but continued till the planned end. The lower and upper boundaries calculated from Eq. 20 are $(a_1, a_2, a_3) = (-0.298, 0.124, 1.645)$ and $(b_1, b_2, b_3) = (1.979, 2.199, 1.645)$, respectively. The acceptance and rejection nominal critical significance levels are (0.6615, 0.4415, 0.050) and (0.0028, 0.0044, 0.050), respectively.

CONCLUSION

A multistage group sequential procedure is presented under the Weibull model based on the SCPRT procedure. The advantages of the proposed sequential procedure are: 1) the maximum sample size of the sequential test is the same as that of the reference fixed sample test; and 2) the probability of discordance, or the probability that the conclusion of the sequential test would be reversed if the experiment was not stopped according to the stopping rule but continued to the planned end, can be controlled to an arbitrarily small level.

ACKNOWLEDGMENTS

The work was supported in part by the National Cancer Institute support grant P30CA021765 and ALSAC. This entry previously appeared as "Power and sample size for randomized phase III survival trials under the Weibull model" in the Journal of Biopharmaceutical Statistics volume 25.^[11] The author acknowledges the permission granted by Taylor & Francis Group for this republication.

REFERENCES

1. Lan, K.K.G.; DeMets, D.L. Discrete sequential boundaries for clinical trials. *Biometrika* **1983**, *70*, 659–663.
2. Whitehead, J.; Stratton, I. Group sequential clinical trial with triangular continuation regions. *Biometrics* **1983**, *39*, 227–236.
3. Xiong, X. A class of sequential conditional probability ratio tests. *J. Am. Stat. Assoc.* **1995**, *15*, 1463–1473.
4. Jennison, C.; Turnbull, B.W. *Group Sequential Methods with Applications to Clinical Trials*; Chapman and Hall: New York, 2000.
5. Tsiatis, A.A.; Boucher, H.; Kim, K. Sequential methods for parametric survival models. *Biometrika* **1995**, *70*, 165–173.
6. Wu, J.; Xiong, X. Group sequential design for randomized phase III trials under the Weibull model. *J. Biopharm. Stat.* **2015**, *25*, 1190–1205.
7. Cox, D.R.; Oakes, D.V. *Analysis of Survival Data*; Chapman and Hall: London, 1984.
8. Xiong, X.; Tan, M.; Boyett, J. A sequential procedure for monitoring clinical trials against historical controls. *Stat. Med.* **2007**, *26*, 1497–1511.
9. Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*; Wiley: New York, 1958.
10. Xiong, X.; Tan, M.; Boyett, J. Sequential conditional probability ratio tests for normalized test statistic on information time. *Biometrics* **2003**, *59*, 624–631.
11. Wu, J. Power and sample size for randomized phase III survival trials under the Weibull model. *J. Biopharm. Stat.* **2015**, *25*, 16–28.

Z-Score

Jiang-Ming Johnny Wu

DOV Pharmaceutical, Inc., Somerset, New Jersey, U.S.A.

Abstract

This entry discusses topics related to the z-score, including Gaussian distribution, the standard normal distribution, and the central limit theorem, and provides applications for the z-score.

Keywords: Central limit theorem; Gaussian distribution; Standard normal distribution; Standard score; Standardization.

INTRODUCTION

The z-score is the most commonly used standard score.^[1] It is derived from the standard normal distribution with mean value 0 and standard deviation 1. Therefore, a z-score of an observation represents how far a value deviates from the mean value in terms of standardized or normalized units, more commonly called its standard deviation.^[2] The z-score is commonly used in social science studies, for example, in educational or psychological testing. With a calculated z-score, two different test scores can be compared. However, use of the z-score is not restricted to social studies. Its application can be applied to most areas of scientific study. One such example is a clinical trial for schizophrenia research.^[3]

After a brief introduction, we will discuss topics related to the z-score, including the Gaussian distribution, the standard normal distribution, and the central limit theorem. Finally, the z-score will be formally discussed and its application will be presented.

GAUSSIAN DISTRIBUTION

Another popular name for the Gaussian distribution is the *normal distribution*. It has many other names as well, such as the Laplacian, the Gauss–Laplace or the Laplace–Gauss distribution, and the second law of Laplace.^[2] It is a continuous distribution with a probability density function (pdf) expressed by the formula:

$$f(x) = 1/[\sigma(2\pi)^{1/2}] * e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty \leq x \leq \infty, \quad (1)$$

where μ is the mean and σ is the standard deviation, with $-\infty < \mu < \infty$, and $\sigma > 0$.

The normal distribution is symmetric with its center at the mean that is also the mode and the median. The normal distribution is actually a family of distributions in that

there are an infinite number of normal distributions with a unique correspondence to the (μ, σ) combination. Given a value for the (μ, σ) combination one can obtain the corresponding normal distribution. In other words, the normal distribution is completely characterized and determined by these two parameters, μ and σ . In many standard mathematical statistics text books^[4–6] a random variable X that has a normal distribution with mean μ and standard deviation σ is thus denoted by $N(\mu, \sigma^2)$. We use the notation $X \sim N(\mu, \sigma^2)$ to represent this situation. A general diagram for the normal distribution is presented in Fig. 1.

An important property of the normal distribution is that any linear transformation of a normally distributed random variable is still normally distributed. This property is summarized in the following theorem.

Theorem. If $X \sim N(\mu, \sigma^2)$ and a and b are constants then $a + bX \sim N(a + b\mu, b^2\sigma^2)$.

A reader interested in the proof for this theorem may consult the textbooks mentioned above or any standard mathematical statistics textbooks.

STANDARD NORMAL DISTRIBUTION

If $X \sim N(\mu, \sigma^2)$, then when $\mu = 0$ and $\sigma = 1$, the corresponding normal distribution is called the *standard normal distribution*, denoted by $N(0, 1)$. The pdf, $f(x)$, of a standard normal distribution is given by the following expression,

$$f(x) = 1/(2\pi)^{1/2} * e^{-x^2/2}, \quad -\infty \leq x \leq \infty, \quad (2)$$

and its corresponding cumulative distribution function (cdf), $F(x)$, is given by

$$F(x) = \frac{1}{2\pi^{1/2}} \int_{-\infty}^x e^{-y^2/2} dy; \quad (3)$$

this is often denoted by $\Phi(x)$.

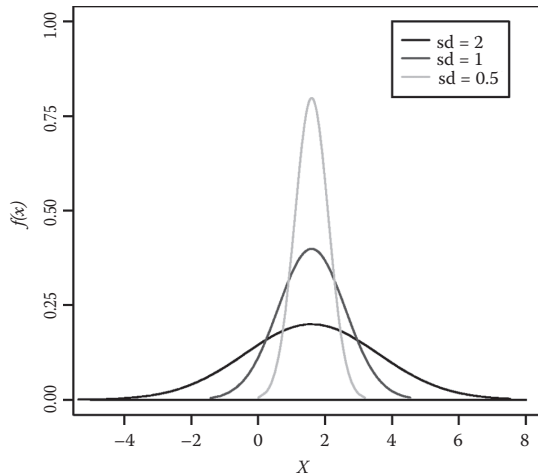


Fig. 1 Normal distribution

Some of the basic properties of $f(x)$ can be found in standard statistics handbook,^[7] for example:

$$f(-x) = f(x) \quad \text{and} \quad F(-x) = 1 - F(x). \quad (4)$$

$f'(x)$ is the first derivative of $f(x)$ with respect to x :

$$f' = \frac{-x}{2\pi^{1/2}} e^{-x^2/2} = -xf(x). \quad (5)$$

$f''(x)$ is the second derivative of $f(x)$ with respect to x :

$$f'' = \frac{x^2 - 1}{2\pi^{1/2}} e^{-x^2/2} = x^2 - 1f(x). \quad (6)$$

$f'''(x)$ is the third derivative of $f(x)$ with respect to x :

$$f''' = \frac{3x - x^3}{2\pi^{1/2}} e^{-x^2/2} = (3x - x^3)f(x). \quad (7)$$

$f^{iv}(x)$ is the fourth derivative of $f(x)$ with respect to x :

$$f^{iv} = \frac{x^4 - 6x^2 + 3}{2\pi^{1/2}} e^{-x^2/2} = (x^4 - 6x^2 + 3)f(x). \quad (8)$$

A table that contains all of the above related functions of $f(x)$ can also be found in the reference work edited by Beyer.^[7] Other forms of the standard normal tables can be found in the books by Conover^[8] and Panago.^[9]

The standard normal distribution has many applications because of the famous central limit theorem that can be found in most standard mathematical statistics books.^[4-6]

Central Limit Theorem. If $X_1, X_2, \dots, X_n$ is a random sample drawn from a distribution that has mean μ and positive variance σ^2 . Then the random variable $Y_n = n^{1/2}(\bar{X} - \mu)/\sigma$ has a limiting distribution, as n approaches infinity, that is normal with mean 0 and variance 1. Here, $\bar{X}$ is the sample mean of the random sample, i.e., $\bar{X} = \sum_{i=1,n} X_i/n$.

Simply stated in mathematical terms, the central limit theorem asserts that the random variable Y_n defined above converges to z , where $z \sim N(0, 1)$, in distribution; or equivalently, the sample mean, $\bar{X}$, converges to the population mean, μ , in distribution.

There are many applications to the central limit theorem because the population distribution for the random variable X is not specified. When random variables have a finite mean and a positive variance, then the central limit theorem applies. For example, if X has a Bernoulli distribution with parameter p , then the central limit theorem becomes the normal approximation to the binomial distribution that can be found in many standard mathematical statistics textbooks.

Another application of the standard normal distribution is for Monte Carlo studies where the standard normal distribution is used to generate any arbitrary normally distributed random variable with a specified mean μ and variance σ^2 . In SAS® programming language^[10] there is a random number generator function, RANNOR(seed), and a corresponding CALL routine, CALL RANNOR(seed, x), that can generate random numbers from a standard normal distribution.

Using the theorem presented in Section II with $\mu = 0$ and $\sigma = 1$, and selecting $a = \mu$ and $b = \sigma$ then that theorem becomes the following:

Lemma. If $X \sim N(0, 1)$ then $\mu + \sigma X \sim N(\mu, \sigma^2)$.

One can easily generate a set of random samples that is normally distributed with mean μ and variance σ^2 using this lemma and the random number generator function RANNOR(seed). Such an example has been illustrated in Fan et al.^[11]

DEFINITION OF Z-SCORE

As mentioned earlier, the z -score describes how far a sample value has deviated from its mean value in terms of its standard deviation units. A formal definition of the z -score, presented by Pagano,^[9] states that:

A z -score is a transformed score that designates how many standard deviation units the corresponding raw score is above or below the mean.

In terms of the standard normal distribution, the z -score is indeed the abscissa under the standard normal curve. Knowing the z -score for a sample value, one can use the standard normal table (e.g., Table A from Pagano^[9]) to tell the relative standing, in terms of percentage, of this value of the whole sample. For example, a z -score of 1.645 (using linear interpolations) means that this value is in the 95th percentile of the whole sample or that 95% of the values in this sample are less than or equal to this value. To obtain the z -score, say z , of a sample value, say x , from a normally distributed random sample with mean μ and variance σ^2 ,

simply subtract the mean value from x and then divide by the standard deviation. That is,

$$z = (x - \mu) / \sigma. \quad (9)$$

The process of using Equation 9 to find the z -score of a sample value in a normally distributed random sample is often referred to as the process of standardization. This process allows two different values, coming from different normal populations, to be compared to each other, including those with different units for each distribution. This is accomplished using the standard normal table and their corresponding z -scores obtained in this manner. To illustrate, consider the following example. You want to compare your height with your weight to see if you are relatively taller or relatively heavier than others in a particular group. Assume that your height and weight measurements are 76 in and 180 lb, respectively, that the height measurement in that particular group follows a normal distribution with mean = 68 in, and SD = 4 in, and that the weight measurement in that group follows another normal distribution with mean = 160 lb, and SD = 20 lb. Given this information, we can easily calculate the z -scores for both your height and weight measurements using Eq. 9. The z -score for your height is $(76-68)/4 = 2$, and the z -score for your weight is $(180-160)/20 = 1$. Therefore, using the standard normal table, your height ranks 97.72% ($z = 2$) relative to that particular group, although your weight only ranks 84.13% ($z = 1$) relative to the same group. Your height has a much higher ranking than your weight compared to that particular group. Consequently, you know you are much taller and heavier than that group's average.

APPLICATIONS

The application of a z -score is many fold. Because the z -score is tightly related to the standard normal distribution, any application of the latter can be seen as an application of the former. For example, in elementary statistical textbooks the hypothesis testing procedure using the normal distribution (sometimes referred to as the z -test) is an application of the standard normal distribution. The test statistic, z , in such a procedure is indeed a z -score. Therefore, the hypothesis testing procedure is also an application of the z -score. To extend this logic, any transformation that results in a standard normal distribution as the limiting distribution may be viewed as an application of the z -score. This is because such a transformation eventually relies on the z -score to infer whether it is for hypothesis testing or for an estimation of the confidence limits or bounds.

We have previously illustrated by example how to use the z -score to compare two different things with different units. By transforming different item scores into z -scores, one can easily compare these item scores through their z -scores. This procedure of standardization allows you to

“compare apples with oranges.” Without the use of z -scores such a comparison is not possible.

The central limit theorem mentioned earlier is a powerful tool that can transform a random variable, derived from a random sample, so that it has a limiting standard normal distribution. Therefore, as mentioned above, this will also be an application of the z -score.

Also mentioned in the previous section is use of the standard normal distribution to generate a random sample from any normal distribution in Monte Carlo studies. In a broad sense, this can also be treated as an application of the z -score.

In an example from a clinical trial that uses the z -score as a way of evaluating efficacy variables, Keefe et al.^[3] studied the effect of BACS (the Brief Assessment of Cognition in Schizophrenia) in patients with schizophrenia. The BACS is a measurement that consists of six domains for cognitive function: verbal memory, working memory, motor speed, attention, executive functions, and verbal fluency. It is found to be fully portable, and is designed to be easily administered by a variety of testers, including nurse clinicians, psychiatrists, neurologists, social workers, and other mental health workers. There are two groups in this study: treatment group and healthy control group. The allocation of patients in these two groups is 3:1 such that with every three patients in the treatment group there will be 1 healthy subject in the control group. The total number of subjects is 200, with 150 in the treatment group and 50 in the control group.

Data analysis for this study has a primary measure from each test (domain) of the BACS. Each measure was standardized by creating z -scores whereby, for test session 1, the healthy control mean was set to 0 and standard deviation set to 1. A composite score was calculated by averaging all of the six standard primary measures from the BACS, and then calculating a z -score of the composite. The relative sensitivity of the composite scores derived from the standard battery and the BACS was determined by comparing each of their between-group differences with t -tests. Other details of this study can be found in Keefe et al.^[3]

This is an example, published in the literature, that used the z -score of the efficacy variable as the primary measurement for data analysis and it serves as a good application of the z -score.

ARTICLES OF FURTHER INTEREST

Confidence Interval and Hypothesis Testing; Hypothesis Testing; Power; P-Values; Statistical Significance.

REFERENCES

1. Vogt, W.P. *Dictionary of Statistics and Methodology*; Sage: Newbury Park, CA, 1993.

2. Marriott, F.H.C. *A Dictionary of Statistical Terms*; Wiley: New York, 1990.
3. Keefe, R.S.E.; Goldberg, T.E.; Harvey, P.D. The brief assessment of cognition in schizophrenia: reliability, sensitivity, and comparison with a standard neurocognitive battery. *Schiz. Res.* **2004**, *68*, 283–297.
4. Hogg, R.V.; Craig, A.T. *Introduction to Mathematical Statistics*; Macmillan Publishing Co: New York, 1986.
5. DeGroot, M.H. *Probability and Statistics*; Yeh-Yeh: Taipei, Taiwan, 1985.
6. Mood, A.M.; Graybill, F.A.; Boes, D.C. *Introduction to the Theory of Statistics*; McGraw-Hill: New York, 1974.
7. *CRC Standard Probability and Statistics Tables and Formulae*; Beyer, W.H., Ed.; CRC Press: Boca Raton, FL, 1991.
8. Conover, W.J. *Practical Nonparametric Statistics*; Wiley: New York, 1980.
9. Pagano, R.R. *Understanding Statistics in the Behavioral Sciences*; West Publishing Co: St. Paul, MN, 1990.
10. SAS Institute Inc. *SAS® 9.1 Language Reference: Dictionary, Vol. 1*; SAS Institute Inc: Cary, NC, 2004.
11. Fan, X.; Felsővályi, Á.; Sivo, S.A.; Keenan, S.C. *SAS® for Monte Carlo Studies-A Guide for Quantitative Researchers*; SAS Institute Inc: Cary, NC, 2002.

Index

A

- Aalen model, 2058
- AB/BA crossover design, 785, 1614
- advantages of, 785–786
 - analysis of covariance in, 786
 - distribution-free random effects model
 - with using a conditional approach, 1615–1617
 - example, 1617
 - normal random effects proportional odds model, 1614
 - simple change score method, 1615
 - uses of crossover designs, 786
- abbreviated new drug application (ANDA), 304, 928, 970, 2199
- ab initio* methods, 229
- absolute bioavailability studies, 263
- clinical pharmacology, 470
- absolute neutrophil count (ANC), 191
- absorption, distribution, metabolism, and excretion (ADME) criteria, 308, 1095
- abuse-deterrence effect, 1121, 1122
- abuse-deterrent opioids, 1121, 1122, 1123
- Academy for Managed Care Pharmacy (AMCP), 1694
- accelerated factor, 558
- Accelerated Life Model (ALM), 2064
- accelerated titration designs (ATDs), 387
- Acceptable Quality Level (AQL), 1, 462–463
- acceptance criteria, 1, 4, 260, 1808
- finished product, 1811
 - content uniformity/dissolution, 1811
 - potency, 1811–1812
 - validation issues, 1812
 - powder blends, 1810
 - average potency, 1811
 - blend uniformity, 1810
- acceptance sampling, 1–6
- ANSI/ASQC sampling procedures, 6
 - common misconceptions, 5–6
 - double sampling, 3–4
 - graphs, 2–3
 - plans
 - for attributes, 2
 - for variables, 4–5
 - single-sided specification limits, 5
 - terminology, 1
 - two-sided specification limits, 5
- accrual bias, 51
- accuracy, 374–375
- acellular pertussis vaccines, 238
- acne, 1666
- active concurrent control, 623
- active control equivalence trials (ACETs), 2201–2202
- active control trials, 8, 14–22
- basic assumptions in, 15
 - design considerations, 20–21
 - hypotheses, 15–18
 - non-inferiority hypotheses
 - with fixed δ margin, 15–16
 - with fraction of control effect to be retained, 17–18
 - objective of, 14
 - planning for, 20
 - relationship between fixed δ -margin hypotheses and fraction retention hypotheses, 18–19
 - remarks, 19–20
 - sample size determination, 21
 - superiority hypotheses, 15
 - validity of, 20
- active (drug) substance
- crystal properties, polymorphs, and solvates, 255
 - intrinsic dissolution rate, 255
 - particle size, surface area, and shape, 255
 - physicochemical characterization, 254–256
 - pK_a , partition coefficient, and permeability, 255
 - salts and prodrugs, 256
 - solubility, 254
 - stability, 256
 - stereochemical characterization, 256
- acute coronary syndromes (ACS), 966
- acute myelogenous leukemia (AML), 763
- adaptation rules, clinical trial simulations for later development phases, 545
- adaptive allocation designs, 1727
- adaptive alpha allocation approach (4A) and consistency ensured methods, 731
- adaptive design, 926, 1572, 1573, 1952
- application and analyzing data from adaptive designs, 1953
 - clinical trials, 28–34
 - case studies, 32–34
 - issues and considerations on design
 - adaptation, 30–32
 - simulations, for later development phases, 544–545
 - with dropping losers
 - clinical trial simulations for later development phases, 550
 - for drugs and biologics, 1573
 - for medical device clinical studies, 1573
 - permitting early stopping and sample size reestimation
 - clinical trial simulations for later development phases, 549
 - sample size of, 1952
 - variability and power of using, 1952
- adaptive enrichment, 965
- adaptive performance, measure of, 2028
- definition, 2028
- adaptive randomization, 1877
- clinical trial simulations for later development phases, 548
- adaptive stratified designs, phase III cancer clinical trials, 395
- adaptive survival trials, 50–51
- optimal allocation approach, 51
 - in practice, 51
 - treatment effect mapping approach, 50
- additive drug, 320
- abuse trials, 316
- Additive Hazards Model (AHM), 2063
- additive hazards rate model, sequential estimation for, 2063–2067
- parameters estimation, 2063
 - data description, 2064
 - estimators for the AHM, 2064–2065
 - large sample properties, 2065
 - regression models, 2063–2064
 - sequential stopping rules, 2065–2066
- administrative interim analyses, 1196
- adverse drug events (ADE) signaling, 1756–1767
- data mining for, 1762
 - multi-item association, 1765
 - pairwise association, 1763–1765
 - high reporting rate, signaling of, 1758–1762
 - monitoring the increase of reporting of, 1757–1758
 - signaling reporting profile difference, 1762
- adverse drug reaction (ADR), 655
- adverse event reporting, 53–58
- background of drug approval process, 53
 - coding dictionaries, 54–55
 - common dictionaries, utilization, 54
 - dictionary versions, 55
 - Medical Dictionary for Regulatory Activities, 54f–55
 - need to code, 54
 - statistical analyses, 55–57
 - individual study report, 55–56
 - integrated summary of safety, 57
 - as secondary endpoint, 57
- Adverse Event Reporting System (AERS), 1756
- adverse events (AEs), 1506–1507, 1665
- adverse events of special interest (AESI), 814
- adverse reactions (ARs), 1665
- aerosol droplet size and distribution, 260
- Affymetrix array, 1378
- Affymetrix GeneChip™, 1378
- aggregate-data study, ecologic inference, 947
- aggregated criteria
 - vs. disaggregated criteria
 - individual bioequivalence, 1164
 - for similarity, 306
- aggregate-level confounding due to individual-level effects, 941–943
- aggregation level, problems in, ecologic inference, 949
- agranulocytosis, 191
- Akaike information criteria (AIC), 1271, 1330, 170, 1312
- all-comer design, 1572
- all-comer population, 964, 965
- allele, 1057
- allele-specific polymerase chain reaction for screening point mutation (example), 1025
- Allen *et al.* method, 1903
- product requiring reconstitution prior to use, 1903–1904
 - product with degradation, 1903
 - product with no degradation, 1903

- “all randomized subjects”, 417
- α -adjustment
 - clinical trial simulations for later development phases, 547–548
- alpha interferons, 237
- α -level AVE test, 553, 554
- α -level MAX test, 553
- α -level min test, 553
- alpha spending function, 59–66, 1198
 - application, 63–65
 - design and analysis of interim trial using, 63, 64
 - issues in using alpha-spending functions, 65
 - background, 59
 - choice of $\alpha^*(t)$, 62
 - computing software, 62
 - concept of, 60
 - derivation of original spending function, 61
 - information fraction t , 62
 - interim analysis, 59
 - origin of, 60
 - standard group sequential methods, 59–60
 - types of error spending functions, 61–62
- alternative designs, 926
- Altshuler’s estimator, 557
- Alzheimer’s disease, 840, 841, 961, 1666
- Alzheimer’s Disease Assessment Scale (ADAS), 960
 - cognitive subscale (ADAS-cog), 841
- AMBER force fields, 229
- Ames assay, 2233
- Ames, Bruce, 67
- Ames test, 67–73, 1155, 2233
 - analysis methods for, 69
 - background, 67
 - biologically based models, 71–72
 - description of conduct of, 68f
 - design, 68–69
 - discussion, 72–73
 - linear regression on revertant counts, 71
 - modified two- and threefold rules, 70
 - non-linear regression on revertant counts, 71
 - non-parametric methods, 70–71
 - parametric modeling, 71
 - reasons for, 67
 - understanding data, 69–70
- amniocentesis, 1057
- analyses of covariance (ANCOVA), 180, 705, 705, 912, 1422, 1587, 1624, 1634
 - discussion, 128
 - and pooling tests, 126–128
 - for premarketing shelf life determination with multiple factors, 125–128
- analysis methods, combination drug clinical trial, 552
- analysis of variance (ANOVA), 83, 100–101, 104–115, 264, 338, 375, 417, 436, 441, 442, 553, 634, 705, 833, 878, 910, 912, 1382, 1384, 1407, 1436, 1587, 1934, 1966, 2237
 - dose proportionality, 877
 - linear model of, 109–110
 - mathematical problem, 104
 - model assumptions
 - for one-way ANOVA, 105
 - for two-way ANOVA, 107
 - notations, 105, 108
 - interpretation, 105
 - parameterization, 1330
 - with rank transformations, 2240
 - remarks, 106, 114–115
 - theorem, 106
 - Total PROtein in serum (TPRO) example, 104f–105, 106–107, 108–109, 110–114
 - treatment effects, under breached blindness, 339
- analysis population, 1123
- analytes, 130
 - in complex matrices, 136–137
- analytical factors, 823
- analytical method, linearity of, 146
- analytical method validation (AMV), 1101
- analytical similarity assessment, 116–123
 - FDA’s current thinking on scientific input, 122
- fundamental similarity assumption, 117–119
- heterogeneity within/between lots within/between test and reference products, 121
- inconsistent results between tiered approaches, 120
- issues of fixed margin/range in tiered approach, 120
- primary assumptions for the tiered approach, 118
 - Tier 1 equivalence test, 118
 - Tier 2 quality range approach, 118–119
 - Tier 3 raw data and graphical comparison, 119
- sample size requirement, 121
- similarity limit and variability, relationship between, 122
- statistical properties of FDA’S Tier 1 equivalence test, 119–120
- tiered approach of, 2219–2224
 - equivalence margin, determination of, 2221–2223
 - graphic and raw data comparison for Tier 3 CQAs, 2221
 - issues of the proposed statistical approaches for Tier 1 CQAs, 2223–2224
 - quality range approach for Tier 2 CQAs, 2221
 - statistical approaches used for assessing similarity, 2219–2220
 - statistical equivalence test for Tier 1 CQAs, 2220–2221
 - unified tiered approach, 122–124
- Anderson–Gill (AG) model, 396, 1494, 1496, 2034
- animal model vs. human model, 1398
- annotation, 230
- anti-obesity agent, 536
- aortic stenosis, 1791
- ApoAI data, local FDR on, 1033
- appropriate estimation method, 1330
- approximate inference procedures, 1925
- Aranda–Ordaz model, 955, 956, 957
- Arbuthnot’s probability, 1845
- arcsine-square-root transformation, 1233
- area-aggregated rates, 2089
 - spatial models with random effects, 2089–2090
 - spatio-temporal models, 2090–2091
- area under blood/plasma concentration-time curve, 928
- Area Under Effect Curve (AUEC) values, 267
- area under the curve (AUC), 824, 2101
- area under the receiver operating characteristic curve, 1963–1964
- artificial neural network (ANN), 808–809, 1738
- Asia-Pacific Region (Japan, Korea, and China), 287
 - China Food and Drug Administration (CFDA), 288
 - Korea Food and Drug Administration (KFDA), 288
 - Ministry of Health, Labour and Welfare (MHLW), 287–289
- assay control, 130
- assay, defined, 130
- assay development, 130–137
 - assay optimization, 135–136
 - assay technologies, 131
 - biological assays and assays of analytes in complex matrices, 136–137
 - calibration, 131–135
 - linear calibration, 131–133
 - nonlinear calibration, 133–134
 - other calibration conventions, 134–135
 - terminology, 130
- assay sensitivity, 15, 20
- assay validation, 1103
 - accuracy, 374–375
 - calibration application in, 374–376
 - performance characteristics, 375–376
 - precision, 375
- assay validation, linearity evaluation in, 146–153
 - average deviation from linearity (ADL), 148
 - coefficient of variation of the deviations from linearity (CVDL), 148–149
 - EP6-A’s criterion, 147–148
 - experiment design, 147
 - numerical examples, 151
 - estimation method vs. TOST method, 151
 - GPQ-SSDL method vs. GPQ-ADL method vs. GPQ-CVDL method, 152–153
 - simulation study, 150–151
 - statistical testing procedures, 149, 150
 - estimators of SSDL, ADL, and CVDL, 149
 - generalized pivotal quantity approach, 149–150
 - two one-sided tests (TOST) procedures, 149
 - sum of squares of deviation from linearity (SSDL), 148
- asthma, 1666
- asymptotically correct tests, 1299
 - example, 1300–1301
- asymptotically unbiased model, 373
- asymptotic approach, 2283–2284

- asymptotic approximation, 169
 - asymptotic covariance matrix, 318
 - asymptotic inferences, odds ratio, 1555
 - asymptotic power, measuring agreement, 1338
 - asymptotic properties, generalized estimating equation, 1047
 - asymptotic relative efficiency (ARE), 787, 1048
 - Australian phenytoin incident, 259
 - autologous products, 236
 - “available information” assumption, 428
 - average BE with expanding limits (ABEL), 2038
 - average bioavailability, 306
 - average bioequivalence (ABE), 288–289, 936, 1160, 1432, 1740
 - average cost-effectiveness ratio (ACER), 753
 - average coverage criterion (ACC), 195
 - average deviation from linearity (ADL), 146, 148
 - “averaged success probability”, 523
 - average length criterion, 195
 - Average Outgoing Quality (AOQ), 3f
 - Average Outgoing Quality Limit (AOQL), 3
 - average performance score (APS), 2028
 - “average power”, 523
 - average ratio vs. incremental ratio, 1710–1713
 - independent programs, 1711
 - mutually exclusive programs, 1711–1713
 - average run length (ARL), 2138
 - Average Sample Number (ASN) curve, 3–4f
 - average superiority, 1474
 - average time to signal (ATS), 2138
 - Average Total Inspection (ATI) graph, 3f
 - AVE test, 553
 - azidothymidine (AZT), 1728
- B**
- Bacillus Calmette–Guerin (BCG) vaccine, 604
 - clinical trials, 608
 - BACS (the Brief Assessment of Cognition in Schizophrenia), 2331
 - Balaam’s design, 298, 565, 925–926
 - balanced linear models, tolerance intervals for, 1055
 - bar chart, 481
 - bar graphs, 482
 - Bartholomew’s test, for determining drug effect, 880
 - baseline correction, 2239
 - base rate problem, 1242
 - Basic Local Alignment Search Tool (BLAST), 225
 - basket rack assembly, 260
 - batch-to-batch similarity, expiration dating period, 982
 - Bayes factor, 181, 192, 1126, 1271
 - Bayes Gibbs sampling approach, 94
 - Bayesian adaptive design, 200–202, 1357
 - Bayesian analysis, 1269, 1318, 190, 316
 - correlated probit model, 746
 - Dale’s model, 92–93
 - latent mixed model, 94
 - Bayesian approach, 203, 387, 517, 542, 1043, 1044, 1516–1517, 1533
 - clustered categorical data, 91
 - correlated probit model, 743
 - crossover design, 781
 - early phases of clinical trials, simulation in, 542
 - in evaluating medical devices, 1356
 - generalizability probability, 1043
 - parallel-group designs, 520–521
 - to stability analysis, 180–184
 - example, 184
 - hierarchical model, 181
 - preliminaries and notations, 181
 - shelf-life estimation, 183
 - testing for poolability of batches, 182
 - two samples with equal variances, 517–520
 - two samples with unequal variances, 519–520
 - Bayesian averaging, 946
 - Bayesian calculations, 1358
 - Bayesian clinical trial, analyzing, 199
 - hypothesis testing and interval estimation, 199
 - interim analysis, 199–200
 - model checking, 200
 - planning, 199
 - Bayesian design, operating characteristics of, 199
 - determining the sample size, 199
 - prior distribution, 199
 - posterior distribution, summary of, 199
 - sensitivity analysis, 200
 - Bayesian clinical trial simulation, 524
 - Bayesian computing, 195–196
 - Bayesian confidence interval, 294
 - Bayesian Cox model, 594
 - Bayesian design, 1732
 - for incorporating historical data, 202
 - for phase II oncology clinical trials, 154–162
 - adaptations of designs, 161
 - designs, 154–156
 - discussion, 161–162
 - dual threshold design, 157–158
 - probabilities of early termination and expected sample sizes, 159
 - single threshold design, 156–157
 - using triplet combination, 160–161
 - for using adult data for pediatric indication, 202–203
 - Bayesian estimate of concordance correlation coefficient (CCC), 164
 - estimation for basic setup, 169
 - execution of, 169
 - extension to address other practically relevant issues, 169
 - motivation for, 164
 - potential limitations and future research of, 170
 - Bayesian factor analysis models, 1010
 - Bayesian framework, 169, 523
 - Bayesian hierarchical modeling, 1358
 - Bayesian hypothesis test, 1126, 1322
 - Bayesian inference Using Gibbs Sampling (BUGS), 176, 195–196, 1358, 170
 - Bayesian information criteria (BIC), 1271, 1330, 170, 1312
 - Bayesian interim analyses, 1357
 - Bayesian logistic regression model (BLRM), 1561–1562
 - model, 1562
 - overdose control, 1562
 - practical notes, 1562–1563
 - Bayesian Markov Chain Monte Carlo (MCMC) methods, 1372
 - Bayesian method, 25, 1327, 294–295
 - inference with, 426
 - in meta-analysis, 173–178
 - applications of, 173–174
 - bias, accounting for, 177
 - computation, 176–177
 - extending simple hierarchical model, 177
 - models for, 175–176
 - prior distributions, 176
 - simple hierarchical model, 175
 - steps in, 174–175
 - stochastic curtailment and, 1109
 - Bayesian model, 1533–1534, 1563
 - Bayesian noninferiority testing, 1357
 - Bayesian paradigm, clinical pharmacology, 471
 - Bayesian procedure, 411
 - Bayesian sample size, 1357
 - Bayesian single-stage approach, phase II cancer clinical trials, 393
 - Bayesian statistics, 190–196, 1356
 - basic notions and Bayes theorem, 190–192
 - example, 191–192
 - Bayesian computing, 195–196
 - choice of prior distributions, 193–194
 - clinical trial monitoring, 194
 - hypothesis testing, 192
 - example, 192
 - inference, 1703
 - program, 1358
 - sample size determination, 194–195
 - skepticism and, 1504–1505
 - Bayesian trees, forests and, 1788
 - Bayes theorem, 190–192
 - Bazett’s correction, 2207
 - Behrens–Fisher (BF) problem, 1115
 - bench testing, 824
 - Benefit-Risk Action Team (BRAT), 474, 476
 - benefit–risk assessment (BRA), in clinical research, 472–482
 - challenges, 482–483
 - CIRS UMBRA Initiative, 474
 - decision makers, 473
 - definitions, 472–473
 - EFSPI Benefit-Risk Special Interest Group and QSPI Benefit-Risk Working Group, 475
 - European Medicines Agency, 474
 - Food and Drug Administration, 473–474
 - frameworks, 475
 - BRAT framework, 476
 - EMA effects table, 475–476
 - FDA sB/R framework, 476–477
 - ProACT-URL, 475
 - Innovative Medicines Initiative and European Federation of Pharmaceutical Industries and Associations, 474–475
 - International Council for Harmonisation, 474
 - ISPOR Risk-Benefit Management Working Group, 475

- benefit–risk assessment (BRA), in clinical research, (*Cont.*)
 - quantitative methodologies, 477
 - general methods to deal with uncertainties, 480–481
 - multi-decision criteria analysis, 478–480
 - number needed to treat/number needed to harm, 477–478
 - visualization, 481
 - bar graphs, 482
 - effects table, 481
 - forest plots, 482
 - tree diagrams, 481–482
- Benjamini–Hochberg adjusted p -value, 1190
- Bergum approach
 - finished product, 1811
 - process validation, 1806, 1810
 - effect of sample size, 1807
 - statistical basis, 1806
- Bernoulli distribution, 2330
- Bernoulli process, 1845
- Bernoulli trial, 993, 2031
- best-fit model, 148
- best linear unbiased predictor (BLUP), 505
- β -HCG, measurement of, 151
- Beta-Blocker Heart Attack Trial (BHAT), 64–65 f
 - maximum duration trial, 65
 - maximum information trial, 65
- beta blockers, 1722
- beta receptors, 1722
- beta spending function, 63
- between-treatment-groups analysis,
 - pharmacodynamics with covariates, 1679–1680
- bias considerations, diagnostic imaging, 831–832
 - for a blinded-read trial, 832
 - for an open-label trial, 831
- bias reduction methods, 1450
 - adaptive Lasso and elastic net, 1451
 - combined penalty (CP) function, 1450–1451
 - relaxed Lasso, 1451
 - SCAD and its oracle property, 1450
 - two-step strategy, 1451
- binary 2×2 crossover trials, 206–212
 - crossover design, 206
 - discussion, 212
 - Hills–Armitage test, 211–212
 - Mainland–Gart test, 208–210
 - McNemar test, 207–208
 - Prescott test, 210–211
 - tests for 2×2 crossover design, 206–207 f
- binary blends, 1947
- binary data
 - without batch-to-batch variation, expiration dating period, 982–983
 - effective doses, 952
- binary endpoint, 279, 1579
- binary outcome, 82
 - variable case, 2011, 2012
- binary-valued test, 824
- binomial distribution, sample size calculation based on, 1981
- binomial (multinomial) models with extra variation, 992
- binormal receiver operating characteristic curve, 1963
- bioassay, 131, 213–215
 - and assays of analytes in complex matrices, 136–137
 - parallel-line bioassays, 213–214
 - quantal response and tolerance distribution, 214–215
- bioavailability (BA)
 - definition, 216, 263
 - determination of, 263
- bioavailability/bioequivalence (BA/BE) studies, 216–222
 - clinical endpoint BE study, 221
 - clinical endpoints, 221
 - statistical approaches, 221–222
 - study design, 221
 - discussion, 222
 - pharmacokinetic (PK) bioequivalence study, 217
 - highly variable drugs, 220
 - narrow therapeutic index (NTI) drugs, 220
 - PK measures, 218
 - statistical approaches for BE assessment, 218–220
 - study design, 217–218
- bioequivalence (BE), 928, 1160
 - application of MLS method to, 1431–1432
 - clinical pharmacology, 470
 - definition, 216, 263
 - design and implementation of, 264
 - determination of, 263
 - determining limits, 1191
 - meta-analysis for, 1369
 - statistical analysis, 264–268
 - food-effect bioavailability and bioequivalence studies, 265–266
 - individual bioequivalence, 266
 - multiple-dose studies, 265
 - narrow-therapeutic-range drugs in Canada, 266
 - pharmacodynamic measures of bioequivalence, 266–268
 - single-dose crossover study, 264–265
 - structural equation model applied to, 2149–2150
 - testing, 345–346
- bioequivalence criteria and conceptual issues, 1250–1251
 - design requirements, 1252
 - desirable properties for distance, 1251
 - generalization to exponential family of distributions, 1252
 - generalization to multivariate case, 1252–1253
 - interpretation on the original scale, 1252
 - logical hierarchy of conclusions, 1252
 - scaling and symmetry, 1251
- bioinformatics, 224–231
 - analyzing biological literature, 229–230
 - comparative genomics, 226
 - extracting relationships, 230
 - finding sequence motifs, 225–226
 - linking biological data to text, 230
 - microarrays, 226–227
 - pairwise sequence comparison, 225
- proteomics, 227
 - sequence analysis, 224–225
 - structural, 228–229
 - systems biology, 227–228
 - translational, 230
 - understanding relationships between genotype to phenotype, 228
- biological data, linking to text, 230
- biological literature, analyzing, 229–230
- biologically based models, of mutagenicity and toxicity, 71–72
- Biologic License Application (BLA), 629
- biologics, 233–234
 - bioterrorism products, 241
 - issues
 - in biologic therapeutics, 237
 - in blood and blood products, 234–236
 - low prevalence conditions, 236
 - safety evaluation, 236
 - statistical issues in assessing biologic test kits, 235–236
 - in vaccines, 237–240
 - pharmacogenetics and drug development, 240–241
 - regulatory process, 234
 - generic products, 234
 - international harmonization, 234
 - product development and licensure, 234
 - websites for FDA information and guidance documents, 241
- biologics license application (BLA), 286, 305
- Biologics Price Competition and Innovation (BPCI) Act, 281, 564–565, 933
- biologic test kits, statistical issues in assessing, 235–236
- biologic therapeutics, issues in, 237
- biomarker-based designs, 964
- biomarker-driven designs, 1572
- biomarker-negative, 964
 - subgroups, 967, 1576
- biomarker/pharmacodynamic data, designs that incorporate, 1567
- biomarker-positive, 964
 - enrichment, 965
 - subgroups, 965, 967, 1576
- biomarker predictive value, 1651
- biomarkers, 1649
- biomarker-stratified design, 1651
 - enrollment strategy for, 1575
- Biometric and Epidemiology Methodology Advisory Committee (BEMAC), 786
- biopharmaceutics, 254–271
 - active (drug) substance
 - crystal properties, polymorphs, and solvates, 255
 - intrinsic dissolution rate, 255
 - particle size, surface area, and shape, 255
 - physicochemical characterization, 254–256
 - pK_a , partition coefficient, and permeability, 255
 - salts and prodrugs, 256
 - solubility, 254
 - stability, 256
 - stereochemical characterization, 256
 - bioavailability and bioequivalence, 263–268
 - definitions, 263

- design and implementation of bioequivalence studies, 264
- determination of, 263
- statistical analysis of bioequivalence studies, 264–268
- compendial monograph tests and standards, 259–263
 - aerosol droplet size and distribution, 260
 - content uniformity, 259–260
 - disintegration, 260–261
 - dissolution, 261–263
 - particle size, 260
- dosage forms, 258–259
 - capsules, 259
 - nonoral dosage, 259
 - solutions, 258
 - suspensions, 259
 - tablets, 259
- formulation components, 257–258
 - disintegrating agents, 257
 - excipients/fillers, 257
 - formulation of inhalants, 258
 - granulating agents and binders, 257
 - lubricants, 257
 - ointments and skin preparations, 257
 - surfactants (wetting agents), 257
- in vitro/in vivo correlations, 270–271
- manufacturing process, 258
 - compression, 258
 - granulation, 258
 - mixing, 258
- properties, 254–263
 - compendial monograph tests and standards, 259–263
 - dosage forms, 258–259
 - excipients, formulation factors, and manufacture, 256–258
 - physicochemical characterization, 254–256
- scale-up and postapproval changes (SUPAC) guidelines, 268–270
 - common elements, 268–269
 - comparison of profiles, 270
 - SUPAC-IR, 269–270
 - SUPAC-MR, 270
 - SUPAC-SS, 270
- biosimilarity assessment, 281
 - Asia-Pacific Region (Japan, Korea, and China), 287
 - China Food and Drug Administration (CFDA), 288
 - Korea Food and Drug Administration (KFDA), 288
 - Ministry of Health, Labour and Welfare (MHLW), 287–289
 - European Union (EU), 284
 - European experiences, 286
 - key principles and basic concept, 285
 - nonclinical and clinical evaluation, 285
 - product class-specific guidelines, 285–287
 - quality, 285
 - reference biotherapeutic product, 285
 - North America, 286
 - Canada (Health Canada), 287
 - US Food and Drug Administration, 286–287
- power calculation for sample size, 297–298
- statistical methods, 292
 - Bayesian methods, 294–295
 - confidence interval approach, 293
 - interval hypotheses, 292
 - Schuirmann's two one-sided tests procedure, 293–294
 - Wilcoxon–Mann–Whitney two one-sided tests procedure, 295–297
- unified approach for, 298–300
- World Health Organization, 282
 - characterization, 284
 - key principles and basic concept, 283
 - manufacturing process, 284
 - nonclinical and clinical studies, 284
 - quality, 284
 - reference biotherapeutic product, 283–285
- biosimilarity, classic methods for assessing, 292
 - Bayesian methods, 294–295
 - confidence interval approach, 293
 - Schuirmann's two one-sided tests procedure, 293–294
- biosimilarity, criteria for, 288
 - average bioequivalence (ABE), 288–289
 - dissolution profile comparison, similarity factor for, 291–292
 - population/individual bioequivalence, 289
 - masking effect, 290
 - profile analysis for in vitro bioequivalence testing, 290–291
- biosimilars
 - application to, 932
 - of follow-on biologics, *see* follow-on biologics, biosimilarity of
- biosimilar studies, 274–276, 277–279
 - binary end points, 279
 - equivalence trials based on two-arm parallel design, 274–275
 - exact and approximate power, 275
 - comparison of, 275–276
 - statistical tests for biosimilarity, 278–279
 - three-arm parallel design, 277–278
- biostatistics and good clinical practice, 1079–1080
- biotechnology company, contract research organization (CRO)
 - increase of outsourcing by, 718
- Biotech Validation Suite For Protein Structures, 229
- bioterrorism products, 241
- bipolar VAS, 1123
- birth and death urns, 1729
- bisection method, 2033
- bivariate double exponential distribution, 792
- bivariate kernel density estimation,
 - mathematical program for, 1775
- bivariate model, 2108
- bivariate normal distribution, 791–792
- bivariate skew t distributions (BVSTs), 166, 792
- bivariate zero-inflated Poisson (ZIP) model, 316, 317
 - determination of sample size, 319
 - example, 320
- likelihood ratio tests for comparison of treatments, 319
- blinded-read component, diagnostic imaging, 831
- blinded-read trial, assessing reader agreement for diagnostic imaging, 832–833
- blinding, 336–341
 - analysis under breached blindness, 338–340
 - clinical trials, 573
 - example, 340–341
 - integrity, 337
 - introduction, 336
 - types, 336–337
 - double blind, 336
 - open label, 336
 - single blind, 336
 - triple blind, 337
- blocked randomization, 573
- blocking, 1622
- blood and blood products, 234, 236
 - evaluating safety of, 236
 - low prevalence conditions, 236
 - statistical issues in assessing biologic test kits, 235–236
- blood glucose measurement, 825
- BMDP-5V, 196
- bone density, 822
- bone mineral density (BMD), 730
- Bonferroni adjustment tests, dose response analysis in clinical trials, 880
- Bonferroni correction, 1468
- Bonferroni–Dunn test, 1689
- Bonferroni–Holm procedure, 881
 - dose response analysis in clinical trials, 881
- Bonferroni procedure, 1028, 1407, 729, 1638
 - modified, 2209
- bootstrap accelerated bias-corrected percentile method, 344
- bootstrapping linear models, 343
- bootstrapping residuals, 343, 345
- bootstrap resampling, 1444
- bootstrap technique, 342–346, 1470, 1700
 - applications in bioequivalence testing, 345–346
 - data sets, 343
 - generating bootstrap data, 343–344
 - bootstrapping imputed data, 344
 - bootstrapping linear models, 343
 - longitudinal bootstrap, 343
 - nonparametric bootstrap, 343
 - parametric bootstrap, 343
 - without-replacement bootstrap, 344
 - using bootstrap data, 344–345
 - percentile type confidence sets, 344
 - t-type confidence sets, 344–345
 - variance estimators, 344
 - validity of, 345
- Box–Behnken designs (BBD), 1944f
- Box–Cox transformation, 168, 1643
- bracketing design, 359
- brain potentials, application to, 1550–1551
- brain volume, 822
- brand-name drug, 928, 929
- breached blindness, 338–340

- breast cancer, 825, 1652
 - and alcohol use, 2286
 - inference in a single study, 2286–2287
 - meta-analysis and trend estimation, 2287
 - local FDR on breast cancer data set, 1033
 - stem cell, 965
 - treatment, 1648
- Breslow estimator, 1500
- bridging study, 238–239, 635, 636, 1573
- Brief Psychiatric Rating Scale (BPRS), 192
- British Institute of Radiology (BIR) trials of radiotherapy in cancer of the larynx and the pharynx, 1306–1308
- British Medical Journal, 424
- British Pharmacopoeia (BP), 259
- bronchial asthma data, analysis of, 1928–1929
- Brownian motion, 545, 2326
 - process, 60, 61
 - structure, 1233
- Brown–Mood test, 1289
- B-splines, 1544, 1545
- B-type natriuretic peptide, 825
- Buckley–James estimator, 558
- budget, 978
- budget impact analysis (BIA), 756
- Buehler test, 2231
- BUGS (Bayesian inference using the Gibbs sampling technique), 176
- Buonaccorsi, 374
- C**
- Caco-2, 255
- calcium, measurement of, 152
- calibration, 370–376
 - and analytical results, 373
 - application in assay validation, 374–376
 - accuracy, 374–375
 - other performance characteristics, 375–376
 - precision, 375
 - example, 373–374
 - models, 312
 - standard, 130
 - standard curve, 370–372
 - model selection, 370–371
 - weight selection, 372
- calibration assay, 131f–135
 - linear calibration, 131–133
 - nonlinear calibration, 133–134
 - other calibration conventions, 134–135
- cancer screening data, analysis of
 - cluster trials, 653–654
- cancer trials, 390–396
 - phase I clinical trials, 390–392
 - continual reassessment method and modifications, 391–392
 - discussion, 392
 - new designs, 391
 - traditional design, 390
 - phase II clinical trials, 392–394
 - Bayesian approach, 393
 - classical approach, 393
 - designs, 392–394
 - with multiple experimental arms, 395
 - phase III clinical trials, 394–395
 - adaptive stratified designs, 395
 - with multiple experimental arms, 395
- stratified designs, 394
- unstratified designs, 394
- statistical methods, 395–396
- canonical analysis, 1946
- capsules, in dosage forms, 259
- carcinogenicity of pharmaceuticals, 398
 - adjustment for effect of multiple tests, 410–411
- adjustment of tumor rates for intercurrent mortality, 399–400
- analysis of data from studies using dual control groups, 409–410
- contexts of observation of tumor types, 400
- evaluation of validity of designs of negative studies, 398–413
- interpretation of results of statistical tests, 410
- tests for trend and difference in tumor incidence in chronic studies
 - using Peto tests, 399–405
 - using poly-*k* tests, 405–409
- tests of data
 - exact tests, 403–405
 - fatal tumors, 403
 - incidental tumors, 400–402
 - mortality-independent tumors, 403
 - tumors observed in incidental and fatal contexts, 403
- use of historical control data, 411–412
- carcinogenicity studies, 2233
- Cardiac Arrhythmia Suppression Trial (CAST), 574, 576, 961, 962
- Cardiac Safety Research Consortium (CSRC), 2215
- cardiovascular system, 1666
 - congestive heart failure, 1666
 - coronary artery disease/angina, 1666
 - hyperlipidemia, 1666
 - hypertension, 1666
- Carrière procedure, 1925
- carry-forward analysis, 416–421
 - discussion, 420–421
 - numerical example, 418–420
 - regulatory requirements, 416–417
 - statistical methods, 417–418
- carryover effect, 206, 777
 - clinical pharmacology, 470
- cascade impaction technique
 - in vitro* bioequivalence testing, 1142
- cascade impactor (CI), 1144, 1817
- case-control studies, inference in, 422–432
 - inference under full population information, 425–426
 - inference with Bayesian prior assumptions, 426–427
 - inference without full population information, 427–432
 - inference without population information, 424–425
 - quantities of interest, 423–424
 - sampling designs, 424
- case report forms (CRFs), 444, 462–463
- CATMOD procedure, 115
- Cauchy–Schwarz inequality, 2034
- causal inference, medical devices, 1352
- causal interpretation, problems in ecologic inference, 949
- cause-and-effect diagram, 2136f
- CD4 lymphocyte counts, 2175
- cDNA microarray, 1378
- celebravascular artery (CVA), 1762
- censoring, 50, 51, 755
- center-by-treatment interaction, medical devices, 1352
- Center for Biologics Evaluation and Research (CBER), 233
- Center for Devices and Radiological Health (CDRH), 233, 1348
- Center for Drug Evaluation and Research (CDER), 233, 407, 1184
- Center for Food Safety and Applied Nutrition (CFSAN), 233
- Center for Veterinary Medicine (CVM), 233
- center sizes, 438
- center weighting in multicenter trials, 435–442
 - comparisons of weighting schemes, 437
 - construction of composite centers, 439–440
 - extensions to non-normal data, 441–442
 - in random-effect models, 440–441
 - regulatory guidance, 437–438
 - sample size considerations, 438–439
 - statistical framework, 435–437
- central composite design (CCD), 1943f, 1944
- centralized risk-based monitoring, 820
- central laboratories, 1262–1263
- Central Limit Theorem (CLT), 342, 489
- Centre for Innovation in Regulatory Science (CIRS) UMBRA Initiative, 474
- chain Monte Carlo algorithms, factor analysis, 1010
- change control, global database and system, 1073
- change management and training, global database and system, 1072–1073
- Chang's method, 2290
- CHARMM force fields, 229
- chemistry, manufacturing, and controls (CMC) information, 1394
- chemo-protective agent (CPA), 763
- chemotherapy, 889
- child abuse, 320
- child injury due to accidental falls, 2095f, 2097f
- China Food and Drug Administration (CFDA), 288
- China, good laboratory practice (GLP), 1085
- Chinese hamster lung (CHL) cells, 1155
- Chinese hamster ovary (CHO) cells, 1155, 1156
- Chinese medicine, *see* traditional Chinese medicine (TCM)
- Chinese Pharmacopoeia (CP), 2262
- chi-square goodness-of-fit tests for chromosome data, 999
- chi-square statistic, 1602
 - Cochran–Armitage, 2284
 - of homogeneity, 1365
- chi-square test, 56, 75, 77, 82, 83, 221–222
 - for individual-level data
 - cluster trials, 651
 - maximal correlation and, 76
 - regression and, 77
 - statistic, 1691

- chlorhexidine, 655
- Choice of Control Group and Related Issues in Clinical Trials, 1080
- cholesterol education and research trial (CEART), 649, 650
- CHRM2* gene, 1064
- chromosomal aberration test in vivo, 2234
- chromosome data, 999
 - chi-square goodness-of-fit tests for, 999
- chronic obstructive pulmonary disease, 2031
- Chuang's life table approach, 2227
- Cicchetti–Alison weight, 833
- classical confidence intervals, 429
- classical linkage analysis, 1057–1058
- classical single-stage approach, phase II cancer clinical trials, 393
- classic case-control, 424–426, 427–430
- classification accuracy, 1654
- Classification and Regression Trees (CART), 1783
- classification model, 827
- class membership, posterior probability of, 1270
- clinical abuse potential studies, 1121
- clinical component, diagnostic imaging, 830–831
- clinical correlations, laboratory analyses, 1263–1264
- Clinical Data Interchange Standards Consortium (CDISC) standards, 1089
- clinical data management (CDM), 444–450
 - processes, 445–449
 - case report form design, 445–446
 - closing a database, 449
 - database design, 446–447
 - data-handling standards, 446–448
 - data validation, 448–449
 - regulations and guidance relevant to, 444–445
 - 21 CFR Part 11, 444
 - 21 CFR Part 312, 444
 - 21 CFR Part 314, 445
 - bioresearch monitoring, 445
 - guidance for industry, 445
 - International Conference on Harmonization, 445
 - systems, 450
- clinical decision rules, 734–735
- clinical development, 911
- clinical efficacy variables, 1722
- Clinical Global Assessment scales, 1625
- clinical inspection, 462–468
 - FDA 483 actions, 465
 - FDA 483 observations, 464–465
 - inspection process, 463
 - before inspection, 463
 - exit interview, 464
 - inspection, 463–466
 - scope of inspection, 463
 - statistical procedure, 465
 - hypergeometric distribution, 465–466
 - single-stage sampling plan, 466–467
 - two-stage sampling plan, 467–468
- Clinical Laboratory Improvement Act (CLIA), 1349
- Clinical Laboratory Standard Institute (CLSI) EP6-A guideline
 - acceptance of linearity by, 146
 - failure of linearity by, 146
- clinical lot consistency studies, 238
- clinical outcome assessments (COAs), 316, 1207–1208
- clinical pharmacology, 469–471
 - absolute bioavailability studies, 470
 - bioequivalence studies, 470
 - dose proportionality studies, 470
 - drug interaction studies, 470
 - fed/fasted studies, 470
 - first human dose studies, 469
 - multiple-dose escalation studies, 470
 - pharmacokinetic and pharmacodynamic work, 470
 - recent developments, 471
 - special population studies, 470
 - statistical design, methods, and analysis, 470–471
- clinical practice, *see* good clinical practice (GCP)
- clinical relevance of laboratory findings, 1263
- clinical research associates (CRA), 336
 - clinical data management guidelines, 446
- clinical research, benefit–risk assessment in, *see* benefit–risk assessment (BRA), in clinical research
- clinical research, comparing variabilities in, 486–500
 - comparing intersubject variabilities, 491
 - comparing intrasubject coefficients of variation, 489
 - conditional random effects model, 490–491
 - test for equality, 491
 - test for noninferiority/superiority, 491
 - test for similarity, 491
 - intrasubject variabilities, comparing, 487
 - noninferiority/superiority, test for, 493
 - test for similarity, 494
 - parallel designs without replicates, 494
 - test for equality, 494
 - test for noninferiority/superiority, 494–495
 - test for similarity, 495
 - parallel design with replicates, 487, 491, 495
 - test for equality, 487, 491–492, 495
 - test for noninferiority/superiority, 487, 492, 495–496
 - test for similarity, 487–488, 492–493, 496
 - replicated $2 \times 2m$ crossover design, 497
 - test for equality, 497
 - test for noninferiority/superiority, 497
 - test for similarity, 497–498
 - replicated crossover design, 488, 493
 - test for equality, 488–489
 - test for noninferiority/superiority, 489
 - test for similarity, 489
 - sample size
 - calculation, 498
 - for *F*-test, 498–499
 - for modified large sample method, 499–500
 - for *Z*-test, 499
 - simple random effects model, 489
 - test for equality, 489
- test for noninferiority/superiority, 489–490
 - test for similarity, 490
- standard 2×2 crossover design, 496
 - test for equality, 496
 - test for noninferiority/superiority, 496
 - test for similarity, 496–497
- total variabilities, comparing, 494
- clinical significance vs. statistical significance, laboratory analyses, 1264
- Clinical Supply Group, 1218
- Clinical Trial Applications (CTAs), 1343
- clinical trial assay (CTA), 1652
- clinical trial design, 523, 1651
 - application of assurance, 525
 - assurance by simulation, 524
 - background, 523
 - basic example, 523
 - with prospective patient population enrichment, 526–529
 - placebo lead-in design, 527–528
 - placebo nonresponse, enrichment on the basis of, 527
 - randomized withdrawal trial design, 528–529
 - sequential parallel comparison design (SPCD), 528
 - two-way enriched design (TED), 529
- clinical trial monitoring, Bayesian analysis for, 190, 194
- clinical trial process, 532–535
 - assessing the end of the process, 533
 - assessing the intermediate outputs or milestones, 534
 - assessing the starting point for the process, 533
 - assessing the tasks involved in the process, 533
 - continuous work process improvement, 534
 - gathering information, 532–533, 534
 - defining intermediate outputs or milestones, 533
 - defining the end of the process, 532
 - defining the starting point for the process, 532
 - defining the tasks in a process, 533
 - preparing a statement of problems, 534
 - work processes
 - monitoring, 534
 - understanding, 532
- clinical trials, 59, 564–571, 572–578, 580–587, 589–595, 1669
 - appropriate hypotheses for clinical investigation, 581
 - blinding, 573
 - complete *N*-of-1 design, 565, 566
 - with four periods, 566
 - with four periods vs. RRRR/RTRT design, 567–569
 - power analysis, 570–571
 - relative efficiency, 566–567
 - with three periods, 565–566
 - with three periods vs. 2×3 dual design, 569–570
 - with two periods, 565
 - with two periods vs. 2×2 crossover design, 569

- clinical trials, (*Cont.*)
 - conducting, in foreign regions, 977
 - content of the statistical section of a protocol for, 1842
 - primary analyses, 1842
 - primary statistical objective, 1842
 - primary variable(s) and analyses, 1842
 - primary variable definition, 1842
- data analyses, 577
- data monitoring committee (DMC)
 - challenge of the independence of, 586–587
- definition, 572
- endpoint selection, clinical strategy for, 583
- good clinical practice (GCP), 578
- handling missing values and outliers, 1670
- interim data analyses, 576–577
- interval-censored data, software for analysis of, 594
- key statistical principles, 2130
- loss to follow-up, 575
- missing data imputation, validity of, 585
- multicenter trials, 576
- multiplicity in, 585
- nature and purpose of the protocol for, 1841–1842
- notation, assumptions, and likelihood function, 590
- patient population included in statistical analysis, 1670
 - all randomized patients, 1670
 - per-protocol patients, 1670
- protocol amendments, impact and sensitivity of, 584–585
- randomization, 572–573, 582–583
- recruitment procedures, 574–575
- regulatory requirements, 2130–2131
- sample size calculation, instability of, 581–582
- selection and assessment of endpoints, 574
- statistical analysis, 2131
 - analysis sets, 2132
 - interim analyses, 2132–2133
 - missing data, 2132
 - of multivariate interval-censored data, 593
 - of other interval-censored data, 594
 - outliers, 2132
 - of univariate interval-censored data, 591–592
- statistical design, 2131
 - blinding, 2131
 - randomization, 2131
- two-stage adaptive design
 - flexibility and feasibility of, 586
- type of comparison, 1669
 - equivalence trials, 1669
 - non-inferiority trials, 1669
 - superiority trials, 1669
- clinical trial simulation (CTS), 536–537, 539–543
 - $A + B$ escalation with dose deescalation, 540
 - $A + B$ escalation without dose deescalation, 540
 - applications and some tools, 537
 - continual reassessment method (CRM), 541
 - simulation example using, 542–543
- dose-escalation algorithms, evaluation of, 541
- dose-escalation trial, 539, 542–543
- dose-level selection, 539
- dose-limited toxicity and maximum tolerated dose, 539
- dose-toxicity modeling, 541–542
- for later development phases, 544–551
 - adaptation rules, 545
 - adaptive design, 544–545
 - adaptive design permitting early stopping and sample size reestimation (example), 549
 - adaptive design with dropping losers (example), 550
 - α -adjustment, 547–548
 - bias due to adaptive randomization RPW (1,1), 548
 - conventional design with multiple treatment groups (example), 549
 - conventional design with two parallel treatment groups (example), 548
 - design with RPW randomization (example), 548
 - discussion of simulation results, 550
 - dose-response model, 545
 - dose response trial design (example), 550
 - early stopping rules, 546
 - flexible design with sample size reestimation (example), 548
 - group sequential design with one interim analysis (example), 549
 - MUM, 547
 - null model vs. model approach, 547
 - randomization rules, 545–546
 - response-adaptive design with multiple treatment groups (example), 549–550
 - response-adaptive randomizations, 546
 - RPW, 546
 - rules for dropping losers, 546
 - sample size adjustment, 546
 - simulation examples, 548
 - test statistic, 547
 - UOM, 547
 - utility-based trial objective, 545
- likelihood function, 542
- next patient, assignment of, 542
- objective of, 536
- parameter, assessment of, 542
 - Bayesian approach, 542
 - least square approach, 542
 - maximum likelihood approach, 542
- parameter, prior distribution of, 542
- phase I clinical trial, objectives of, 539
- population for treatment, 539
- in regulatory decisions, 537
- sample size per dose level, 540
- simulation example using CRM, 542–543
- simulation example using TER and TSER, 541
- traditional dose-escalation design, 540
- two-stage dose-escalation design, 540–541
- clinical trial validation, 822, 823, 2208
- closed-form sample size formula, 2033
- closed testing procedure, 732–733, 1122
- closeness, 1096
- clozapine, 191
- cluster analysis and pattern recognition, 1386
- dimension reduction analysis and visualization, 1389–1390
- gene expression matrix map with hierarchical clustering, 1386–1388
- gene network modeling and integrated system biology, 1390
- non-hierarchical clustering, 1388–1389
- clustered binary data, 82–86
 - analysis methods, 82–84
 - contingency table for, 83, 84
 - example, 85
 - general estimating equations, 85
 - hypothesis testing, 83–85
- clustered categorical data, 88–96
 - analysis of clustered multivariate ordinal models, 92–94
 - Dale's model for binary set up, 92–93
 - latent mixed model, 93–94
 - discussion, 96
 - illustrative example, 94–95
 - oral hygiene data, 95
 - simulation study, 94–95
 - tibolone data, 94
 - methods for, 91–92
 - Bayesian approach, 91
 - GEE approach, 92
 - likelihood based approach, 91
 - models for, 88–91
 - latent mixed models, 89
 - multivariate clustered ordinal model, 90–91
 - proportional odds mixed models, 89
- clustered multivariate ordinal models, 92–94
 - Dale's model for binary set up, 92–93
 - latent mixed model, 93–94
- clustering, 226
- cluster trials, 648–656
 - adjusted Chi-square test for individual-level data, 651
 - cancer screening data, analysis of, 653–654
 - cluster-level analysis, 650–651
 - common designs, 650
 - conducting, 649
 - covariates, adjusting for, 655
 - finite sample performance, 655
 - individual-level analysis with mixed effects models, 652–653
 - individual-level analysis with population-averaged models, 651–652
 - multiple time points, 655
 - non-randomized trials, 655
 - in pharmaceutical studies, 655
 - randomized, 648
 - sample size, determining, 653
 - statistical analysis of, 650
- cluster-unit trials, 648
- cocaine project, 800–801
- Cochran–Armitage–Mantel trend test, 79
- Cochran–Armitage test, 406, 1934, 634
 - for determining drug effect, 880
- Cochran–Mantel–Haenszel (CMH) test, 442, 1321, 825
 - pharmacodynamics with covariates, 1682
- Cochran's Q statistic, 1324

- Code of Federal Regulations (CFR), 233, 1095
 - 21 CFR Part 11, 444
 - 21 CFR Part 312, 444
 - 21 CFR Part 314, 445
- coding dictionaries
 - adverse event reporting, 54–55
 - common dictionaries, utilization, 54
 - dictionary versions, 55
 - Medical Dictionary for Regulatory Activities, 54–55
 - need to code, 54
- coefficient of variation (CV), 486, 1143, 2316
- coefficient of variation of the deviation from linearity (CVDL), 146, 148–149
- Cognitive Interview Technique, 1169
- Cognitive Laboratory, 1169
- Collaborative Study on the Genetics of Alcoholism (COGA) project, 1062
- collinearity, test for, 1311–1312
- colloquialisms, 1168
- comarketing, 717
- combination drug clinical trial, 552–554
 - fixed-dose combination, 552–553
 - multiple-dose combinations, 553–554
 - statistical designs and analysis methods, 552
- therapeutic synergism, 554
- combination-treatment studies, dose response analysis in clinical trials, 882
- combination vaccines, 238
 - study, 635, 637
- combined penalty (CP) function, 1450–1451
- Commission E, 2262
- Committee for Medicinal Products for Human Use (CHMP), 938, 1328, 285
- Committee for Proprietary Medicinal Products (CPMP), 438, 1095
- COMMIT trial, 649, 650, 650, 651
- Common Terminology Criteria (CTC), 539
- communality and unique components, exploratory factor analysis, 986
- communication, ethnic factors, 978
- community trials, 648
- companion diagnostic devices, 827
- comparative fit index (CFI), 1173
- comparative genomics, 226
- compendial monograph tests and standards, 259–263
 - aerosol droplet size and distribution, 260
 - content uniformity, 259–260
 - disintegration, 260–261
 - dissolution, 261–263
 - particle size, 260
- competing risks, 2182
- complete clinical data package (CCDP), 2279
- complete design, 926
- completely randomized design, 1623
 - comparing two treatments, 2168
 - covariate, role of, 2170
 - individual treatment effects and potential outcomes, 2168
 - model for binary outcomes and consequences of S-T interaction, 2169–2170
 - model for continuous outcomes and consequences of S-T interaction, 2169
- complete *N*-of-1 design
 - with four periods, 566
 - vs. RRRR/RTRT design, 567–569
 - with three periods, 565–566
 - vs. 2 × 3 dual design, 569–570
 - with two periods, 565
 - vs. 2 × 2 crossover design, 569
- complete null hypothesis, 728
- complete randomization, 1876
- complete response (CR), 2296
- “completers” subgroup analysis, 903
- compliance, patient, 1627–1631
 - effect of poor compliance on sample size requirements in clinical trials, 1629
 - effect of poor compliance on the clinical status of patients, 1629–1630
 - improving compliance in clinical trials, 1630–1631
 - magnitude and pattern of poor compliance, 1628
 - measuring, 1627–1628
- Compliance Program Guidance Manual, 1076
- composite centers, 439–440
- compositional data, 1821
- compounding symmetry (CS), 2012, 2213
- compound items, 1168
- compression, 258
- compulsive drug seeking, 320
- computational algorithm
 - of double threshold design, 158
 - of single threshold design, 156–157
- Computational Systems Biology conference (CSB), 224
- computer assisted diagnostic devices (CAD), 1355
- computer-assisted trial design (CATD), clinical trial simulation, 536
- computer-generated designs, 1944–1945
- computerized adaptive testing (CAT), 1214
- concentration-QTc (C-QTc) modeling, 2206, 2207, 2212, 2213
- concentration–response study, 887
- concordance correlation, 1910–1911
- concordance correlation coefficient (CCC), 164
 - inter-CCC, 170
 - intra-CCC, 170
 - measuring agreement, 1337
- concurrent controls (CCTs), 572
- concurrent validity, 1175
- conditional across-trial type I error rate, 1540–1541
- conditional and unconditional methods, 1904
 - conditional rule, release limits based on, 1904–1905
 - decision rules, 1904
 - conditional rule, 1904
 - unconditional rule, 1904
 - expected loss function approach, 1906
 - unconditional rule, release limits based on, 1905–1906
- conditional autoregressive (CAR) specification, 2090
- conditional error function approach, 1574
- conditional error probability, 1574
- conditional exact approach, 2284–2285
- conditional random effects model, 490–491
 - test for equality, 491
 - test for noninferiority/superiority, 491
 - test for similarity, 491
- conditional type I error function, 2023
- condition numbers, 1444
- confidence bound approach, 516–517
- confidence interval (CI), 432, 699, 929, 1323–1324, 1914, 2032
 - approaches, 293, 972
 - construction of, 699–700
 - definitions of, 699
 - finished product, 1811
 - following a group sequential test, 1110
 - process validation, 1805
 - testing procedure, 16, 18–19
- confidence limits and bands, 1233
- confidence sets, 344–345
- confirmatory factor analysis (CFA), 1008, 1009, 1167, 1172–1173
 - example, 1010–1012
 - extensions, 1010
- confounding, 703–704, 707
 - controlling through data analysis, 705–706
 - controlling through design, 704–705
 - evaluating, 703–704
 - regulatory requirements, 707
- congestive heart failure, 825, 1666
- conjugacy, 194
- conjugate prior distribution, 176
- consistency index, 307
 - assessment of, 1513–1514
- consistency lots study, 634–635, 636
- consistency requirement, 1579, 1580
- Consortium of Innovation and Quality in Pharmaceutical Development (IQ), 2215
- constrained full likelihood LDA model (cLDA), 1624
- construct validity, 1175–1176
- consumer's risk, 2, 1097
- content uniformity (CU), 709–711, 1807
 - USP tests, 259–260
- content validity, 1168
- content validity index (CVI), 1168
- contingency table approach, 2227
- continual reassessment method (CRM), 387–388, 545, 1561
 - continual reassessment method and modifications, in phase I cancer clinical trials, 391–392
 - early phases of clinical trials, simulation in, 541
 - in phase I cancer clinical trials, 391–392, 1717–1718
 - simulation example using, 542–543
- continuous outcomes, 613, 615
 - two-treatment three-period designs, 615–617
 - two-treatment two-period designs, 615
- continuous outcome variable case, 2011, 2012
- continuous work process improvement, 534
- contract research organization (CRO)
 - industry, 716–719, 1076, 812, 1084, 721–724
- emergence of, 717
 - in public market, 717

- contract research organization (CRO)
 - industry, (*Cont.*)
 - research and development spending and duration of marketing exclusivity, 717–718
- future, in pacific rim, 718
 - critical issues to ensure the quality of clinical trials, 719
 - market size for pharmaceuticals, 718
 - traditional contract research organization services, extension of, 719
- growth of, 718
 - full-service contract research organizations, growth of, 718
 - increase of outsourcing by pharmaceutical and biotechnology companies, 718
 - upstream and downstream growth, 718
- industry background, 716
 - alternate approaches to pharmaceutical research and development, 717
 - pressures on pharmaceutical industry from early 1990s, 716–717
 - role of the biostatistician within, 722–724
- contract sales organizations (CSOs), 719
- contrast drug products, diagnostic imaging, 829–830
- contrasts, 879
 - and trend tests, group sequential procedures, 1118–1119
- control charts, 2137–2140
- control effect, 16f
- controlled gene-expression method, 932
- Controlled Substances Act (CSA), 1122
- control rate, therapeutic trials, meta-analysis of, 1370
- control selection in clinical trials, 622–626
 - assay sensitivity, historical evidence of sensitivity-to-drug-effects, and noninferiority margin, 624–625
 - selection of concurrent control, 626
 - types of controls, 622
 - active concurrent control, 623
 - dose–response concurrent control, 623
 - external (historical) control, 623
 - no-treatment concurrent control, 622
 - placebo concurrent control, 623
 - useful alternatives and design modifications, 625–626
- control treatment, 318
- conventional “all-comer design”, 964
- conventional design with multiple treatment groups
 - clinical trial simulations for later development phases, 549
- convergent validity, 1176
- Cook’s approach, 1295
- Cook’s distance, 506, 508f
- coordinating center (CC), 576
- Copy Number Variants (CNVs), 228, 1444
- coronary artery disease/angina, 1666
- Coronary Drug Project (CDP) Research Group, 575
- coronary heart disease (CHD) mortality, 1629
- corrective and preventive action (CAPA) plan, 462
- correlated measures, percentile charts on, 1643–1647
 - Box-Cox transformation, 1643
 - correlated response variables, 1644–1645
 - example, 1645–1646
 - quantile regression, 1644
- correlated probit model, 741–749
 - correlated probit model, 743–744
 - developmental toxicity application, 746–747
 - estimation, types of, 746
 - implied correlation structure, 744
 - inference, 746
 - interpretation, 744
 - marginal maximum likelihood estimation via adaptive Gaussian quadrature, 745–746
 - maximum likelihood estimation, general approaches for, 744–745
 - multivariate probit model, 742
 - probit model for uncorrelated data, 741–742
 - relationship to other models, 744
 - software, 748
- correlated random effects, 1418
- correlation coefficient (ICC), 1335
- correlation matrix, 85
 - principal components based on, 1799
- correlations, 1444
 - structure, 2013
- corticosteroid guidelines, 266, 268
- costart, 54
- cost-benefit analysis (CBA), 754, 1695, 1707
- cost-consequences analysis, 1695
- cost-effectiveness (CE), 525
 - CE acceptability curve, 753
 - CE plane, 752
- cost-effectiveness analysis (CEA), 751–758, 1695, 1696, 1707, 1708–1709
 - alternatives, 1710
 - average ratio vs. incremental ratio, 1710–1713
 - independent programs, 1711
 - mutually exclusive programs, 1711–1713
- controversies, 756
 - lambda, silence of, 757
 - one vs. multiple methods and analyses, 757
 - perspective, 757
 - quality-adjusted life year (QALY), 756–757
- design, methods, and statistics-related issues, 753
 - budget impact analysis (BIA), 756
 - censoring, 755
 - discounting, 754
 - modeling and sensitivity analyses, 755–756
 - sample size, 756
 - skewness, 754
 - study design, 754
 - types of CEA, 753–754
- economic evaluation, types of, 1707–1708
- meta-analysis in, 1368
- performing, key concepts in, 1709
- statistical measures, 752
 - average cost-effectiveness ratio (ACER), 753
- cost-effectiveness (CE) acceptability curve, 753
- cost-effectiveness plane, 752
- incremental cost-effectiveness ratio (ICER), 752
 - incremental net benefit (INB), 753
- cost-effectiveness ratio (C/E), 1708, 1711
- costing methods, 1695–1696
- cost-of-illness (COI) study, 1708
- cost-utility analysis (CUA), 1695, 1707
- Council for International Organizations of Medical Sciences (CIOMS), 1345
- counseling, 240
- covariance, 772–775
 - adjusting for baseline values, 772–773
 - choice of covariates, 773–774
 - interaction, 774–775
 - nonlinear models, 774
 - other covariates, 773
- covariance matrix, 1927
- covariance structure, 2214
- covariate-adjusted adaptive dose-finding, 760, 761
 - dose acceptability, 761–762
 - dose desirability, 762
 - illustration, 763–767
 - probability model, 760
 - prior specification, 761
 - regression models for efficacy and toxicity, 760–761
 - trial conduct, 763
- covariate-adjusted response-adaptive (CARA), 1727
- covariate balance assessment, 1794, 1826
- covariates, 1731
 - adjustment for, 772–775
 - cluster trials, 655
 - control, ecologic inference, 944–945
 - pharmacodynamics with, 1673–1684
 - between-treatment-groups analysis, 1679–1680
 - Cochran-Mantel-Haenszel test, 1682
 - Cox proportional hazards analysis, 1683–1684
 - crossover design, 1676–1677
 - logistic regression, 1682–1683
 - one-way analysis of covariance, 1681–1682
 - one-way analysis of variance with repeated measures, 1677–1678
 - simple linear regression, 1678
 - two-way analysis of variance, 1673–1676
 - within-treatment-group analysis, 1678–1679
- coverage probability (CP), 954, 2213
 - measuring agreement, 1338
- Cox-Aalen model, 594, 2060f, 2061
- Cox model, 557, 1835, 1836
- Cox proportional hazards analysis, 559, 561, 1494, 1496, 1591, 1684
 - pharmacodynamics with covariates, 1683–1684
 - regression model, 395–396
- Cox regression model, 425, 2058, 2061f
- Cramér–von Mises-type test statistics, 1235, 1590
- credible intervals, 169, 199, 294

- credible sets, 169
 - Critical Path Initiative, 537
 - critical quality attributes (CQAs), 116, 2219
 - Cronbach's α coefficient, 1176
 - cross-level bias, 946
 - cross-level confounding, 946
 - crossover design, 206, 776
 - baselines, 781
 - Bayesian approaches, 781
 - dose response analysis in clinical trials, 882
 - general linear model, analysis using, 777–778
 - hybrid designs, 781–782
 - modeling carryover, 779
 - pharmacodynamics with covariates, 1676–1677
 - testing for carryover, 780
 - cross-over trial, 776
 - crystal properties, 255
 - cubic logistic model, effective doses, 956, 957
 - cumulative case-control sampling, 423
 - cumulative dosing, 1721
 - cumulative hazard rate, 423
 - cumulative incidence function, 2182
 - estimating, 1498
 - cumulative meta-analysis
 - vs. expert opinion, 1368
 - of therapeutic trials, 1371
 - current good manufacturing practice (cGMP), 1, 370, 1095, 1101
 - curse of dimensionality, 227
 - curtailment, 3
 - curves, analyzing sample of, 1546
 - dynamic time warping, 1547–1549
 - shape-invariant modeling, 1547
 - structural analysis, 1547
 - cutoff designs, 798
 - cocaine project, 800–801
 - combining RD and RE designs, 800
 - functional form, specifying, 800
 - modeling and analyzing, 801–802
 - regression-discontinuity (RD) design, 798–799
 - validity of, 799
 - relative sample sizes needed in, 802
 - cytochalasin B blocks cell division, 1155
 - cytogenetic tests in vivo, 2234
 - cytokine products, 237
 - cytosine arabinoside, 763
- D**
- DAKO C-KIT pharmDx, 1651
 - Dale's model, for binary set up
 - analysis of, 92–93
 - Bayesian analysis, 92–93
 - classical likelihood based analysis, 92
 - Dali, structural alignment methods, 228
 - data accrual, 1792
 - data adequacy, 1792
 - data analyses
 - clinical trials, 577
 - controlling confounding through, 705–706
 - Data- and Safety-Monitoring Committee (DSMC), 573, 576, 577
 - data assurance, 1792
 - database development process, 446, 447
 - data collection principles affecting analyses, 1345
 - data components, global database and system, 1071
 - data-conversion issues, laboratory analyses, 1262
 - data coordinators, 446
 - data-driven methods, 524
 - data gathering, 534
 - data-handling process, clinical data
 - management guidelines, 448
 - data-integration issues, laboratory analyses, 1262
 - data management plan (DMP), 447
 - data management process traditional data
 - collection, 448
 - data management software applications, 449
 - data mining and biopharmaceutical research, 805
 - artificial neural networks, 808–809
 - data mining and polypharmacy (case study), 809
 - data-mining methods, 806
 - data-mining process, 805
 - genetic programming, 807–808
 - kernel density estimation, 806
 - link analysis, 806
 - rule induction, 807
 - text analysis, 807
 - data mining and polypharmacy (case study), 809
 - data monitoring, 60
 - data monitoring committee (DMC), 580, 811–820, 1109–1110
 - bias, 816–817
 - at the analysis level, 817
 - reporting, 817
 - treatment assignment, knowledge of, 817
 - challenge of the independence of, 586–587
 - clinical issues, 814
 - goals of safety analysis, 814
 - multiregional (global) trials, clinical issues in, 815
 - safety data, 815
 - decisions, 817
 - decision-making environment, 818
 - final meeting, 818–820
 - interim analysis for infant pharma sponsors, 819
 - planned interim analyses regarding efficacy, 818
 - types of DMC decisions, 817–818
 - emerging issues, 819
 - adaptive designs, 819
 - biosimilar designs, 819–820
 - centralized risk-based monitoring, 820
 - fraud detection, 820
 - mergers and licensing, 820
 - novel designs in oncology, 819
 - training of DMC members, 820
 - final meeting, 818–820
 - history, 811–812
 - interim analysis for infant pharma
 - sponsors, 819
 - meetings, 813
 - data review meetings, 814
 - DMC Charter, 813
 - orientation meeting, 813–814
 - patient protection, 812
 - place of DMCs in the development cycle, 812
 - planned interim analyses regarding efficacy, 818
 - when safety issue arises, 818
 - safety monitoring program in confirmatory trials, 812
 - conflicts of interest, 813
 - DMC creation, 813
 - liability and indemnification, 813
 - members of the safety monitoring team, 812–813
 - sponsor size, 812
 - statistical issues, 815
 - goal of statistical analysis, 815
 - hypothesis testing, 816
 - statistical description, 816
 - useful data displays, 815–816
 - data preprocessing and normalization, 1381–1382
 - DCLHb, 1338
 - D-dimer/fibrinogen ratio, 825
 - death-rate method, 403
 - decision makers, 473
 - decision making, 534
 - Declaration of Helsinki, 1075–1076
 - deep-vein thrombosis (DVT), 1368
 - degree of closeness, 1096
 - Delphi interview method, 1772
 - delta limits, Lilly Reference Ranges, 1277
 - density case-control, 425, 426, 430–432
 - sampling design, 424
 - density estimation, non-parametric regression, 1551
 - dermal irritation, testing for, 2231
 - Derogatis Interview for Sexual Function-Self Report® (DISF-SR21), 1175
 - DerSimonian-Laird's method of moments (MM), 605
 - descriptive tools, 164
 - design
 - allocation strategy, 612–613
 - controlling confounding through, 704–705
 - driven individual bioequivalence criteria, 1253–1254
 - Design BD, 387
 - design effect (DE), 648, 993
 - “design to progress”, 1725
 - “design to stop”, 1725
 - detailing, 1089
 - detectors, 131, 267
 - determinant, 1444
 - deterrent effect, test for, 1122
 - developmental toxicity application, correlated probit model, 746–747
 - development cycle, place of DMCs in, 812
 - deviance information criterion (DIC), 170, 200
 - deviation from linearity at concentration, 148
 - device performance, 822
 - device registries, 1792
 - diabetes, 1667
 - diagnostic accuracy, 822, 824, 827
 - Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV), 1267

- diagnostic device evaluation, 822
 - clinical performance, 824
 - design theory for diagnostic tests, 826
 - measurement validation, 823
 - training and validation, 827
 - diagnostic devices, 1353
 - analysis issues for, 1354
 - diagnostic design issues, 1353–1354
 - measurement method comparison for
 - diagnostic tests, 1354–1355
 - diagnostic imaging, 829–835
 - assessing reader agreement for blinded-read trial, 832–833
 - bias considerations, 831–832
 - for a blinded-read trial, 832
 - for an open-label trial, 831
 - blinded-read component, 831
 - clinical component, 830–831
 - contrast drug products, 829–830
 - devices, 823
 - diagnostic performance measures, 833–834
 - diagnostic radiopharmaceuticals, 829
 - regulatory requirements, 829
 - study design considerations, 830
 - Diagnostic Interview Schedule (DIS), 1267
 - diagnostic likelihood ratio, 826
 - diagnostic procedure, 836–837
 - practice, remarks on, 837
 - sensitivity and specificity, 836–837
 - diagnostic radiopharmaceuticals, 829
 - diagnostic tool kit, 1444
 - diastolic blood pressure (DBP), 1909
 - dichotomous outcomes, 613–614, 618
 - two-treatment three-period designs, 619
 - two-treatment two-period designs, 618–619
 - diethylhexyl phthalate (DEHP), 1935
 - differentially expressed genes, identifying, 1382
 - assigning a significance level, 1383–1384
 - gene ranking, 1382
 - dimensionality, curse of, 227
 - dimension reduction analysis and visualization
 - microarray gene expression, 1389–1390
 - direct assays, 213
 - direct-by-carryover interaction, 2210
 - direct costs, 1695
 - direct data capture devices, clinical data management guidelines, 447
 - direct method, expiration dating period, 981
 - Dirichlet process, 92
 - disaggregated criteria for similarity, 306
 - discounting, 754, 1697
 - to present values, 1697
 - discrete-continuous estimation, mathematica program for, 1775
 - discrete wavelet transform (DWT), 1546
 - discriminant validity, 1175
 - discrimination ability, 1654
 - disease modification effects, 839–842
 - design considerations, 839–842
 - parallel group designs, 842
 - randomized start/withdrawal designs, 839–841
 - methods for data analysis, 842
 - slope analysis, 842
 - survival analysis, 842
 - disintegrating agents, 257
 - disintegration test, 260–261
 - dispersion parameter, 993, 2032, 2033, 2034
 - treatment-specific, 2036
 - dissimilarity index as an efficacy measure, 2265–2266
 - dissociation constant, 255
 - dissolution, 261–263, 270
 - Dissolution of Semisolid Dosage Forms, 261
 - dissolution profiles
 - measures of difference/similarity between
 - two dissolution profiles, 1148–1150
 - similarity test based on similarity factor, 1150
 - bias of the estimate of similarity factor, 1151
 - estimation and confidence interval of similarity factor, 1151
 - theoretical considerations, 1150–1151
 - similarity testing, 1150
 - dissolution profiles comparison, 854–858
 - discussion, 857–858
 - model-independent MSD methods for
 - dissolution profile comparison, 855
 - Hotelling's T^2 -based approaches, 855
 - Sarandasa and Krishnamoorthy's (SK) method, 855–856
 - similarity factor for, 291–292, 307
 - simulation study, 856
 - distance function, 226
 - distribution-free random effects model
 - with using a conditional approach, 1618–1619
 - divergent validity, 1176
 - Divine Providence, 1845
 - Division of Biologics Standards (DBS), 233
 - DNA (deoxyribonucleic acid), 1057, 1377
 - indicator tests, 2234
 - microarray experiment, 1377, 1378f
 - documentation in programming, 1092
 - domestic violence, 320
 - D-optimality, 1945
 - dosage forms, 258–259
 - capsules, 259
 - nonoral dosage, 259
 - solutions, 258
 - suspensions, 259
 - tablets, 259
 - dose acceptability, covariate-adjusted adaptive
 - dose-finding, 761–762
 - dose desirability, covariate-adjusted adaptive
 - dose-finding, 762
 - dose-dumping, 265
 - dose-effect information, 552, 554
 - dose-escalation algorithms, evaluation of
 - early phases of clinical trials, simulation in, 541
 - dose-escalation trial
 - early phases of clinical trials, simulation in, 539, 542–543
 - dose–exposure–response model, 1331
 - Dose Finding R package, 1332
 - dose independence, *see* dose proportionality (DP)
 - dose-level selection
 - early phases of clinical trials, simulation in, 539
 - dose-limiting toxicity (DLT), 1561, 1715
 - early phases of clinical trials, simulation in, 539
 - interval designs, 1563
 - Bayesian model, 1563
 - interval definitions, 1563
 - modified toxicity probability interval, 1563
 - unit probability mass, 1563
 - model-based designs, 1561
 - Bayesian logistic regression model, 1562–1563
 - rule-based designs, 1561
 - dose proportionality (DP), 876–878
 - clinical pharmacology, 470
 - design of dose-proportionality study, 878
 - slope approach for assessment of, 2084–2088
 - discussion, 2088
 - example, 2087
 - sample size calculation, 2086–2087
 - statistical model and slope approach, 2085
 - test for departure from linearity, 2087
 - tests for, 2085–2086
 - statistical models for assessing, 876–878
 - analysis of variance, 877
 - linear regression, 876–877
 - power model, 877–878
 - dose response analysis in clinical trials, 879–882
 - combination-treatment studies, 882
 - drug effect, tests for determining, 879–880
 - Bartholomew's test, 880
 - Cochran–Armitage test, 880
 - contrasts, 879
 - Jonckheere's test, 880
 - linear trend, 879
 - overall F-test, 879
 - William's test, 879
 - effective doses, 880–881
 - Bonferroni adjustment, 880
 - Bonferroni–Holm procedure, 881
 - Dunnett's procedure, 881
 - Fisher's least significant difference, 880–881
 - fixed sequential testing procedure, 881
 - Hochberg sequential procedure, 881
 - minimally effective dose and maximum tolerated dose, 880
 - mixed-effects models, 880
 - nonparametric regression, 880
 - orthogonal polynomials, 880
 - sigmoid models, 880
 - ordinal endpoints, 882
 - regulatory guidelines, 879
 - safety endpoints, 882
 - trial design, analysis considerations based on, 881–882
 - crossover designs, 882
 - parallel group designs, 882
 - titration designs, 882
- dose–response concurrent control, 623
- dose–response curve, 1326, 1327, 1328
 - characterization of, 1398
- dose–response data, combining, 2286
 - meta-analysis for trend estimation, 2286

- dose–response estimation, 1326
- dose response immunogenicity study, 634
- dose–response information, 1330
- dose–response model, 1326, 1327, 1330, 1331
 - clinical trial simulations for later development phases, 545
 - parameter, 1330
- dose–response signal, 1326, 1328, 1330
 - testing, 1328
- dose–response study design, 885–889
 - dose–selection process for a new drug, 885–886
 - drug/disease specific considerations, 888–889
 - drugs developed for prevention, 889
 - drugs for special patient population, 889
 - drugs need new formulation, 889
 - drugs with narrow therapeutic window, 888
 - life-threatening diseases, 888–889
 - general design considerations, 886–888
 - dose allocation, dose spacing, 888
 - fixed-dose vs. dose-titration designs, 887
 - number of doses to be tested, 887
 - optimal designs, 888
 - sample size considerations, 888
 - selection of control, 887
- dose–response testing, 1326
- dose–response trial design
 - clinical trial simulations for later development phases, 550
- dose spacing, 1331
- dose titration study, 2225
- dose–toxicity modeling, 542
 - early phases of clinical trials, simulation in, 541–542
- double blind trials, 336, 573
 - examples, 337
- double data entry with blind/interactive verification
 - clinical data management guidelines, 447
- double sampling acceptance plan, 3–4
- doubly adaptive biased coin design (DBCD), 51, 1731, 1951–1952
- Down syndrome, 1057
- Draft Guidance documents, 1792
- droplet size distribution, 1142
 - cascade impaction technique, 1142
 - laser diffraction, 1142
- dropout process, 901–906
 - analysis populations and accompanying methods, 903
 - Diggle and Kenward, 904–905
 - extensions and software, 906
 - Wu and Bailey, 905–906
- drop-the-loser (DL)
 - allocation, 1729
 - design, 1573
 - rule, 1950–1952
- drug abuse
 - addition of, 316
 - potential, 1123
- drug approval process, 53
- drug characteristics for different dosage forms, 980
- drug development, 544, 908–913
 - nonclinical development, 909–910
 - drug formulation development, 910
 - pharmacology, 909–910
 - toxicology/drug safety, 910
 - and pharmacogenetics, 240–241
 - postmarketing clinical development, 912–913
 - postmarketing phase of, 772
 - premarketing clinical development, 911–913
 - new drug application, 912
 - phase 1 clinical trials, 911–912
 - phase 2/3 clinical trials, 912
 - process, 1572
 - regulatory requirements, 908–909
 - drug–drug interaction, 1332, 1723
 - drug effect, tests for determining, 879–880
 - Bartholomew’s test, 880
 - Cochran–Armitage test, 880
 - contrasts, 879
 - Jonckheere’s test, 880
 - linear trend, 879
 - overall F-test, 879
 - William’s test, 879
 - Drug Efficacy Study Implementation (DESI)
 - drugs, 1098
 - drug formulation, 909
 - drug-high, 1123
 - drug interaction
 - in a clinical study (example), 1026
 - studies, clinical pharmacology, 470
 - drug interchangeability, 923, 929, 1432
 - alternative criterion based on reverse of test and reference product, 925
 - criteria for PBE and IBE, 923
 - individual bioequivalence, 923
 - population bioequivalence, 923
 - criterion for, 923–924
 - design for, 925
 - adaptive designs, 926
 - alternative designs, 926
 - Balaam’s design, 925–926
 - complete design, 926
 - modified Balaam’s design, 926
 - two-sequence dual design, 926
 - for generic products, 935–938
 - average bioequivalence, 936
 - individual bioequivalence (IBE) for switchability, 937
 - population bioequivalence (PBE) for prescribability, 936–937
 - power analysis for sample size, 926
 - sample size determination, 927
 - testing for PBE/IBE, 926
 - safety margins for monitoring, 931
 - sample size, 927
 - scaled criterion for, 918–924
 - “drug lag” problem, 1573
 - drug-liking, 1123
 - drug management system, 1216
 - drug planning and development, therapeutic trials, meta-analysis of, 1370
 - drug prescribability, 929
 - Drug Price Competition and Patent Term Restoration Act, 304, 313, 928, 2199
 - drug product, 489
 - drug substance, stability of, 125
 - drug supply and packaging, 1218
 - drug switchability, 929, 1160
 - drug-taking again, 1123
 - DSMB (Data and Safety Monitoring Board), 60
 - dual control groups, 409–410
 - dual threshold design (DTD), 157–158
 - prior distribution, computational algorithm, and sample sizes, 158
 - Dunnett’s procedure, 731
 - dose response analysis in clinical trials, 881
 - dynamic time warping, 1547–1549

E

 - early stopping rules, clinical trial simulations for later development phases, 546
 - East (software), 1199
 - Eastern Cooperative Oncology Group (ECOG) performance status, 1639
 - ECG records, 2207
 - ECG tracing, 2206
 - ECG waves, 2206
 - ecological analysis, 940
 - ecological regression model, 947
 - ecologic inference, 940–950
 - aggregate-level confounding due to individual-level effects, 941–943
 - aggregation level, problems in, 949
 - assumptions and fallacies in ecological analyses, 943
 - basic model, extensions of, 946–947
 - causal interpretation, problems in, 949
 - covariate control, 944–945
 - ecological regression, 947
 - ecological vs. individual-level sources of BIAS, 941
 - measurement error, 948
 - modeling assumptions, 945
 - multi-level model theory for, 945
 - multi-level studies, 947–948
 - non-comparability among ecological analysis variables, 948
 - observational studies, 944
 - randomization assumptions, 943–944
 - special cases, 946
 - special problems of ecological analysis, 948
 - economic evaluation
 - cost-effectiveness analysis, 1707–1708
 - importance of, 1693–1694
 - effective doses, 952–959
 - interval estimation, four methods of, 953–954
 - numerical experiments, 954–955
 - robustness of interval estimators of ED_{50} , 955
 - asymmetric Aranda–Ordaz model, 956
 - cubic logistic model, 956
 - probit model, 955
 - robustness of interval estimators of ED_{90} , 957
 - asymmetric Aranda–Ordaz model, 957
 - cubic logistic model, 957
 - probit model, 957
 - statistical model, 953
 - effective sample size (ESS), 190, 761
 - effect modification, 706
 - effect-size calculation, 608–609
 - effects table, 475, 481
 - effects tree diagrams, 481

- efficacy information, integration of, 1180f
- efficacy, integrated summary of, 1179
- dose rationale, 1179–1180
 - efficacy data, 1180
 - meta-analysis, 1180–1181
- efficacy lower bounding function, 761
- efficacy measurements, 416
- efficacy response, 1723
- efficient development program, 1724
- EFSPi Benefit-Risk Special Interest Group and QSPI Benefit-Risk Working Group, 475
- EGFR tyrosine kinase inhibitors, 825
- 80/125 decision rule, 929
- elastic-net penalty, 1449
- elderly population, pharmacodynamic issues, 1670
- electrocardiogram, 824
- electroencephalogram (EEG), as a quantitative endophenotype, 1062–1064
- electronic data collection system, 446
- clinical data management guidelines, 447
- electronic records, clinical data management guidelines, 444
- electronic signatures, clinical data management guidelines, 444
- electrophoresis methods, 227
- Electrophysiologic Study vs. Electrocardiographic Monitoring (ESVEM), 961, 962
- elicitation process, 193
- Emax model, 1328, 1331
- EM iterative convex minorant (EMICM) algorithm, 591
- Empirical Bayes (EB) estimates, 2107
- empirical Bayes methods, 2091
- empirical log-odds ratio, trend estimation by using, 2285
- empirical power curve, 2190f
- employment loss, 320
- endocrine system, 1667
- diabetes, 1667
 - osteoporosis, 1667
- End-of-Phase-2A (EOP2A) pilot program, 536, 537
- endometrial hyperplasia (EH), 730
- endorsement frequency, 1170
- endpoint adjudication committee, 812
- endpoints, 1467
- clinical strategy for selection of, 583
 - selection and assessment of, 574
- enriched population, 964
- enrichment decision-making, 965
- enrichment design, 960–962, 1572, 964
- analysis aspect, 965
 - design aspect, 964
 - draft guidance on, 965
 - examples, 960–962
 - methods, 965, 966
 - overall treatment effect in all-comer population, 966
 - trial design and sample size considerations, 967
 - weighted estimator of overall treatment effect, 967
 - with patient population augmentation, 964
- entropy penalty function, 1449
- enzyme immunoassay (EIA), 131
- EP6-A's criterion, 147–148
- Epanechnikov kernel, 1545
- epidemiologic models, 1697
- epidermal growth factor receptor (EGFR), 825
- epileptics, 2031
- equal precision (EP) bands, 1233
- equipotence, 135
- equivalence, 525
- margin, determination of, 2221–2223
- equivalence acceptance criterion (EAC), 118
- equivalence trials, 970–974, 2031, 2033
- based on two-arm parallel design, 274–275
 - confidence interval approaches, 972
 - equivalence limits, 972–973
 - hypotheses in, 1126
 - population bioequivalence, 973–974
 - regulatory concerns about the conduct of, 1669
- Erbix (cetuximab), 1648, 1651
- error criteria for multiple comparisons, 1028
- definition, 1028–1029
- error spending function, *see also* alpha spending function
- types of, 61–62
- erythropoietin, 237
- escalation with overdose control (EWOC), 1719
- Escherichia coli*, 68
- estimated parameter-based allocation using link functions, 1729
- ethics-based target allocation functions, 1730
- link function–based design for continuous responses, 1729–1730
- target allocation functions
- combining ethics and precision measures, 1730–1731
 - implementing, 1731
- estimated regional treatment effect, 1578
- estimating equations (GEE), 1046, 1420
- estimating function, 1046
- estimation
- vs. TOST method, 151
 - types of, 746
- estimators of SSdL, ADL, and CVDL, 149
- ethical and scientific quality of clinical research, 1075
- Declaration of Helsinki, 1075–1076
 - quality assurance and quality control, 1076
- ethics-based target allocation functions, 1730
- Ethics Committee, 812
- ethnic factors, 976–978
- defining, 976
 - in designing safe and effective trials, 977
 - in foreign populations time, 978
 - budget, 978
 - communication, 978
 - stigmatization, 978
 - implications for human subjects review, 977–978
 - implications for those conducting clinical trials in foreign regions, 977
- ethnicity, 976
- European Agency for the Evaluation of Medicinal Products (EMA), 585
- European Bioinformatics Institute (EBI), 229
- European Commission of the European Union, 1202
- European Community (EU) Pharmaceutical Review, 308
- European Federation of Pharmaceutical Industries and Associations (EFPIA), 474, 1202
- European Medicinal Evaluation Agency (EMA), 191
- European Medicine Agency, 304, 474, 933, 2040–2041
- effects table, 475–476
- European Organization for Research and Treatment of Cancer (EORTC), 597
- European Pharmacopeia (EP), 260, 709
- European Public Assessment Report (EPAR), 474
- European Statistics Code of Good Practice, 1098–1099
- European Union (EU), 284
- European experiences, 286
 - good laboratory practice (GLP), 1084–1086
 - history of good clinical practice implementation by, 1077
 - key principles and basic concept, 285
 - nonclinical and clinical evaluation, 285
 - product class-specific guidelines, 285–287
 - quality, 285
 - reference biopharmaceutical product, 285
- event-related oscillations (EROs), 1062
- event-related potentials (ERPs), 1062
- event time, 556
- evidence and replication, 726–727
- exact and approximate power, 275
- comparison of, 275–276
- Exact Inferences, odds ratio, 1555–1556
- exact permutation trend test, 403–405
- exact tests, for $2 \times K$ table, 77
- exchangeability, 174
- excipients/fillers, 256, 257
- exit interview, 464
- expectation and maximization (EM)
- algorithm, 594, 1270, 1354, 2122
 - correlated probit model, 743
- “expected power”, 523
- expected sample sizes (EN), 159
- Expert Panel on the Theory of Reference Values (EPTRV), 1276
- expiration dating period, 979–983, *see also* shelf-life
- regulatory requirements, 979
 - batch-to-batch similarity, 982
 - binary data without batch-to-batch variation, 982–983
 - direct method, 981
 - FDA's approach, 981
 - inverse method, 981–982
 - multiple components, 983
 - random batches, 983
 - shelf-life estimation with discrete responses, 982
 - statistical methods, 981
 - two-phase shelf-life estimation, 982

- explanatory variable, 1685
 exploratory factor analysis (EFA), 985, 1008
 common and unique factors, 985–986
 communality and unique components, 986
 computation, 1009
 factor loadings, 986
 model, 986
 number of factors, 1009
 orthogonal vs. oblique rotation, 986
 parameter estimation, 1009
 principal component analysis, comparison with, 987
 steps in conducting, 987
 step 1: determine the number of retained factors, 987–988
 step 2: rotate the chosen factors, 988–989
 step 3: interpret the factors, 988–989
 theoretic model, 1008–1009
 exponential distribution, 2180
 exponential model, 1152
 sample size calculation for, 1981–1982
 extended Cox model, 1839
 extended generalized least squares (EGLS), 1411, 1412
 Extended Mantel–Haenszel test, 56
 Extensible Markup Language (XML), 1093
 external (historical) control, 623
 external validation, 827
 Extracorporeal Life Support Organization, 1791
 extracorporeal membrane oxygenation (ECMO), 798
 study, 1953
 trial, 1732
 extra-Poisson variation, 73
 extra variation models, 992–1005
 Generalized Estimating Equations, 1001–1003
 goodness-of-fit test for, 1000–1001
 likelihood models, 997–1000
 quasi-likelihood models, 993–997
 random effect models, 1004–1005
 eye irritation, testing for, 2231
- F**
- factanal, 1009
 factor, 1009
 factor analysis, 1008–1012, *see also* exploratory factor analysis (EFA)
 background, 1008
 confirmatory factor analysis, 1009
 example, 1010–1012
 extensions, 1010
 methods, 1171
 results, depiction of, 989
 factorial designs, 1014–1026
 history, 1014
 mixed factorial design in biomedical research, application of, 1025
 orthogonal designs for 2^n factorial experiments, 1015
 interactions between and within subgroups (case), 1020
 interactions between subgroups (case), 1017–1020
 interactions within subgroups (case), 1016–1017
 orthogonal designs for 3^m factorial experiments, 1020–1022
 orthogonal designs for mixed $2^n \times 3^m$ factorial experiments, 1022–1025
 orthogonal designs for s^m factorial experiments, 1015
 factorization models, 1415
 asymmetry in treatment of outcomes, 1417
 direction of conditioning, 1417
 factor loadings, exploratory factor analysis, 986
 factor structure, determination of, 1170–1172
 failure-time model, 556
 accelerated failure-time models, 558
 estimation, 558–559
 issues, 561
 proportional hazards models, 557
 estimation, 557–558
 proportional odds models, 559
 estimation, 560–561
 false classification rates (FCR), 1609
 false discovery rate (FDR), 1028–1034, 1189, 1190–1191, 1471, 1625
 controlling methods, 1506
 control of, 1029
 error criteria for multiple comparisons, 1028
 definition, 1028–1029
 estimation of, 1029–1030
 based on marginal density of the statistics, 1030–1031
 local FDR, 1031
 on ApoAI data, 1033
 on breast cancer data set, 1033
 on Golub data set, 1032
 false negative, 833
 false negative rate (FNR), 2213, 2214
 coverage probability for models with, 2216
 false non-discovery rate (FNR), 1029
 false positive, 833
 false-positive error, 410–411, 727
 false-positive fraction, 826
 false-positive rate (FPR), 11–12, 2101, 2101
 non-inferiority test procedure, 11
 simulation, 12
 superiority test based on non-inferiority hypothesis, 12
 family disintegration, 320
 familywise error (FWE), 1383
 familywise error rate (FWER), 728, 1028, 1189–1190, 1502–1503, 1625
 and its control, 728, 1328
 fatal cardiac arrhythmia, 2206
 fatal tumors, 403
 FEATURE, 228
 Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA), 1084
 fed/fast studies, clinical pharmacology, 470
 field goal approach
 clinical pharmacology, 471
 Fieller's confidence set, 2144, 2145
 Fieller's method, 955, 956
 Fieller's theorem, 1700, 1702
 file drawer problem, *see* publication bias
 final analysis (FA), 1575
 finished product, 1811
 content uniformity/dissolution, 1811
 Bergum approach, 1811
 confidence interval, 1811
 simulation, 1811
 tolerance interval, 1811
 potency, 1811–1812
 sampling, 1804
 validation issues, 1812
 finite sample performance, cluster trials, 655
 Finney, D.J., 213
 first human dose studies, clinical pharmacology, 469
 first-order Taylor series approximation, 134
 fishbone diagram, *see* cause-and-effect diagram
 Fisher's combination of independent p -values, 545
 Fisher's dispersion test, 70
 Fisher's exact test, 56, 83, 221–222, 1384, 1556, 1558
 pharmacodynamics with no covariates, 1690–1691
 Fisher's information matrix (FIM), 193, 318, 994, 1604, 1730, 2032
 Fisher's least significant difference (LSD) test, 1689
 dose response analysis in clinical trials, 880–881
 Fisher's P -value and significance testing, 1125
 Fisher's scoring algorithm, 91
 fixed and flexible fixed-sequence (fallback) methods, 730–731
 fixed correction methods, 2207
 fixed-dose combination
 combination drug clinical trial, 552–553
 vs. dose-titration designs, 887
 vs. flexible dose, 2256
 fixed-effects model, 175
 therapeutic trials, meta-analysis of, 604–605, 609–610, 1364–1367
 fixed-margin method, 15, 16, 18, 1364–1366
 two confidence interval method
 with a fixed margin, 1536–1537
 with a random margin, 1537
 fixed parameter value, 524
 fixed sequence (FS) method, 730
 fixed sequential testing procedure, dose response analysis in clinical trials, 881
 fixed target values, 1336
 Fleiss–Cohen weight, 833
 Fleming multi-stage design, 1485–1486
 flexible design with sample size reestimation, clinical trial simulations for later development phases, 548
 flexible fixed-sequence (FFS) methods, 730
 flexible modeling of survival data, 2058–2059
 flexible sample size design, 2026
 objective of, 2027
 flow-through apparatuses, 263
 fluoroquinolone, 2209
 fluoxetine, surveillance of, 1774
 Follmann–Proschan–Geller approach, 1118
 Follmann's test, 1474
 follow-on biologics, biosimilarity of, 304–314
 assessing similarity using genomic data, 313
 criteria for similarity, 305–307

- follow-on biologics, biosimilarity of, (*Cont.*)
 absolute change vs. relative change, 305–306
 aggregated vs. disaggregated, 306
 consistency in manufacturing process/quality control, 307
 moment based criteria vs. probability based criteria, 306
 similarity factor for dissolution profile comparison, 307
 defined, 304
 factors considered by FDA, 305
 regulatory requirements, 305
 scientific issues, 308–312
 biosimilarity in biological activity, 308
 manufacturing process, 308
 problem of immunogenicity, 308
 similarity in size and structure, 308
 statistical considerations, 309–312
 statistical considerations, 309–312
 alternative criteria for biosimilarity, 311
 fundamental biosimilarity assumption, 309
 methods, 311–312
 study design, 309–310
- Food and Drug Administration (FDA), 191, 444, 473–474, 488, 552, 840, 964, 1089, 1095, 1121, 1184, 1259, 1326, 1648, 2040
- Critical Path Initiative, 536
 current Good Manufacturing Practice (cGMP), 370
 draft stability guidance, 359
 guidance, 1577
 Modernization Act of 1997, 537
 Pharmaceutical Science Advisory Committee, 537
 process validation, 1810
 sB/R framework, 476–477
 websites, guidance documents relating to biologics, 241
- food-effect bioavailability and bioequivalence studies, 265–266
- forced expiratory volume in 1 sec (FEV1), 1666
- force fields, 229
- force of mortality, 556
- forest plots, 482
- forests and Bayesian trees, 1788
- formularies, 1693
- formulation components, 257–258
 disintegrating agents, 257
 excipients/fillers, 257
 formulation of inhalants, 258
 granulating agents and binders, 257
 lubricants, 257
 ointments and skin preparations, 257
 surfactants (wetting agents), 257
- Fourier transform, 1545
- four-treatment four-period crossover design, 1619
- fractional factorial designs, response surface methodology, 1940–1941
- fraction retention method, 15, 17, 18, 19, 20
- frailty models, 1839–1840
- fraud detection, 820
- frequentist approach
 fixed-effects, 1531–1532
 generalizability probability, 1042–1043
 frequentist operational characteristics (FOC)-based approaches, 1593
 randomized trial with time to event end point, 1594–1595
 randomized two-arm study with normal outcome, 1594
 single arm study with binary end point, 1593–1594
 frequentist random-effects models, 1532–1533
 “friction cost” approach, 1696
 Fridericia’s correction method, 2208
 friendliness, remodeling
 good programming practice, 1091
 frozen drug products, 2111
 frozen study, 2111
F-test, 126, 127, 128, 436, 486, 491, 879, 1686, 1689
 sample size for, 498–499
 full Bayesian methods, 2092
 full risk set, 425
 full-service contract research organizations, growth of, 718
 functional mapping, statistical genetics, 2123
 Fundamental Bioequivalence Assumption, 304, 928, 929, 933
- ## G
- GABA_A gene cluster, 1063
 Galbraith plot, 1366
 Galois fields, 1014
 gastroduodenal ulcers, 316
 gastrointestinal (GI) studies, 316
 gastrointestinal system, 1667
 peptic ulcer disease, 1667
 gatekeeping testing strategies, 734
 Gauss-Hermite quadrature, 91
 Gaussian copula, 760
 Gaussian distribution, 651–652, 2066, 2329
 Gaussian endpoints, 1474
 Gaussian quadrature, correlated probit model, 745
 Gauss-Newton procedure, 994
 GEEs, *see* generalized estimating equation (GEE)
 Gehan estimating function, 559
 Gehan statistic, 2181
 Gehan’s test, 50
 Gehan’s two-stage design, 1485
 GenBank, 224
 gender differences, pharmacodynamic issues, 1671
 gene expression matrix map with hierarchical clustering, 1386–1388
 GENEHUNTER, 1061
 gene network modeling and integrated system biology, 1390
 Gene Ontology, 230
 generalizability probability, in clinical research, 1042
 applications, 1043
 Bayesian approach, 1043
 example, 1044
 frequentist approach, 1042–1043
 generalized confidence intervals (GCIs), 1054, 1055
 generalized estimating equation (GEE), 835, 1001–1003, 1046–1052, 1936, 2010–2014
 approach, 82, 85, 655, 743, 746
 asymptotic properties, 1047
 binary outcome variable case, 2011
 clustered categorical data approach, 92
 computation, 1052
 continuous outcome variable case, 2011
 correlation structure, 2013
 examples, 2013
 GENISOS (binary outcome case), 2014
 labor pain study (continuous outcome case), 2013–2014
 extensions of, 1051
 joint modeling, 1051
 missing data, 1051
 nonparametric GEE, 1051
 genetic application of, 1051–1052
 missing pattern, 2013
 sample size calculation, 2011
 binary outcome variable case, 2012
 continuous outcome variable case, 2012
 test statistics, 1049
 likelihood ratio-type, 1050
 score-type statistics, 1049–1050
 Wald-type statistics, 1049
 using GEE with caution, 1050
 working correlation, 1047
 efficiency, 1048
 estimation of correlation, 1048–1049
 generalized inference, 1054–1056
 area under the ROC curve, 1055
 balanced linear models, tolerance intervals for, 1055
 generalized *p*-value and generalized confidence interval, 1054
 limitations, 1055–1056
 generalized least-squares (GLS) approach, 1423, 1531, 784
 generalized linear latent and mixed models (GLLAMM), 744
 generalized linear mixed model (GLMM), 652, 655, 1004, 1005, 502
 outlier detection under, 505
 formulation, 505
 individual observations, diagnostics for, 506–507
 outlying observations, detection of, 505–506
 generalized linear models (GLM), 993, 1047, 502, 1330
 framework, 1004
 generalized partial credit model, 1212
 generalized pivotal quantities (GPQs), 149–150, 1054, 1056
 generalized test variables (GTVs), 1054, 1056
 general linear model (GLM), 104, 126
 general location model (GLOM), 1416
 general mixed-data model (GMDM), 1416
 gene ranking, 1382
 generic drugs, 1187–1194
 adjusted indirect comparison for generics’ bioequivalence, 1188–1189
 under 2 × 2 crossover design without carryover effects, 1193–1194
 bioequivalence limits, determining, 1191

- companies, 928, 929
- multiple comparisons, 1189–1190
- real example analysis, 1191
 - HIV/AIDS generics, 1191–1192
 - malaria generics, 1191
- restricted confidence intervals, 1189
- similarity assumption, 1190–1192
- simultaneous comparison of three generics, 1190
- generic products, drug interchangeability for, 935–938
 - average BE, 936
 - individual BE (IBE) for switchability, 937
 - population BE (PBE) for prescribability, 936–937
- gene set analysis (GSA), 1384
- Gene Set Enrichment Analysis (GSEA), 1384
- genetic application, of generalized estimating equation (GEE), 1051–1052
- genetic biomarkers, 1649
- genetic distance between two markers, 1058
- genetic effect, 98, 99
- genetic linkage and linkage disequilibrium analysis, 1057–1066
 - classical linkage analysis, 1057–1058
 - complex diseases, genetics of, 1062
 - EEG as a quantitative endophenotype, 1062–1064
 - Study on the Genetics of Alcoholism (COGA), 1062
- quantitative trait loci (QTL), mapping, 1058, 1060
 - combined linkage and LD analysis, 1061
 - heritability, 1058
 - IBD estimation, 1060
 - LD and fine mapping, 1060–1061
 - linkage analysis, 1059–1060
 - LOD score, 1060
 - software, 1061
- genetic mapping, 1058, 1122
 - by incorporating mechanistic pathways, 2125–2126
- genetic marker, 1057
- genetic programming, 807–808
- Genetics vs Environment In Scleroderma Outcome Study (GENISOS), 2010
- genetic toxicology tests, 67, 73, 1155
- genetic/transcriptional regulatory networks (GRNs), 227
- GENISOS (binary outcome case), 2014
- Genomes to Life Program, 1444
- genome-wide association studies (GWAS), 1448, 2184
- genomic data, assessing similarity using, 313
- genomics, comparative, 226
- genotoxicity studies, 2233
 - chromosome aberrations, tests for detection of, 2234
 - chromosomal aberration test in vivo, 2234
 - micronucleus test, 2234
 - DNA indicator tests, 2234
 - gene mutations, methods for detection of, 2233–2234
- genotype and phenotype, 228
- Genstat package, 954
- geographical information system (GIS), 2094
- geometric mean ratio (GMR), 219, 288–289, 929, 1189–1190
- Gibbs sampling method, 1358
- Gleevec (Imatinib), 1651
- global/complete null hypothesis, 728
- global database and system, 1069–1073
 - data components, 1071
 - maintenance, 1072
 - change control, 1073
 - change management and training, 1072–1073
 - requirements analysis, 1069–1070
 - system architecture, 1070–1071
 - system development, 1071
 - methodology, 1071–1072
 - validation, 1072
 - transition, 1073
- global dynamic balance/harmony, 2254
- globalization of clinical development, 1080
- global regulatory efforts, medical devices, 1350
 - analysis, 1352
 - causal inference, 1352
 - center-by-treatment interaction, 1352
 - choice of active control, 1351
 - choice of noninferiority margin, 1351
 - historical controls, 1350–1351
 - implants, 1350
 - masking (blinding), 1350
 - monitoring, 1352
 - non-diagnostic medical device studies,
 - design issues in, 1350
 - noninferiority trials, 1351
 - propensity scores, 1352–1353
 - repeated measurements, 1352
 - sensitivity analysis, 1353
 - sham controls (the placebo effect), 1350
 - subgroup analysis, 1352
 - surrogates, 1351
- global testing, 732
- “go” decision, 1723
- gold standard, 822
- Golub data set, local FDR on, 1032
- Gompertz–Makeham survival analysis
 - regression models, 1495
- Gompertz model, 1152, 1547
- “go/no go” decision, 1722, 1723
- good clinical practice (GCP), 535, 578, 719, 978, 1073, 1075–1079, 1095, 580
 - clinical research and, 1075
 - ethical and scientific quality of clinical research, 1075
 - Declaration of Helsinki, 1075–1076
 - quality assurance and quality control, 1076
 - future perspectives, 1079
 - biostatistics and good clinical practice, 1079–1080
 - globalization of clinical development, 1080
 - history of implementation of, 1076
 - by European Union, 1077
 - by Japan, 1077–1078
 - by United States, 1076–1077
 - by World Health Organization (WHO), 1078
- International Harmonization of Good Clinical Practice, 1078–1079
- good correction method, 2207
- good laboratory practice (GLP), 1082–1086, 1095
 - China, 1085
 - European Union, 1084–1086
 - good statistics practice and, 1086
 - Japan, 1085
 - laboratory accreditation quality system and, 1086
 - mutual recognition, 1085–1086
 - nonclinical laboratory study and scope of, 1083
 - principles, 1083
 - US Environmental Protection Agency, 1084
 - US Food and Drug Administration, 1084
 - World Health Organization (WHO), 1085
- Goodman’s identification condition, 944
- goodness-of-fit, 131
 - assessment, 170
 - diagnostics, logistic regression, 1305–1306
 - statistic to test, 993
 - test for extra variation models, 1000–1001
- goodness-of-fit index (GFI)
 - factor analysis, 1010
- good programming practice, 1088–1094
 - coding, 1089
 - designing, 1088–1089
 - documentation and organization, 1092
 - economy, 1091–1092
 - extendibility, 1091
 - keeping updated, 1093
 - programming etiquette on a shared project, 1089–1094
 - quality, 1088
 - remodeling friendliness, 1091
 - robustness, 1089–1091
 - security plan, 1093
 - validation, 1092–1093
- good regulatory practice (GRP), 1095
- good statistics practice (GSP), 580, 1086, 1095–1099
 - drug development, role of statistics in, 1095–1096
 - implementation of GSP, 1099
 - statistical principles, 1096
 - bias and variability, 1096
 - confounding and interaction, 1096
 - European Statistics Code of Good Practice, 1098–1099
 - one-sided test vs. two-sided test, 1098
 - randomization, 1097
 - sample size determination/justification, 1097
 - statistical difference and scientific difference, 1097–1098
 - type I error, significance level, and power, 1096–1097
- good validation practice (GVP), 1101–1103
 - analytical procedures, validation of, 1102
 - analytical life cycle and method transfer, 1103
 - harmonized characteristics, 1103
 - regulated requirement, 1102–1103
 - process validation (PV) of pharmaceutical industry, 1102

- good validation practice (GVP), (*Cont.*)
 quality-by-design (QbD) quality-by-design concept and PV, 1102
 regulated requirement, 1102
 technology transfer (TT) of pharmaceutical process, 1102
 statistics and good validation practice, 1103
- GPQ-SSDL method vs. GPQ-ADL method vs. GPQ-CVDL method, 152–153
- graded response model, 1212–1213
- granulating agents and binders, 257
- granulation, 258
- granulocyte colony-stimulating factors (G-CSF), 237
- graphic and raw data comparison for Tier 3 CQAs, 2221
- gravimetric mass, 130
- Graybill-Deal (GD) estimator, 2292
- Greenberg Report, 811
- Greenwood's formula, 557
- grid search method, 498
- griseofulvin, 260
- group-by-dependence interaction, 207
- grouped data, trend estimation with, 2285–2286
- grouped jackknife method, 1701
- group-randomized trials, 648
- group sequential methods/procedure, 59, 1105–1111, 1114, 2327
 confidence intervals following a group sequential test, 1110
- Data Monitoring Committee, 1109–1110
- multiarmed trials, extension to, 1116–1117
- with one interim analysis, clinical trial simulations for later development phases, 549
- recent developments, 1110
 group sequential methods, 1110
 monitoring design specifications to extend a trial, 1111
 sequential monitoring trials with multiple endpoints, 1110
- regulatory requirements, 1105
- related issues, 1117
 contrasts and trend tests, 1118–1119
 multiple comparisons, 1117–1118
- sample size calculation in, 1983–1984
- and sample size re-estimation designs, 2027–2028
- standard, 59–60
- statistical methods, 1106
 equally spaced group sequential boundaries, 1106
 fundamentals, 1106
 information time/fraction and maximum duration trial vs. maximum information trial, 1108–1109
 partial sum process with independent increments, 1108
 stochastic curtailment and Bayesian methods, 1109
 type I error spending/use function approach, 1106–1108
 under variance heterogeneity, 1115–1116
- growth data, application to, 1549
- guidance documents, for biologics, 241
- H**
- Haemophilus influenzae* type B vaccine, 631
- Haldane-Anscombe 1/2 correction estimator, 1555
- Haldane's model of recombination, 1058
- half-life of drug, clinical pharmacology, 470
- Hall and Wellner (HW) bands, 1233, 2059
- Halobacterium salinarum*, 2126
- haloperidol, 192
- Hamilton Depression (HAM-D) score, 1589, 2201
- hard gelatin capsules, 259
- harmful multicollinearity, 1444
- HARTS, 54
- Haseman and Elston method (H-E method), 1052, 1059, 1061
- Haseman decision rule, 410
- Hatch-Waxman Act, 281, *see also* Drug Price Competition and Patent Term Restoration Act
- Haybittle-Peto boundaries, 59, 60f, 1106
- hazard rates, 556
 survival functions and, 2180
- hazard reduction, 1579
- health-care assessment, application of IRT in, 1214
- health maintenance organizations (HMOs), 716
- Health Protection Branch (HPB), 908
- healthy years equivalent (HYE), 1697
- heart muscles, 1722
- heart rate, 822
- Helicobacter pylori* treatments, surveillance of, 1773
 cost-effectiveness analysis, 1774
 drug interactions, 1773
 excluded populations, 1773–1774
 long-term follow-up, 1773
- Hellinger distance, 1255
- Helsinki, Declaration of, 1075–1076
- hemophilia, 235
- HER2+ breast cancer, 964
- HER2 gene amplification, 965
- HER2-targeting agents, 965
- HerceptTest™, 1648
- HERCEPTIN®, 241
- herd immunity phenomenon, 238
- heritability
 analysis of, 98–102
 defined, 98
 discussion, 102
 hypothesis tests, 102
 interval estimation, 101
 point estimation, 100–101
 analysis of variance estimation, 100–101
 maximum likelihood and restricted maximum likelihood estimation, 101
 statistical modeling of, 99–100
- heterogeneous studies, 2102
- heterogeneous variances, issue of, 2236–2237
- hierarchical Bayesian model, 174, 175f
- hierarchical clustering tree (HCT), 1386, 1388
 Flipping effects of intermediate nodes of, 1387f
- hierarchical modeling, 1357
- hierarchical testing principle/framework, 733
- high density lipoproteins (HDL) cholesterol levels, 1033
- highest posterior density (HPD), 195
 interval, 294
- High Level Group Term (HLGT), 1342
- High Level Term (HLT), 1342
- highly variable (HV) drugs, 217
 Scaled average bioequivalence (SABE) for, 2039–2040
- high-performance liquid chromatography (HPLC), 131
- high reporting rate, signaling of, 1758–1762
- high-throughput molecular profiling, 1648
- Hill factor, 880
- Hills-Armitage test, 211–212
 SAS code and output for, 211–212f
- H index, 606–607
- histidine, 68, 69
- histograms, 2135–2136f
- historical control data, 411–412
- historical controls (HCTs), 572
- historical data, 524
- HIV/AIDS generics, 1191–1192
- Hochberg's method, 729, 730, 1469, 1638, 2209
 dose response analysis in clinical trials, 881
- Hodges-Lehmann estimator, 295–296, 2143
- holistic approach, process validation, 1810
- Holm's procedure, 729, 2209
- Hommel and Hochberg tests, 730
- homogeneity assessment of treatment effects
 therapeutic trials, meta-analysis of, 1365–1366
- homogeneous studies, 2102
- homologous sequences, 225
- homology modeling methods, 229
- hormone receptor status, 1649
- Hosmer-Lemeshow goodness-of-fit test, 1653
- Hospital Anxiety and Depression Scale (HADS), 985
- Hotelling's T^2 test, 504–505, 855, 856, 1474, 1638
- human abuse potential (HAP) study, 1121, 1122
- human capital method, 1695, 1696
- human central nervous system (CNS), 1121
- human drug abuse trials, 1121
 human abuse potential (HAP) study, 1122
 study for abuse-deterrent opioids, 1121
- human epidermal growth factor receptor 2 (HER-2), 1648
- Human Genome Project (HCP), 313, 962
- human immunodeficiency virus (HIV), 1610
- human lymphocytes (HL), 1155
- hybrid designs, 781–782
- hydrocodone, 1121
- 3-hydroxy-3-methylglutaryl coenzyme A (HMG CoA) reductase inhibitors, 1258
- hypergeometric distribution, 209, 211
 clinical inspection, 465–466
- hyperlipidemia, 1666
- hyperparathyroidism, diagnostic methods for, 85, 86
- hypertension, 1666
- hypoactive sexual desire, 1168
- hypoactive sexual desire disorder (HSDD), 1167

hypotheses in active control non-inferiority trials, 8
 with fixed δ margin, 15–16
 with fraction of control effect to be retained, 17–18
 linear transformation, 10
 non-inferiority hypotheses, transformation of, 10
 parameter space, 9
 ratio non-inferiority hypotheses, 9–10
 superiority hypotheses, 15
 for treatment and control effect, 9
 union non-inferiority hypotheses, 10–11
 hypothesis testing, 1124–1126, 1235, 816
 Bayesian, 1126
 clustered binary data, 83–85
 and confidence interval, 2283
 asymptotic approach, 2283–2284
 conditional exact approach, 2284–2285
 in equivalence trials, 1126
 Fisher's *P*-value and significance testing, 1125
 heritability, 102
 Neyman–Pearson formulation of, 1124–1125
 in noninferiority trial, 1125
 relationship with, 701
 in superiority trial, 1125
 in super superiority trial, 1125
 hypothetical gene–protein couples, 2127f
 hypoxanthine-guanine
 phosphoribosyltransferase, 2233
 Hyslop method, 346

I

I^2 index, 607
 idarubicin (IDA), 763
 identical by descent (IBD) genes, 1052, 1059
 identifiability, 1269
 local, 1269
 weak, 1269
 immune globulins, 241
 immune response, 239
 immunoassays, 131, 133
 immunogenicity, 308
 immunohistochemistry (IHC) test systems, 823
 immunologic persistence study, 635
 imperfect multicollinearity, 1444
 implantable wireless monitoring system, 822
 implied correlation structure, 744
 improvement, between covariates and outcome, 774
 imputations
 case analysis, 1927–1928
 methods, 417
 multiple imputations, 1926–1927
 repeated measures data, 1926–1927
 single imputations, 1926
 imputed data, bootstrapping, 344, 345
 incidental tumors, 400–402
 data analysis, 407
 incomplete block crossover design, 1619
 incomplete design approach, 2227
 incomplete repeated measures data analyses, 1919–1921
 incontinence episode frequency (IEF), 1221

incremental cost-effectiveness ratio (ICER), 752, 1699
 incremental net benefit (INB), 753
 incremental ratio, average ratio vs., 1710–1713
 independent programs, 1711
 mutually exclusive programs, 1711–1713
 Independent Data Monitoring Committee (IDMC), 1109
 independent double data entry, clinical data management guidelines, 447
 Independent Ethics Committees (IECs), 1077
 independent missing type, 2013
 independent programs, cost-effectiveness analysis, 1711
 indirect costs, 1696
 individual bioequivalence (IBE), 266, 306, 345, 922, 929, 973, 1160–1166, 1250, 1253, 1432, 1740
 average bioequivalence, limitations of, 1160–1161
 criterion (IBC), 1161–1162
 design-driven individual bioequivalence criteria, 1253–1254
 review of 2001 FDA guidance, 1164
 aggregated criteria vs. disaggregated criteria, 1164
 interpretation of the criteria, 1164
 masking effect, 1164
 outlier detection, 1165
 power and sample size determination, 1165
 study design, 1165
 two-stage test procedure, 1165
 statistical method, 1162–1163
 for switchability, 937
 individual correction methods, 2207
 individual difference ratios (IDRs), 219, 1161
 individual equivalence, medical devices, 1355
 analytical studies, 1356
 Bayesian adaptive design, 1357
 Bayesian approach in evaluating medical devices, 1356
 Bayesian calculations, 1358
 Bayesian hierarchical modeling, 1358
 Bayesian interim analyses, 1357
 Bayesian noninferiority testing, 1357
 Bayesian sample size, 1357
 Bayesian statistics, 1356
 biomarkers, diagnostic tests, and microarrays, 1356
 diagnostic imaging, 1355
 historical controls as prior information, 1357
 monitoring devices, 1356
 valid prior information, 1357
 individual heterogeneity, 1839
 individual-level analysis, cluster trials
 with mixed effects models, 652–653
 with population-averaged models, 651–652
 individual minimum effective dose (IMED), 1406
 infectious diseases (ID), 1667
 clinical trials, 316
 inference, correlated probit model, 746
 inference, in case-control studies, 422–432
 with Bayesian prior assumptions, 426–427
 case-control sampling designs, 424

without full population information
 classic case-control, 427–430
 density case-control, 430–432
 without population information, 424–425
 classic case-control, 424–425
 density case-control, 425
 quantities of interest, 423–424
 inflation factor, 84, 85
 inflation of the type I error rate, 728
 influential observations, assessment of, 1498–1499
 information assessment, clinical trial process, 533–534
 assessing end of the process, 533
 assessing intermediate outputs/milestones, 534
 assessing starting point for the process, 533
 assessing tasks involved in the process, 533
 information gathering, clinical trial process, 532–533, 534
 defining intermediate outputs or milestones, 533
 defining the end of the process, 532
 defining the starting point for the process, 532
 defining the tasks in a process, 533
 step, 532
 information technology (IT), 719
 information time/fraction and maximum duration trial vs. maximum information trial, 1108–1109
 informative censoring, 1235
 informative prior distribution, 176
 informed consent, contract research organization, 719
 inhalants, formulation of, 258
 innovative drug companies, 928
 Innovative Medicines Initiative and European Federation of Pharmaceutical Industries and Associations, 474–475
 inspection process, 463–466
 exit interview, 464
 before inspection, 463
 scope of inspection, 463
 installation qualification (IQ), 1072
 instantaneous death rate, 556
 institution effect, 1221
 Instituto Nacional de Estadística (INE), 1098
 instrument calibration, 370
 instrument development and validation, 1167–1177
 concurrent validity, 1175
 confirmatory factor analysis, 1172–1173
 construct validity, 1175–1176
 data quality, 1170
 determination of factor structure, 1170–1172
 instrument validation, 1174
 item generation, 1168–1169
 item response theory, 1173–1174
 methods of analysis, 1170
 qualitative content validation, 1169–1170
 quantitative instrument development, 1170
 scoring, 1174
 test-retest reliability and internal consistency, 1176
 intangible benefits, 1696

- integrated summary of efficacy (ISE), 1179
- integrated summary of safety (ISS), 57, 1179, 1181, 1343
- integrated summary report, 1179–1181
 - of efficacy, 1179
 - dose rationale, 1179–1180
 - efficacy data, 1180
 - meta-analysis, 1180–1181
 - of safety, 1181
- integrity of blinding, 337
- intellectual property, contract research organization, 719
- intelligence, 1008
- intelligence quotient (IQ), 1266
- Intelligent Systems for Molecular Biology (ISMB), 224
- intensity rate, 556
- intention-to-treat (ITT) analyses, 417, 577, 705, 903, 1183–1186, 1296, 2203
 - regulatory requirements and guidelines, 1183–1184
 - scientific issues, 1185–1186
 - statistical issues, 1185
- intent-to-diagnose (ITD), 826
- interaction, 703, 706
 - covariate, 774–775
 - evaluating, 706
 - regulatory requirements, 707
- interactive voice response system (IVRS)/
 - interactive web response system (IWRS) for randomization, 1216–1223
- applications, 1222
- clinical data management guidelines, 446, 447
- drug supply and packaging, 1218
- mapping by, 1219–1220
- minimization, 1221–1222
- nonstratified randomization, 1221
- patient randomization using, 1220
- randomization schedule, 1219
- relational database for randomization and drug assignment, 1219f
- stratified randomization, 1221
- traditional approach to randomization and drug management, 1216–1217
 - problems associated with, 1217
- working of, 1218
- interactive web response system (IWRS), 1216
- Interagency Registry for Mechanically Assisted Circulatory Support, 1791
- interchangeability, 973
- interchangeability, generic drugs, 1187–1194
 - under 2×2 crossover design without carryover effects, 1193–1194
 - adjusted indirect comparison for generics' bioequivalence, 1188–1189
 - bioequivalence limits, determining, 1191
 - multiple comparisons, 1189–1190
 - real example analysis, 1191
 - HIV/AIDS generics, 1191–1192
 - malaria generics, 1191
 - restricted confidence intervals, 1189
 - similarity assumption, 1190–1192
 - simultaneous comparison of three generics, 1190
- intercurrent mortality, adjustment of tumor rates for, 399–400
- interim analysis, 59, 1195–1199
 - clinical trials, 576–577
 - design and analysis, using alpha spending function, 63, 64
 - motivation for the use of, 1195–1196
 - procedural issues and concerns, 1196–1198
 - software, 1199
 - statistical issues, 1196
 - statistical methodology, 1196, 1198–1199
- intermediate endpoint, 2174
- internal validation, 827
- International Classification of Diseases
 - ICD-9, 54
 - ICD-10, 1267
- International Conference on Harmonization (ICH), 54, 234, 313, 359, 408, 417, 445, 585, 719, 909, 976, 1042, 1078, 1202–1206, 1259, 1344
 - guidelines, 1202–1205
 - ICH-E9 document, 59, 438, 442
 - ICH E14 analysis, 2206, 2208, 2215
 - ICH E14 guidance, 2206, 2208, 2210, 2211, 2214
 - incidence of adverse events, 56–57
 - organization, 1202
- International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, 125
- International Council for Harmonization (ICH), 474, 2206
- International Federation of Pharmaceutical Manufacturers Association (IFPMA), 1202
- International Harmonization of Good Clinical Practice, 1078–1079
- International Index of Erectile Function (IIEF), 1175
- International Personality Item Pool (IPIP), 1168
- International Society for Pharmacoeconomics and Outcomes Research (ISPOR), 475
- International Society of Computational Biology (ISCB), 224
- International System of Units, 1259
- interpretation
 - correlated probit model, 744
 - logistic regression, 1305
- intersection hypothesis of dimension, 732
- intersection null hypothesis of dimension, 728
- intersection–union principle (IUT), 971, 1161, 2150, 2208, 2209
- Interstitial Cystitis (IC), factor analysis, 1010
- intersubject (or between-subjects) variability, 486
 - comparing, 491
- interval-censored data
 - software for analysis of, 594
 - statistical analysis of, 594
- interval-censored failure time data, 589–595
 - interval-censored data, software for analysis of, 594
 - notation, assumptions, and likelihood function, 590
 - statistical analysis
 - of multivariate interval-censored data, 593
 - of other interval-censored data, 594
 - of univariate interval-censored data, 591–592
- interval estimation, 101
- interval hypotheses, 292
 - of equivalence, 971
- intraclass and weighted kappa coefficients, interchanging, 1243
- intraclass correlation, 1911–1912
- intraclass correlation coefficient (ICC), 833, 1176
- intraclass kappa, 1238–1239
- intracluster correlation coefficient (ICC), 83, 648, 993
- intrarater, interrater, and test–retest reliability, 1912
- intrasubject (or within-subject) variability, 486
- intra-subject rank-based method, 787
 - estimation of τ , 789
 - rank tests, 788–789
 - R estimation, 787–788
- intra-subject variability (ISV), 289–290, 487
- intrinsic dissolution rate, 255
- inverse gamma (IG) distributions, 176
- inverse method, expiration dating period, 981–982
- inverse normal combination test, 1578
- inverse probability of treatment weighting (IPTW), 1825
- investigational data presentation, 1344
- Investigational Device Exemption (IDE), 1349
- investigational new drug (IND), 180, 908, 911, 1343
 - application, 234, 721
 - clinical data management guidelines, 444
- in vitro* assay, 68
- in vitro* bioequivalence testing, 1141–1146
 - droplet size distribution, 1142
 - cascade impaction technique, 1142
 - laser diffraction, 1142
 - plume geometry, 1142
 - priming, emitted dose uniformity, priming/repriming, and tail-off profile, 1142
 - profile analysis for, 290–291
 - recent development, 1145
 - nonprofile analysis, 1145
 - profile analysis, 1145–1146
 - sample size, 1146
 - regulatory requirements and statistical methods, 1143
 - noncomparative analysis, 1143
 - nonprofile analysis, 1143–1144
 - profile analysis, 1144–1145
 - spray pattern, 1142
 - study design, 1141
- in vitro* “companion diagnostic” devices, 1648, 1649
- in vitro* diagnostic (IVD) devices, 823, 1349, 822
- in vitro* dissolution profile comparison, 1148–1153
 - alternate approaches, 1152–1153
 - dissolution profile similarity test based on similarity factor f_2 , 1150
 - bias of the estimate of similarity factor, 1151

- estimation and confidence interval of similarity factor, [1151](#)
 - theoretical considerations, [1150–1151](#)
 - dissolution profile similarity testing based on D_M , [1150](#)
 - measures of difference/similarity between two dissolution profiles, [1148–1150](#)
 - modeling approach, [1152](#)
 - selection of measures, [1150](#)
 - similarity criterion, [1151–1152](#)
 - in vitro* micronucleus (IVM) test, [1155–1159](#)
 - design, [1156](#)
 - need for, [1155](#)
 - results, [1157](#)
 - statistical analyses, [1156–1157](#)
 - in vivo* assay, [68](#)
 - invoked population model, [1875](#)
 - I-optimality, [1945](#)
 - Irwin's toxicity study, [1414](#)
 - Iso-chi-square test, [79, 80](#)
 - ISPOR Risk-Benefit Management Working Group, [475](#)
 - Item Characteristic Curve (ICC), [1173f](#) and categorical probability curve, [1209–1210](#)
 - item generation, [1168–1169](#)
 - item nonrespondents, imputation with, [1128–1134](#)
 - example, [1131–1132](#)
 - statistical inference, [1130](#)
 - simulation study, [1130–1132](#)
 - testing hypothesis, [1130](#)
 - variance estimates, [1130](#)
 - statistical model, [1129](#)
 - multiple-imputation classes, [1129–1130](#)
 - single-imputation class, [1129](#)
 - item response theory (IRT), [1167, 1173–1174, 1207–1214](#)
 - application, in health-care assessment, [1214](#)
 - background, [1207–1208](#)
 - basic assumptions and terminology, [1209](#)
 - item characteristic curve and categorical probability curve, [1209–1210](#)
 - item fit, [1211](#)
 - item parameters, [1210–1211](#)
 - local independence, [1209](#)
 - model fit, [1211](#)
 - sample size considerations, [1211](#)
 - unidimensionality, [1209](#)
 - brief history of, [1208–1209](#)
 - choosing an IRT model, [1214](#)
 - factor analysis, [1010](#)
 - generalized partial credit model, [1212](#)
 - graded response model, [1212–1213](#)
 - models, [1173](#)
 - partial credit model, [1212](#)
 - Rasch model, [1211](#)
 - 2PL model, [1211–1212](#)
 - iteratively reweighted least squares (IRLS), [134, 1304](#)
- J**
- Japan, [1517](#)
 - good laboratory practice (GLP), [1085](#)
 - history of good clinical practice implementation by, [1077–1078](#)
 - Japanese MHLW guidance, [1578](#)
 - Japanese Ministry of Health, Labor, and Welfare (MHLW), [1202](#)
 - Japanese Pharmaceutical Manufacturers Association (JPMA), [1202](#)
 - Japanese Pharmacopeia (JP), [709](#)
 - Jeffreys–Good–Lindley paradox, [1849](#)
 - Jeffreys prior, [193](#)
 - Jennison–Turnbull confidence interval, [2303](#)
 - Jensen's inequality, [2034](#)
 - joint mixed-outcome models, [1415](#)
 - Jonckheere–Terpstra test, [2240](#)
 - Jonckheere test, [70, 72](#)
 - for determining drug effect, [880](#)
- K**
- Kaplan–Meier curves, [1653](#)
 - Kaplan–Meier estimator, [556, 557, 558, 1231–1235, 1699, 1838, 2181](#)
 - descriptive statistics, [1234–1235](#)
 - example, [1233–1234](#)
 - formulation of, [1231–1232](#)
 - hypothesis testing, [1235](#)
 - informative censoring, [1235](#)
 - Nelson–Aalen estimator, [1232](#)
 - self-consistent estimator, [1232](#)
 - survival function, inference on, [1232](#)
 - confidence limits and bands, [1233](#)
 - variance estimation, [1232](#)
 - Kaplan–Meier graphical method, [816–817](#)
 - Kaplan–Meier Product-Limit method, [1589, 1590f](#)
 - Kaplan–Meier proportion, [1690](#)
 - Kaplan–Meier survival curves, [2179, 2180f](#)
 - kappa coefficients in medical research, [1237–1243](#)
 - intraclass kappa, [1238–1239](#)
 - mathematical basis, [1237–1238](#)
 - problems with kappa, [1242](#)
 - base rate problem, [1242](#)
 - interchanging the intraclass and weighted kappa coefficients, [1243](#)
 - kappa for ordered categories, [1243](#)
 - multicategory kappa, [1242](#)
 - weighted kappa, [1239](#)
 - in the context of measures of 2×2 association, [1240–1242](#)
 - kappa statistic, [832, 833](#)
 - Kefauver-Harris amendment of FDC Act of [726, 1962](#)
 - Kendall's test of independence, [1977](#)
 - example, [1978–1979](#)
 - fixed alternative, [1977](#)
 - local alternative, [1977](#)
 - simulation study, [1977–1978](#)
 - kernel density estimation, [806, 2161](#)
 - and multivariate regression, [1551](#)
 - density estimation, [1551](#)
 - multivariate regression, [1551](#)
 - kernel estimators, [1543–1544](#)
 - Kim and DeMets error spending function, [62](#)
 - Kittleson and Emerson group sequential method, [62](#)
 - K-mean algorithms, [1389](#)
 - K-means clustering, [1388](#)
 - “known groups validity”, [1175](#)
 - Kolmogorov-Smirnov tests, [311, 1235, 1966](#)
 - Korea Food and Drug Administration (KFDA), [288](#)
 - Krebs, Hans, [227](#)
 - Kruskal–Wallis test, [70, 1689, 1245–1248](#)
 - discussions, [1247–1248](#)
 - general power and sample size formula, [1246](#)
 - MOG-induced EAE mouse study, [1247](#)
 - for ordered categorical data, [1245](#)
 - pharmacodynamics with no covariates, [1689](#)
 - power and sample size formula for ordered categorical data, [1246–1247](#)
 - Kullback–Leibler divergence (KLD), [1249–1256](#)
 - bioequivalence criteria and conceptual issues, [1250–1251](#)
 - design requirements, [1252](#)
 - desirable properties for distance, [1251](#)
 - generalization to exponential family of distributions, [1252](#)
 - generalization to multivariate case, [1252–1253](#)
 - interpretation on the original scale, [1252](#)
 - logical hierarchy of conclusions, [1252](#)
 - scaling and symmetry, [1251](#)
 - Hellinger distance, [1255](#)
 - individual bioequivalence, [1253](#)
 - design-driven individual bioequivalence criteria, [1253–1254](#)
 - simulations, [1255–1256](#)
- Kullback–Leibler information number, [1643](#)
- L**
- L'Abbe plot, [1366](#)
 - labeling extension values (LEVs), [1382](#)
 - laboratory analyses, [1258–1264](#)
 - analysis of outliers, [1260–1261](#)
 - grouping of, [1259](#)
 - laboratory-related adverse events, [1261](#)
 - means, mean percentage change, and medians, [1260](#)
 - measured, [1258](#)
 - reference range of, [1259](#)
 - scientific issues, [1261](#)
 - central laboratories, [1262–1263](#)
 - clinical correlations, [1263–1264](#)
 - clinical relevance of laboratory findings, [1263](#)
 - clinical significance vs. statistical significance, [1264](#)
 - data-conversion issues, [1262](#)
 - data-integration issues, [1262](#)
 - laboratory measurement errors, [1262](#)
 - real laboratory findings vs. laboratory error or artifact, [1263](#)
 - study population considerations, [1263](#)
 - subgroup analyses, [1262](#)
 - shift tables, [1260](#)
 - types of, [1259](#)
 - units of, [1259](#)
 - laboratory-based *in vitro* manipulation and extraction studies, [1121](#)
 - labor pain study (continuous outcome case), [2013–2014](#)
 - Lack-of-Fit test, [147](#)
 - Lagrange multipliers approach, [855](#)

- lag time, 1980
- lambda, silence of, 757
- Lan and DeMets spending function, 60, 61
- LANDEM software, 62
- language, translation in, 2270
- large-scale meta-analysis, 1327
- Lars method, 1449
- laser diffraction, *in vitro* bioequivalence testing, 1142
- Lasso and its extensions, 1449–1450
- Lasso method, 1449, 2272
- last observation carried forward (LOCF), 416, 418, 585–586, 903, 1128, 1296–1301, 1422, 1425, 1624
 - asymptotically correct tests, 1299
 - example, 1300–1301
 - one-way ANOVA model, validity of LOCF test under, 1297–1298
 - two-way ANOVA model, validity of LOCF ANOVA test under, 1298
 - with interaction, 1299
 - without interaction, 1298–1299
- latent class analysis, 1266–1273
 - assumptions, 1268
 - estimation methods, 1270
 - likelihood-based approaches, 1270
 - Markov chain Monte Carlo (MCMC) approach, 1271
 - identifiability, 1269
 - local, 1269
 - weak, 1269
 - likelihood function, 1268
 - mental health example, 1266–1267
 - model evaluation, 1271
 - assessing model assumptions, 1272
 - choosing a model and assessing fit, 1271–1272
 - notation, 1267–1268
 - posterior probability of class membership, 1270
 - practical application, 1272–1273
 - reparameterization, 1270
 - sample size considerations, 1269–1270
- latent class model (LCM), 1266
- latent mixed model, 93–94
 - Bayesian analysis, 94
 - for clustered categorical data, 89, 90–91
 - multilevel models, 89
- late-phase confirmatory trials, 2031
- Law of Iterated Logarithm (LIL), 1371
- LBE (Location Based Estimator), 1031
- ldBands() function, 62
- least square approach
 - early phases of clinical trials, simulation in, 542
- least squared estimators, 149
- Leber, Paul, 840
- left- and right-tail error coverage probabilities, 700
- Lehmann class, 557
- “level I” changes, data validation, 448
- Lexis diagram, 2064, 2065f
- licensing, mergers and, 820
- lidocaine, 574
- life table approach, 2227
- life-threatening diseases, 888–889
- likelihood, 174
 - likelihood-based approaches, 1270
 - likelihood based clustered categorical data approach, 91
 - likelihood displacement, 1293
 - likelihood distance test, 503–504
 - likelihood function, 1268, 2180
 - early phases of clinical trials, simulation in, 542
 - likelihood models, 997–1000
 - likelihood principle, violation of, 1849–1850
 - likelihood ratio chi-square, 76
 - likelihood ratio test (LRT), 317, 319, 1473, 1603–1604
 - for qualitative interaction, 2192–2194
 - likelihood ratio-type, generalized estimating equation, 1050
- Lilly Reference Ranges, 1276–1291
 - methods, 1276
 - delta limits, 1277
 - results, 1288–1291
 - selection of reference sample group, 1277
 - selection of sample size and statistical methods for estimating reference limits, 1277
- limiting allocation proportion (LAP), 1729
- limit of detection, 376
 - and quantitation, 2319
- limit of quantitation, 376, 2319
- linear calibration, 131–133
- linear contrasts within an ANOVA, 2240
- linear estimators, non-parametric regression, 1543
 - kernel estimators, 1543–1544
 - local polynomials, 1545
 - smoothing splines, 1544–1545
 - wavelet estimators, 1545–1546
- linear fixed-effect model, 181
- linearity assessment, 823
- linearity of an analytical method, 146, 376
- linear–linear fit, 131, 131–132f
- linear minimum mean squared error (LMMSE), 905
- linear minimum variance unbiased (LMVUB), 905
- linear mixed models (LMM), 502
- linear model, of ANOVA, 109–110
- linear pharmacokinetics, *see also* dose proportionality (DP)
 - properties, clinical pharmacology, 470
- linear predictor, 994
- linear regression
 - dose proportionality, 876–877
 - model, 371
 - on revertant counts, 71
- linkage disequilibrium (LD), 1444, 2124
- link function, 994
 - based design for continuous responses, 1729–1730
- Lipid Research Clinics Coronary Primary Prevention Trial, 575
- Lipid Research Clinics Primary Prevention Trial, 575
- litter effect, 1933
- liver function, 1671
- LLT level, 54
- local FDR
 - on ApoAI data, 1033
 - on breast cancer data set, 1033
 - on Golub data set, 1032
- local identifiability, 1269
- local influence analysis, 1292–1295
 - applications and extensions, 1294–1295
 - example, 1294
 - likelihood displacement, 1293
 - local influence, 1293–1294
 - perturbation schemes, 1292–1293
- locally most powerful (LMP) test, 102
- local polynomials, 1545
- logic errors, 1089
- logistic curve, 1152
- logistic-normal model, 1416
 - vs. probit-normal models, 1417
- logistic regression in three-point designs, 1315–1318
 - choice of design points, 1316
 - data configurations, 1316
 - complete separation of responders from nonresponders, 1316
 - only one response type represented in data, 1316
 - overlap of the two response types, 1317
 - properties of the maximum likelihood estimator illustrated, 1317
 - quasi-complete separation of responders from nonresponders, 1316–1317
 - discussion, 1318
 - maximum likelihood analysis, 1315–1316
 - results, 1317–1318
- logistic regression model, 76, 78, 82, 85, 89, 1303–1308, 1793, 1313
 - adjusted odds ratio in, 1557
 - British Institute of Radiology (BIR) trials of radiotherapy in cancer of the larynx and the pharynx, 1306–1308
 - four-parameter, 133f, 134
 - goodness-of-fit diagnostics, 1305–1306
 - interpretation, 1305
 - maximum likelihood estimation (MLE), 1304
 - pharmacodynamics with covariates, 1682–1683
 - and receiver operating characteristic curve, 1964
- logistic regression process, 1309–1313
 - model building, 1311
 - model diagnosis and internal validation, 1312
 - model selection, 1312
 - multivariate analysis, 1312
 - test for collinearity, 1311–1312
 - univariate analysis, 1311
 - model generalizability, 1312–1313
 - propensity score matching, 1310–1311
- logit analysis, 214
- logit confidence interval, 1555
- logit model, 426
- logit-threshold fixed effect model for summary receiver operating characteristic curve, 2102–2103
- log-likelihood function, 1603, 1684
- log-linearity assumption, assessment of, 1499
- log–log fit, 131, 132f
- log-rank test, 2179, 2181, 2181, 2182

pharmacodynamics with no covariates, 1689–1690
 statistic, 1574
 log scale of HR, 1579
 log scale of relative risk, 1579
 log-transformed data, 929
 longest vertical distance (LVD), 1142
 longitudinal bootstrap, 343, 346
 longitudinal data analysis (LDA) model, 416, 901, 1624
 longitudinal responses, 1732
 longitudinal studies, 901
 long-term toxicity study, 1723
 long-term variation, 2140
 losers, rules for dropping
 clinical trial simulations for later development phases, 546
 loss to follow-up, clinical trials, 575
 Lot Tolerance Percent Defective (LTPD), 1
 lovastatin, 1698
 lower control limit (LCL), 2137
 lubricants, 257

M

macular degeneration study, 1414
 magnetic resonance spectroscopy (MRS), 1610
 Mahalanobis distance, 855
 Mainland-Gart test, 208–210, 212
 SAS code and output for, 210f
 major adverse cardiac event (MACE), 815
 major depressive disorder (MDD), 1736
 malaria generics, 1191
 Malik's method, 1856
 mammography, 825, 2164–2165
 Mann–Whitney test, 79, 80
 Manski's "ignorance" results, 427
 Manski's risk difference bounds, 423
 Mantel formulation, 50
 Mantel–Haenszel test, 56, 83, 88, 705, 1555, 1982, 1618
 Mantel–Haenszel type estimator, 1958
 manufacturing process, 258
 compression, 258
 consistency in, 307, 308
 granulation, 258
 mixing, 258
 marginal density, estimation based on statistics, 1030–1031
 marginal maximum likelihood estimation via adaptive Gaussian quadrature
 correlated probit model, 745–746
 marginal models, for clustered categorical data, 90
 marker prognostic value, 1651
 Marker Sequential Test (MaST) design, 1577f
 marketing authorization application (MAA), 912
 Marketing Authorizations (MAs), 1343
 Markov chain Monte Carlo (MCMC)
 approach, 91, 169, 176, 183, 195, 226, 1271, 1358, 1372, 1411, 1697, 1983, 2092, 2107
 martingale residual-based estimators, 560
 Martingale residuals, 1839
 masking, 1793, *see also* blinding
 medical devices, 1350

mass spectrometry, 227
 Master's partial credit model, 1212
 matched-pair cluster randomized trials, 650
 matched-pairs design, 2170
 binary outcomes, 2171–2172
 continuous outcomes, 2171
 matching, 650
 Mathematica®, 113
 maximal tolerated dose (MTD), 1721
 maximal utility (MU) design, 1581
 maximum effective dose (MaxED), 885
 maximum likelihood (ML), 904, 1410, 1330
 analysis, 1315–1316
 early phases of clinical trials, simulation in, 542
 procedures, 100, 101
 maximum likelihood estimate (MLE), 316, 412, 953, 994, 1108, 1315, 1316, 1410, 1422, 1479, 1602, 2032, 2091, 2169, 2260, 2273, 2298
 correlated probit model, 744–745
 logistic regression, 1304
 method, 605
 properties of, 1317
 maximum tolerable dose (MTD) for cancer
 chemotherapy, 390, 539, 880, 885, 384–388, 1397, 1406, 1715
 continual reassessment method, 387–388
 definition, 384
 early phases of clinical trials, simulation in, 539
 general considerations, 384–385
 single-stage design, 385–386
 two-stage design, 387
 accelerated titration designs (ATDs), 387
 Design BD, 387
 maximum utility model (MUM)
 clinical trial simulations for later development phases, 547
 McNemar's test, 82, 83, 207–208, 212, 834, 1320–1324, 1691, 2155
 alternative tests, 1321–1322
 applications of, 1321
 calculation of the test statistic, 1320–1321
 characteristics of, 1323
 example, 1321
 generalizations and extensions of, 1324
 McNemar's test, applications of, 1321
 pharmacodynamics with no covariates, 1691
 related confidence intervals, 1323–1324
 sample size determination, 1322–1323
 SAS code and output for, 208f
 mean absolute deviation (MAD), 1784
 means, mean percentage change, and medians, 1260
 mean squared error (MSE), 614
 mean square deviation (MSD)
 measuring agreement, 1337
 measurement bias, 823
 measurement error
 consequences of, 1908–1909
 ecologic inference, 948
 Measurement Performance Validation of Diagnostic Devices, 1649, 1650, 1650
 measurement precision, 823
 measurement validation, 822, 823

measuring agreement, 1335–1340
 assay validation example, 1339
 methods, 1337
 asymptotic power, 1338
 concordance correlation coefficient (CCC), 1337
 coverage probability, 1338
 mean square deviation (MSD), 1337
 proportional error case, 1338
 simulation results, 1338
 total deviation index, 1337–1338
 methods comparison example, 1338
 models, 1336
 fixed target values, 1336
 random target values, 1336
 precision and accuracy, 1335f
 traditional approaches, 1335–1336
 Medical Device Innovation Consortium (MDIC), 473–474
 medical devices, 1348–1352
 diagnostic devices, 1353
 analysis issues for, 1354
 diagnostic design issues, 1353–1354
 measurement method comparison for diagnostic tests, 1354–1355
 regulation, 1349
 global regulatory efforts, 1350
 active control, choice of, 1351
 analysis, 1352
 causal inference, 1352
 center-by-treatment interaction, 1352
 historical controls, 1350–1351
 implants, 1350
 masking (blinding), 1350
 monitoring, 1352
 non-diagnostic medical device studies, design issues in, 1350
 noninferiority trials, 1351
 propensity scores, 1352–1353
 repeated measurements, 1352
 sensitivity analysis, 1353
 sham controls (the placebo effect), 1350
 subgroup analysis, 1352
 surrogates, 1351
 individual equivalence, 1355
 analytical studies, 1356
 Bayesian adaptive design, 1357
 Bayesian approach in evaluating medical devices, 1356
 Bayesian calculations, 1358
 Bayesian hierarchical modeling, 1358
 Bayesian interim analyses, 1357
 Bayesian noninferiority testing, 1357
 Bayesian sample size, 1357
 Bayesian statistics, 1356
 biomarkers, diagnostic tests, and microarrays, 1356
 diagnostic imaging, 1355
 historical controls as prior information, 1357
 monitoring devices, 1356
 valid prior information, 1357
 investigational device exemption, 1349
 vs. pharmaceutical drugs in the United States, 1349
 premarket approval applications, 1349
 premarket notifications, 1349

- medical devices, (*Cont.*)
 - regulation of medical devices vs. pharmaceutical drugs in the United States, 1349
 - regulatory history in the United States, 1348
 - statistical issues in the postmarket, 1358
- Medical Dictionary for Regulatory Activities (MedDRA), 815, 1342–1346
 - coding system, 53, 54, 54f–55
 - current state of MedDRA structure, 1342
 - data collection principles affecting analyses, 1345
 - investigational data presentation, 1344
 - MedDRA-encoded data and the concept of versioning, 1343–1344
 - post-marketing data presentation, 1344–1345
- medical literature, examples from, 2163
 - comparison of subgroups within the population, 2163
 - identification of subgroups to investigate, 2163–2164
 - mammography screening, 2164–2165
- medical professionals, attitude of, 719
- medicinal product development, 1721
- Medicines and Healthcare Products Regulatory Agency (MHRA), 1350
- MEDLINE, 229
- meiosis, 1058
- Mendelian diseases, 1062
- Mendelian inheritance, 1059
- mental health example, 1266–1267
- mergers and licensing, 820
- meta-analyses of individual patient data (MIPD), 1370, 1371
- meta-analysis, 1370, 1671, 603–611
 - based on data subsets, 1531
 - based on separate outcomes, 1530
 - based on summary outcomes, 1530–1531
 - basic steps in performing, 1367f
 - Bayesian methods in, 173–178
 - BCG vaccine clinical trials, 608
 - for bioequivalence, 1369
 - combining odds ratios in, 1558
 - in cost-effectiveness analysis, 1368
 - defined, 1362
 - of diagnostic test data, 2101–2102
 - effect-size calculation, 608–609
 - fixed-effects, 604–605, 609–610
 - of individual patient data, 1370–1371
 - meta-regression, 607, 610–611
 - for optimal dosing, 1368–1369
 - quantifying heterogeneity in, 606
 - H index, 606–607
 - I² index, 607
 - τ^2 index, 606
 - random-effects, 605–606, 610
 - for safety monitoring, 929, 930
 - steps in, 1529
 - correlation estimation, 1530
 - fixed or random effects, 1529–1530
 - summary statistics and the sources of variations, 604
 - and trend estimation, 2286, 2287
- metabolic networks, 227
- MetaCyc, 227
- meta-regression analysis, 607, 610–611
- therapeutic trials, meta-analysis of, 1366
- metastatic nasopharyngeal carcinoma
 - paclitaxel/carboplatin/gemcitabine in, 160–161
- metered-dose inhalers (MDIs), 970
- method-of-moment estimators, 652
- method of steepest ascent, 1939
- metrics, 534
- microarray gene expression, 1377–1390
 - classification and prediction, 1385
 - model building, 1385
 - performance assessment, 1385
 - cluster analysis and pattern recognition, 1386
 - dimension reduction analysis and visualization, 1389–1390
 - gene expression matrix map with hierarchical clustering, 1386–1388
 - gene network modeling and integrated system biology, 1390
 - non-hierarchical clustering, 1388–1389
 - data preprocessing and normalization, 1381–1382
 - DNA microarray experiment, 1377, 1378f
 - experimental design, 1379
 - sample size estimation, 1381
 - types of experimental design, 1379–1380
 - variability, replication, and experimental unit, 1379
 - gene set analysis (GSA), 1384
 - identifying differentially expressed genes (gene selection), 1382
 - assigning a significance level, 1383–1384
 - gene ranking, 1382
 - microarray technology, 1378
 - cDNA microarray, 1378
 - oligonucleotide microarray, 1378
- microarrays, 226–227
 - based comparative genomic hybridization (array-CGH), 1379
- microarray technology, 1378
 - cDNA microarray, 1378
 - oligonucleotide microarray, 1378
- microdose vs. placebo, 1397
- microdosing studies, 1394
 - prediction of microdose, 1398
 - predictability index, 1398–1399
 - sample size requirement, 1399
- regulatory perspectives, 1395
 - FDA guidance, 1397
 - ICH guideline, 1395–1397
- scientific factors, 1397
 - animal model vs. human model, 1398
 - dose–response curve, characterization of, 1398
 - false negative and false positive, 1398
 - microdose vs. placebo, 1397
 - therapeutic dose, selection of, 1397–1399
 - variability associated with dose, 1398
- micronucleated, binucleated (MNBN) cells, 1156
- micronucleus test, 2234
- Miettinen's asymptotic formula, 1322
- minimal total sample size (MTSS) design, 1581
- minimization procedure, 705, 1401–1404
 - comparison with other procedures, 1403
 - discussion, 1403–1404
 - procedure and algorithm, 1402–1403
- minimum distance estimators, 560
- minimum effective dose (MED), 736, 885, 1397, 1406–1407, 1722
 - design and analysis, 1406–1407
- Ministry of Health, Labour and Welfare (MHLW), 287–289
- MinRMS, structural alignment methods, 228
- min test, 552
- missing at random (MAR), 902, 1422, 1423, 1918, 1919, 1924, 1925
- missing completely at random (MCAR), 902, 1422, 1639, 1918, 1924
- missing-data mechanism, 416, 902, 1732
- missing not at random (MNAR), 1422
- missing pattern, 2013
- missing values
 - analysis of repeated measures data with, 1924–1929
 - in repeated measurement designs (RMD), 1918–1922
 - general approaches to missing data, 1919
 - incomplete repeated measures data analyses, 1919–1921
 - missing data mechanism, 1918–1919
 - use of proxy information, 1921–1922
- mixed carryover effect, 613
- mixed-effects model repeated measurement (MMRM) methodology, 903, 1422–1426
 - estimating G and R in, 1423
 - estimating β and γ in, 1424
 - example, 1424–1425
 - general linear mixed model, 1423
 - statistical properties and inferences, 1424
- mixed-effects models, 1330, 1409–1412
 - clinical pharmacology, 470, 471
 - dose response analysis in clinical trials, 880
 - individual-level analysis with cluster trials, 652–653
 - linear models, 1409
 - background, 1409–1410
 - extended generalized least squares (EGLS), 1411
 - maximum likelihood estimation, 1410
 - REML estimation, 1411
 - non-linear models, 1412
 - small sample adjustments, 1411–1412
 - software, 1412
- mixed factorial design in biomedical research, application of, 1025
 - allele-specific polymerase chain reaction for screening point mutation (example), 1025
 - drug interaction in a clinical study (example), 1026
- mixed logistic model, 652
- mixed model for repeated measurements (MMRM), 842
- mixed-outcome data, 1414–1420
 - factorization models, 1415
 - asymmetry in treatment of outcomes, 1417
 - direction of conditioning, 1417

- joint mixed-outcome models, 1415
 - model estimation, 1420
 - Plackett-Dale distribution, 1419
 - random-effects models, 1417
 - correlated random effects, 1418
 - random effects plus correlated errors, 1418–1419
 - shared random effects, 1417–1418
 - mixed-up randomization codes, effect of, 1877–1878
 - mixing, 258
 - mixture experiments, 1946–1947
 - ModBase, 228
 - model diagnosis and internal validation, 1312
 - model generalizability, 1312–1313
 - model-independent MSD methods for
 - dissolution profile comparison, 855
 - Hotelling's T^2 -based approaches, 855
 - Sarandasa and Krishnamoorthy's (SK) method, 855–856
 - model selection, 1312
 - calibration, 370–371
 - Modernization Act of 1997, 537
 - modified Balaam's design, 926
 - modified large sample (MLS) method, 486, 494, 1427, 1742
 - application of, 1431–1432
 - confidence intervals, 1256
 - extension
 - to difference of variance components, 1428–1430
 - when estimators are dependent, 1430–1431
 - to ratio of variance components, 1430
 - main idea of, 1428
 - modified two- and threefold rules, 70
 - moment based criteria, for similarity, 306
 - Monitoring Committee (DMC), 1197
 - monotherapy maximum tolerated dose,
 - designs to identify, 1561
 - designs based on time to first toxicity, 1564
 - discussion and recommendations, 1564–1565
 - dose level and schedule, designs that optimize, 1564
 - dose-limiting toxicity (DLT), designs based on, 1561
 - interval designs, 1563
 - model-based designs, 1561–1563
 - rule-based designs, 1561
 - ordinal toxicity, designs based on, 1563–1564
- Monte Carlo EM (MCEM) algorithm, 91
- Monte Carlo (MC) methods, 91, 519, 521, 1059, 1973, 1975, 1978
- Monte-Carlo Simulation-Based simulations, 524
- mortality end points, 1696
- mortality-independent tumors, 403
- Mosteller's data, 208
- most powerful test, 1124–1125
- moving range (MR) chart, 2138
- moxifloxacin, 2209
- MRC Biostatistics Unit, 195
- multiarmed trials, extension to, 1116–1117
- multiaxial coding, 54–55f
- multicategory kappa, 1242
- multicenter trials, 576, 1434–1437
 - organizational structure and decision making, 1435
 - protocol development and implementation, 1434–1435
 - recent developments, 1436
 - regulatory requirements, 1435
 - scientific issues, 1436
 - statistical analysis, 1436
 - statistical design, 1435
- multicollinearity, 1439–1448
 - accounting for, 1444
 - alternative ways to go about issues in, 1444
 - prognostic factor analysis using QOL measures, 1444–1448
 - bootstrap resampling, 1444
 - condition numbers, 1444
 - correlations, 1444
 - determinant, 1444
 - diagnostic tool kit, 1444
 - discussion, 1448
 - genetic association studies and epistasis, 1444, 1448
 - harmful, 1444
 - imperfect, 1444
 - prognostic factor analysis using quality of life (QOL) measures, 1439–1444
 - two faces of the same coin, 1444
 - variance inflation factors (VIFs), 1444
- multi-decision criteria analysis, 478–480
- multidimensional data analysis, 1448
 - bias reduction methods, 1450
 - adaptive Lasso and elastic net, 1451
 - combined penalty (CP) function, 1450–1451
 - relaxed Lasso, 1451
 - SCAD and its oracle property, 1450
 - two-step strategy, 1451
 - traditional penalized methods, 1448
 - extensions or alternative to Lasso, 1450
 - L_0 penalty, 1448–1449
 - L_1 penalty, 1449
 - Lasso and its extensions, 1449–1450
 - tuning and feature selection issues, 1452
- multi-dimensional scaling (MDS), 1389
- multilevel modeling (MLM) framework, 1534
- multi-level model theory for ecological inference, 945
- multiple ascending dose (MAD), 2212
- multiple comparison procedures and modeling (MCP-Mod), 1326, 1724
 - design aspects, 1331
 - determination of total sample size, 1331
 - selection of candidate set of models, 1331
 - selection of doses to utilize in trial, 1331
 - dose-finding and, 1326
 - for general data, 1330
 - literature review and historical overview, 1328
 - for normally distributed data, 1328
 - scope of, 1332
 - software implementation, 1332
- multiple comparisons procedures (MCPs), 880, 1502, 1625, 2237–2238
 - error criteria for, 1028
 - definition, 1028–1029
 - group sequential procedures, 1117–1118
- multiple-dose bioequivalence studies, 265, 1477
 - results and discussion, 1479–1480
 - statistical methods, 1478–1479
- multiple-dose combinations, combination drug clinical trial, 553–554
- multiple-dose escalation studies, clinical pharmacology, 470
- multiple endpoints, 1467, 1625
 - graphical technique, 1472
 - inferring the global superiority of the treatment from combined evidence of all endpoints, 1471–1475
 - P -value-based techniques, 1468–1469
 - regulatory agencies and the multiple-endpoint issue, 1475
 - techniques based on distribution/correlation structure of endpoint, 1470
 - techniques to analyze categorical and ordinal endpoints, 1470–1471
 - false discovery rate, 1471
- multiple hypotheses testing principles, 732
 - closed testing principle, 732–733
 - gatekeeping testing strategies, 734
 - global testing, 732
 - hierarchical testing principle/framework, 733
 - partitioning principle, 733
 - union-intersection (UI) and intersection-union (IU) testing principles, 732
- multiple imputation (MI) methods, 1422, 1926–1927
 - statistical model, 1129–1130
- multiple-look problem, 577
- multiple sclerosis, 2031
- multiple-stage designs for phase II cancer trials, 1481–1491
 - binomial response, 1485
 - Fleming multi-stage design, 1485–1486
 - Gehan's two-stage design, 1485
 - randomized discontinuation design (RDD), 1489
 - Simon's two-stage designs, 1486–1487
 - three-stage designs, 1487
 - hypothesis testing, 1482
 - binomial response, 1482
 - trinomial response, 1482–1483
 - single-stage design, 1483
 - binomial response, 1483
 - trinomial response, 1483–1484
 - trinomial response, 1489–1491
- multiple testing, 1028
- multiple time points, 1625
 - cluster trials, 655
- multiplicative intensity (MI) models, 1494
 - estimation and hypotheses testing, 1496
 - estimating the cumulative intensity functions, 1498
 - parametric multiplicative intensity models, 1496–1497
 - semiparametric multiplicative intensity models, 1497–1498
 - with frailty, 1500
- model assessment and diagnostics, 1498
 - checking parametric baseline intensity form, 1500
 - influential observations, assessment of, 1498–1499

- multiplicative intensity (MI) models, (*Cont.*)
 - log-linearity assumption, assessment of, 1499
 - multiplicative intensity models with frailty, 1500
 - omnibus goodness-of-fit methods, 1499–1500
 - proportional intensity assumption, 1499
- parametric multiplicative intensity models, 1495–1496
- semiparametric multiplicative intensity models, 1496
- multiplicity, 727–728, 1852
 - in clinical trials, 585, 734, 1502–1509
 - approaches for reducing the burden of multiplicity, 735
 - Bayesian statistics, 1504–1505
 - clinical decision rules, 734–735
 - closed testing methods and applications to clinical trials, 1507
 - composite endpoint issues, 735
 - controversies, 1505–1508
 - dose–response and dose–control comparison trials, 736
 - drug combination trials, 736
 - fixed sequence of or a priori ordered clinical objectives, 1507–1508
 - gatekeeping procedures for primary and secondary endpoints, 1508
 - primary and secondary endpoint concepts, 735
 - problem, 1503
 - skepticism, 1503–1504
 - subgroup analysis, 736–737
 - three-arm “gold standard” trials, 736
 - control procedures, 1578
 - methods for addressing, 729
 - adaptive alpha allocation approach (4A) and the consistency ensured methods, 731
 - Bonferroni test, 729
 - Dunnett’s test, 731
 - exact tests of contrasts, 731
 - fixed and flexible fixed-sequence (fallback) methods, 730–731
 - Holm’s test, 729
 - Hommel and Hochberg tests, 730
 - prospective alpha allocation scheme (PAAS) and Šidák’s test, 730
 - Scheffe’s procedure, 731
 - Šidák’s test, extensions of, 730
 - Tukey and Tukey–Kramer procedures, 731
 - penalty, 965
- Multipoint Engine for Rapid Likelihood Inference (MERLIN), 1061
- multireader multicase (MRMC) study, 825
- Multi-Regional clinical trials (MRCT), 603, 1511–1520, 1573, 1578
 - adaptive design for, 1581
 - consistency requirement for regional approval, 1578
 - discussions, 1520
 - example, 1518–1520
 - issues in, 1511
 - multicenter trials, 1512
 - multiregional, multi-center trials, 1512–1513
 - optimal design for, 1580
 - statistical methods, 1513
 - achieving reproducibility and/or generalizability, 1515–1516
 - applicability of those approaches, 1517–1518
 - Bayesian approach, 1516–1517
 - consistency index, assessment of, 1513–1514
 - consistency, test for, 1513
 - Japanese approach, 1517
 - sensitivity index, evaluation of, 1514–1515
- multistage liquid impringer (MSLI), 1144
- Multitrait/Multi-item Analysis Program—Revised (MAP-R), 1176
- Multivariate analysis of variance (MANOVA), 912, 1934, 1312
- Multivariate clustered ordinal model
 - for clustered categorical data, 90–91
 - latent mixed models, 90–91
 - marginal models, 90
- Multivariate interval-censored data, statistical analysis of, 593
- Multivariate meta-analysis, 1528–1534
 - Multivariate approaches, 1531
 - Bayesian models, 1533–1534
 - frameworks, 1534
 - frequentist fixed-effects approaches, 1531–1532
 - frequentist random-effects models, 1532–1533
 - notation, 1531
 - sources of multivariate data, 1528–1529
 - steps in a meta-analysis, 1529
 - correlation estimation, 1530
 - fixed or random effects, 1529–1530
- univariate approaches, 1530
 - based on data subsets, 1531
 - based on separate outcomes, 1530
 - based on summary outcomes, 1530–1531
- Multivariate normal distribution, 169
- Multivariate probit model, 742
- Multivariate regression, non-parametric regression, 1551
- Multivariate responses, 1732
- Multivariate skew normal (MVSN), 165
- Multivariate skew t (MVST) distribution, 165, 169
- Multivariate statistical distance (MSD) tests, 854
- Multivariate ZIP (MZIP) model, 317
- Musculoskeletal system, 1666
 - osteoarthritis, 1666
 - rheumatoid arthritis, 1666
- Mutagenicity, 67, 68, 71, 72, 73
- Mutagenic mechanisms, 67
- Mutation tests, 825
- MutDB function, 228
- Mutual Acceptance of data (MAD), 1086
- Mutually exclusive programs, cost-effectiveness analysis, 1711–1713
- Myelin oligodendrocyte glycoprotein (MOG)-induced experimental autoimmune encephalomyelitis (EAE) mouse study, 1245, 1247
- Myocardial infarction data, prognostic effects for, 2059
- Myoinositol, 1610
- N
 - naive analysis, 966
- Narrow therapeutic index (NTI) drugs, 220, 2038
 - mean comparison, 220–221
 - scaled average bioequivalence (SABE) for, 2042
 - within-subject variability comparison, 220–221
- Narrow-therapeutic-range (NTR) drugs, in Canada, 266
- Narrow therapeutic window, drugs with, 888
- National Academy of Sciences, 1648
- National and international registries, 1791
 - control formed from registries, 1793
 - covariate balance diagnosis, 1794
 - covariate selection, 1793
 - missing data in covariates, 1794
 - sample size and power, 1794
 - selection of subjects from registry, 1794
- general considerations, 1792
 - data quality and relevance, 1792
 - registry vs. clinical study, 1792
 - potential opportunities, 1791
- National Cancer Institute (NCI), 539
- National Center for Biotechnology Information (NCBI), 224
- National Center for Toxicological Research (NCTR), 233
- National Committee for Clinical Laboratory Standards (NCCLS), 1276
- National Death Index, 575
- National Institute of Health (NIH), 224, 230, 233
- National Regulatory Authorities (NRAs), 283–284
- National Toxicological Program (NTP), 408, 410
- Naturally occurring cell death, 2123
- Negative binomial (NB) regression, 2031
 - alternative approaches, 2035
 - Cook’s formulae for NI trials, 2035
 - sample size assuming equal follow-up times, 2035
 - sample size for NI trials based on rate difference, 2035
 - sample size under heterogeneous dispersion, 2036
 - sample size using variance under null hypothesis, 2035
 - analysis of count data, 2032
 - NB distribution, 2031
 - non-inferiority, NB rates in, 2031
 - numerical illustrations, 2036
 - sample size bounds, 2034
 - sample size for equivalence trials, 2033
 - sample size for NI trials, 2033
 - sample size for superiority trials, 2032
 - superiority, NB rates in, 2031
- Negative likelihood ratio (NLR), 822, 824
- Negative percent agreement (NPA), 824
- Negative predictive values (NPV), 822, 824, 2184

- Nelson–Aalen estimator, 556, 557, 557, 1232, 1232, 2181
- nervous system, 1666
- Alzheimer's disease, 1666
 - Parkinson's disease, 1666
- “nested-crossover” alternative design, 2211
- neural network, 808
- new chemical entity (NCE), 1671
- new drug application (NDA), 57, 125, 180, 305, 721, 886, 912, 970, 1076, 1343
- new drug approvals, 1622
- new drug, dose-selection process for, 885–886
- new drug submission (NDS), 912
- new formulation, drugs and, 889
- Newman–Keuls test, 1689
- Newton–Raphson algorithm, 498, 994, 2011
- Neyman–Pearson formulation, 102, 1125, 1846
- of hypothesis testing, 1124–1125
- “no difference in pharmacokinetic properties”, 471
- no-effect dose, clinical pharmacology, 469
- N*-of-1 design analysis, 564–571
- complete *N*-of-1 design, 565
 - with four periods, 566
 - with four periods vs. RRRR/RTRT design, 567–569
 - power analysis, 570–571
 - relative efficiency, 566–567
 - with three periods, 565–566
 - with three periods vs. 2×3 dual design, 569–570
 - with two periods, 565
 - with two periods vs. 2×2 crossover design, 569
 - “no go” decision, 1723
- nominal category, 75
- nonantiarrhythmic drugs, 2206
- non-Bayesian flexible methods, 1327
- noncentral F-distribution, 1626
- nonclinical development, 908
- nonclinical drug development, 909–910, 1721
- drug formulation development, 910
 - pharmacology, 909–910
 - toxicology/drug safety, 910
- non-diagnostic medical device studies, design issues in, 1350
- non-hierarchical clustering, microarray gene expression, 1388–1389
- nonignorable missing data (NMD), 1918, 1919
- nonignorable missing (NIM) mechanism, 1924, 1925
- non-inferiority (NI), 525
- noninferiority analysis in active controlled clinical trials, 23
- challenges, 25–26
 - non-inferiority trial, objectives of, 23–24
 - statistical methods, 24–25
- non-inferiority hypotheses
- of active control trails
 - with fixed δ margin, 15–16
 - with fraction of control effect to be retained, 17–18
 - transformation of, 10
- non-inferiority margin, 15, 16, 18, 19, 634, 2032
- medical devices, 1351
- noninferiority/superiority
- similarity, test for, 494
 - test for, 493
- non-inferiority test procedure, 11
- non-inferiority trial, 238, 1536–1541, 2032
- Cook's formulae for, 2035
 - differences between the three methods, 1537–1538
- fixed-margin method
- two confidence interval method with a fixed margin, 1536–1537
 - two confidence interval method with a random margin, 1537
- hypotheses in, 1125
- regulatory concerns about the conduct of, 1669
- sample size for, 2033
- synthesis method, 1537
- type I error rates, 1538
- conditional across-trial type I error rate, 1540–1541
 - unconditional across-trial type I error rate, 1539–1540
 - within-trial type I error rate, 1538–1539
- non-informative prior distribution, 176, 193, 193
- nonlinear calibration, 133–134
- nonlinear models, 774
- non-linear regression, on revertant counts, 71
- NONMEM (nonlinear mixed-effects modelling), 1750
- non-monotonicity, 72
- nonoral dosage, 259
- non-parametric Bayesian methods, 1327
- nonparametric bootstrap, 343, 345
- nonparametric Dirichlet process, 92
- non-parametric maximum likelihood estimator (NPMLE), 591
- non-parametric methods, of mutagenicity and toxicity, 70–71, 72
- non-parametric regression, 1543–1552
- analyzing sample of curves, 1546
 - dynamic time warping, 1547–1549
 - shape-invariant modeling, 1547
 - structural analysis, 1547
 - application to brain potentials, 1550–1551
 - application to growth data, 1549
 - dose response analysis in clinical trials, 880
 - kernel density estimation and multivariate regression, 1551
 - density estimation, 1551
 - multivariate regression, 1551
- linear estimators, 1543
- kernel estimators, 1543–1544
 - local polynomials, 1545
 - smoothing splines, 1544–1545
 - wavelet estimators, 1545–1546
- software, 1552
- non-parametric survival models, 2181
- nonprofile analysis, *in vitro* bioequivalence testing, 1145
- non-randomized confirmatory study, 1826
- non-randomized trials, cluster trials, 655
- non-small-cell lung cancer, 825
- non-ST elevation myocardial infarction (NSTEMI), 966
- nonstratified randomization, 1221
- No Observed Effect Level (NOEL), 2233
- normally distributed data, detection of outlier in, 502
- test for extreme values, 502–503
 - test for residuals in regression models, 503
- normal random effects proportional odds model, 1617
- North America, 286
- Canada (Health Canada), 287
 - US Food and Drug Administration, 286–287
- Northern California Kaiser Permanente Medical Care Program, 638
- “no safety concern”, 471
- not missing at random (NMAR), 842
- no-treatment concurrent control, 622
- nuclear magnetic resonance (NMR), 131
- Nucleic Acid Database, 228
- nucleic acid testing (NAT) methods, 235
- null hypothesis, 18, 21, 319, 524, 2209
- null model vs. model approach
- clinical trial simulations for later development phases, 547
- number needed to harm (NNH), 477
- number needed to treat (NNT), 477, 1957
- O**
- O'Brien–Fleming boundary, 1108, 1110, 1198
- O'Brien–Fleming group sequential method, 60f, 61, 62, 63f–64f
- O'Brien's test, 1473, 1638
- observational postmarket study, 1826
- observational studies, of covariance, 772
- observation carried forward (BOCF), 1422
- observed information matrix, 1837
- “obvious changes”, data validation, 448
- Ochi and Prentice's method, 743
- odds ratio, 423, 432, 1241, 1242, 1243, 1425, 1554–1558
- asymptotic inferences, 1555
 - combining in stratified or meta-analyses, 1558
 - definition, 1554–1555
 - Exact Inferences, 1555–1556
 - local and cumulative odds ratio, 1558
 - in logistic regression models, 1557
 - numerical example, 1556
 - relationship with other measures, 1556
 - sample size calculations, 1557
 - in therapeutic equivalence clinical trials, 1557
- ointments and skin preparations, 257
- olanzapine, 191, 192, 194
- oligonucleotide microarray, 1378
- omnibus goodness-of-fit methods, 1499–1500
- Omnitrope®, 305, 309
- omphalitis, 655
- oncology, 1667
- drug development, 1573
 - novel designs in, 819
- oncology clinical trials, 1560–1568
- designs based on ordinal toxicity, 1563–1564
 - designs based on time to first toxicity, 1564
 - designs that optimize dose level and schedule, 1564

- oncology clinical trials, (*Cont.*)
 designs to identify the combination MTD, 1567
 example of partial orders, 1567–1568
 partial order CRM, 1567
 designs to identify the recommended phase II dose, 1565
 designs that balance toxicity and efficacy, 1566
 designs that incorporate biomarker/pharmacodynamic data, 1567
 discussion and recommendations, 1567
 discussion and recommendations, 1564–1565, 1568
 dose-limiting toxicity (DLT), designs based on, 1561
 interval designs, 1563
 model-based designs, 1561–1563
 rule-based designs, 1561
 recent trends in dose finding, 1568
 oncology trials, development strategy, 1572
 adaptive strategies for, 1573
 specific issues with time-to-event endpoint, 1574
 SSR design, 1573
 biomarker-driven design with established biomarkers, 1575
 modified stratified design, 1577
 phase III adaptive design strategy, 1577
 phase III fixed design strategy, 1575
 stratified designs, 1575
 Multi-Regional clinical trials (MRCT), 1578
 adaptive design for, 1581
 consistency requirement for regional approval, 1578
 optimal design for, 1580
 one-parameter gamma family, 62
 one-parameter logistic (IPL) model, 1211
 one-sample Wilcoxon's test, 1972
 example, 1973–1975
 fixed alternative, 1973
 local alternative, 1973
 simulation study, 1973–1974
 one-sided equivalence and non-inferiority trials, 2199–2201
 one-sided modified T^2 test, 1474
 one-sided testing vs. two-sided testing, 1098, 2238
 one-sided/two-sided P -values, 1852
 one-way analysis of covariance, 1681–1682
 one-way ANOVA model, 105, 1688–1689
 with repeated measures, 1677–1678
 validity of LOCF test under, 1297–1298
 onset of action, 1585–1591
 statistical considerations and approaches, 1585
 prespecified early time point, analysis at, 1585–1588
 repeated measures analysis, 1588
 survival analysis, 1588–1589
 survival analysis using mixture models, 1590–1591
 onset-rate method, 403
 open label trials, 336
 operating characteristic (OC) curve, 2*f*, 5*f*, 709, 1807, 2311
 operational qualification (OQ), 1072
 opioid drugs, 1121
 OPLS force fields, 229
 optical character recognition (OCR) process, clinical data management guidelines, 447
 optimal allocations, 51, 1951
 optimal dosing, meta-analysis for, 1368–1369
 optimal futility interim design, 1593–1600
 discussions, 1600
 frequentist operational characteristics (FOC)-based approaches, 1593
 randomized trial with time to event end point, 1594–1595
 randomized two-arm study with normal outcome, 1594
 single arm study with binary end point, 1593–1594
 probability of success (POS) based approaches, 1595
 decision rule based on CPOS, 1596
 decision rule based on PPOS, 1596
 probability of success, 1595–1596
 randomized trial with normal end point, 1598–1599
 randomized trial with time-to-event end point, 1596–1598
 single arm study with binary end point, 1599–1600
 optimality, 373
 optimal therapeutic effect, 317
 optimal variable screening, 2271–2272
 oral hygiene data (case study), 95
 ordered categorical data, Kruskal–Wallis test for, 1245
 ordered categories, kappa for, 1243
 ordered multiple class receiver operating characteristic (ROC) analysis, 1608–1611
 application, 1610
 ROC curve definition and construction, 1608
 ROC surface and hypersurface, 1609–1610
 ordinal data under a crossover design, 1613–1619
 AB/BA crossover design, 1614
 distribution-free random effects model with using a conditional approach, 1615–1617
 example, 1617
 normal random effects proportional odds model, 1614
 simple change score method, 1615
 four-treatment four-period crossover design, 1619
 incomplete block crossover design, 1619
 three-treatment three-period crossover design, 1617
 distribution-free random effects model with using a conditional approach, 1618–1619
 example, 1619
 normal random effects proportional odds model, 1617
 simple change score method, 1617–1618
 ordinal endpoints, dose response analysis in clinical trials, 882
 ordinal outcomes, 89
 ordinary differential equations (ODE), 2126
 ordinary least squares (OLS) approach, 784
 organization, 1092
 orgasm, 1171
 orthogonal arrays, 2045
 orthogonal designs, 1014
 for 2^n factorial experiments, 1015
 interactions between and within subgroups (case), 1020
 interactions between subgroups (case), 1017–1020
 interactions within subgroups (case), 1016–1017
 for 3^m factorial experiments, 1020–1022
 for mixed $2^n \times 3^m$ factorial experiments, 1022–1025
 for s^m factorial experiments, 1015
 orthogonal main effect plans, 2045
 orthogonal polynomials, dose response analysis in clinical trials, 880
 orthogonal vs. oblique rotation, exploratory factor analysis, 986
 ossification data, 996
 osteoarthritis, 1666
 osteoporosis, 1667
 outlier detection in clinical research, 502–509
 in crossover model, 503
 likelihood distance test, 503–504
 test based on Hotelling T^2 statistic, 504–505
 example, 507–509
 under generalized linear mixed models, 505
 formulation of GLMM, 505
 individual observations, diagnostics for, 506–507
 outlying observations, detection of, 505–506
 in normally distributed data, 502
 test for extreme values, 502–503
 test for residuals in regression models, 503
 outliers, analysis of, 1260–1261
 overall coherence rate, 2212
 overall survival (OS) event analysis, 1574
 overdispersion, 70, 992, 2031
 over-representation analysis (ORA), 1384
 overruling, 62, 65
 over-the-counter (OTC) market, 717
 oxycodone, 1121
 oxymorphone, 1121
- P**
 Pacific Symposium on Biocomputing (PSB), 224
 paddle apparatus, 261
 pain management, 1121
 paired t -test, pharmacodynamics with no covariates, 1686–1687
 pairwise sequence comparison, 225
 parallel design, 1621–1626
 analysis issues, 1624
 baseline information, 1624–1625
 combination treatments, 1625
 multiple comparison, 1625
 multiple endpoints, 1625
 multiple time points, 1625
 sample size, 1626

- definition, 1621–1622
 - examples of, 1623
 - completely randomized design, 1623
 - designs with prognostic factors, 1623–1624
 - experimental designs, three basic concepts in, 1622–1623
 - with replicates, 487, 491, 495
 - test for equality, 487, 491–492, 495
 - test for noninferiority/superiority, 487, 492, 495–496
 - test for similarity, 487–488, 492–493, 496
 - without replicates, 494
 - test for equality, 494
 - test for noninferiority/superiority, 494–495
 - test for similarity, 495
 - parallel flats designs (PFD), 1014
 - parallel group designs, 842
 - dose response analysis in clinical trials, 882
 - parallel-groups, fixed-dose study, 2225
 - parallel-line assay, 135, 137, 311
 - parallel-line bioassays, 213–214
 - parallel testing strategy, 1576
 - parameter, assessment of
 - early phases of clinical trials, simulation in, 542
 - parameter, prior distribution of
 - early phases of clinical trials, simulation in, 542
 - parametric baseline intensity form, checking, 1500
 - parametric bootstrap, 343, 345, 346
 - parametric estimation problems, 1330
 - parametric method vs. nonparametric method, 2238
 - parametric modeling, of mutagenicity and toxicity, 71, 73
 - parametric multiplicative intensity models, 1495–1497
 - parametric time-to-event models, 1330
 - Pareto charts, 2136f
 - Parkinson's disease, 1666
 - partial credit model, 1212
 - partial likelihood, 1836
 - partially stochastic analyses, 1699
 - particle size, 255, 260
 - particle size distribution (PSD), 1817
 - partition coefficient, 255
 - partitioning principle, 733
 - Partners for Prevention Project (PARTNERS), 649, 650
 - Pasteur, Louis, 227
 - Patel's data without the outlier, 794, 795
 - inferences in the presence of baseline values, 794–795
 - inferences without the baseline values, 794–795
 - pathogen strains, evaluation of vaccines against, 240
 - Patient Centered Benefit-Risk (PCBR), 473–474
 - Patient-Centered Outcomes Research Institute (PCORI), 756
 - patient characteristics, 1653
 - patient preference information (PPI), 473–474
 - patient randomization using interactive voice/web response system, 1220
 - patient registry, 1791
 - patient-reported outcomes (PROs), 1633–1641
 - analytical considerations, 1640
 - cross-sectional analysis, 1634–1635
 - defining the endpoint, 1634
 - longitudinal analysis, 1637
 - missing data, 1638–1640
 - multiple testing, 1637–1638
 - in regulatory setting, 1633–1634
 - reporting of results, 1640–1641
 - summary measures and statistics, 1635
 - area under the curve, 1635
 - responder analysis, 1636
 - time to significant worsening or improvement, 1636–1637
 - Patient Satisfaction with Insulin Therapy (PSIT) Questionnaire, 987
 - Pearson correlation coefficient, 1175, 1176, 1335f, 1337, 1386
 - Pearson residuals, 1305
 - Pearson's chi-square test, 337, 651, 994, 997, 1555, 1558, 583, *see also* chi-square test
 - goodness-of-fit, 999
 - pharmacodynamics with no covariates, 1691–1692
 - pediatric population, pharmacodynamic issues, 1670
 - peptic ulcer disease, 1667
 - percent accepted mutation (PAM) matrix, 225
 - percentage reduction, 1122
 - percentile charts on correlated measures, 1643–1647, 1646f
 - Box-Cox transformation, 1643
 - correlated response variables, 1644–1645
 - example, 1645–1646
 - quantile regression, 1644
 - percentile type confidence sets, 344
 - perfect match (PM) probe, 1378
 - performance characteristics, 375–376
 - performance comparison and design optimization, 2028–2029
 - performance qualification (PQ), 1072
 - performing, key concepts in, 1709
 - period effects, clinical pharmacology, 470
 - Perlman's likelihood ratio test, 1474
 - permeability, 255
 - permuted-block randomization, 1876–1877
 - per protocol subjects, 417
 - personalized medicine, 964, 1648
 - Measurement Performance Validation of Diagnostic Devices, 1650
 - precision medicine vs., 1780–1781
 - risk prediction, 1652
 - treatment selection, 1650
 - adaptive signature, 1652
 - biomarker-stratified design, 1651
 - bridging study, 1652
 - prescreening bias, 1652
 - targeted design, 1651
 - PEST software, 62
 - Peto-Peto Prentice Wilcoxon test, 50, 407
 - pharmaceutical flow chart, 2137
 - pharmaceutical industry, contract research organization (CRO)
 - increase of outsourcing by, 718
 - pressures from early 1990s, 716–717
 - Pharmaceutical Research and Manufacturers of America (PhRMA), 1202, 1327, 474
 - pharmaceutical studies, cluster trials in, 655
 - pharmacodynamic issues, 1665–1671
 - cardiovascular system, 1666
 - congestive heart failure, 1666
 - coronary artery disease/angina, 1666
 - hyperlipidemia, 1666
 - hypertension, 1666
 - clinical trial, regulatory concerns about the conduct of, 1669
 - all randomized patients, 1670
 - equivalence trials, 1669
 - handling missing values and outliers, 1670
 - non-inferiority trials, 1669
 - per-protocol patients, 1670
 - superiority trials, 1669
 - distribution of, 1668
 - endocrine system, 1667
 - diabetes, 1667
 - osteoporosis, 1667
 - factors affecting the selection of statistical methods for, 1667
 - frequency of evaluation of the, 1668–1669
 - gastrointestinal system, 1667
 - peptic ulcer disease, 1667
 - infectious diseases, 1667
 - measurements of, 1669
 - meta-analysis, 1671
 - model assumptions, 1668
 - musculoskeletal system, 1666
 - osteoarthritis, 1666
 - rheumatoid arthritis, 1666
 - nervous system, 1666
 - Alzheimer's disease, 1666
 - Parkinson's disease, 1666
 - oncology, 1667
 - pharmacokinetic/pharmacodynamic relationships, 1671
 - respiratory system, 1666
 - asthma, 1666
 - skin, 1666
 - acne, 1666
 - psoriasis, 1666
 - study objective and design, 1667
 - subgroup considerations, 1670
 - elderly population, 1670
 - gender differences, 1671
 - impaired renal/liver function, patients with, 1671
 - pediatric population, 1670
 - racial differences, 1670–1671
 - surrogate markers, 1670
 - therapeutic area of study drug, 1668
 - transplantation, 1667
 - type and number of explanatory variables and response variables, 1667–1668
- pharmacodynamic measures, of
 - bioequivalence, 264, 266–268
- pharmacodynamic model, 1751
 - structural model, choice of, 1751–1752
- pharmacodynamic property, 886
- pharmacodynamics (PD), 1121, 1721
 - clinical pharmacology, 470

- pharmacodynamics with covariates, 1673–1684
 - statistical methods
 - between-treatment-groups analysis, 1679–1680
 - Cochran-Mantel-Haenszel test, 1682
 - Cox proportional hazards analysis, 1683–1684
 - crossover design, 1676–1677
 - logistic regression, 1682–1683
 - one-way analysis of covariance, 1681–1682
 - one-way analysis of variance with repeated measures, 1677–1678
 - simple linear regression, 1678
 - two-way analysis of variance, 1673–1676
 - within-treatment-group analysis, 1678–1679
- pharmacodynamics with no covariates, 1685–1692
 - statistical methods, 1685
 - (Pearson's) Chi-square test, 1691–1692
 - Fisher exact test, 1690–1691
 - Kruskal-Wallis test, 1689
 - log-rank test (Mantel-Cox test), 1689–1690
 - McNemar's test, 1691
 - one-way analysis of variance, 1688–1689
 - paired *t*-test, 1686–1687
 - two-sample *t*-test, 1685–1686
 - Wilcoxon's rank-sum test, 1687
 - Wilcoxon's signed-rank test, 1687–1688
- pharmacoeconomics, 1693–1704
 - design considerations, 1695
 - costing methods, 1695–1696
 - discounting, 1697
 - outcome measures, 1696–1697
 - study perspective, 1695
 - importance of economic evaluation, 1693–1694
 - interpreting the results, 1702
 - Bayesian statistical inference, 1703
 - methods of analysis, 1698
 - analysis of cost data, 1699
 - cost-effectiveness ratios, 1699–1701
 - net benefits, 1701
 - Scandinavian Simvastatin Survival Study, 1701–1702
 - research methods, 1697
 - clinical-economic trials, 1697–1698
 - modeling studies, 1697
 - types of analyses, 1694–1695
- Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium (PROTECT)
 - pharmacogenetics, 1057
 - and drug development, 240–241
- pharmacogenomics, 228
- pharmacokinetic (PK) bioequivalence study, 217
 - highly variable drugs, 220
 - narrow therapeutic index (NTI) drugs, 220
 - mean comparison, 220–221
 - within-subject variability comparison, 220–221
 - PK measures, 218
 - statistical approaches for BE assessment, 218
 - average BE, 218–219
 - individual BE, 219–220
 - population BE, 219
 - study design, 217–218
- pharmacokinetic information, 2209
- pharmacokinetic/pharmacodynamic (PD) effect, 309, 1748–1749
- pharmacokinetic property, 886
- pharmacokinetics, 1721, 2211
 - clinical pharmacology, 470
 - model, 1750–1751
- pharmacokinetic studies, 1121
- The Pharmacological Basis of Therapeutics*, 469
- pharmacopeial harmonization, 709
- Pharmacopoeial Tests and Acceptance Criteria, 259
- pharmacovigilance functions, 1344
- pharmacovigilance (PV) group, 813
- PharmGKB project, 228
- phase I cancer clinical trials, 390–392, 1715–1719
 - continual reassessment method (CRM), 391–392, 1717–1718
 - new designs, 391, 1719
 - modified up-and-down designs, 391
 - pharmacologically guided designs, 391
 - simulation techniques, 539
 - traditional design, 390, 1715–1717
 - dose escalation rule, 390
 - dose levels selection, 390
 - issues with traditional design, 390
- phase IIb dose-ranging study, 1722
- phase II cancer clinical trials, 392–394, 2296–2304
 - Bayesian designs for, 154–162
 - designs, 392–394
 - Bayesian single-stage approach, 393
 - classical single-stage approach, 393
 - sequential and multistage designs, 393–394
 - with multiple experimental arms, 395
 - optimal two-stage designs, 2297–2298
 - p*-value calculation, 2301–2304
 - response rate, inference on, 2298
 - confidence interval, 2300–2301
 - estimation of RR, 2298–2300
 - of vaccines with behavioral component, 240
- phase II clinical development, 1721
 - design phase IIa PoC trial, 1722
 - design phase IIb dose-ranging trial, 1724
 - impact of PoC decisions, 1723
 - importance of, 1725
 - risks associated with PoC study, 1723
- phase III cancer clinical trials, 394–395
 - adaptive stratified designs, 395
 - design, 1575
 - with multiple experimental arms, 395
 - stratified designs, 394
 - unstratified designs, 394
- “Phase I/II” design, 760
- phenotype and genotype, 228
- phenotypic variance of a trait, 1058
- phenylpropanolamine (PPA), 1763
- phenytoin, 995
- Phi coefficient, 1240
- Physicians Health Study, 576
- physiological measurement, 822
- piecewise exponential distribution, 2181
- piggyback studies, 1697
- Pitman-Morgan two one-sided tests procedure, 971
- placebo
 - concurrent control, 623
 - controls, titration with, 2228
 - defined, 1735
- placebo effect, 1735–1738
 - attempts at predicting placebo response in depression, 1738
 - factors influencing, 1736
 - regulatory requirements, 1736
 - statistical issues, 1737
 - design considerations, 1737–1738
 - factors associated with the placebo response, 1737
- Plackett-Burman design, 1941–1942, 2045
- Plackett-Dale distribution, 1419
- play-the-winner (PW) rule, 1728, 1950, 1953
- plug-in method, 169
- plume angle (ANG), 1142
- plume geometry, *in vitro* bioequivalence testing, 1142
- pneumococcal vaccines, 240
- Pneumocystis carinii*, 1369
- Pocock boundaries, 1198
- Pocock group sequential method, 60f, 61, 62
- point estimation, 100–101
 - analysis of variance estimation, 100–101
 - maximum likelihood and restricted maximum likelihood estimation, 101
- Poisson assumption, 1496
- Poisson distribution, 69–70, 71, 72, 316, 318, 995, 996, 1496, 2031
- Poisson model, 316, 508f, 992, 993, 996, 2091
- Poisson-normal model with correlated random effects, 1418
- Poisson rate ratio, 816
- poly-3 and poly-6 tests, 405
- poly-*k* tests, 405–409
- polymerase chain reaction (PCR), 131, 823, 1014
- polymorphism, 255
- poolability testing, 180, 181, 182
 - testing equality of intercepts with common slope, 182
 - testing equality of slopes with individual intercepts, 182
- pooling control groups, 2238
- pooling tests, and ANCOVA approach, 126–128
- poor compliance
 - on clinical status of patients, 1629–1630
 - magnitude and pattern of, 1628
 - on sample size requirements in clinical trials, 1629
- population average (PA) models, 651, 1004
 - individual-level analysis with cluster trials, 651–652
- population-based correction method, 2207
- population bioequivalence (PBE), 289, 306, 345, 922, 929, 936–937, 973–974, 1160, 1249, 1255f, 1432, 1740–1745

- masking effect, 290
- regulatory review, 1741
 - bioequivalence criterion, 1741–1742
 - concept, 1741
 - sample size calculation, 1742
 - statistical model, 1741
 - study design, 1742
- statistical analysis, 1742
 - alternative modified large sample method, 1744
 - FDA 2001 method, 1743–1744
 - general background, 1742–1743
 - new method that considers correlation among variance components, 1744
- population difference ratio (PDR), 1161, 1741, 219
- population equivalence, 970
- population model, 1875
- population PK/PD analysis, 1747–1755
 - definitions, 1747–1748
 - drug action, 1752–1753
 - pharmacodynamic model, 1751
 - structural model, choice of, 1751–1752
 - pharmacokinetic model, 1750–1751
 - population pharmacokinetics, 1747–1748
 - simulation of admissible dosage regimens, 1754–1755
 - simulation of PK toxicity index, 1753
 - software for, 1750
- population selection, phase I clinical trial, 539
- positive and negative syndrome scale (PANSS), 1044, 1128, 1625
- positive false discovery rate (pFDR), 1029
- positive likelihood ratio (PLR), 822, 824
- positive percent agreement (PPA), 824
- positive predictive values (PPV), 822, 824, 2184
- positron emission tomography (PET), 85
- postanalytical factors, 823
- posterior distribution, 176, 191, 294
- posterior odds ratio, 192
- postmarketing clinical drug development, 912–913
- post-marketing data presentation, 1344–1345
- postmarketing safety reporting, 53
- postmarketing surveillance, 1769–1775
 - data collection, 1769–1770
 - fluoxetine, surveillance of, 1774
 - Helicobacter pylori* treatments, surveillance of, 1773
 - cost-effectiveness analysis, 1774
 - drug interactions, 1773
 - excluded populations, 1773–1774
 - long-term follow-up, 1773
- postmarketing questions, 1770
 - cost-effectiveness analysis, 1772–1773
 - drug interactions, 1770–1772
 - excluded populations, 1772
 - trials, 886
- potency, 1811–1812
 - assays, 130
 - estimation of, 213
- powder blends, 1810
 - average potency, 1811
 - blend uniformity, 1810
 - Bergum approach, 1810
 - Food and Drug Administration approach, 1810
 - holistic approach, 1810
 - standard deviation prediction interval, 1810
 - tolerance interval approach, 1810
- sampling, 1803–1804
- power, 1777–1778
 - analysis, clinical pharmacology, 471
 - inadequately powered study, 1778
 - model, dose proportionality, 877–878
 - statistical power, correlates of, 1777
 - of the test, 1096
 - and variability, 1732
- power and sample size, 1777, 2033, 297–298
 - determination, individual bioequivalence, 1165
 - for ordered categorical data, 1246–1247
- power approach, estimated, 511
 - equal variances, two samples with, 511–512
 - parallel-group designs, 515–516
 - unequal variances, two samples with, 512–515
- practice, remarks on, 837
- preanalytical factors, 823
- precision, 375, 1650
- precision medicine, 1648, 1649, 1779–1781
 - future perspectives, 1780–1781
 - vs. personalized medicine, 1780–1781
- “preclinical” development, 909
- prediction interval, process validation, 1805
- prediction models, 1698
- prediction trees, 1782–1788
 - basic notations and concepts, 1783
 - for class variable and for continuous variable, 1783
 - algorithms for constructing CART, 1785
 - pruning in CART, 1784–1785
 - recursive partitioning in CART, 1783–1784
 - forests and Bayesian trees, 1788
 - generalized prediction trees, 1785
 - trees for more general outcomes, 1786–1787
 - tree-structured classification and regression, 1786
- predictive clustering, 1788
- trees with soft nodes, 1787
- predictive biomarker, 964
- predictive clustering, 1788
- predictive distribution, 191
- predictive enrichment, 964
 - selection maneuvers, 964
- predictive models, 1653
- predictive probabilities, 1199
- predictive tests, 1649
- predictor variables, 1653
- preliminary sample size, 1794
- premarket approval (PMA), 198
- premarketing clinical drug development, 911–913
 - new drug application, 912
 - phase 1 clinical trials, 911–912
 - phase 2/3 clinical trials, 912
- Prentice–William–Peterson (PWP) models, 396
- preparation contrast, 214
- Prescott test, 210–211, 212
- prescribability, 973
 - population BE (PBE) for, 936–937
- prespecified early time point, analysis at, 1585–1588
- prevalence method, 400
- prevention, drugs developed for, 889
- preventive vaccines, 239
- PREVNAR®, 240
- primary analysis, 437
- primary efficacy analysis, 239
- primary regulatory guidance, for dose finding, 1326
- principal component analysis (PCA), 1389, 1796
 - based on correlation matrix, 1799
 - computation of, 1797
 - data visualization, 1798
 - exploratory factor analysis, 987
 - interpretation of, 1799–1800
 - regression, 1798–1799
 - related inference on, 1797
 - sample, 1797
 - selecting principal components for analysis, 1798
- principal components, 189–190, 1796–1797
 - computational methods, 190
 - data analysis and study design, applications to, 190
 - cross-tabulated counts from small samples, 188–190
 - prior effective sample size, 190
 - two-agent dose–response model, 188–189
- principal investigator (PI), 598
- prior distribution, 174, 176
 - choice of, 193–194
 - of double threshold design, 158
 - of single threshold design, 156
- prior odds, 192
- ProACT-URL, 475
- probabilities of early termination (PET), 159
- probability based criteria, for similarity, 306
- probability density function, 2180
- probability distribution, 523
- probability model, 760
 - prior specification, 761
 - regression models for efficacy and toxicity, 760–761
- probability of success (POS) based approaches, 1595
 - decision rule based on CPOS, 1596
 - decision rule based on PPOS, 1596
 - probability of success, 1595–1596
 - randomized trial with normal end point, 1598–1599
 - randomized trial with time-to-event end point, 1596–1598
 - single arm study with binary end point, 1599–1600
- probit model, 1152, 1416, 1417
 - effective doses, 955, 957
 - with shared random effect, 1417
 - for uncorrelated data, 741–742
- process capability, 2140
- process flowchart, 2136–2137

- process validation, 1102, 1802–1815
 acceptance criteria, comparison of, 1808
 finished product, 1811–1812
 powder blends, 1810–1811
 examples, 1812–1815
 future directions, 1812
 history and overview, 1802
 QbD quality-by-design concept and PV, 1102
 regulated requirement, 1102
 sampling, 1803
 finished product sampling, 1804
 powder blend sampling, 1803–1804
 statistical techniques/approaches, 1805
 Bergum approach, 1806–1807
 confidence interval, 1805
 prediction interval, 1805
 simulation, 1805
 tolerance interval, 1805
 variance components, 1805
 technology transfer (TT) of pharmaceutical process, 1102
 types of validation, 1803
 PROC GENMOD, 71, 85
 PROC GLM, 436
 Procheck program, 229
 PROC MIXED, in SAS, 196, 1925, 1928, 1929
 produgs, 256
 producer's risk, 1, 2, 1097
 product development process, 1721
 Product License Application (PLA), 970
 Product-Limit estimator, 1231
 profile analysis, 1817–1822
 discussion, 1822
in vitro bioequivalence testing, 1145–1146
 model-based testing methods, 1821–1822
 simulation study and an example, 1820–1821
 study designs and be criteria, 1817–1818
 testing procedures for profile analysis, 1818
 FDA's approach, 1819
 two-stage random sampling approach, 1819
 profile likelihood approach, 560
 Profile of Female Sexual Function® (PFSF), 1167, 1175f
 Profile of Mood States, 1121
 prognostic biomarkers, 1649
 prognostic effects for acute myocardial infarction data, 2059
 prognostic factors, designs with, 1623–1624
 programmed cell death (PCD), 2123
 programmed death-ligand 1 (PD-L1), 825
 programming etiquette on a shared project, 1089–1094
 progression-free survival (PFS), 63, 155, 155, 1574
 PROMIS (Patient-Reported Outcomes Measurement Information System), 1168
 proof of concept (PoC) study, 1722
 “proof of mechanism” (PoM), 1722
 propensity score, 1823, 1825
 adjusted analysis, 1794
 applications, 1826
 control group selection, 1827
 covariate balance diagnostics, 1826
 covariate selection and identification, 1827
 missing data in covariates, 1827
 objective study design, 1826
 outcome analysis, 1827
 in regulatory settings, 1823–1824
 sample size estimation and power consideration, 1827
 estimation, 1794, 1826
 matching, 1310–1311, 1825, 1826, 1827
 medical devices, 1352–1353
 methodology, 1793, 1825
 modeling, 1825
 stratification technique, 1794, 1825, 1826, 1827
 proportional error case, measuring agreement, 1338
 Proportional Hazard Model (PHM), 1834–1840, 2063
 estimation, 1836
 extensions, 1839
 frailty models, 1839–1840
 time-dependent covariates, 1839
 model, 1835–1836
 model adequacy, assessment of, 1838
 Martingale residuals, 1839
 Schoenfeld residuals, 1838
 regression coefficients, estimation of
 for distinct event time data, 1836–1837
 in the presence of ties, 1837
 survival function, estimation of, 1837–1838
 proportional intensity assumption, 1499
 proportional interactions model (PIM), 203
 proportional odds mixed models, for clustered categorical data, 89
 proportional odds (PO) model, 1564–1565
 proportion of similar responses (PSR), 2168
 proportion of treatment effect (PTE), 1830–1833
 statistical concept and methods, 1830
 information separation, 1832
 single model, PTE estimation based on, 1831–1832
 two separate models, PTE estimation based on, 1830–1831
 Prosa II program, 229
 PROSITE, 225
 prospective alpha allocation scheme (PAAS) and Šidák's test, 730
 prostate-screening antigen (PSA) tests, 1348
 Protein Data Bank, 228
 proteins, 228
 structure, prediction methods, 229
 proteomics, 227
 protocol amendments, impact and sensitivity of, 584–585
 protocol development, 1841–1844
 analysis subsets, 1842–1843
 content of the statistical section of a protocol for a clinical trial, 1842
 primary analyses, 1842
 primary statistical objective, 1842
 primary variable(s) and analyses, 1842
 primary variable definition, 1842
 decision rules, 1842
 design and number of subjects, 1843
 error control, 1842
 interim analyses, 1843
 nature and purpose of the protocol for a clinical trial, 1841–1842
 problem data, 1843
 randomization and blinding, 1843
 safety data, 1843–1844
 secondary variable(s) and analyses, 1843
 protocol-specific databases, clinical data management guidelines, 446
 protocol-specific draft CRF, 446
 Prove program, 229
 proxy information, use of, 1921–1922
 pseudo-likelihood, 905
 pseudomaximum likelihood estimator, 560
 PSI-BLAST profile alignments, 229
 psoriasis, 1666
 psychometric instrument, 1167
 publication bias, 177
 Public Health Service (PHS) Act, 2219
 pulmonary artery pressure, 822
 pulmonary embolism, 825
 pure blends, 1947
p-value, 651, 725–726, 1054, 1125, 1126, 1383, 1469, 1602, 1845–1853
 -based techniques, 1468–1469
 calculation, 2301–2304
 clinical pharmacology, 471
 combining, 2024
 criticisms, 1848
 Jeffreys–Good–Lindley paradox, 1849
 distribution of, 1846–1848
 history, 1845–1846
 issues affecting the use of *P*-values in practice, 1850
 adaptive designs, 1853
 approaches to combining *P*-values, 1851
 discrete data, problems with, 1852
 multiplicity, 1852
 one-sided/two-sided *P*-values, 1852
 two-trials rule, 1850–1851
 likelihood principle, violation of, 1849–1850
 penalizing, 2024
 probability density of, 1847f
 repeatability of, 1850
- Q**
 QT analysis, 1855–1863
 clinical trials, analysis of ECG data from, 1856
 fixed QT correction analysis, 1857–1858
 model specification, 1857
 off-drug QT correction analysis, 1858–1859
 off-on-drug QT correction analysis, 1859
 QT correction method overview, 1856–1857
 QT-Interdisciplinary Review Team (QT-IRT), 2206
 QTc circadian rhythm, 2213
 QTc effect, 2206, 2209
 QTc interval, 2206
 Q-twist method, 600f–601
 qualitative analysis, 111

- qualitative assays, evaluation of, 2319
 - matrix effects, 2322–2323
 - receiver operating characteristic (ROC) curve
 - applications of, 2320–2321
 - sensitivity and specificity, 2321–2322
 - qualitative content validation, 1169–1170
 - qualitative interaction, testing for, 2192–2197
 - likelihood ratio test, 2192–2194
 - range test, 2194–2195
 - simultaneous inference, test procedure based on, 2195–2197
 - qualitative trait, 1058
 - quality-adjusted life-years (QALYs), 754, 756–757, 1696, 1640, 1708
 - quality assurance and quality control, 1076
 - quality assurance units (QAUs), 1076
 - quality attributes (QAs), 2219–2220
 - Quality-by-design (QbD) concept and PV, 1102
 - quality control (QC)
 - consistency in, 307
 - method, 2259
 - quality-of-life (QOL) instruments
 - comparison of life factors, 599f
 - developing, 597–599
 - existing, 597
 - prognostic factor analysis using QOL measures, 1439–1444
 - QOL-like instrument, 2248
 - validation of, 2248
 - Q-twist method, 600–601f
 - survival differences between control and treatment, 600f
 - quality range approach for Tier 2 CQAs, 2221
 - quantal response
 - assays, 136f
 - and tolerance distribution, 214–215
 - Quantile-Quantile plot, 2187f
 - quantile regression, 1644
 - quantitative analysis, 111
 - quantitative assays, evaluation of, 2314
 - accuracy, 2314
 - method comparison, 2314–2316
 - recovery study, 2314
 - key difference and similarity between qualitative assays and, 2313–2314
 - limits of detection (LOQ) and quantitation, 2319
 - linearity and range, 2318
 - study designs and statistical analysis, 2318
 - matrix effects, 2317–2318
 - precision, 2316
 - study designs and statistical inferences, 2316–2317
 - quantitative devices, 825
 - quantitative instrument
 - development, 1170
 - validation of, 2257–2258
 - quantitative interaction, 1436
 - quantitative trait loci (QTL), mapping, 1057, 1058, 1060, 2122, 2123
 - combined linkage and LD analysis, 1061
 - heritability, 1058
 - IBD estimation, 1060
 - LD and fine mapping, 1060–1061
 - linkage analysis, 1059–1060
 - LOD score, 1060
 - software, 1061
 - quantitative trait nucleotides (QTNs), 2123, 2124–2125
 - quantities of interest, 423–424
 - quasi-complete separation of responders from nonresponders, 1316–1317
 - quasi-likelihood models, 993–997
 - Question Understanding Aid (QUAID), 1168
 - Quigley-Hein index, 94
 - quinidine, 574
 - Q value, 1152
 - Q wave, 2206, 2207
- R**
- racial differences, pharmacodynamic issues, 1670–1671
 - radioimmunoassay (RIA), 131
 - random censoring, 1235
 - random coefficient-selection model, 905
 - random-effects meta-analysis, 605–606, 610
 - random-effects meta-regression, 610
 - random-effects model, 99, 180, 175, 374, 1004–1005, 1417
 - center weighting in, 440–441
 - correlated random effects, 1418
 - random effects plus correlated errors, 1418–1419
 - shared random effects, 1417–1418
 - therapeutic trials, meta-analysis of, 1367
 - randomization, 798, 1097
 - based inferences, 943–944
 - clinical trials, 572–573
 - of clusters, 648
 - defined, 1216
 - and drug management, traditional approach to, 1216–1217
 - integrity of randomization/blinding, 582–583
 - methods, 1876
 - adaptive randomization, 1877
 - complete randomization, 1876
 - permuted-block randomization, 1876–1877
 - models, 1875, 1876
 - invoked population model, 1875
 - population model, 1875
 - problems associated with, 1217
 - rules for clinical trial simulations for later development phases, 545–546
 - schedule, 1219
 - randomized block design, 1623
 - randomized clinical trial (RCT), 174, 572, 1791, 1956, 2265
 - example, 2179–2180
 - randomized discontinuation design (RDD), 1489
 - randomized, double-blind, placebo-controlled, fixed-dose, parallel dose-response design, 886
 - randomized play-the-winner (RPW)
 - allocation, 1728
 - rule, 545, 1950
 - randomized start/withdrawal designs, 839–841
 - randomized trials, 773
 - randomized two-period crossover designs,
 - clinical pharmacology, 470
 - randomized withdrawal trial design, 528–529
 - random-play-the-winner (RPW), 546
 - clinical trial simulations for later development phases, 546
 - design with RPW randomization
 - clinical trial simulations for later development phases, 548
 - random target values, 1336
 - range, of analytical method, 376
 - range test, for qualitative interaction, 2194–2195
 - rank-based robust analysis for crossover design, 784–796
 - AB/BA crossover designs, 785
 - advantages of, 785–786
 - analysis of covariance in, 786
 - uses of crossover designs, 786
 - asymptotic comparisons, 791
 - when baseline values are absent, 793
 - bivariate double exponential distribution, 792
 - bivariate normal distribution, 791–792
 - bivariate *t* distribution, 792
 - in presence of baseline values, 792–793
 - intra-subject rank-based method, 787
 - estimation of τ , 789
 - rank tests, 788–789
 - R* estimation, 787–788
 - model and assumptions, 785
 - Patel's data without the outlier, 794, 795
 - inferences in the presence of baseline values, 794–795
 - inferences without the baseline values, 794–795
 - using period differences, 789
 - estimation of τ_1 , 791
 - rank-based shelf-life determination, 1889
 - shelf-life for a single batch, 1889
 - shelf-life for multiple batches, 1889
 - rank regression in stability analysis, 1888–1891
 - alternative approaches, 1890–1891
 - numerical studies, 1890–1891
 - preliminary test for batch similarity, 1890
 - rank-based shelf-life determination, 1889
 - shelf-life for a single batch, 1889
 - shelf-life for multiple batches, 1889
 - semiparametric regression models, 1888
 - rare events assumption, 422
 - Rasch model, 1211, 1214
 - rate, 432
 - rate difference, 432
 - ratio non-inferiority hypotheses, 9–10
 - ratio trend test, 406
 - real laboratory findings vs. laboratory error or artifact, 1263
 - receiver operating characteristic (ROC)
 - analysis, 1175, 1354, 1608–1611
 - application, 1610
 - plot, 1654
 - ROC surface and hypersurface, 1609–1610
 - receiver operating characteristic (ROC) curve, 236, 824, 1962–1968, 2101, 2321f
 - applications of, 2320–2321
 - area under, 1963–1964

- receiver operating characteristic (ROC) curve, (*Cont.*)
 - biopharmaceutical applications of curve, 1964
 - characterizing accuracy with summary measures, 1964–1965
 - choosing a cut-off point for positive result for confirmatory trials, 1964
 - combining sensitivities and specificities from several clinical trials, 1966
 - combining separate measures of accuracy and diagnostic confidence, 1965
 - comparing treatments/methods using receiver operating characteristic summary measures, 1965–1966
 - construction of, 1963
 - definition and construction, 1608
 - example, 1967
 - fitting a binormal, 1963
 - logistic regression and, 1964
 - partial area under, 1964
 - regulatory requirements, 1962
 - sample size calculation, 1966
 - scientific issues, 1967–1968
 - single points on, 1964
- reciprocating cylinder, 263
- recombinant DNA method, 932
- recombination, defined, 1058
- RECPAM approach, 1785, 1786, 1786
- recruitment procedures, clinical trials, 574–575
- rectification inspection programs, 3
- red blood cells (RBCs), 825, 1263
- Reed and Muench method, 137
- reference-based pricing, 1693
- reference biotherapeutic product (RBP), 283
- reference limits, estimating Lilly Reference Ranges, 1277
- reference prior, 176, 193
- reference sample group, selection of Lilly Reference Ranges, 1277
- reference-scaled average BE (RSABE), 220
- region-level adaptive design, 1582
- region of optimum, 1939
- registry database, 1791
- regression analyses, 912
- regression coefficients, 1940
 - estimation
 - for distinct event time data, 1836–1837
 - in the presence of ties, 1837
- regression contrast, 214
- regression-discontinuity (RD) design, 798–799
 - validity of, 799
- regression model covariates, 772
- regressors, 110
- regulatory decisions, clinical trial simulation in, 537
- regulatory guidance, 193, 437–438
 - dose response analysis in clinical trials, 879
- regulatory process, of biologics, 234
 - generic products, 234
 - international harmonization, 234
 - product development and licensure, 234
- regulatory requirements, 707, 1143
 - for biosimilars, 305
- diagnostic imaging, 829
- drug development, 908–909
- intention-to-treat analyses (ITT), 1183–1184
 - noncomparative analysis, 1143
 - nonprofile analysis, 1143–1144
 - profile analysis, 1144–1145
- relational database management system (RDBMS), 1070
- relative asymptotic efficiency, 999
- relative potency (RP), 134*f*, 135
- relative sample sizes needed in cutoff designs, 802
- relative standard deviation (RSD), 133, 709
- release targets, 1903–1906
 - Allen *et al.* method, 1903
 - product requiring reconstitution prior to use, 1903–1904
 - product with degradation, 1903
 - product with no degradation, 1903
 - conditional and unconditional methods, 1904
 - conditional rule, release limits based on, 1904–1905
 - decision rules, 1904
 - expected loss function approach, 1906
 - unconditional rule, release limits based on, 1905–1906
- reliability, 1176, 1908–1916
 - agreement, measuring, 1915–1916
 - analysis, 1908
 - concordance correlation, 1910–1911
 - confidence intervals, 1914
 - different aspects of, 1909–1910
 - estimation, 1913
 - improving, 1914–1915
 - intraclass correlation, 1911–1912
 - intrarater, interrater, and test–retest reliability, 1912
 - measurement error, consequences of, 1908–1909
- renal/liver function, 1671
- reparameterization, 1270
- repeated confidence intervals, 1199
- repeated measurement designs (RMD),
 - missing values in, 1918–1922
 - general approaches to missing data, 1919
 - incomplete repeated measures data analyses, 1919–1921
 - missing data mechanism, 1918–1919
 - use of proxy information, 1921–1922
- repeated measures analysis, 1588
 - analysis with missing values, 1924–1929
 - available data analysis, 1925
 - bronchial asthma data analysis, 1928–1929
 - design, 1624
 - imputations, 1926–1927
 - case analysis, 1927–1928
 - multiple, 1926–1927
 - single, 1926
 - missing data mechanisms and implications, 1924–1925
- replicated $2 \times 2m$ crossover design, 497
 - test for equality, 497
 - test for noninferiority/superiority, 497
 - test for similarity, 497–498
- replicated crossover design, 488, 493
 - test for equality, 488–489
 - test for noninferiority/superiority, 489
 - test for similarity, 489
- replication, 1622
- reproducibility of a study drug, 500
- reproducibility probability, 510–521, 523
 - application and extensions, 521
 - binary data, 521
 - reproducibility probabilities for K trials, 521
 - substantial evidence with a single trial, 521
- Bayesian approach, 517
 - parallel-group designs, 520–521
 - two samples with equal variances, 517–520
 - two samples with unequal variances, 519–520
- confidence bound approach, 516–517
- estimated power approach, 511
 - parallel-group designs, 515–516
 - two samples with equal variances, 511–512
 - two samples with unequal variances, 512–515
- ratio, 1516
 - for superiority test, 1516
- reproductive/developmental studies, 910, 1931–1937
 - developmental studies, 1931
 - regulatory requirements, 1931
 - three-segment design, 1932
 - scientific issues, 1935–1936
 - statistical design/method/analysis, 1932
 - analysis, 1935–1936
 - experimental design, 1932–1933
 - methods, 1934–1935
- reproductive toxicity studies, 2235
- requirements analysis, global database and system, 1069–1070
- research hypothesis, 1124
- Research in Computational Molecular Biology (RECOMB), 224
- residual distribution function, 1330
- residual draw (RD) method, 1926
- residual effect, 206
- residual maximum-likelihood (REML), 1004, 1411
- residual plots, 132*f*
- respirable volume, 255
- respiratory system, 1666
 - asthma, 1666
- response-adaptive allocation, 1727–1728
- response-adaptive designs (RADs), 612–621, 1727–1733, 1950–1953
 - adaptive design, 1952
 - and analyzing data from adaptive designs, 1953
 - sample size of, 1952
 - variability and power of using, 1952
 - continuous outcomes, 615
 - two-treatment three-period designs, 615–617
 - two-treatment two-period designs, 615
 - design allocation strategy, 612–613
 - design model, 613
 - continuous outcomes, 613
 - dichotomous outcomes, 613–614

- dichotomous outcomes, 618
 - two-treatment three-period designs, 619
 - two-treatment two-period designs, 618–619
- different issues, 1731
 - Bayesian designs, 1732
 - covariates, 1731
 - longitudinal responses, 1732
 - missing data, 1732
 - multivariate responses, 1732
 - power and variability, 1732
 - real-life applications, 1732
 - sample size calculation, 1732–1733
- estimated parameter-based allocation using
 - link functions, 1729
 - ethics-based target allocation functions, 1730
 - implementing the target allocation function, 1731
 - link function–based design for continuous responses, 1729–1730
 - target allocation functions combining ethics and precision measures, 1730–1731
- with multiple treatment groups
 - clinical trial simulations for later development phases, 549–550
- optimal allocations, 1951
- target-based, 1951
 - doubly adaptive biased coin design, 1951–1952
 - sequential maximum likelihood procedure, 1951
- treatment effective mapping, 1952
- urn models, 1950
 - drop-the-loser rule, 1950–1951
 - play-the-winner rule, 1950
 - randomized play-the-winner rule, 1950
- response-adaptive randomization, 546
 - clinical trial simulations for later development phases, 546
- response and allocation history-based allocations, 1728
 - binary responses, 1728
 - birth and death urns, 1729
 - play-the-winner (PW) allocation, 1728
 - randomized PW (RPW) allocation, 1728
 - categorical responses, 1729
 - continuous responses, 1729
- response factor plot, 132
- response measurements, 417
- response rate (RR)
 - inference on, 2298
 - confidence interval, 2300–2301
 - estimation of RR, 2298–2300
- response surface, 1939
 - statistical analysis of, 1945–1946
- response surface methodology (RSM), 1939–1948, 1940f
 - 2^k factorial designs, 1940
 - center runs in, 1942
 - computer-generated designs, 1944–1945
 - experimental designs for fitting a second-order response surface, 1943–1944
 - fractional factorial designs, 1940–1941
 - mixture experiments, 1946–1947
 - model building, 1939–1940
 - robust parameter design (RPD), 1948
 - screening designs (Plackett–Burman), 1941–1942
 - statistical analysis of a response surface, 1945–1946
 - steepest ascent/descent, 1942–1943f
 - response variables, 418, 1685
 - R estimation, 787–788
 - restricted maximum likelihood estimation (RMLE), 1479, 1742, 2091
 - restricted maximum likelihood (REML) method, 100, 101, 375, 605, 795, 905, 1410
 - resubstitution bias, 827
 - retinal thickness, 822
 - reverse mutation assay, *see* Ames test
 - revertant counts, 68, 70, 71, 72–73
 - linear regression on logarithmically transformed, 71
 - non-linear regression on, 71
 - rheumatoid arthritis, 1666
 - ridge analysis, 1946
 - right-censored event, 1980
 - “right-censoring” process, 905
 - right truncation, 2182
 - rising ridge, 1947f
 - risk, 431
 - risk difference, 431
 - risk prediction, 1652, 1653
 - risk ratio analysis, 431, 1956–1960
 - definition and characteristics of, 1956
 - estimation and asymptotic properties, 1957
 - estimation under stratified Poisson sampling, 1959
 - relations between RR and other indices, 1956–1957
 - stratified binomial sampling, 1957–1958
 - estimation in large stratum size, 1958
 - estimation in sparse data, 1958
 - test of homogeneity of RR, 1958–1959
 - risk reduction, 1579
 - risk score, 826
 - risk sets, 424, 425
 - Robit-normal model, 1419
 - robust Bayesian interpretation, 429
 - robustness, good programming practice, 1089–1091
 - robustness of interval estimators
 - of ED₅₀, 955
 - asymmetric Aranda–Ordaz model, 956
 - cubic logistic model, 956
 - probit model, 955
 - of ED₉₀, 957
 - asymmetric Aranda–Ordaz model, 957
 - cubic logistic model, 957
 - probit model, 957
 - robust parameter design (RPD), 1948
 - Rom’s method, 1474
 - root-mean-square error of approximation (RMSEA), 1173
 - Rosetta program, 229
 - rotating basket method, 261
 - R software, 62, 180, 183
 - ruggedness, 375
 - Rutter and Gatsonis model, 2107
 - R waves, 2207

S

 - S + Seq Add-on package for S-plus software, 62
 - S + SeqTrial, 1199
 - Saccharomyces cerevisiae*, 1390
 - saddle point, 1945f, 1946f
 - Safety Assessment Committees (SAC), 812
 - safety data, 55–56
 - safety endpoints, dose response analysis in clinical trials, 882
 - safety, integrated summary of, 1181
 - safety monitoring procedures, 928
 - safety monitoring program in confirmatory trials, 812
 - conflicts of interest, 813
 - DMC creation, 813
 - liability and indemnification, 813
 - members of the safety monitoring team, 812–813
 - safety pharmacology studies, 2232
 - safety profile, 67
 - Salmonella typhimurium*, 67, 2233
 - tester strains, 68
 - salts, 256
 - sampled risk sets, 425
 - sample matrix, 130
 - sample size, 438–439
 - adjustment
 - clinical trial simulations for later development phases, 546
 - bounds, 2034
 - considerations, 1269–1270
 - determination of, 21, 1097, 1604, 1827, 2032, 2187
 - in Bayesian analysis, 190, 194–195
 - cluster trials, 653
 - of double threshold design, 158
 - early phases of clinical trials, simulation in, 540
 - for fixed sample test, 2326–2327
 - in vitro* bioequivalence testing, 1146
 - for modified large sample method, 499–500
 - planning, 1794
 - of single threshold design, 157
 - sample size calculation, 498, 1732–1733, 1980–1984, 2011, 2031
 - based on the binomial distribution, 1981
 - binary outcome variable case, 2012
 - continuous outcome variable case, 2012
 - for exponential model, 1981–1982
 - in group sequential trials, 1983–1984
 - odds ratio, 1557
 - for semiparametric test, 1982–1983
 - statistical software for, 1984
 - for two-stage seamless adaptive design, 2293–2294
 - sample size calculation based on nonparametric statistics, 1970–1979
 - Kendall’s test of independence, 1977
 - example, 1978–1979
 - fixed alternative, 1977
 - local alternative, 1977
 - simulation study, 1977–1978
 - one-sample Wilcoxon’s test, 1972
 - example, 1973–1975
 - fixed alternative, 1973

- sample size calculation based on
 - nonparametric statistics, (*Cont.*)
 - local alternative, 1973
 - simulation study, 1973–1974
 - sign test, 1971
 - example, 1971–1972
 - fixed alternative, 1971
 - local alternative, 1971
 - simulation study, 1971–1972
 - two-sample Wilcoxon's test, 1975
 - example, 1975–1977
 - fixed alternative, 1975
 - local alternative, 1975
 - simulation study, 1975–1976
- sample size modification, 2021–2024
 - fixed sample size design, 2021
 - issues on, 2021
 - efficiency, 2022
 - inflation of type I error rate, 2022
 - method for projecting new sample size, 2022
 - operational issue, 2023
 - statistical test based on, 2023
 - combining normal scores through weighting, 2023
 - combining p values, 2024
 - conditional type I error function, 2023
 - penalizing p value, 2024
- sample size reestimation (SSR) design, 30, 1573
 - with robust power, 2026–2029
 - adaptive performance, measure of, 2028
 - flexible sample size design, objective of, 2027
 - group sequential and sample size re-estimation designs, 2027–2028
 - literature review, 2026
 - performance comparison and design optimization, 2028–2029
- sample space, 1845
- sampling, acceptance, *see* acceptance sampling
- Sarandasa and Krishnamoorthy's (SK) method, 855–856
- SBA (summary basis of approval), clinical trial simulation, 537
- Sbaves software, 196
- scaled agreement indices, 164
- scaled average bioequivalence (SABE), 289, 2038–2042
 - for highly variable (HV) drugs, 2039–2040
 - for narrow therapeutic index (NTI) drugs, 2042
 - regulatory implementations of, 2040
 - European Medicines Agency, 2040–2041
 - FDA and EMA, comparison of the approaches of, 2041–2042
 - Food and Drug Administration, 2040
 - unscaled ABE, 2038–2039
- scaled criterion for drug interchangeability (SCDI), 922
- scale-up and postapproval changes (SUPAC) guidelines, 268, 307
 - common elements, 268–269
 - changes in batch size (scale-up/scale-down), 269
 - changes in components and composition, 269
 - changes in site of manufacture, 269
 - equipment changes, 269
 - process changes, 269
 - test documentation, 269
- comparison of profiles, 270
- SUPAC-IR, 269–270
- SUPAC-MR, 270
- SUPAC-SS, 270
- Scandinavian Simvastatin Survival Study, 1701–1702
- Scheffe's procedure, 731, 1689
- schizophrenia, 2331
- Schoenfeld residuals, 1838
- Schuurmann's two one-sided tests (TOSTs), 293–294, 297, 217
- scientific difference, 1097–1098
- scientific issues, intention-to-treat analyses (ITT), 1185–1186
- score-type statistics, generalized estimating equation, 1049–1050
- screening design, 2045–2046
- SDS gel electrophoresis methods, 227
- seamless adaptive trial design, 2047–2057
 - dose selection, criteria for, 2048
 - conditional power, 2048
 - confidence interval, precision analysis based on, 2048
 - probability of being the best dose, 2049
 - success, predictive probability of, 2048–2049
- practical implementation and example, 2049
 - co-primary endpoints, 2049–2051
 - example, 2051
 - single primary endpoint, 2049
- simulation studies, 2051
 - co-primary endpoints, 2054–2057
 - single primary endpoint, 2052–2054
- secondary statistical analysis, 1122
- secondary variable(s) and analyses, 1843
- second-order response surface, experimental designs for fitting, 1943–1944
- second-phase shelf-life, 2112–2113
- security plan, good programming practice, 1093
- segmental function error spending function, 62
- selection tests, 1649
- selectivity, 376
- self-administered instrument, 1128
- self carryover effect, 613
- self-consistent estimator, 1232
- self-organizing maps (SOM), 1388
- semiparametric accelerated failure model, 558
- semiparametric multiplicative intensity models, 1496, 1497–1498
- semiparametric priors distribution, 176
- semiparametric proportional odds model, 559
- semiparametric regression model, 557
- semiparametric test, sample size calculation for, 1982–1983
- semi-parametric time-varying regression models, 2058–2062
- applications, 2059
- prognostic effects for acute myocardial infarction data, 2059
- treatment effect for CSL1 study, 2060–2062
- flexible modeling of survival data, 2058–2059
- software, 2062
- sensitivity, 2212
 - and specificity, 836–837
- sensitivity analysis, 79, 2118
 - medical devices, 1353
 - one analysis approach
 - testing poolability, 2119
 - without testing poolability, 2118
- sensitivity index, evaluation of, 1514–1515
- sequence alignment, 225
- sequence analysis, 224–225
- sequence motifs, 225–226
- sequential and multistage designs
 - phase II cancer clinical trials, 393–394
- sequential maximum likelihood procedure, 1731, 1951
- sequential monitoring trials with multiple endpoints, 1110
- Sequential Oligogenic Linkage Analysis Routines (SOLAR), 1061
- sequential parallel comparison design (SPCD), 528
- sequential stopping rules, 2065–2066
- sequential testing, 967
- sequential test statistics, 2325–2326
- serious adverse event (SAE)
 - data flow, 813
 - reconciliation process, 447
- sertraline, 1736
- Sexual Arousal, 1171
- Sexual Desire, 1169, 1174f
- sexually transmitted diseases (STDs), 943
- Sexual Pleasure, 1171
- sham controls, medical devices, 1350
- shape-invariant modeling, 1547
- shared project, programming etiquette on, 1089–1094
- shared random effects, 1417–1418
- shelf-life, 979
 - defined, 180
 - determination, based on ANCOVA modeling, 125–128
 - of drug product, 2111
 - estimation, 183
 - with discrete responses, expiration dating period, 982
 - for multiple batches, 1889
 - for single batch, 1889
- shifted null hypotheses, testing, 2142
 - proof of equivalence, 2144–2145
 - proof of noninferiority, 2144
 - proof of superiority, 2142–2144
 - wider applications, 2145
- shift tables, 1260
- short- or medium-term rodent test, 398
- Short Tandem Repeat, 1057
- Šidák's test
 - extensions of, 730
 - prospective alpha allocation scheme (PAAS) and, 730
- sieve maximum likelihood approach, 560

- sigmoid Emax model, 1327
- sigmoid models
 - dose response analysis in clinical trials, 880
- signaling reporting profile difference, 1762
- signal transduction networks, 227
- significance analysis of microarrays (SAM), 1030
- significance level, 1096
- significance tests, 59
- sign test, 1971
 - example, 1971–1972
 - fixed alternative, 1971
 - local alternative, 1971
 - simulation study, 1971–1972
- similar biotherapeutic products (SBPs), 283–284
- similarity, between drug products, 305–307
 - absolute change vs. relative change, 305–306
 - aggregated vs. disaggregated, 306
 - consistency in manufacturing process/quality control, 307
 - dissolution profile comparison, similarity factor for, 307
 - moment based criteria vs. probability based criteria, 306
- similarity limit and variability, relationship between, 122
- Simon's two-stage designs, 1486–1487
- simple change score method, 1617–1618
- simple linear regression
 - pharmacodynamics with covariates, 1678
- simple ordering, 75
- simple random effects model, 489
 - test for equality, 489
 - test for noninferiority/superiority, 489–490
 - test for similarity, 490
- Simpson's paradox, 704
- simulated data, 418
- simulating clinical trials, *see* clinical trial simulation (CTS)
- simulation, 12
 - approach, 342
 - early phases of clinical trials, simulation in, 541
 - finished product, 1811
 - methods, 1927
 - process validation, 1805
 - results, measuring agreement, 1338
- simulation scenarios and values of covariance, 2213
- simulation study, 94–95, 931
 - parameter specification for, 931
 - safety margins for, 931
 - statistical inference, 1130–1132
- simultaneous inference, test procedure based on, 2195–2197
- single active ingredient vs. multiple components, 2256
- single ascending dose (SAD), 2212
- single blind trials, 336, 573
- single contrast test, 1329
- single-dose crossover bioequivalence study, 264–265
- single-dose escalation
 - clinical pharmacology, 469
- single imputations, 1926
 - statistical model, 1129
- single nucleotide polymorphism (SNP), 228, 1057, 1448, 2123–2124, 2184
- single photon emission CT (SPECT) scan, 85
- single sampling acceptance plan, 2
- single-sided specification limits, 5
- single-stage sampling plan, clinical inspection, 466–467
- single threshold design (STD), 156–157
 - computational algorithm and sample sizes, 156–157
 - predictive version of, 161
 - prior distribution, 156
- site management organizations (SMOs), 719
- skepticism, 1503–1504
 - and Bayesian statistics, 1504–1505
- skewness, 754
- skin, 1666
 - acne, 1666
 - psoriasis, 1666
- skin preparations, ointments and, 257
- skin sensitization, test for, 2231–2232
- slang terms, 1168
- “sledgehammer” approach, 1722, 1724
- slope analysis, 842
- Smith-Waterman algorithm, 225
- smoothing splines, 1544–1545
- Snapinn's fifth test, 552
- soft gelatin capsules, 259
- software, interim analysis, 1199
- solubility, 254
- solutions, in dosage forms, 258
- somatropin, 305
- space–time interaction model, 2091
- sparsity-of-effects principle, 1941
- spatial restraints, 229
- spatio-temporal modeling, 2089–2097
 - for area-aggregated rates, 2089
 - spatial models with random effects, 2089–2090
 - discrete mixture models, 2092
 - estimation, 2091
 - empirical Bayes methods, 2091
 - full Bayesian methods, 2092
 - example, 2093–2094
 - models for point pattern processes, 2092
 - models for spatial Gaussian processes, 2092
- Spearman and Kärber method, 137
- Spearman's correlation coefficient, 1175
- Spearman's theory, 1008
- special patient population, drugs for, 889
- special population studies
 - clinical pharmacology, 470
- Special Search Categories (SSCs), 1345
- specifications, 2099–2100
- specificity, 376, 909, 2212
 - sensitivity and, 836–837
- spending function, for alpha, 59–66
- splitting type I error rate, 965
- Splus, 2094
- sponsor size, 812
- Spontaneous Reporting System (SRS), 1756
- spray pattern, *in vitro* bioequivalence testing, 1142
- stability, 256
 - testing, 125
- stability analysis based on rank regression, 1888
 - semiparametric regression models, 1888
- stability analysis for frozen drug products, 2111–2114
 - example, 2113
 - statistical model, 2111
 - second-phase shelf-life, 2112–2113
 - two-phase regression analysis, 2112
- stability matrix, 2116–2120
 - analysis, 2118
 - background, 2118
 - comparison of designs, 2119
 - designs, 2116
 - background, 2116
 - basic matrix 2/3 on time design, 2116–2117
 - matrix 1/3 on time design, 2117–2118
 - matrix 2/3 on time design with multiple packages, 2117
 - matrix 2/3 on time design with multiple packages and multiple strengths, 2117
 - matrix on batch $\times$ strength $\times$ package combinations, 2117–2118
 - uniform matrix design, 2118
 - factors acceptable to matrix, 2119–2120
 - history, 2116
 - separate analysis approach, 2118
 - one analysis approach, without testing poolability, 2118
 - one analysis: testing poolability, 2119
- stable unit treatment value assumption (SUTVA), 2169
- standard 2×2 crossover design, 496
 - test for equality, 496
 - test for noninferiority/superiority, 496
 - test for similarity, 496–497
- standard curve, 131
 - in calibration, 370–372
 - model selection, 370–371
 - statistical model, 370
 - weight selection, 372
- standard deviation, 523
- standard deviation prediction interval (SDPI), 1805
 - process validation, 1810
- standard group sequential methods, 59–60
- standardization, 1069
- standardized MedDRA queries (SMQs), 815
- standardized regression coefficients, 989
- STandardized RESidual Sum of Squares (STRESS), 1390
- Standardized Root Mean Square Residual (SRMR), 1173
- Standard MedDRA Queries (SMQs), 1345
- standard normal distribution, 319, 2329–2330
- standard operating procedures (SOPs), 463, 813, 1076, 1093, 2247
- statins, 1258
- stationary point, 1945
- stationary ridge, 1946
- statistical analysis
 - for adverse event data, 55–57
 - adverse events as secondary endpoint, 57
 - individual study report, 55–56
 - integrated summary of safety, 57
 - of cluster trials, 650

- statistical analysis, (*Cont.*)
 - for continuous data, 1366
 - fixed-effects model, 1366–1367
 - random-effects model, 1367
 - for discrete data, 1364
 - assessing homogeneity of treatment effects, 1365–1366
 - fixed-effects and random-effects models, 1364–1366
 - meta-regression analysis, 1366
 - patient population included in, 1670
 - all randomized patients, 1670
 - per-protocol patients, 1670
- statistical analysis plan (SAP), 1422, 1826
- Statistical Analysis System (SAS), 1366
 - code
 - Hills-Armitage test, 212f
 - Mainland-Gart test, 209f
 - McNemar test, 208f
 - SAS PROC ANOVA, 126
 - program, 104, 2155–2156
 - system, 1088
- statistical approaches used for assessing similarity, 2219–2220
- statistical considerations and approaches, 1585
 - prespecified early time point, analysis at, 1585–1588
 - repeated measures analysis, 1588
 - survival analysis, 1588–1589
 - using mixture models, 1590–1591
- statistical decision theory, 826, 827
- statistical designs and analysis methods
 - clinical pharmacology, 470–471
 - combination drug clinical trial, 552
- statistical difference, 1097–1098
- statistical equivalence test for Tier 1 CQAs, 2220–2221
- statistical genetics, 2122–2126
 - functional mapping, 2123
 - future direction, 2125–2126
 - genetic mapping, 2122
 - by incorporating mechanistic pathways, 2125–2126
 - from quantitative trait loci (QTLs) to quantitative trait nucleotides (QTNs), 2123
 - quantitative trait nucleotides, 2124–2125
 - single nucleotide polymorphisms, 2123–2124
- statistical inference, 1130
 - simulation study, 1130–1132
 - testing hypothesis, 1130
 - variance estimates, 1130
- statistical interaction, 706
- statistical issues, 815
 - goal of statistical analysis, 815
 - hypothesis testing, 816
 - intention-to-treat analyses (ITT), 1185
 - in postmarket, 1358
 - statistical description, 816
 - useful data displays, 815–816
- statistical methods, 1122
 - biosimilarity assessment, 292
 - and cancer trial endpoints, 395–396
 - classic methods for assessing biosimilarity, 292
 - Bayesian methods, 294–295
 - confidence interval approach, 293
 - Schuirman's two one-sided tests procedure, 293–294
 - Wilcoxon–Mann–Whitney two one-sided tests procedure, 295–297
- and considerations, 1792
- expiration dating period, 981
- interval hypotheses, 292
- of non-inferiority analysis, 24–25
- for pharmacodynamic evaluations, 1667
- statistical model, 424, 1129
 - for assessing dose proportionality, 876–878
 - analysis of variance, 877
 - linear regression, 876–877
 - power model, 877–878
 - of heritability, 99–100
 - multiple-imputation classes, 1129–1130
 - single-imputation class, 1129
- statistical power, 523
 - correlates of, 1777
 - and sample size, 1777
- statistical principles for clinical trials, 2130–2133
 - key statistical principles, 2130
 - regulatory requirements, 2130–2131
- statistical analysis, 2131
 - analysis sets, 2132
 - interim analyses, 2132–2133
 - missing data, 2132
 - outliers, 2132
- statistical design, 2131
 - blinding, 2131
 - randomization, 2131
- statistical principles, good statistics practice (GSP), 1096
 - bias and variability, 1096
 - confounding and interaction, 1096
 - European Statistics Code of Good Practice, 1098–1099
 - one-sided test vs. two-sided test, 1098
 - randomization, 1097
 - sample size determination/justification, 1097
 - statistical difference and scientific difference, 1097–1098
 - type I error, significance level, and power, 1096–1097
- statistical process control (SPC), 2135–2141
 - cause-and-effect diagram, 2136f
 - control charts, 2137–2140
 - histograms, 2135–2136f
 - Pareto charts, 2136f
 - process capability, 2140
 - process flowchart, 2136–2137
- statistical significance, 2142–2145
 - shifted null hypotheses, testing, 2142
 - proof of equivalence, 2144–2145
 - proof of noninferiority, 2144
 - proof of superiority, 2142–2144
 - wider applications, 2145
 - vs. toxicological relevance, 2236
- statistical software for sample size calculation, 1984
- statistical techniques/approaches, process validation, 1805
 - Bergum approach, 1806
 - effect of sample size, 1807
 - statistical basis, 1806
- in bioassay, 213
- confidence interval, 1805
- prediction interval, 1805
- simulation, 1805
- tolerance interval, 1805
- variance components, 1805
- statistical testing procedures, assay validation, 149, 150
 - estimators of SSDL, ADL, and CVDL, 149
 - generalized pivotal quantity approach, 149–150
 - two one-sided tests (TOST) procedures, 149
- statistical tests, 2239
 - ANOVA with rank transformations, 2240
 - for biosimilarity, 278–279
 - Jonckheere–Terpstra test, 2240
 - linear contrasts within an ANOVA, 2240
 - Student's *t*-test, 2239
 - Welch *t*-test, 2239
 - Wilcoxon test, 2240
 - Williams test, 2239–2240
- statistics and good validation practice, 1103
- steepest ascent/descent, 1942–1943f
- steering committee, 812
- Stein's two-stage procedure, 2026
- stereochemical characterization, 256
- stewardship, 811
- stigmatization, ethnic factors, 978
- stochastic curtailment, 1199
 - and Bayesian methods, 1109
- stochastic ordering, 80
- stratification factors, 1401
- stratified analysis, 435
 - combining odds ratios in, 1558
- stratified binomial sampling, 1957–1958
 - estimation in large stratum size, 1958
 - estimation in sparse data, 1958
 - test of homogeneity of risk ratio, 1958–1959
- stratified designs, phase III cancer clinical trials, 394
- stratified Poisson sampling, estimation under, 1959
- stratified randomization, 573, 1221, 1401
- Stroke Prevention in Atrial Fibrillation (SPAF) trial, 1731
- structural alignment methods, 228
- structural average curve, 1547
- structural bioinformatics, 228–229
- structural equation model, 2147–2152
 - applied to bioequivalence, 2149–2150
 - example, 2150–2152
 - model, 2147–2149
- structural error-in-equation model (SEEM), 2152
- Stuart–Maxwell test, 834, 1691, 2154–2158
 - application to clinical trials, 2156
 - clinical study, 2156–2157
 - discussion, 2157
 - algorithm to find zero row and zero column for A matrix, 2157–2158
 - SAS® program, 2155–2156
- Student's *t*-test, 2142, 2239
- study design considerations, diagnostic imaging, 830

- Study on the Genetics of Alcoholism (COGA), 1062
- study population considerations, laboratory analyses, 1263
- subchronic and chronic toxicity studies, 2232–2233
- subdistribution function, *see* [cumulative incidence function](#)
- subgroup analysis, 2159–2165
 - alternative methods of analysis, 2160
 - Kernel density estimation, 2161
 - mixing distributions, 2161
 - laboratory analyses, 1262
 - medical devices, 1352
 - medical literature, examples from, 2163
 - comparison of subgroups within the population, 2163
 - identification of subgroups to investigate, 2163–2164
 - mammography screening, 2164–2165
 - need for, 2159–2160
 - problems with, 2161
 - effect size, 2163
 - finding significance where none exists, 2161–2162
 - insufficient power, 2162–2163
 - sample size, 2163
 - type I (or a) error, 2162
 - type II (or b) error, 2162–2163
- subjective drug effects, 1121
- subject-specific/mixed effects model approach, 652
- subject-treatment (S-T) interaction, 2167–2172
 - completely randomized design comparing two treatments, 2168
 - covariate, role of, 2170
 - individual treatment effects and potential outcomes, 2168
 - model for binary outcomes and consequences of S-T interaction, 2169–2170
 - model for continuous outcomes and consequences of S-T interaction, 2169
 - individual treatment effects in two other designs, 2170
 - matched-pairs design, 2170–2172
 - two-period crossover design, 2172
- subsequent repolarization, 2206, 2207
- substance abuse, 320
- substantial evidence, 726
- sudden cardiac death, 2206
- summary receiver operating characteristic (SROC) curve, 2101–2109
 - area under the curve and Q , 2104
 - numerical tabulations, 2105
 - empirical behavior of, 2103–2104
 - fixed-effect model, issues with, 2106–2107
 - hierarchical (mixed) effects models, 2107
 - bivariate model, 2108
 - Rutter and Gatsonis model, 2107
 - hierarchical models, advantages of, 2108
 - logit-threshold fixed effect model for, 2102–2103
 - meta-analysis of diagnostic test data, 2101–2102
 - standard errors of area under the curve and Q , 2105–2106
- summary statistics meta-analysis (SS-MA), 604
- sum of squares of deviation from linearity (SSDL), 146, 148
- superiority hypotheses, of active control trials, 15
- superiority immunogenicity study, 634
- superiority, noninferiority, and equivalence trials
 - hypotheses in, 1125–1126
- superiority trials
 - hypotheses in, 1125
 - regulatory concerns about the conduct of, 1669
 - sample size for, 2032
 - super, 1125
- surface area, 255
- surfactants (wetting agents), 257
- surrogate endpoint, 574, 2174–2178
 - limitations of using, 2175
 - recent developments, 2177–2178
 - regulatory requirements, 2176
 - scientific issues, 2175
 - statistical methods, 2176–2177
- surrogate markers, 1670, 2174
- surrogates, medical devices, 1351
- survival-adjusted prevalence method, 399, 400
- survival analysis, 842, 1588–1589, 2179–2183
 - comparing survival curves, 2181–2182
 - competing risks, 2182
 - issues in, 2182–2183
 - non-parametric survival models, 2181
 - parametric survival models, 2180–2181
 - randomized clinical trial example, 2179–2180
 - survival functions and hazard rates, 2180
 - using mixture models, 1590–1591
- survival curves, comparing, 2181–2182
- survival function
 - estimation of, 1837–1838
 - inference on, 1232
 - confidence limits and bands, 1233
 - variance estimation, 1232
- survival time, 556, 2179
- survival trials, adaptive, 50–51
- suspected adverse reaction (SUSAR), 814
- suspensions, in dosage forms, 259
- switchability, 973
 - individual BE (IBE) for, 937
- syntax errors, 1089
- synthesis method, 17, *see also* [fraction retention method](#)
- system architecture
 - global database and system, 1070–1071
- system organ class (SOC), 54, 55, 1342, 1344, 815
- system qualification, 1072
- system requirement specification (SRS), 1071
- systems biology, 227–228
- systolic blood pressure (SBP), 1722
- Systolic Hypertension in the Elderly Program, 575
- T**
- τ^2 index, 606
- tablets, in dosage forms, 259
- Tamhane's formula, 1470
- Tang's method, 1474
- Tan-Machin designs, 156
- target allocation functions
 - combining ethics and precision measures, 1730–1731
 - implementing, 1731
- target-based response-adaptive designs, 1951
 - doubly adaptive biased coin design, 1951–1952
 - sequential maximum likelihood procedure, 1951
- target condition, 822
- targeted clinical trials, 2184–2191
 - current approaches, 2185
 - discussion, 2187–2191
 - EM procedures, 2186
 - sample size determination, 2187
 - simulation study, 2186–2189
- Taylor approximation, 945
- Taylor series, 1700
 - approximation, 134
- Taylor's expansion, 342, 489
- technology transfer (TT) of pharmaceutical process, 1102
- temazepam, 1766f
- ternary blend, 1947
- test for ordered categorical data, 1602–1606
 - examples, 1605–1606
 - model and assumptions, 1603
 - likelihood ratio test, 1603–1604
 - sample size, determination of, 1604
 - numerical results, 1604–1605
- testing hypothesis, statistical inference, 1130
- testing individual equivalence ratio (TIER), 1161
- test–retest reliability, 1909, 1914, 824
 - and internal consistency, 1176
- tests of validity, 213
- test statistic
 - clinical trial simulations for later development phases, 547
- test treatment (T), 318
- thawed study, 2111
- therapeutic dose, selection of, 1397–1399
- therapeutic equivalence, 2199–2203
 - active control equivalence trials (ACETs), 2201–2202
 - equivalence limit, 2202
 - odds ratio in therapeutic equivalence clinical trials, 1557
 - one-sided equivalence and non-inferiority trials, 2199–2201
 - recent development, 2203
- therapeutic index (TI), 885
- therapeutic synergism, combination drug clinical trial, 554
- therapeutic trials, meta-analysis of, 1362–1372
 - applications, 1368
 - cumulative meta-analysis vs. expert opinion, 1368
 - drug planning and development, 1370
 - meta-analysis for bioequivalence, 1369
 - meta-analysis for optimal dosing, 1368–1369
 - meta-analysis in cost-effectiveness analysis, 1368

- therapeutic trials, meta-analysis of, (*Cont.*)
 benefits, 1362–1363
 control rate, 1370
 cumulative meta-analysis, 1371
 indirect comparisons, 1371–1372
 large trials and meta-analyses, 1370
 limitations, 1363–1364
 meta-analyses of individual patient data, 1370–1371
 methods, software, education, 1372
 performing and evaluating a meta-analysis, 1367–1368
 statistical analysis for continuous data, 1366
 fixed-effects model, 1366–1367
 random-effects model, 1367
 statistical analysis for discrete data, 1364
 assessing homogeneity of treatment effects, 1365–1366
 fixed-effects and random-effects models, 1364–1366
 meta-regression analysis, 1366
 therapeutic vaccines, 237
 therapeutic window, 885
 therapy-by-center interaction, 1436
therascreen KRAS RGQ PCR kit, 1648, 1651
 thorough QT (TQT) studies, 2206
 assessment of treatment-induced prolongation of QTC interval, 2208
 lessons learned from hundreds of, 2211–2212
 new approach to C-QTC modeling, 2212–2216
 QT interval corrected, by heart rate, 2207
 sample size and design issues of, 2210
 validation test, 2208
 three-arm “gold standard” trials, 736
 three-arm parallel design, 277–278
 three-parameter logistic (3PL) model, 1210
 three-stage designs, 1487
 three-treatment three-period crossover design, 1617
 distribution-free random effects model with using a conditional approach, 1618–1619
 example, 1619
 normal random effects proportional odds model, 1617
 simple change score method, 1617–1618
 tibolone data (case study), 94
 Tier 1 equivalence test, 118
 statistical properties of, 119–120
 Tier 2 quality range approach, 118–119
 Tier 3 raw data and graphical comparison, 119
 time-dependent covariates, 561, 1839
 time-frequency representations (TFR), 1064
 time period, 1724
 time point, 115, 126
 time-to-event CRM (TTE-CRM), 1564
 time-to-event data, 395
 time-to-event endpoint, 155, 1574, 1579
 randomized trial with, 1596–1598
 time-to-market launch, 1721
 TIMI (Thrombolysis in Myocardial Infarction), 966
 titration, 2225–2228
 advantages of, 2226–2227
 clinical rationale, 2226
 disadvantages of, 2226
 titration design, 2227
 analysis of safety data, 2228
 dose response analysis in clinical trials, 882
 types of analysis, 2227
 contingency table approach, 2227
 incomplete design approach, 2227
 life table approach, 2227
 titration with placebo controls, 2228
T method, 503
 tolerance distribution, 214–215
 tolerance interval
 finished product, 1811
 process validation, 1805, 1810
 Torsades de Pointes (TdP), 2206
 total deviation index
 measuring agreement, 1337–1338
 total sleep time (TST), 165
 total sum of square (sst), 105
 total variabilities, comparing, 494
 toxic dose level, 71
 toxicity-based endpoint, 398
 toxicity index (TI), 1753
 toxicity probability intervals (TPIs), 1561
 toxicity upper bounding function, 761
 toxicological studies, 2230–2242
 acute toxicity, 2231
 single-dose toxicity, 2231
 test for skin sensitization, 2231–2232
 testing for acute dermal irritation, 2231
 testing for acute eye irritation, 2231
 carcinogenicity studies, 2233
 genotoxicity studies, 2233
 chromosomal aberration test in vivo, 2234
 DNA indicator tests, 2234
 methods for detection of gene mutations, 2233–2234
 micronucleus test, 2234
 normal range comparisons, 2240
 power of analysis, techniques to improve, 2237
 baseline correction, 2239
 multiple comparison procedures, 2237–2238
 one-sided testing vs. two-sided testing, 2238
 parametric method vs. nonparametric method, 2238
 pooling control groups, 2238
 trend test vs. pairwise comparisons, 2238
 regulatory requirements, 2230
 reproductive toxicity studies, 2235
 safety pharmacology studies, 2232
 statistical analysis, aims and issues in, 2235
 aim of the analysis, 2235
 equivalence tests, 2235–2236
 heterogeneous variances, issue of, 2236–2237
 statistical significance vs. toxicological relevance, 2236
 statistical tests, 2239
 ANOVA with rank transformations, 2240
 Jonckheere–Terpstra test, 2240
 linear contrasts within an ANOVA, 2240
 Student’s *t*-test, 2239
 Welch *t*-test, 2239
 Wilcoxon test, 2240
 Williams test, 2239–2240
 subchronic and chronic toxicity studies, 2232–2233
 two concepts for analyzing chronic/subchronic toxicity studies, 2241
 body weights, organ weights, and food consumption, 2241
 concepts, comparison of, 2241
 first concept, 2241
 second concept, 2241
 traditional Chinese medicine (TCM), 2244–2251, 2252–2264, 2265–2269
 assessing treatment effect through dissimilarity index, 2266–2267
 clinical endpoint, development of, 2247–2248
 clinical trials, 2257
 clinical endpoint, 2258
 matching placebo, 2258
 quantitative instrument, validation of, 2257–2258
 sample size calculation, 2258–2259
 study design, 2257
 consortium for globalization of Chinese medicine, 2262–2263
 discussion, 2269
 dissimilarity index as an efficacy measure, 2265–2266
 matching placebo, preparation of, 2247
 medical practice, 2254–2255
 medical theory and mechanism, 2253–2254
 perspectives of TCM clinical development, 2245
 practical issues, 2249
 animal studies, 2250
 stability analysis, 2250
 test for consistency, 2249–2250
 quality-of-life-like instrument, validation of, 2248
 research and development, 2259
 animal studies, 2262
 indication/label, 2262
 regulatory requirements, 2262
 stability analysis, 2261–2262
 test for consistency, 2259–2261
 sample size determination, 2267
 sample size estimation, power calculation for, 2249
 selection of study design, 2245–2247
 simulation, 2267–2268
 techniques of diagnosis, 2255–2256
 treatment, 2256
 fixed dose vs. flexible dose, 2256
 single active ingredient vs. multiple components, 2256
 validated instrument, calibration of, 2248–2249
 traditional designs, in phase I cancer clinical trials, 390, 391
 traditional dose-escalation design
 early phases of clinical trials, simulation in, 540
 traditional escalation rules (TER), 540
 traditional (frequentist) clinical trial design, 523

- traditional penalized methods, 1448
 - extensions or alternative to Lasso, 1450
 - L_0 penalty, 1448–1449
 - L_1 penalty, 1449
 - Lasso and its extensions, 1449–1450
- traditional randomized clinical trials (RCTs), 1572
- training data set, 827
- trandolapril cardiac evaluation trial (TRACE)
 - study group, 2059
- transcatheter aortic valve replacement, 1791
- transcatheter valve therapy registry, 1791
- transdermal delivery systems (TDSs), 222
- transdermal systems, dissolution of, 262
- translational bioinformatics, 230
- translational medicine, 2270–2281
 - animal model vs. human model, 2277
 - biomarker development, 2271
 - model selection and validation, 2272
 - optimal variable screening, 2271–2272
 - bridging studies, 2279
 - consistency, test for, 2280
 - reproducibility and generalizability, test for, 2280–2281
 - similarity, test for, 2281
 - one-way translation, 2273
 - example, 2275
 - measures of closeness, 2274–2275
 - study endpoints, translation in, 2277
 - example, 2278
 - power analysis and sample size calculation, 2278
 - translation, lost in, 2276
 - two-way translation, 2275
 - example, 2276
 - two-way translational process, validation of, 2275–2276
- Transparent Reporting of a multivariate model for Individual Prognosis Or Diagnosis (TRIPOD), 827
- transplantation, 1667
- Trastuzumab*, 964, 965, 1648
 - adjuvant trial, 965
- treatment by-center interaction, 435
- treatment-by-covariate interaction, 774
- treatment effect, 16f
- treatment effect hazard ratio (HR), 1578
- treatment effect mapping approach, 50
- treatment-emergent adverse events, 814
- treatment-induced QTc prolongation, 2209
- treatment-induced QT prolongation, 2215
- treatment $\times$ period interaction, 206, 207, 207
- tree diagrams, 481–482
- trees
 - for more general outcomes, 1786–1787
 - with soft nodes, 1787
- tree-structured classification and regression, 1786
- trend estimation, 2283–2288
 - association between breast cancer and alcohol use, 2286
 - inference in a single study, 2286–2287
 - meta-analysis and trend estimation, 2287
 - combining dose–response data, 2286
 - meta-analysis for trend estimation, 2286
 - with grouped data, 2285–2286
- hypothesis testing and confidence interval, 2283
 - asymptotic approach, 2283–2284
 - conditional exact approach, 2284–2285
 - meta-analysis for, 2286
 - by using an empirical log-odds ratio, 2285
- trend test, 79
 - vs. pairwise comparisons, 2238
- trial simulations, 32
- triazolam, 1766f
- trichloropropene oxide, 995
- triple blind trials, 337, 573
- true class rates (TCR), 1608, 1609, 1609
- “true” endpoints, 574
- true negative, 833
- true-negative fraction, 824
- “true” overall treatment effect, 1578
- true positive, 833
- true-positive fraction, 824, 826
- true positive rate (TPR), 2101
- “true” regional treatment effect, 1578
- TRW, Inc., 55
- tryptophan, 68, 69
- t -test, 210, 317, 419, 436, 650, 651, 912, 2187, 2239
- t -type confidence sets, 344–345
- tuberculosis (TB), 604
- Tucker Lewis index (TLI), 1173
- Tukey and Tukey–Kramer procedures, 731
- Tukey’s honestly significant difference (HSD) test, 1689
- tumor incidence, tests for trend and difference in
 - using Peto tests, 399–405
 - using poly- k tests, 405–409
- tuning and feature selection issues, 1452
- T wave, 2206, 2207
- two-agent dose–response model, 188–189
- two-arm parallel design, equivalence trials based on, 274–275
- $2 \times K$ contingency table, 76
- $2 \times K$ nominal tables, 75–76
 - exact tests, 77
 - example, 77
 - likelihood ratio chi-square, 76
 - logistic regression, 76, 78
 - maximal correlation and Pearson’s chi-square, 76
 - Pearson’s chi-square test, 75
 - regression and Pearson’s chi-square, 77
- $2 \times K$ ordered tables, 78–80
 - choice of scores, 79
 - example, 80
 - Iso-chi-square test, 79
 - methods without scores, 79
 - methods with scores, 79
 - stochastic ordering, 80
- two confidence interval (TCI) method, 1536
- two one-sided tests (TOST) procedure, 149, 217, 854, 936, 2039, 2220
- two-period crossover design, 2172
- two-phase regression analysis, 2112
- two-phase shelf-life estimation
 - expiration dating period, 982
- 2PL model, 1211–1212
- two rank-based tests, 70
- two-sample t -test
 - pharmacodynamics with no covariates, 1685–1686
- two-sample Wilcoxon’s test, 1975
 - example, 1975–1977
 - fixed alternative, 1975
 - local alternative, 1975
 - simulation study, 1975–1976
- two-sequence dual design, 926
- two-sided equivalence, 971
- two-sided specification limits, 5f
- two-stage adaptive design, 1577
 - flexibility and feasibility of, 586
- two-stage design (phase II cancer clinical trials), 2296–2304
 - inference on RR, 2298
 - confidence interval, 2300–2301
 - estimation of RR, 2298–2300
 - optimal two-stage designs, 2297–2298
 - p -value calculation, 2301–2304
- two-stage escalation algorithm (TSER)
 - early phases of clinical trials, simulation in, 540–541
- two-stage random sampling approach, 1819
- two-stage sampling plan, clinical inspection, 467–468
- two-stage seamless adaptive design, 2289–2294
 - analysis of, 2290
 - Category I designs, analysis of, 2290–2291
 - Category II designs, analysis of, 2291–2292
 - Category III and IV designs, analysis for, 2292–2293
 - types, 2290
 - sample size calculation for, 2293–2294
 - vs. traditional approach, 2289–2290
- two-stage study design process, 1793
- two-stage test procedure, individual bioequivalence, 1165
- two-step sample size determination strategy, 967
- two-trials rule, 1850–1851
- two-way ANOVA model, 107, 1673–1676
 - validity of LOCF ANOVA test under, 1298
 - with interaction, 1299
 - without interaction, 1298–1299
- two-way enriched design (TED), 529
- type I error, 59
- inflation, 1573
- type I error rate, 727, 1538
 - conditional across-trial type I error rate, 1540–1541
 - inflation of, 728, 2022
 - unconditional across-trial type I error rate, 1539–1540
 - within-trial type I error rate, 1538–1539
- type I error spending/use function approach, 1106–1108
- type II error, 60
 - rate, 727
- type I sums of squares, 126, 127, 128
- typical phase III program, 1725
- U
- uncertainties, general methods to deal with, 480–481
- unconditional across-trial type I error rate, 1539–1540

unconditional distribution, 524
 uncorrelated data, probit model for, 741–742
 unidimensionality, 1209
 uniform error spending function, 61
 uniformly minimum variance unbiased estimator (UMVUE), 1487, 2298
 uniformly most powerful (UMP) test, 1115, 1124–1125
 uninformative prior distribution, 176
 union-intersection (UI) and intersection-union (IU) testing principles, 732
 union non-inferiority hypotheses, 10–11
 United Network for Organ Sharing (UNOS) registry data, 1791
 United States Environmental Protection Agency
 good laboratory practice (GLP), 1084
 United States Federal Food, Drug, and Cosmetic Act (FDCA), 928
 United States Food and Drug Administration (FDA), 1202, 1792, 928
 good laboratory practice (GLP), 1084
 United States Pharmacopeia and National Formulary (USP/NF), 709, 933, 1095
 United States Pharmacopeia (USP) tests, 255, 1436, 2311–2312
 univariate analysis, 1311
 univariate interval-censored data, statistical analysis of, 591–592
 univariate kernel density estimation, mathematica program for, 1775
 univariate Poisson binomial distribution, 317
 unscaled agreement indices, 164
 unscaled average bioequivalence, 2038–2039
 unstratified designs, phase III cancer clinical trials, 394
 unstructured (UN) covariance structure, 1425
 unstructured variance-covariance, 2213
 unweighted average is used (UWLE), 905
 unweighted estimator, 436
 unweighted inferences, 436
 upper control limit (UCL), 2137
 urn models, 1950
 -based allocations, 1729
 drop-the-loser rule, 1950–1951
 play-the-winner rule, 1950
 randomized play-the-winner rule, 1950
 use function, *see* **alpha spending function**
 user acceptance testing, 1070, 1072
 USP Dissolution Test Acceptance Criteria, 262
 uterine contractions, 822
 utility-based trial objective
 clinical trial simulations for later development phases, 545
 utility-offset model (UOM), 545
 clinical trial simulations for later development phases, 547

V

Vaccine Adverse Event Reporting System (VAERS), 638
 vaccine efficacy, 237, 629
 based on ratio of two binomial proportions, 630
 based on ratio of two rates, 630–631
 estimates, 631

sample size and power calculation for
 vaccine efficacy trials, 632
 trials to estimate relative, 631
 vaccine immunogenicity analysis
 different types of, 634
 bridging study, 635
 combination/multivalent vaccine study, 635
 consistency lots study, 634–635
 dose response immunogenicity study, 634
 immunologic persistence study, 635
 superiority immunogenicity study, 634
 general approaches to, 632
 geometric mean titers, approaches for analyzing, 633–634
 immune response rates, approaches for analyzing, 633
 immunogenicity endpoints, 632
 noninferiority/equivalence margins, considerations for, 634
 sample size and power considerations for, 636
 bridging study, 636
 combination/multivalent vaccine study, 637
 consistency lot study, 636
 vaccine licensure, regulatory process for, 629
 vaccines, 234
 economic evaluation of, 640–641
 immune surrogate and correlate of protection for, 640
 issues in, 237–240
 vaccine safety, assessing, 637–639
 vaccine trials, additional statistical issues for analyzing, 639
 analyses with stratification variables, 639
 intention-to-treat or per-protocol analysis, 639
 interim analyses, 639
 missing/coarse data and loss to follow-up, 639
 Vaginal Infections and Prematurity (VIP) Study, 1628
 vague conjugate priors, 171
 validation, defined, 1101, 1802
 validation test, 1122, 1123
 validity, 1174
 of bootstrap, 343, 345, 346
 concurrent, 1175
 construct, 1175–1176
 Valtropin®, 309
 value tree, 481
 variable acceptance sampling plans, 1
 variance-balanced design, 2210
 variance-component method/model, 1059, 1061
 variance components
 extension of MLS method to difference of, 1428–1430
 extension of MLS method to ratio of, 1430
 process validation, 1805
 variance-covariance matrix, 997, 1316, 2148
 variance estimation, 344, 1232
 statistical inference, 1130
 variance inflation factors (VIFs), 648, 1444, 1448

variation, comparing intrasubject coefficients of, 489
 Vascular Quality Initiative, 1791
 Venclexta® (venetoclax), 1648
 ventricular depolarization, 2206, 2207
 ventricular fibrillation, 2206
 ventricular myocytes, 2206
 ventricular premature contractions (VPC), 961
 Veterans Administration lung cancer trial, 559
 Veterans' Health Administration health performance monitoring study, 1414
 visual analogue scales (VAS), 1121
 visualization, 481
 bar graphs, 482
 effects table, 481
 forest plots, 482
 tree diagrams, 481–482
 volume under the ROC surface (VUS), 1608
 Vysis CII Fish Probe kit, 1648

W

Wald statistic, 1683, 2284
 Wald test, 955, 2325, 2033
 Wald-type statistics, generalized estimating equation, 1049
 Wald-type test procedures, 1411
 Waller-Duncan test, 1689
 Wang-Tsatis delta family group sequential method, 62
 washout period, clinical pharmacology, 470
 wavelet estimators, 1545–1546
 weak identifiability, 1269
 web-based electronic data capture system
 clinical data management guidelines, 446
 Weibull curves, 1152
 Weibull distribution, 556, 2180f, 1330
 Weibull model, 558, 1590, 2325–2328
 group sequential procedure, 2327
 sample size for fixed sample test, 2326–2327
 sequential test statistics, 2325–2326
 weighted analyses, 439
 weighted empirical odds functions, 560
 weighted estimator, 436
 weighted inferences, 436
 weighted kappa, 1239
 in the context of measures of 2×2 association, 1240–1242
 weighted least-squares (WLS) method, 372, 1958
 weighted log-rank statistics, 559
 weighted regression, 134
 weighting schemes, 437
 weight selection, calibration, 372
 Wei-Lin-Westfield (WLW) model, 396
 Welch's correction, 1116, 1118
 Welch *t*-test, 2238, 2239
 Westernization of TCM, 2252
 Western medicines (WM), 2252, 2253, 2256
 wetting agents, 257
 WhatIf program, 229
 Whoart, 54
 wholly stochastic analysis, 1699
 widest horizontal distance (WHD), 1142
 Wilcoxon-Mann-Whitney rank sum test, 296, 317

- Wilcoxon–Mann–Whitney test, [296](#), [781](#)
 - Wilcoxon–Mann–Whitney two one-sided tests, [295–297](#)
 - Wilcoxon–Mann–Whitney-type statistic, [1729](#)
 - Wilcoxon rank sum statistic, [2143](#)
 - Wilcoxon rank sum test, [295–296](#), *see also* [Mann–Whitney test](#)
 - pharmacodynamics with no covariates, [1687](#)
 - Wilcoxon scores, [793](#)
 - Wilcoxon-signed rank test, [1122](#)
 - pharmacodynamics with no covariates, [1687–1688](#)
 - Wilcoxon test, [561](#), [1973](#), [1589](#), [1590](#), [2142](#), [2240](#)
 - Wilks Theorem, [319](#)
 - Williams square, [1123](#), [2210](#), [2210](#), [2214](#)
 - William’s test, [2239–2240](#)
 - for determining drug effect, [879](#)
 - “willingness-to-pay” method, [1695](#)
 - Wilson’s score interval, [700](#)
 - WinBUGS software, [180](#), [183](#), [176](#), [2094](#)
 - Wishart distribution, [203](#)
 - within-treatment-group analysis
 - pharmacodynamics with covariates, [1678–1679](#)
 - within-trial type I error rate, [1538–1539](#)
 - without-replacement bootstrap, [344](#)
 - working correlation, [1047](#)
 - efficiency, [1048](#)
 - estimation of correlation, [1048–1049](#)
 - work processes, clinical trial process
 - continuous improvement in, [534](#)
 - monitoring, [534](#)
 - understanding, [532](#)
 - World Health Organization (WHO), [282](#), [1085](#)
 - characterization, [284](#)
 - history of good clinical practice implementation by, [1078](#)
 - key principles and basic concept, [283](#)
 - manufacturing process, [284](#)
 - nonclinical and clinical studies, [284](#)
 - quality, [284](#)
 - reference biotherapeutic product, [283–285](#)
 - worst case carried forward, [418](#)
 - worst observation carry forward (WOOF), [1422](#)
 - worst outcome criterion, [195](#)
 - WP2uvrA (*Escherichia coli* tester strain), [68](#)
- Z**
- zero-inflated Poisson (ZIP) model, [316](#), [317](#)
 - determination of sample size, [319](#)
 - development of, [316](#)
 - example, [320](#)
 - likelihood ratio tests for comparison of treatments, [319](#)
 - zero-inflation, concept of, [316](#)
 - zero-mixing probability, [316](#)
 - z-score, [2329–2331](#)
 - applications, [2331](#)
 - definition of, [2330–2331](#)
 - Gaussian distribution, [2329](#)
 - normal distribution, [2330f](#)
 - standard normal distribution, [2329–2330](#)
 - Z-test, [1131](#), [1469](#)
 - sample size for, [499](#)
 - Zurich growth data, [1549](#)
 - Z-value, [1116](#)
 - Zyprexa®, *see* [olanzapine](#)